

Cross-Lingual Ontology Matching and Enrichment

Kumulative Dissertation
zur Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Shimaa Khaled Shaker Ibrahim
aus
Alexandria, Ägypten

Bonn, 2024

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen
Friedrich-Wilhelms-Universität Bonn

1. Gutachter / Betreuer:	Prof. Dr. Jens Lehmann
2. Gutachter:	Prof. Dr. Stefan Wrobel
Tag der Promotion:	08.01.2025
Erscheinungsjahr:	2025

Abstract

The Semantic Web has emerged as a foundational paradigm for representing and interlinking data in a semantically meaningful manner. Central to this paradigm are ontologies, pivotal for encapsulating spatial knowledge and fostering a standardized comprehension of information. However, the multilingual nature of global communication presents challenges to the seamless interoperability and effective utilization of ontologies, such as semantic ambiguity across languages and the limited availability of bilingual dictionaries.

This thesis aims to tackle these challenges by developing innovative approaches to utilizing multilingual data in order to enrich ontologies across different languages. The ultimate goal is to facilitate semantic interoperability and the exchange of knowledge across ontologies in a multilingual context. Key steps towards this goal have been taken by proposing approaches for cross-lingual ontology matching and enrichment, accompanied by the development of a comprehensive metrics suite to assess the quality of the enrichment process.

In order to achieve this goal, a cross-lingual ontology matching approach is introduced to align ontologies cohesively before the enrichment process. This approach provides a robust framework for evaluating various similarity measures across languages, empowering researchers to select the most suitable measures for their endeavors.

Cross-lingual ontology enrichment is identified as a cornerstone, enabling the representation of domain-specific knowledge written in different languages through enriching ontologies with additional linguistic and semantic content. We proposed approaches for enriching ontologies by leveraging information from unstructured multilingual text corpora and diverse ontological resources, producing multilingual ontologies. These approaches promote the efficient integration and sharing of knowledge across systems that operate in multilingual contexts.

One of the main obstacles to improving the quality of the enrichment process is ensuring the accuracy and relevance of additional information. Therefore, a thorough analysis of factors influencing the enrichment process has been conducted, resulting in the innovation of a comprehensive quality metrics suite.

Through empirical evaluations and real-world use cases, the efficacy and applicability of the proposed approaches are demonstrated, highlighting their potential to facilitate seamless cross-lingual communication, enhance knowledge sharing, promote semantic interoperability, and promote cultural diversity in the Semantic Web.

Acknowledgments

I would not have been able to complete this work without the support and guidance of many individuals throughout my Ph.D. journey. I want to thank the Ministry of Higher Education of Egypt for providing me with a scholarship to conduct my Ph.D.

My deepest gratitude goes to my supervisor, Prof. Dr. Jens Lehmann, for giving me the opportunity to join the Smart Data Analytics (SDA) research group. His insightful guidance and unwavering support have been instrumental throughout my Ph.D. studies. I am also immensely thankful to my mentor, Dr. Hajira Jabeen, for her support during this thesis. Her constant guidance helped me explore new ideas when tackling the research problem. Additionally, I thank my second supervisor, Prof. Dr. Stefan Wrobel, for his assistance during the transition phase in the SDA and for providing a conducive environment for furthering my research activities.

I extend my heartfelt thanks to my friends, especially Heba Mohamed, who has been like a sister to me since we became friends in 2002. I am very grateful for her presence in my life, especially during my Ph.D. I would also like to thank Dr. Said Fathallah for being a constant co-author and for always offering valuable insights that improved our work.

To my dear family, I am profoundly grateful for your unwavering support and belief in me. Thank you for always being there for me and loving me unconditionally. To my mother, thank you for your love and care throughout my life. Your belief in my potential has always motivated me to thrive in my studies. To my beloved father, whom I lost during my doctoral journey, thank you for being my backbone. To my lovely sister, thank you for always standing by my side and being my strongest supporter.

I dedicate this thesis to my dear late father.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Problem Definition and Challenges	3
1.3	Research Questions	5
1.4	Thesis Overview	6
1.4.1	Contributions	6
1.4.2	List of Publications	8
1.4.3	Thesis Structure	9
2	Preliminaries	11
2.1	Semantic Web Technologies	11
2.1.1	Resource Description Framework	12
2.1.2	Ontologies	14
2.2	Ontology Matching	15
2.3	Similarity Measures	16
2.3.1	Syntactic Similarity	16
2.3.2	Semantic Similarity	16
2.4	Ontology Enrichment	17
2.5	Dataset	18
3	Multilingual Ontology Matching	19
4	Ontology Enrichment from Multilingual Unstructured Sources	23
5	Ontology Enrichment from Multilingual Structured Sources	27
6	Multilingual Ontology Merging	31
7	Quality Assessment of Ontology Enrichment	35
8	Conclusion and Future Work	39
8.1	Conclusion	39
8.2	Future Work	41
	Bibliography	43

A	Toward the multilingual Semantic Web: Multilingual ontology matching and assessment	51
B	Cross-lingual ontology enrichment based on multi-agent architecture	71
C	CLOE: a cross-lingual ontology enrichment using multi-agent architecture	85
D	OECM: A Cross-Lingual Approach for Ontology Enrichment	107
E	From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment	113
F	Multilingual ontology merging using cross-lingual matching	131
	List of Figures	141
	List of Tables	143
	Acronyms	145

Introduction

One of the biggest innovations of the digital era is the rise of the Semantic Web. Tim Berners-Lee presented this vision to transform the World Wide Web into a platform that is more intelligent and machine-readable [1]. The Semantic Web was developed to give computers the ability to understand and interpret the meaning of data and information available on the Internet. A shared vocabulary that can be used to define concepts, relationships, and instances within a particular domain is provided by the concept of ontology, which forms the foundation for the Semantic Web. Ontologies are used by machines to help comprehend the context of data, which makes it simpler for machines to get involved in automatic reasoning and data processing. This has substantial implications for a wide variety of applications in various fields of science, such as Biology [2], Engineering [3], and Medicine [4].

1.1 Motivation

Due to the fast growth of multilingual data on the Semantic Web, many ontologies accessible in various natural languages have been developed. Based on [Linked Open Vocabularies \(LOV\)](https://lov.linkeddata.es/)¹, English stands out as the dominant language in the Semantic Web, with a majority of ontologies being in English. Nonetheless, there are plenty of ontologies in other Indo-European languages, such as German and French. With a total of 851 vocabularies found in [LOV](https://lov.linkeddata.es/), 652 are in English, 68 in French, and 45 in German (see Figure 1.1). There are a few ontologies listed in [LOV](https://lov.linkeddata.es/) that are in non-Indo-European languages. For example, ten are in Chinese, and seven are in Arabic.

Monolingual ontologies with labels or local names in a specific natural language pose significant challenges to supporting semantic interoperability and knowledge exchange across linguistic barriers. Cross-lingual ontology enrichment offers a promising solution to overcome these challenges by enriching ontologies with cross-lingual information, which facilitates seamless communication and collaboration across linguistic boundaries [5, 6]. It fosters data exchange between systems operating in multilingual environments. Initiating this process requires establishing mappings between concepts and relationships in various languages from different resources (known as cross-lingual matching), resulting in enriching the target ontology with additional concepts and semantic relationships from external resources in various languages [7].

¹ <https://lov.linkeddata.es/dataset/lov/vocabs> - accessed on 21.06.2024

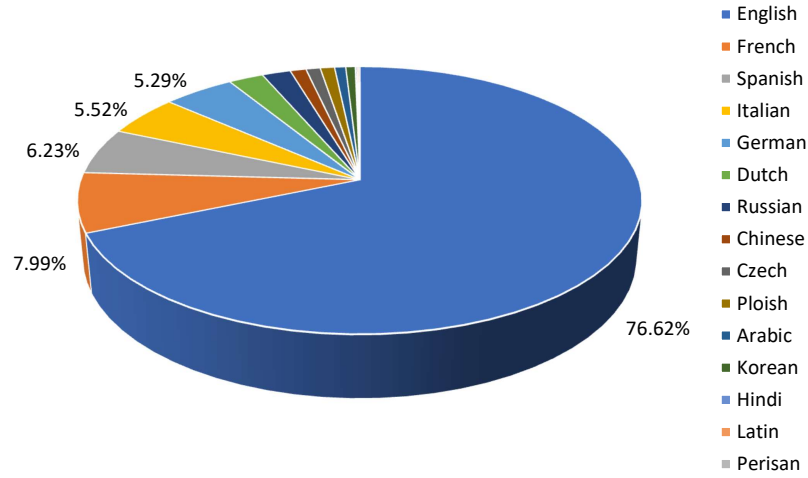


Figure 1.1: Language distribution of ontologies according to [LOV](#).

Cross-lingual ontology enrichment improves information retrieval, knowledge sharing, and collaboration across diverse linguistic and cultural contexts. It accomplishes one of the design goals for [Web Ontology Language \(OWL\)](#), i.e., internationalization, ensuring that an ontology can be utilized across different languages and cultures (also known as a multilingual ontology) [8]. In fact, multilingual ontology enables individuals and systems to understand and exchange information, regardless of language barriers. For instance, during the COVID-19 pandemic, there were demands for integrating data at the international level [9]. When merging various data sources and considering the multitude of languages spoken worldwide, multilingual ontologies are crucial in this scenario. Furthermore, cross-lingual ontology enrichment is indispensable for the advancement of [Natural Language Processing \(NLP\)](#) technologies, particularly in tasks such as machine translation and cross-lingual information retrieval [10, 11]. The resulting enriched ontologies facilitate the translation of words and phrases into their semantic equivalents and the retrieval of information in various languages. Such translation is achieved by frameworks that map linguistic expressions to their semantic counterparts. Moreover, cross-lingual ontology enrichment helps in safeguarding and spreading cultural heritage and linguistic diversity [12]. It is essential for preserving and promoting cultural identities in an increasingly interconnected world by enabling the representation and sharing of cultural knowledge and traditions across different languages.

As mentioned before, cross-lingual ontology matching is the first step in enrichment. Therefore, there is a need to develop new algorithms that can handle data in multiple languages and align ontologies in different natural languages [13]. There has been a significant amount of study conducted on monolingual ontology matching; nevertheless, there have been relatively few researchers who have shed light on cross-lingual ontology matching [13–15]. Several approaches have been created to tackle this problem, each providing distinct methodologies and techniques. For instance, Abu Helou and Palmonari [16] introduced a technique that uses cross-lingual lexical matching to align lexically-rich resources, such as WordNet, which uses translation tools including Google Translate and BabelNet. Similarly, Musyaffa et al. [17] introduced a framework incorporating machine translation and similarity measures to interlink heterogeneous multilingual open fiscal data. Fu et al. [18] contributed algorithms considering various factors such as semantics and user feedback for matching ontologies across

languages, whereas Hertling and Paulheim [19] explored the use of Wikipedia’s interlanguage links for ontology alignment. Trojahn et al. [20] presented a multi-agent architecture that uses translation agents and lexical databases to match ontologies across languages. Furthermore, Khiat et al. [21, 22] proposed SimCat and Crolom, employing translation services and semantic resources for lexical matching.

Despite continuous attempts to develop different methods, no definitive solution has been proposed for tackling the cross-lingual matching problem. Most of the previously mentioned approaches depend solely on one-to-one translation between terms across different natural languages, which fails to capture cultural variations and contextual nuances, ultimately leading to inaccurate mappings. Additionally, there is a notable absence of standardized benchmarks and metrics to validate the effectiveness and reliability of these techniques. Further investigation is needed to improve cross-language ontology matching techniques in order to achieve more accurate matching results as well as improve enrichment outcomes. According to the aforementioned studies, machine translation tools and pre-processing techniques have a substantial impact on the quality of the matching output [15]. Consequently, we shall investigate the optimal way to leverage the integration of machine translation tools, pre-processing techniques, and similarity measures in order to achieve superior outcomes in comparison to the current state-of-the-art approaches.

1.2 Problem Definition and Challenges

This thesis is driven by a research problem encapsulated in the question: *How can ontologies be efficiently enriched through the utilization of existing multilingual data?* Throughout the journey to reach the goal, problems related to cross-lingual ontology matching, cross-lingual ontology enrichment, and quality assessment for the enrichment process have been confronted. Successful cross-lingual ontology matching demands careful attention and creative solutions to align different ontological structures across language barriers. Concurrently, enriching ontologies in a cross-lingual context needs flexible techniques that can efficiently deal with the diversity and semantic heterogeneity of multilingual sources and produce a consistent, rich ontology. In addition, pursuing quality assessment in enriched ontologies highlights the importance of robust quality metrics to assess different quality dimensions in the enrichment process, such as accuracy, completeness, coherence, and consistency.

In the following, we present the main challenges encountered by this thesis:

Challenge 1: Identifying mappings between similar concepts and relationships from different ontologies in different natural languages.

Generating cross-lingual ontology mappings between ontologies in different natural languages poses various challenges due to the complex nature of the lexical information in ontologies [13, 14, 23]. Lexical information primarily derives from natural language data, such as words and short sentences, found in the entities and comments of the ontology [23]. These challenges can be summarized as:

1. *Semantic ambiguity*, where an expression has multiple meanings across languages, e.g., the word “angel” is used in both English and Dutch, but in Dutch, it can also mean “sting” [24].
2. *Synonymy*, where multiple equivalent terms for the same concept can be found, e.g., the concept “president” has many synonyms such as “chair”, “chairman”, “director”, “head”, and “manager” [25].

3. *Bilingual resources availability*, where aligned parallel corpora or bilingual dictionaries, such as BabelNet², are necessary for accurately generating cross-lingual mappings [26]. However, such resources may be limited or unavailable for certain language pairs,
4. *Senses combination*, where concept labels consist of many words. For example, consider the concept pair to be matched ⟨Conference, Conference Chair⟩. Such two concepts are different and may cause erroneous correspondences [23], and
5. *Senses fuzziness*, where senses are not correlated but express the same concepts, such as the two relations `publishedBy` and `hasPublisher` [23]. In this example, “published” and “publisher” are not the same, but they express the same role of publisher.

Therefore, a holistic, robust cross-lingual ontology matching approach that uses linguistic analysis techniques to overcome these obstacles is required.

Challenge 2: Preserving consistency while enriching the ontology.

Enriching an ontology with new information, such as classes and relationships, results in a change in the hierarchy of the ontology, i.e., changing the structure of the ontology [27]. This change includes adding, updating, deleting, and moving classes and relationships. Preserving consistency while performing these changes to enrich the ontology is a challenging task. It is essential to find an appropriate position in the ontology hierarchy to add new information, thereby producing a consistent, enriched ontology.

Another challenge is the diversity and semantic heterogeneity of multilingual sources that are utilized for enrichment. These multilingual sources, which include structured and unstructured data (e.g., text corpora), require robust methodologies for extracting, processing, matching, and enriching information to enhance semantic interoperability and effectively resolve conflicts that may arise after enrichment [28].

Dealing with these challenges requires new methods for cross-lingual enrichment to foster robust and coherent knowledge representations that can be understood across different languages.

Challenge 3: Assessing the quality of the enrichment process.

One of the main challenges to improving the quality of ontology enrichment is ensuring the precision and relevance of the additional information. Ontologies are specifically created to semantically represent domain knowledge, therefore, any contradictions that occur throughout the process of adding additional information could affect their reliability. The quality of the input ontologies and the matching procedure between these ontologies significantly impact the quality of the enrichment. For instance, in terms of richness, the more knowledge the input ontologies provide, the richer the enriched ontology is. Most of the research efforts have focused on evaluating the quality of ontologies from multiple dimensions, such as syntactic, semantic, structural, functional, content, and schema [29–32]. Every dimension has its own set of metrics to quantitatively assess the characteristics of an ontology [33]. Few researchers have addressed the quality of the matching process [34]. Therefore, it is challenging to identify the quality measures that assess the ontology enrichment process.

² <https://babelnet.org/>

1.3 Research Questions

Based on the motivation and identified challenges, we define the main research question in this thesis as follows:

Is it possible to enrich the existing ontologies using information from external sources in various natural languages?

This research question is split into three specific research questions. Each challenge is mapped to a research question and altogether contributes to the overall research problem definition tackled throughout this thesis.

RQ1: Do string similarity measures affect the generation of cross-lingual ontology mappings between two ontologies in different natural languages?

In order to answer this question, we break it down into two sub-questions: 1) What is the most effective string similarity measure(s) for matching multilingual terms across different ontologies? and 2) How strong is the performance of the similarity measure when the language of the input ontologies is changed? To address these questions, we investigate the use of different string similarity techniques, proposed in [35], to match ontology's entities. These techniques, known as string-based techniques, consider strings as alphabetical sequences of letters. We apply the similarity measures to different ontologies in Indo and non-Indo European languages and analyze the matching results to determine the most effective similarity measures for this task. We calculate the correlation between language pairs in order to demonstrate the relationship between them and the similarity measures. The findings from the investigation of research question RQ1 enable us to tackle Challenge 1.

RQ2: How can we enrich an ontology using knowledge from external sources in different natural languages to produce a rich, consistent ontology?

In order to address this question, we explore the utilization of various data sources, both structured and unstructured, to enrich the ontologies and produce a consistent, rich ontology. We divide this question into three sub-questions: 1) What would be a suitable approach to enriching ontologies from unstructured text in a different language? 2) How can an ontology be enriched using another ontology in a different natural language? and 3) Can cross-lingual ontology merging be a more informative method for enrichment? To address these questions, we investigate the use of multilingual text corpora as a source for enriching ontologies. Extracting concepts and relations from multilingual text poses a significant challenge due to the variations in pre-processing techniques across various languages. In addition, we explore the use of existing ontologies in various natural languages as a source for enrichment. To overcome problems arising from linguistic ambiguity and synonymy caused by language diversity, we use semantic similarity. We study how to make the enriched ontology more rich and informative by merging two ontologies in different natural languages. The findings from these investigations enable us to tackle Challenge 2.

RQ3: How can we assess the quality of the ontology enrichment process?

In order to answer this question, we analyze the factors that impact the enrichment process. These factors encompass the quality of the input ontology, the quality of the matching process, and the quality of the produced enriched ontology. Based on the quality metrics defined in [29–31], we introduce a comprehensive quality metrics suite that assesses the factors influencing the quality of the enrichment process. The findings from the investigation of research question RQ3 enable us to tackle Challenge 3.

1.4 Thesis Overview

This section provides an overview of the primary contributions made during this thesis (see Figure 1.2), the research areas investigated, our scientific publications supporting this work, and the thesis structure.

1.4.1 Contributions

Contribution 1: Designing a flexible framework for ontology matching across languages

In order to answer **RQ1**, we propose a [Multilingual Ontology Matching \(MoMatch\)](#) approach for matching ontologies in multiple natural languages. It uses several string-matching techniques to detect similar or identical classes and properties across ontologies in different natural languages. We provide an in-depth analysis of several similarity measures across various languages, allowing researchers to choose the best similarity measures for their specific purpose. We present three use cases describing the usability of [MoMatch](#) in different domains. More detailed information on this contribution is given in Chapter 3 and the publication [36].

Contribution 2: Enriching ontologies from unstructured multilingual sources

In order to answer **RQ2**, we propose a [Cross-Lingual Ontology Enrichment \(CLOE\)](#) approach, a domain-independent approach for enriching ontologies from a multilingual text based on a multi-agent architecture. In this architecture, each agent is responsible for enriching the ontology of a particular domain in a specific natural language. We develop a novel communication algorithm that enables agents to learn from each other about new classes or properties. The main shortcoming of this approach is that it produces monolingual ontologies rather than multilingual ones, which lose naturalization, one of the most important design goals for [OWL](#) [8]. Furthermore, [CLOE](#), a semi-automated approach, requires ontology experts to validate candidate classes and properties prior to their incorporation into the ontology. Such validation is a resource-demanding and time-consuming task for heavy-weight ontologies. Chapter 4 and publications [37, 38] provide more detailed information on this contribution. The [CLOE](#) approach contributes to answering the research question **RQ2**.

Contribution 3: Enriching ontologies from structured multilingual sources and producing consistent multilingual ontologies

To continue answering **RQ2**, we propose a fully automated approach known as [Ontology Enrichment using Cross-lingual Matching \(OECM\)](#). [OECM](#) enriches a monolingual ontology by incorporating

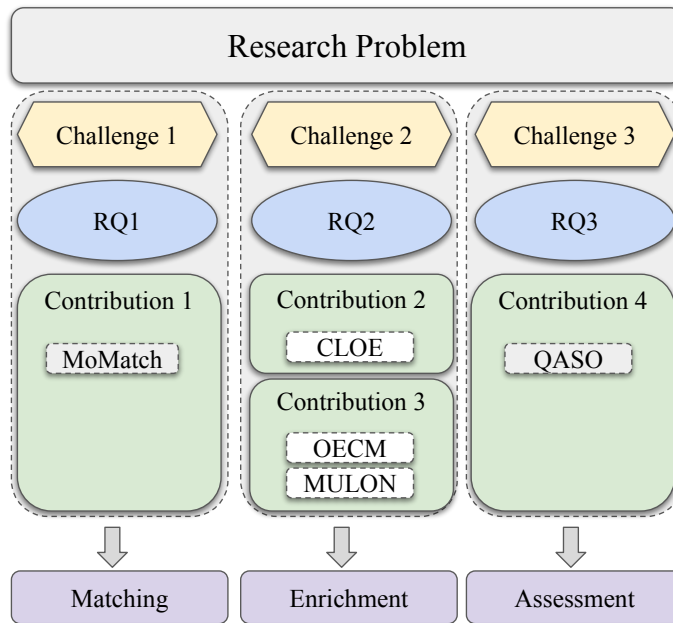


Figure 1.2: The distribution of thesis contributions among the defined research questions.

information from another ontology in a different natural language, utilizing a cross-lingual matching technique. The objective of this method is to identify additional information (classes) from the source ontology in order to enrich the target ontology. **OECM** chooses the best translation from a pool of existing translations for the ontology’s classes (known as one-to-many translation), depending on their degree of semantic similarity with the target ontology’s classes. **OECM** involves an iterative triple retrieval procedure to determine which information from the source ontology can be used to enrich the target ontology and where such information should be placed in the ontology hierarchy. The output is a consistent, multilingual, and enriched ontology in two different natural languages. **OECM** focuses exclusively on classes throughout the enrichment process. As a result, we introduce a more informative technique for enrichment known as **MULTilingual Ontology mergiNg (MULON)**. Using a cross-lingual matching technique, **MULON** merges monolingual ontologies in different natural languages, resulting in a multilingual merged ontology. We develop a merging approach instead of an enrichment one because some ontologies independently model pre-existing domains, resulting in ontologies that contain redundant and overlapping information. Hence, we develop a merging approach to exploit and reconcile such information, improving the richness of ontologies. In the merging process, **MULON** does not consider only classes but also properties. **MULON** shows its usability by presenting three use cases in various domains, including fiscal data, food, and scholarly communication. We use both Indo and non-Indo-European language ontologies in the merging process to demonstrate **MULON**’s robustness. More details about **OECM** and **MULON** can be found in Chapters 5 and 6, respectively, and in publications [39–41]. This contribution covers **RQ2**.

Contribution 4: A metrics suite for assessing the quality of the enrichment process

In order to answer **RQ3**, we analyze the characteristics that affect the enrichment process. According to the findings of this analysis, a metrics suite, i.e., [Quality Assessment Suite for Ontologies \(QASO\)](#), has been proposed. This suite consists of 17 quality metrics that provide a comprehensive evaluation of different factors that impact the enrichment process. Out of these 17 metrics, seven are used to evaluate the quality of ontologies at the schema level, seven to evaluate the quality of the matching process, and three to evaluate the enriched/merged ontology. More details about this contribution is presented in Chapter 7 and in publications [36, 41].

1.4.2 List of Publications

This thesis builds upon the following publications:

- *Journal Articles (peer reviewed)*

1. **Shimaa Ibrahim**, Said Fathalla, Jens Lehmann, and Hajira Jabeen. *Toward the multilingual Semantic Web: Multilingual ontology matching and assessment*. IEEE Access, vol. 11, pp. 8581-8599, 2023. DOI: <https://doi.org/10.1109/ACCESS.2023.3238871>

Shimaa Ibrahim developed the idea of multilingual ontology matching approach that uses different string similarity measures. She implemented all parts of the source code. She wrote the manuscript. All authors revised the manuscript.

2. Mohamed Ali, Said Fathalla, **Shimaa Ibrahim**, Mohamed Kholief, and Yasser F. Hassan. *CLOE: a cross-lingual ontology enrichment using multi-agent architecture*. Enterprise Information Systems 13, no. 7-8, pp. 1002-1022, 2019. DOI: <https://doi.org/10.1080/17517575.2019.1592232>

This articles is a joint work with Mohamed Ali. Mohamed Ali was a Ph.D. student at the University of Alexandria, Egypt. In this article, Mohamed Ali and Said Fathalla developed the idea of enriching ontologies from a multilingual text corpora. Shimaa Ibrahim was responsible for implementing the proposed approach, designing and preparing the case study, conducting the experiments, and evaluating the results. Mohamed Ali, Said Fathalla, and Shimaa Ibrahim wrote the manuscript.

- *Conference Papers (peer reviewed)*

3. **Shimaa Ibrahim**, Said Fathalla, Jens Lehmann, and Hajira Jabeen. *Multilingual ontology merging using cross-lingual matching*. In Proceedings of IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 113-120. IEEE, 2020. DOI: <https://doi.org/10.1109/WIIAT50758.2020.00020>

The idea of merging ontologies in different natural languages was developed by Shimaa Ibrahim and further conceptualized with Said Fathalla. Shimaa was responsible for implementing the source code, performing the experiments, evaluating the results, and writing the manuscript. All authors revised the manuscript.

4. **Shimaa Ibrahim**, Said Fathalla, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. *From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment*. In Proceedings of 15th International Conference on Semantic Systems

(SEMANTiCS), Research & Innovation, pp. 215-230. Springer International Publishing, 2019. DOI: https://doi.org/10.1007/978-3-030-33220-4_16

The idea of enhancing and expanding our previous work in [42] was developed by Shimaa Ibrahim and further conceptualized with Said Fathalla. Shimaa implemented all parts of the source code, conducted the experiments, evaluated the results, and wrote the manuscript. All authors revised the manuscript.

5. Mohamed Ali, Said Fathalla, **Shimaa Ibrahim**, Mohamed Kholief, and Yasser Hassan. *Cross-lingual ontology enrichment based on multi-agent architecture*. Procedia Computer Science 137 (SEMANTiCS), Research & Innovation, pp. 127-138, 2018. DOI: <https://doi.org/10.1016/j.procs.2018.09.013>

This article is a joint work with Mohamed Ali. Mohamed Ali was a Ph.D. student at the University of Alexandria, Egypt. In this article, Mohamed Ali and Said Fathalla developed the idea of enriching ontologies from a multilingual text corpora. Shimaa Ibrahim was responsible for implementing the proposed approach, designing and preparing the case study, conducting the experiments, and evaluating the results. Mohamed Ali, Said Fathalla, and Shimaa Ibrahim wrote the manuscript.

- *Poster Papers (peer reviewed)*

6. **Shimaa Ibrahim**, Said Fathalla, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. *OECM: A cross-lingual approach for ontology enrichment*. In European Semantic Web Conference (ESWC), pp. 100-104. Springer International Publishing, 2019. DOI: https://doi.org/10.1007/978-3-030-32327-1_20

The idea to exploit ontologies in different natural languages to enrich ontologies was proposed by Shimaa Ibrahim and further developed together with Said Fathalla. Shimaa implemented the source code, conducted the experiments, evaluated the results, and wrote the manuscript. All authors revised the manuscript.

1.4.3 Thesis Structure

We present an overview of the thesis to prepare the reader for the upcoming chapters. The thesis is structured in eight chapters. Chapter 1 introduces the overall motivation, challenges, research questions, scientific contributions addressing research questions, and a list of published scientific papers describing these contributions. Chapter 2 provides preliminary information about the terminologies, tools, datasets, and technologies used in the work presented in this thesis. Chapter 3 presents our research efforts for addressing **RQ1**. We introduce the **MoMatch** framework for matching ontologies in various natural languages. Chapter 4 investigates the crucial area of enriching ontologies through the use of unstructured sources in several languages (e.g., multilingual text corpora) by introducing **CLOE**. Chapter 5 examines the utilization of structured data, such as other ontologies in other languages, as a source to enrich ontologies by presenting **OECM**. Chapter 6 presents a merging approach, **MULON**, that produces a rich ontology compared to earlier approaches. Chapter 4, Chapter 5, and Chapter 6 address **RQ2**. Chapter 7 discusses several aspects that have a substantial influence on the enrichment process and presents a metrics suite (**QASO**) that evaluates the quality of the enrichment process. This chapter addresses **RQ3**. Chapter 8 summarizes the contributions presented in this thesis and provides directions for future research.

Preliminaries

In this chapter, we define some commonly used terms in this thesis to lay the ground for the next subsequent chapters. Section 2.1 gives an overview of the Semantic Web technologies used in the thesis. Section 2.2 explain the meaning of ontology matching. Section 2.3 outlines the similarity measures that are used in the ontology matching process. Section 2.4 gives an introduction to ontology enrichment. Section 2.5 provides a comprehensive summary of the dataset that used in this thesis.

2.1 Semantic Web Technologies

Despite the rapid expansion of digital data, the efficient management, integration, and usage of information on the Internet have become crucial obstacles. The use of Semantic web technologies is a crucial answer to this imperative since they provide a structured method for providing web resources with explicit semantics. This, in turn, enables automated reasoning and intelligent processing. According to Tim Berners-Lee [1], the Semantic Web paradigm expands the traditional web of documents to include a web of data, where information is not only accessible by humans but also comprehensible by machines. The evolution in perspective is supported by a set of standards and procedures designed to facilitate the representation, sharing, and retrieval of structured data on the web. Figure 2.1 illustrates the several layers of technology associated with the Semantic Web. The top layers are still undergoing changes, while the lower layers have been developing for a considerable period. The layers are categorized as follow¹:

1. **Hypertext Web technologies.** The bottom layers encompass the technologies that describe and identify semantic web resources. They consist of three components: [Uniform Resource Identifier \(URI\)](#), which provides a unique definition for the resource; Unicode, which facilitates the representation and manipulation of text in multiple languages; and semantic technologies, which utilize [Extensible Markup Language \(XML\)](#) and [XML Namespaces](#) to add semantics to the data and establish connections between namespaces from various sources.
2. **Standardized Semantic Web technologies.** The middle layers consist of technologies that have been standardized by the [World Wide Web Consortium \(W3C\)](#) in order to facilitate the

¹ https://en.wikipedia.org/wiki/Semantic_Web_Stack

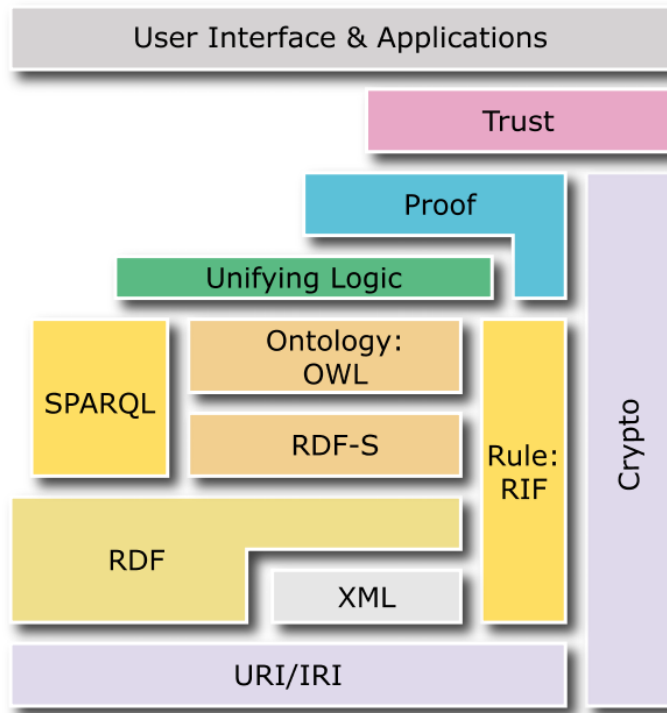


Figure 2.1: Semantic Web Stack (known as Semantic Web Cake) [43].

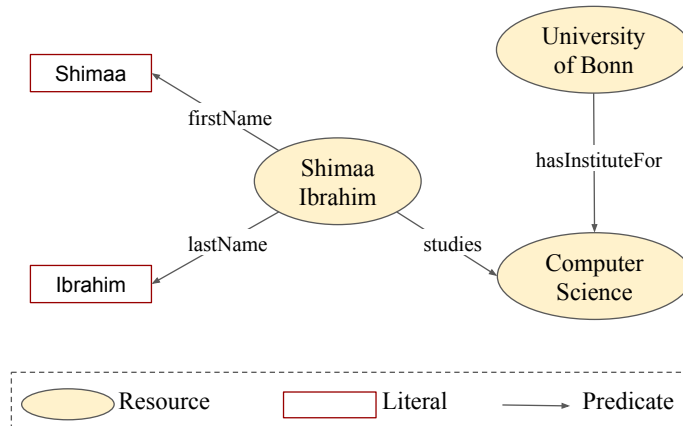
development of semantic web applications. They include languages used to represent the semantics of the data, such as [Resource Description Framework \(RDF\)](#), [Resource Description Framework Schema \(RDFS\)](#), and [OWL](#). It also includes languages used to represent rules, such as [Rule Interchange Format \(RIF\)](#) and [Semantic Web Rule Language \(SWRL\)](#). Additionally, languages used to query and retrieve data, such as [Simple Protocol and RDF Query Language \(SPARQL\)](#), are part of this layer.

3. **Unrealized Semantic Web technologies.** The top layers include technologies that support the realization of the semantic web; however, these technologies are not yet standardized. These technologies (e.g., logic, cryptography, trust, and proof) are employed to verify the source and logic of the semantic web data (i.e., trusted sources). The user interface serves as the ultimate layer that facilitates human interaction with semantic web applications.

The semantic technologies that are used in this thesis are described in more details in the following sub-sections.

2.1.1 Resource Description Framework

[RDF](#) is a data model developed by the [W3C](#) to describe resources over the World Wide Web [44]. The fundamental concepts of RDF are resources, properties, and statements (known as triples) [45]. A resource is a fact or a thing that can be described and identified, such as authors, books, publishers, places, people, hotels, rooms, etc. An [RDF](#) resource is uniquely identified by a [URI](#) reference that can

Figure 2.2: An **RDF** graph represents statements a–d.

be a **Uniform Resource Locator (URL)** or another type of unique identifier [45]. Properties (known as predicates) are another kind of resource that are used to indicate relationships between resources, e.g., "published by", "gender", "first name", etc. Statements (known as **RDF** triples) describe metadata about resources. An **RDF** triple is defined as (s, p, o) , where s , p , and o represent *subject*, *predicate*, and *object*, respectively. The *subject* is a resource, the *predicate* is the relationship that exists between the *subject* and the *object*. The *object* can be either a literal or a resource. Literals are composed of either a string together with its language tag, or a value along with its data type. Subjects and objects can also be blank nodes. A set of **RDF** triples forms a directed graph, vertices can be either subjects or objects, and the edges are the predicates. **RDF** is serialized in several formats, such as N-Triples², Turtle³, and RDF/XML⁴. In order to gain a comprehensive understanding of how **RDF** can effectively represent information, let us consider a simple example. Consider the following sentence: “*Shimaa Ibrahim studies at the Institute of Computer Science at the University of Bonn*”. It is important to first identify the resources, properties, and statements associated with this sentence. In this example, the resources are Shimaa Ibrahim, Institute of Computer Science, and the University of Bonn, while the corresponding statements are: a) the first name of Shimaa Ibrahim is Shimaa, b) the last name of Shimaa Ibrahim is Ibrahim, c) Shimaa Ibrahim studies computer science, and d) the University of Bonn has an Institute for Computer Science. Figure 2.2 shows the **RDF** graph representation for these statements. Each statement is represented by a single **RDF** triple as shown in Listing 2.1, where the **URI** of the identified resources is `http://www.ShimaaIbrahim.org`.

```
<http://www.ShimaaIbrahim.org/ShimaaIbrahim> <http://www.ShimaaIbrahim.org/firstName> "Shimaa" .
<http://www.ShimaaIbrahim.org/ShimaaIbrahim> <http://www.ShimaaIbrahim.org/lastName> "Ibrahim" .
<http://www.ShimaaIbrahim.org/ShimaaIbrahim> <http://www.ShimaaIbrahim.org/studies>
<http://www.ShimaaIbrahim.org/ComputerScience> .
<http://www.ShimaaIbrahim.org/UniversityOfBonn> <http://www.ShimaaIbrahim.org/hasInstituteFor>
<http://www.ShimaaIbrahim.org/ComputerScience> .
```

Listing 2.1: The set of **RDF** triples uses the N-Triples serialization to represent statements a–d.

² <https://www.w3.org/TR/n-triples/>

³ <https://www.w3.org/TR/turtle/>

⁴ <https://www.w3.org/TR/rdf-syntax-grammar/>

2.1.2 Ontologies

Ontologies are considered one of the significant cornerstones of representing information more meaningfully, providing machine-understandable semantics of knowledge. They are employed to formally express metadata within the semantic web stack and act as containers for capturing semantic information within specific domains, facilitating the sharing and reusing of knowledge within those domains [7, 46]. They define concepts and the interrelationships between them within a given domain, providing a formal representation that specifies classes, individuals, attributes, and relations within that domain [47]. In this context, a class denotes a concept or category, whereas an individual represents a specific instance of that class. An attribute defines the properties of classes and individuals, and a relation defines the relationship between classes and individuals. Formally, an ontology can be defined as follows:

Definition 1 (Ontology). An ontology can be represented as a tuple of five components [48]: $O = \langle C, R, P, I, X \rangle$, where C is the set of classes/concepts, R is the set of relationships between classes (known as object properties), P is the set of data properties (a specific type of relation whose domain is a class and the range is a data type), I is the set of class instances, and X is the set of axioms and rules used for checking and verifying ontology consistency and new knowledge inference.

Ontologies store data hierarchically which facilitates inferring new information from data and their interrelationships. Ontologies are being widely used in various fields of computer science such as Knowledge Management [49], Data Integration [50], and Information Retrieval [51], and in other fields of science, including Healthcare [52], Biology and Bioinformatics [2, 53] Engineering [3], and Social Sciences [54].

RDF is not able to describe a group of related resources and their interrelationships, such as classes, sub-classes, and properties, since it can only define resources in a form of triples, i.e., (s, p, o) . Thus, another data modeling schema should capture the formal semantics of **RDF** resources and interlinking relationships to help intelligent agents efficiently handle this information. The two most widely used schema languages for ontologies are **RDFS** and **OWL**.

Resource Description Framework Schema

RDFS is a semantic extension of **RDF**, which provides a data-modeling vocabulary for **RDF** data [44]. **RDFS** serves as a framework to define the vocabulary used to describe **RDF** data. Within the **RDFS** framework, resources are organized into distinct groups known as classes. These classes, denoted by the `rdfs:Class` property, define the characteristics of resources. Classes can be organized in hierarchies. As an illustration example, let us consider a class X is a specialization of a superclass Y . In such a scenario, X is considered a subclass of Y and can be formally defined by utilizing the `rdfs:subClassOf` property. The instances or individuals of a class are defined using `rdf:type`. Properties of class are defined using `rdf:Property` and `rdfs:subPropertyOf` which are used to define the relation between properties. In order to increase human-readability, `rdfs:comment` and `rdfs:label` are used to provide descriptions about classes.

The Web Ontology Language

OWL is an ontology language for the Semantic Web, serves as an extension to **RDFS** [55]. This extension involves the addition of language primitives that are specifically designed to enhance the

expressiveness of **RDFS** [45]. It enables the formal definition of resources with diverse axioms within the domain in a more comprehensive manner. The **OWL** formal semantics describes classes, properties, and their instances to deduce its logical consequences, i.e. facts are not explicitly stated in the ontology but entailed by the semantics [56]. Formal semantics allows reasoning about the knowledge presented in the ontology. The elementary modeling capabilities of **RDFS** allow only organizing concepts and properties in typed hierarchies, specifying domain and range restrictions, and identifying instances. However, **OWL** allows for more advanced reasoning about classes and properties by offering a wider range of capabilities compared to **RDFS**. Some examples of these capabilities are: describing classes using many description logic terms, such as `owl:complementOf`, `owl:intersectionOf`, and `owl:unionOf`. It also employs `owl:disjointWith` to state that there are no shared instances between two classes, for example, the disjoint classes “Male” and “Female”. Furthermore, `owl:equivalentClass` is utilized to identify two classes as equivalent, whereas `owl:sameAs` specifies two individuals as identical.

2.2 Ontology Matching

Ontology matching is a complex process that aims to bridge the semantic gap between multiple representations of the same domain. [57]. The matching process identifies correspondences (i.e., an alignment) between semantically related entities, such as classes, properties, and instances, of two or more ontologies that satisfy specific criteria [58]. The result of the matching process (i.e., correspondences) can be used for various tasks, such as ontology enrichment [59], ontology merging [60], query answering [61], or data translation [10]. Formally, ontology matching can be defined as follows:

Definition 2 (*Ontology Matching*). “a function f which, from a pair of ontologies to match O_1 and O_2 , an input alignment A , a set of parameters p and a set of resources r , returns an alignment A' between these ontologies: $A' = f(O_1, O_2, A, p, r)$ ” [62].

Definition 3 (*Alignment*). the alignment between two ontologies, O_1 and O_2 , is a set of correspondences between pairs of entities belonging to O_1 and O_2 , respectively [62]. It is the output of the matching process.

Definition 4 (*Correspondence*). given two ontologies O_1 and O_2 , the correspondence between two entities e_{11} and e_{12} , where $e_{11} \in O_1$ and $e_{12} \in O_2$, is the relation that produces according to a matching algorithm between e_{11} and e_{12} [62]. It can be represented by the triple: $\langle e_{11}, e_{12}, r \rangle$, where r is the relation between the two entities e_{11} and e_{12} .

For instance, consider the two classes “Chairman” and “President” in O_1 and O_2 respectively. These classes can be identified as equivalent. Similarly, the class “Chairman” can be marked as a subclass of “Person” in the second ontology.

Matching ontologies is the same natural language known as *monolingual ontology matching* [13]. With the increasing amount of multilingual data available on the Web and the subsequent development of ontologies in various natural languages, there is a increasing demand for *cross-lingual ontology matching*.

Definition 5 (*Cross-lingual ontology matching*). the process of matching the two ontologies O_1 and O_2 , which are in different natural languages, i.e., $L_1 \neq L_2$, where L_1 is the natural language of O_1 and L_2 is the natural language of O_2 [13].

Various techniques for ontology matching have been proposed, which can be classified as [57]:

- (a) *Element-level*: these matching techniques focus on identifying correspondences by analyzing entities in the ontologies individually, without considering their relationships with other entities or instances.
- (b) *Structure-level*: these matching techniques involve analyzing the relationships between entities and their instances in the ontology to identify correspondences.

In this thesis, we use element-level matching techniques by employing several similarity measures described in the next section.

2.3 Similarity Measures

Similarity measures assess the degree of similarity between ontology entities based on various criteria [62]. In this thesis, we utilize syntactic and semantic similarity measures.

2.3.1 Syntactic Similarity

Syntactic similarity focuses on the lexical representations of terms. It uses string-based techniques to match names, labels, and comments of ontology entities [62]. These techniques consider strings as alphabetical sequences of letters. The greater the similarity between the strings, the more likely they represent identical concepts. String-based techniques are classified as [35]:

- (a) *Sequence-based techniques*: these techniques consider input strings as a sequence of characters, such as Hamming distance, Jaro, Jaro Winkler, Levenshtein, Partial ratio, Partial token sort, Ratio, and Token sort. Given two strings, the primary goal is to identify and quantify how one string can be transformed into the other one through operations such as insertions, deletions, or substitutions. The fewer operations conducted, the higher the probability of the two strings being a match. The sequence-based similarity measures that are utilized in this thesis are described in detail in Chapter 3 and Appendix A (Table 1).
- (b) *Set-based techniques*: these techniques consider input strings as sets or multiple sets of tokens, such as Cosine, Dice, Jaccard, Overlap coefficient, and Tversky Index. The main objective is to determine how much two strings overlap in terms of their constituent parts, which can be characters, words, or n-grams (a sequence of n words). The set-based similarity measures that are used in this thesis are thoroughly explained in Chapter 3 and Appendix A (Table 2).

2.3.2 Semantic Similarity

Semantic similarity focuses on the meaning of terms. It calculates the degree of relatedness between terms exploiting external knowledge sources such as WordNet⁵. WordNet [63] is a large lexical database of English designed to support NLP and computational linguistics tasks by providing a rich repository of word meanings and relationships. It organizes words into sets of synonyms known as

⁵ <https://wordnet.princeton.edu/>

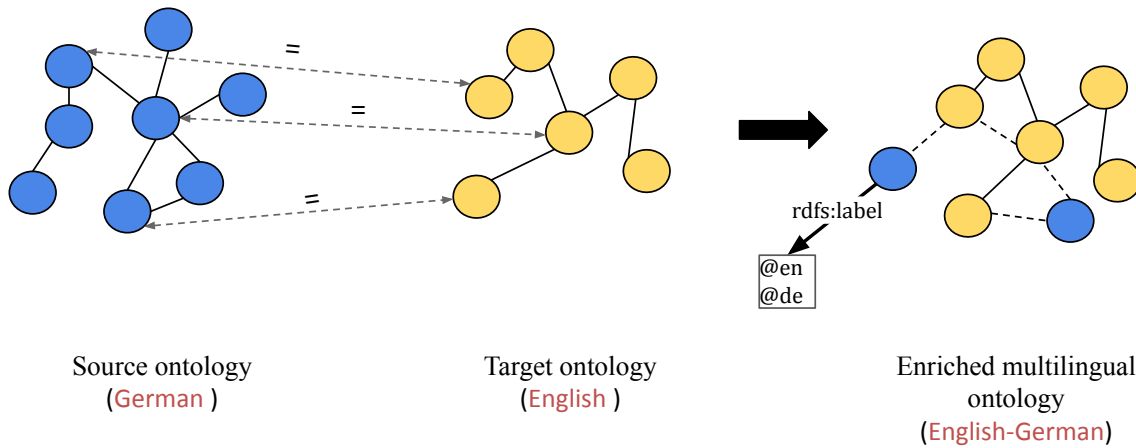


Figure 2.3: An illustration of enriching English ontology from German ontology and producing a multilingual enriched ontology in English and German.

synsets, and identifies various semantic relations among these synsets. Words that share the same meaning (i.e., semantically equivalent) within a context are grouped into these synsets. For example, the words “car” and “automobile” are grouped into a single synset because they have the same meaning. WordNet uses path-based measures which calculate the similarity based on the shortest path between synsets in the WordNet hierarchy [64]. If two words are semantically equivalent, i.e., belonging to the same WordNet synset, the path distance is 1.00.

2.4 Ontology Enrichment

Expanding an ontology with additional information extracted from external sources is called *ontology enrichment* [7]. It is important to place the new information at the correct positing in the ontology. A monolingual ontology with labels or local names in a single natural language makes semantic interoperability and knowledge sharing across linguistic borders difficult.

Cross-lingual ontology enrichment extend the ontology with additional multilingual information extracted from multilingual sources, helping to promote semantic interoperability across languages [5, 6]. It enables to produce not only monolingual ontologies, but also multilingual ones (see Figure 2.3). Multilingual ontologies represent entities in more than one language by adding *rdfs:label* for each language using language-tagged strings, such as the languages-tags @en, @de, and @ar are used for English, German, and Arabic respectively.

Some ontologies have been designed to model already existing domains in an independent way, which leads to ontologies with overlapping and redundant information. Enriching ontologies by integrating these ontologies (known as *ontology merging*) into a coherent ontology allows this information to be optimally exploited and reconciled. Generally, there are two types of ontology merging [31]:

- (a) *Symmetric merging*; which aims to integrate all resources (classes and properties) in the input ontologies to a single ontology, i.e. preserves all input resources (which is used in Chapter 6).

- (b) *Asymmetric merging*; which considers one of the input ontologies as the target ontology and merges the remaining non-redundant resources of the input ontologies into the target ontology.

2.5 Dataset

In order to evaluate the effectiveness of the cross-lingual ontology matching process, we use the MultiFarm benchmark⁶ [65]. Since 2012, MultiFarm has been included in the [Ontology Alignment Evaluation Initiative \(OAEI\)](#) to evaluate and compare the performance of multilingual matching approaches. [OAEI](#)⁷ is an annual international ontology matching competition. MultiFarm comprises seven ontologies (*Cmt*, *Conference*, *ConfOf*, *Edas*, *Ekaw*, *Iasted*, *Sigkdd*), derived from the Conference track of [OAEI](#), in addition to the corresponding cross-lingual alignments between them. Furthermore, it includes the translation of these ontologies into nine different languages (Chinese, Czech, Dutch, French, German, Portuguese, Russian, Spanish, and Arabic). The statistics for MultiFarm ontologies are presented in Table 2.1. We compare the matching outcomes from the proposed approaches in this thesis with the results from the state-of-the-art approaches presented in [OAEI 2018](#)⁸, [2019](#)⁹, and [2020](#)¹⁰.

Table 2.1: Statistics of ontologies in the MultiFarm benchmark.

Ontology Name	Classes	Object Properties	Data Properties
Ekaw	74	33	0
Edas	104	30	20
Conference	60	46	18
ConfOf	38	13	23
Cmt	36	49	10
Iasted	140	38	3
Sigkdd	49	17	11

⁶ <https://www.irit.fr/recherches/MELODI/multifarm/>

⁷ <https://oaei.ontologymatching.org/>

⁸ <https://oaei.ontologymatching.org/2018/results/multifarm/index.html>

⁹ <https://oaei.ontologymatching.org/2019/results/multifarm/index.html>

¹⁰ <https://oaei.ontologymatching.org/2020/results/multifarm/index.html>

Multilingual Ontology Matching

This chapter presents the following publication (see Appendix A):

Shimaa Ibrahim, Said Fathalla, Jens Lehmann, and Hajira Jabeen. “Toward the multilingual Semantic Web: Multilingual ontology matching and assessment”. IEEE Access, vol. 11, pp. 8581-8599, 2023. DOI: <https://doi.org/10.1109/ACCESS.2023.3238871>

Authors’ contributions Shimaa Ibrahim developed the idea of multilingual ontology matching approach that uses different string similarity measures. She implemented all parts of the source code. She wrote the manuscript. All authors revised the manuscript.

Summary

Ontologies are considered one of the significant cornerstones of representing information more meaningfully, providing machine-understandable semantics of knowledge. It is a container for capturing semantic information of a particular domain, which allows sharing and reusing knowledge in this domain [7, 46]. Ontology matching is one of the major tasks within ontology development. In recent years, the number of multilingual ontologies has increased gradually due to the rapid growth of multilingual data on the Semantic Web. Since monolingual ontologies that include labels or local names presented in a particular language are not understandable to people who speak other languages, this hinders semantic interoperability. Therefore, cross-lingual ontology matching has become a prime approach to bridging the semantic gaps between monolingual ontologies by identifying their correspondences. Despite the vast research in monolingual ontology matching, few researchers have shed light on cross-lingual ontology matching [13, 14]. Nonetheless, a clear winner has yet to emerge in solving the cross-lingual matching problem. Further investigations are still needed to advance cross-language ontology matching techniques to achieve good results compared to monolingual approaches. As a result, a flexible framework for ontology matching across languages is required, which is not limited to one particular language.

In the publication, *Toward the multilingual Semantic Web: Multilingual ontology matching and assessment*, we propose **MoMatch** for matching ontologies in different natural languages. **MoMatch** uses sequence-based and set-based string similarity measures in addition to machine translation to match classes and properties across ontologies. The input to **MoMatch** is two ontologies, which may be expressed in the same natural language or two different ones. The output is the alignment between the input ontologies. **MoMatch** involves the following four phases: 1) Resource extraction: the process of extracting all resources (classes and properties) from the input ontologies; 2) Translation: the task of translating the extracted resources; 3) Pre-processing: utilizing **NLP** techniques such as tokenization and POS-tagging to clean and prepare the translated resources for the matching phase; and 4) Matching: the identification of correspondences between the input ontologies. **MoMatch** can match not only ontologies in different natural languages (i.e., cross-lingual matching) but also ontologies in the same natural language (i.e., monolingual matching). The research questions addressed in this publication are: RQ1) What is the most effective string similarity measure(s) for matching multilingual terms across different ontologies? and RQ2) How strong is the performance of the similarity measure when the language of the input ontologies is changed?

We show the usability of **MoMatch** by addressing a use case for matching cross-lingual ontologies from the scholarly communication domain (the scientific events ontology [66]) and one for matching monolingual ontologies from the biomedical domain (Ontologies from BioPortal¹). **MoMatch** can be easily adapted for other use cases and domains.

We design a series of experiments employing MultiFarm benchmark² [65] (see Section 2.5 for more details) from the T-Box/Schema matching track of **OAEI 2020**³. The experimental outcomes are presented in order to conduct a comprehensive analysis pertaining to the comparison of 13 similarity measures across various languages. Out of the similarity measures considered, eight are string-based, including Hamming distance, Levenshtein, and Token sort, while five are set-based, including Cosine,

¹ <https://biportal.bioontology.org/>

² <https://www.irit.fr/recherches/MELODI/multifarm/>

³ <http://oaei.ontologymatching.org/2020/>

Jaccard, and Overlap coefficient. Ontologies in Indo and non-Indo-European languages are used. Ten language pairs have been tested in [MoMatch](#). Finally, we compare [MoMatch](#) with the state-of-the-art approaches presented in [OAEI 2020](#).

We find that the precision and recall of the matching process are significantly influenced by the string similarity measures that are employed. It is crucial to consider the characteristics of the ontologies being matched when choosing a similarity measure, as well as whether precision or recall is more relevant to the matching process. We observe that Levenshtein, Hamming, and Jaccard similarity measures have the highest precision and F-measure, while partial ratio, partial token sort, and overlap coefficient have the highest recall for matching multilingual ontologies (RQ1). Accordingly, we sort the 13 similarity measures into two lists according to precision and recall to support choosing the most appropriate one for the matching process. The correlation between language pairs is consistently high and positive (RQ2). This finding shows that, regardless of language pairs, the resulting precision, recall, and F-measure scores produced from the combination of several similarity measures and threshold values remain highly correlated. [MoMatch](#) outperforms most other state-of-the-art systems in terms of precision, recall, and F-measure when using Jaccard and Levenshtein similarities.

Based on the findings from [OAEI 2022](#)⁴ and [2023](#)⁵, which were conducted after our publication, it is evident that the performance of [MoMatch](#) remains competitive when compared to state-of-the-art approaches as shown in Table 3.1. [MoMatch*](#) represents results when considering new correspondences that were missing from the gold standard. CIDER-LM [67] and Matcha [68] employ [Large Language Model \(LLM\)](#)s such as Sentence-BERT [69]. [LLMs](#) are experiencing rapid evolution and have recently shown their ability in the ontology matching process [70]. Despite that, Matcha could not provide sufficient results because they had a character encoding issue [68]. CIDER-LM [67] achieved average results in terms of F-measure while surpassing other approaches in terms of recall.

We believe that [MoMatch](#) established a significant step toward a multilingual Semantic Web. Based on our in-depth analysis for comparing similarity measures across different languages, researchers can choose the most efficient similarity measure for their task.

Table 3.1: Aggregated results for the state-of-the-art techniques in [OAEI 2022](#) and [2023](#) compared to [MoMatch](#). Bold entries are the top scores.

Approaches	Precision	F-measure	Recall
CIDER-LM [67]	0.16	0.25	0.58
LSMatch [71, 72]	0.24	0.04	0.02
LSMatch Multilingual [71, 72]	0.68	0.47	0.36
LogMap [73, 74]	0.72	0.44	0.31
LogMapLt [73, 74]	0.24	0.04	0.02
Matcha [68]	0.37	0.08	0.04
MoMatch	0.52	0.38	0.31
MoMatch*	0.89	0.45	0.31

⁴ <https://oaei.ontologymatching.org/2022/results/multifarm/index.html>

⁵ <https://oaei.ontologymatching.org/2023/results/multifarm/index.html>

Ontology Enrichment from Multilingual Unstructured Sources

This chapter presents the following publications (see Appendix B and Appendix C) for enriching ontologies from a multilingual text.

1. Mohamed Ali, Said Fathalla, **Shimaa Ibrahim**, Mohamed Kholief, and Yasser Hassan. “Cross-lingual ontology enrichment based on multi-agent architecture”. *Procedia Computer Science* 137 (SEMANTiCS), Research & Innovation, pp. 127-138, 2018. DOI: <https://doi.org/10.1016/j.procs.2018.09.013>
2. Mohamed Ali, Said Fathalla, **Shimaa Ibrahim**, Mohamed Kholief, and Yasser F. Hassan. “CLOE: a cross-lingual ontology enrichment using multi-agent architecture”. *Enterprise Information Systems* 13, no. 7-8, pp. 1002-1022, 2019. DOI: <https://doi.org/10.1080/17517575.2019.1592232>

Authors’ contributions: These articles are a joint work with Mohamed Ali. Mohamed Ali was a Ph.D. student at the University of Alexandria, Egypt. In these articles, Mohamed Ali and Said Fathalla developed the idea of enriching ontologies from a multilingual text corpora. Shimaa Ibrahim was responsible for implementing the proposed approach, designing and preparing the case study, conducting the experiments, and evaluating the results. Mohamed Ali, Said Fathalla, and Shimaa Ibrahim wrote the manuscript.

Summary

The exponential growth of unstructured data in different natural languages has opened new opportunities for the development of the Semantic Web [75]. This includes transforming such data into machine-readable and understandable formats, resulting in the development of numerous ontologies in different natural languages. Ontology, being one of the crucial components of the Semantic Web, plays a vital role in capturing and organizing information and providing a standardized understanding of concepts, relationships, and instances within a particular domain. Ontology development and enrichment are very challenging, especially when it comes to multilingual data. Innovative algorithms capable of processing multilingual information and enriching an ontology using external resources, such as unstructured text or ontology, in different natural languages are essential to utilizing these resources [14]. However, multilingual and cross-lingual ontology enrichment still needs extensive improvements because of the growing interest in ontologies and the Semantic Web. Researchers are focusing on exploring methods to enrich ontologies from external resources in a specific language by extracting relevant information and mapping it to the appropriate concepts in ontology [7, 76, 77]. Their ultimate objective is to create a comprehensive ontology that can facilitate knowledge sharing and discovery in various domains. Therefore, it is critical to establish a standardized multilingual framework that can use external resources in various natural languages and facilitate effective cross-lingual ontology enrichment.

In the publication “*Cross-lingual ontology enrichment based on multi-agent architecture*”, we propose a domain-dependent approach to enriching ontologies from multilingual text or ontology using a multi-agent architecture. The subsequent publication, “*CLOE: a cross-lingual ontology enrichment using multi-agent architecture*”, builds upon the previous approach and offers a domain-independent solution. Therefore, this chapter presents the second publication (CLOE), as it provides a more holistic approach than the first.

CLOE is a multi-agent architecture approach that enriches ontologies in different domains and languages using multilingual, unstructured text. This approach has several challenges that require careful consideration: 1) the diversity of pre-processing techniques used for multilingual text; 2) matching concepts between ontologies in different languages; 3) handling communication between agents; 4) eliminating unnecessary text from the input; and 5) determining how agents can learn from each other.

CLOE takes text corpora, which can be in various natural languages, and the target ontologies, which need to be enriched, as input and generates the enriched ontologies. CLOE comprises five phases. The first phase is text pre-processing and concept extraction, which involves cleaning and preparing the text using various NLP techniques. The second phase is domain identification, which classifies the text into the appropriate domain using a classification approach. The third phase is candidate sentence selection, which eliminates unnecessary or irrelevant sentences from the text. The fourth phase is cross-lingual ontology enrichment, which enriches ontologies using a multi-agent model. We propose an innovative enrichment algorithm that enables intelligent agents to learn about their ontologies through their interaction with each other using a communication scheme. The multi-agent model has three types of agents: master agent (MA), domain agent (DA), and language agent (LA). We build a three-level hierarchical multi-agent model (see Figure 4.1) with a single master agent MA at the top of the hierarchy, a set of domain agents DA_i at the middle level, where $i = 1, 2, \dots, n$ and n represents the number of domain agents, and a set of language agents LA_{ij} at the bottom level, where $j = 1, 2, \dots, m$ and m represents the number of language agents. For example, a domain agent DA_1

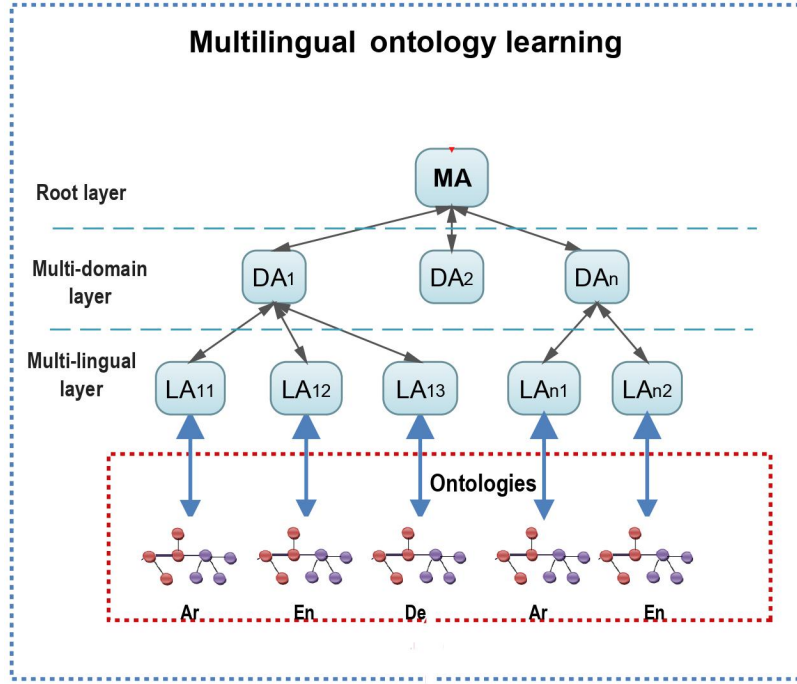


Figure 4.1: The three-level hierarchical multi-agent model from the architecture of CLOE [37].

manages three language agents LA_{11} , LA_{12} , and LA_{13} . Language agents can learn about concepts, relations, and instances for a particular ontology in a specific language. The number of domain agents depends on the number of domains in the ontologies to be enriched, and the number of language agents depends on the number of languages in the ontologies in a particular domain. The master agent manages and stores information about domain and language agents involved in the hierarchy to be able to decide which domain agent is responsible for learning the contents of each incoming text based on the language of the text. We develop a new communication algorithm that enables agents to learn from each other about new concepts or relations. The fifth and final phase is ontology verification and validation, which ensures the accurate extraction of candidate relations and concepts (verification) and confirms that they are the correct relations and concepts for the target ontology (validation). This phase employs ontology experts to validate or modify the learned concepts and relations.

We show the practicality of CLOE through a use case that entailed enriching an ontology from the Information Technology domain with a multilingual text in English, German, and Arabic. The objective is to demonstrate the complete process, starting with submitting the multilingual text and culminating in extracting concepts and relations and integrating them into the ontology.

We assess the effectiveness of CLOE using two evaluation methods. The first method involves a satisfaction questionnaire¹ filled out by 24 ontology engineers. The questionnaire consists of 12 questions to evaluate the different phases of CLOE. Out of the 24 participants, 20 of them note that this approach is a step towards developing a domain-independent approach for ontology enrichment. The second evaluation method is a gold standard-based evaluation to determine the accuracy of the concepts and relations used in the learning process. To conduct this evaluation, we designed an

¹ <https://goo.gl/DnnPnv>

experiment using the 20 Newsgroups dataset², which contains approximately 20,000 newsgroup posts covering 20 different categories such as computers, religion, science, and politics. CLOE outperforms the four state-of-the-art approaches involved in the evaluation in terms of recall and F-measure.

CLOE presents a groundbreaking domain-independent approach that simultaneously enriches multiple ontologies in different languages by processing multilingual text. This approach has successfully addressed and resolved the challenges mentioned earlier. We believe that this work is the first attempt in this field, and it represents a valuable tool for ontology engineers. The most notable feature of CLOE is its ability to foster knowledge sharing and learning among agents, which enables them to learn from each other. This feature is particularly relevant in light of the growing need for effective communication and collaboration across different languages and domains.

² <http://qwone.com/~jason/20Newsgroups/>

Ontology Enrichment from Multilingual Structured Sources

This chapter presents the following publications (see Appendix D and Appendix E) for enriching ontologies from other ontologies in different natural languages.

1. **Shimaa Ibrahim**, Said Fathalla, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. “OECM: A cross-lingual approach for ontology enrichment”. In European Semantic Web Conference (ESWC), pp. 100-104. Springer International Publishing, 2019. DOI: https://doi.org/10.1007/978-3-030-32327-1_20
2. **Shimaa Ibrahim**, Said Fathalla, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. “From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment”. In Proceedings of 15th International Conference on Semantic Systems (SEMANTiCS), Research & Innovation, pp. 215-230. Springer International Publishing, 2019. DOI: https://doi.org/10.1007/978-3-030-33220-4_16

Authors’ contributions: The idea to exploit ontologies in different natural languages to enrich ontologies was proposed by Shimaa Ibrahim and further developed together with Said Fathalla. Shimaa implemented the source code, conducted the experiments, evaluated the results, and wrote the manuscript. All authors revised the manuscript.

Summary

The rapid increase in multilingual data on the Semantic Web poses a challenge to the creation of multilingual ontologies from monolingual ones. The lack of methods for cross-lingual ontology enrichment further exacerbates this challenge. Cross-lingual ontology enrichment, however, can significantly enhance semantic interoperability among different ontologies in various natural languages [13]. Multilingual ontologies can be built through cross-lingual ontology enrichment techniques that expand the target ontology with additional concepts and semantic relations extracted from external resources in other languages [7]. Attempting to achieve such enrichment through human labor is a time-consuming and error-prone task. Therefore, fully automated cross-lingual ontology enrichment approaches are highly desired. In the light of recent research on ontology enrichment, there is a considerable concern about enriching English ontologies from English resources rather than non-English ones, which is known as monolingual enrichment [7]. This raises a pivotal question; How can an ontology be enriched using another one in a different natural language? Most of the literature has focused on enriching monolingual ontologies with multilingual information in order to translate or localize these ontologies (i.e., adapting an ontology to a specific language and cultural community) [78, 79]. As mentioned before in Section 1.1, cross-lingual ontology matching is the first step in enrichment. Few researchers have tried to develop different techniques for cross-lingual ontology matching which focus on one-to-one translation between ontology terms [13]. As a result, unsuitable translations degrade the effectiveness of the matching procedure [13]. Hence, it is crucial to create novel techniques that can enrich ontologies by identifying the best translation out of all those that are available (also known as one-to-many translations) for a specific term.

In the publication *OEEM: A cross-lingual approach for ontology enrichment*, we propose a prototype semi-automated cross-lingual ontology enrichment approach (OEEM) to enrich ontologies from other ontologies in different natural languages. The subsequent publication, *From Monolingual to Multilingual Ontologies: The Role of Cross-Lingual Ontology Enrichment*, extends the OEEM approach by providing a comprehensive solution for creating consistent multilingual ontologies. Therefore, this chapter presents the second publication.

OEEM is a fully automated ontology enrichment approach that creates multilingual ontologies from monolingual ones using cross-lingual matching techniques. The aim of OEEM is to find the complementary information $\mathcal{T}_e = S - (S \cap T)$ from the source ontology S in order to enrich the target ontology T , where S and T are in two different natural languages. OEEM is characterized by its notable feature of selecting the best translation from the pool of available translations when matching classes across ontologies. This selection process enhances the quality of the matching process. OEEM considered only class labels, or local names, and three standard relations: `rdfs:subClassOf`, `owl:equivalentClass`, `owl:disjointWith`. OEEM comprises six phases: 1) Translation: translates the class labels of the source ontology and considers all available translations for each class, 2) Pre-processing: applies a variety of NLP techniques to process the class labels of both the target and translated source ontology, 3) Terminological matching: discovers potential correspondences between the source and the target ontologies by selecting the best translation for each class in the source ontology that matches the corresponding class in the target ontology through the utilization of lexical and semantic similarity measures, 4) Triple retrieval: uses an iterative procedure to determine which information from the source ontology can be used to enrich the target ontology and where such information should be placed in the target ontology hierarchy, 5) Enrichment: enriches the target ontology by including the multilingual information retrieved from the source ontology, and 6)

Validation: validates the enriched ontology to detect inconsistencies and syntax errors.

We demonstrate the usability of **OECD** by presenting a use case of how it can be utilized to enrich the Scientific Events Ontology [66] in English. We use ontologies in German and Arabic as sources for enrichment. The resulting ontology is a consistent multilingual ontology in English, German, and Arabic.

We conduct experiments using the MultiFarm benchmark [65] (see Section 2.5 for more details) to evaluate the effectiveness of the cross-lingual matching process in **OECD**. We thoroughly analyzed **OECD**'s matching outcomes by comparing them with the results provided in the **OAEI** 2018¹. We compare **OECD**'s matching results with the reference alignment provided in the MultiFarm benchmark, with four state-of-the-art approaches. We assess the quality of the enrichment process by comparing the resulting enriched ontology against a predefined reference ontology created manually by ontology experts.

Our findings indicate that incorporating semantic similarity into **OECD** and considering all available translations for a class significantly enhances the matching results. Therefore, the cross-lingual matching process in **OECD** achieved high precision, recall, and F-measure compared to the state-of-the-art approaches. The use of the iterative triple retrieval process allows **OECD** to retrieve all classes and their related subclasses from the source ontology hierarchy and place them in the right position in the target ontology, producing a consistent, enriched ontology. Thus, the performance of the enrichment process is promising, confirming the efficacy of our approach to cross-lingual ontology enrichment.

OECD presents a highly effective and practical solution for developing multilingual ontologies through the utilization of monolingual ones. Our approach is characterized by its robustness, which is reflected in its ability to incorporate both Indo-European and non-Indo-European languages, making it a highly versatile and adaptable methodology.

Despite the success of **OECD**, we have identified two limitations. The first limitation is that **OECD** only adds classes from the source ontology to the target ontology and does not consider properties. The second limitation is that expert input is required to create a reference ontology for assessing the enrichment process, which may not be feasible for extensive ontologies. These limitations may impede the efficiency and effectiveness of the ontology enrichment process, and further research is required to address these limitations and improve the methodology. We address these limitations in Chapter 6 and Chapter 7.

¹ <https://oaei.ontologymatching.org/2018/results/multifarm/index.html>

Multilingual Ontology Merging

This chapter presents the following publication (see Appendix F):

Shimaa Ibrahim, Said Fathalla, Jens Lehmann, and Hajira Jabeen. “Multilingual ontology merging using cross-lingual matching”. In Proceedings of IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 113-120. IEEE, 2020. DOI: <https://doi.org/10.1109/WIIAT50758.2020.00020>

Authors’ contributions: The idea of merging ontologies in different natural languages was developed by Shimaa Ibrahim and further conceptualized with Said Fathalla. Shimaa was responsible for implementing the source code, performing the experiments, evaluating the results, and writing the manuscript. All authors revised the manuscript.

Summary

The Semantic Web has experienced a significant increase in multilingual data, resulting in the development of several ontologies in various natural languages that represent the same domain. Consequently, many ontologies contain duplicated and overlapping information. Ontology merging optimally exploits and reconciles such duplicated and overlapping information by consolidating all relevant information from various input ontologies into a single, coherent ontology [80]. While the Web of Data includes information in several languages and the OWL standard aims to support numerous natural languages, there are no effective techniques for automatically utilizing and reconciling this multilingual information [8, 14]. Innovative methodologies for merging existing ontologies to create multilingual ontologies must be developed to construct a multilingual web and foster semantic interoperability among ontologies in different natural languages [13, 14]. Merging ontologies in different natural languages is a complex process that demands meticulous adaptation and rigorous techniques to control various steps of the merging process. The first step in the merging process is identifying mappings between the multilingual input ontologies, which is crucial for successful ontology merging [81]. However, most existing research on ontology merging only considers English ontologies, with no regard for those in other natural languages [80, 82, 83]. To the best of our knowledge, prior efforts have not been made to create multilingual ontologies by merging ontologies in different natural languages. Nonetheless, in our previous work (see Chapter 5), we have developed, **OECM** [40], a fully automated ontology enrichment approach, which creates a multilingual ontology by enriching a monolingual ontology with information from another one in a different natural language. The enrichment process of **OECM** focuses solely on classes and does not take properties into consideration. Therefore, we address this limitation in this chapter.

In the publication, *Multilingual ontology merging using cross-lingual matching*, we propose the **MULON** approach for creating a multilingual ontology by merging two ontologies in different natural languages. **MULON** consists of three modules. The first module is the preparation module, which extracts and processes all resources, such as classes and properties, from the two input ontologies. The second module is the merging module, which identifies potential cross-lingual matches using machine translation tools and lexical and semantic similarity measures. The input ontologies are then merged based on linguistic and structural knowledge, and redundant information is removed. We use symmetric merging to integrate all resources (classes and properties) in the input ontologies into a single ontology, i.e., preserve all input resources. The third module is the assessment module, which validates the merged ontology to detect inconsistencies and syntax errors that might be introduced during the merging process. The quality of the merged ontology is evaluated using a set of quality metrics, described in detail in Chapter 7.

In order to demonstrate the effectiveness of **MULON** and its impact on the Semantic Web community, we present three real-world use cases across different domains, including fiscal data, food, and scholarly communication. **MULON** has successfully addressed challenges, such as multilingualism, previously encountered by projects in the mentioned domains. Furthermore, it is adaptable and easily customized to suit other domains beyond those presented.

As part of our evaluation process, we used the MultiFarm benchmark [65] (see Section 2.5 for more details) to assess the quality of the cross-lingual matching process in **MULON**. We compare our cross-lingual matching results with the results from the **OAEI 2019**¹. We deliberately use ontologies

¹ <https://oaei.ontologymatching.org/2019/results/multifarm/index.html>

in non-Indo-European languages such as Arabic to demonstrate the great potential of **MULON** across different languages. The performance of **MULON** is impressive, as it outperformed three state-of-the-art approaches in terms of precision, recall, and F-measure in the matching process.

MULON presents a revolutionary approach that enables researchers to integrate ontologies in multiple natural languages effortlessly. It is a pivotal tool in building a multilingual Semantic Web. Moreover, **MULON**'s adaptability and ease of customization make it a versatile solution tailored to any domain.

Quality Assessment of Ontology Enrichment

This chapter presents the quality assessment part in the following publications (see Appendix A and Appendix F):

1. **Shimaa Ibrahim**, Said Fathalla, Jens Lehmann, and Hajira Jabeen. *Multilingual ontology merging using cross-lingual matching*. In Proceedings of IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 113-120. IEEE, 2020. DOI: <https://doi.org/10.1109/WIIAT50758.2020.00020>
2. **Shimaa Ibrahim**, Said Fathalla, Jens Lehmann, and Hajira Jabeen. *Toward the multilingual Semantic Web: Multilingual ontology matching and assessment*. IEEE Access, vol. 11, pp. 8581-8599, 2023. DOI: <https://doi.org/10.1109/ACCESS.2023.3238871>

Authors' contributions: see Chapter 3 and Chapter 6.

Summary

Improving the quality of ontology enrichment presents a multifaceted challenge that requires ensuring the precision and relevance of the additional information incorporated into ontologies. In fact, the critical nature of ontologies lies in their ability to accurately represent semantic information of a particular domain, which allows sharing and reusing knowledge in this domain [7, 46]. Hence, any inconsistencies or contradictions introduced during the process of enriching ontologies with additional information could significantly compromise their reliability and utility. These contradictions might arise due to mismatches in terminology, conflicting definitions, or inaccuracies in the representation of domain knowledge. Thus, we thoroughly investigate the various factors that influence the enrichment process. Our investigation reveals that the quality of the enrichment process is highly dependent on the quality of the input ontologies and the quality of the matching result.

Ontology matching plays a pivotal role in the enrichment process as it enables the identification and resolution of semantic heterogeneities. The matching process entails an analysis of the structure, content, and context of two or more ontologies to identify correspondences. As a result, the quality of the matching process is heavily dependent on the quality of the input ontologies. Therefore, it is crucial to ensure that the ontologies are correctly structured, semantically rich, and error-free.

Developing a high-quality ontology is a multi-faceted process that requires careful consideration of various dimensions, such as logic, reasoning, structure, and domain knowledge, that need to be modeled based on the specific tasks at hand [84]. Thus, it is important to have ontology quality criteria that can be used to evaluate the different dimensions of an ontology. These dimensions may have several measures (i.e., metrics) that can provide ontology developers with objective and quantitative insights into the quality of their ontology. By analyzing these metrics, developers can identify areas that need improvement, such as inconsistencies, redundancies, or incompleteness.

In this thesis, our goal is to assess the quality of the enrichment process; therefore, we focus on presenting metrics that evaluate certain dimensions of ontologies and their potential for knowledge representation (e.g., richness) rather than describing whether an ontology is effective or ineffective. In addition, we present metrics that evaluate the matching process and the resulting enriched ontology. Most of the research efforts have focused on evaluating the quality of ontologies from multiple dimensions, such as syntactic, semantic, structural, functional, content, and schema [29–32]. Every dimension has its own set of metrics to quantitatively assess the characteristics of an ontology [33]. Few researchers have addressed the quality of the matching process [34]. Therefore, determining the quality metrics that have a significant impact on ontology quality and consequently influence the ontology enrichment process is a challenging task.

In the publication, *Toward the Multilingual Semantic Web: Multilingual ontology matching and assessment*, we propose QASO, a metrics suite consisting of 17 quality metrics that provide a more comprehensive evaluation of the process of enriching ontologies. Out of these metrics, seven are employed to evaluate the quality of ontologies at the schema level, seven are utilized to evaluate the matching process, and three are employed to evaluate the enriched ontology.

The quality of ontologies has been addressed in several dimensions, such as richness, breadth, depth, and height in the schema inheritance tree, to evaluate their potential for rich knowledge representation. A well-designed, rich ontology should possess a harmonious combination of these dimensions in order to guarantee its ability to be a suitable source for enrichment. For example, a deep and narrow inheritance tree can make it challenging to find relevant information, while a broad and shallow tree can result in redundancy and confusion.

To assess the quality of the matching process, we approach it from two dimensions: completeness and correctness. Completeness (recall) refers to the ability of the matching process to retrieve all relevant correspondences, whereas correctness (precision) refers to the ability of the matching process to identify the correct correspondences between the input ontologies. A successful matching process should achieve high recall and precision, indicating that all relevant correspondences have been retrieved and are correct. However, recall and precision require a reference alignment, which can be expensive, particularly for larger ontologies. To address this issue, [QASO](#) includes two metrics, namely match coverage and match ratio. These metrics provide rough approximations for recall and precision based on the matching results obtained. Using these two metrics, we can evaluate the quality of the matching process without the need for a reference alignment, making the process more efficient and cost-effective. We successfully assessed the quality of the matching process in [MoMatch](#) using [QASO](#).

In the publication, *Multilingual ontology merging using cross-lingual matching*, we propose three metrics to assess the quality of the merged/enriched ontology (merging is a kind of enrichment as mentioned in Chapter 2). These metrics are intended to measure the relative coverage of information preserved and the level of compactness achieved in the merged ontology. A crucial requirement for merging ontologies is to ensure that the information represented in the input ontologies is preserved in the merged result, which these metrics have measured. Using [QASO](#), we successfully assess the factors that influence the quality of the enrichment process.

Conclusion and Future Work

This chapter provides a concise overview of the research conducted in this thesis, emphasizing the key findings and outlining potential avenues for future investigation.

8.1 Conclusion

In this thesis, the challenges that are related to the growth of ontologies across a variety of natural languages on the Semantic Web have been addressed comprehensively and systematically. The primary objective is to establish a robust foundation for semantic interoperability among different linguistic and domain-specific ontologies. The contributions of this thesis can be split into three major parts: 1) creating a flexible framework to study the use of various similarity measures in matching ontologies across diverse natural languages; 2) enriching ontologies by incorporating information from both unstructured and structured sources; and 3) evaluating the quality of the enrichment process.

Chapter 1 describes the research problem and the challenges to be addressed. Preliminary information about the terminologies, data models, technologies, and dataset used in this thesis is presented in Chapter 2.

In Chapter 3, we introduce [MoMatch](#), an approach for matching ontologies in multiple languages. [MoMatch](#) uses sequence-based and set-based string similarity metrics in conjunction with machine translation to align ontologies that are represented in distinct natural languages. This solution offers a thorough approach to matching ontologies, whether they are in different languages or in the same language.

The use of [MoMatch](#) in scholarly communication and the biomedical fields serves as an example of its practical applicability and emphasizes its adaptability across various domains. The experiments carried out provide an in-depth analysis of 13 similarity measures across different languages. We discover that string similarity metrics considerably affect matching precision and recall. When selecting a similarity metric, it is crucial to take into account the characteristics of ontologies and determine if precision or recall have greater significance. In order to select the most suitable similarity measure for the matching process, we arrange the 13 similarity measures into two categories based on their precision and recall. Besides, it is worth noting that the results for precision, recall, and F-measure obtained from the combination of various similarity measures and threshold values remain highly correlated, regardless of the language pairs being used. [MoMatch](#) outperforms the majority

of state-of-the-art techniques in terms of precision, recall, and F-measure by utilizing Jaccard and Levenshtein similarities.

In Chapter 4, we explore enriching ontologies using unstructured sources in several languages. We highlight the difficulties that arise from the enormous amount of unstructured data available in various natural languages.

This chapter presents two publications, both of which contribute to the overall objective of enriching ontology in an effective manner. The initial publication presents a domain-dependent approach that relies on a multi-agent architecture. In contrast, the later publication, **CLOE**, expands upon this framework to provide a solution that is not limited to any specific domain (a domain-independent approach). **CLOE**'s multi-agent architecture addresses various issues, including the use of different text pre-processing techniques, cross-lingual concept matching, inter-agent communication, and the removal of redundant text.

The five phases of **CLOE**, from text pre-processing to ontology verification and validation, establish a comprehensive framework for enriching ontologies from multiple domains in different languages. The effective use of **CLOE** in enhancing an Information Technology ontology through the incorporation of multilingual text demonstrates its practical utility.

The effectiveness of **CLOE** is validated using evaluation approaches such as a satisfaction questionnaire and gold standard-based assessment. **CLOE** demonstrates superior performance compared to four state-of-the-art techniques in terms of recall and F-measure, thus emphasizing its value in the context of ontology enrichment.

In Chapter 5, we address how cross-lingual ontology enrichment can enhance semantic interoperability among different ontologies in various natural languages. We present **OECM**, a fully automated approach for enriching ontologies by utilizing information from other ontologies in diverse languages, producing a consistent, multilingual ontology. **OECM** aims to find complementary information from the source ontology to enrich the target ontology. The iterative triple retrieval process in **OECM** allows relevant information to be identified from the source ontology and to determine its appropriate location in the target ontology hierarchy.

We show the usability of **OECM** by using German and Arabic ontologies to enrich a scholarly communication ontology that describes scientific events. The output of the enrichment process is a consistent ontology in English, German, and Arabic.

The innovative function in **OECM** to choose the optimal translation from a set of choices using lexical and semantic similarity measures has greatly enhanced the results of the matching process. Consequently, **OECM** surpasses the state-of-the-art techniques in terms of precision, recall, and F-measure.

Despite the success of **OECM**, it has certain limitations. These limitations include a narrow emphasis on classes rather than properties, as well as the requirement for expert input in order to develop reference ontologies for evaluating the enrichment process's quality. This might be unfeasible for large-scale ontologies. We address these limitations in the next chapters.

Our research revealed that some ontologies have overlapping and redundant information because they are designed independently to model the already existing domains. Therefore, in Chapter 6, we have put forward **MULON** as a viable resolution to this issue. The three components of **MULON** (preparation, merging, and assessment) effectively collaborate to generate a consistent merged ontology. The use of machine translation techniques, lexical similarity, and semantic similarity improves the merging results.

MULON's effectiveness in tackling multilingual difficulties across several domains is demonstrated

by its practical applications in fiscal data, food, and scholarly communication. The versatility of **MULON** is highlighted, emphasizing its adaptation and modification capabilities that extend beyond the discussed domains. **MULON**'s performance across many languages is shown through the comprehensive evaluation process. **MULON** demonstrates superior performance compared to three state-of-the-art approaches in terms of precision, recall, and F-measure. The tool's adaptability and simplicity in customization make it a great asset for researchers in diverse fields.

In Chapter 7, we describe different factors that greatly influence the ontology enrichment process. We introduce **QASO**, which includes 17 quality metrics that assess these factors. Among these 17 quality metrics, seven are used to evaluate the quality of ontologies at the schema level, seven to evaluate the quality of the matching process, and three to evaluate the enriched/merged ontology.

The quality of ontologies have been assessed in terms of several dimensions such as richness. **QASO** incorporates metrics to evaluate the matching process, taking into account both the completeness and correctness of the matching result. Match coverage and match ratio metrics provide cost-effective option to improve efficiency, eliminating the requirement for a reference alignment. Three metrics are presented in **QASO** to assess the quality of the enriched/merged ontology. These metrics assess the extent of coverage and compactness, guaranteeing the preservation of information during merging.

The research findings and analysis outlined in this thesis provide a foundation for future investigation and advancement, representing a notable advancement in the pursuit of achieving semantic coherence within the continuously expanding multilingual Semantic Web.

8.2 Future Work

There are several directions to further explore in the context of cross-lingual ontology enrichment and assessment. In this following, we outline the possible directions that can be considered for the further improvement of the various contributions offered by this thesis.

- *Including instances in the enrichment process.* In this thesis, we focus solely on the ontology's schema (i.e., classes and properties) for enrichment. Ontologies enriched with instances can provide more detailed and precise information. However, ensuring the accuracy and consistency of these instances poses a significant challenge. Therefore, further developments are required for the proposed cross-lingual matching and enrichment techniques to incorporate ontology instances. Moreover, it is crucial to develop quality metrics aimed at guaranteeing the accuracy and consistency of instances.
- *Designing scalable ontology enrichment techniques.* Large-scale ontologies require scalable and efficient frameworks to process, integrate, and analyze vast amounts of data. This is critical for ontology enrichment because of the complexity and volume of information involved. Big data frameworks such as Apache Spark¹ can be utilized to handle massive datasets distributed across clusters of computers [85]. In big data frameworks, the distribution computing paradigms allow for parallel execution of tasks across multiple nodes in a cluster. This parallelism accelerates computation-intensive tasks in ontology enrichment, such as ontology matching.

¹ <https://spark.apache.org/>

- *Utilizing Large Language Models.* LLMs in NLP have seen a significant rise and rapid development, making them a valuable tool for investigating their potential in the cross-lingual ontology enrichment process. The effectiveness of LLMs depends on the quality and diversity of the training data [67]. It is crucial to train models on comprehensive and representative multilingual datasets. Giglou et al., [86] investigate the use of 11 LLMs for ontology learning. They found that LLMs are not adequately suited for ontology learning, which requires advanced reasoning abilities and domain expertise. However, with proper adjustments, they could potentially serve as helpful aides in overcoming the bottleneck of knowledge acquisition for ontology construction [86]. They also investigate the use of LLMs in the monolingual ontology matching task [70]. They discovered that employing LLMs in conjunction with retriever models such as TFIDF [87] can outperform traditional ontology matching systems. These findings highlight the considerable potential of LLMs, which opens up opportunities for additional investigation into the process of cross-lingual ontology matching.

Bibliography

- [1] T. Berners-Lee, J. Hendler and O. Lassila, *Web Semantic*, Scientific American **284** (2001) 34 (cit. on pp. [1](#), [11](#)).
- [2] G. O. Consortium, *The gene ontology resource: 20 years and still GOing strong*, Nucleic acids research **47** (2019) D330 (cit. on pp. [1](#), [14](#)).
- [3] S. El Kadiri et al., “Ontology in engineering applications”, *Formal Ontologies Meet Industry: 7th International Workshop, FOMI 2015, Berlin, Germany, August 5, 2015, Proceedings 7*, Springer, 2015 126 (cit. on pp. [1](#), [14](#)).
- [4] E. Ong et al., *Modelling kidney disease using ontology: insights from the Kidney Precision Medicine Project*, Nature Reviews Nephrology **16** (2020) 686 (cit. on p. [1](#)).
- [5] E. Montiel-Ponsoda, G. A. De Cea, A. Gómez-Pérez and W. Peters, *Enriching ontologies with multilingual information*, Natural language engineering **17** (2011) 283 (cit. on pp. [1](#), [17](#)).
- [6] C. Soria et al., *Exploring interoperability of language resources: The case of cross-lingual semi-automatic enrichment of wordnets*, Language Resources and Evaluation **43** (2009) 87 (cit. on pp. [1](#), [17](#)).
- [7] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara and E. Zavitsanos, “Ontology population and enrichment: State of the art”, *Knowledge-driven multimedia information extraction and ontology evolution*, Springer-Verlag, 2011 134 (cit. on pp. [1](#), [14](#), [17](#), [20](#), [24](#), [28](#), [36](#)).
- [8] M. Keet, *An introduction to ontology engineering*, vol. 1.5, College Publications, 2020 (cit. on pp. [2](#), [6](#), [32](#)).
- [9] F. Gillis-Webber and C. M. Keet, *A Survey of Multilingual OWL Ontologies in BioPortal*, (2022) (cit. on p. [2](#)).
- [10] H. Sun et al., *Unsupervised neural machine translation with cross-lingual language representation agreement*, IEEE/ACM Transactions on Audio, Speech, and Language Processing **28** (2020) 1170 (cit. on pp. [2](#), [15](#)).
- [11] S. Ren, Y. Wu, S. Liu, M. Zhou and S. Ma, *Explicit cross-lingual pre-training for unsupervised machine translation*, arXiv preprint arXiv:1909.00180 (2019) (cit. on p. [2](#)).

- [12] J. Monti, M. Monteleone, M. P. Di Buono and F. Marano, “Cross-lingual information retrieval and semantic interoperability for cultural heritage repositories”, *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, 2013 483 (cit. on p. 2).
- [13] C. Trojahn, B. Fu, O. Zamazal and D. Ritze, “State-of-the-art in multilingual and cross-lingual ontology matching”, *Towards the Multilingual Semantic Web*, Springer, 2014 119 (cit. on pp. 2, 3, 15, 20, 28, 32).
- [14] J. Gracia et al., *Challenges for the multilingual web of data*, *Web Semantics: Science, Services and Agents on the World Wide Web* **11** (2012) 63 (cit. on pp. 2, 3, 20, 24, 32).
- [15] T. Ivanova, “Cross-lingual and multilingual ontology mapping-survey”, *Proceedings of the 19th International Conference on Computer Systems and technologies*, 2018 50 (cit. on pp. 2, 3).
- [16] M. A. Helou and M. Palmonari, “Cross-lingual lexical matching with word translation and local similarity optimization”, *Proceedings of the 11th International Conference on Semantic Systems*, 2015 97 (cit. on p. 2).
- [17] F. A. Musyaffa, M.-E. Vidal, F. Orlandi, J. Lehmann and H. Jabeen, *IOTA: Interlinking of heterogeneous multilingual open fiscal DaTA*, *Expert Systems with Applications* **147** (2020) 113135 (cit. on p. 2).
- [18] B. Fu, R. Brennan and D. O’Sullivan, *A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes*, *Web Semantics: Science, Services and Agents on the World Wide Web* **15** (2012) 15 (cit. on p. 2).
- [19] S. Hertling and H. Paulheim, *WikiMatch: using wikipedia for ontology matching*, *Ontology Matching* **946** (2012) (cit. on p. 3).
- [20] C. Trojahn, P. Quaresma and R. Vieira, *A framework for multilingual ontology mapping*, (2008) (cit. on p. 3).
- [21] A. Khiat, E. A. Ouhiba, M. A. Belfedhal and C. E. Zoua, “SimCat results for OAEI 2016.”, *OM@ ISWC*, 2016 217 (cit. on p. 3).
- [22] A. Khiat, *Crolom: Cross-lingual ontology matching system*, *Ontology Matching* **146** (2016) (cit. on p. 3).
- [23] X. Liu, Q. Tong, X. Liu and Z. Qin, *Ontology matching: State of the art, future challenges, and thinking based on utilized information*, *IEEE Access* **9** (2021) 91235 (cit. on pp. 3, 4).
- [24] T. Degani and N. Tokowicz, *Semantic ambiguity within and across languages: An integrative review*, *Quarterly Journal of Experimental Psychology* **63** (2010) 1266 (cit. on p. 3).
- [25] G. Bella, F. Giunchiglia and F. McNeill, *Language and domain aware lightweight ontology matching*, *Journal of Web Semantics* **43** (2017) 1 (cit. on p. 3).

-
- [26] B. Fu, R. Brennan and D. O'sullivan, *Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web.*, MSW **571** (2010) 13 (cit. on p. 4).
- [27] A. M. Khattak, R. Batool, Z. Pervez, A. M. Khan and S. Lee, *Ontology Evolution and Challenges.*, J. Inf. Sci. Eng. **29** (2013) 851 (cit. on p. 4).
- [28] I. Osman, S. B. Yahia and G. Diallo, *Ontology integration: approaches and challenging issues*, Information Fusion **71** (2021) 38 (cit. on p. 4).
- [29] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth and B. Aleman-Meza, *OntoQA: Metric-based ontology quality analysis*, (2005) (cit. on pp. 4, 6, 36).
- [30] S. Mc Gurk, C. Abela and J. Debatista, "Towards Ontology Quality Assessment.", *MEPDaW/LDQ@ ESWC*, 2017 94 (cit. on pp. 4, 6, 36).
- [31] S. Raunich and E. Rahm, "Towards a benchmark for ontology merging", *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, 2012 124 (cit. on pp. 4, 6, 17, 36).
- [32] R. S. I. Wilson, J. S. Goonetillake, A. Ginige and W. A. Indika, "Ontology Quality Evaluation Methodology", *International Conference on Computational Science and Its Applications*, Springer, 2022 509 (cit. on pp. 4, 36).
- [33] R. Wilson, J. S. Goonetillake, W. Indika and A. Ginige, "Analysis of Ontology Quality Dimensions, Criteria and Metrics", *International Conference on Computational Science and Its Applications*, Springer, 2021 320 (cit. on pp. 4, 36).
- [34] X. Xue, Y. Wang, W. Hao and J. Hou, *Optimizing ontology alignments through NSGA-II without using reference alignment*, Computing and Informatics **33** (2014) 857 (cit. on pp. 4, 36).
- [35] A. Doan, A. Halevy and Z. Ives, *Principles of data integration*, Elsevier, 2012 (cit. on pp. 5, 16).
- [36] S. Ibrahim, S. Fathalla, J. Lehmann and H. Jabeen, *Toward the Multilingual Semantic Web: Multilingual Ontology Matching and Assessment*, IEEE Access **11** (2023) 8581 (cit. on pp. 6, 8).
- [37] M. Ali, S. Fathalla, S. Ibrahim, M. Kholief and Y. F. Hassan, *CLOE: a cross-lingual ontology enrichment using multi-agent architecture*, Enterprise Information Systems (2019) 1 (cit. on pp. 6, 25).
- [38] M. Ali, S. Fathalla, S. Ibrahim, M. Kholief and Y. Hassan, *Cross-lingual ontology enrichment based on multi-agent architecture*, Procedia Computer Science **137** (2018) 127 (cit. on p. 6).
- [39] S. Ibrahim, S. Fathalla, H. S. Yazdi, J. Lehmann and H. Jabeen, "OECM: a cross-lingual approach for ontology enrichment", *The Semantic Web: ESWC 2019 Satellite Events: ESWC 2019 Satellite Events, Portorož, Slovenia, June 2–6, 2019, Revised Selected Papers 16*, Springer, 2019 100 (cit. on p. 7).

- [40] S. Ibrahim, S. Fathalla, H. Shariat Yazdi, J. Lehmann and H. Jabeen, “From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment”, *Semantic Systems. The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9–12, 2019, Proceedings 15*, Springer International Publishing, 2019 215 (cit. on pp. 7, 32).
- [41] S. Ibrahim, S. Fathalla, J. Lehmann and H. Jabeen, “Multilingual Ontology Merging Using Cross-lingual Matching”, *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, IEEE, 2020 113 (cit. on pp. 7, 8).
- [42] S. Ibrahim, S. Fathalla, H. S. Yazdi, J. Lehmann and H. Jabeen, “OECM: A Cross-lingual Approach for Ontology Enrichment”, *European Semantic Web Conference*, Springer, 2019 In Press (cit. on p. 9).
- [43] S. Bratt, *Semantic web and other W3C technologies to watch*, Talks at W3C, January (2007) (cit. on p. 12).
- [44] W. W. W. Consortium et al., *RDF 1.1 Primer*, (2014) (cit. on pp. 12, 14).
- [45] G. Antoniou and F. Van Harmelen, *A semantic web primer*, MIT press, 2004 (cit. on pp. 12, 13, 15).
- [46] M. Florence, *Building a multilingual ontology for education domain using monto method*, *Computer Science and Information Technologies* **1** (2020) 47 (cit. on pp. 14, 20, 36).
- [47] N. Guarino, D. Oberle and S. Staab, *What is an ontology?*, *Handbook on ontologies* (2009) 1 (cit. on p. 14).
- [48] A. Guedes, F. Baiao and K. Revoredo, “On the identification and representation of ontology correspondence antipatterns”, *Proceedings of the 8th International Workshop on Modular Ontologies (WoMO14)*, vol. 1248, CEUR-WS, 2014 (cit. on p. 14).
- [49] S. C. Brandt et al., *An ontology-based approach to knowledge management in design processes*, *Computers & Chemical Engineering* **32** (2008) 320 (cit. on p. 14).
- [50] A. Buccella, A. Cechich and N. Rodríguez Brisaboa, *An ontology approach to data integration*, *Journal of Computer Science & Technology* **3** (2003) (cit. on p. 14).
- [51] M. Fernández et al., *Semantically enhanced information retrieval: An ontology-based approach*, *Journal of Web Semantics* **9** (2011) 434 (cit. on p. 14).
- [52] O. Bodenreider, *The unified medical language system (UMLS): integrating biomedical terminology*, *Nucleic acids research* **32** (2004) D267 (cit. on p. 14).
- [53] P. G. Baker et al., *An ontology for bioinformatics applications.*, *Bioinformatics (Oxford, England)* **15** (1999) 510 (cit. on p. 14).
- [54] F. W. Geels, *Ontologies, socio-technical transitions (to sustainability), and the multi-level perspective*, *Research policy* **39** (2010) 495 (cit. on p. 14).

-
- [55] W. O. W. Group et al., *OWL 2 web ontology language document overview*, <http://www.w3.org/TR/owl2-overview/> (2009) (cit. on p. 14).
 - [56] H. Wu and A. Yamaguchi, *Semantic Web technologies for the big data in life sciences*, *Bioscience trends* **8** (2014) 192 (cit. on p. 15).
 - [57] L. Otero-Cerdeira, F. J. Rodríguez-Martínez and A. Gómez-Rodríguez, *Ontology matching: A literature review*, *Expert Systems with Applications* **42** (2015) 949 (cit. on pp. 15, 16).
 - [58] C. Trojahn, R. Vieira, D. Schmidt, A. Pease and G. Guizzardi, *Foundational ontologies meet ontology matching: A survey*, *Semantic Web* **13** (2022) 685 (cit. on p. 15).
 - [59] S. Ahmetaj et al., “Ontology-enriched query answering on relational databases”, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 17, 2021 15247 (cit. on p. 15).
 - [60] F. Ocker, B. Vogel-Heuser and C. J. Paredis, *A framework for merging ontologies in the context of smart factories*, *Computers in Industry* **135** (2022) 103571 (cit. on p. 15).
 - [61] A. Solimando, E. Jiménez-Ruiz and C. Pinkel, “Evaluating Ontology Alignment Systems in Query Answering Tasks.”, *ISWC (Posters & Demos)*, 2014 301 (cit. on p. 15).
 - [62] J. Euzenat and P. Shvaiko, *Ontology matching*, 2nd, Heidelberg (DE): Springer-Verlag, 2013 (cit. on pp. 15, 16).
 - [63] C. Fellbaum, *WordNet: An electronic lexical database*, MIT press, 1998 (cit. on p. 16).
 - [64] J. Ferlež and M. Gams, *Shortest-path semantic distance measure in wordnet v2. 0*, *Information Society in 2004* **381** (2004) (cit. on p. 17).
 - [65] C. Meilicke et al., *MultiFarm: A benchmark for multilingual ontology matching*, *Journal of web semantics* **15** (2012) 62 (cit. on pp. 18, 20, 29, 32).
 - [66] S. Fathalla, S. Vahdati, C. Lange and S. Auer, “SEO: A scientific events data model”, *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II* **18**, Springer, 2019 79 (cit. on pp. 20, 29).
 - [67] J. Vela and J. Gracia, “Cross-lingual ontology matching with CIDER-LM: results for OAEI 2022.”, *OM@ ISWC*, 2022 158 (cit. on pp. 21, 42).
 - [68] D. Faria et al., *Results for Matcha and Matcha-DL in OAEI 2023*, (2023) (cit. on p. 21).
 - [69] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, arXiv preprint arXiv:1908.10084 (2019) (cit. on p. 21).
 - [70] H. B. Giglou, J. D’Souza and S. Auer, *LLMs4OM: Matching Ontologies with Large Language Models*, arXiv preprint arXiv:2404.10317 (2024) (cit. on pp. 21, 42).
 - [71] A. Sharma, A. Patel and S. Jain, *Lsmatch and lsmatch-multilingual results for oaei*, **3324** (2022) 191 (cit. on p. 21).

- [72] A. Sharma and S. Jain, “LSMatch and LSMatch-multilingual results for OAEI 2023.”, *Proceedings of the 18th International Workshop on Ontology Matching (OM 2023)*, vol. 3591, 2023 159 (cit. on p. 21).
- [73] E. Jiménez-Ruiz, “Logmap family participation in the oaei 2022”, *Proceedings of the 17th International Workshop on Ontology Matching (OM 2022)*, vol. 3324, CEUR-WS, 2022 188 (cit. on p. 21).
- [74] E. Jiménez-Ruiz, “Logmap family participation in the oaei 2023”, *Proceedings of the 18th International Workshop on Ontology Matching (OM 2023)*, vol. 3591, CEUR-WS, 2023 157 (cit. on p. 21).
- [75] V. Iyer, L. Mohan, M. Bhatia and Y. R. Reddy, “A survey on ontology enrichment from text”, *Proceedings of the 16th International Conference on Natural Language Processing*, 2019 95 (cit. on p. 24).
- [76] D. Spohr, L. Hollink and P. Cimiano, *A machine learning approach to multilingual and cross-lingual ontology matching*, The Semantic Web–ISWC 2011 (2011) 665 (cit. on p. 24).
- [77] L. M. Sanagavarapu, V. Iyer and R. Reddy, *A deep learning approach for ontology enrichment from unstructured text*, Cybersecurity and High-Performance Computing Environments: Integrated Innovations, Practices, and Applications (2022) 261 (cit. on p. 24).
- [78] M. Espinoza, A. Gómez-Pérez and E. Mena, “Enriching an ontology with multilingual information”, *The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings 5*, Springer, 2008 333 (cit. on p. 28).
- [79] D. W. Embley, S. W. Liddle, D. W. Lonsdale and Y. Tijerino, “Multilingual ontologies for cross-language information extraction and semantic search”, *Conceptual Modeling–ER 2011: 30th International Conference, ER 2011, Brussels, Belgium, October 31–November 3, 2011. Proceedings 30*, Springer, 2011 147 (cit. on p. 28).
- [80] S. Raunich and E. Rahm, “ATOM: Automatic target-driven ontology merging”, *2011 IEEE 27th international conference on data engineering*, IEEE, 2011 1276 (cit. on p. 32).
- [81] M. Fahad, *Merging of axiomatic definitions of concepts in the complex OWL ontologies*, Artificial Intelligence Review 47 (2017) 181 (cit. on p. 32).
- [82] N. F. Noy, M. A. Musen et al., “Algorithm and tool for automated ontology merging and alignment”, *Proceedings of the 17th National Conference on Artificial Intelligence. Available as SMI technical report SMI-2000-0831*, vol. 115, sn, 2000 (cit. on p. 32).
- [83] J. de Bruijn, F. Martin-Recuerda, D. Manov and M. Ehrig, *D4. 2.1 state-of-the-art-survey on ontology merging and aligning v1*, SEKT Project deliverable D 4 (2004) 2 (cit. on p. 32).

-
- [84] M. McDaniel and V. C. Storey,
Evaluating domain ontologies: clarification, classification, and challenges,
ACM Computing Surveys (CSUR) **52** (2019) 1 (cit. on p. 36).
- [85] S. Salloum, R. Dautov, X. Chen, P. X. Peng and J. Z. Huang,
Big data analytics on Apache Spark,
International Journal of Data Science and Analytics **1** (2016) 145 (cit. on p. 41).
- [86] H. Babaei Giglou, J. D’Souza and S. Auer,
“LLMs4OL: Large language models for ontology learning”,
International Semantic Web Conference, Springer, 2023 408 (cit. on p. 42).
- [87] “TF-IDF”, *Encyclopedia of Machine Learning*, ed. by C. Sammut and G. I. Webb,
Boston, MA: Springer US, 2010 986, ISBN: 978-0-387-30164-8,
URL: https://doi.org/10.1007/978-0-387-30164-8_832 (cit. on p. 42).

**Toward the multilingual Semantic Web:
Multilingual ontology matching and assessment**

Received 22 November 2022, accepted 13 January 2023, date of publication 23 January 2023, date of current version 27 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3238871

RESEARCH ARTICLE

Toward the Multilingual Semantic Web: Multilingual Ontology Matching and Assessment

SHIMAA IBRAHIM^{1,2}, SAID FATHALLA³, JENS LEHMANN^{4,5}, AND HAJIRA JABEEN⁶

¹Institute of Computer Science, University of Bonn, 53115 Bonn, Germany

²Institute of Graduate Studies and Research, University of Alexandria, Alexandria 5422004, Egypt

³Faculty of Science, University of Alexandria, Alexandria 5413213, Egypt

⁴Institute for Applied Informatics Association e.V. (InfAI), 01069 Dresden, Germany

⁵Amazon, 01097 Dresden, Germany (work done outside of Amazon)

⁶Big Data Analytics, KTS, GESIS-Leibniz Institute for the Social Sciences, 50667 Cologne, Germany

Corresponding author: Shima Ibrahim (ibrahim@cs.uni-bonn.de)

ABSTRACT The amount of multilingual data on the Web proliferates; therefore, developing ontologies in various natural languages is attracting considerable attention. In order to achieve semantic interoperability for the multilingual Web, cross-lingual ontology matching techniques are highly required. This paper proposes a Multilingual Ontology Matching (MoMatch) approach for matching ontologies in different natural languages. MoMatch uses machine translation and various string similarity techniques to identify correspondences across different ontologies. Furthermore, we propose a Quality Assessment Suite for Ontologies (QASO) that comprises 14 metrics, out of which seven metrics are used to assess the quality of the matching process and seven metrics are used to evaluate the quality of the ontology. We present an in-depth comparison of different string similarity techniques across various languages to get the most effective similarity measure(s) between multilingual terms. To illustrate the applicability of our approach and how it can be used in different domains, we present two use cases. MoMatch has been implemented using Scala and Apache Spark under an open-source license. We have compared our results with the results from the Ontology Alignment Evaluation Initiative (OAEI 2020). MoMatch has achieved significantly high precision, recall, and F-measure compared to five state-of-the-art matching approaches.

INDEX TERMS Cross-lingual matching, knowledge management, multilingual web, ontology engineering, ontology matching, string similarity.

I. INTRODUCTION

Ontologies are being widely used in various fields of science other than computer science, including Biology [1], Engineering [2], and Medicine [3]. With the rapid expansion of multilingual data on the Semantic Web, more ontologies have become available in different natural languages. According to Linked Open Vocabularies (LOV),¹ English is by far the most prominent language, i.e., most ontologies in the Semantic Web are in English; however, many ontologies in other Indo-European languages (e.g., German) also exist. Specifically, out of a total of 782 vocabularies found in LOV, 584 are in

English, 65 in French, and 41 in German. Few ontologies registered in LOV are in non-Indo-European languages, e.g., ten are in Chinese, and seven are in Arabic. In fact, monolingual ontologies that include labels or local names presented in a particular language are not understandable to people who speak other languages, hindering semantic interoperability. Therefore, to improve semantic interoperability between monolingual ontologies, innovative algorithms need to be developed to digest multilingual data and match ontologies in different natural languages [4]. Identifying correspondences between ontologies in different natural languages is called *Cross-lingual ontology matching*.

Cross-lingual ontology matching plays a crucial role in various research areas involving ontology enrichment/merging, data warehouses, linked data, peer-to-peer information

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed ¹.

¹<https://lov.linkeddata.es/dataset/lov/vocabs>

sharing, and web service composition [5]. The main idea is that all ontologies should be mapped to a core set of related ontologies in advance when a system is attempting to automatically search for a specific piece of information from or exchange knowledge with other systems.

To date, there is no clear winner in solving the cross-lingual matching problem. Further research is required to advance cross-lingual ontology matching techniques to obtain better results compared to monolingual approaches. Therefore, we can conclude that there is a need to create a flexible framework for matching ontologies in different languages without relying on a specific language.

In this paper, we propose a Multilingual Ontology Matching (MoMatch) approach for matching ontologies in different natural languages. MoMatch uses different string similarity techniques and machine translation to match classes and properties across ontologies. MoMatch comprises four phases: 1) *resource extraction*: extracting all resources (classes and properties) from the input ontologies, 2) *pre-processing*: preparing the extracted resource for the subsequent phases, 3) *translation*: translating the extracted resources, and 4) *matching*: identifying potential matches between the input ontologies.

The quality of the matching process greatly depends on the quality of the input ontologies to be matched. Existing approaches defined in the literature have addressed the quality evaluation of ontology from different dimensions, such as syntactic, semantic, pragmatic, social, structural, functional, content, schema, and usage [6], [7], [8], [9]. Each dimension has its own set of criteria, and each criterion has its own set of metrics to measure the characteristics of an ontology that can be represented formally [10]. A vast number of metrics have been addressed from different aspects. As a result, determining what quality metrics that affect the quality of the ontology matching process is complex. In our previous work MULON [11], an approach for merging monolingual ontologies in different natural languages, we proposed seven quality metrics to assess the quality of ontologies on the schema level. In this paper, we extend the quality metrics, by proposing the Quality Assessment Suite for Ontologies (QASO). QASO comprises 14 quality metrics for assessing the quality of the matching process in addition to the quality of the input ontologies' schema that influences the matching process (more details in subsection V-A). Out of the 14 metrics, seven are used to assess the quality of the input ontologies and seven to evaluate the quality of the matching process. We address the following research questions: *RQ1) What is the most effective string similarity measure(s) for matching multilingual terms across different ontologies? RQ2) How strong is the performance of the similarity measure when the language of the input ontologies is changed? and RQ3) How can the quality of the ontology matching process be measured using a set of metrics?*

The contributions of this work can be summarized in the following points:

- MoMatch can efficiently match ontologies in any natural language compared to state-of-the-art (cf. subsection VII-C),
- Ten language pairs, including Indo and Non-Indo European languages, have been tested in this approach,
- We present a comparative analysis of 13 different string similarity measures,
- A metric suit (QASO) has been designed for assessing the quality of the matching process, which also can be used to assess the quality of the ontology,
- Two use cases demonstrating the usability of the MoMatch in matching multilingual as well as monolingual ontologies from different domains are presented, and
- MoMatch empirically showed significantly better performance when compared to five state-of-the-art approaches.

We believe that MoMatch is a crucial step towards realizing the multilingual Semantic Web as it supports the integration of ontologies in different languages. MoMatch² and QASO³ are available in two separate public repositories in GitHub, in which the source code is documented, describing each configurable parameter and function.

The rest of this paper is organized as follows: In section II, we provide background information about our research topic as well as formal definitions of ontology and the ontology matching process. In section III, we provide an overview of related work. The proposed approach is described in detail in section IV. The quality metric suite is presented in section V. Two use cases are presented in section VI to demonstrate possible applications of MoMatch. The results of experiments and evaluations are presented in section VII. Finally, in section VIII, we summarize the main conclusions with an outline of future work.

II. PRELIMINARIES AND DEFINITIONS

Ontology is considered one of the significant cornerstones of representing information more meaningfully, providing machine-understandable semantics of knowledge. It is a container for capturing semantic information of a particular domain, which allows sharing and reusing knowledge in this domain [17], [18].

Definition 1 (Ontology): Formally, an ontology can be represented as a tuple of five components [19]: $O = \langle C, R, P, I, X \rangle$, where C is the set of classes/concepts, R is the set of relationships between classes (object properties), P is the set of data properties (a specific type of relation whose domain is a class and the range is a data type), I is the set of class instances (concrete objects), and X is the set of axioms and rules used for checking and verifying ontology consistency and new knowledge inference.

²<https://github.com/SmartDataAnalytics/MoMatch>

³<https://github.com/SmartDataAnalytics/QASO>

TABLE 1. An overview of eight sequence-based string similarity measures utilized in MoMatch, as well as their formulas.

Name	Description	Formula
Hamming distance	Measures the minimum number of substitutions required to change one string x into the other y [12].	The hamming distance normalized by the length of the longest string is given by [5]: $Hamm(x, y) = \frac{(\sum_{i=1}^{\min(x , y)} x[i] \neq y[i]) + x - y }{\max(x , y)} \quad (1)$
Jaro	Counts the number of common characters c between two strings x, y and the number of transpositions t required to make these common characters have the same sequence [5], [13].	$Jaro(x, y) = \frac{1}{3} \times (\frac{c}{ x } + \frac{c}{ y } + \frac{c - \frac{t}{2}}{c}) \quad (2)$
Jaro Winkler	Captures cases where two strings x, y have a low Jaro score but share a prefix. It enhances Jaro by including the length l of the longest common prefix and a weight w given to that prefix [5], [13].	$JW(x, y) = (1 - l \times w) \times Jaro(x, y) + l \times w \quad (3)$
Levenshtein	Counts the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string x into the other y [5], [14]. Also known as Edit Distance.	$Lev(x, y) = 1 - \frac{\delta(x, y)}{\max(\text{length}(x), \text{length}(y))} \quad (4)$ where $\delta(x, y)$ is the edit distance between strings x, y .
Ratio	Calculates the ratio of the number of matching characters m in the two strings x, y to the total number of characters t in both strings [15].	$R = \text{round}(2.0 \times \frac{m}{t} \times 100) \quad (5)$
Partial ratio	Finds the similarity between the shorter string x and every sub-string y_i of length m of the longer string y , and returns the maximum of that similarity measures [15].	$PR = \max(\sum_{i=1}^n \text{ratio}(x, y_i)) \quad (6)$ where n is the number of sub-strings.
Token sort	Splits the two strings x, y into tokens, sorting the tokens, and calculates the ratio similarity of the transformed strings [16].	
Partial token sort	Splits the two strings x, y into tokens, sorting the tokens, and calculates the partial ratio similarity of the transformed strings [16].	

Ontology matching is a complicated procedure to bridge the semantic gap between multiple representations of the same domain [20]. It identifies correspondences between the entities (classes and properties) of two or more ontologies that satisfy specific conditions.

Definition 2 (Ontology Matching): “a function f which, from a pair of ontologies to match O_1 and O_2 , an input alignment A , a set of parameters p and a set of resources r , returns an alignment A' between these ontologies: $A' = f(O_1, O_2, A, p, r)$ ” [5].

Definition 3 (Alignment): the alignment between two ontologies, O_1 and O_2 , is a set of correspondences between pairs of entities belonging to O_1 and O_2 , respectively [5]. It is the output of the matching process.

Definition 4 (Correspondence): given two ontologies O_1 and O_2 , the correspondence between two entities e_{11} and e_{12} , where $e_{11} \in O_1$ and $e_{12} \in O_2$, is the relation that produces according to a matching algorithm between e_{11} and e_{12} [5]. It can be represented by the triple: $\langle e_{11}, e_{12}, r \rangle$, where r is the relation between the two entities e_{11} and e_{12} .

Definition 5 (Monolingual ontology matching): the process of matching ontologies in the same natural language, i.e., $L_1 = L_2$, where L_1 is the natural language of O_1 , and L_2 is the natural language of O_2 [4].

The growing amount of multilingual data on the Web and the resulting development of ontologies in different natural languages has increased the demand for cross-lingual ontology matching.

Definition 6 (Cross-lingual ontology matching): the process of matching ontologies in different natural languages, i.e., $L_1 \neq L_2$ [4].

Ontology matching techniques are used to identify the correspondence between two ontologies' entities, including the analysis of subsumption between classes and the similarity between the entity names. Different ontology matching techniques have been proposed, which can be classified as [20]: a) Element-level matching techniques: identify the correspondences by analyzing entities in the ontologies in isolation, ignoring the structure, i.e., ignoring their relations with other entities or instances. b) Structure-level matching techniques: identify the correspondences by analyzing the structure of entities in the ontology, i.e., considering the relations between entities and their instances. In this paper, we propose an element-level matching technique.

String-based techniques (as element-level techniques) are utilized to match names and name descriptions of ontology entities [5]. These techniques consider strings as alphabetical sequences of letters. The more similar the strings are, the

TABLE 2. An overview of five set-based string similarity measures utilized in MoMatch, as well as their formulas. The input is two sets X and Y , each containing a set of tokens for the two strings x and y , respectively. All the resulting values are in the range of $[0,1]$.

Name	Description	Formula
Cosine	Computes a variant of cosine measure known as Ochiai coefficient, which calculates the intersection between two sets X, Y ⁴ [16].	$Cosine(X, Y) = \frac{ X \cap Y }{\sqrt{ X \cdot Y }} \quad (7)$
Dice	Twice the common terms (intersection) between two strings over the total number of terms in both strings ⁴ [16].	$Dice(X, Y) = \frac{2 \times X \cap Y }{ X + Y } \quad (8)$
Jaccard	The size of the common terms between two strings divided by the size of the union of all terms in the two strings ⁴ [16].	$Jaccard(X, Y) = \frac{ X \cap Y }{ X \cup Y } \quad (9)$
Overlap coefficient	Measures the overlap between two sets by dividing the intersection between them by the smaller sizes of the two sets ⁴ [21]. If one string is a subset of another, it is considered a full match.	$OC(X, Y) = \frac{ X \cap Y }{\min(X , Y)} \quad (10)$
Tversky Index	The size of the common tokens between two sets divided by: the sum of intersection between sets, the number of items only available on the first token set multiplied by a coefficient α , and the number of elements only available in the second token sets multiplied by a coefficient β [22].	$T(X, Y) = \frac{ X \cap Y }{ X \cap Y + \alpha X - Y + \beta Y - X } \quad (11)$ where α and $\beta > 0$.

more likely they represent the same concepts. Inspired by `py_stringmatching` library⁴ and Doan et al. [13] string-based techniques are categorized as follows:

- **Sequence-based:** Input strings are considered as a sequence of characters. Such as Hamming distance, Jaro, Jaro Winkler, Levenshtein, Partial ratio, Partial token sort, Ratio, and Token sort.
- **Set-based:** Input strings are considered as sets or multi-sets of tokens. Such as Cosine, Dice, Jaccard, Overlap coefficient, and Tversky Index.
- **Bag-based:** Input strings are considered as bags, i.e., collections of tokens, in which a token appears multiple times. Such as TF/IDF.
- **Hybrid-based:** It combines sequence-based and token-based similarity measures. Such as generalized Jaccard, soft TF/IDF, and Monge-Elkan.
- **Phonetic-based:** It matches strings based on their sound instead of appearances. This similarity measure effectively matches names, which often appear in different ways that sound the same, such as Meyer, Meier, and Mire; Smith, Smithe, and Smythe. Such as Soundex.

MoMatch utilizes a set of string similarity measures to search for the most effective string similarity measures that can be used to find identical matches between entities. The utilized string-based techniques are described in Table 1 and Table 2 with their formulas.

III. RELATED WORK

According to a recent review of the multilingual Web of Data literature, fewer researchers have addressed cross-lingual

ontology matching [23]. Cross-lingual ontology matching techniques are mainly used for matching linguistic information across ontologies in different natural languages [4], [23].

A. CROSS-LINGUAL MATCHING APPROACHES

Abu Helou and Palmonari [24] proposed a cross-lingual lexical matching technique to map lexically-rich language resources such as WordNet. The results of word translations are used as evidence to map concepts lexicalized in different languages. Google Translate and BabelNet are used as external resources for translation. Four language versions of WordNet (Arabic, Italian, Slovene, and Spanish) are mapped to the English WordNet. Musyaffa et al. [25] proposed a framework for interlinking heterogeneous multilingual open fiscal data. Machine translation and similarity measures are used to map concepts across different languages.

Fu and Brennan [26], [27] proposed an algorithm for matching English and Chinese ontologies that takes into account the target ontology's semantics, mapping intent, operating domain, time and resource constraints, and user feedback. Hertling and Paulheim [28] proposed an approach for finding corresponding ontology elements that makes use of Wikipedia's interlanguage links. Lin and Krizhanovsky [29] proposed an approach that uses Wiktionary⁵ as a resource for background information to match English and French ontologies.

Trojahn et al. [30] proposed a multi-agent architecture for cross-lingual ontology matching. The translation agent translates the source ontology into the target language using a dictionary. Lexical databases and thesaurus have been used

⁴http://anhaidgroup.github.io/py_stringmatching/v0.4.x/index.html

⁵<https://www.wiktionary.org/>

for identifying mappings. Tigrine et al. [31] presented an approach for matching several ontologies in different natural languages that uses the multilingual semantic network Babel-Net⁶ as a source of background information.

Ali et al. [32] proposed a multi-agent architecture-based cross-lingual ontology enrichment approach to enrich ontologies from multilingual text or ontologies. Ibrahim et al. [33], [34] proposed a fully automated ontology enrichment approach based on cross-lingual matching that creates a multilingual ontology by enriching a monolingual one from another in a different language. They used lexical similarity (Jaccard) and semantic similarity (based on WordNet) to filter the equivalent classed. All translations produced by Google Translate for each class are considered during the matching process. They proposed another approach for merging monolingual ontologies in various natural languages to produce a multilingual ontology [11]. First, the alignments between input ontologies are identified using the cross-lingual matching techniques, then adds them to the merged multilingual ontology by adding `rdfs:label` for each language (using language-tagged strings). Offline dictionaries have been built using Yandex translate API for concept translations, and lexical (Jaccard) and semantic similarity are used to find the alignment between input ontologies.

SimCat [35] and Crolom [36] proposed a lexical matching technique that uses Yandex translator and WordNet to compute the semantic similarity between concepts. They first apply NLP techniques to normalize ontology entities, translate all entities, and compute the similarity between them. Such order affects the quality of the translation because NLP techniques eliminate and normalize words, which can greatly affect the translation quality and significantly reduce the alignment quality.

B. CROSS-LINGUAL MATCHING TOOLS IN OAEI 2020

In the context of OAEI 2020⁷ results of the ontology alignment evaluation initiative 2020 campaign for evaluating ontology matching technologies, VeeAlign [37], AML [38], LogMap [39], and Wiktionary Matcher [40] provide high-quality alignments for the cross-lingual matching task. These approaches heavily rely on the lexical matching technique, except VeeAlign [37], which discovers alignments using a supervised deep learning approach. VeeAlign [37] proposes a two-step model which uses contextualized representations of concepts to discover alignments based on semantic and structural aspects of an ontology. AML [38] is based on lexical and structural matching algorithms. It utilizes background knowledge and machine translation tools, such as Microsoft Translator, before starting the matching process. LogMap [39] implements optimized data structures for lexically and structurally indexing the input ontologies for the matching process. It is an iterative process that begins with initial mappings (i.e., almost exact lexical correspondences) and proceeds to

the discovery of new mappings. Therefore, LogMap cannot find matches between ontologies that lack sufficient lexical information. LogMapLt [39] is a “lightweight” version of LogMap, which only applies string matching techniques. Wiktionary Matcher [40] is an element-level, label-based matcher which uses multiple language versions of Wiktionary as an external background knowledge source. Meilicke et al. [41] developed a benchmark dataset (MultiFarm) based on manual translations of a set of ontologies from the conference domain into eight natural languages. This dataset is widely utilized to evaluate cross-lingual matching approaches [37], [38], [39], [40]. Good literature on the state-of-the-art approaches in cross-lingual ontology matching is provided in [4].

Despite ongoing efforts to develop various techniques, no clear winner has emerged in the solution to the cross-lingual matching problem. Further investigations are still needed to advance cross-language ontology matching techniques not only to obtain good results but also to assess them.

IV. MOMATCH: THE PROPOSED APPROACH

In this section, we will describe in detail the four phases of our approach (cf. Figure 1). The input consists of two ontologies O_1 and O_2 , which can be in two different natural languages (for example, $L_1 = fr$ and $L_2 = de$) or in the same language. The output is the alignment between the input ontologies in addition to the assessment sheet for the input ontologies and the resultant alignment. In the following subsections, we describe each of these phases in detail.

A. RESOURCE EXTRACTION

This phase aims to extract all resources (including both classes and properties) from the two input ontologies and store them in the resources matrix R . R is a two-dimensional matrix ($n \times 6$), where n is the number of extracted resources. Each row in R is represented as a tuple of `(resource, type, source, language, translation, pre-processed translation)`, which contains the resource label, the type of the resource ('C' for a class and 'P' for a property), the source ontology of this resource, the language tag, the translation of this resource, and the pre-processed translation. Currently, the translation and the pre-processed translation for each resource will be NULL and will be assigned later in the following phases. For more illustration, consider the tuple `(Comité de programme, C,  $O_1$ , fr, NULL, NULL)`, in which the resource “Comité de programme” is a class extracted from O_1 and its language is French. Similarly, the tuple `(ist geschrieben von, P,  $O_2$ , de, NULL, NULL)`, in which the resource “ist geschrieben von” is a property extracted from O_2 and its language is German. The output of this phase is R .

B. TRANSLATION

In order to match resources from different languages, all of them should be translated into a common natural language.

⁶<https://babelnet.org/>

⁷<http://oaei.ontologymatching.org/2020/results/multifarm/index.html>

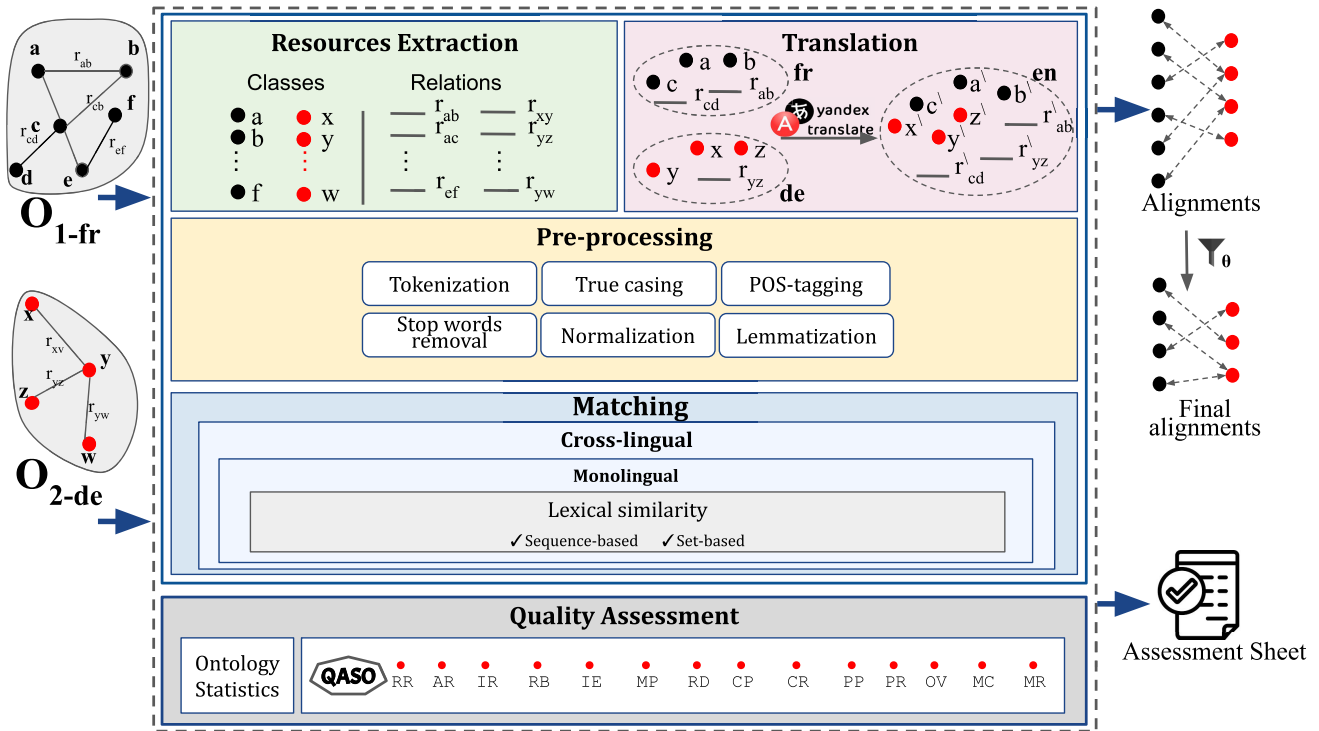


FIGURE 1. MoMatch architecture.

Two translation paths could be followed. The first path is to translate one ontology's resources into the other's language. For example, if the input ontologies are in German and French, then either the resources of the German ontology are translated to French or the resources of the French ontology are translated to German. The second path is to translate both of the input ontologies into a chosen pivot language, like English, for example. In our approach, we chose the second path because we found that most of the Natural Language Processing (NLP) techniques for English text outperform others in other languages. Furthermore, most of the external knowledge sources used in the matching process, such as dictionaries and thesaurus, are available in English.

According to Abu Helou et al. [42], machine translation tools can return proper translations for a very large number of resources in the cross-lingual matching task. We use Yandex translate API⁸ to translate each resource in the resource matrix R and add the resource's translation to R . For example, consider the same tuple from the previous example in subsection IV-A (Comité de programme, C, O_1 , fr, Program Committee, NULL), in which the English translation "Program Committee" is attached to the French class "Comité de programme". The output of this phase is R with the translations of all resources.

C. PRE-PROCESSING

This phase aims to clean and prepare the translated resources for the matching phase by employing the following NLP techniques.

⁸<https://tech.yandex.com/translate/>

- *Tokenization*: divides the resource name into a set of tokens. Tokens are separated by delimiters such as whitespace characters.
- *True casing*: recognizes resources with camel cases and adds a space between lower-case and upper-case letters such as "BestPosterAward" and "isSponsorOf" became "Best Poster Award" and "is Sponsor Of" respectively.
- *POS-tagging*: classifies tokens into their parts of speech, depending on their definition and context, then assigns them a special label or tag (such as Adjective (ADJ), Conjunction (CONJ), Verb (V), and Preposition (PREP)) accordingly.
- *Stop words removal*: removes tokens with high frequencies of occurrence and have no contribution to the subject of a text, such as pronouns, prepositions, and conjunctions.
- *Normalization and regular expressions*: transforms a token into a standard form by removing non-alphanumeric characters and additional white spaces.
- *Lemmatization*: removes inflections in a token and maps it to its root form. For example, "learning" is mapped to "learn" and "played" to "play".

The output of this phase is R with the pre-processed translated resources.

D. MATCHING

This phase aims to identify correspondences (matched resources) between the two input ontologies. We perform a pairwise lexical similarity between the translated pre-

TABLE 3. Ontology statistics metrics in MoMatch.

Metric	Description
getNumTrip(O)	Get the number of triples in the ontology.
getNumRes(O)	Get the number of all resources (classes and properties) in the ontology.
getNumClass(O)	Get the number of classes in the ontology.
getNumProp(O)	Get the number of all properties in the ontology.
getNumObjProp(O)	Get the number of object properties in the ontology.
getNumDataProp(O)	Get the number of data properties in the ontology.
getNumAnnProp(O)	Get the number of annotation properties in the ontology.
getNumSubClassOf(O)	Get the number of subClassOf relations in the ontology.
getNumHRD(O)	Get the number of human-readable descriptions (HRD) such as comment, label, and description.

processed resources of the two input ontologies. We use 13 string matching measures described in Table 1 and Table 2. We take different threshold values to select the top-matched resources in the matching process. MoMatch matches not only ontologies in different natural languages (i.e., cross-lingual matching) but also ontologies in the same natural language (i.e., monolingual matching). In monolingual matching, the translation phase is skipped, and the matching occurs between the pre-processed resources of the two ontologies.

The output of this phase is the matched resources stored in a matrix M . M is a two-dimensional matrix ($m \times 4$), where m is the number of matched resources. Each row in M is represented as a tuple of $(resource1, resource2, type, simScore)$, which contains resource1 from O_1 , resource2 from O_2 , type of both resources, and the similarity score between them. For example, $(Car, , C, 1.00)$, i.e., “Car” in English and “ ” in Arabic are two resources of a type class, from O_1 and O_2 , respectively, with a similarity score of 1.00.

V. QUALITY ASSESSMENT

We propose the metric suite QASO for assessing the quality of the matching process. These metrics need to have prior statistical information about the ontologies. We calculate the statistics of the input ontologies using a distributed in-memory approach for statistical computing calculations of large-scale RDF datasets using Apache Spark [43]. A sample for the statistics metrics in QASO is described in Table 3.

A. QASO SUITE

QASO comprises seven metrics for assessing the quality of the ontologies and seven for evaluating the quality of the matching process. We adapt and reformulate the metrics defined in [6], [7], and [8].

- **Relationship richness (RR)** [6]: refers to the diversity of relations and their position in the ontology. The more relations the ontology has (except `rdfs:subClassOf` relation), the richer it is. The quality score function $f_{RR} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{RR}(O) = \frac{|P_{obj}|}{|P_{subClassOf}| + |P_{obj}|} \in [0, 1] \quad (12)$$

where P_{obj} represents the relationships (i.e., object properties) and $P_{subClassOf}$ represents the `rdfs:subClassOf` relations in O .

- **Attribute Richness (AR)** [6]: refers to how much knowledge about classes is in the schema. The more attributes are defined, the more knowledge the ontology provides. The quality score function $f_{AR} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{AR}(O) = \frac{|P_{attr}|}{|C|} \in [0, +\infty] \quad (13)$$

where C represents the ontology classes and P_{attr} represents all classes' attributes (i.e., data properties).

- **Inheritance Richness (IR)** [6]: refers to how well knowledge is distributed across different levels in the ontology. The more `rdfs:subClassOf` relations, the broader range of general knowledge the ontology provides. The quality score function $f_{IR} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{IR}(O) = \frac{|P_{subClassOf}|}{|C|} \in [0, +\infty] \quad (14)$$

- **Readability (RB)** [6]: refers to the existence of human-readable descriptions (HRD) in the ontology, such as comments, labels, or descriptions. The more HRD exists, the more readable the ontology is. The quality score function $f_{RB} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{RB}(O) = \frac{|HRD|}{|R|} \in [0, +\infty] \quad (15)$$

where $HRD \in \{label, comment, description\}$ and R represents the ontology resources.

- **Isolated Elements (IE)** [7]: refers to classes and properties which are defined but not connected to the rest of the ontology, i.e., not used. The quality score function $f_{IE} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{IE}(O) = \frac{|R_{isolated}|}{|R|} \in [0, 1] \quad (16)$$

where $R_{isolated}$ represents resources defined but not used in O .

- **Missing Domain or Range in Properties (MP)** [7]: refers to missing information about properties. The less information about properties is missing, the more complete the ontology. The quality score function $f_{MP} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{MP}(O) = \frac{|P_{incomplete}|}{|P|} \in [0, 1] \quad (17)$$

where $P_{incomplete}$ represents properties that do not have domain or range.

- **Redundancy (RD)** [8]: refers to how many redundant resources exist. Resources which are syntactically (e.g. “isMemberOf” and “is_member_of”) or semantically (e.g. “Chair” and “Chairman”) close are considered as redundant resources. The quality score function $f_{RD} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{RD}(O) = \frac{|R_r|}{|R|} \in [0, +\infty] \quad (18)$$

where R_r represents the redundant resources in O .

All the previous metrics can be used to assess the quality of any ontology. To assess the quality and effectiveness of a matching process, we need to verify whether all relevant correspondences have been retrieved (high recall) and correct (high precision). The following metrics are adapted to assess the quality of the matching process not only by using the reference alignment (*Ref*) but also without it.

- **Class Precision (CP)**: refers to the fraction of relevant matched classes among the retrieved ones. The more relevant results retrieved, the more precision the matching process has. The quality score function $f_{CP} : O \rightarrow \mathbb{R}$ for the matching process M is defined as follows:

$$f_{CP}(M) = \frac{|Ref_c| \cap |C_{match}|}{|C_{match}|} \in [0, 1] \quad (19)$$

where C_{match} is the retrieved matched classes by MoMatch, and Ref_c is the matched classes in the reference alignment.

- **Class Recall (CR)**: refers to the fraction of relevant matched classes retrieved by the system. The more relevant results retrieved, the more recall the matching process has. The quality score function $f_{CR} : O \rightarrow \mathbb{R}$ for the matching process M is defined as follows:

$$f_{CR}(M) = \frac{|Ref_c| \cap |C_{match}|}{|Ref_c|} \in [0, 1] \quad (20)$$

- **Property Precision (PP)**: refers to the fraction of relevant matched properties among the retrieved ones. The more relevant results retrieved, the more precision the matching process has. The quality score function $f_{PP} : O \rightarrow \mathbb{R}$ for the matching process M is defined as follows:

$$f_{PP}(M) = \frac{|Ref_p| \cap |P_{match}|}{|P_{match}|} \in [0, 1] \quad (21)$$

where P_{match} is the retrieved matched properties by MoMatch, and Ref_p is the matched properties in the reference alignment.

- **Property Recall (PR)**: refers to the fraction of relevant matched properties retrieved by the system. The more relevant results retrieved, the more recall the matching process has. The quality score function $f_{PR} : O \rightarrow \mathbb{R}$ for the matching process M is defined as follows:

$$f_{PR}(M) = \frac{|Ref_p| \cap |P_{match}|}{|Ref_p|} \in [0, 1] \quad (22)$$

- **Degree of Overlap (OV)**: refers to how many common resources exist between the input ontologies. Resources that are syntactically or semantically close are considered common resources. The quality score function $f_{OV} : O \rightarrow \mathbb{R}$ for the matching process M is defined as follows:

$$f_{OV}(M) = \frac{|R_{match}|}{|R_1| \cup |R_2|} \in [0, 1] \quad (23)$$

where R_{match} represents the set of found correspondences in a match result produced when matching O_1 and O_2 ontologies. R_1 and R_2 represent the resources of the input ontologies O_1 and O_2 respectively.

If the reference alignment is unavailable, we use rough approximations for recall and precision based on the relative quality of the obtained matching results produced by MoMatch [44].

- **Match Coverage (MC)** [44]: estimation for recall refers to the fraction of resources that exist in at least one correspondence in the matching results in comparison to the total number of resources in the input ontologies. The quality score function $f_{MC} : O \rightarrow \mathbb{R}$ for the matching process M is defined as follows:

$$f_{MC}(M) = \frac{|R_{O_1-match}| \cup |R_{O_2-match}|}{|R_1| \cup |R_2|} \in [0, 1] \quad (24)$$

where $R_{O_1-match}$ and $R_{O_2-match}$ represents the set of matched resources of ontologies O_1 and O_2 respectively.

- **Match Ration (MR)** [44]: estimation for precision refers to the ratio between the number of found correspondences and the number of matched resources in the input ontologies. The closer the ratio is to 1.00, the better precision the match result has. In other words, when a match result is not loosely matched to many other resources but only the most similar ones, the match result is better. The quality score function $f_{MR} : O \rightarrow \mathbb{R}$ for the matching process M is defined as follows:

$$f_{MR}(M) = \frac{2 \times |R_{match}|}{|R_{O_1-match}| \cup |R_{O_2-match}|} \in [1, +\infty] \quad (25)$$

VI. USE CASES

In this section, we show the applicability of MoMatch in monolingual and cross-lingual ontology matching in different domains.

A. USE CASE 1: CROSS-LINGUAL MATCHING IN SCHOLARLY COMMUNICATION DOMAIN

In this use case, we use ontologies from the scholarly communication domain in various natural languages. We use an example scenario to match the *SEO_{en}*⁹ ontology (with 106 classes and 88 properties), in English, with *Conference_{de}* ontology (60 classes and 64 properties), in German, from the

⁹<https://w3id.org/seo>

TABLE 4. Use case 1: a sample output of each phase for matching SEO_{en} with $Conference_{de}$, starting from resource extraction to matching results.

Phase	Output
Resource extraction	$\langle \text{Konferenzbeteiligung}, C, O_1, de \rangle$ $\langle \text{Vorsitzender}, C, O_1, de \rangle$ $\langle \text{hat Beitrag eingreicht}, P, O_1, de \rangle$ $\langle \text{student registration}, C, O_2, en \rangle$
Translation	$\langle \text{Tutorium}, C, O_1, de, \text{Tutorial} \rangle$ $\langle \text{Thema}, C, O_1, de, \text{Topic} \rangle$ $\langle \text{hat Beitrag eingreicht}, P, O_1, de, \text{has contribution submitted} \rangle$
Pre-processing	SizeOrDuration $\rightarrow$ size duration WorkshopProposals $\rightarrow$ workshop proposal InvitedSpeaker $\rightarrow$ invite speaker In-useTrack $\rightarrow$ use track
Matching	$\langle \text{Gutachter}, \text{reviewer}, C, 1.00 \rangle$ $\langle \text{eingeladener Referent}, \text{invited speakers}, C, 1.00 \rangle$ $\langle \text{eingeladener Referent}, \text{keynote speaker}, C, 0.57 \rangle$ $\langle \text{person}, \text{person}, C, 1.00 \rangle$ $\langle \text{Tutorium}, \text{tutorial proposals}, C, 0.78 \rangle$ $\langle \text{Koferenzdokument}, \text{license document}, C, 0.61 \rangle$ $\langle \text{Gutachten}, \text{reviewer}, C, 0.75 \rangle$ $\langle \text{proceedings}, \text{proceedings}, C, 1.00 \rangle$ $\langle \text{Organisator}, \text{organizer}, C, 1.00 \rangle$ $\langle \text{wichtiges Datum}, \text{important dates}, C, 1.00 \rangle$ $\langle \text{Herausgeber}, \text{publisher}, C, 1.00 \rangle$ $\langle \text{Leiter der Workshops}, \text{workshop chair}, C, 1.00 \rangle$ $\langle \text{hat wichtiges Datum}, \text{has important dates}, P, 1.00 \rangle$ $\langle \text{hat Abfolge}, \text{hasTrack}, P, 1.00 \rangle$

MultiFarm dataset (see section VII). The goal of this use case is to demonstrate the entire process, from submitting the input ontologies to producing the alignment. Here, O_1 is the German ontology $Conference_{de}$ and O_2 is the English ontology SEO_{en} . Table 4 demonstrates the matching process and shows each phase's sample output. The relevant matching results are identified in the matching phase using Jaccard similarity with a threshold $\theta \geq 0.90$. MoMatch identified eight matched classes and two matched properties.

B. USE CASE 2: MONOLINGUAL MATCHING IN BIOMEDICAL DOMAIN

In this use case, we use ontologies from the biomedical domain. We match monolingual ontologies from BioPortal¹⁰ – a web-based application for accessing and sharing biomedical ontologies and providing alignments between them. We match the Sample Processing and Separation Techniques Ontology (SEP)¹¹ with several ontologies from the Bio-portal such as Plant Experimental Conditions (PECO),¹²

TABLE 5. Statistics of BioPortal ontologies (R_{match} = number of matched resources, i.e., correspondences) and degree of overlap (OV) between SEP ontology and other ontologies.

Ontology Name	Classes	Properties	R_{match} in BioPortal	R_{match} by MoMatch	OV ¹⁶ by MoMatch
PECO	3,119	124	11	46	1.17%
PO	2,019	111	1	39	1.38%
PTO	1,659	1	0	37	2%
UO	634	2	448	434	34%
SEP	658	2	–	–	–

Plant (PO),¹³ Plant Trait (PTO),¹⁴ and Units of Measurement (UO)¹⁵ ontologies. Table 5 shows the statistics of the Bioportal ontologies and the degree of overlap between SEP with each ontology. The degree of overlap refers to how many common resources exist between ontology pairs as defined in subsection V-A. MoMatch identified 434 out of 448 correspondences found in Bioportal for matching SEP $\times$ UO with a 34% degree of overlap. Surprisingly, MoMatch found new correspondences which were missing in BioPortal. In matching SEP $\times$ PECO, MoMatch found 35 new correspondences. Similarly, in matching SEP $\times$ PO, MoMatch found 38 new correspondences. In matching SEP $\times$ PTO, MoMatch found 37 new correspondences.

VII. EVALUATION

In this section, we show the results of two experiments to provide an in-depth analysis for comparing similarity measures across different languages. In addition, we evaluate the quality of the matching process using QASO.

A. DATASET AND EXPERIMENT SETUP

1) DATASET

We use MultiFarm benchmark¹⁷ from the T-Box/Schema matching track of OAEI 2020.¹⁸ The OAEI 2020 competition is an annual international ontology matching competition. MultiFarm is a cross-lingual ontology matching system evaluation benchmark. It consists of seven ontologies (*Cmt*, *Conference*, *ConfOf*, *Edas*, *Ekaw*, *Iasted*, *Sigkdd*) derived from the Conference benchmark of OAEI, their translation into nine languages (Chinese, Czech, Dutch, French, German, Portuguese, Russian, Spanish and Arabic), and the corresponding cross-lingual alignments between them. Statistics for Mulltifarm ontologies are presented in Table 6. Classification of the nine languages according to Wikipedia¹⁹ is presented in Table 7. We could not use the dataset from OAEI 2021 [45] because the results are not available in detail (for each language pair and each ontology pair) as in OAEI 2020.

¹⁰<https://bioportal.bioontology.org/>

¹¹<https://bioportal.bioontology.org/ontologies/SEP>

¹²<https://bioportal.bioontology.org/ontologies/PECO>

¹³<https://bioportal.bioontology.org/ontologies/PO>

¹⁴<https://bioportal.bioontology.org/ontologies/PTO>

¹⁵<https://bioportal.bioontology.org/ontologies/UO>

¹⁷<https://www.irit.fr/recherches/MELODI/multifarm/>

¹⁸<http://oaei.ontologymatching.org/2020/>

¹⁹https://en.wikipedia.org/wiki/Language_family

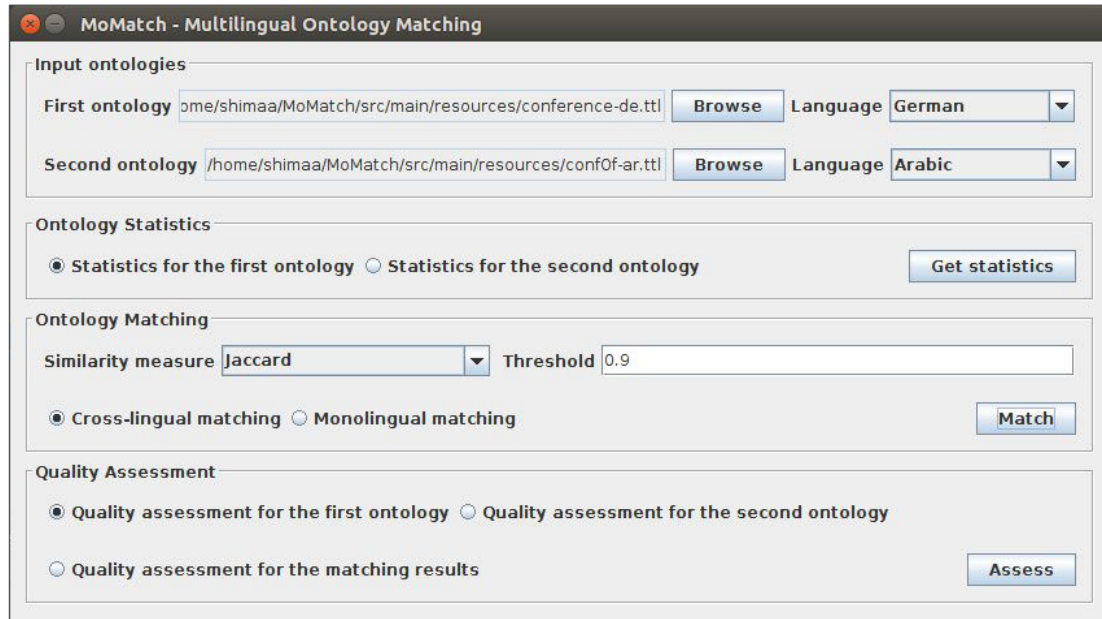


FIGURE 2. MoMatch GUI.

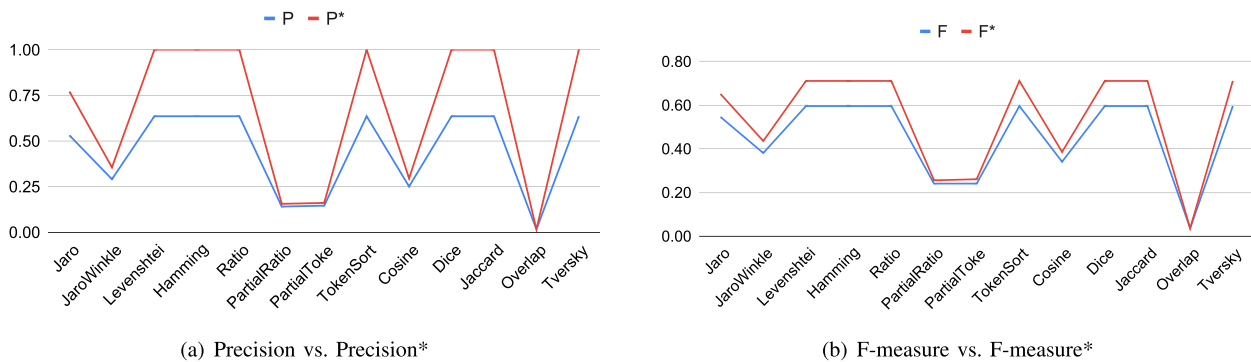
FIGURE 3. Matching French-German ontologies where $\theta \geq 0.90$. P^* and F^* represent the improvement in precision and F-measure results when considering the new correspondences.

TABLE 6. Statistics of MultiFarm ontologies.

Ontology Name	Classes	Object Properties	Data Properties
<i>Ekaw</i>	74	33	0
<i>Edas</i>	104	30	20
<i>Conference</i>	60	46	18
<i>ConfOf</i>	38	13	23
<i>Cmt</i>	36	49	10
<i>Iasted</i>	140	38	3
<i>Sigkdd</i>	49	17	11

2) EXPERIMENT SETUP

Scala and Apache Spark²⁰ were used to implement all phases of MoMatch. A graphical user interface of MoMatch is created (cf. Figure 2). To parse and manipulate the input ontologies (as RDF triples), SANSARDF library²¹ [46] with

TABLE 7. Classification of languages in Multifarm according to Wikipedia and number of LOV in each category.

Language	# of LOV	Category	
English (en)	584	Indo-European	Germanic
German (de)	41		Germanic
Dutch (nl)	20		Germanic
French (fr)	65		Italic
Portuguese (pt)	19		Italic
Spanish (es)	49		Italic
Russian (ru)	17		Balto-Slavic
Czech (cz)	9		Balto-Slavic
Chinese (cn)	10		Tai-Kadai
Arabic (ar)	7	Afro-Asiatic	

Apache Jena framework²² are used. To process the resource labels, the Stanford CoreNLP²³ [47] is used. All experiments

²⁰<https://spark.apache.org/>

²¹<https://github.com/SANSARDF/SANSARDF>

²²<https://jena.apache.org/>

²³<https://stanfordnlp.github.io/CoreNLP/>

TABLE 8. Average values (for all ontology pairs) for precision (P), recall (R), and F-measure (F) of matching multilingual ontologies. P* and F* describe the results with adjusted precision and F-measure. Red, green, and blue entries are the top scores for precision, recall, and F-measure for each similarity measure per row. Similar values are merged.

Similarity measure	1					0.95					0.90					0.85					0.80				
	P	P*	R	F	F*	P	P*	R	F	F*	P	P*	R	F	F*	P	P*	R	F	F*	P	P*	R	F	F*
Jaro											0.53	0.77		0.55	0.65	0.33	0.42		0.42	0.48	0.22	0.26		0.31	0.35
JaroWinkler											0.29	0.36		0.38	0.44	0.19	0.23		0.28	0.32	0.14	0.15		0.22	0.24
Levenshtein	0.64	1.00	0.56	0.60	0.71	0.64	1.00	0.56	0.60	0.71	0.64		0.56	0.60		0.64	1.00	0.56	0.60	0.71	0.64	1.00	0.56	0.60	0.71
Hamming											0.64	1.00		0.60	0.71	0.64	1.00		0.60	0.71	0.64	1.00		0.60	0.71
Ratio											0.64			0.60		0.57	0.83		0.56	0.67	0.46	0.63		0.51	0.59
PartialRatio	0.15	0.16	0.73	0.24	0.26	0.15	0.16	0.73	0.24	0.26	0.14	0.16	0.73	0.24	0.26	0.14	0.15	0.73	0.23	0.25	0.20	0.24	0.73	0.31	0.35
PartialTokenSort	0.15	0.16	0.73	0.24	0.26	0.15	0.16	0.73	0.24	0.26	0.15	0.16	0.73	0.24	0.26	0.14	0.16	0.73	0.24	0.26	0.11	0.12	0.73	0.19	0.20
TokenSort											0.64	1.00		0.60	0.71	0.57	0.83	0.56	0.56	0.67	0.46	0.63	0.56	0.51	0.59
Cosine	0.64	1.00	0.56	0.60	0.71	0.64	1.00	0.56	0.60	0.71	0.25	0.30		0.34	0.39	0.08	0.08	0.67	0.14	0.14	0.04	0.04	0.78	0.07	0.07
Dice											0.64	1.00	0.56	0.60	0.71	0.53	0.77	0.56	0.55	0.65	0.28	0.33	0.56	0.37	0.42
Jaccard											0.64	1.00		0.60	0.71	0.64	1.00	0.56	0.60	0.71	0.64	1.00	0.56	0.60	0.71
Overlap	0.02	0.02	0.73	0.05	0.05	0.02	0.02	0.73	0.04	0.04	0.02	0.02	0.73	0.04	0.04	0.01	0.01	0.73	0.02	0.02	0.01	0.01	0.78	0.01	0.01
Tversky	0.64	1.00	0.56	0.60	0.71	0.64	1.00	0.56	0.60	0.71	0.64	1.00	0.56	0.60	0.71	0.64	1.00	0.56	0.60	0.71	0.60	0.92	0.56	0.58	0.69

(a) French-German

Similarity measure	1					0.95					0.90					0.85					0.80				
	P	P*	R	F	F*	P	P*	R	F	F*	P	P*	R	F	F*	P	P*	R	F	F*	P	P*	R	F	F*
Jaro											0.59	0.80		0.48	0.54	0.35	0.41		0.38	0.41	0.27	0.31		0.32	0.35
JaroWinkler											0.30	0.33		0.34	0.37	0.16	0.17		0.23	0.24	0.09	0.10		0.15	0.15
Levenshtein	0.71	1.00	0.41	0.52	0.59	0.71	1.00	0.41	0.52	0.59	0.71		0.41	0.52		0.71	1.00	0.41	0.52	0.59	0.71	1.00	0.41	0.52	0.59
Hamming											0.71	1.00		0.52	0.59	0.71	1.00		0.52	0.59	0.71	1.00		0.52	0.59
Ratio											0.71			0.52		0.59	0.78		0.48	0.54	0.40	0.47		0.40	0.44
PartialRatio	0.13	0.13	0.54	0.20	0.21	0.13	0.13	0.54	0.20	0.21	0.13	0.13	0.54	0.20	0.21	0.12	0.12	0.54	0.19	0.20	0.10	0.10	0.54	0.17	0.17
PartialTokenSort	0.13	0.13	0.54	0.20	0.21	0.13	0.13	0.54	0.20	0.21	0.13	0.13	0.54	0.20	0.21	0.12	0.13	0.54	0.19	0.20	0.10	0.10	0.54	0.17	0.17
TokenSort											0.71	1.00		0.52	0.59	0.59	0.78	0.41	0.48	0.54	0.40	0.47	0.41	0.40	0.44
Cosine	0.71	1.00	0.41	0.52	0.59	0.71	1.00	0.41	0.52	0.59	0.22	0.25		0.27	0.29	0.06	0.06	0.47	0.10	0.11	0.02	0.02	0.54	0.04	0.04
Dice											0.71	1.00	0.41	0.52	0.59	0.54	0.70	0.41	0.47	0.52	0.24	0.26	0.41	0.30	0.32
Jaccard											0.71	1.00		0.52	0.59	0.71	1.00	0.41	0.52	0.59	0.71	1.00	0.41	0.52	0.59
Overlap	0.02	0.02	0.54	0.04	0.04	0.02	0.02	0.54	0.04	0.04	0.01	0.01	0.54	0.03	0.03	0.01	0.01	0.54	0.01	0.01	0.00	0.00	0.54	0.01	0.01
Tversky	0.71	1.00	0.41	0.52	0.59	0.71	1.00	0.41	0.52	0.59	0.71	1.00	0.41	0.52	0.59	0.71	1.00	0.41	0.52	0.59	0.59	0.78	0.41	0.48	0.54

(b) German-Arabic

are run on Ubuntu 16.04 LTS with an Intel Core i7-4600U CPU @ 2.10GHz x 4 CPU and 10 GB of memory.

3) EVALUATION METRICS

The precision, recall, and F-measure metrics, inspired by the information retrieval community, can be used to evaluate the effectiveness of the matching process. We use the gold standard alignments between each pair of ontologies in Multifarm to compute precision, recall, and F-measure. Precision is the fraction of retrieved resources that are relevant, while recall is the fraction of relevant matched resources retrieved by MoMatch. The F-measure is the harmonic mean of precision and recall. Formally, precision is defined as $TP/(TP + FP)$, and recall is defined as $TP/(TP + FN)$, where TP (true positive) is a matching result that MoMatch retrieves and exists in the gold standard, FP (false positive) is a matching result that MoMatch retrieves and does not exist in the gold standard, and FN (false negative) is a matching result that exists in the gold standard, and MoMatch could not retrieve it.

4) EXPERIMENTAL CONFIGURATION

In our experiments, we choose one natural language for each language category in Table 7. Regarding categories with more than one language, such as Italic, we select the most widespread language according to Linked Open Vocabularies (LOV)¹. For example, the Italic family has three languages where 65 vocabularies are in French, 19 in Portuguese, and 49 in Spanish, i.e., French is selected, which has the most extensive vocabulary in the Italic family. We did not include English in our experiments because it is the most prominent language, and a lot of work has been done in this language. Therefore, we used German, French, Russian, Chinese, and Arabic in the experiments.

The Multifarm benchmark is composed of 55 pairs of languages, with 49 matching tasks for each of them, taking into account the alignment direction (e.g., $Cmt_{en} \rightarrow Edas_{de}$ and $Cmt_{de} \rightarrow Edas_{en}$ are distinct matching tasks), i.e., 2,695 matching task [48]. In MoMatch, we used 13 similarity measures, so we needed to perform a 35,035 matching task. To decrease the matching tasks without losing their value, we design our experiments to 1) evaluate the effectiveness of the cross-lingual matching process in MoMatch using different similarity measures compared to the reference alignment provided in the MultiFarm benchmark, 2) compare MoMatch matching results with five state-of-the-art approaches, and 3) assess the quality of the matching process using QASO.

B. EFFECTIVENESS OF MOMATCH

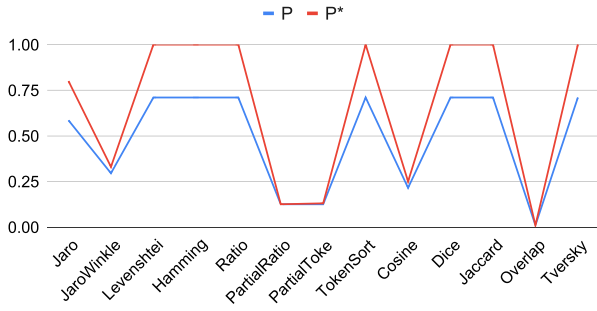
In this experiment, we use all similarity measures listed in both Table 1 and Table 2. Different threshold values ($\theta = \{1.00, 0.95, 0.90, 0.85, 0.80\}$) are used to select the top-matched results according to the resulting similarity scores between every two resources' labels. We select two ontologies (*Iasted* and *ConfOf*) from the Multifarm benchmark and two language pairs from different families: French–German

and German–Arabic. In matching French–German ontologies, we match the resources' labels of the French version of *Iasted_{fr}* ontology with the resources' labels of the German version of *ConfOf_{de}* (i.e., $Iasted_{fr} \times ConfOf_{de}$). Similarly, we match the resources' labels of the French version of *ConfOf_{fr}* ontology with the resources' labels of the German version of *Iasted_{de}*. The same procedure has been followed in matching German–Arabic ontologies. The resulting alignments are compared with the reference alignments as a gold standard provided in the benchmark for each pair of ontologies. Table 8(a) and Table 8(b) show the average values (for all ontology pairs) for precision, recall and F-measure for matching French–German and German–Arabic ontologies respectively. Surprisingly, we found new correspondences missing in the gold standard alignments, such as the correspondence $\langle \text{écrit, schreibt, } \equiv, 1.00 \rangle$, which implies that the French and German properties “écrit” and “schreibt” are identical with a similarity score = 1.00. P^* and F^* represent adjusted precision and F-measure results when considering the new correspondences. Therefore, P^* and F^* represent results that are not false positives in practice. Figure 3(a) and Figure 3(b) present the significant improvement of P^* and F^* over P and F , respectively, in all similarity measures for matching French–German ontologies, where $\theta \geq 0.90$. The precision and F-measure results are improved by an average of 23% and 8%, respectively. Similarly, Figure 4(a) and Figure 4(b) present a significant improvement in matching German–Arabic ontologies. The precision and F-measure results are improved by an average of 18% and 5%, respectively. To address our research questions, we study these results in terms of two dimensions: 1) similarity measures (RQ1) and 2) language pairs (RQ2).

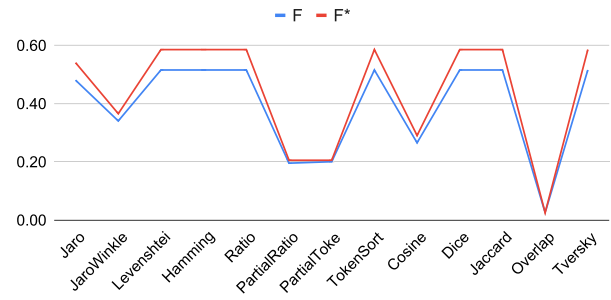
1) SIMILARITY MEASURES

As shown in Table 8, the precision is directly proportional to the threshold values, which means the precision increases when the similarity measure value increases, especially with $\theta \geq 0.90$. While the recall is inversely proportional to the threshold values, e.g., the highest recall is achieved with $\theta = 0.80$.

In matching French–German ontologies, Levenshtein, Hamming, Jaccard, and Tversky have achieved the best precision of 100% for all thresholds except for $\theta = 0.80$. Tversky performs 92% (see Figure 5(a)). Similarly, Levenshtein, Hamming, Jaccard, and Tversky have achieved the best F-measure of 71% for all thresholds except for $\theta = 0.80$. Tversky reaches 69% (see Figure 5(c)). Cosine, Overlap coefficient, Partial ratio, and Partial token sort have achieved the best recall of 78%, 78%, 73%, and 73%, respectively, where $\theta = 0.80$ (see Figure 5(b)). In matching German–Arabic ontologies, Levenshtein, Hamming, Jaccard, and Tversky have achieved the best precision of 100% for all thresholds (except for $\theta = 0.80$, Tversky gains 78%) (see Figure 6(a)). Levenshtein, Hamming, Jaccard, and Tversky have achieved the best F-measure of 59% for all thresholds (except for $\theta = 0.80$, Tversky performs 54%) (see Figure 6(c)). Cosine,

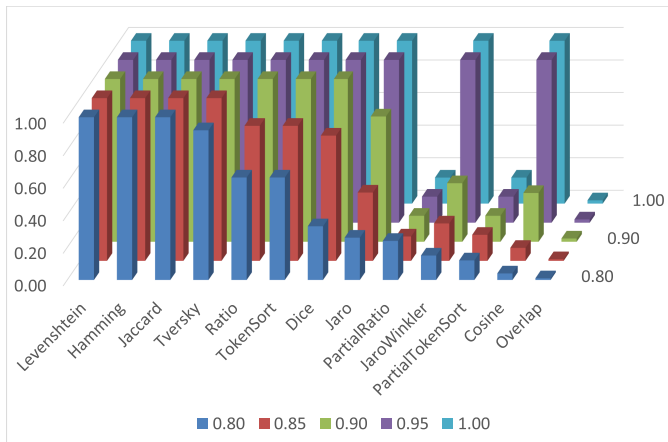


(a) Precision vs. Precision*

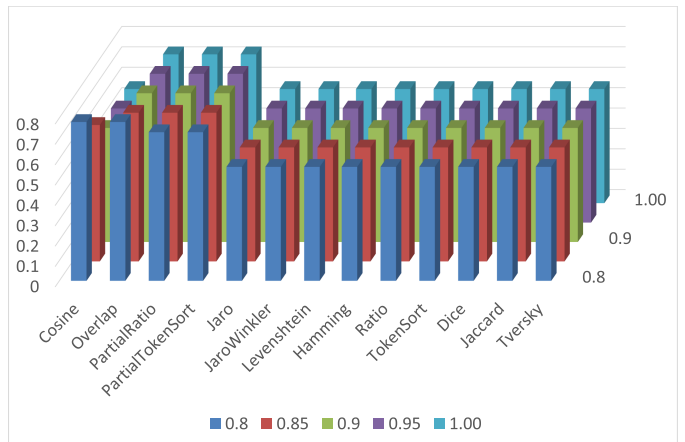


(b) F-measure vs. F-measure*

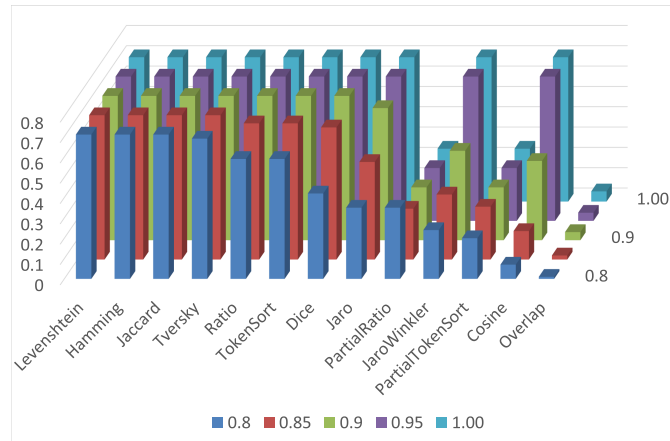
FIGURE 4. Matching German–Arabic ontologies where $\theta \geq 0.90$. The improvement of P^* and F^* have similar patterns with Figure 3(a) and Figure 3(b).



(a) Precision.



(b) Recall.



(c) F-measure.

FIGURE 5. Comparing precision, recall, and F-measure of different similarity measures over all thresholds for matching French–German ontologies, as shown in Table 8. Similarity measures are sorted from left (the highest value) to right (the lowest value) according to precision, recall, and F-measure. The more threshold values, the more precision and F-measure scores are achieved. In contrast, the lower the threshold values, the more recall scores are achieved.

Overlap coefficient, Partial ratio, and partial token sort have achieved the best recall of 54% where $\theta = 0.80$ (see Figure 6(b)).

Across the two language pairs, French $\times$ German and German $\times$ Arabic, precision and F-measure in sequence-based similarity measures are sensitive to variations in the

threshold (see Figure 7(a) and Figure 7(e)). Levenshtein, Hamming, Token sort, and Ratio provide the best precision and F-measure scores. In contrast, recall remains constant across all threshold values (see Figure 7(c)). Partial ratio and Partial token sort provide the highest recall of 64% across all threshold values. Similarly, precision and F-measure in

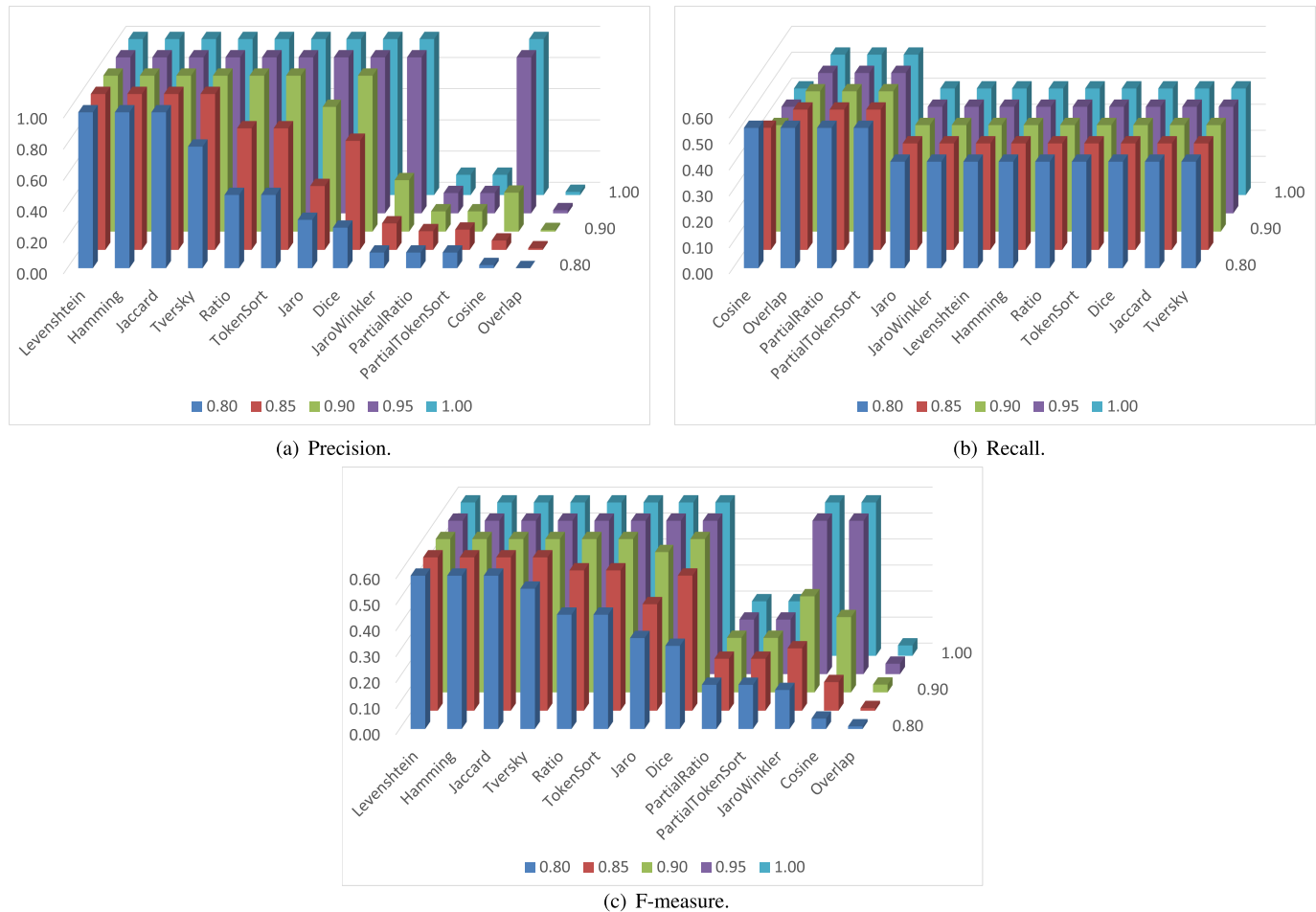


FIGURE 6. Precision, recall, and F-measure charts of different similarity measures over all thresholds for matching German-Arabic ontologies, as shown in Table 8(b). There is no significant difference between these charts and the charts for matching French-German ontologies (Figure 5).

set-based similarity measures are also sensitive to variations in the threshold (see Figure 7(b) and Figure 7(f)). Jaccard, Tversky, and Dice have the best precision and F-measure scores. The recall is relatively stable against threshold changes (see Figure 7(d)). The overlap coefficient provides the best recall of 64% and 66% for $\theta \geq 0.80$.

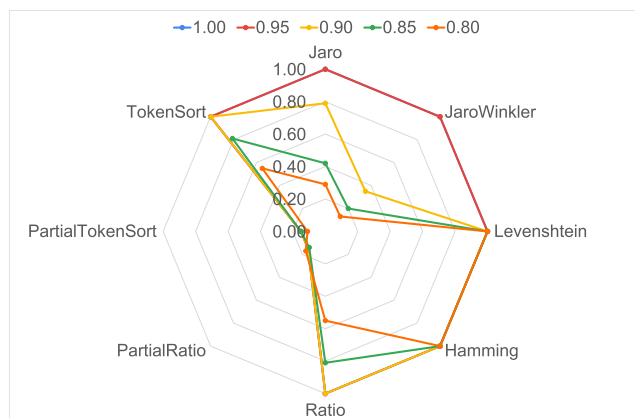
The choice of string similarity measure greatly influences the precision and recall of the matching process. When selecting a similarity measure to be used in ontology matching, it is crucial to examine the features of the ontologies being matched and whether the precision or the recall is more significant to the matching technique. Levenshtein, Hamming, and Jaccard achieve the best precision. If one wants to use more than one similarity measure, then one can choose from the following list, sorted in descending order of precision: 1) Levenshtein, Hamming, and Jaccard, 2) Tversky, 3) Token sort and Ratio, 4) Dice, 5) Jaro, 6) Jaro Winkler, 7) Cosine, 8) Partial ratio, 9) Partial token sort, 10) Overlap coefficient. On the other hand, the Overlap coefficient achieves the best recall. Similarly, if one wants to use more than one similarity measure, then one can select from the following list sorted in descending order of recall: 1) Overlap coefficient, 2) Partial

ratio and Partial token sort, 3) Cosine, 4) Jaro, Jaro Winkler, Levenshtein, Hamming, Ratio, Token sort, Dice, Jaccard, and Tversky.

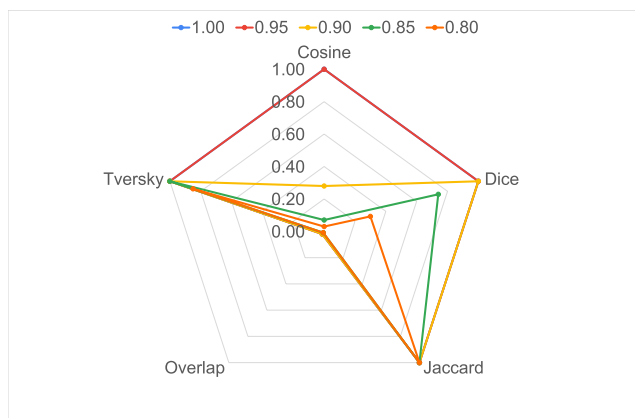
2) LANGUAGE PAIRS

The language pair French-German has better results than German-Arabic (see Figure 8). Two native speakers of Arabic found that the reason behind that is the linguistic mistakes found in the Arabic ontologies, which negatively affect the translation and the matching results. We correct these mistakes and make them available at the MoMatch GitHub repository².

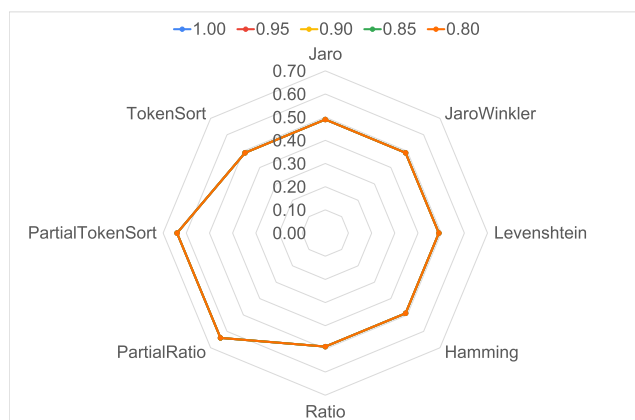
In order to answer the research question *RQ2*, we calculate the Spearman correlation between the two language pairs for precision, recall, and F-measure where $\theta = \{1.00, 0.95, 0.90, 0.85, 0.80\}$. All threshold values have the highest correlation of 1.00 except for $\theta = 0.80$; the correlation values are 0.99, 0.80, and 0.98 for precision, recall, and F-measure, respectively. As a result, the two language pairs are positively and strongly correlated to each other, i.e., the higher similarity measure is ranked in French-German, the higher similarity measure is ranked in German-Arabic, and



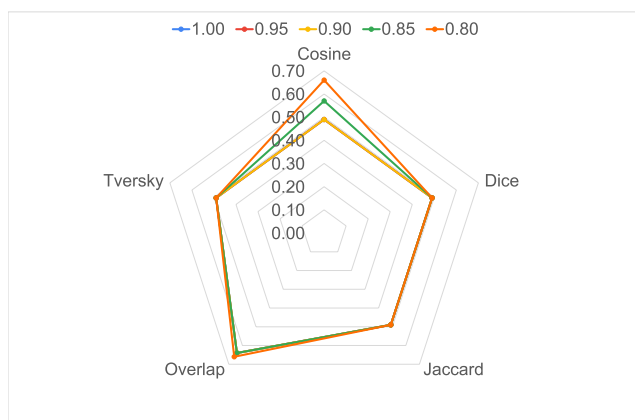
(a) Precision – Sequence-based



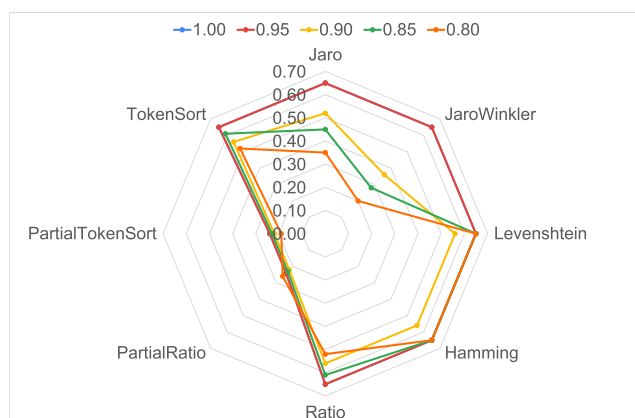
(b) Precision – Set-Based



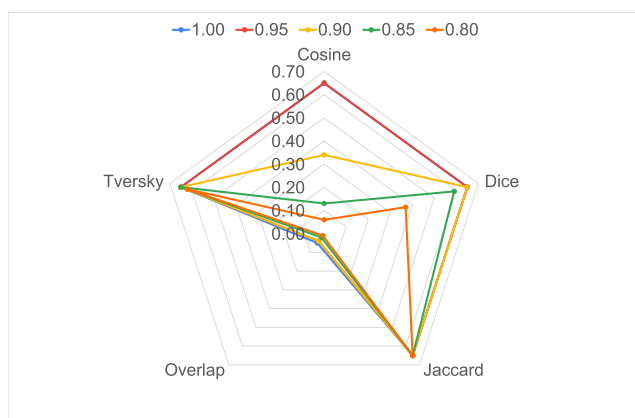
(c) Recall – Sequence-Based



(d) Recall – Set-Based



(e) F-measure – Sequence-Based



(f) F-measure – Set-Based

FIGURE 7. Average precision, recall, and F-measure on different similarity measures and thresholds across the two language pairs French–German and German–Arabic. The greater the distance between the threshold line and the polygon's center, the higher the precision, recall, and F-measure scores.

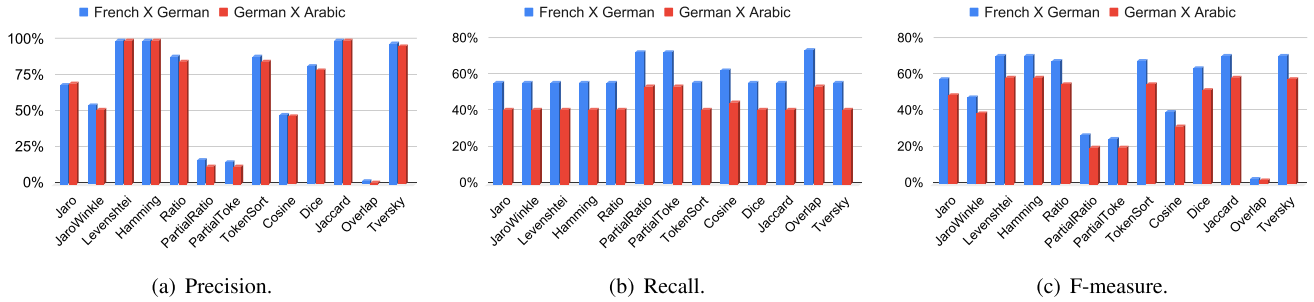


FIGURE 8. Average values (for all thresholds) for precision, recall, and F-measure for matching French–German and German–Arabic ontologies.

TABLE 9. State-of-the-art comparison results using Jaccard, Levenshtein, and Overlap coefficient similarity measures. Red, green, and blue entries are the top scores for precision, recall, and F-measure for each language pair per row.

(a) Using Jaccard similarity.																				
Language pairs	AML [38]			LogMap [39]			LogMapLt [39]			VeeAlign [37]			Wiktionary [40]			MoMatch			MoMatch*	
	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P*	F*
German × French	0.59	0.38	0.29	0.67	0.38	0.27	1.00	0.05	0.03	0.25	0.00	0.00	0.61	0.19	0.11	0.60	0.47	0.39	0.97	0.55
German × Arabic	0.56	0.30	0.21	0.62	0.33	0.23	1.00	0.00	0.00	0.54	0.17	0.10	1.00	0.05	0.03	0.62	0.47	0.39	0.94	0.55
French × Arabic	0.51	0.26	0.18	0.42	0.17	0.11	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.58	0.42	0.33	0.96	0.49
Russian × Arabic	0.41	0.18	0.12	0.67	0.35	0.24	1.00	0.00	0.00	0.18	0.05	0.03	1.00	0.05	0.03	0.52	0.38	0.31	0.79	0.44
German × Russian	0.50	0.29	0.21	0.81	0.35	0.23	1.00	0.00	0.00	0.38	0.03	0.02	0.82	0.25	0.16	0.54	0.43	0.37	0.80	0.50
French × Russian	0.61	0.37	0.27	0.67	0.35	0.25	1.00	0.00	0.00	0.31	0.09	0.06	0.58	0.19	0.12	0.50	0.38	0.31	0.83	0.44
Chinese × Arabic	0.50	0.21	0.14	0.39	0.13	0.08	1.00	0.00	0.00	0.75	0.00	0.00	1.00	0.00	0.00	0.48	0.31	0.24	0.91	0.37
French × Chinese	0.59	0.34	0.24	0.53	0.20	0.13	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.46	0.31	0.24	0.95	0.37
German × Chinese	0.50	0.26	0.18	0.63	0.17	0.10	1.00	0.00	0.00	0.25	0.00	0.00	0.64	0.27	0.17	0.50	0.36	0.30	0.92	0.43
Russian × Chinese	0.65	0.40	0.29	0.72	0.32	0.21	1.00	0.00	0.00	0.75	0.02	0.01	1.00	0.00	0.00	0.40	0.26	0.21	0.84	0.31

(b) Using Levenshtein similarity.																				
Language pairs	AML [38]			LogMap [39]			LogMapLt [39]			VeeAlign [37]			Wiktionary [40]			MoMatch			MoMatch*	
	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P*	F*
German × French	0.59	0.38	0.29	0.67	0.38	0.27	1.00	0.05	0.03	0.25	0.00	0.00	0.61	0.19	0.11	0.55	0.44	0.38	1.00	0.55
German × Arabic	0.56	0.30	0.21	0.62	0.33	0.23	1.00	0.00	0.00	0.54	0.17	0.10	1.00	0.05	0.03	0.58	0.50	0.44	1.00	0.61
French × Arabic	0.51	0.26	0.18	0.42	0.17	0.11	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.58	0.41	0.32	1.00	0.48
Russian × Arabic	0.41	0.18	0.12	0.67	0.35	0.24	1.00	0.00	0.00	0.18	0.05	0.03	1.00	0.05	0.03	0.52	0.36	0.29	0.83	0.42
German × Russian	0.50	0.29	0.21	0.81	0.35	0.23	1.00	0.00	0.00	0.38	0.03	0.02	0.82	0.25	0.16	0.55	0.43	0.36	0.80	0.49
French × Russian	0.61	0.37	0.27	0.67	0.35	0.25	1.00	0.00	0.00	0.31	0.09	0.06	0.58	0.19	0.12	0.50	0.37	0.30	0.83	0.43
Chinese × Arabic	0.50	0.21	0.14	0.39	0.13	0.08	1.00	0.00	0.00	0.75	0.00	0.00	1.00	0.00	0.00	0.50	0.30	0.23	1.00	0.36
French × Chinese	0.59	0.34	0.24	0.53	0.20	0.13	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.48	0.31	0.24	1.00	0.37
German × Chinese	0.50	0.26	0.18	0.63	0.17	0.10	1.00	0.00	0.00	0.25	0.00	0.00	0.64	0.27	0.17	0.53	0.37	0.30	0.97	0.43
Russian × Chinese	0.65	0.40	0.29	0.72	0.32	0.21	1.00	0.00	0.00	0.75	0.02	0.01	1.00	0.00	0.00	0.41	0.26	0.21	0.83	0.31

(c) Using Overlap coefficient similarity.																				
Language pairs	AML [38]			LogMap [39]			LogMapLt [39]			VeeAlign [37]			Wiktionary [40]			MoMatch			MoMatch*	
	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P*	F*
German × French	0.59	0.38	0.29	0.67	0.38	0.27	1.00	0.05	0.03	0.25	0.00	0.00	0.61	0.19	0.11	0.02	0.03	0.56	0.02	0.03
German × Arabic	0.56	0.30	0.21	0.62	0.33	0.23	1.00	0.00	0.00	0.54	0.17	0.10	1.00	0.05	0.03	0.03	0.05	0.69	0.03	0.05
French × Arabic	0.51	0.26	0.18	0.42	0.17	0.11	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.03	0.06	0.52	0.03	0.06
Russian × Arabic	0.41	0.18	0.12	0.67	0.35	0.24	1.00	0.00	0.00	0.18	0.05	0.03	1.00	0.05	0.03	0.03	0.06	0.48	0.03	0.06
German × Russian	0.50	0.29	0.21	0.81	0.35	0.23	1.00	0.00	0.00	0.38	0.03	0.02	0.82	0.25	0.16	0.03	0.06	0.49	0.03	0.06
French × Russian	0.61	0.37	0.27	0.67	0.35	0.25	1.00	0.00	0.00	0.31	0.09	0.06	0.58	0.19	0.12	0.03	0.05	0.44	0.03	0.05
Chinese × Arabic	0.50	0.21	0.14	0.39	0.13	0.08	1.00	0.00	0.00	0.75	0.00	0.00	1.00	0.00	0.00	0.03	0.06	0.42	0.03	0.06
French × Chinese	0.59	0.34	0.24	0.53	0.20	0.13	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.02	0.04	0.38	0.02	0.04
German × Chinese	0.50	0.26	0.18	0.63	0.17	0.10	1.00	0.00	0.00	0.25	0.00	0.00	0.64	0.27	0.17	0.03	0.05	0.43	0.03	0.05
Russian × Chinese	0.65	0.40	0.29	0.72	0.32	0.21	1.00	0.00	0.00	0.75	0.02	0.01	1.00	0.00	0.00	0.03	0.05	0.34	0.03	0.05

vice versa. This result indicates that regardless of language pairs, the resulting scores for precision, recall, and F-measure obtained from the combination of different similarity measures and threshold values stay highly correlated.

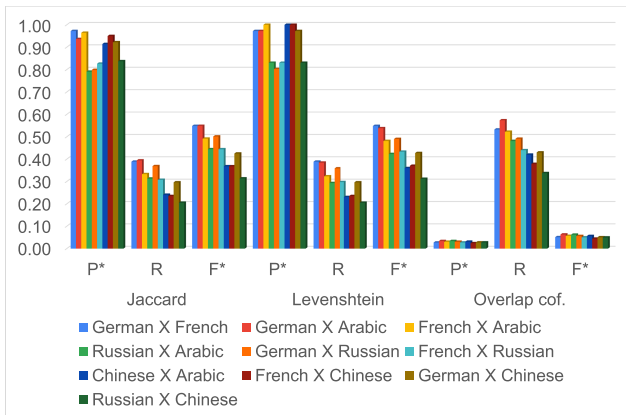
C. COMPARISON WITH THE STATE-OF-THE-ART

In this experiment, we identified five of the related approaches (AML [38], LogMap [39], LogMapLt [39],

VeeAlign [37], Wiktionary [40]) to be included in our evaluation. The other related works neither publish their code nor their evaluation datasets [26], [27], [31]. In order to show the applicability of MoMatch, we use different language pairs. We select the broadest language from each category in Table 7, i.e., German (Germanic), French (Italic), Russian (Balto-Slavic), Chinese (Tai-Kadai), and Arabic (Afro-Asiatic). Therefore, we have ten pairs of language as follows:

TABLE 10. Quality assessment results for input ontologies and the matching results. Bold entries are the top scores.

Input Ontologies	RR ¹²	AR ¹³	IR ¹⁴	RB ¹⁵	IE ¹⁶	MP ¹⁷	RD ¹⁸
<i>Conference_{de}</i>	0.48	0.78	0.83	1.17	0.11	0.02	0.00
<i>Edas_{fr}</i>	0.26	0.29	0.82	0.08	0.77	0.00	0.00
<i>Conference_{ru}</i>	0.48	0.78	0.83	1.17	0.11	0.02	0.00
<i>Ekaw_{ar}</i>	0.32	0.45	0.97	0.02	0.02	0.21	0.00
<i>ConfOf_{de}</i>	0.28	0.34	0.87	1.45	0.04	0.03	0.00
<i>Iasted_{ar}</i>	0.22	0.27	0.94	1.02	0.08	0.10	0.00
Ontology pairs	CP ¹⁹	CR ²⁰	PP ²¹	PR ²²	OV ²³	MC ²⁴	MR ²⁵
<i>Conference_{de} × Edas_{fr}</i>	0.86	0.46	0.50	0.67	0.05	0.09	1.00
<i>Conference_{ru} × Ekaw_{ar}</i>	0.44	0.17	0.33	0.50	0.06	0.11	1.04
<i>ConfOf_{de} × Iasted_{ar}</i>	1.00	0.44	0.00	N/A	0.03	0.05	1.00

**FIGURE 9.** Precision, recall, and F-measure for matching ontologies in ten language pairs using Jaccard vs. Levenshtein vs. Overlap coefficient similarity measures.

German × French, German × Russian, German × Chinese, German–Arabic, French × Russian, French × Chinese, French × Arabic, Russian × Chinese, Russian × Arabic, and Chinese × Arabic. We choose the most effective similarity measures in terms of precision, recall, and F-measure from the previous experiment. Therefore, we select Jaccard and Levenshtein, which has achieved the best precision and F-measure, and Overlap coefficient, which has achieved the best recall. In order to compare our results with the state-of-the-art, we match *Conference* (the middle ontology in Table 6) with *Edas* and *Ekaw* ontologies as mentioned in the results of OAEI 2020⁷. Therefore, there are four ontology pairs in each language pair. For example, in matching French × Chinese, the ontology pairs are: *Conference_{fr} × Edas_{cn}*, *Conference_{cn} × Edas_{fr}*, *Conference_{fr} × Ekaw_{cn}*, and *Conference_{cn} × Ekaw_{fr}*. We evaluate the quality of the matching process by calculating precision, recall, and F-measure as in the previous experiment. Table 9(a), Table 9(b), and Table 9(c) show a comparison between MoMatch’s results for matching ontologies in ten pairs of languages against five state-of-the-art systems using Jaccard, Levenshtein, and Overlap coefficient respectively. We found new correspondences for all ontology pairs which were missing in the gold standard alignments. MoMatch* represents results when considering the new correspondences. It provides the matching results with the adjusted precision and F-measure. LogMapLt

achieves the highest precision of 100% but the lowest recall and F-measure of 0% for all language pairs. Similarly, Wiktionary achieves the highest precision of 100% but the lowest recall and F-measure of 0% for all language pairs with Arabic or Chinese, except for German × Chinese. MoMatch outperforms most other systems in terms of precision, recall, and F-measure when using Jaccard and Levenshtein similarities and not considering the new correspondences as false positives. For instance, in matching German × French ontologies, MoMatch outperforms AML, the highest precision (after LogMapLt), recall, and F-measure among the others in matching German × French ontologies, by 38%, 10%, and 17% in terms of precision, recall, and F-measure respectively when using Jaccard similarity. Similarly, MoMatch outperforms AML by 41%, 9%, and 17% in precision, recall, and F-measure, respectively, when using Levenshtein similarity.

The use of Jaccard and Levenshtein similarity measures give relatively similar results in precision, recall, and F-measure (see Figure 9). While the use of Overlap coefficient similarity in MoMatch achieves the highest recall among the other systems for matching all language pairs. These results confirm our findings from the previous experiment where Jaccard and Levenshtein similarity measures can achieve the best precision and F-measure while the Overlap coefficient can accomplish the best recall.

We calculate the Spearman correlation between F-measure values for Jaccard, Levenshtein, and Overlap coefficient produced by MoMatch, and the ten language pairs. All language pairs achieve the highest correlation of 1.00 except the two language pairs, French × Chinese and Russian × Chinese achieve 0.87. In addition, the correlation between German × Chinese and French × Chinese achieve 0.87 as well. As a result, the ten language pairs are positively and strongly correlated to each other. These results confirm our results from the previous experiment (see subsection VII-B), where the effects of different similarity measures are independent of the language pair.

D. EVALUATING THE QUALITY OF THE MATCHING PROCESS USING QASO

In this experiment, we randomly choose three pairs of ontologies from the previous experiments (*Conference_{de} × Edas_{fr}*, *Conference_{ru} × Ekaw_{ar}*, and *ConfOf_{de} × Iasted_{ar}*). We assess

the quality of the matching process using QASO described in subsection V-A. Table 10 shows the assessment results for each input ontology in addition to the assessment for the matching results. The quality metric results for $Conference_{de}$ and $Conference_{ru}$ are identical because it is the same ontology but in two different natural languages (German and Russian). $Conference$ has the highest results of 48% and 78% in terms of relationship richness and attribute richness, respectively. In terms of inheritance richness, $Ekaw_{ar}$ achieves the highest results of 97%. $Ekaw_{ar}$ is the most ontology that suffers from missing properties information by 21%.

MoMatch achieves the highest class precision of 100% for matching $ConfOf_{de} \times Iasted_{ar}$ and the highest class recall of 46% for matching $Conference_{de} \times Edas_{fr}$. In terms of property precision and recall, MoMatch achieves the highest results of 50% and 67%, respectively, in matching $Conference_{de} \times Edas_{fr}$. The property recall for matching $ConfOf_{de} \times Iasted_{ar}$ is N/A because there are no properties retrieved in the reference alignment ($Ref_p = 0$). MoMatch identifies the highest degree of overlap of 6% in matching $Conference_{ru} \times Ekaw_{ar}$, where it has the highest results for match coverage and match ratio of 11% and 104% respectively.

VIII. CONCLUSION

We propose the MoMatch approach, which matches ontologies in different natural languages. We show a comparative analysis of 13 different string similarity measures. Additionally, we present QASO – a metrics suite for assessing the quality of any ontology and the quality of the matching process. We test the performance of MoMatch over different ontologies in different natural languages, including Indo and non-Indo-European languages. We find that Levenshtein, Hamming, and Jaccard similarity measures have the highest precision and F-measure, while partial ratio, partial token sort, and overlap coefficient have the highest recall for matching multilingual ontologies (RQ1). Accordingly, we sorted the 13 similarity measures into two lists according to precision and recall to support choosing the most appropriate one for the matching process. The correlation between language pairs is consistently high and positive (RQ2). The results of the cross-lingual matching process in MoMatch are found to be promising compared to five state-of-the-art approaches. We assess the quality of the matching process using QASO (RQ3). We show the usability of MoMatch by presenting two use cases in scholarly communication and biomedical domains for both cross-lingual and monolingual ontology matching. MoMatch can be easily adapted for other use cases and domains. In conclusion, MoMatch established the first step toward a multilingual Semantic Web.

In the future, we intend to further; 1) consider individuals in the matching process, 2) include other similarity measures, such as string-based structural measures which consider the entity's neighbors in the matching process, and 3) develop scalable approaches to match large-scale ontologies and knowledge graphs efficiently.

REFERENCES

- [1] G. O. Consortium, "The gene ontology resource: 20 years and still going strong," *Nucleic Acids Res.*, vol. 47, no. 1, pp. 330–338, 2019.
- [2] S. El Kadiri, W. Terkaj, E. N. Urwin, C. Palmer, D. Kiritsis, and R. Young, "Ontology in engineering applications," in *Proc. Int. Workshop Formal Ontologies Meet Industries*. Cham, Switzerland: Springer, 2015, pp. 126–137.
- [3] E. Ong, L. L. Wang, J. Schaub, J. F. O'Toole, B. Steck, A. Z. Rosenberg, F. Dowd, J. Hansen, L. Barisoni, S. Jain, and I. H. De Boer, "Modelling kidney disease using ontology: Insights from the kidney precision medicine project," *Nature Rev. Nephrology*, vol. 16, no. 11, pp. 686–696, 2020.
- [4] C. Trojahn, B. Fu, O. Zamazal, and D. Ritze, "State-of-the-art in multilingual and cross-lingual ontology matching," in *Towards the Multilingual Semantic Web*. Berlin, Germany: Springer, 2014, pp. 119–135.
- [5] J. Euzenat and P. Shvaiko, *Ontology Matching*, 2nd ed. Heidelberg, Germany: Springer-Verlag, 2013.
- [6] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth, and B. Aleman-Meza, "OntoQA: Metric-based ontology quality analysis," in *Proc. IEEE Workshop Knowl. Acquisition From Distrib., Auton., Semantically Heterogeneous Data Knowl. Sources (ICDM)*, Boston, MA, USA, 2005, pp. 45–53.
- [7] S. Mc Gurk, C. Abela, and J. Debatista, "Towards ontology quality assessment," in *Proc. MEPDaWLDQ ESWC*, 2017, pp. 94–106.
- [8] S. Raunich and E. Rahm, "Towards a benchmark for ontology merging," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.* Cham, Switzerland: Springer, 2012, pp. 124–133.
- [9] R. S. I. Wilson, J. S. Goonetillake, A. Ginige, and W. A. Indika, "Ontology quality evaluation methodology," in *Proc. Int. Conf. Comput. Sci. Appl.* Cham, Switzerland: Springer, 2022, pp. 509–528.
- [10] R. Wilson, J. S. Goonetillake, W. Indika, and A. Ginige, "Analysis of ontology quality dimensions, criteria and metrics," in *Proc. Int. Conf. Comput. Sci. Appl.* Cham, Switzerland: Springer, 2021, pp. 320–337.
- [11] S. Ibrahim, S. Fathalla, J. Lehmann, and H. Jabeen, "Multilingual ontology merging using cross-lingual matching," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, Dec. 2020, pp. 113–120.
- [12] A. Bookstein, V. A. Kulyukin, and T. Raita, "Generalized Hamming distance," *Inf. Retr.*, vol. 5, no. 4, pp. 353–375, 2002.
- [13] A. Doan, A. Halevy, and Z. Ives, *Principles of Data Integration*. Amsterdam, The Netherlands: Elsevier, 2012.
- [14] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys.-Dokl.*, vol. 10, no. 8, pp. 707–710, 1966.
- [15] G. A. Rao, G. Srinivas, K. VenkataRao, and P. P. Reddy, "A partial ratio and ratio based fuzzy-wuzzy procedure for characteristic mining of mathematical formulae from documents," *IJSC-ICTACT J. Soft. Comput.*, vol. 8, no. 4, pp. 1728–1732, 2018.
- [16] L. Bulygin, "Combining lexical and semantic similarity measures with machine learning approach for ontology and schema matching problem," in *Proc. Int. Conf. Data Anal. Manag. Data Intensive Domains (DAM-DID/RCDL)*, 2018, pp. 245–249.
- [17] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos, "Ontology population and enrichment: State of the art," in *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*. Berlin, Germany: Springer-Verlag, 2011, pp. 134–166.
- [18] M. Florence, "Building a multilingual ontology for education domain using monto method," *Comput. Sci. Inf. Technol.*, vol. 1, no. 2, pp. 47–53, Jul. 2020.
- [19] A. Guedes, F. Baiao, and K. Revoredo, "On the identification and representation of ontology correspondence antipatterns," in *Proc. 8th Int. Workshop Modular Ontologies (WoMO)*, vol. 1248, 2014, pp. 38–48.
- [20] L. Otero-Cerdeira, F. J. Rodríguez-Martínez, and A. Gómez-Rodríguez, "Ontology matching: A literature review," *Exp. Syst. Appl.*, vol. 42, no. 2, pp. 949–971, Feb. 2015.
- [21] M. K. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Mach. Learn. Appl. Int. J.*, vol. 3, no. 2, pp. 19–28, 2016.
- [22] A. Tversky, "Features of similarity," *Psychol. Rev.*, vol. 84, no. 4, p. 327, Jul. 1977.
- [23] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae, "Challenges for the multilingual web of data," *J. Web Semantics*, vol. 11, pp. 63–71, Mar. 2012.
- [24] M. A. Helou and M. Palmonari, "Cross-lingual lexical matching with word translation and local similarity optimization," in *Proc. 11th Int. Conf. Semantic Syst.*, Sep. 2015, pp. 97–104.
- [25] F. A. Musyaffa, M.-E. Vidal, F. Orlandi, J. Lehmann, and H. Jabeen, "IOTA: Interlinking of heterogeneous multilingual open fiscal DaTA," *Exp. Syst. Appl.*, vol. 147, Jun. 2020, Art. no. 113135.

- [26] B. Fu and R. Brennan, "Cross-lingual ontology mapping and its use on the multilingual semantic web," in *Proc. 1st Workshop Multilingual Semantic Web*, vol. 571, Jan. 2010, pp. 13–20.
- [27] B. Fu, R. Brennan, and D. O'Sullivan, "A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes," *J. Web Semantics*, vol. 15, pp. 15–36, Sep. 2012.
- [28] S. Hertling and H. Paulheim, "WikiMatch: Using Wikipedia for ontology matching," *Ontology Matching*, vol. 946, pp. 1–47, Nov. 2012.
- [29] F. Lin and A. Krizhanovsky, "Multilingual ontology matching based on wiktionary data accessible via SPARQL endpoint," in *Proc. RCDL*, 2011, pp. 1–8.
- [30] C. Trojahn, P. Quaresma, and R. Vieira, "A framework for multilingual ontology mapping," in *Proc. 6th Int. Lang. Resour. Eval. (LREC)*. Marrakesh, Morocco: European Language Resources Association, 2008.
- [31] A. N. Tigrine, Z. Bellahsene, and K. Todorov, "Light-weight cross-lingual ontology matching with LYAM++," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.* Cham, Switzerland: Springer, 2015, pp. 527–544.
- [32] M. Ali, S. Fathalla, S. Ibrahim, M. Kholief, and Y. F. Hassan, "CLOE: A cross-lingual ontology enrichment using multi-agent architecture," *Enterprise Inf. Syst.*, vol. 13, nos. 7–8, pp. 1002–1022, Sep. 2019.
- [33] S. Ibrahim, S. Fathalla, H. S. Yazdi, J. Lehmann, and H. Jabeen, "OECM: A cross-lingual approach for ontology enrichment," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2019, pp. 100–104.
- [34] S. Ibrahim, S. Fathalla, H. S. Yazdi, J. Lehmann, and H. Jabeen, "From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment," in *Proc. Int. Conf. Semantic Syst.* Cham, Switzerland: Springer, 2019, pp. 215–230.
- [35] A. Khiaat, E. A. Ouhiba, M. A. Belfedhal, and C. E. Zoua, "SimCat results for OAEI 2016," in *Proc. OM ISWC*, 2016, pp. 217–221.
- [36] A. Khiaat, "Crolom: Cross-lingual ontology matching system," in *Proc. 15th Int. Semantic Web Conf. (ISWC)*, 2016, pp. 146–152.
- [37] V. Iyer, A. Agarwal, and H. Kumar, "VeeAlign: A supervised deep learning approach to ontology alignment," in *Proc. OM ISWC*, 2020, pp. 216–224.
- [38] B. Lima, D. Faria, F. M. Couto, I. F. Cruz, and C. Pesquita, "OAEI 2020 results for AML and AMLC," in *Proc. OM ISWC*, 2020, pp. 154–160.
- [39] E. Jimenez-Ruiz, "Logmap family participation in the OAEI 2020," in *Proc. 15th Int. Workshop Ontology Matching (OM)*, vol. 2788, 2020, pp. 201–203.
- [40] J. Portisch and H. Paulheim, "Wiktionary matcher results for OAEI 2020," in *Proc. CEUR Workshop*, vol. 2788. Aachen, Germany: RWTH, 2020, pp. 225–232.
- [41] C. Meilicke, R. García-Castro, F. Freitas, W. R. van Hage, E. Montiel-Ponsoda, R. R. De Azevedo, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, A. Tamilin, C. Trojahn, and S. Wang, "MultiFarm: A benchmark for multilingual ontology matching," *J. Web Semantics*, vol. 15, pp. 62–68, Sep. 2012.
- [42] M. A. Helou, M. Palmonari, and M. Jarrar, "Effectiveness of automatic translations for cross-lingual ontology mapping," *J. Artif. Intell. Res.*, vol. 55, pp. 165–208, Jan. 2016.
- [43] G. Sejdiu, I. Ermilov, J. Lehmann, and M. N. Mami, "DistLODStats: Distributed computation of RDF dataset statistics," in *Proc. Int. Semantic Web Conf.* Cham, Switzerland: Springer, 2018, pp. 206–222.
- [44] X. Xue, Y. Wang, W. Hao, and J. Hou, "Optimizing ontology alignments through NSGA-II without using reference alignment," *Comput. Informat.*, vol. 33, no. 4, pp. 857–876, 2014.
- [45] M. A. N. Pour, "Results of the ontology alignment evaluation initiative 2021," in *Proc. CEUR Workshop*, vol. 3063, 2021, pp. 62–108.
- [46] J. Lehmann, G. Sejdiu, L. Buhmann, P. Westphal, C. Stadler, I. Ermilov, S. Bin, N. Chakraborty, M. Saleem, A. C. N. Ngomo, and H. Jabeen, "Distributed semantic analytics using the SANSa stack," in *Proc. Int. Semantic Web Conf.* Cham, Switzerland: Springer, 2017, pp. 147–155.
- [47] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60. [Online]. Available: <http://www.aclweb.org/anthology/P/P14/P14-5010>
- [48] N. Pour, A. Algergawy, R. Amini, D. Faria, I. Fundulaki, I. Harrow, S. Hertling, E. Jimenez-Ruiz, C. Jonquet, N. Karam, and A. Khiaat, "Results of the ontology alignment evaluation initiative 2020," in *Proc. 15th Int. Workshop Ontology Matching (OM)*, vol. 2788, 2020, pp. 92–138.



SHIMAA IBRAHIM received the B.Sc. and M.Sc. degrees from the University of Alexandria, Egypt. She is currently pursuing the Ph.D. degree with the Smart Data Analysis Group, University of Bonn, Germany. Her current research interests include semantic web technologies, natural language processing, machine learning, and knowledge graphs.



SAID FATHALLA received the M.Sc. degree in computer science from the University of Alexandria, Egypt, and the Ph.D. degree in information systems from the University of Bonn, Germany, in 2021. He has several conferences and journals publications in the semantic web, text classification, bioinformatics, scholarly communication, and knowledge engineering. He has published over 40 papers in peer-reviewed journals and international conferences. His current research interests include semantic web technologies, scholarly communication, knowledge graph construction, metadata management, and natural language processing.



JENS LEHMANN received the Ph.D. degree (summa cum laude) from the University of Leipzig and the joint master's degree in computer science from the Technical University of Dresden (TU Dresden) and the University of Bristol. He is currently a Principal Scientist with Amazon, working at Alexa AI on conversational AI, knowledge graphs, and generalized intelligence. He is also an Honorary Professor with TU Dresden. Previously, he was a Full Professor at the University of Bonn and Fraunhofer IAIS. His academic activities at TU Dresden and InfAI support the Smart Data Analytics Research Group. He has authored more than 300 publications, which were cited more than 25000 times. His research interests include semantic web technologies, question answering, machine learning, and knowledge graph analysis. He was selected as a fellow of ELLIS. He is a member of InfAI. He has received 16 international best papers awards, including multiple "Test of Time" awards.



HAJIRA JABEEN received the Ph.D. degree in computer science from the National University of Computing and Emerging Sciences, Islamabad, Pakistan. She is currently a Team Leader in big data analytics with the GESIS-Leibniz Institute for the Social Sciences and an Associate Researcher with the University of Bonn. Her research interests include artificial intelligence, data analytics, knowledge graphs, and scalable machine learning. She has authored many journals articles and conference papers in these areas.

...

Cross-lingual ontology enrichment based on multi-agent architecture

SEMANTiCS 2018 – 14th International Conference on Semantic Systems

Cross-Lingual Ontology Enrichment Based on Multi-Agent Architecture

Mohamed Ali^a, Said Fathalla^{a,b,*}, Shima Ibrahim^{b,c}, Mohamed Kholief^d, Yasser Hassan^a

^a*Faculty of Science, University of Alexandria, Egypt*

^b*Smart Data Analytics (SDA), University of Bonn, Germany*

^c*Institute of Graduate Studies and Research, University of Alexandria, Egypt*

^d*College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt*

Abstract

The proliferation of ontologies and multilingual data available on the Web has motivated many researchers to contribute to multilingual and cross-lingual ontology enrichment. Cross-lingual ontology enrichment greatly facilitates ontology learning from multilingual text/ontologies in order to support collaborative ontology engineering process. This article proposes a cross-lingual ontology enrichment (CLOE) approach based on a multi-agent architecture in order to enrich ontologies from a multilingual text or ontology. This has several advantages: 1) an ontology is used to enrich another one, written in a different natural language, and 2) several ontologies could be enriched at the same time using a single chunk of text (*Simultaneous Ontology Enrichment*). A prototype for the proposed approach has been implemented in order to enrich several ontologies using English, Arabic and German text. Evaluation results are promising and showing that CLOE performs well in comparison with four state-of-the-art approaches.

© 2018 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

Peer-review under responsibility of the scientific committee of the SEMANTiCS 2018 – 14th International Conference on Semantic Systems.

Keywords: cross-lingual ontology enrichment, multi-agent, knowledge management, ontology learning.

1. Introduction

The next decade is likely to witness a considerable rise in ontology learning and enrichment due to the massive utilization of ontologies in different fields of science. In particular, ontologies are widely used in various areas of computer science such as information retrieval, text classification [13], scholarly communication [17, 15, 16] and bioinformatics [14, 12]. Getting into that, more and more ontological resources will become usable on the Semantic Web in different languages. Therefore, it is necessary to develop innovative algorithms which are capable of digesting multilingual information, as well as enriching an ontology using a well-formed one in a different natural language. Most of the existing ontologies have been developed and maintained with a human intervention which is extremely laborious, costly process and time-consuming.

* Corresponding author. Tel.: +4-9163-8260501.

E-mail address: fathalla@cs.uni-bonn.de

Generally, the problem of ontology learning still needs a lot of work [3], in particular, multilingual and cross-lingual ontology learning. According to Gracia et al. [19], multilingual and cross-lingual ontology enrichment is accomplished by using cross-lingual ontology matching. Multilingual and cross-lingual ontology matching process is used for mapping and establishing relationships between ontological resources in different natural languages [26]. Most studies have only tended to focus on learning a specific language. Therefore, we can conclude that there is a big need to create a standard framework for learning and matching ontologies from different languages. This raises a key question: what would be a suitable approach to simultaneously enrich ontologies using multilingual text or another ontology written in a different language? This article proposes a cross-lingual ontology enrichment (CLOE) approach based on a multi-agent architecture in order to enrich ontologies from a multilingual text or ontology. Multi-agent systems have been widely employed in many domains due to the distributed processing offered, efficiency, maintainability and scalability. We introduce a new terminology, *Simultaneous Ontology Enrichment* (SOE), which refers to enriching several ontologies simultaneously. We propose two novel algorithms, for simultaneous ontology enrichment and agents communication. The most prominent feature of the proposed approach is that agents could learn from each other, using a predefined communication scheme, to get the benefit of already-learned concepts found in the input ontologies. In addition, the enrichment of an ontology is performed using another ontology or text in different languages. A prototype for the proposed approach has been implemented in order to enrich several ontologies using English, Arabic and German text. We choose such three languages because they are much familiar to the authors. However, concepts and relationships extraction tasks become more challenging when it concerns the Arabic language. In Arabic, the complexity of spelling, morphology, and semantics makes pre-processing and extraction tasks quite difficult compared to Latin languages [8].

The remainder of this article is structured as follows: we present an overview of related work in [Section 2](#). The proposed approach is described in [Section 3](#). A case study is presented in [Section 4](#). Prototype implementation and preliminary evaluation are presented in [Section 5](#). Finally, we conclude with an outline of the future research in [Section 6](#).

2. Related Work

A recent review of the literature on ontology learning found that most studies tended to focus on learning ontologies from monolingual data sources (e.g., French [18], Arabic [6], German [20], and English [29]). Cross-lingual ontology mapping approaches are used for the automatic construction and enrichment of multilingual or large lexical ontologies [21, 10]. Since more than a million datasets have been published online as linked open data (LOD) from 43 countries in 24 different natural languages [22], cross-lingual ontology mapping becomes a challenging area in the field of ontology learning. Yang and Callan [30] presented a technique to extract concepts from a corpus of public comments. Benabdallah et al. [6] proposed an approach for automatic ontology construction from Arabic corpus by using statistical method for extracting simple and complex terms. In addition, they link these terms by semantic relations using automatic learning of lexical and syntactic markers. Albukhitan and Helmy [1] proposed an ontology learning system from unstructured Arabic text based on NLP tools. Concepts are extracted based on several statistical and data mining based algorithms. Beseiso et al. [7] proposed a multilingual ontology learning algorithm from email. Albukhitan and Helmy [2] presented an interesting automatic ontology maintenance system using multi-agent based approach for multilingual ontologies for two domains: food and health. The maintenance tasks are distributed among a set of agents. Furthermore, agents are used for finding and mapping ontologies. Validating the extracted concepts is performed using the explicit semantic analysis. Rafi et al. [25] proposed a multi-agent based approach towards supporting ontology maintenance involving ontology enrichment and integration. One of the limitations of their approach is that it does not support the maintenance of the ontologies. A good description of the available literature in the field of multilingual ontologies is provided in [27]. A key problem with much of the literature on ontology enrichment is focusing on learning ontologies from monolingual data sources. We address this limitation in our proposed approach by supporting the process of enriching ontologies using multilingual text.

3. Cross-Lingual Ontology Enrichment (CLOE)

The proposed approach comprises three phases: pre-processing and concept extraction, candidate sentence selection, and multi-agent based enrichment. The input is the multilingual text and ontologies to be enriched, and the output is the enriched ontologies. At the end, experts validate the extracted concepts and relations before ingestion to the ontology. The main phases of the proposed architecture are (see Figure 1):

3.1. Pre-processing and Concept Extraction

Text pre-processing is the process of cleaning and preparing the text for further processing. The main aim of this phase is to process input text and make it ready for the next steps, by employing a variety of NLP techniques, *OpenNLP*¹ packages are used. Due to the multilingualism of the input text, it might contain text written in two or more natural languages, so it is important to detect and separate text from different languages.

Language identification: plays a key role in several NLP applications. Depending on the language of the text, appropriate techniques are used for stop words and unnecessary word removal, part-of-speech (POS) tagging, and text annotation. For instance, given an input text $T = \text{"مشغل الاقراص لا يعمل"}$ your DVD drive may not be found", T will be divided into two sentences $T_1 = \text{"مشغل الاقراص لا يعمل"}$ and $T_2 = \text{"your DVD drive may not be found"}$ with labels $L_1 = Ar$ and $L_2 = En$ of the identified languages respectively. Various approaches have been made to master this task [24]. *Apache Tika*² package is used for language identification.

Sentence splitter: splits the text into a list of sentences. It is worth to mention that we keep each sentence as a block of text in the further preprocessing tasks.

Tokenization: divides each sentence to a set of tokens. Tokens separated by delimiters such as white-space characters.

Stop words and unnecessary words filter:, remove words with high frequencies of occurrence, but have no contribution to the subject of text such as pronouns, prepositions, conjunctions, and unnecessary words such as “very”, “really”, and non-alphanumeric contents.

POS tagger, analyzes the text to find out phrase structures and group words according to their syntactic and semantic property. There are nine default tags: Adjective (ADJ), Adverb (ADV), Conjunction (CONJ), Determiner (DET), Noun (N), Number (NO), Preposition (PREP), Pronoun (PRO), and Verb (V). Token verbs are identified to represent a relation. At this stage, the identified verbs in each sentence are counted as a two-argument relation with nouns in the same sentence.

3.2. Candidate Sentence Selection

In this phase, we filter pre-processed sentences in order to eliminate unnecessary/useless sentences. Iterate on sentences to count the number of nouns and noun phrases in a sentence. Then, candidate sentences selected according to the following rules:

- **Rule 1:** Single-concept sentences are ignored, for instance “*Cydia has been installed successfully*”. This sentence has only “*Cydia*” as a noun and does not has any useful information to be kept, therefore it will be ignored.
- **Rule 2:** If the sentence has two POS or more as nouns or noun phrases, then keep the sentence, otherwise discard it. For example “*Cydia has been installed successfully on the iOS device*”. This sentence has two nouns: “*Cydia*” and “*iOS device*”, therefore it will be kept. This is required to extract the relations between concepts.

¹<https://opennlp.apache.org/docs/1.8.2/apidocs/opennlp-tools/overview-summary.html>

²<https://tika.apache.org/1.17/api/>

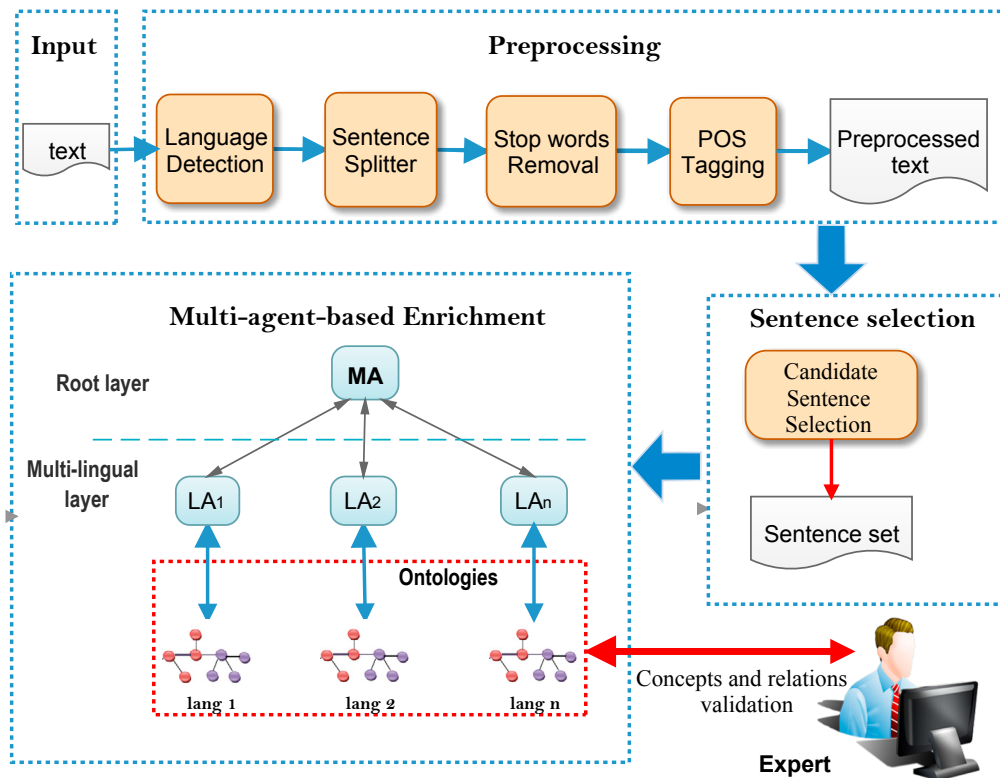


Figure 1: CLOE architecture

3.3. Multi-agent based enrichment

We proposed an innovative enrichment algorithm that allows a set of intelligent agents to interact with each other to learn about their ontologies. There are two types of agents: Master Agent (MA) and Language Agent (LA). Two-level hierarchical multi-agent model has been built where each node is modeled as an agent. The language agents LA_i are at the lowest hierarchy supervised by a single master agent MA as shown in Figure 1. In other words, language agents have the task of learning concepts, relations, and instances for a particular ontology language. For instance, the master agent MA is responsible for managing a set of n language agents ($LA_{i,i=1,2,...,n}$). The number of language agents is the number of languages of the ontologies.

3.3.1. Agents Registration

At birth, LAs register themselves at MA by sending the language of the associated ontology, then MA creates the look-up table LT and gives each one of them a unique id . The look-up table is a 2-dimensional table consists of tuples of $\langle id, Language \rangle$ which is used by MA to decide which LA is responsible for learning the contents of each incoming text based on the language of the text. For each incoming text, MA takes: candidate sentences S , text language L , and look-up table LT as input from the previous phase and then selects the appropriate LA based on L (calling FindLA). A sketch of the main routine is reported in algorithm 1 assuming the following assumptions:

- *Assumption 1:* candidate sentences are vectors of $\langle word, POS \rangle$ pairs,
- *Assumption 2:* input text might have more than one language,
- *Assumption 3:* LAs are registered at MA,
- *Assumption 4:* each LA is responsible for learning concepts and relations between those concepts (taxonomic and non-taxonomic) for a specific language.

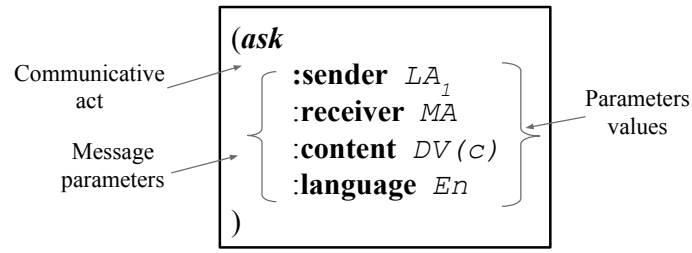


Figure 2: Components of the ACL message

Each LA maintains a comprehensive *translation table* (*TT*). Each translation table lists the concepts that the agent knows and maps them to the corresponding concepts in the associated ontology. English is the common language between all translation tables of all agents. For example, translation table of LA_{Gr} translates concepts from German to English and vice-versa, while the translation table of LA_{Ar} translates concepts from Arabic to English and vice-versa. Only translations that are credible will be recorded in the translation table. Google Cloud Translation API ³ for Java is used by MA to provide the translations for concepts and relations of the given ontologies to be enriched. Currently, Google Translate is considered as the best free online translator and can also be used offline [4]. One of the most prominent features in the proposed approach is that agents can exchange queries and messages to learn about incoming concepts and relations using ontologies, in different languages, of the other agents.

3.3.2. Agents Communication

Communications between agents are carried out using the Foundation for Intelligent Physical Agents (FIPA) Agent Communication Language (ACL) [23]. ACL is based on speech act theory: messages are communicative acts sent by the sender agent to the receiver agents in order to perform some actions. Language agents can learn using two ways: 1) users can teach them by supplying a pre-built ontology and its translation table, and 2) an agent can learn a new experience through its communications with its neighbors by *querying another agent for a certain concept*.

Querying Another Agent for a Certain Concept: The language agent sends out a query to the master agent when it needs to obtain some information about an incoming concept. Agents communication actions are defined in Algorithm 2. For instance, in Figure 2, LA_1 is querying agent MA about the English concept c_i . If the queried agent responds positively with its own semantics, description vector(s) (DV) of matched concept(s), the querying agent takes two actions: 1) enrich the associated ontology by the DV of that concept and 2) update its translation table. When the queried agent processes the query, it first matches the concept to its translation table. If a credible translation is found, then the queried agent simply sends back a description vector about the associated concept translated ($trans_{En}(DV(C))$). Figure 3 shows the complete communication actions taken by two language agents, LA_1 and LA_2 , and the master agent MA when LA_1 wants to get some information about an incoming concept C.

There are two possible dialogues of communication between agents. First, language agent asks the master agent about a certain concept. Then the master agent replies with the DV of this concept. Second, the master agent asks other language agents, except the asking one, about the DV of the concept. If a positive reply is received, then the master agent asks the language agent to *stream-out* the available n DV of this concept. With the completion of these actions, we have succeeded in achieving the goal, i.e. simultaneous ontology enrichment.

³<https://developers.google.com/api-client-library/java/apis/translate/v2>

4. A Case Study

In this case study, we will use an example scenario of a multilingual text, containing English, German and Arabic, from the Information Technology domain. This case study aims to show the whole process starting from submitting the text to the system till getting the candidate concepts and relations. In this scenario, the user submits the following text: "Netbeans fits the pieces together. The license of Netbeans software is available online. Die Firma produziert Netbeans software ist Oracle. تطور البرمجيات بواسطة المبرمج". Then, the system processes the text and produces a set of candidate concepts and relations, taxonomic and non-taxonomic. Finally, experts edit and validate them before adding them to the ontology, to preserve the quality. Table 1 demonstrates the output of each phase starting from the pre-processing phase to producing the set of candidate concepts and relations. Figure 4 shows a small fragment from the enriched ontologies before and after submitting a multilingual text chunk containing English, German and Arabic text. As shown in Figure 4, the German concept (Firma) detected in the input text and had a match in the Arabic ontology with (شركة), then LA_{ar} returns the DV which describes it to the master agent which in turn translates it to German as (Firma) and sends it to LA_{de} to use it to enrich the German ontology. Similarly, the concept (license), which found in the Arabic ontology as (رخصة), is used to enrich the German ontology after translating it to German (Lizenz).

5. Implementation and Preliminary Evaluation

For performing the evaluation, we have implemented all phases of the proposed architecture in Java. In particular, java agent development framework (JADE) is used for the development of the intelligent agent [5]. In our experiments, we consider only text from three languages: English, German and Arabic. All experiments are carried out on Ubuntu 16.04 LTS operating system using Java (JDK 1.8) with an Intel Core i7-4600U CPU @ 2.10GHz x 4 CPU and 10 GB of memory. Within this framework, the expert will make the final decision to select relevant concepts and relations.

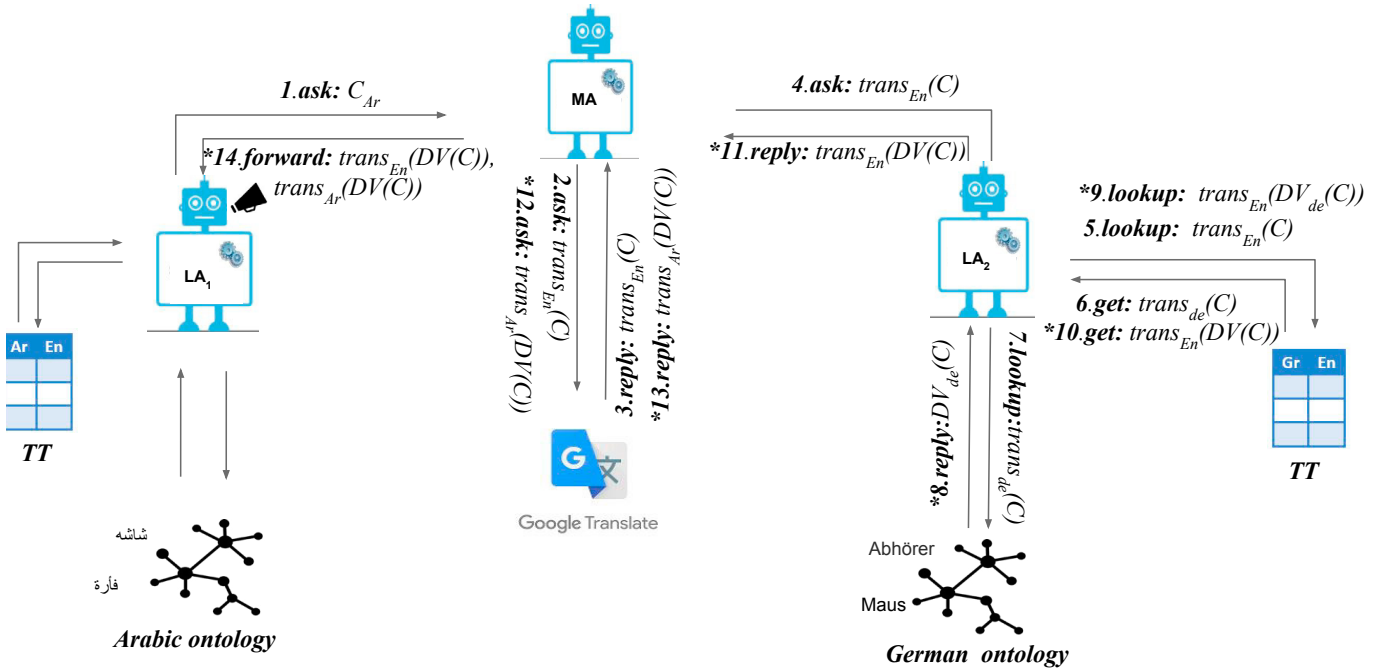


Figure 3: Agent communication actions when LA_1 wants to learn about a new incoming Arabic concept C . The asterisk "*" means that this step might repeat several times if there are more than one DV for a concept.

Table 1: Case study: the output of each phase starting from pre-processing to produce the set of candidate concepts and relations.

Phase		Output
Pre-processing	Lang. identification	En, De, Ar
	Language separation	En: Netbeans fits the pieces together. The license of Netbeans software is available online, DE: Oracle-Firma produziert Netbeans-Software, AR: تطور البرمجيات بواسطة المبرمج
	Sentence splitter and Tokenization	1) Netbeans fits the pieces together. 2) The license of Netbeans software is available online. 3) Oracle-Firma produziert Netbeans-Software. 4) تطور البرمجيات بواسطة المبرمج
	Stop words filter	1) Netbeans fits 2) license Netbeans software is available online. 3) Oracle Firma produziert Netbeans Software. 4) تطور البرمجيات بواسطة المبرمج
	POS tagging	1) Netbeans(N) fits(V) 2) license(N) Netbeans(N) software(N) is(V) available (ADJ) online(N). 3) Oracle(N) Firma(N) produziert(V) Netbeans(N) Software(N). 4) بواسطة (N) البرمجيات (V) تطور (PPREP) المبرمج (N)
Candidate sentence selection		sentence 1 has been removed because it has just one POS as noun.
Ontology enrichment		1. LA_{de} ask MA for $DV_{de}(Firma)$
		2. MA translates $Firma$ to English: $Company$
		3. MA asks LA_{ar} for $Company$
		4. LA_{ar} looks up for $Company$ in its TT and get the translation, in case of found, to Arabic: شركة
		5. LA_{ar} lookup(شركة) in its Arabic ontology and get its DV, in RDF triples as: $\langle \text{http://.../برمجيات} \rangle \langle \text{http://.../منإنتاج} \rangle \langle \text{http://.../شركة} \rangle$
		6. LA_{ar} gets the English translation of these DVs using its TT and sends it back to MA as: $\langle \text{http://.../software} \rangle \langle \text{http://.../producedBy} \rangle \langle \text{http://.../company} \rangle$
		7. MA translates the DV of $Company$, using an online translator, to German: $\langle \text{http://.../software} \rangle \langle \text{http://.../produziertVon} \rangle \langle \text{http://.../Firma} \rangle$
		8. MA sends $DV_{de}(Firma)$ and $DV_{En}(Company)$ to LA_{de}
		9. LA_{de} updates its TT and then enriches its German ontology with the incoming DVs.

Algorithm 1 Ontology Enrichment

```

1: procedure ONTOLOGYENRICHMENT( $S[], L, LT[] []$ )
2:   MA forwards input data to LA;
3:    $LA = FindLA(LT[] [], L);$ 
4:    $LA.Enrich(S[]);$ 
5: end procedure
6: procedure LA.ENRICH( $S[]$ )
7:   for each sentence  $s_i$  in  $S[]$  do
8:     for each concept  $c_i$  in  $s_i$  do
9:        $found = SearchFor(c_i, LA.TT)$ 
10:      if ( $found = True$ )
11:        Skip this concept because it is already exist;
12:      else
13:         $DV[] = AskMA(c_i)$ 
14:        if ( $!DV.isEmpty$ )
15:          Enrich its ontology with the returned DV;
16:          update LA.TT;
17:        else
18:          Use any of the state-of-the-arts tool for learning;
19:        end if
20:      end if
21:    end for
22:  end for
23: end procedure

```

Generally, evaluation of ontology learning systems is the assessment of the resulting ontologies which are used to guide and refine the learning process. According to [11], there are three ways to evaluate a learned ontology. The resulting ontology can be evaluated: 1) in an executable application; 2) by experts or even by 3) comparing it with a predefined reference ontology (gold standard). We followed two strategies for our evaluation: satisfaction questionnaire evaluation and Gold standard-based evaluation.

Satisfaction Questionnaire Evaluation: We first succinctly introduce the evaluation setup and then discuss the result. A total of 24 ontology engineers were recruited for this questionnaire. At the beginning of the evaluation, all the participants have completely understood the approach by giving them a presentation about our approach and a case study describes it. Then, we asked them to fill in a satisfaction questionnaire⁴ with 12 questions to assess the distinct phases of the proposed approach. Figure 5 shows the results of the satisfaction questionnaire analysis.

By analyzing the responses to the questionnaire, we have observed the following observations:

- Among all participants, 70.8% strongly agree that our approach would help ontology engineers.
- About 87.5% of the participants believe that it is *very hard* to manually develop ontologies from a multilingual text, while only 4.2% responds with *Neutral*.
- 62.5% of the participants pointed out that they believe that this approach is considered as a step towards a language-independent approach for ontology enrichment.
- 79% of the participants were satisfied with the *completeness* of the candidate relations and concepts, involving 33% of them are *strongly satisfied*.
- Finally, the overall satisfaction is : 60.9% of the participants were *strongly satisfied*, 21.7% of them were *satisfied* and 17.4% were *Neutral*.

⁴<https://goo.gl/DnnPnv>

Algorithm 2 Agents Communications

```

1: procedure ASKMA( $c_i$ )
2:    $C_{en} = \text{GoogleAPI.trans}_{en}(c_i)$ ; //translate concept  $c_i$  to english
3:   for each  $LA$  connected to  $MA$  do
4:      $DV[] = \text{AskLA}(C_{en})$ ;
5:     if ( $!DV.isEmpty$ )
6:       break;
7:     end if
8:   end for
9:   return  $DV[]$ 
10: end procedure
11: procedure ASKLA( $C_{en}$ )
12:    $found = \text{SearchInTT}(C_{en}, LA.TT)$ ;
13:   if ( $found = True$ )
14:      $C_{LA.lang} = \text{Trans}_{LA.lang}(C_{en}, LA.TT)$ ;
15:      $DV_{lang} = \text{SearchInOntology}(C_{LA.lang})$ ;
16:      $DV_{en} = \text{TranslateFromTT}(C_{LA.lang}, LA.TT)$ ;
17:     return  $DV_{en}$ ;
18:   else
19:     return null
20:   end if
21: end procedure
22: procedure FINDLA( $LT[], L$ )
23:   for each integer  $i$  in  $LA$  do
24:     if ( $LT[i][1] = L$ )
25:       return  $i$ ;
26:     end for
27:   return null
28: end procedure

```

Gold standard-based evaluation: The objective of the gold standard-based evaluation is to determine whether the concepts/relations used in the learning process are correct after comparing them with the gold standard results. We have used a corpus of texts from the 20 Newsgroups dataset⁵. It is a common benchmark used for many learning tasks [9]. The 20 Newsgroups dataset contains approximately 20,000 newsgroup posts, partitioned across 20 different newsgroups which are organized into more extensive categories involving computers, religion, science, and politics. We consider only the computers category (2936 documents). We have divided this category into two subsets. The first one is used for manually, by ontology experts, creating a reference standard ontology. Then, we have fed the system with this subset and saved the output. The second subset is used for evaluating CLOE.

We randomly selected 100 documents from the second subset. Fifty documents out of the 100 were translated using Google API to Arabic and German to be able to submit multilingual text to the system. Inspired by the information retrieval community, the quality of the learned ontology, without expert validation, can also be measured by Precision, Recall and F-measure metrics. The lexical precision (LP) and lexical recall (LR) reflect how good the extracted terms covering the target domain. Lexical precision (LP) is the fraction of retrieved concepts and relations that are relevant, while lexical recall (LR) is the fraction of relevant items retrieved by the system [28]. F-measure (F) is the harmonic mean of precision and recall.

⁵<http://qwone.com/~jason/20Newsgroups/>

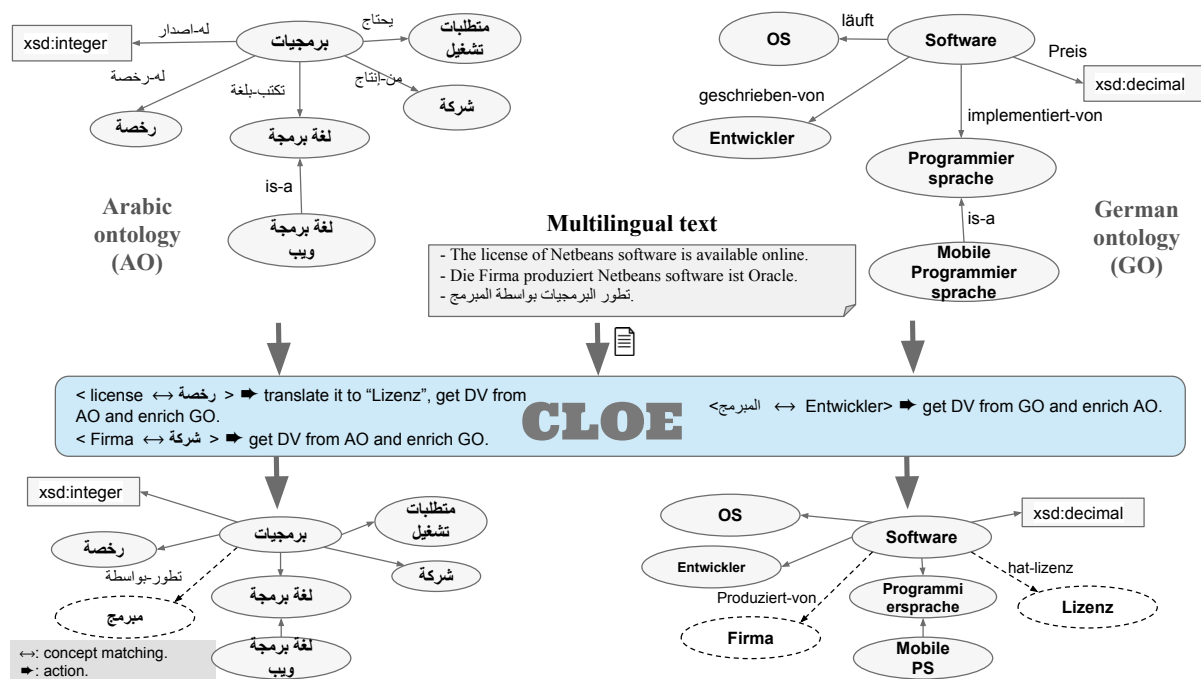


Figure 4: Small fragment from the enriched ontologies before and after submitting a multilingual text chunk containing English, German and Arabic text

We have identified four state-of-the-art approaches from the literature as the comparison approaches. The average of precision, recall, and F-measure are used to evaluate CLOE in comparison to the identified approaches. Table 2 shows the comparison results with the four identified approaches. CLOE outperforms all other systems in recall and F-measure. In average, CLOE achieves a high recall of 99% and a high F-measure of 94% among all the other four systems. CLOE achieved F-measure from 86% to 100% for the extraction of concepts and relations. In 2014, Albukhitan and Helmy [2] achieve a high precision by using explicit semantic analysis to validate the extracted concepts while CLOE outperforms it by 22% in terms of recall and 9% in terms of F-measure. In their paper in 2016, Albukhitan and Helmy [1] do not achieve

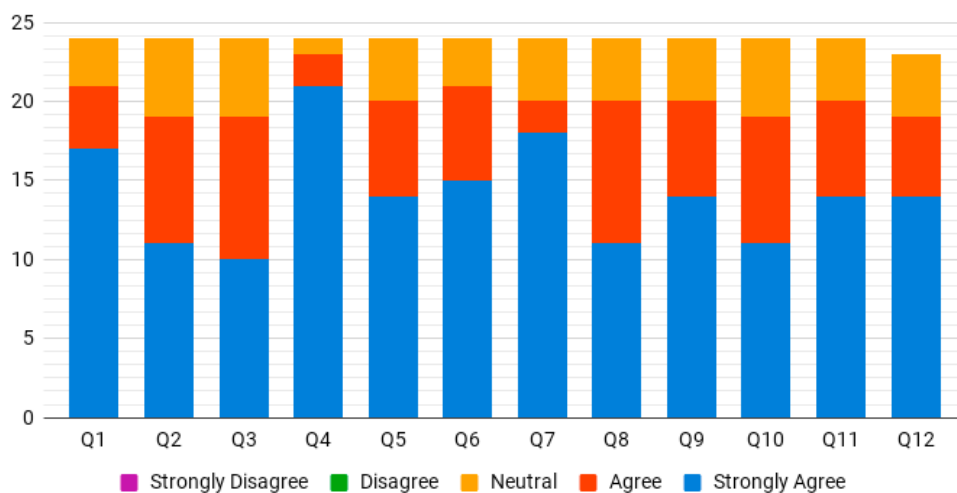


Figure 5: Results of the satisfaction questionnaire analysis

Table 2: Comparison with the state-of-the-art systems.

Metrics	CLOE	Benabdallah et al. [6]	Albukhitan and Helmy 2014 [2]	Albukhitan and Helmy 2016 [1]	Beseiso et al. [7]
LP	0.90	0.90	0.95	0.76	0.89
LR	0.99	0.71	0.77	0.45	0.73
F	0.94	0.79	0.85	0.56	0.80

satisfactory results by using statistical measures for concept recognition based on the frequency of the word in the corpus. Overall, the evaluation results are promising and show a considerable precision, recall, and F-measure. This finding confirms the usefulness of multi-agents in simultaneous ontology enrichment.

6. Conclusion

We present a novel multi-agent-based approach (CLOE) in order to enrich several ontologies (Simultaneous Ontology Enrichment) from multilingual data sources. Each agent is responsible for enriching an ontology for a particular language. Two-level hierarchical multi-agent model has been built where each node is modeled as an agent. The most prominent feature of the proposed approach is that agents could learn from each other, using a predefined communication scheme, to get the benefit of already-learned concepts found in the input ontologies. The main contribution of this work is the usage of a text segment to enrich several ontologies from different languages than the language of the input text. The results are satisfying compared to four state-of-the-art approaches. CLOE outperforms all other systems in terms of lexical precision and lexical recall, and consequently F-measure. In conclusion, this study has gone some way towards cross-lingual ontology enrichment.

Interesting lines of the future research are adapting the proposed approach to be domain independent approach and translate concepts based on language entries in DBpedia instead of Google Translation. Furthermore, techniques to generate multilingual ontologies from monolingual ones will also to be considered.

Acknowledgement

Said Fathalla and Shima Ibrahim would like to thank and acknowledge the Ministry of Higher Education (MoHE) of Egypt for providing scholarships to conduct this study.

References

- [1] Saeed Albukhitan and Tarek Helmy. ‘Arabic Ontology Learning from Un-structured Text’. In: *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*. IEEE. 2016, pp. 492–496.
- [2] Saeed Albukhitan and Tarek Helmy. ‘Multi-agent Based System for Multilingual Ontologies Maintenance’. In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. Vol. 1. IEEE. 2014, pp. 419–423.
- [3] Mohamed Ali et al. ‘The problem learning Non-Taxonomic Relationships of Ontologies from unstructured data sources’. In: *23rd International Conference on Automation and Computing (ICAC)*. 2017, pp. 1–6.
- [4] Gabor Bella, Fausto Giunchiglia and Fiona McNeill. ‘Language and domain aware lightweight ontology matching’. In: *Web Semantics: Science, Services and Agents on the World Wide Web* 43 (2017), pp. 1–17.
- [5] Fabio Bellifemine et al. ‘JADE—a java agent development framework’. In: *Multi-Agent Programming*. 2005, pp. 125–147.
- [6] Ali Benabdallah, Mohammed AlaEddine Abderrahim and Mohammed El-Amine Abderrahim. ‘Extraction of terms and semantic relationships from Arabic texts for automatic construction of an ontology’. In: *International Journal of Speech Technology* (2017), pp. 1–8.

- [7] Majdi Beseiso et al. 'Multilingual Ontology Learning Algorithm for Emails'. In: *Soft Computing Applications and Intelligent Systems*. Springer, 2013, pp. 229–244.
- [8] Ibrahim Bounhas, Wiem Lahbib and Bilel Elayeb. 'Arabic domain terminology extraction: A literature review'. In: *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer. 2014, pp. 792–799.
- [9] Ming-Wei Chang et al. 'Importance of Semantic Representation: Dataless Classification'. In: *AAAI*. Vol. 2. 2008, pp. 830–835.
- [10] Gerard De Melo and Gerhard Weikum. 'Towards a universal wordnet by learning from combined evidence'. In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 513–522.
- [11] Klaas Dellschaft and Steffen Staab. 'On how to perform a gold standard based evaluation of ontology learning'. In: *International Semantic Web Conference*. Vol. 4273. Springer. 2006, pp. 228–241.
- [12] Said Fathalla. 'Detecting Human Diseases Relatedness: A Spreading Activation Approach Over Ontologies'. In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 14.3 (2018), pp. 120–133.
- [13] Said M Fathalla, Yasser F Hassan and Maged El-Sayed. 'A hybrid method for user query reformation and classification'. In: *Computer Theory and Applications (ICCTA), 2012 22nd International Conference on*. IEEE. 2012, pp. 132–138.
- [14] Said Fathalla and Yaman Kannot. 'A Bidirectional-Based Spreading Activation Method for Human Diseases Relatedness Detection Using Disease Ontology'. In: *Conference on Computational Collective Intelligence Technologies and Applications*. Springer. 2017, pp. 14–23.
- [15] Said Fathalla et al. 'Analysing Scholarly Communication Metadata of Computer Science Events'. In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2017, pp. 342–354.
- [16] Said Fathalla et al. 'SemSur: A Core Ontology for the Semantic Representation of Research Findings'. In: *Proceedings of the 14th International Conference on Semantic Systems*. ACM. 2018. In press.
- [17] Said Fathalla et al. 'Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles'. In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2017, pp. 315–327.
- [18] David Faure and Thierry Poibeau. 'First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX'. In: *Proceedings of the ECAI workshop on Ontology Learning*. 2000.
- [19] Jorge Gracia et al. 'Challenges for the multilingual web of data'. In: *Web Semantics: Science, Services and Agents on the World Wide Web* 11 (2012), pp. 63–71.
- [20] Udo Hahn and Martin Romacker. 'The SYNDIKATE text Knowledge base generator'. In: *Proceedings of the first international conference on Human language technology research*. 2001, pp. 1–6.
- [21] Mamoun Abu Helou and Matteo Palmonari. 'Cross-lingual lexical matching with word translation and local similarity optimization'. In: *Proceedings of the 11th International Conference on Semantic Systems*. ACM. 2015, pp. 97–104.
- [22] *International Open Government Data Search*. https://logd.tw.rpi.edu/iogds_analytics_2. [Online; accessed: 2017-12].
- [23] Yannis Labrou, Tim Finin and Yun Peng. 'Agent communication languages: The current landscape'. In: *IEEE Intelligent Systems and Their Applications* 14.2 (1999), pp. 45–52.
- [24] Marco Lui, Jey Han Lau and Timothy Baldwin. 'Automatic detection and language identification of multilingual documents'. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 27–40.
- [25] Muhammad Rafi, Hilal Qureshi and Hasina Khatoun. 'Ontology Maintenance via Multi-agents'. In: *INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on*. IEEE. 2009, pp. 955–959.
- [26] Dennis Spohr, Laura Hollink and Philipp Cimiano. 'A machine learning approach to multilingual and cross-lingual ontology matching'. In: *The Semantic Web-ISWC 2011* (2011), pp. 665–680.
- [27] Cássia Trojahn et al. 'State-of-the-art in multilingual and cross-lingual ontology matching'. In: *Towards the Multilingual Semantic Web*. Springer, 2014, pp. 119–135.
- [28] Wilson Wong, Wei Liu and Mohammed Bennis. 'Ontology learning from text: A look back and into the future'. In: *ACM Computing Surveys (CSUR)* 44.4 (2012), p. 20.
- [29] Takahira Yamaguchi. 'Acquiring Conceptual Relationships from Domain-Specific Texts.' In: *Workshop on Ontology Learning*. Vol. 38. 2001, pp. 69–113.
- [30] Hui Yang and Jamie Callan. 'Ontology generation for large email collections'. In: *Proceedings of the 2008 international conference on Digital government research*. Digital Government Society of North America. 2008, pp. 254–261.

CLOE: a cross-lingual ontology enrichment using multi-agent architecture

CLOE: a cross-lingual ontology enrichment using multi-agent architecture, Mohamed Ali, Said Fathalla, Shima Ibrahim, Mohamed Kholief, and Yasser F. Hassan, Enterprise Information Systems, and © copyright 2019, reprinted by permission of Informa UK Limited, trading as Taylor & Taylor & Francis Group, <http://www.tandfonline.com>. DOI: <https://doi.org/10.1080/17517575.2019.1592232>

CLOE: A Cross-Lingual Ontology Enrichment Using Multi-Agent Architecture

Mohamed Ali^a, Said Fathalla^{a b}, Shima Ibrahim^{b c}, Mohamed Kholief^d and Yasser F. Hassan^a

^aFaculty of Science, Alexandria University, Alexandria, Egypt; ^bSmart Data Analytics (SDA), Bonn University, Bonn, Germany; ^cInstitute of Graduate Studies and Research, Alexandria University, Alexandria, Egypt; ^dCollege of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt

ARTICLE HISTORY

Compiled February 2, 2021

ABSTRACT

Nowadays, the increasing quantity of multilingual data on the Web, and the consequent development of ontologies in different natural languages has motivated many researchers to work on multilingual and cross-lingual ontology enrichment. This article proposes a cross-lingual ontology enrichment (CLOE) approach based on a multi-agent architecture, in order to enrich ontologies from a multilingual text or ontology. A three-level hierarchical multi-agent model has been built, where each node is modelled as an agent. The most prominent feature of the proposed approach is that agents can learn from each other, using a predefined communication scheme, to benefit from the already-learned concepts found in the input ontologies. This has several advantages: 1) a well-formed ontology could be used to enrich another one written in a different natural language, 2) several ontologies could be enriched at the same time using multilingual text (simultaneous ontology enrichment) and 3) it is a domain-independent approach. We have implemented a prototype for the proposed approach in order to enrich several ontologies using English, Arabic and German texts. Evaluation results are promising, showing considerable precision and recall scores. We believe that this approach is the first such attempt to enrich multiple ontologies using a domain-independent approach.

KEYWORDS

Cross-lingual Ontology enrichment, Natural Language Processing, Semantic Web, Knowledge Management, Ontology Mapping.

1. Introduction

As the amount of unstructured data increases per day or per hour or even per minute as the motivation of representing the knowledge in this data in the machine-readable and understand form is increased. Ontologies occupy a key role in capturing space knowledge and providing a standard understanding of knowledge. Most of the existing ontologies are developed and maintained via human intervention, which is an extremely laborious, costly and time-consuming process. A massive research has been done on building ontology taxonomies semi-automatically (Fortuna et al. 2006; Wang et al. 2006) and automatically (Cimiano et al. 2004; Cimiano and Völker 2005). The main challenge in ontology building is concepts and

CONTACT Fathalla: fathalla@cs.uni-bonn.de, Ali: M_hassan@nbe.com.eg, Shima: ibrahim@cs.uni-bonn.de, Kholief: kholief@aast.edu and Fouad: y.fouad@alexu.edu.eg

relationships extraction. The next decade is likely to witness a considerable rise in ontology learning and enrichment due to the massive utilization of ontologies in different fields of science. Hence, more and more ontological resources will become available on the Semantic Web in different languages. It is necessary to develop innovative algorithms which are capable of digesting multilingual information, as well as of enriching an ontology using a well-formed one in a different natural language. The problem of multilingual and cross-lingual ontology enrichment still needs a considerable of work. According to (Gracia et al. 2012), the multilingual and cross-lingual ontology enrichment is accomplished by using cross-lingual ontology matching. The multilingual and cross-lingual ontology matching process is used for mapping and establishing relationships between ontological resources in different natural languages (Spohr et al. 2011). Few researchers have addressed the problem of multilingual and cross-lingual ontology enrichment. Most studies have only tended to focus on learning in a specific language. Therefore, we can conclude that there is a pressing need to create a standard framework for learning and matching ontologies in different language.

The main objective of this work is to enrich ontologies from multilingual text using domain-independent approach. The key questions are: what would be a suitable approach to enrich an ontology using another ontology written in a different language? Moreover, what would be a suitable approach to simultaneously enrich ontologies from different domains using multilingual text? In this paper, we present an innovative approach that relies on a multi-agent architecture in order to enrich several ontologies, in different languages, belong to different domains, from a multilingual text. The main contribution of this study is simultaneously enrich several ontologies in different languages from multilingual text using a domain-independent approach. A striking feature of the proposed work is that the enrichment of an ontology is performed using another ontology or text in different languages. Our research challenges the problems of:

- Preprocessing multilingual text because each language has different preprocessing techniques. For example, stop word elimination is different for each language,
- Matching concepts between ontologies in different languages. For example, “*Firma*” and “Company”,
- Handling the communication between agents,
- Eliminating unnecessary text from the input (i.e. candidate sentence selection),
- How can agents learn from each other?

The remainder of this article is structured as follows, We provide background information about the topic of our research in section 2. We present an overview of related work in section 3. The proposed approach for cross-lingual ontology enrichment is described in section 4. A case study is presented in section 5. Prototype implementation and preliminary evaluation are presented in section 6. Finally, we conclude with an outline of the future research in section 7.

2. Background

The explosive growth of data makes information retrieval, information extraction or any other applications use this data a very hard task. This data contains very useful information, however it is a human-understandable not machine-understandable (Browarnik and Maimon 2015). Therefore, there is a pressing need to represent this data in such a way to be understood by both human and machine. Ontology is considered as one of the major cornerstones of representing the information in a more meaningful way. Basically, ontologies provide a machine-understandable semantics of knowledge to enhance the search process and used

for many targets that associated with their usability of data, data analysis, and data sharing (Sharma and Rastogi 2014). Ontologies occupy a key role in capturing space knowledge and providing a machine standard understanding the semantics of knowledge in many computer science areas, such as data mining Fathalla et al. (2012a), scholarly communication Fathalla et al. (2017); Fathalla and Lange (2018); Fathalla et al. (2018) and bioinformatics Fathalla and Kannot (2017); Fathalla (2018).

Definition 1. Formally, an ontology can be represented by a tuple of six components (Sabou 2004): $O = (C, H, I, R, P, A)$ where, C is the set of ontology domain entities (i.e. classes /concepts) of the ontology, H is the set of taxonomic relations between classes, I is the set of relations between ontology classes and its instances, R is the set of ontology relations which are neither *is-a* nor *type-of*, i.e. non-taxonomic relations, P is the set of properties of the classes and A is the set of axioms and rules used for checking and verifying ontology consistency and new knowledge inference.

The manual development of ontologies is an expensive and time-consuming task, because domain experts and knowledge engineers are required for this task. After reviewing the literature, we found that a formal definition of multilingual and cross-lingual ontology enrichment is lacking. In addition, there is no clear distinction between the two notions: multilingual and cross-lingual ontology enrichment (Fu et al. 2009). We define monolingual, multilingual and cross-lingual ontology enrichment based on the general definition of ontology enrichment as the process of extracting ontological concepts from unstructured text in order to put these new concepts in the ontology. Given a source ontology SO , a target ontology TO and a text T , and a set of natural languages of the source and target ontologies respectively $L(SO)$ and $L(TO)$. Inspired by related definitions in established fields such as multilingual and cross-lingual information retrieval (Peters et al. 2012), this section defines monolingual, multilingual and cross-lingual ontology enrichment as follows.

Definition 2. Monolingual ontology enrichment is the process of enriching TO by extracting concepts from T or SO , both written in a single natural language. More formally, $(L(SO) = L(TO)) \wedge (L(SO) = L(T)) \wedge (|L(SO)| = |L(TO)| = |L(T)|)$.

Definition 3. Multilingual ontology enrichment is the process of enriching TO by extracting concepts from T or SO , but they share a common natural language(s), but at least one of them has two or more natural languages within itself. More formally, $(L(SO) \cap L(TO) \cap L(T) \neq \phi) \wedge (|L(SO)| \geq 2 \vee |L(TO)| \geq 2 \vee |L(T)| \geq 2)$.

Definition 4. Cross-Lingual ontology enrichment is the process of enriching TO by extracting concepts from T or SO , but both share no common natural language. More formally, $(L(SO) \cap L(TO) \cap L(T) = \phi)$.

Definition 5. Ontology matching is “ a function f which, from a pair of ontologies to match SO and TO , an input alignment A , a set of parameters p and a set of resources r , returns an alignment A' between these ontologies: $A' = f(SO, TO, A, p, r)$ ” (Euzenat et al. 2007).

3. Related Work

A recent review of the literature on ontology learning found that most studies tended to focus on learning ontologies from monolingual data sources for domain-specific ontologies, i.e domain-dependent (Table 1). Learning ontologies from multilingual data sources using a domain-independent approach has become an intriguing area in the field of ontology learning. Cross-lingual ontology mapping approaches are used for the automatic construction and enrichment of multilingual or large lexical ontologies (De Melo and Weikum 2009; Helou and Palmonari 2015). Since more than a million datasets have now been published online as linked

Table 1.: A comparison between the state of the art tools for ontology learning

Tool	Lang.	Learned items	Approach
ASIUM (Faure and Poibeau 2000)	French	verb frames, concepts, taxonomic relations	conceptual clustering
DODDLE-II (Yamaguchi 2001)	English	taxonomic, non-taxonomic relations	analysis of lexical co-occurrence statistics, association rule algorithm
SYNDIKATE (Hahn and Romacker 2001)	German	concepts, taxonomic, non-taxonomic relations	association rules, formal concept analysis, hierarchical clustering analysis
HASTI (Shamsfard and Barforoush 2002)	Persian	concepts, taxonomic, non-taxonomic relations, basic axioms	linguistic-based, semantic analysis, logical reasoning, syntactic templates
TEXT2ONTO (Montoyo et al. 2005)	Spanish, German	concepts, taxonomic, non-taxonomic relations	linguistic preprocessing, NLP, RTF/TF-IDF metrics
CRCTOL (Jiang and Tan 2010)	English	concepts, taxonomic and non-taxonomic relations	NLP, statistical and lexico-syntactic methods, rule based algorithm and association rule mining
OntoCmaps (Zouaq et al. 2011)	English	taxonomic and non-taxonomic relations	syntactic patterns, TF-IDF metrics and NLP
PROMINE (Gillani Andleeb 2015)	English	concepts, basic axioms, taxonomic and non-taxonomic relations	linguistic preprocessing, NLP, text mining and statistical and semantic measures
(Albukhitan and Helmy 2016)	Arabic	concepts, taxonomic and non-taxonomic relations	GATE NLP text analysis, RTF/TF-IDF metrics
(Benabdallah et al. 2017)	Arabic	concepts, taxonomic relations	NLP, statistical measures, linguistic markers

open data (LOD), from 43 countries in 24 different natural languages (IOGDS 2017), cross-lingual ontology mapping has become a challenging area in the field of ontology learning. Park et al. (2010) have divided the ontology learning tools into three categories: 1) Ontology editing tools, which provide the help to ontology engineers for acquiring, organizing, and visualizing domain knowledge; 2) Ontology merging tools, which create one coherent ontology from two or more existing ontologies; and 3) Ontology enrichment tools, which extract concepts and/or relations from unstructured data by using NLP and/or machine learning techniques. Yang and Callan (2008) presented a technique to extract concepts from a corpus of public comments. Candidate concepts were identified by n-gram mining. WordNet (Fellbaum 1998) and surface text pattern matching (Ravichandran and Hovy 2002) are used to identify relationships among the concepts and to guide the organization of concepts into intended hierarchical relationships. The research efforts that support ontology construction for the Arabic language on the semantic web still restricted because of the complexity related to the NLP of Arabic data resources (Al-Khalifa and Al-Wabil 2007). There are two categories of research

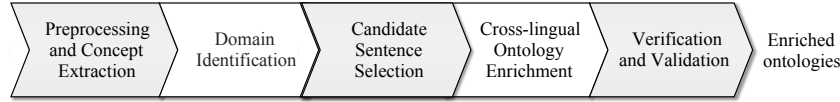


Figure 1.: The main phases of CLOE

regarding the Arabic language in the semantic web (Al-Zoghby et al. 2013). First, the development of Arabic ontologies (Ishkewy et al. 2014; Soudani et al. 2014). Second, the exploitation of ontologies for rising Arabic named entity extraction (Boudabous et al. 2013; Zaidi et al. 2012). Benabdallah et al. (2017) proposed an approach for automatic ontology construction from an Arabic corpus using a statistical method for extracting simple and complex terms. In addition, they link these terms by semantic relations using automatic learning of lexical and syntactic markers. Albukhitan and Helmy (2016) proposed an ontology learning system from unstructured Arabic text based on NLP tools. Concepts are extracted based on several statistical and data mining-based algorithms. Beseiso et al. (2013) proposed a multilingual ontology learning algorithm for email. Albukhitan and Helmy (2014) present an interesting automatic ontology maintenance system using a multi-agent based approach for multilingual ontologies for two domains: food and health. The maintenance tasks are distributed among a set of agents. Furthermore, agents are used for finding and mapping ontologies. Validating the extracted concepts is performed using the explicit semantic analysis. Rafi et al. (2009) proposed a multi-agent based approach for supporting ontology maintenance involving ontology enrichment and integration. Individual agents are for solving each specific problem of ontology maintenance such as merging and integrating ontologies, aligning ontologies and mapping ontologies. However, one of the limitations of their approach is that it does not support the maintenance of multiple domains. We are addressing this limitation in our proposed approach by supporting the process of enriching ontologies from multiple domains and languages. An interesting work in linguistic enrichment of ontologies is presented in (Buitelaar et al. 2009; Montiel-Ponsoda et al. 2011). A good description of the available literature in the field of multilingual ontologies is provided in (Fu et al. 2010; Trojahn et al. 2014).

What additionally distinguishes our work from the related work mentioned above is that our approach can enrich several ontologies using multilingual data sources and other ontologies, even in a different natural language, using a domain-independent approach.

4. The Proposed Approach

The proposed approach comprises five phases: text preprocessing and concept extraction, domain identification, candidate sentence selection, cross-lingual ontology enrichment, and ontology verification and validation. The input is a text that might be written in different natural languages and the ontologies to be enriched, and the output is the enriched ontologies. The main phases of the proposed architecture are shown in Figure 1 and explained below in detail.

4.1. Text Pre-processing and Concept Extraction

Text preprocessing is the process of cleaning and preparing the text for further processing. The main aim of this phase is to process input text by employing a variety of NLP techniques

to make it ready for the next steps. *OpenNLP*¹ packages are used in this work for several preprocessing tasks. This phase involves five sub-tasks:

- (1) **Language identification and separation:** as the input text might contain text written in two or more different natural languages, it is important to detect and separate the text in different languages. Language identification plays a key role in several NLP applications. Depending on the language of the text, appropriate techniques are used for stop words and unnecessary word removal, part-of-speech (POS) tagging, and text annotation. For instance, given an input text $T = \text{"مشغل الاقراص لا يعمل"}$ your DVD drive may not be found", T will be divided into two sentences $T_1 = \text{"مشغل الاقراص لا يعمل"}$ and $T_2 = \text{"your DVD drive may not be found"}$ with labels $L_1 = Ar$ and $L_2 = En$ for the identified languages respectively. Various approaches have been used to master this task (Lui et al. 2014; Souter et al. 2017). *Apache Tika*² package is used for language identification.
- (2) **Sentence Splitter:** detects the sentence boundaries in a raw input text.
- (3) **Tokenization:** each sentence is divided into a set of tokens. Tokens separated by delimiters such as white-space characters.
- (4) **Stop words and unnecessary words filter:** Remove words, i.e stop words, that have a high frequency of occurrence, but make no contribution to the subject of the text, such as pronouns, prepositions, conjunctions, and unnecessary words such as "very", "really" and non-alphanumeric contents.
- (5) **Part-of-Speech (POS) tagging:** analyses the text to determine the phrase structures and to group words according to their syntactic and semantic properties.

There are nine default tags: adjective (ADJ), adverb (ADV), conjunction (CONJ), determiner (DET), noun (N), number (NO), preposition (PREP), pronoun (PRO) and verb (V). Token verbs are identified as relations. At this stage, the identified verbs in each sentence are counted, as a two-argument relation with nouns in the same sentence. A Java implementation of the log-linear part-of-speech taggers is described in (Toutanova et al. 2003).

4.2. Domain Identification

In this phase, we have adopted the classification approach proposed by (Fathalla et al. 2012b) in order to identify the domain of the input text. The input text T is possibly classified to more than one domain $D = \{d_1, d_2, \dots, d_n\}$ with percentage of relevance using a fuzzy relation $\mu_d(w) : T \times D \rightarrow [0, 1]$. Equation 1 shows the possible values of membership. The larger the value, the more relevant is the text to the domain.

$$\mu_d(w) = \begin{cases} 0 & , w \notin d \\ (0, 1) & , w \in d_1, d_2, \dots, d_n \\ 1 & , w \in d \end{cases} \quad (1)$$

Let $TS = \{(t_1, c(t_1)), (t_2, c(t_2)), \dots, (t_n, c(t_n))\}$ be a set of training dataset containing n pre-classified texts t_i along with its classification $c(t_i)$. Each text in the training set is represented by a set of term-frequency pairs: $t = \{(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)\}$ where f_j is the occurrence frequency of the word w_j in t . The membership value $\mu_d(w_k)$ is calculated

¹<https://opennlp.apache.org/docs/1.8.2/apidocs/opennlp-tools/overview-summary.html>

²<https://tika.apache.org/1.17/api/>

by

$$\mu_d(w_k) = N_d(w_k)/N_D(w_k) \quad (2)$$

Given that, N is the number of words in t and $\mu_d(w_k)$ (calculated by Equation 2), the classification of t is calculated by

$$C_d(t) = \sum_{k=1}^N \mu_d(w_k) \quad (3)$$

Since the text might be classified to more than one domain, i.e. M domains with a degree of relevance, the dominant classification of the text, denoted by $C_{dom}(t)$, is determined by Equation 4.

$$C_{dom}(t) = \max_{(i=1 \dots M)} \{C_i(t)\} \quad (4)$$

For a special case, if the input text is classified to only one class, i.e. $M=1$, then the dominant classification is the resultant domain.

4.3. Candidate Sentence Selection

The following steps are carried out during candidate sentence selection phase in order to filter the sentence set to eliminate unnecessary/useless sentences:

- (1) Single-concept sentences are ignored because there is no sense or useful information to be kept.
- (2) Iteration is performed on sentences to count the number of nouns and noun phrases in the sentence. If the sentence has two or more POSs which are nouns or noun phrases, then keep the sentence; otherwise, discard it. This is required in order to extract the relations between concepts.

4.4. Cross-lingual Ontology Enrichment

Multi-agent systems have been widely adopted in many application domains due to the distributed processing offered and their efficiency, maintainability and scalability (Balaji and Srinivasan 2010). Multi-agent systems allow intelligent agents to interact with each other using a communication scheme. In order to achieve the goal, we proposed an innovative enrichment algorithm that allows a set of intelligent agents to interact with each other to learn about their ontologies. There are three types of agents: a master agent (MA), domain agents (DA), and language agent (LA). A three-level hierarchical multi-agent model was built, where each node is modelled as an agent. The language agents LA_{ij} are in the lowest hierarchy supervised by a set of domain agents DA_i in a higher level in the hierarchy and finally a single master agent MA in the top of hierarchy supervises the domain agents as shown in Figure 2. In other words, language agents have the task of learning concepts, relations and instances for a particular ontology and language. For instance, the domain agent DA_1 is responsible for managing a set of n language agents $LA_{1,i=1,2,\dots,n}$. The number of domain agents depends on the number of domains in the ontologies to be enriched and the number of language agents depends on the number of languages of the ontologies in a particular domain. Initially, DAs

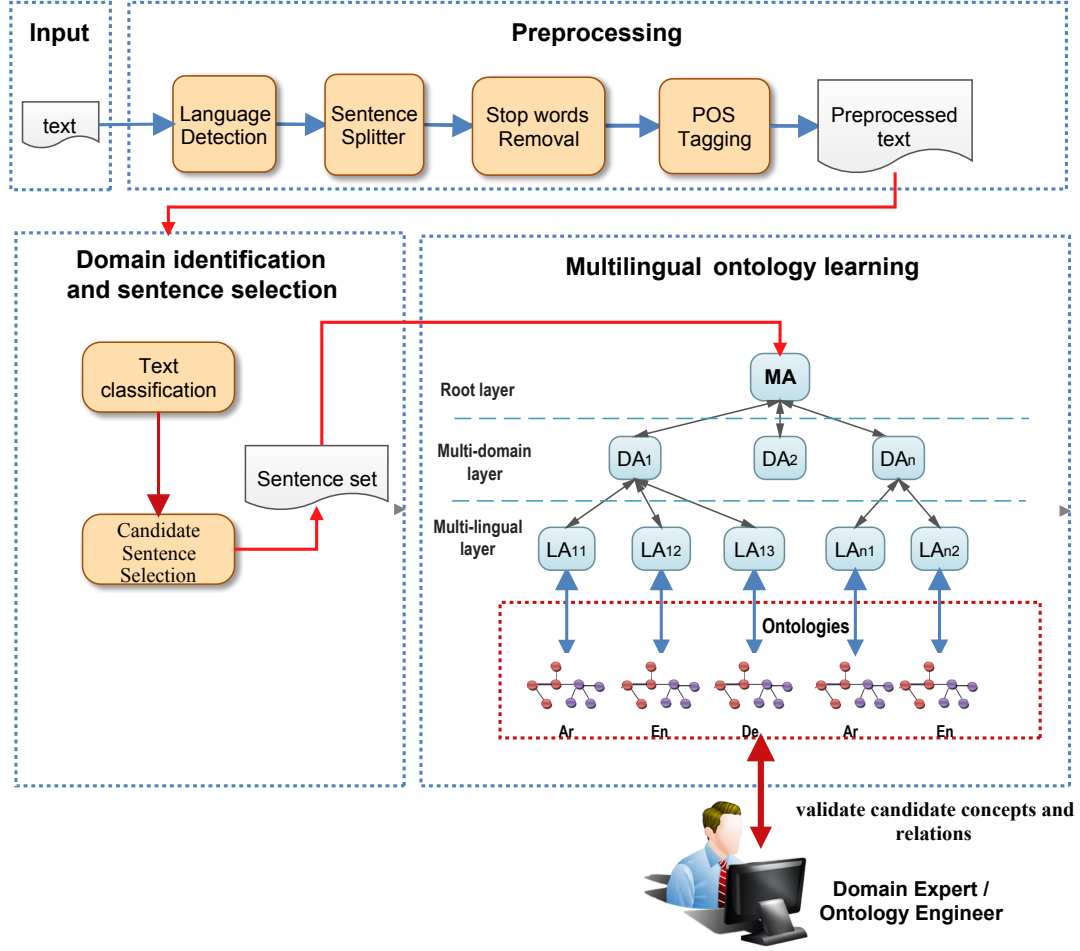


Figure 2.: CLOE architecture

and LAs register themselves with the MA by sending the domain and the language of the associated ontology; then the MA creates the look-up table LT and gives each LA a unique id . Communications between agents were carried out using the Foundation for Intelligent Physical Agents (FIPA) Agent Communication Language (ACL) (Labrou et al. 1999).

4.4.1. Agents Registration

The look-up table is a two-dimensional table consisting of tuples of $\langle id, domain, Language \rangle$. It is used by the MA to manage and store information about agents involved in the hierarchy. This table is used by the MA to decide which DA is responsible for learning the contents of each incoming text, based on the language of the text. Figure 3 shows the sequence diagram for agents registration. For each incoming text, the MA takes the candidate sentences S , text classification D , text language L and look-up table LT as inputs from the previous phase and then selects the appropriate DA based on D (calling $FindDA$ in Algorithm 1). Similarly, domain agents then select the appropriate LA based on the L (calling $FindLA$). A sketch of the main routine is reported in Algorithm 1, assuming the following six hypotheses.

- Candidate sentences consist of a set of sentences each of which consists of vectors of $\prec word, POS \succ$ pairs.

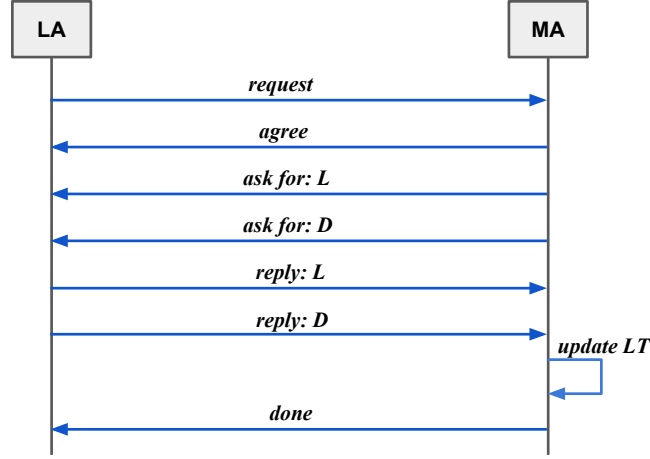


Figure 3.: Agents registration sequence diagram

- Input text might be classified to different domains with a percentage of relevance.
- The domain and the language of the text are provided.
- DAs and LAs register themselves with the MA by sending the associated ontology domain and language, then the MA creates the look-up table and gives each *LA* a unique ID.
- Each *LA* is responsible for learning concepts and the relations between those concepts (taxonomic and non-taxonomic) for a specific domain and language.
- The input text might contain texts written in more than one language.

Algorithm 1 Ontology Enrichment

```

1: procedure ONTOLOGYENRICHMENT( $S, D, L, LT$ )
2:    $DA = FindDA(LT, D)$ 
3:    $MA$  forward input data to  $DA$ 
4:    $LA = FindLA(LT, L)$ 
5:    $LA.Enrich(S)$ 
6: end procedure
7: procedure ENRICH( $S$ )
8:   for each sentence  $s_i$  in  $S$  do
9:     for each concept  $c_i$  in  $s_i$  do
10:       $found = SearchFor(c_i, LA.TT)$ 
11:      if ( $found = True$ )
12:         $continue$ 
13:      else
14:         $DV = AskDA(c_i)$ 
15:        if ( $!DV.isEmpty$ )
16:           $OntologyEnrich(LA.ontology, DV)$ 
17:           $LA.updateTT(LA.TT)$ 
18:        end if
19:      end if
20:    end for
21:  end for
22: end procedure
  
```

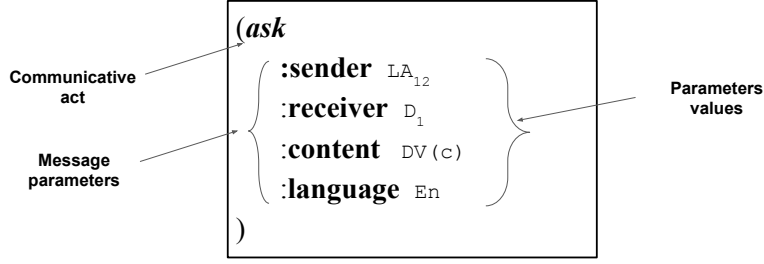


Figure 4.: Components of the ACL message

Each LA maintains a comprehensive *translation table* (TT). Each translation table lists the concepts that the agent knows and maps them to the corresponding concepts in the associated ontology. English is the common language for all translation tables of all agents. For example, the translation table of the agent of German ontologies translates concepts from German to English and vice versa, while the translation table of the agent of Arabic ontologies translates concepts from Arabic to English and vice versa. Only translations that are credible will be recorded in the translation table. Google Cloud Translation API ³ for Java is used by domain agents in order to provide the translations for the concepts and relations of the given ontologies to be enriched. Google Translate has been widely used in cross-lingual ontology mapping as a machine translation service. Currently, Google Translate is considered to be the best free online translator. It can also be used offline (Bella et al. 2017). One of the most prominent features in the proposed approach is that agents can exchange queries and messages to learn about incoming concepts and relations, using ontologies in the various languages of the other agents.

4.4.2. Agents Communication

All agent communication is carried out based on FIPA ACL messages (FIPA 2002). ACL is based on speech act theory, messages are the communicative acts sent by the sender agent to the receiver agents in order to perform some actions. Language agents can learn in two ways: 1) users can teach them by supplying a pre-built ontology and its translation table, and 2) an agent can learn from a new experience through its communications with its neighbours by *querying another agent for a certain concept*.

Querying another agent for a certain concept: The language agent sends out a query to the domain agent when it needs to obtain some information about an incoming concept. For instance, in Figure 4, agent LA_{12} asks D_1 about the English concept c_i . If the queried agent responds positively with its own semantics and description vector(s) ($DV(s)$) of matched concept(s), the querying agent takes two actions: 1) it enriches the associated ontology by the DV of that concept and 2) it updates its translation table. When the queried agent processes the query, it first matches the concept to its translation table. If a credible translation is found, then the queried agent simply sends back a description vector about the associated translated concept ($trans_{En}(DV(C))$). Figure 5 shows the complete set of communication actions taken by two language agents, LA_{11} and LA_{12} , and the domain agent D_1 , when LA_{11} wants to obtain information about an incoming concept C . As there might be more than one DV for each concept, steps marked with "*" refer to the repetition of this step for each of the returned DV . There are two possible dialogues of communication between agents. First, the language agent asks the domain agent about a certain concept and the domain agent replies with the DV of this concept, Dialogue (a) in Figure 6. Second, the domain agent asks other the language

³<https://developers.google.com/api-client-library/java/apis/translate/v2>

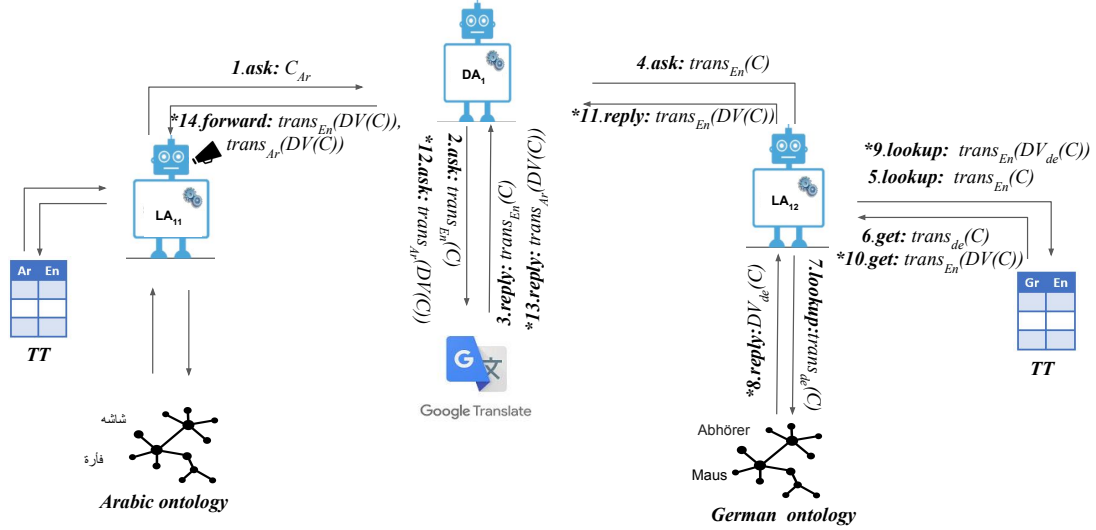


Figure 5.: Agent communication actions when LA_{11} wants to learn about a new incoming Arabic concept C . The asterisk "*" means that this step might be repeated several times if there is more than one DV for a concept.

agents (except the originator of the query) for the DV of the concept. If a positive reply is received, then the domain agent asks the language agent to *stream out* the available n DVs of this concept.

4.5. Verification and Validation

Ontology evaluation is defined in the context of two interesting concepts: verification and validation (Vrandečić 2009). Verification is the process of making sure that the candidate relations and concepts are extracted correctly, while validation is the process of making sure that they are the correct relations and concepts for the target ontology. Ontology engineers/experts are able to edit or validate the learned concepts and relations. (?) proposed a metric suite, which contains ten metrics for ontology evaluation: accuracy, adaptability, clarity, cohesion, completeness, computational efficiency, conciseness, consistency, coupling and coverage. Actually, ontology verification and validation play a key role in the ontology learning process, whose performance greatly affects the system's overall performance for building domain ontologies.

Dialogue (a):	
<pre>(ask :sender LA₁₁ :receiver D₁ :language En :reply-with q1: :content (DV (c1)))</pre>	<pre>(reply :sender D₁ :receiver LA₁₁ :language En :in-reply-to q1: :content (= (DV c1) (DV1)))</pre>
Dialogue (b):	
<pre>(stream-about :sender D₁ :receiver LA₁₂ :language En :reply-with q1: :content (DV (c1))) (tell :sender LA₁₂ :receiver D₁ :language En :in-reply-to q1: :content (= (DV c1) (DV1))) ...</pre>	<pre>(tell :sender LA₁₂ :receiver D₁ :language En :in-reply-to q1: :content (= (DV c1) (DVn))) (eos :sender LA₁₂ :receiver D₁ :language En :in-reply-to q1)</pre>

Figure 6.: Agent communication actions scenario using ACL messages in two different dialogues.

Algorithm 2 Agents Communications

```

1: procedure ASKDA( $c_i$ )
2:    $C_{en} = \text{GoogleAPI.trans}_{en}(c_i)$     //translate concept  $c_i$  to english
3:   for each  $LA$  connected to  $DA$  do
4:      $DV.add(AskLA(C_{en}))$ 
5:     if ( $\neg DV.isEmpty$ )
6:       break
7:     end if
8:   end for
9:   return  $DV$ 
10: end procedure
11: procedure ASKLA( $C_{en}$ )
12:    $found = SearchTT(C_{en}, LA.TT)$ 
13:   if ( $found = True$ )
14:      $C_{LA.lang} = Trans_{LA.lang}(C_{en}, LA.TT)$ 
15:      $DV_{lang} = SearchInOntology(C_{LA.lang})$ 
16:      $DV_{en} = TranslateFromTT(C_{LA.lang}, LA.TT)$ 
17:     return  $DV_{en}$ 
18:   else
19:     return null
20:   end if
21: end procedure
22: procedure FINDDA( $LT, D$ )
23:   for integer  $i = 0$  to  $LT.length$  do
24:     if ( $LT[i][1] = D$ )
25:       return  $i$ 
26:     end for
27:   return null
28: end procedure

```

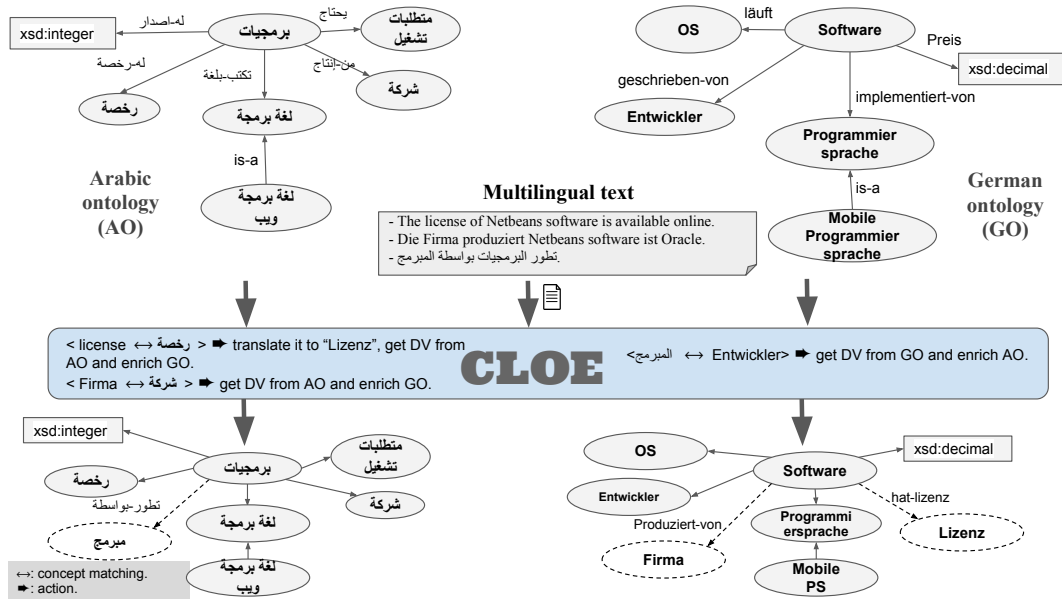


Figure 7.: Small fragment from the enriched ontologies before and after submitting a multilingual text chunk containing English, German and Arabic text

5. A Case Study

In this case study, we will use an example scenario of a multilingual text, containing English, German and Arabic, from the information technology domain. This case study aims to show the whole process, starting from submitting the text to the system up to obtaining the candidate concepts and relations. Figure 7 shows a small fragment from the enriched ontologies before and after submitting a multilingual text chunk containing English, German and Arabic text. In this scenario, the user submits the following text: "Netbeans fits the pieces together. The license for Netbeans software is available online. Die Firma produziert Netbeans software ist Oracle. *تطور البرمجيات بواسطة المبرمج*". Then, the system processes the text and produces a set of candidate concepts and relations, both taxonomic and non-taxonomic. Finally, experts edit and validate them before adding them to the ontology. Table 2 demonstrates the output of each phase starting from the preprocessing phase up to producing the set of candidate concepts and relations. In this case study, LA_{de} can not find any description for the concept (Firma) in its German ontology. Therefore, LA_{de} ask DA_{IT} for the DV_{de} of the concept (Firma). DA_{IT} in turn translates (Firma), to the common language between agents which is the English language, and get the translated concept (Company). As shown in Figure 7, the German concept "Firma" has a match in the Arabic ontology with (شركة). Thus, DA_{IT} ask LA_{ar} for the DV of the concept (Company). LA_{ar} translates (Company) to (شركة) by using its TT, and searches for the translated concept (شركة) in its Arabic ontology to get its DV as RDF triple (see table 2. Ontology enrichment (5)). Then, LA_{ar} translates such DV to the common language (English) using its TT and returns it to DA_{IT} . DA_{IT} in turn translates the received DV to German DV_{de} (Firma) and sends it to LA_{de} . Finally, the LA_{de} uses DV_{de} (Firma) to update its TT and enrich its German ontology. Similarly, the concept (license), which found in the Arabic ontology as (رخصة), is used to enrich the German ontology after translating it to German (Lizenz).

Table 2.: Case study: the output of each phase starting from preprocessing up to producing the set of candidate concepts and relations.

Phase		Output
Pre-processing	Lang. identification	En, De, Ar
	Language separation	En: Netbeans fits the pieces together. The licence for Netbeans software is available online, DE: Oracle-Firma produziert Netbeans-Software, AR: تطور البرمجيات بواسطة المبرمج
	Sentence splitter and Tokenization	1) Netbeans fits the pieces together. 2) The licence for Netbeans software is available online. 3) Oracle-Firma produziert Netbeans-Software. 4) تطور البرمجيات بواسطة المبرمج
	Stop words filter	1) Netbeans fits 2) licence Netbeans software is available online. 3) Oracle Firma produziert Netbeans Software. 4) تطور البرمجيات بواسطة المبرمج
	POS tagging	1) Netbeans(N) fits(V) 2) license(N) Netbeans(N) software(N) is(V) available (ADJ) online(N). 3) Oracle(N) Firma(N) produziert(V) Netbeans(N) Software(N). 4) تطور (V) البرمجيات (N) بواسطة (PPREP) المبرمج (N)
Domain identification		Information Technology (IT)
Candidate sentence selection		sentence 1 has been removed because it has just one POS which is a noun.
Ontology enrichment		(1) LA_{de} ask DA_{IT} for $DV_{de}(Firma)$ (2) DA_{IT} translates <i>Firma</i> to English: <i>Company</i> (3) DA_{IT} asks LA_{ar} for <i>Company</i> (4) LA_{ar} looks up for <i>Company</i> in its TT and obtains the translation, in case of found, to Arabic: شركة (5) LA_{ar} looks up(شركة) in its Arabic ontology and get its DV, in RDF triples as: $\langle \text{http://.../برمجيات} \rangle$ $\langle \text{http://.../شركة} \rangle$ (6) LA_{ar} obtains the English translation of these DVs using its TT and sends it back to DA_{IT} as: $\langle \text{http://.../software} \rangle$ $\langle \text{http://.../producedBy} \rangle$ $\langle \text{http://.../company} \rangle$ (7) DA_{IT} translates the DV of <i>Company</i> , using an online translator, to German: $\langle \text{http://.../software} \rangle$ $\langle \text{http://.../produziertVon} \rangle$ $\langle \text{http://.../Firma} \rangle$ (8) DA_{IT} sends $DV_{de}(Firma)$ and $DV_{En}(Company)$ to LA_{de} (9) LA_{de} updates its TT and then enriches its German ontology with the incoming DVs.

6. Implementation and Preliminary Evaluation

Generally, the evaluation of ontology learning systems involves the assessment of the resulting ontologies, which leads to guiding and refining the learning process. According to (Dellschaft and Staab 2006), there are three ways to evaluate a learned ontology: 1) the resulting ontology can be evaluated in an executable application, 2) it can be evaluated by domain experts or 3) it may even be compared with a predefined reference ontology (gold standard).

To perform the evaluation, we have implemented all phases of the proposed architecture in Java. In our experiments, we consider only text in only three languages, English, German and Arabic. All experiments are carried out on an Ubuntu 16.04 LTS operating system using Java (JDK 1.8) with an Intel Core i7-4600U CPU with 2.10GHz $\times$ 4 CPU and 10 GB of memory. This paper uses three domain ontologies in English, Arabic, and German covering the information technology domain (IT). Within this framework, the domain expert will make the final decision to select relevant concepts and relations.

6.1. Evaluation strategies

We followed two strategies for our evaluation: a satisfaction questionnaire evaluation and a gold standard-based evaluation.

Satisfaction Questionnaire Evaluation: We first briefly introduce the evaluation set-up and then discuss the result. A total of 24 ontology engineers were recruited for this questionnaire. At the beginning of the evaluation, it was insured that all the participants have completely understood the approach by giving them a presentation about the approach and a case study describing it. Then, we asked them to fill in a satisfaction questionnaire⁴, with 12 questions, to assess the distinct phases of the proposed approach. Figure 8 shows the results of the satisfaction questionnaire analysis.

By analysing the responses to the questionnaire, we made the following observations:

- Among all participants, 70.8% strongly agree that our approach would help ontology engineers by reducing time and effort.
- About 87.5% of the participants believed that it is *very hard* to manually develop ontologies from a multilingual text, while only 4.2% responded with *neutral*.
- Twenty out of the 24 participants pointed out that they believe that this approach can be considered as a step towards a domain-independent approach for ontology enrichment.
- A total of 79% of the participants were satisfied with the *completeness* of the candidate relations and concepts, including 33% who were *strongly satisfied*.
- Finally, for the overall satisfaction is : 60.9% of the participants were *strongly satisfied*, 21.7% of them were *satisfied* and 17.4% were *neutral*.

Gold standard-based evaluation: The objective of the gold standard-based evaluation was to determine whether the concepts/relations used in the learning process were correct, after comparing them with the gold standard results. We used a corpus of texts from the 20 Newsgroups dataset⁵. This is a common benchmark used for many learning tasks (Chang et al. 2008). We have developed three small ontologies, from the IT domain, for this experiment in English, German, and Arabic. These ontologies describe the same concepts, but in three different languages. For instance, there are some classes, such as *Software*, *Programming Language*, and *License* and relations, such as *hasVersion*, *producedBy*, and *writtenIn*. The 20 Newsgroups dataset contains approximately 20,000 newsgroup posts, partitioned across 20

⁴<https://goo.gl/DnnPnv>

⁵<http://qwone.com/~jason/20Newsgroups/>

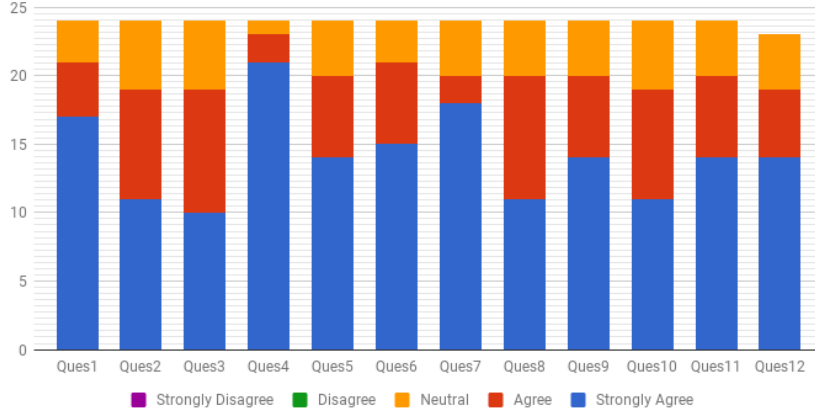


Figure 8.: Results of the satisfaction questionnaire analysis

different newsgroups which are organised into more extensive categories involving computers, religion, science and politics. We considered only the computers category (2,936 documents) which is the most related category to the IT domain. In this experiment, we divided the dataset into two subsets. The first one was used manually, by ontology experts, creating a reference standard ontology (Gold standard-based evaluation). Then, we have fed this subset into the system and saved the output. The second subset was used for evaluating CLOE. We randomly selected 100 documents from the second subset. Fifty documents out of the 100 were translated using Google API into Arabic and German, in order to submit multilingual text to the system. Inspired by the information retrieval community, the quality of the learned ontology can also be measured through precision, recall and F-measure metrics. The lexical precision (LP) and lexical recall (LR) reflect how well the extracted terms covering the target domain. Lexical precision (LP) is the fraction of the retrieved concepts and relations that are relevant, while lexical recall (LR) is the fraction of relevant items retrieved by the system. The F-measure is the harmonic mean of precision and recall. Formally, LR and LP are defined in Equation 7 given the number of relevant terms extracted (E_r), the total number of relevant terms in the gold standard ontology (N_r), and the total number of terms extracted (N_t).

$$\text{Lexical Precision} = \frac{E_r}{N_t} \quad (5)$$

$$\text{Lexical Recall} = \frac{E_r}{N_r} \quad (6)$$

$$F - \text{measure} = \frac{2 \times LP \times LR}{LP + LR} \quad (7)$$

We have identified four state-of-the-art approaches from the literature as the comparison approaches. The average of the precision, recall, and F-measure are used to evaluate CLOE in comparison to the identified approaches. Table 3 shows the comparison of the results with those of the four identified approaches. According to Table 3, CLOE outperforms all other systems with respect to recall and F-measure. On average, CLOE achieves a high recall of 99% and a high F-measure of 94% among all the other four systems. For instance, consider the

Table 3.: Comparison with the state-of-the-art systems.

Metrics	CLOE	(Benabdallah et al. 2017)	(Albukhitan and Helmy 2014)	(Albukhitan and Helmy 2016)	(Beseiso et al. 2013)
LP	0.90	0.90	0.95	0.76	0.89
LR	0.99	0.71	0.77	0.45	0.73
F	0.94	0.79	0.85	0.56	0.80

query, in German, “Die *DB-Sicherung* ist in der *Testumgebung* erfolgreich. *DB-Backup* wurde bis Ende dieser Woche auf den neuen *Server verschoben*”, $E_r = 6$, $N_r = 6$ and $N_t = 6$ which means that CLOE retrieves all the six relevant concepts, appear in italic, identified by expert which gives F-measure of 1.0. CLOE achieved an F-measure between 86% and 100% for the extraction of concepts and relations. In 2014, (Albukhitan and Helmy 2014) achieved high precision by using explicit semantic analysis to validate the extracted concepts. CLOE outperformed this by 22% in terms of recall and 9% in terms of F-measure. In 2016, (Albukhitan and Helmy 2016) did not achieve satisfactory results by using statistical measures for concept recognition based on the frequency of the word in the corpus. Overall, the evaluation results of CLOE are promising and show good precision, recall, and F-measure values.

7. Conclusion

In this paper, we have presented a novel multi-agent-based approach (CLOE) for enriching ontologies from multilingual and multi-domain data sources. Each agent is responsible for enriching an ontology in a particular language. The most prominent feature of our approach is that agents can learn from each other to benefit from the already-learned concepts found in the input ontologies. The main contribution of this work is the usage of a text block to enrich several ontologies from languages other than the language of the input text. Furthermore, the approach is domain-independent. Evaluations on existing state of the art tools show that our approach achieved an average precision of 89.5% with an average recall of 99.7%. By analysing the evaluation results, we draw the following conclusions:

- CLOE is considered as a very helpful tool for ontology engineers.
- The completeness and the correctness of the candidate relations and concepts are promising and encouraging.
- It is very hard to develop ontologies from multilingual text, which agrees with our research motivation.
- CLOE is a domain-independent approach for ontology enrichment.
- CLOE achieves the best recall and F-measure among all methods involved in the evaluation. These results are satisfying compared to four state-of-the-art tools.

In conclusion, this study has gone some way towards cross-lingual and multi-domain ontology enrichment. Our research challenges are solved and we believe that future research in this area can improve the learning process. An interesting line for future research is to use machine learning techniques in order to enhance the iterative ontology learning process and

to semantic rules from existing ontologies. In addition, techniques to generate multilingual ontologies from monolingual ontologies, should also be considered. Finally, we are planning to evaluate the performance of the system using heavy-weight ontologies.

Acknowledgement

Said Fathalla and Shima Ibrahim would like to thank and acknowledge the Ministry of Higher Education (MoHE) of Egypt for providing scholarships to conduct this study as well as the University of Bonn, Germany for offering all facilities required to do this work.

References

- Al-Khalifa, H. and A. Al-Wabil (2007). The arabic language and the semantic web: Challenges and opportunities. In *The 1st int. symposium on computer and Arabic language*.
- Al-Zoghby, A. M., A. S. E. Ahmed, and T. T. Hamza (2013). Arabic semantic web applications: A survey. *Journal of Emerging Technologies in Web Intelligence* 5(1), 52–69.
- Albukhitan, S. and T. Helmy (2014). Multi-agent based system for multilingual ontologies maintenance. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, Volume 1, pp. 419–423. IEEE.
- Albukhitan, S. and T. Helmy (2016). Arabic ontology learning from un-structured text. In *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, pp. 492–496. IEEE.
- Balaji, P. and D. Srinivasan (2010). An introduction to multi-agent systems. In *Innovations in multi-agent systems and applications-1*, pp. 1–27. Springer.
- Bella, G., F. Giunchiglia, and F. McNeill (2017). Language and domain aware lightweight ontology matching. *Web Semantics: Science, Services and Agents on the World Wide Web* 43, 1–17.
- Benabdallah, A., M. A. Abderrahim, and M. E.-A. Abderrahim (2017). Extraction of terms and semantic relationships from arabic texts for automatic construction of an ontology. *International Journal of Speech Technology*, 1–8.
- Beseiso, M., A. R. Ahmad, R. Ismail, and M. Taher (2013). Multilingual ontology learning algorithm for emails. In *Soft Computing Applications and Intelligent Systems*, pp. 229–244. Springer.
- Boudabous, M. M., N. C. Kammoun, N. Khedher, L. H. Belguith, and F. Sadat (2013). Arabic wordnet semantic relations enrichment through morpho-lexical patterns. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pp. 1–6. IEEE.
- Browarnik, A. and O. Maimon (2015). Ontology learning from text: Why the ontology learning layer cake is not viable. *International Journal of Signs and Semiotic Systems (IJSSS)* 4(2), 1–14.
- Buitelaar, P., P. Cimiano, P. Haase, and M. Sintek (2009). Towards linguistically grounded ontologies. *The semantic web: research and applications*, 111–125.
- Chang, M.-W., L.-A. Ratinov, D. Roth, and V. Srikumar (2008). Importance of semantic representation: Dataless classification. In *AAAI*, Volume 2, pp. 830–835.
- Cimiano, P., A. Hotho, and S. Staab (2004). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pp. 435–439. IOS Press.
- Cimiano, P. and J. Völker (2005). A framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), Lecture Notes in Computer Science, Springer*, Volume 3513, pp. 227–238. Springer.
- De Melo, G. and G. Weikum (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 513–522. ACM.
- Dellschaft, K. and S. Staab (2006). On how to perform a gold standard based evaluation of ontology learning. In *International Semantic Web Conference*, Volume 4273, pp. 228–241. Springer.

- Euzenat, J., P. Shvaiko, et al. (2007). *Ontology matching*, Volume 18. Springer.
- Fathalla, S. (2018). Detecting human diseases relatedness: A spreading activation approach over ontologies. *International Journal on Semantic Web and Information Systems (IJSWIS)* 14(3), 120–133.
- Fathalla, S. and Y. Kannot (2017). A bidirectional-based spreading activation method for human diseases relatedness detection using disease ontology. In *Conference on Computational Collective Intelligence Technologies and Applications*, pp. 14–23. Springer.
- Fathalla, S. and C. Lange (2018). Eventskg: a knowledge graph representation for top-prestigious computer science events metadata. In *International Conference on Computational Collective Intelligence*, pp. 53–63. Springer.
- Fathalla, S., S. Vahdati, S. Auer, and C. Lange (2017). Towards a knowledge graph representing research findings by semantifying survey articles. In *International Conference on Theory and Practice of Digital Libraries*, pp. 315–327. Springer.
- Fathalla, S., S. Vahdati, S. Auer, and C. Lange (2018). Semsur: a core ontology for the semantic representation of research findings. *Procedia Computer Science* 137, 151–162.
- Fathalla, S. M., Y. F. Hassan, and M. El-Sayed (2012a). A hybrid method for user query reformation and classification. In *Computer Theory and Applications (ICCTA), 2012 22nd International Conference on*, pp. 132–138. IEEE.
- Fathalla, S. M., Y. F. Hassan, and M. El-Sayed (2012b). A hybrid method for user query reformation and classification. In *Computer Theory and Applications (ICCTA), 2012 22nd International Conference on*, pp. 132–138. IEEE.
- Faure, D. and T. Poibeau (2000). First experiments of using semantic knowledge learned by asium for information extraction task using intex. In *Proceedings of the ECAI workshop on Ontology Learning*.
- Fellbaum, C. (1998). *WordNet*. Wiley Online Library.
- FIPA, A. (2002). Fipa acl message structure specification. *Foundation for Intelligent Physical Agents*, <http://www.fipa.org/specs/fipa00061/SC00061G.html> (30.6. 2004).
- Fortuna, B., D. Mladenich, and M. Grobelnik (2006). Semi-automatic construction of topic ontologies. pp. 121–131. Springer.
- Fu, B., R. Brennan, and D. O’sullivan (2010). Cross-lingual ontology mapping and its use on the multilingual semantic web. *MSW* 571, 13–20.
- Fu, B., R. Brennan, and D. O’Sullivan (2009). Cross-lingual ontology mapping—an investigation of the impact of machine translation. *The Semantic Web*, 1–15.
- Gillani Andleeb, S. (2015). *From text mining to knowledge mining: An integrated framework of concept extraction and categorization for domain ontology*. Ph. D. thesis, Budapesti Corvinus Egyetem.
- Gracia, J., E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae (2012). Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web* 11, 63–71.
- Hahn, U. and M. Romacker (2001). The syndikate text knowledge base generator. In *Proceedings of the first international conference on Human language technology research*, pp. 1–6. Association for Computational Linguistics.
- Helou, M. A. and M. Palmonari (2015). Cross-lingual lexical matching with word translation and local similarity optimization. In *Proceedings of the 11th International Conference on Semantic Systems*, pp. 97–104. ACM.
- IOGDS (2017). International open government data search. https://logd.tw.rpi.edu/iogds_analytics_2. [Online; accessed: 2017-12-3].
- Ishkewy, H., H. Harb, and H. Farahat (2014). Azhary: An arabic lexical ontology. *arXiv preprint arXiv:1411.1999*.
- Jiang, X. and A.-H. Tan (2010). Crctol: A semantic-based domain ontology learning system. *Journal of the Association for Information Science and Technology* 61(1), 150–168.
- Labrou, Y., T. Finin, and Y. Peng (1999). Agent communication languages: The current landscape. *IEEE Intelligent Systems and Their Applications* 14(2), 45–52.
- Lui, M., J. H. Lau, and T. Baldwin (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics* 2, 27–40.
- Montiel-Ponsoda, E., G. A. de Cea, A. Gómez-Pérez, and W. Peters (2011). Enriching ontologies with

- multilingual information. *Natural language engineering* 17(3), 283–309.
- Montoyo, A., R. Muñoz, and E. Métais (2005). *Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15-17, Proceedings*, Volume 3513. Springer.
- Park, J., W. Cho, and S. Rho (2010). Evaluating ontology extraction tools using a comprehensive evaluation framework. *Data & Knowledge Engineering* 69(10), 1043–1061.
- Peters, C., M. Braschler, and P. Clough (2012). *Multilingual information retrieval: From research to practice*. Springer Science & Business Media.
- Rafi, M., H. Qureshi, and H. Khatoon (2009). Ontology maintenance via multi-agents. In *INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on*, pp. 955–959. IEEE.
- Ravichandran, D. and E. Hovy (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 41–47. Association for Computational Linguistics.
- Sabou, M. (2004). Extracting ontologies from software documentation: a semi-automatic method and its evaluation.
- Shamsfard, M. and A. Barforoush (2002). An introduction to hasti: an ontology learning system. In *Proceedings of the iasted international conference artificial intelligence and soft computing, Acta Press, Calgary, Canada*, pp. 242–247.
- Sharma, A. G. and M. V. Rastogi (2014). Filter for spamming email by using ontology. *Innovative Science Engineering and Technology 1*.
- Soudani, N., I. Bounhas, B. ElAyeby, and Y. Slimani (2014). Toward an arabic ontology for arabic word sense disambiguation based on normalized dictionaries. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, pp. 655–658. Springer.
- Souter, C. et al. (2017). Natural language identification using corpus-based models. *HERMES-Journal of Language and Communication in Business* 7(13), 183–203.
- Spohr, D., L. Hollink, and P. Cimiano (2011). A machine learning approach to multilingual and cross-lingual ontology matching. *The Semantic Web–ISWC 2011*, 665–680.
- Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 173–180. Association for Computational Linguistics.
- Trojahn, C., B. Fu, O. Zamazal, and D. Ritze (2014). State-of-the-art in multilingual and cross-lingual ontology matching. In *Towards the Multilingual Semantic Web*, pp. 119–135. Springer.
- Vrandečić, D. (2009). Ontology evaluation. In *Handbook on Ontologies*, pp. 293–313. Springer.
- Wang, Y., J. Völker, and P. Haase (2006). Towards semi-automatic ontology building supported by large-scale knowledge acquisition. In *AAAI Fall Symposium On Semantic Web for Collaborative Knowledge Acquisition*, Volume 6, pp. 06.
- Yamaguchi, T. (2001). Acquiring conceptual relationships from domain-specific texts. In *Workshop on Ontology Learning*, Volume 38, pp. 69–113.
- Yang, H. and J. Callan (2008). Ontology generation for large email collections. In *Proceedings of the 2008 international conference on Digital government research*, pp. 254–261. Digital Government Society of North America.
- Zaidi, S., A. Abdelali, F. Sadat, and M.-T. Laskri (2012). Hybrid approach for extracting collocations from arabic quran texts. In *Workshop Organizers*, pp. 102.
- Zouaq, A., D. Gasevic, and M. Hatala (2011). Towards open ontology learning and filtering. *Information Systems* 36(7), 1064–1081.

OECM: A Cross-Lingual Approach for Ontology Enrichment

Reproduced with permission from Springer Nature. Shima Ibrahim, Said Fathalla, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. *OECM: A cross-lingual approach for ontology enrichment*. In European Semantic Web Conference (ESWC), pp. 100-104. Springer International Publishing, 2019. DOI: https://doi.org/10.1007/978-3-030-32327-1_20



OECEM: A Cross-Lingual Approach for Ontology Enrichment

Shimaa Ibrahim^{1,2()}, Said Fathalla^{1,3}, Hamed Shariat Yazdi¹,
Jens Lehmann^{1,4}, and Hajira Jabeen¹

¹ Smart Data Analytics (SDA), University of Bonn, Bonn, Germany
{ibrahim,fathalla,shariat,jens.lehmann,jabeen}@cs.uni-bonn.de

² Institute of Graduate Studies and Research, University of Alexandria,
Alexandria, Egypt

³ Faculty of Science, University of Alexandria, Alexandria, Egypt

⁴ Enterprise Information Systems Department, Fraunhofer IAIS,
Sankt Augustin, Germany

Abstract. Due to the rapid expansion of multilingual data on the web, development of approaches to enrich ontologies has become an interesting and active subject of research. In this paper, we propose a cross-lingual matching approach for ontology enrichment (OECEM) in order to enrich an ontology using another one in a different natural language. A prototype for the proposed approach has been implemented and evaluated using the MultiFarm benchmark. Evaluation results show higher precision and recall in comparison to other four state-of-the-art approaches.

Keywords: Cross-lingual matching · Ontology enrichment · Machine translation

1 Introduction

The increasing amount of multilingual data on the Semantic Web has motivated many researchers to develop ontologies in various natural languages. In fact, ontologies can be enriched by adding additional classes and/or relations extracted from other resources, even in another natural language [7]. Such enrichment is a resource demanding and time-consuming task. Therefore, automated or semi-automated ontology enrichment approaches are highly desired. Most research efforts pay attention to enrich English ontologies from English resources rather than non-English ones, by applying ontology matching techniques [7]. This raises a key question; How can an ontology be enriched using another ontology in a different natural language? In order to enrich ontologies from multilingual resources, most of the recent efforts in developing different techniques for cross-lingual ontology matching focus on one-to-one translation between ontology concepts [8]. Consequently, inappropriate translations negatively affect the quality of the matching process [8]. Therefore it is important to develop innovative approaches which are capable of enriching ontologies by selecting the best

translation among all available translations (i.e. one-to-many translations) for a particular term. To the best of our knowledge, only our previous work [1] has addressed the problem of enriching ontologies from multilingual text. In this paper, we propose a new approach (OECM) to enrich an ontology, i.e. the target ontology T , using another one, i.e. the source ontology S , in a different natural language. The prominent feature of the proposed approach is the selection of the best translation between all available translations when matching classes among ontologies. This selection significantly improves the quality of the matching process. Furthermore, the usage of ontologies as the source for the enrichment process can significantly reduce the cost of data pre-processing and parsing of the data used being used for the enrichment. To evaluate OECM, we compare the cross-lingual ontology matching process with four state-of-the-art approaches. The implementation of OECM and the datasets used for evaluation are publicly available at <https://github.com/shmkhaled/OECM>.

2 The Proposed Approach

Goal: Given two ontologies S and T , in two different natural languages L_1 and L_2 respectively, as RDF triples $\langle s, p, o \rangle \in \mathcal{C} \times \mathcal{R} \times (\mathcal{C} \cup \mathcal{L})$ where $\mathcal{C}$ is the set of ontology domain entities (i.e. classes), $\mathcal{R}$ is the set of relations, and $\mathcal{L}$ is the set of literals. We aim at finding the complementary information $\mathcal{T}_e = S - (S \cap T)$ from S in order to enrich T . The methodology of the proposed approach comprises three phases: (1) pre-matching, (2) matching, and (3) enrichment. We have considered only class labels, or local names, and three standard relations: `rdfs:subClassOf`, `owl:equivalentClass`, `owl:disjointWith`.

(1) Pre-matching: T and S are prepared before starting the matching phase by performing two tasks: **(a) Pre-processing:** The aim of this task is to prepare the local names and/or labels of classes of S and T by employing several natural language processing techniques, such as tokenization, normalization, stop words removal and POS-tagging. The output of this task is two sets of pre-processed classes $\mathcal{C}'_S$ and $\mathcal{C}'_T$ for S and T respectively, **(b) Translation:** Each class in $\mathcal{C}'_S$ is translated using Google Translator to the language of T (i.e. L_2). A list of translations is associated with each class, for example, the class label “Thema” in German, has a list of two English translations: “Subject” and “Topic”. The best translation will be selected in the next phase.

(2) Matching: In order to identify which, and where the new information will be added to T , potential matches between S and T should be identified. We use two types of matching: *Terminological matching* and *Structural matching*. **(a) Terminological matching:** This task is used to identify which information can be added to T . In order to choose the best translation for each class that matches the corresponding one in T , we perform a pairwise string matching between them. We chose Jaccard similarity as a string similarity metric because it has achieved the best score in terms of precision in the experiments conducted

Table 1. State-of-the-art comparison results

Approaches	Conference _{de} × Ekaw _{en}			Conference _{de} × Edas _{en}		
	Precision	Recall	F-measure	Precision	Recall	F-measure
AML [3]	0.56	0.20	0.29	0.86	0.35	0.50
KEPLER [6]	0.33	0.16	0.22	0.43	0.18	0.25
LogMap [5]	0.71	0.20	0.31	0.71	0.29	0.42
XMap [9]	0.18	0.16	0.17	0.23	0.18	0.20
OECM	0.75	0.67	0.71	0.93	0.76	0.84

for the ontology alignment task in the MultiFarm benchmark¹ [2]. We consider similarity scores greater than or equal to a specific threshold θ to get the best matches. After running the experiments for ten times, we obtained the best value of θ which gives the best matching results. If no match is found, this class is considered as a new class, which is added to T . At the end, matched classes are validated by experts in order to confirm that the best translation is selected for each class. **(b) Structural matching:** It is used to identify where the new information can be added to T . Each class in S is replaced by its best translation found in the previous matching in order to get a translated ontology S_{trans} . We apply a pairwise triple comparison between S_{trans} and T to get the set of triples to be enriched $\mathcal{T}_e$, which is represented by $\langle s, p, o, F \rangle$. Each triple is associated with a flag F , with a value either ‘E’ for enrichment or ‘A’ for addition. For a particular triple, if $s \in \mathcal{C}'_T$ and $o \notin \mathcal{C}'_T$, then $F = \text{‘E’}$, i.e. this triple is needed to enrich the existing information in T , while if $s \notin \mathcal{C}'_T$ and $o \in \mathcal{C}'_T$, then $F = \text{‘A’}$, i.e. this triple is needed to add a new class to T .

(3) Enrichment: $\mathcal{T}_e$ is used to enrich T according to the flags associated with each triple. We enrich the Scientific Events Ontology (SEO_{en}) [4], which has 49 classes in English, using the Conference_{de} ontology from the MultiFarm dataset (see Sect. 3), which has 60 classes in German. OECM has identified new 15 triples to enrich SEO_{en}. For instance, $\langle \text{ConferenceContributor}, \text{subClassOf}, \text{Person}, \text{‘A’} \rangle$ is used to add a new class `ConferenceContributor`, as a `subClassOf` `Person`, to SEO_{en}. In addition, $\langle \text{KeynoteSpeaker}, \text{subClassOf}, \text{ConferenceContributor}, \text{‘E’} \rangle$ is used to enrich SEO_{en} with additional information, i.e. adding a new relation `subClassOf` between the two classes. The complete 15 triples can be found at the GitHub repository. We have successfully enriched SEO_{en} by 93.75% of the triples identified by an expert.

3 Evaluation

We use ontologies in the MultiFarm benchmark to measure the quality of the cross-lingual matching process. MultiFarm consists of seven ontologies, their

¹ <https://www.irit.fr/recherches/MELODI/multifarm/>.

translation into nine languages, and the corresponding cross-lingual alignments between them (i.e. the gold standard). We compare our results with four state-of-the-art approaches (see Table 1) for matching *Conference_{de}* with *Ekaw_{en}* and *Conference_{de}* with *Edas_{en}* ontologies. OECM outperforms all other systems in terms of precision, recall, and F-measure. For AML [3], authors include pre-computed dictionaries with translations, to overcome the query limit of Microsoft Translator which decrease the efficiency of their approach. LogMap [5] depends mainly on the initial mappings to discover new mappings, which decreased after performing the translation. XMap [9] did not achieve satisfactory results because of many internal exceptions. Surprisingly, we found seven new alignments, which did not exist in the gold standard, when matching *Conference_{de}* with *Ekaw_{en}*, for instance, (*<LeiterDerWorkshops>_{de}*, *<Workshop_Chair>_{en}*).

4 Conclusion

We present a new approach (OECM) for enrichment of ontologies using other ontologies in different natural languages. Terminological and structural matching have been used in order to identify which, and where, information from the source ontology, can be used to enrich the target ontology. We consider all available translations for each term and select the best translation that matches the corresponding term in the target ontology. Such selection has significantly improved the quality of the matching process. It is worth to mentioning that OECM has also found new alignments, which were missing in the gold standard. OECM outperforms most of the state of the art systems in terms of precision, recall, and F-measure. We are in the process of investigating the usage of semantic similarity between terms in the matching process, in addition to considering other non-standard semantic relations and individuals in the enrichment process.

References

1. Ali, M., Fathalla, S., Ibrahim, S., Kholief, M., Hassan, Y.: Cross-lingual ontology enrichment based on multi-agent architecture. *Proc. Comput. Sci.* **137**, 127–138 (2018)
2. Cheatham, M., Hitzler, P.: String similarity metrics for ontology alignment. In: Alani, H., et al. (eds.) *ISWC 2013. LNCS*, vol. 8219, pp. 294–309. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41338-4_19
3. Faria, D., et al.: Results of AML participation in OAEI 2018. In: *Proceedings of the 13th International Workshop on Ontology Matching*, pp. 125–131. CEUR-WS (2018)
4. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: The scientific events ontology of the openresearch.org curation platform. In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pp. 2311–2313. ACM (2019)
5. Jiménez-Ruiz, E., Grau, V.C.: LogMap family participation in the OAEI 2018. In: *Proceedings of the 13th International Workshop on Ontology Matching*, pp. 187–191. CEUR-WS (2018)

6. Kachroudi, M., Diallo, G., Yahia, S.B.: OAEI 2018 results of KEPLER. In: Proceedings of the 13th International Workshop on Ontology Matching, pp. 173–178. CEUR-WS (2018)
7. Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E.: Ontology population and enrichment: state of the art. In: Paliouras, G., Spyropoulos, C.D., Tsatsaronis, G. (eds.) Knowledge-Driven Multimedia Information Extraction and Ontology Evolution. LNCS (LNAI), vol. 6050, pp. 134–166. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20795-2_6
8. Trojahn, C., Fu, B., Zamazal, O., Ritze, D.: State-of-the-art in multilingual and cross-lingual ontology matching. In: Buitelaar, P., Cimiano, P. (eds.) Towards the Multilingual Semantic Web, pp. 119–135. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-43585-4_8
9. Djeddi, W.E., Yahia, S.B., Khadir, M.T.: XMap results for OAEI 2018. In: Proceedings of the 13th International Workshop on Ontology Matching, pp. 210–215. CEUR-WS (2018)

**From monolingual to multilingual ontologies:
The role of cross-lingual ontology enrichment**



From Monolingual to Multilingual Ontologies: The Role of Cross-Lingual Ontology Enrichment

Shimaa Ibrahim^{1,2} (✉), Said Fathalla^{1,3}, Hamed Shariat Yazdi¹,
Jens Lehmann^{1,4}, and Hajira Jabeen¹

¹ Smart Data Analytics (SDA), University of Bonn, Bonn, Germany
{ibrahim,fathalla,shariat,jens.lehmann,jabeen}@cs.uni-bonn.de

² Institute of Graduate Studies and Research, University of Alexandria,
Alexandria, Egypt

³ Faculty of Science, University of Alexandria, Alexandria, Egypt

⁴ Enterprise Information Systems Department, Fraunhofer IAIS, Sankt Augustin,
Germany

Abstract. While the multilingual data on the Semantic Web grows rapidly, the building of multilingual ontologies from monolingual ones is still cumbersome and hampered due to the lack of techniques for cross-lingual ontology enrichment. Cross-lingual ontology enrichment greatly facilitates the semantic interoperability between different ontologies in different natural languages. Achieving such enrichment by human labor is a time-consuming and error-prone task. Thus, in this paper, we propose a fully automated ontology enrichment approach using cross-lingual matching (OECM) approach, which builds a multilingual ontology by enriching a monolingual ontology from another one in a different natural language using a cross-lingual matching. OECM selects the best translation among all available translations of ontology concepts based on their semantic similarity with the target ontology concepts. We present a use case of our approach for enriching English Scholarly Communication Ontologies using German and Arabic ontologies from the MultiFarm benchmark. We have compared our results with the results from the Ontology Alignment Evaluation Initiative (OAEI 2018). Our approach has higher precision and recall in comparison to five state-of-the-art approaches. Additionally, we recommend some linguistic corrections in the Arabic ontologies in Multifarm which have enhanced our cross-lingual matching results.

Keywords: Cross-lingual ontology enrichment · Cross-lingual matching · Multilingual ontology · Ontology engineering · Knowledge management

1 Introduction

The wide proliferation of multilingual data on the Semantic Web results in many ontologies scattered across the web in various natural languages. According to

the Linked Open Vocabularies (LOV)¹, the majority of the ontologies in the Semantic Web are in English, however, ontologies in other Indo-European languages also exist. For instance, out of a total 681 vocabularies found in LOV, 500 are in English, 54 in French, 39 in Spanish, and 33 in German. Few ontologies exist in non-Indo-European languages, such as 13 in Japanese and seven in Arabic. Monolingual ontologies with labels or local names presented in a certain language are not easily understandable to speakers of other languages. Therefore, in order to enhance semantic interoperability between monolingual ontologies, approaches for building multilingual ontologies from the existing monolingual ones should be developed [26]. Multilingual ontologies can be built by applying *cross-lingual ontology enrichment* techniques, which expand the target ontology with additional concepts and semantic relations extracted from external resources in other natural languages [23]. For example, suppose we have two ontologies; Scientific Events Ontology in English (SEO_{en}) and Conference in German (Conference_{de}). Both SEO_{en} and Conference_{de} have complementary information, i.e. SEO_{en} has some information which does not exist in Conference_{de} and vice versa. Let us consider a scenario where a user wants to get information from both SEO_{en} and Conference_{de} to be used in an ontology-based application. This may not be possible without a cross-lingual ontology enrichment solution, which enrich the former by the complementary information in the latter. Manual ontology enrichment is a resource demanding and time-consuming task. Therefore, fully automated *cross-lingual* ontology enrichment approaches are highly desired [23]. Most of the existing work in ontology enrichment focus on enriching English ontologies from English sources only (monolingual enrichment) [23]. To the best of our knowledge, only our previous work [1, 14] has addressed the cross-lingual ontology enrichment problem by proposing a semi-automated approach to enrich ontologies from multilingual text or from other ontologies in different natural languages.

In this paper we address the following research question; *how can we automatically build multilingual ontologies from monolingual ones?* We propose a fully automated ontology enrichment approach in order to create multilingual ontologies from monolingual ones using cross-lingual matching. We extend our previous work [14] by: (1) using the semantic similarity to select the best translation of class labels, (2) enriching the target ontology by adding new classes in addition to all their related subclasses in the hierarchy, (3) using ontologies in non-Indo-European languages (e.g., Arabic), as the source of information, (4) building multilingual ontologies, and (5) developing a fully automated approach. OECM comprises six phases: (1) *translation*: translate class labels of the source ontology, (2) *pre-processing*: process class labels of the target and the translated source ontologies, (3) *terminological matching*: identify potential matches between class labels of the source and the target ontologies, (4) *triple retrieval*: retrieve the new information to be added to the target ontology, (5) *enrichment*: enrich the target ontology with new information extracted from the source ontology, and (6) *validation*: validate the enriched ontology. A noticeable feature of

¹ <https://lov.linkeddata.es/dataset/lov/vocabs>.

OECM is that we consider multiple translations for a class label. In addition, the use of semantic similarity has significantly improved the quality of the matching process. We present a use case for enriching the Scientific Events Ontology (SEO) [9], a scholarly communication ontology for describing scientific events, from German and Arabic ontologies. We compare OECM to five state-of-the-art approaches for cross-lingual ontology matching task. OECM outperformed these approaches in terms of precision, recall, and F-measure. Furthermore, we evaluate the enriched ontology by comparing it against a Gold standard created by ontology experts. The implementation of OECM and the datasets used in the use case are publicly available².

The remainder of this paper is structured as follows: we present an overview of related work in Sect. 2. Overview of the proposed approach is described in Sect. 3. In order to illustrate possible applications of OECM, a use case is presented in Sect. 4. Experiments and evaluation results are presented in Sect. 5. Finally, we conclude with an outline of the future work in Sect. 6.

2 Related Work

A recent review of the literature on multilingual Web of Data found that the potential of the Semantic Web for being multilingual can be accomplished by techniques to build multilingual ontologies from monolingual ones [12]. Multilingual enrichment approaches are used to build multilingual ontologies from different resources in different natural languages [5, 6, 24]. Espinoza et al. [6] has proposed an approach to generate multilingual ontologies by enriching the existing monolingual ontologies with multilingual information in order to translate these ontologies to a particular language and culture (ontology localization). In fact, ontology enrichment depends on matching the target ontology with external resources, in order to provide the target ontology with additional information extracted from the external resources.

All the literature have focused on the cross-lingual ontology matching techniques which are used for matching different natural languages of linguistic information in ontologies [12, 26]. Meilicke et al. [20] created a benchmark dataset (MultiFarm) that results from the manual translations of a set of ontologies from the conference domain into eight natural languages. This dataset is widely used to evaluate the cross-lingual matching approaches [7, 15, 16, 28]. Manual translation of ontologies can be infeasible when dealing with large and complex ontologies. Trojahn et al. [27] proposed a generic approach which relies on translating concepts of source ontologies using machine translation techniques into the language of the target ontology. In the translation step, they depend on getting one translation for each concept (one-to-one translation), then they apply monolingual matching approaches to match concepts between the source ontologies and the translated ones. Fu et al. [10, 11] proposed an approach to match English and Chinese ontologies by considering the semantics of the target ontology, the mapping intent, the operating domain, the time and resource

² <https://github.com/shmkhaled/OECM>.

constraints and user feedback. Hertling and Paulheim [13] proposed an approach which utilizes Wikipedia’s inter-language links for finding corresponding ontology elements. Lin and Krizhanovsky [18] proposed an approach which use Wiktionary³ as a source of background knowledge to match English and French ontologies. Tigrine et al. [25] presented an approach, which relies on the multilingual semantic network BabelNet⁴ as a source of background knowledge, to match several ontologies in different natural languages. In the context of OAEI 2018 campaign⁵ for evaluating ontology matching technologies, AML [7], KEPLER [16], LogMap [15] and XMap [28] provide high-quality alignments. These systems use terminological and structural alignments in addition to using external lexicon, such as WordNet⁶ and UMLS-lexicon⁷ in order to get the set of synonyms for the ontology elements. In order to deal with multilingualism, AML and KEPLER rely on getting (one-to-one translation) using machine translation technologies, such as Microsoft translator, before starting the matching process. LogMap and XMap do not provide any information about the utilized translation methodology. Moreover, LogMap is an iterative process, that starts from initial mappings (‘almost exact’ lexical correspondences) to discover new mappings. It is mentioned in [15] that the main weakness of LogMap is that it can not find matching between ontologies which do not provide enough lexical information as it depends mainly on the initial mappings. A good literature of the state-of-the-art approaches in cross-lingual ontology matching is provided in [26].

Most of the literature have focused on enriching monolingual ontologies with multilingual information in order to translate or localize these ontologies. In addition, in the cross-lingual ontology matching task, there is a lack of exact one-to-one translation between terms across different natural languages which negatively affects the matching results. We address this limitations in our proposed approach by building multilingual ontologies, where a class label is presented by several natural languages, from monolingual ones. Such approach support the ontology matching process with multiple translations for a class label in order to enhance the matching results.

3 The Proposed Approach

Goal: Given two ontologies S and T , in two different natural languages L_s and L_t respectively, as RDF triples $\langle s, p, o \rangle \in \mathcal{C} \times \mathcal{R} \times (\mathcal{C} \cup \mathcal{L})$ where $\mathcal{C}$ is the set of ontology domain entities (i.e. classes), $\mathcal{R}$ is the set of relations, and $\mathcal{L}$ is the set of literals. We aim at finding the complementary information $\mathcal{T}_e = S - (S \cap T)$ from S in order to enrich T .

The proposed approach comprises six phases (Fig. 1): translation, pre-processing, terminological matching, triple retrieval, enrichment, and validation.

³ <https://www.wiktionary.org/>.

⁴ <https://babelnet.org/>.

⁵ <http://oaei.ontologymatching.org/2018/results/multifarm/index.html>.

⁶ <https://wordnet.princeton.edu/>.

⁷ <https://www.nlm.nih.gov/research/umls/>.

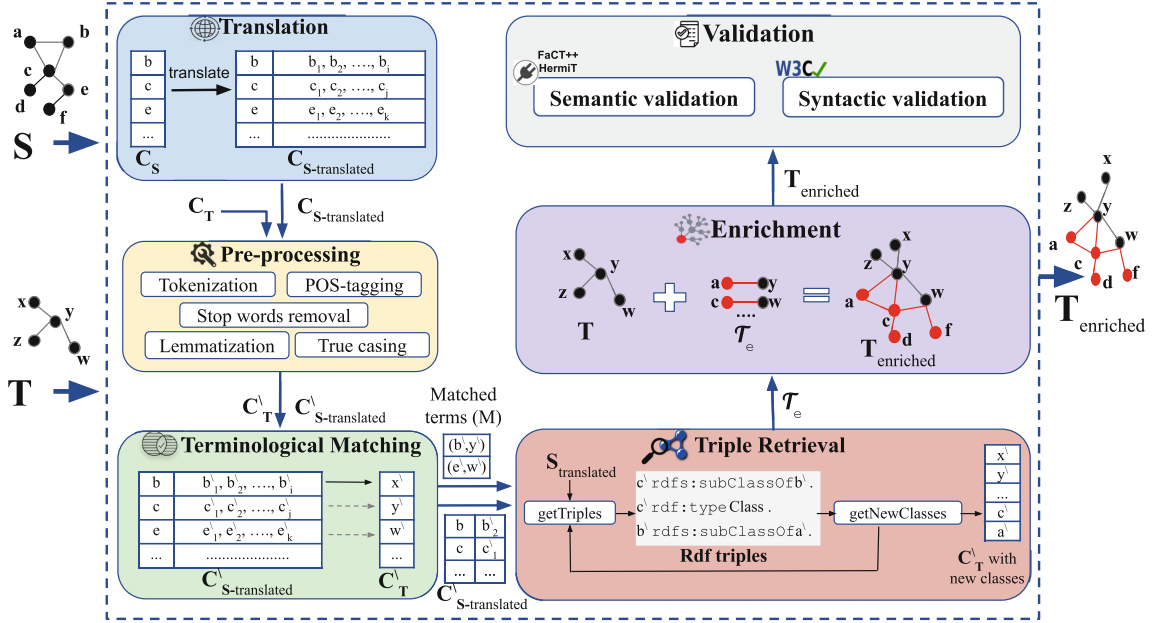


Fig. 1. The workflow of OECM.

The input is the two ontologies in two different natural languages, i.e. the target ontology T and the source ontology S . The output is the multilingual enriched ontology $T_{enriched}$ in two different natural languages L_1 and L_2 . In the following subsections, we describe each of these phases in details.

3.1 Translation

Let $\mathcal{C}_S$ and $\mathcal{C}_T$ be the set of classes in S and T respectively. Each class is represented by a label or a local name. The aim of this phase is to translate each class in $\mathcal{C}_S$ to the language of T (i.e. L_t). Google Translator⁸ is used to translate classes of source ontologies. All available translations are considered for each class. Therefore, the output of the translation is $\mathcal{C}_{S-translated}$ which has each class, in S , associated with a list of all available translations. For example, the class **Thema** in German has a list of English translations (Subject and Topic), and the class label “مراجعة” in Arabic has a list of English translations such as “Review, Revision, Check”. The best translation will be selected in the terminological matching phase (Subsect. 3.3).

3.2 Pre-processing

The aim of this phase is to process classes of $\mathcal{C}_T$ and lists of translations in $\mathcal{C}_{S-translated}$ by employing a variety of natural language processing (NLP) techniques, such as tokenization, POS-tagging (part-of-speech tagging), and lemmatization, to make it ready for the next phases. In order to enhance the similarity

⁸ <https://translate.google.com/>.

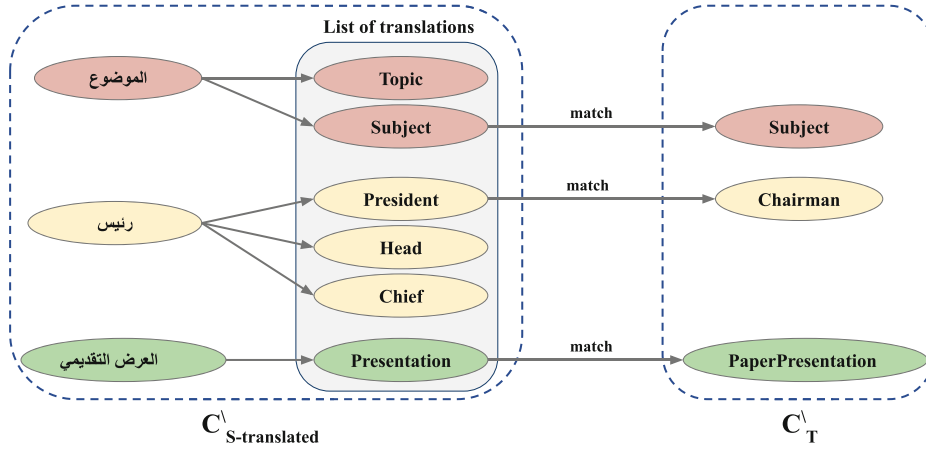


Fig. 2. Illustration of a terminological matching between list of translations, in English, for every concept in $C'_{S-translated}$, in Arabic, and C'_T in English

results between C_T and $C_{S-translated}$, stop words are removed and normalization methods and regular expressions are used to remove punctuation, symbols, additional white spaces, and to normalize the structure of strings. Furthermore, our pre-processing is capable of recognizing classes such as camel cases “ReviewArticle” and adds a space between lower-case and upper-case letters “Review Article” (i.e. true casing technique). The output of this phase is C'_T , which has pre-processed translations of classes in T , and $C'_{S-translated}$, which has pre-processed translations for each class in S .

3.3 Terminological Matching

The aim of this phase is to identify potential matches between class labels of S and T . We perform a pairwise lexical and/or semantic similarity between the list of translations of each class in $C'_{S-translated}$ and C'_T to select the best translation for each class in S that matches the corresponding class in T (see Algorithm 1). Jaccard similarity [22] is used to filter the identical concepts instead of using semantic similarity from the beginning because there is no need for extra computations to compute semantic similarity between two identical classes. The reason behind choosing the Jaccard similarity is that according to the experiments conducted for the ontology alignment task for the MultiFarm benchmark in [2], Jaccard similarity has achieved the best score in terms of precision. For non-identical concepts, we compute the semantic similarity using the path length measure, based on WordNet, which returns the shortest path between two words in WordNet hierarchy [3]. If two words are semantically equivalent, i.e., belonging to the same WordNet synset, the path distance is 1.00. We use a specific threshold θ in order to get the set of matched terms (matched classes) M . We obtained the best value of $\theta = 0.9$ which has the best matching results after running the experiments for ten times. If no match is found, we consider this class as a new class that can be added to T and we consider its list of translations as synonyms for that class. Generally, class labels have more than one

Algorithm 1. Terminological Matching

Data: $\mathcal{C}'_{S-translated}$, $\mathcal{C}'_T$, θ similarity threshold
Result: M matched terms, $\mathcal{C}'_{S-translated}$

```

1 foreach  $c_s \in \mathcal{C}'_{S-translated}$ ,  $t \in listOfTranslations$ ,  $c_t \in \mathcal{C}'_T$  do
2    $similarityScore \leftarrow getSimilarity(t, c_t)$ 
3   if  $similarityScore \geq \theta$  then
4      $M ::= (t, c_t)$ 
5      $\mathcal{C}'_{S-translated} = update(\mathcal{C}'_{S-translated}, M)$ 
6 Function  $getSimilarity(sentence1, sentence2)$ :double
7    $similarity \leftarrow getJaccardSimilarity(sentence1, sentence2)$ 
8   if  $similarity \neq 1$  then
9      $similarity \leftarrow (sentenceSimilarity(sentence1, sentence2)$ 
10       $+ sentenceSimilarity(sentence2, sentence1))/2$ 
11   return  $similarity$ 
12 Function  $sentenceSimilarity(sentence1, sentence2)$ :double
13    $simScore \leftarrow 0.0$ 
14    $count \leftarrow 0.0$ 
15   foreach  $w_i \in sentence1.split(" ")$  do
16     foreach  $w_j \in sentence2.split(" ")$  do
17        $pathSim ::= getPathSimilarity(w_i, w_j)$ 
18        $simScore += pathSim.max$ 
19        $count += 1$ 
20    $simScore \leftarrow simScore / count$ 
21   return  $simScore$ 

```

word, for example “InvitedSpeaker”, therefore, the semantic similarity between sentences presented in [21] is adapted as described in Algorithm 1 - line 9. Given two sentences *sentence1* and *sentence2*, the semantic similarity of each sentence with respect to the other is defined by: for each word $w_i \in sentence1$, the word w_j in *sentence2* that has the highest path similarity with w_i is determined. The word similarities are then summed up and normalized with the number of similar words between the two sentences. Next, the same procedure is applied to start with words in *sentence2* to identify the semantic similarity of *sentence2* with respect to *sentence1*. Finally, the resulting similarity scores are combined using a simple average. Based on the similarity results, the best translation is selected and $\mathcal{C}'_{S-translated}$ is updated. For example, in Fig. 2, the class “رئيس” in Arabic, has a list of English translations such as “President, Head, Chief”. After computing the similarity between $\mathcal{C}'_{S-translated}$ and $\mathcal{C}'_T$, “President” has the highest similarityScore of 1.00 with the class “Chairman”, in $\mathcal{C}'_T$, because they are semantically equivalent. Therefore, “President” is selected to be the best translation for “رئيس”. The output of this phase is the list of matched terms M between $\mathcal{C}'_T$ and the updated $\mathcal{C}'_{S-translated}$.

Algorithm 2. Triple Retrieval

Data: $S, \mathcal{C}'_{S-translated}, \mathcal{C}'_T, M$
Result: T_e triples to be enriched

- 1 $S_{translated} \leftarrow \text{translateOntologyClasses}(S, \mathcal{C}'_{S-translated})$
- 2 $newClasses \leftarrow M$
- 3 **while** $!newClasses.isEmpty()$ **do**
- 4 $tempTriples \leftarrow \text{getTriplesForNewClasses}(S_{translated}, newClasses)$
- 5 $newClasses \leftarrow \text{getClasses}(tempTriples).subtract(newClasses)$
- 6 $newTriples \leftarrow newTriples.union(tempTriples)$
- 7 $otherLangTriples \leftarrow \text{getOtherLangTriples}(newTriples, \mathcal{C}'_{S-translated})$
- 8 $T_e \leftarrow newTriples.union(foreignLanguageTriples)$

3.4 Triple Retrieval

The aim of this phase is to identify which and where the new information can be added to T . Each class in S is replaced by its best translation found in $\mathcal{C}'_{S-translated}$ from the previous phase in order to get a translated ontology $S_{translated}$ (see Algorithm 2). We design an iterative process in order to obtain $\mathcal{T}_e$, which is represented by $\langle s, p, o \rangle$, that has all possible multilingual information from S to be added to T . We initiate the iterative process with all matched terms ($newClasses = M$) in order to get all related classes, if exist. The iterative process has three steps: (1) for each class $c \in newClasses$, all triples $tempTriples$ are retrieved from $S_{translated}$ where c is a subject or an object, (2) a new list of new classes is obtained from $tempTriples$, (3) $tempTriples$ is added to $newTriples$ which will be added to T . These three steps are repeated until no new classes can be found ($newClasses.isEmpty() = true$). Next, we retrieve all available information from the other language for each class in $newTriples$ such as $\langle \text{president}, \text{rdfs:label}, \text{"رئيس"@ar} \rangle$. The output of this phase is $\mathcal{T}_e$ which contains all multilingual triples (i.e., in L_s and L_t languages) to be added to T .

3.5 Enrichment

The aim of this phase is to enrich T using triples in $\mathcal{T}_e$. By using OEMCM, the target ontology can be enriched from several ontologies in different natural languages sequentially, i.e. one-to-many enrichment. In this case, $T_{enriched}$ can have more than two natural languages. For instance, English T can be enriched from a German ontology, then the enriched ontology can be enriched again from a different Arabic ontology, i.e. the final result for $T_{enriched}$ is presented in English, German, and Arabic. With the completion of this phase, we have successfully enriched T and create a multilingual ontology from monolingual ones.

3.6 Validation

The aim of this phase is to validate the enriched ontology, which is a crucial step to detect inconsistencies and syntax errors, which might be produced during

```

### https://w3id.org/seo#Publisher
seo:Publisher rdf:type owl:Class ;
    rdfs:subClassOf <http://xmlns.com/foaf/0.1/Organization> ;
    rdfs:comment "The publisher of the event proceedings."@en ;
    rdfs:label "Publisher"@en .
            "Herausgeber"@de .

### http://conference_de#CommitteeMember
conference_de:CommitteeMember rdf:type owl:Class ;
    rdfs:subClassOf <http://xmlns.com/foaf/0.1/Person> ;
    rdfs:label "committee member"@en .
            "Angehörige des Ausschusses"@de .

### https://w3id.org/seo#Chair
seo:Chair rdf:type owl:Class;
    rdfs:subClassOf conference_de:CommitteeMember ;
    rdfs:label "Chair"@en .
            "Vorsitzender"@de .

```

Fig. 3. Small fragment from SEO_{en-de} ontology after the enrichment. The newly added information is marked in bold.

the enrichment process [8]. There are two types of validations: syntactic and semantic validation. In the syntactic validation, we validate $T_{enriched}$ to conform with the W3C RDF standards using the online RDF validation service⁹ which detects syntax errors, such as missing tags. For semantic validation, we use two reasoners, FaCT++ and HermiT, for detecting inconsistencies in $T_{enriched}$ [8].

4 Use Case: Enriching the Scientific Events Ontology

In this use case, we use an example scenario to enrich the SEO_{en} ¹⁰ ontology (with 49 classes), in English, using the MultiFarm dataset (see Sect. 5). We use the Conference ontology (60 classes) and the ConfOf ontology (38 classes), in German and Arabic respectively, as source ontologies. This use case aims to show the whole process starting from submitting the source and target ontologies until producing the enriched multilingual ontology. Here, the source ontology is the German ontology $Conference_{de}$ and the target ontology is the English ontology SEO_{en} . The output is the enriched ontology SEO_{en-de} , which becomes a multilingual ontology in English and German. Table 1 demonstrates the enrichment process for SEO_{en} from $Conference_{de}$ and shows the output sample of each phase starting from the translation phase to the produced set of triples which are used to enrich SEO_{en} . In the terminological matching task, the relevant matching results (with similarity scores in bold) are identified with $\theta \geq 0.9$. The iterative process, in the triple retrieval phase, is initiated with the identified matched terms, for example, *person* class. At the first iteration, six triples

⁹ <https://www.w3.org/RDF/Validator/>.

¹⁰ <https://w3id.org/seo>.

Table 1. Use case: the sample output of each phase, from translation to triple retrieval.

Phase	Output
Translation	$(\text{Thema})_{de} \rightarrow (\text{subject, topic})_{en}$ $(\text{Gutachter})_{de} \rightarrow (\text{reviewer, expert})_{en}$ $(\text{Herausgeber})_{de} \rightarrow (\text{publisher, editor})_{en}$ $(\text{Fortschritte der Konferenz})_{de} \rightarrow (\text{Progress of the conference})_{en}$
Pre-processing	$\text{SizeOrDuration} \rightarrow \text{size duration}$ $\text{WorkshopProposals} \rightarrow \text{workshop proposal}$ $\text{InvitedSpeaker} \rightarrow \text{invite speaker}$ $\text{In-useTrack} \rightarrow \text{use track}$
Terminological matching score results	$(\text{invited speaker, keynote speaker}, 0.57)$ $(\text{person, person}, \mathbf{1.00})$ $(\text{tutorial, tutorial proposals}, 0.78)$ $(\text{prize, award}, \mathbf{1.00})$ $(\text{conference document, license document}, 0.61)$ $(\text{publisher, publisher}, \mathbf{1.00})$ $(\text{conference series, event series}, 0.79)$ $(\text{conference series, symposium series}, 0.75)$ $(\text{proceedings, proceedings}, \mathbf{1.00})$ $(\text{poster, posters track}, 0.78)$
Triple Retrieval (Iterative process)	1^{st} Iteration: $\langle \text{conference contributor}, \text{rdfs:subClassOf}, \text{person} \rangle$ $\langle \text{committee member}, \text{rdfs:subClassOf}, \text{person} \rangle$ 2^{nd} Iteration: $\langle \text{committee member}, \text{rdf:type}, \text{Class} \rangle$ $\langle \text{chairman}, \text{rdfs:subClassOf}, \text{committee member} \rangle$ $\langle \text{conference contributor}, \text{rdf:type}, \text{Class} \rangle$ $\langle \text{invited speaker}, \text{rdfs:subClassOf}, \text{conference contributor} \rangle$ $\langle \text{regular author}, \text{rdfs:subClassOf}, \text{conference contributor} \rangle$
Triple Retrieval ($\mathcal{T}_e$)	$\langle \text{committee member}, \text{rdf:type}, \text{Class} \rangle$ $\langle \text{committee member}, \text{rdfs:label}, \text{"committee member"@en} \rangle$ $\langle \text{committee member}, \text{rdfs:label}, \text{"Angehörige des Ausschusses"@de} \rangle$ $\langle \text{chairman}, \text{rdfs:subClassOf}, \text{committee member} \rangle$

(not all results are exist in the table because of the limited space) are produced such as $\langle \text{conference contributor}, \text{rdfs:subClassOf}, \text{person} \rangle$, where the matched term **person** is located at the object position. New classes are determined from the produced triples such as **conference contributor** and **committee member** (in bold). At the second iteration, all triples that have these new classes, as subject or object, are retrieved, for example; for the **committee member** class, the triples $\langle \text{committee member}, \text{rdf:type}, \text{Class} \rangle$ and $\langle \text{chairman}, \text{rdfs:subClassOf}, \text{committee member} \rangle$ are retrieved. This process is repeated again and new classes are identified from the produced triples such as **chairman**. The iterative process ended at the fifth iteration where three triples are produced without any new classes. The output of this phase is $\mathcal{T}_e$ which has 40 new triples (with 20 new classes and their German labels), to be added to SEO_{en} and produce SEO_{en-de} . Figure 3 shows a small fragment of the enriched

ontology SEO_{en-de} , in Turtle, after completing the enrichment process. The resulting multilingual ontology contains a newly added class `CommitteeMember` with its English and German labels, a new relation `rdfs:subClassOf` between the two classes `CommitteeMember` and `Chair`, and new German labels such as `Herausgeber` and `Vorsitzender` for classes `Publisher` and `Chair` respectively. Similarly, SEO_{en-de} is enriched from the Arabic ontology $ConfOf_{ar}$, where all classes with English labels in SEO_{en-de} are matched with class labels in $ConfOf_{ar}$. The produced $SEO_{en-de-ar}$ has 113 new triples with 37 new classes with their Arabic labels. Final output results can be found at the OECM GitHub repository.

5 Evaluation

The aim of this evaluation is to measure the quality of the cross-lingual matching process in addition to the enrichment process. We use ontologies in MultiFarm benchmark¹¹, a benchmark designed for evaluating cross-lingual ontology matching systems. MultiFarm consists of seven ontologies (*Cmt*, *Conference*, *ConfOf*, *Edas*, *Ekaw*, *Iasted*, *Sigkdd*) originally coming from the Conference benchmark of OAEI, their translation into nine languages (Chinese, Czech, Dutch, French, German, Portuguese, Russian, Spanish and Arabic), and the corresponding cross-lingual alignments between them.

Experimental Setup. All phases of OECM have been implemented using Scala and Apache Spark¹². SANSARDF library¹³ [17] with Apache Jena framework¹⁴ are used to parse and manipulate the input ontologies (as RDF triples). In order to process the class labels, the Stanford CoreNLP¹⁵ [19] is used. All experiments are carried out on Ubuntu 16.04 LTS operating system with an Intel Corei7-4600U CPU @ 2.10 GHz x 4 CPU and 10 GB of memory. In our experiments, we consider English ontologies as target ontologies to be enriched from German and Arabic ontologies.

Our evaluation has three tasks: (1) evaluating the effectiveness of the cross-lingual matching process in OECM compared to the reference alignment provided in the MultiFarm benchmark, (2) comparing OECM matching results with four state-of-the-art approaches, in addition to our previous work (OECM 1.0) [14], and (3) evaluating the quality of the enrichment process.

Effectiveness of OECM. In this experiment, we use the English version of *Cmt* ontology as the source ontology, and German and Arabic versions of *Conference*, *ConfOf*, and *Sigkdd* ontologies as target ontologies. We match class labels in *Cmt* ontology with class labels of German and Arabic versions of *Conference*, *ConfOf*, and *Sigkdd* ontologies separately. The resulting alignments are compared

¹¹ <https://www.irit.fr/recherches/MELODI/multifarm/>.

¹² <https://spark.apache.org/>.

¹³ <https://github.com/SANSARDF/SANSARDF>.

¹⁴ <https://jena.apache.org/>.

¹⁵ <https://stanfordnlp.github.io/CoreNLP/>.

Table 2. Precision, recall and F-measures for the cross-lingual matching

Ontology pairs	German $\times$ English			Arabic $\times$ English					
	Precision	Recall	F-measure	Precision		Recall		F-measure	
				Before	After	Before	After	Before	After
Conference $\times$ Cmt	1.00	0.38	0.56	1.00	1.00	0.33	0.42	0.50	0.59
ConfOf $\times$ Cmt	1.00	0.70	0.82	1.00	1.00	0.30	0.60	0.46	0.75
Sigkdd $\times$ Cmt	1.00	0.90	0.95	1.00	1.00	0.40	0.80	0.57	0.89

with the reference alignments, as a gold standard, provided in the benchmark for each pair of ontologies. Table 2 shows the precision, recall and F-measure of the matching process for each pair of ontologies. OECD achieves the highest precision of 1.00 for all pair of ontologies. Meanwhile, OECD achieves the highest recall and F-measure of 0.90 and 0.95 respectively for matching the German *Sigkdd* with the English *Cmt*. As two authors of this work are native speakers of Arabic, we found some linguistic mistakes in the Arabic ontologies which negatively affect the translation and the matching results. Therefore, we correct these mistakes and make it available at the OECD GitHub repository. Matching results *before* and *after* the corrections are presented in the table, where such corrections have greatly improved the matching results in terms of recall and F-measure. For instance, in matching the Arabic *Sigkdd* with the English *Cmt*, recall and F-measure are enhanced by 40% and 32% respectively.

Comparison with the State-of-the-Art. We identified four of the related approaches (AML, KEPLER, LogMap, and XMap) to be included in our evaluation in addition to OECD 1.0. The other related work, neither publish their code, nor their evaluation datasets [10, 11, 25]. In order to compare our results with the state-of-the-art, we use German (*Conference_{de}*) and Arabic (*Conference_{ar}*) versions of the *Conference* ontology as the source ontologies, and *Ekaw_{en}* and *Edas_{en}* ontologies as the target English ontologies. We choose *Ekaw_{en}* and *Edas_{en}* ontologies in this evaluation because they are used in the state-of-the-art systems for evaluation, as mentioned in the results of OAEI 2018. We generate the gold standard alignments between each pair of ontologies using the Alignment API 4.9¹⁶, as used by the state-of-the-art systems, in order to compute precision, recall, and F-measures. Table 3 shows the comparison between our results against four state-of-the-art approaches and OECD 1.0 (results for matching English and German ontologies only). In addition, we add the updated Arabic ontology (*Conference'_{ar}*) with our linguistic correction in the matching process in order to show the effectiveness of such corrections. The current version of OECD (OECD 1.1) outperforms all other systems in precision, recall and F-measure. For instance, when matching *Conference_{de}* $\times$ *Ekaw_{en}*, OECD 1.1 outperforms LogMap, the highest precision, recall and F-measure among the others, by 29%, 60% and 58% in terms of precision, recall and F-measure respectively. The use of semantic similarity in OECD 1.1 significantly improves the matching results

¹⁶ <http://alignapi.gforge.inria.fr/>.

Table 3. State-of-the-art comparison results. Bold entries are the top scores.

Approaches	Conference _{de} × Ekaw _{en}			Conference _{de} × Edas _{en}		
	Precision	Recall	F-measure	Precision	Recall	F-measure
AML [7]	0.56	0.20	0.29	0.86	0.35	0.50
KEPLER [16]	0.33	0.16	0.22	0.43	0.18	0.25
LogMap [15]	0.71	0.20	0.31	0.71	0.29	0.42
XMap [28]	0.18	0.16	0.17	0.23	0.18	0.20
OECM 1.0 [14]	0.75	0.67	0.71	0.93	0.76	0.84
OECM 1.1	1.00	0.80	0.89	1.00	0.78	0.88
	Conference _{ar} × Ekaw _{en}			Conference _{ar} × Edas _{en}		
	Precision	Recall	F-measure	Precision	Recall	F-measure
AML [7]	0.64	0.39	0.28	0.71	0.42	0.29
KEPLER [16]	0.40	0.30	0.24	0.40	0.30	0.24
LogMap [15]	0.40	0.13	0.08	0.40	0.18	0.12
XMap [28]	1.00	0.0	0.0	1.00	0.00	0.00
OECM 1.1	1.00	0.50	0.67	0.86	0.67	0.75
OECM 1.1	Conference' _{ar} × Ekaw _{en}			Conference' _{ar} × Edas _{en}		
	0.88	0.70	0.78	1.00	0.78	0.88

compared to the results of OECM 1.0. For instance, when matching Conference_{de} × Ekaw_{en}, matching results in OECM 1.0 have been enhanced by 25%, 13%, and 18% in terms of precision, recall and F-measure respectively. When matching Conference_{ar} × Edas_{en}, XMap outperform OECM by 14% in terms of precision, while OECM outperforms it in both recall and f-measure. It is observed that the precision of OECM slightly decreased because of the linguistic mistakes found in Conference_{ar}. When considering Conference'_{ar}, which has the linguistic correction, as a source ontology in this matching, the matching results are improved.

Evaluating the Enrichment Process. According to [4], the enriched ontology can be evaluated by comparing it against a predefined reference ontology (Gold standard). In this experiment, we evaluate the enriched ontology SEO_{en-de} (cf. Sect. 4). A gold standard ontology has been manually created by ontology experts. By comparing SEO_{en-de} with the gold standard, OECM achieves 1.00, 0.80, and 0.89 in terms of precision, recall, and F-measure respectively. This finding confirms the usefulness of our approach in cross-lingual ontology enrichment.

6 Conclusion

We present a fully automated approach, OECM, for building multilingual ontologies. The strength of our contribution lies on building such ontologies from monolingual ones using cross-lingual matching between ontologies concepts. Indo and non-Indo-European languages resources are used for enrichment in order to

illustrate the robustness of our approach. Considering multiple translations of concepts and the use of semantic similarity measures for selecting the best translation have significantly improved the quality of the matching process. Iterative triple retrieval process has been developed to determine which information, from the source ontology, can be added to the target ontology, and where such information should be added. We show the applicability of OECM by presenting a use case for enriching an ontology in the scholarly communication domain. The results of the cross-lingual matching process are found promising compared to five state-of-the-art approaches, involving the previous version of OECM. Furthermore, evaluating the quality of the enrichment process emphasizes the validity of our approach. Finally, we propose some linguistic corrections for the Arabic ontologies in the MultiFarm benchmark that used in our experiment, which considerably enhanced the matching results. In conclusion, our approach provides a springboard for a new way to build multilingual ontologies from monolingual ones. In the future, we intend to further consider properties and individuals in the enrichment process. In addition, we aim to apply optimization methods in order to evaluate the efficiency of OECM when enriching very large ontologies.

Acknowledgments. This work has been supported by the BOOST EU project no. 755175. Shima Ibrahim and Said Fathalla would like to acknowledge the Ministry of Higher Education (MoHE) of Egypt for providing scholarships to conduct this study.

References

1. Ali, M., Fathalla, S., Ibrahim, S., Kholief, M., Hassan, Y.F.: CLOE: a cross-lingual ontology enrichment using multi-agent architecture. In: Enterprise Information Systems, pp. 1–21 (2019)
2. Cheatham, M., Hitzler, P.: String similarity metrics for ontology alignment. In: Alani, H., et al. (eds.) ISWC 2013. LNCS, vol. 8219, pp. 294–309. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41338-4_19
3. Cross, V.: Semantic similarity: a key to ontology alignment. In: Ontology Matching: OM-2018: Proceedings of the ISWC Workshop, p. 61 (2018)
4. Dellschaft, K., Staab, S.: On how to perform a gold standard based evaluation of ontology learning. In: Cruz, I., et al. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 228–241. Springer, Heidelberg (2006). https://doi.org/10.1007/11926078_17
5. Embley, D.W., Liddle, S.W., Lonsdale, D.W., Tijerino, Y.: Multilingual ontologies for cross-language information extraction and semantic search. In: Jeusfeld, M., Delcambre, L., Ling, T.-W. (eds.) ER 2011. LNCS, vol. 6998, pp. 147–160. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24606-7_12
6. Espinoza, M., Gómez-Pérez, A., Mena, E.: Enriching an ontology with multilingual information. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 333–347. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-68234-9_26
7. Faria, D., et al.: Results of AML participation in OAEI 2018. In: Proceedings of the 13th International Workshop on Ontology Matching, pp. 125–131. CEUR-WS (2018)

8. Fathalla, S., Lange, C., Auer, S.: EVENTSKG: a 5-star dataset of top-ranked events in eight computer science communities. In: Hitzler, P., et al. (eds.) ESWC 2019. LNCS, vol. 11503, pp. 427–442. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21348-0_28
9. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: SEO: a scientific events data model. In: International Semantic Web Conference. Springer (2019, in Press)
10. Fu, B., Brennan, R.: Cross-lingual ontology mapping and its use on the multilingual semantic web. *MSW* **571**, 13–20 (2010)
11. Fu, B., Brennan, R., O’Sullivan, D.: A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. *Web Semant. Sci. Serv. Agents World Wide Web* **15**, 15–36 (2012)
12. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. *Web Semant. Sci. Serv. Agents World Wide Web* **11**, 63–71 (2012)
13. Hertling, S., Paulheim, H.: Wikimatch: using Wikipedia for ontology matching. In: *Ontology Matching*, vol. 946 (2012)
14. Ibrahim, S., Fathalla, S., Yazdi, H.S., Lehmann, J., Jabeen, H.: OECM: a cross-lingual approach for ontology enrichment. In: European Semantic Web Conference. Springer (2019, in Press)
15. Jiménez-Ruiz, E., Grau, V.C.: LogMap family participation in the OAEI 2018. In: *Proceedings of the 13th International Workshop on Ontology Matching*, pp. 187–191. CEUR-WS (2018)
16. Kachroudi, M., Diallo, G., Yahia, S.B.: OAEI 2018 results of KEPLER. In: *Proceedings of the 13th International Workshop on Ontology Matching*, pp. 173–178. CEUR-WS (2018)
17. Lehmann, J., et al.: Distributed semantic analytics using the SANSA stack. In: d’Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10588, pp. 147–155. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68204-4_15
18. Lin, F., Krizhanovsky, A.: Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint. In: *RCDL* (2011)
19. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60 (2014). <http://www.aclweb.org/anthology/P/P14/P14-5010>
20. Meilicke, C., et al.: MultiFarm: a benchmark for multilingual ontology matching. In: *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 15, pp. 62–68 (2012)
21. Mihalcea, R., Corley, C., Strapparava, C., et al.: Corpus-based and knowledge-based measures of text semantic similarity. *AAAI* **6**, 775–780 (2006)
22. Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S.: Using of Jaccard coefficient for keywords similarity. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1 (2013)
23. Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E.: Ontology population and enrichment: state of the art. In: Paliouras, G., Spyropoulos, C.D., Tsatsaronis, G. (eds.) *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*. LNCS (LNAI), vol. 6050, pp. 134–166. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20795-2_6
24. Spohr, D., Hollink, L., Cimiano, P.: A machine learning approach to multilingual and cross-lingual ontology matching. In: Aroyo, L., et al. (eds.) ISWC 2011. LNCS, vol. 7031, pp. 665–680. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25073-6_42

25. Tigrine, A.N., Bellahsene, Z., Todorov, K.: Light-weight cross-lingual ontology matching with LYAM++. In: Debruyne, C., et al. (eds.) OTM 2015. LNCS, vol. 9415, pp. 527–544. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26148-5_36
26. Trojahn, C., Fu, B., Zamazal, O., Ritze, D.: State-of-the-art in multilingual and cross-lingual ontology matching. In: Buitelaar, P., Cimiano, P. (eds.) Towards the Multilingual Semantic Web, pp. 119–135. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-43585-4_8
27. Trojahn, C., Quaresma, P., Vieira, R.: A framework for multilingual ontology mapping (2008)
28. Warith Eddine Djeddi, S.B.Y., Khadir, M.T.: XMap results for OAEI 2018. In: Proceedings of the 13th International Workshop on Ontology Matching, pp. 210–215. CEUR-WS (2018)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Multilingual ontology merging using cross-lingual matching

© 2020, IEEE. Reprinted, with permission, from Shimaa Ibrahim, Said Fathalla, Jens Lehmann, and Hajira Jabeen, Multilingual ontology merging using cross-lingual matching, In Proceedings of IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), and 06/2021. DOI: <https://doi.org/10.1109/WIIAT50758.2020.00020>.

Multilingual Ontology Merging Using Cross-lingual Matching

Shimaa Ibrahim^{*†}, Said Fathalla^{*‡}, Jens Lehmann^{*§} and Hajira Jabeen[¶]

^{*}Smart Data Analytics (SDA), University of Bonn, Bonn, Germany

Email: {ibrahim,fathalla,jens.lehmann}@cs.uni-bonn.de

[†]Institute of Graduate Studies and Research, University of Alexandria, Alexandria, Egypt

[‡]Faculty of Science, University of Alexandria, Alexandria, Egypt

[§]NetMedia Department, Fraunhofer IAIS, Dresden, Germany

[¶]Cluster of Excellence on Plant Sciences (CEPLAS), University of Cologne, Cologne, Germany

Email: hajira.jabeen@uni-koeln.de

Abstract—With the growing amount of multilingual data on the Semantic Web, several ontologies (in different natural languages) have been developed to model the same domain. Creating multilingual ontologies by merging such monolingual ones is important to promote semantic interoperability among different ontologies in different natural languages. This is a step towards achieving the multilingual Semantic Web. In this paper, we propose MULON, an approach for merging monolingual ontologies in different natural languages producing a multilingual ontology. MULON approach comprises three modules; *Preparation Module*, *Merging Module*, and *Assessment Module*. We consider both classes and properties in the merging process. We present three real-world use cases describing the usability of the MULON approach in different domains. We assess the quality of the merged ontologies using a set of predefined assessment metrics. MULON has been implemented using Scala and Apache Spark under an open-source license. We have compared our cross-lingual matching results with the results from the Ontology Alignment Evaluation Initiative (OAEI 2019). MULON has achieved relatively high precision, recall, and F-measure in comparison to three state-of-the-art approaches in the matching process and significantly higher coverage without any redundancy in the merging process.

Index Terms—ontology merging, cross-lingual matching, multilingual ontology, quality assessment, knowledge management

I. INTRODUCTION

Ontologies are being widely used in various fields of science such as information retrieval, question-answering, document retrieval, and text summarization. With the rapid expansion of multilingual data on the Semantic Web, more ontologies have become available in different natural languages. According to Linked Open Vocabularies (LOV)¹, English is by far the most prominent language, i.e., the majority of ontologies in the Semantic Web are in English, however, there are many ontologies in other Indo-European languages also exist. Specifically, out of a total of 720 vocabularies found in LOV, 533 are in English, 57 in French, and 34 in German. Few ontologies in LOV are in non-Indo-European languages, i.e. nine ontologies in Chinese and seven in Arabic. Some ontologies have been designed to model already existing domains in an independent way, which leads to ontologies with overlapping and redundant

information. Although the Web of Data contains information in various natural languages and internationalization is one of the design goals of OWL standard (i.e., ontologies should handle multiple languages), it still lacks efficient mechanisms to automatically exploit and reconcile such information [19], [13]. Therefore, for building a multilingual web and enhancing semantic interoperability between ontologies in different natural languages, approaches for building multilingual ontologies by merging the existing ones must be developed [13], [29]. In multilingual ontologies, resources (classes and properties) can be published in a language-independent way, associated with language-dependent (linguistic) information, which supports access across various natural languages. Merging means creating a single ontology to provide a unified view of the input ontologies by maintaining all information contained in them [26]. The creation of such ontology is a complex task and requires considerable adaptation and rigorous techniques to control various steps of the creation, especially when merging ontologies in different natural languages. Identification of mappings between multilingual input ontologies, the first step in the merging process, plays a vital role in the ontology merging process [5]. Most of the existing work in ontology merging focuses only on English ontologies [24], [26], [2], without considering other ontologies in different natural languages. To the best of our knowledge, no work has been done for creating multilingual ontologies by merging ontologies in different natural languages. However, our previous work OECM [15], a fully automated ontology enrichment approach, creates a multilingual ontology from ontologies in different natural languages. OECM enriches an ontology from another one in a different natural language. We have identified two limitations in OECM; 1) only classes from the source ontology are added to the target ontology (i.e., properties are not considered), and 2) experts are needed for creating a reference ontology to evaluate the enrichment process, which is an infeasible task for large ontologies.

In this paper, we propose MULON (MULTilingual Ontology mergiNg) approach for creating a multilingual ontology by merging two ontologies, in different natural languages. As an illustration example, first, MULON identifies cross-lingual

¹<https://lov.linkeddata.es/dataset/lov/vocabs>

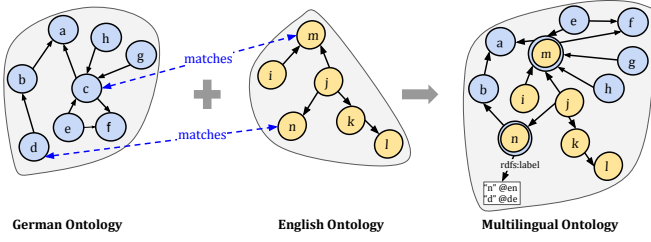


Fig. 1. An illustration of generating a multilingual ontology from merging German and English ontologies.

matches between input ontologies using cross-lingual matching techniques, then adds them to the merged multilingual ontology by adding `rdfs:label` for each language (using language-tagged strings) as shown in Figure 1. Cross-lingual matching helps to lower redundancy in the merged ontology. MULON comprises three modules: 1) *preparation module*, which extracts and processes resources of the input ontologies, 2) *merging module*, in which the match and the merge processes occur, and 3) *assessment module*, which validates and assesses the quality of the merged ontology. By providing such approach that enables researchers to easily integrate ontologies in different natural languages, MULON is an essential step in the direction of a multilingual Semantic Web. MULON is available in a public repository in GitHub², in which the source code is documented, describing each configurable parameter and function.

The contributions of this work can be summarized in the following points:

- We propose MULON to create a multilingual ontology by merging two ontologies in different natural languages based on linguistic and structure knowledge of these ontologies,
- We consider not only classes but also properties,
- MULON adapts a set of quality metrics to automatically assess not only the merged ontology but also any ontology,
- MULON considers ontologies in Indo and non-Indo-European languages,
- We present three real-world use cases demonstrating the usability of the MULON approach and showing its impact to the Semantic Web community, and
- MULON empirically showed significantly better results when compared to three state-of-the-art approaches.

The remainder of this paper is structured as follows: we present an overview of related work in section II. An overview of the proposed approach is described in section III. In order to illustrate the usability of MULON, three use cases are presented in section IV. Experiments and evaluation results are presented in section V. Finally, we conclude with an outline of the future work in section VI.

II. RELATED WORK

A recent review of the literature on multilingual Web of Data found that in order to support Multilingual Semantic Web, techniques for building multilingual ontologies are desired [13]. Bhatt [1] proposed an unsupervised learning algorithm to automatically learn multilingual ontology from unstructured text. Florence and Lourdasamy [10], [18] developed an approach for building multilingual ontologies by 1) collecting terms in different natural languages for a specific domain from different resources, 2) creating identical monolingual ontologies based on these terms for each natural language, and 3) mapping and merging these monolingual ontologies. Merging different ontologies in different natural languages is a good solution for building multilingual ontologies. However, all the literature has focused on merging ontologies in the same natural language [2], without considering other ontologies in different natural languages. PROMPT [24] is one of the earliest tools in ontology merging, which utilizes a semi-automated approach for ontology merging based on linguistic and structural knowledge. PROMPT lets the user choose between merging suggestions. ATOM [26] is an automatic target-driven approach to merge large taxonomies such as product catalogs or web directories. It preserves the structure of the target taxonomy as much as possible. Fahad [5] has created an algorithm to map class expressions of concepts and merge them in an automatic ontology merging process. His algorithm addresses how to combine multiple axiomatic definitions into one compact definition considering the consistency of the merged solution.

In order to merge ontologies in different natural languages, cross-lingual ontology matching is required. Cross-lingual ontology matching techniques are used for matching linguistic information across ontologies [13], [29]. Fu et al. [11], [12] proposed an approach to match English and Chinese ontologies by considering the semantics of the target ontology, the mapping intent, the operating domain, the time and resource constraints, and user feedback. In the context of OAEI 2019 campaign³ for evaluating ontology matching technologies, AML [6], LogMap [17] and Wiktionary Matcher [25] provide high-quality alignments for the matching task. AML [6] is based on lexical and structural matching algorithms. It utilizes machine translation technologies, such as Microsoft Translator, before starting the matching process. AML uses external biomedical ontologies and WordNet⁴ as sources of background knowledge. LogMap [17] implements optimized data structures for lexically and structurally indexing the input ontologies for the matching process. It is an iterative process that starts from initial mappings (almost exact lexical correspondences) till discovering new mappings. Therefore, LogMap is not able to find matching between ontologies that do not provide enough lexical information. Wiktionary Matcher [25] is an element-level, label-based matcher which uses multiple language versions of Wiktionary. Good literature

²<https://smartdataanalytics.github.io/MULON/>

³<http://oaei.ontologymatching.org/2019/results/multifarm/index.html>

⁴<https://wordnet.princeton.edu/>

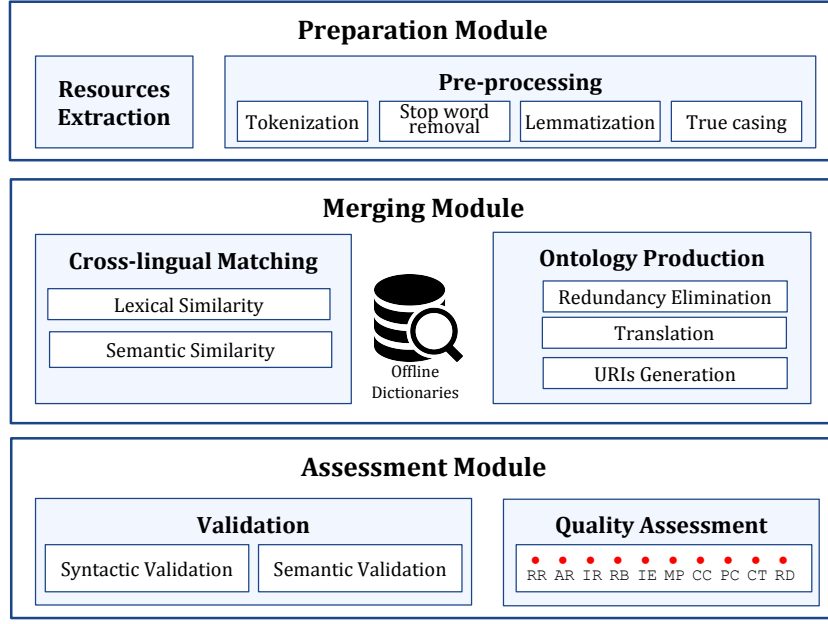


Fig. 2. MULON architecture

of the state-of-the-art approaches in cross-lingual ontology matching is provided in [29].

In summary, the major challenge faced by the ontology merging research is to design an automatic and consistent multilingual ontology merging approach that produces reliable (good quality) and usable multilingual merged ontology. We found that previous work has only focused on merging monolingual ontologies. Consequently, we propose an approach for merging monolingual ontologies to generate a multilingual ontology, where each resource is presented by more than one natural language.

III. MULON ARCHITECTURE

The architecture of MULON is comprised of three modules namely, preparation, merging, and assessment (cf. Figure 2). In the following subsections, we describe each of these modules in detail. The input is two ontologies (O_1 and O_2) in two different natural languages L_1 and L_2 respectively. The output is 1) a multilingual merged ontology (O_m) and 2) an assessment sheet presenting the quality of the merged ontology.

A. Preparation Module

This module is responsible for extracting resources from the input ontologies and prepare them for further steps. Preparation module has two components:

Resources Extraction. The aim of this component is to extract all resources (both classes and properties) from the two input ontologies into the resources matrix R . R is a two-dimensional matrix ($n \times 4$), where n is the number of resources. Each row in R is represented as a tuple of $\langle \text{resource}, \text{type}, \text{source}, \text{language} \rangle$, which contains the resource label, the type of the resource ('C' for

class and 'P' for property), the source ontology (O_1 or O_2), and the language tag. For more illustration, consider the tuple $\langle \text{BestPosterAward}, C, O_1, \text{en} \rangle$, in which the resource "BestPosterAward" is a class extracted from O_1 and its language is English. The output of this component is R .

Pre-processing. To prepare extracted resources for the next modules, all resources are processed by employing a variety of natural language processing techniques, such as stop words removal, tokenization, POS-tagging (part-of-speech tagging), and lemmatization. Normalization methods and regular expressions are used to remove stop words, punctuations, symbols, and additional white spaces. Furthermore, true casing techniques are used to recognize resources with camel cases, e.g., "BestPosterAward" and "isSponsorOf" became "Best Poster Award" and "is Sponsor Of" respectively. The output of this component is the pre-processed resources matrix (R').

B. Merging Module

This module is responsible for merging the two input ontologies into a new multilingual one O_m . Generally, there are two types of ontology merging [27]: a) *Symmetric merging*; which aims to integrate all resources (classes and properties) in the input ontologies to a single ontology, i.e. preserves all input resources (which is used in this paper), and b) *Asymmetric merging*; which considers one of the input ontologies as the target ontology and merges the remaining non-redundant resources of the input ontologies into the target ontology. Merging module has two components:

Cross-lingual Matching. The aim of this component is to identify potential matches between the extracted resources. We divide the ontology resources into two sets: classes and properties. We match only resources of the same set, i.e., classes

in one ontology are matched with the classes of the other one. Additionally, we identify particular threshold values in the matching process to select the top matched results. Cross-lingual matches are identified using machine translation tools and similarity measures. Machine translation tools are able to return useful translations for a very large number of resources in the cross-lingual matching task [14]. We build offline dictionaries for all resources in R using Yandex translate API⁵. Such offline dictionaries reduce extra translation overhead that is produced when translating the same resources several times. Each resource in O_1 is translated from L_1 to L_2 and similarly, each resource in O_2 is also translated (to be used in merging). In order to identify the matched resources between O_1 and O_2 , we perform a pairwise lexical and semantic similarity between resources of O_1 and the translated resources of O_2 (now both are in L_1). First, Jaccard similarity [23] is used to identify the identical resources in order to save extra computations that are produced when computing semantic similarity between already identical resources. We choose Jaccard similarity because it achieved the best score in terms of precision for the ontology alignment task for the MultiFarm benchmark in [3]. Second, semantic similarity is used to compute the similarity between non-identical resources (by Jaccard). For each pair of resources, the path length is computed, based on WordNet⁴, which returns the shortest path between two words in WordNet hierarchy [4]. The output of this component is a matched resources matrix (M). M is a two-dimensional matrix ($m \times 4$), where m is the number of matched resources. Each row in M is represented as a tuple of $\langle \text{resource}(O_1), \text{resource}(O_2), \text{type}, \text{simScore} \rangle$, which contains a resource from O_1 , a resource from O_2 , type of the resources, and the similarity score between them. For example, $\langle \text{Car}, \text{سيارة}, \text{C}, 1.00 \rangle$, i.e. “Car” in English and “سيارة” in Arabic are two resources of type class, from O_1 and O_2 respectively, with similarity score of 1.00.

Ontology Production. The aim of this component is to eliminate redundancy that might occur in the merged ontology. Therefore, the input ontologies have been merged based on linguistic and structural knowledge. First, we select matched resources, from M , with specific thresholds θ (for classes) and α (for properties) in order to get matched classes C_{match} and matched properties P_{match} respectively. Then, for each pair of matched resources (r_1, r_2) , where $r_1 \in \text{resource}(O_1)$ and $r_2 \in \text{resource}(O_2)$, we merge these matched resources by representing them as r_m associated with the two labels of r_1 and r_2 with the corresponding language tags. For instance, consider the match $\langle \text{hatWichtigesDatum}, \text{hasImportantDates}, \text{P}, 1.00 \rangle$, these two properties are merged and associated with English and German labels as $\langle \text{hasImportantDates}, \text{rdfs:label}, \text{"has important dates"@en} \rangle$ and $\langle \text{hasImportantDates}, \text{rdfs:label}, \text{"hat wichtiges datum"@de} \rangle$. For non-matched resources, we translate them, using the offline dictionaries, to the other language. Similarly, a new label

with the other language tag is created. For instance, the non-matched resource “InvitedSpeakers” is translated to “EingeladeneReferenten” and associated with English and German labels as in the previous example. At this step, we have a multilingual ontology containing all information in both input ontologies.

C. Assessment Module

This module is responsible for validating and assessing the quality of the merged ontology.

Validation. The aim of this component is to validate the merged ontology, which is a crucial step to detect inconsistencies and syntax errors, which might be introduced during the merging process [7]. There are two types of validation: syntactic and semantic validation. In the syntactic validation, we validate O_m to conform with the W3C RDF standards using the online RDF validation service⁶ which detects syntax errors, such as missing tags. For semantic validation, we have used FaCT++⁷ reasoner for detecting inconsistencies in O_m .

Quality Assessment. The aim of this component is to assess the quality of the merged ontology. Evaluating the quality of the merged ontology strongly depends on the quality of the input ontologies. We assess the quality of the two input ontologies and the resulting merged ontology by adapting and reformulating the metrics defined in [28], [21], [27].

- **Relationship richness (RR)** [28]: refers to the diversity of relations and their position in the ontology. The more relations the ontology has (except `rdfs:subClassOf` relation), the richer it is. The quality score function $f_{RR} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{RR}(O) = \frac{|P_{obj}|}{|P_{subClassOf}| + |P_{obj}|} \quad (1)$$

where P_{obj} represents the relationships (i.e. object properties) and $P_{subClassOf}$ represents the `rdfs:subClassOf` relations in O .

- **Attribute Richness (AR)** [28]: refers to how much knowledge about classes is in the schema. The more attributes are defined, the more knowledge the ontology provides. The quality score function $f_{AR} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{AR}(O) = \frac{|P_{attr}|}{|C|} \quad (2)$$

where C represents the ontology classes and P_{attr} represents the attributes (i.e., data properties) for all classes.

- **Inheritance Richness (IR)** [28]: refers to how well knowledge is distributed across different levels in the ontology. The more `rdfs:subClassOf` relations, the wider range of general knowledge the ontology provides. The quality score function $f_{IR} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{IR}(O) = \frac{|P_{subClassOf}|}{|C|} \quad (3)$$

⁵<https://tech.yandex.com/translate/>

⁶<https://www.w3.org/RDF/Validator/>

⁷<https://github.com/ethz-asl/libfactplusplus>

- *Readability (RB)* [28]: refers to the existence of human-readable descriptions (HRD) in the ontology, such as comments, labels, or descriptions. The more HRD exists, the more readable the ontology is. The quality score function $f_{RB} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{RB}(O) = \frac{|HRD|}{|R|} \quad (4)$$

where $HRD \in \{label, comment, description\}$ and R represents the ontology resources.

- *Isolated Elements (IE)* [21]: refers to classes and properties which are defined but not connected to the rest of the ontology, i.e. not used. The quality score function $f_{IE} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{IE}(O) = \frac{|R_{isolated}|}{|R|} \quad (5)$$

where $R_{isolated}$ represents resources which are defined but not used in O .

- *Missing Domain or Range in Properties (MP)* [21]: refers to missing information about properties. The less of missing information about properties, the more the ontology is complete. The quality score function $f_{MP} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{MP}(O) = \frac{|P_{incomplete}|}{|P|} \quad (6)$$

where $P_{incomplete}$ represents properties which do not have domain or range.

- *Redundancy (RD)* [27]: refers to how many redundant resources exist. Resources which are syntactically (e.g. “is-MemberOf” and “is_member_of”) or semantically (e.g. “Chair” and “Chairman”) close are considered as redundant resources. The quality score function $f_{RD} : O \rightarrow \mathbb{R}$ for an input ontology O is defined as follows:

$$f_{RD}(O) = \frac{|R_r|}{|R|} \quad (7)$$

where R_r represents the redundant resources in O .

All the previous metrics can be used to assess the quality of any ontology. The following metrics are adapted in order to assess the quality of the merged ontology by comparing it with the input ontologies.

- *Class Coverage (CC)* [27]: refers to how many classes in the input ontologies ($C_1 + C_2$) are preserved in the merged ontology (C_m) excluding matched classes (C_{match}). The more preserved classes exist, the more coverage the ontology provides. The quality score function $f_{CC} : O \rightarrow \mathbb{R}$ for the merged ontology O is defined as follows:

$$f_{CC}(O) = \frac{|C_m|}{|C_1| + |C_2| - |C_{match}|} \quad (8)$$

- *Property Coverage (PC)* [27]: refers to how many properties in the input ontologies are preserved in the merged ontology excluding matched properties P_{match} . The more

preserved properties exist, the more coverage the ontology provides. The quality score function $f_{PC} : O \rightarrow \mathbb{R}$ for the merged ontology O is defined as follows:

$$f_{PC}(O) = \frac{|P_m|}{|P_1| + |P_2| - |P_{match}|} \quad (9)$$

where P_m , P_1 , P_2 are properties of merged and two input ontologies respectively. P_{match} are the matched properties between the two input ontologies.

- *Compactness (CT)* [27]: refers to how much the size of the merged ontology compared to the combination of the two input ontologies. The smaller the size of merged ontology, the more the ontology is compacted, e.g. if some resources are removed in order to avoid redundant resources in the merged ontology. The quality score function $f_{CT} : O \rightarrow \mathbb{R}$ for the merged ontology O is defined as follows:

$$f_{CT}(O) = \frac{|R_m|}{|R_1| + |R_2|} \quad (10)$$

where R_m , R_1 , R_2 are resources of merged and two input ontologies respectively.

At the end of this module, MULON provides the user with the resulting merged ontology O_m in addition to the assessment sheet which has the quality assessment results for the input and merged ontology.

IV. IMPACT AND USE CASES

The main goal of MULON is to develop a generic approach which can merge monolingual ontologies into a multilingual ontology, offering algorithms for cross-lingual ontology matching, and automatically assess the quality of not only the merged ontology but, in principle, any ontology.

Impact on the Semantic Web Community. Ontologies occupy a key role in capturing space knowledge and providing a standard understanding of knowledge across different domains and different natural languages. Considering that ontologies often have overlapping and redundant information, MULON can be applied in numerous applications to effectively merge such ontologies. To validate this, we investigate use case implementations in several domains and projects below.

Fiscal Data. OpenBudgets⁸ has build a platform to upload, visualize, and analyze fiscal data coming from different countries in different natural languages. Fiscal datasets are heterogeneous in nature, since they are published by various administrations. One of the major challenges faced in OpenBudgets, is to link concepts in different languages in order to analyze fiscal data [22]. OpenBudgets is currently investigating MULON for cross-lingual vocabulary matching. The vocabulary matching can allow automated cross-regional budget comparisons which was not easily possible before.

Food. AutoChef [16] is a framework for culinary arts. It takes input from well-rated recipes of different cuisines and invents new recipes by recombining the instructions, spices, and ingredients. AutoChef aims to extend their framework by

⁸<http://openbudgets.eu/>

TABLE I
STATISTICS OF ALL INPUT AND MERGED ONTOLOGIES.

	Ontology Name	Classes	Object Properties	Data Properties
Input	<i>SEO</i>	106	66	23
	<i>Ekaw</i>	74	33	0
	<i>Edas</i>	104	30	20
	<i>Conference</i>	60	46	18
	<i>ConfOf</i>	38	13	23
	<i>Cmt</i>	36	49	10
	<i>Iasted</i>	140	38	3
	<i>Sigkdd</i>	49	17	11
Merging Pairs	$Conference_{de} \times SEO_{en}$	156	103	40
	$Conference_{de} \times Ekaw_{en}$	121	78	18
	$ConfOf_{ar} \times Edas_{en}$	132	41	38
	$Cmt_{ar} \times Ekaw_{en}$	99	78	10
	$Iasted_{fr} \times Edas_{en}$	233	67	23
	$Sigkdd_{fr} \times Ekaw_{en}$	118	51	11

considering different kinds of recipes from different countries, i.e., in different natural languages. Therefore, at present, AutoChef is investigating MULON for merging recipe-ontologies into a multilingual ontology which can assist in the discovery of new ingredient relationships and deliver new recipes.

Scholarly Communication. Scientific events play a key role in publishing scholarly data on the Web. They are considered as the focal point for establishing scientific relations between scholarly objects in the scholarly communication domain. Scholarly events are organized by people from different countries, therefore event’s related information, such as a call for papers and events topics, is published, in some cases, in different natural languages, particularly for local events [8]. To facilitate the management of such data, a multilingual integration approach is desired. MULON has successfully merged the Scientific Events Ontology⁹ (SEO_{en}) [9], in English, which describes the scientific events, and the Conference ontology ($Conference_{de}$), in German, which exists in the MultiFarm dataset (see section V). Statistics for these ontologies and the merged ontology are presented in Table I. MULON was able to find nine matched classes and four matched properties between the two ontologies and remove duplicated resources in the merged ontology. The merged ontology was consistent and free of syntax errors. The merging process was free of redundancy and achieved significantly high results in terms of coverage and compactness. The produced merged ontology and the assessment sheet are available at the GitHub² repository.

V. EVALUATION

The aim of this evaluation is to measure the quality of the cross-lingual matching process in addition to the quality of the merged ontology. We use ontologies in MultiFarm¹⁰, a benchmark designed for evaluating cross-lingual ontology matching systems. MultiFarm consists of seven ontologies (*Cmt*, *Conference*, *ConfOf*, *Iasted*, *Sigkdd*, *Edas*, and *Ekaw*) originally coming from the Conference benchmark of OAEI, translations of the first five ontologies (i.e., *Edas* and *Ekaw*

are available in English only) into nine languages (Chinese, Czech, Dutch, French, German, Portuguese, Russian, Spanish and Arabic), and the corresponding cross-lingual alignments between them.

Experimental Setup. All modules of MULON have been implemented using Scala and Apache Spark¹¹ (available under Apache License 2.0) which is a distributed processing engine for large-scale data. SANSARDF library¹² [20] with Apache Jena framework¹³ are used to parse and manipulate the input ontologies (as RDF triples) in a distributed manner. We used Apache Spark and SANSARDF library because we want to extend this work to merge large-scale ontologies efficiently (see section VI). In order to process the resource labels, the Stanford CoreNLP¹⁴ is used. All experiments are carried out on Ubuntu 16.04 LTS operating system with an Intel Corei7-4600U CPU @ 2.10GHz x 4 CPU and 10 GB of memory. Our evaluation has two tasks: 1) comparing MULON matching results with three state-of-the-art approaches, and 2) evaluating the merged ontology.

Comparison with the state-of-the-art. We identified three state-of-the-art approaches (AML [6], LogMap [17], and Wikitionary Matcher [25]) to be included in our evaluation. The other related works, neither published their codes, nor evaluation data [12], [11]. In order to compare our results with the state-of-the-art, we use the five ontologies *Cmt*, *Conference*, *ConfOf*, *Iasted*, and *Sigkdd* in German, Arabic, and French, in addition to the two ontologies *Edas* and *Ekaw* in English as mentioned in the results of OAEI 2019³. We chose Arabic as a non-Indo European language to show the promising potential of MULON across different languages. *Edas* and *Ekaw* are available in English only because they are used for blind evaluation in OAEI. In order to translate such ontologies into German, Arabic, and French, experts will be needed for each language to validate such translations which makes this task is difficult to achieve. As we didn’t participate in OAEI, we found that there is no need to translate them in order to avoid translation errors. Therefore, we consider all different ontology pairs in different languages where *Edas* or *Ekaw* are in English. Inspired by the information retrieval community, the quality of the matching process can be measured through precision, recall, and F-measure metrics. We use the gold standard alignments between each pair of ontologies which exist in the dataset in order to compute precision, recall, and F-measure. Precision is the fraction of the retrieved matched resources that are relevant, while recall is the fraction of relevant matched resources retrieved by MULON. The F-measure is the harmonic mean of precision and recall. Formally, precision is defined as $TP/(TP + FP)$ and recall is defined as $TP/(TP + FN)$, where TP (true positive) is the matching results which are retrieved by MULON and exist in the gold standard, FP (false positive) is the matching results which are retrieved by MULON and do

⁹<https://w3id.org/seo>

¹⁰<https://www.irit.fr/recherches/MELODI/multifarm/>

¹¹<https://spark.apache.org/>

¹²<https://github.com/SANSARDF/SANSARDF>

¹³<https://jena.apache.org/>

¹⁴<https://stanfordnlp.github.io/CoreNLP/>

TABLE II

STATE-OF-THE-ART COMPARISON RESULTS. MULON* DESCRIBES RESULTS WITH ADJUSTED PRECISION AND F-MEASURE. RED ENTRIES ARE THE TOP SCORES FOR EACH METRIC PER ROW.

Ontology Pairs	AML			LogMap			Wiktionary			MULON			MULON*	
	P	R	F	P	R	F	P	R	F	P	R	F	P	F
$Cmt_{de} \times Edas_{en}$	1.00	0.31	0.47	0.80	0.31	0.44	0.83	0.38	0.53	0.64	0.58	0.61	0.88	0.70
$Cmt_{de} \times Ekaw_{en}$	0.80	0.36	0.50	1.00	0.36	0.53	0.75	0.27	0.40	1.00	0.45	0.63	1.00	0.63
$ConfOf_{de} \times Edas_{en}$	0.93	0.68	0.79	0.86	0.63	0.73	0.82	0.47	0.60	0.81	0.68	0.74	1.00	0.81
$ConfOf_{de} \times Ekaw_{en}$	0.92	0.60	0.73	0.93	0.65	0.76	0.80	0.40	0.53	0.92	0.58	0.71	1.00	0.73
$Conference_{de} \times Edas_{en}$	0.86	0.35	0.50	0.71	0.29	0.42	0.71	0.29	0.42	0.69	0.56	0.62	1.00	0.72
$Conference_{de} \times Ekaw_{en}$	0.56	0.20	0.29	0.71	0.20	0.31	0.83	0.20	0.32	0.85	0.44	0.58	1.00	0.61
$Iasted_{de} \times Edas_{en}$	0.38	0.16	0.22	0.43	0.16	0.23	0.67	0.21	0.32	0.86	0.32	0.46	1.00	0.48
$Iasted_{de} \times Ekaw_{en}$	0.83	0.50	0.63	0.75	0.60	0.67	0.60	0.30	0.40	1.00	0.60	0.75	1.00	0.75
$Sigkdd_{de} \times Edas_{en}$	0.75	0.40	0.52	0.63	0.33	0.43	0.67	0.27	0.38	0.67	0.40	0.50	0.86	0.55
$Sigkdd_{de} \times Ekaw_{en}$	1.00	0.45	0.63	1.00	0.45	0.63	0.83	0.45	0.59	1.00	0.55	0.71	1.00	0.71
$Cmt_{ar} \times Edas_{en}$	1.00	0.38	0.56	0.60	0.23	0.33	1.00	0.00	0.00	0.67	0.67	0.67	0.89	0.76
$Cmt_{ar} \times Ekaw_{en}$	1.00	0.27	0.43	1.00	0.27	0.43	1.00	0.00	0.00	1.00	0.45	0.63	1.00	0.63
$ConfOf_{ar} \times Edas_{en}$	0.86	0.32	0.46	0.86	0.32	0.46	1.00	0.11	0.19	0.73	0.58	0.65	1.00	0.73
$ConfOf_{ar} \times Ekaw_{en}$	1.00	0.60	0.75	1.00	0.60	0.75	1.00	0.10	0.18	0.91	0.56	0.69	1.00	0.71
$Conference_{ar} \times Edas_{en}$	0.71	0.29	0.42	0.40	0.12	0.18	1.00	0.00	0.00	0.62	0.50	0.55	0.89	0.64
$Conference_{ar} \times Ekaw_{en}$	0.60	0.24	0.34	0.40	0.08	0.13	1.00	0.00	0.00	0.78	0.28	0.41	1.00	0.44
$Iasted_{ar} \times Edas_{en}$	0.60	0.32	0.41	0.43	0.16	0.23	1.00	0.05	0.10	0.86	0.32	0.46	1.00	0.48
$Iasted_{ar} \times Ekaw_{en}$	0.57	0.40	0.47	1.00	0.30	0.46	1.00	0.10	0.18	1.00	0.57	0.73	1.00	0.73
$Sigkdd_{ar} \times Edas_{en}$	1.00	0.47	0.64	0.71	0.33	0.45	1.00	0.07	0.13	0.67	0.40	0.50	0.86	0.55
$Sigkdd_{ar} \times Ekaw_{en}$	1.00	0.27	0.43	1.00	0.27	0.43	1.00	0.09	0.17	1.00	0.36	0.53	1.00	0.53
$Cmt_{fr} \times Edas_{en}$	1.00	0.62	0.76	n/a	n/a	n/a	0.60	0.46	0.52	0.88	0.54	0.67	1.00	0.70
$Cmt_{fr} \times Ekaw_{en}$	1.00	0.27	0.43	0.80	0.36	0.50	0.67	0.36	0.47	0.80	0.36	0.50	1.00	0.53
$ConfOf_{fr} \times Edas_{en}$	0.90	0.47	0.62	0.73	0.42	0.53	0.73	0.42	0.53	0.58	0.58	0.58	1.00	0.73
$ConfOf_{fr} \times Ekaw_{en}$	1.00	0.70	0.82	1.00	0.60	0.75	0.67	0.30	0.41	0.91	0.50	0.65	1.00	0.67
$Conference_{fr} \times Edas_{en}$	0.57	0.24	0.33	0.43	0.18	0.25	0.57	0.24	0.33	0.64	0.44	0.52	0.88	0.58
$Conference_{fr} \times Ekaw_{en}$	0.55	0.24	0.33	0.50	0.20	0.29	0.50	0.16	0.24	0.73	0.32	0.44	1.00	0.48
$Iasted_{fr} \times Edas_{en}$	0.67	0.32	0.43	0.50	0.26	0.34	0.42	0.26	0.32	0.64	0.37	0.47	1.00	0.54
$Iasted_{fr} \times Ekaw_{en}$	0.67	0.40	0.50	0.63	0.50	0.56	0.63	0.50	0.56	0.57	0.40	0.47	1.00	0.57
$Sigkdd_{fr} \times Edas_{en}$	0.80	0.53	0.64	0.71	0.33	0.45	0.60	0.40	0.48	1.00	0.50	0.67	1.00	0.67
$Sigkdd_{fr} \times Ekaw_{en}$	1.00	0.55	0.71	1.00	0.36	0.53	0.86	0.55	0.67	1.00	0.45	0.63	1.00	0.63

not exist in the gold standard, and FN (false negative) is the matching results which exist in the gold standard and MULON could not retrieve them. Table II shows a comparison between MULON's results for matching 30 pair of ontologies against three state-of-the-art systems. Surprisingly, we found new alignments for 22 pair of ontologies which were missing in the gold standard, e.g., the German and English properties "ist Mitglied von" and "is a member of". MULON* represents results when considering the new alignments. It provides the matching results with the adjusted precision and F-measure. Therefore, MULON* presents results that are not false positive in practice. MULON outperforms most of the other systems in terms of precision, recall, and F-measure when not considering the new alignments as false positive. For instance, when matching $Conference_{de} \times Ekaw_{en}$, MULON outperforms Wiktionary Matcher, the highest precision, recall and F-measure among the others in matching German and English ontologies, by 17%, 22% and 29% in terms of precision, recall, and F-measure respectively. When matching $ConfOf_{ar} \times Edas_{en}$, MULON outperform AML and LogMap, the highest recall and F-measure among the others in matching Arabic and English ontologies, by 26% and 27% in terms of recall and F-measure respectively. When matching $Iasted_{fr} \times Edas_{en}$, MULON outperform AML, the highest precision, recall, and F-measure among the others in matching French and English ontologies, by 33%, 5%, and 11% in terms of precision, recall, and F-measure respectively.

Evaluating the merged ontology. In order to evaluate the merged ontology, we choose randomly five pairs of ontologies

($Conference_{de} \times Ekaw_{en}$, $ConfOf_{ar} \times Edas_{en}$, $Cmt_{ar} \times Ekaw_{en}$, $Iasted_{fr} \times Edas_{en}$, and $Sigkdd_{fr} \times Ekaw_{en}$) from Table II. We validate and assess the quality of the merging process using the assessment metrics described in subsection III-C. We found that all merged ontologies are consistent and free of syntax errors. Table III shows the assessment results for each output ontology. MULON achieves the highest relationship richness (RR) of 40% for merging $Cmt_{ar} \times Ekaw_{en}$. In terms of inheritance richness (IR), MULON achieves the highest results of 141% for merging $Iasted_{fr} \times Edas_{en}$ where number of `rdfs:subClassOf` relation in the merged ontology (328) is greater than number of classes (233) (see Table I). In terms of readability, MULON achieves significantly high results because it adds HRD (two `rdfs:label` for each resource) to the merged ontology. MULON achieves a full class coverage of 100% for the merging pair of ontologies $Conference_{de} \times Ekaw_{en}$, $Cmt_{ar} \times Ekaw_{en}$, $Iasted_{fr} \times Edas_{en}$, and $Sigkdd_{fr} \times Ekaw_{en}$, and 99% for $ConfOf_{ar} \times Edas_{en}$. In addition, MULON achieves the highest property coverage of 99% for $Conference_{de} \times Ekaw_{en}$ and $Iasted_{fr} \times Edas_{en}$. Some values for class coverage (CC) and property coverage (PC) are less than 100% due to the redundant resources that exist in the input ontologies. For instance, MULON found four redundant properties in $Conference_{de}$. Cross-lingual matching in MULON helps to remove redundant resources, therefore, the merged ontologies contain no redundant resources. MULON compacts ontologies by an average of 5.2%.

TABLE III
QUALITY ASSESSMENT RESULTS FOR THE MERGED ONTOLOGIES. BOLD ENTRIES ARE THE TOP SCORES.

Merging Ontology Pairs	RR	AR	IR	RB	IE	MP	RD	CC	PC	CT
$Conference_{de} \times Ekaw_{en}$	0.37	0.65	1.09	2.01	0.05	0.07	0.00	1.00	0.99	0.94
$ConfOfar \times Edas_{en}$	0.23	0.31	1.05	2.06	0.01	0.00	0.00	0.99	0.98	0.92
$Cmt_{ar} \times Ekaw_{en}$	0.40	0.79	1.17	2.02	0.02	0.08	0.00	1.00	0.96	0.95
$Iasted_{fr} \times Edas_{en}$	0.17	0.29	1.41	2.03	0.02	0.00	0.00	1.00	0.99	0.96
$Sigkdd_{fr} \times Ekaw_{en}$	0.28	0.43	1.12	1.99	0.04	0.11	0.00	1.00	0.98	0.97

VI. CONCLUSION AND FUTURE WORK

We present MULON, an approach for creating multilingual ontologies. The strength of our contribution lies in building such ontologies by merging ontologies in different natural languages using cross-lingual matching techniques. We show the usability of MULON by presenting three use cases in various domains; fiscal data, food, and scholarly communication. Indo and non-Indo-European languages ontologies are used in the merging process in order to illustrate the robustness of MULON. The results of the cross-lingual matching process are found promising compared to three state-of-the-art approaches. MULON not only produces multilingual ontologies but also automatically assesses their quality using a set of adapted quality metrics. This assessment emphasizes the validity of MULON. In conclusion, MULON established the first step towards a multilingual Semantic Web. In the future, we intend to further; 1) explore relationships between classes in the merged ontology that might occur after the merging process, e.g., a class exists in one ontology might be a subclass of another one in the other ontology, 2) resolve inconsistency that may exist in the merged ontology, 3) consider multiple labels for a resource, 4) consider individuals in the merging process, and 5) develop scalable approaches that can efficiently merge large-scale ontologies.

VII. ACKNOWLEDGMENTS

This work has been supported by the EU Horizon2020 projects LAMBDA (GA no. 809965) and PLATOON (GA no. 872592).

REFERENCES

- [1] Bhatt, B.: Unsupervised multilingual ontology learning. In: 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). pp. 1–7. IEEE
- [2] de Bruijn, J., Martin-Recuerda, F., Manov, D., Ehrig, M.: D4. 2.1 state-of-the-art-survey on ontology merging and aligning v1. SEKT Project deliverable D 4, 2–1 (2004)
- [3] Cheatham, M., Hitzler, P.: String similarity metrics for ontology alignment. In: ISWC. pp. 294–309. Springer (2013)
- [4] Cross, V.: Semantic similarity: a key to ontology alignment. In: OM@ ISWC. pp. 61–65 (2018)
- [5] Fahad, M.: Merging of axiomatic definitions of concepts in the complex owl ontologies. Artificial Intelligence Review **47**(2), 181–215 (2017)
- [6] Faria, D., Pesquita, C., Tervo, T., Couto, F.M., Cruz, I.F.: Aml and amlc results for oaei 2019. OM@ ISWC **2536**, 101–106 (2019)
- [7] Fathalla, S., Lange, C., Auer, S.: Eventskg: A 5-star dataset of top-ranked events in eight computer science communities. In: ESWC. pp. 427–442. Springer (2019)
- [8] Fathalla, S., Vahdati, S., Auer, S., Lange, C.: The scientific events ontology of the openresearch. org curation platform. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. pp. 2311–2313 (2019)
- [9] Fathalla, S., Vahdati, S., Lange, C., Auer, S.: Seo: A scientific events data model. In: ISWC. pp. 79–95. Springer (2019)
- [10] Florence, M.: Building a multilingual ontology for education domain using monto method. Computer Science and Information Technologies **1**(2), 47–53 (2020)
- [11] Fu, B., Brennan, R.: Cross-lingual ontology mapping and its use on the multilingual semantic web. MSW **571**, 13–20 (2010)
- [12] Fu, B., Brennan, R., O’Sullivan, D.: A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. Web Semantics: Science, Services and Agents on the World Wide Web **15**, 15–36 (2012)
- [13] Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. Web Semantics: Science, Services and Agents on the World Wide Web **11**, 63–71 (2012)
- [14] Helou, M.A., Palmonari, M., Jarrar, M.: Effectiveness of automatic translations for cross-lingual ontology mapping. Journal of Artificial Intelligence Research **55**, 165–208 (2016)
- [15] Ibrahim, S., Fathalla, S., Yazdi, H.S., Lehmann, J., Jabeen, H.: From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment. In: International Conference on Semantic Systems. pp. 215–230. Springer (2019)
- [16] Jabeen, H., Tahara, N., Lehmann, J.: Evochef: Show me what to cook! artificial evolution of culinary arts. In: International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar). pp. 156–172. Springer (2019)
- [17] Jimenez-Ruiz, E.: Logmap family participation in the oaei 2019. In: CEUR Workshop Proceedings. vol. 2536, pp. 160–163 (2019)
- [18] Joseph, M.F., Lourdasamy, R.: Ontology mediation method for building multilingual ontologies. International Journal of Information Technology **10**(1), 11–19 (2018)
- [19] Keet, M.: An introduction to ontology engineering, vol. 1.5. College Publications (2020)
- [20] Lehmann, J., Sejdin, G., Bühmann, L., Westphal, P., Stadler, C., Ermilov, I., Bin, S., Chakraborty, N., Saleem, M., Ngomo, A.C.N., et al.: Distributed semantic analytics using the sansa stack. In: International Semantic Web Conference. pp. 147–155. Springer (2017)
- [21] Mc Gurk, S., Abela, C., Debattista, J.: Towards ontology quality assessment. In: MEPDaW/LDQ@ ESWC. pp. 94–106 (2017)
- [22] Musyaffa, F.A.: Towards comparative analysis of open fiscal data. In: Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance. pp. 644–647. ACM (2018)
- [23] Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S.: Using of jaccard coefficient for keywords similarity. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. vol. 1 (2013)
- [24] Noy, N.F., Musen, M.A., et al.: Algorithm and tool for automated ontology merging and alignment. In: Proceedings of the 17th National Conference on Artificial Intelligence. Available as SMI technical report SMI-2000-0831. vol. 115. sn (2000)
- [25] Paulheim, H.: Wiktionary matcher. Ontology Matching p. 181 (2019)
- [26] Raunich, S., Rahm, E.: Atom: Automatic target-driven ontology merging. In: 2011 IEEE 27th international conference on data engineering. pp. 1276–1279. IEEE (2011)
- [27] Raunich, S., Rahm, E.: Towards a benchmark for ontology merging. In: OTM Confederated International Conferences” On the Move to Meaningful Internet Systems”. pp. 124–133. Springer (2012)
- [28] Tartir, S., Arpinar, I.B., Moore, M., Sheth, A.P., Aleman-Meza, B.: Ontoqa: Metric-based ontology quality analysis (2005)
- [29] Trojahn, C., Fu, B., Zamazal, O., Ritze, D.: State-of-the-art in multilingual and cross-lingual ontology matching. In: Towards the Multilingual Semantic Web, pp. 119–135. Springer (2014)

List of Figures

1.1	Language distribution of ontologies according to LOV.	2
1.2	The distribution of thesis contributions among the defined research questions.	7
2.1	Semantic Web Stack (known as Semantic Web Cake) [43].	12
2.2	An RDF graph represents statements a–d.	13
2.3	An illustration of enriching English ontology from German ontology and producing a multilingual enriched ontology in English and German.	17
4.1	The three-level hierarchical multi-agent model from the architecture of CLOE [37].	25

List of Tables

2.1	Statistics of ontologies in the MultiFarm benchmark.	18
3.1	Aggregated results for the state-of-the-art techniques in OAEI 2022 and 2023 compared to MoMatch. Bold entries are the top scores.	21

Acronyms

CLOE Cross-Lingual Ontology Enrichment. [6](#), [9](#), [24–26](#), [40](#), [141](#)

LLM Large Language Model. [21](#), [42](#)

LOV Linked Open Vocabularies. [1](#), [2](#), [141](#)

MoMatch Multilingual Ontology Matching. [6](#), [9](#), [20](#), [21](#), [37](#), [39](#), [143](#)

MULON MULtilingual Ontology mergiNg. [7](#), [9](#), [32](#), [33](#), [40](#), [41](#)

NLP Natural Language Processing. [2](#), [16](#), [20](#), [24](#), [28](#), [42](#)

OAEI Ontology Alignment Evaluation Initiative. [18](#), [20](#), [21](#), [29](#), [32](#), [143](#)

OECM Ontology Enrichment using Cross-lingual Matching. [6](#), [7](#), [9](#), [28](#), [29](#), [32](#), [40](#)

OWL Web Ontology Language. [2](#), [6](#), [12](#), [14](#), [15](#)

QASO Quality Assessment Suite for Ontologies. [8](#), [9](#), [36](#), [37](#), [41](#)

RDF Resource Description Framework. [12–14](#), [141](#)

RDFS Resource Description Framework Schema. [12](#), [14](#), [15](#)

RIF Rule Interchange Format. [12](#)

SPARQL Simple Protocol and RDF Query Language. [12](#)

SWRL Semantic Web Rule Language. [12](#)

URI Uniform Resource Identifier. [11–13](#)

URL Uniform Resource Locator. [13](#)

W3C World Wide Web Consortium. [11](#), [12](#)

XML Extensible Markup Language. [11](#)

List of Publications

This thesis builds upon the following publications:

- *Journal Articles (peer reviewed)*
 1. **Shimaa Ibrahim**, Said Fathalla, Jens Lehmann, and Hajira Jabeen. *Toward the multilingual Semantic Web: Multilingual ontology matching and assessment*. IEEE Access, vol. 11, pp. 8581-8599, 2023. DOI: <https://doi.org/10.1109/ACCESS.2023.3238871>
 2. Mohamed Ali, Said Fathalla, **Shimaa Ibrahim**, Mohamed Kholief, and Yasser F. Hassan. *CLOE: a cross-lingual ontology enrichment using multi-agent architecture*. Enterprise Information Systems 13, no. 7-8, pp. 1002-1022, 2019. DOI: <https://doi.org/10.1080/17517575.2019.1592232>
- *Conference Papers (peer reviewed)*
 3. **Shimaa Ibrahim**, Said Fathalla, Jens Lehmann, and Hajira Jabeen. *Multilingual ontology merging using cross-lingual matching*. In Proceedings of IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 113-120. IEEE, 2020. DOI: <https://doi.org/10.1109/WIIAT50758.2020.00020>
 4. **Shimaa Ibrahim**, Said Fathalla, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. *From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment*. In Proceedings of 15th International Conference on Semantic Systems (SEMANTiCS), Research & Innovation, pp. 215-230. Springer International Publishing, 2019. DOI: https://doi.org/10.1007/978-3-030-33220-4_16
 5. Mohamed Ali, Said Fathalla, **Shimaa Ibrahim**, Mohamed Kholief, and Yasser Hassan. *Cross-lingual ontology enrichment based on multi-agent architecture*. Procedia Computer Science 137 (SEMANTiCS), Research & Innovation, pp. 127-138, 2018. DOI: <https://doi.org/10.1016/j.procs.2018.09.013>
- *Poster Papers (peer reviewed)*
 6. **Shimaa Ibrahim**, Said Fathalla, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. *OEEM: A cross-lingual approach for ontology enrichment*. In European Semantic Web Conference (ESWC), pp. 100-104. Springer International Publishing, 2019. DOI: https://doi.org/10.1007/978-3-030-32327-1_20