International Journal of Semantic Computing
Vol. 17, No. 3 (2023) 1–19
©World Scientific Publishing Company
DOI: 10.1142/S1793351X23500034

World Scientific
www.worldscientific.com

Adaptive Semantic Matching in a Multilingual Context

Zhan Liu* and Nicole Glassey Balet†

Institute of Informatics
University of Applied Sciences and
Arts Western Switzerland (HES-SO Valais-Wallis)
Techno-Pole 3, 3960 Sierre, Switzerland
**zhan.liu@hevs.ch*
†nicole.glassey@hevs.ch

Received 12 March 2022
Revised 15 June 2022
Accepted 2 July 2022
Published

In an increasingly multilingual digital world, information management tools must support the simultaneous use and matching of multiple natural languages. A prerequisite for this is that the underlying database engine seamlessly processes multilingual data across languages. However, most natural language processing-based techniques have focused on developing monolingual matching algorithms, often ignoring context knowledge and external domain-based sources, which lead to incomplete and inaccurate matching results in a multilingual environment. The purpose of this study is to propose an adaptive semantic matching method with context knowledge and user involvement as two new dimensions for matching the semantically related entities ontologies. We present a comprehensive evaluation of our solution by applying it in a multilingual e-commerce platform case study, which performed well on matching accuracy.

Keywords: Semantic matching; context knowledge; user involvement; semantic annotation; multilingual.

1. Introduction

With the development of global multilingual digitization, a growing number of information management tools, such as web search engines, e-commerce portals and applications support the simultaneous use of multiple natural languages, the prerequisite for which is that the underlying database engine seamlessly processes multilingual data across languages. As one of the most important tasks in natural language processing, semantic matching is widely used in information retrieval, similarity detection, machine translation and automatic query answering. However, while its solutions commonly include syntactic matching and meaning matching, both of which focus on developing monolingual matching algorithms, they no longer meet the growing number of multilingual resources. Moreover, they often ignore

*Corresponding author.

context knowledge and external domain-based sources, leading to incomplete and inaccurate matching results in a multilingual environment.

In many metadata-intensive applications, a semantic match is a critical technique [1] that may include ontology integration, data warehousing, schema and mapping integration and e-commerce. The match operator function takes two graph-like structures and produces a map between the graph nodes that correspond semantically to each other [2]. Taking different approaches to the matching problem, researchers have divided them into two categories. The first is *syntactic matching*, which exploits the semantic information codified in graphs, either implicitly or explicitly, to match the node labels and searches for similarities between labels using syntax-driven techniques and syntactic similarity measures. These technical approaches compute syntactic similarity coefficients between labels to identify common substrings [3–6], similar soundex [7–9] or expand abbreviations [10–12]. Thus, in syntactic matching, semantics are not analyzed directly but their correspondences are searched for their syntactic features. The second category focuses on semantic relations by analyzing *meaning matching*, which calculates the mappings between schema elements by computing semantic relations [13]. Specifically, these approaches focus on concepts rather than labels, on the assumption that it is not sufficient to consider the meanings of the node labels alone but rather to address the graphs node positions as well [14–17].

Research has largely ignored the issue of user intervention, focusing instead on machine-based algorithms when handling matching tasks. While human knowledge and skills are expected to obtain more accurate results, they are rarely integrated into the matching system. Many research tools use log querying to enhance match candidate generation [18], or user clicks to track history for taxonomy matching [19]. However, because leaving the user to handle large amounts of data increases the tasks difficulty and takes time, we believe that user-friendly interfaces that offer better cognitive support [20] will increase productivity far more than improving accuracy and recall in matching algorithms.

In this study, we propose an adaptive semantic matching method to perform the matching task in a multilingual context. Our method uses data-linking techniques to identify and connect individuals that represent the same real-world object. To increase the performance of multilingual matching and improve the accuracy of the results, we developed two new dimensions to measure similarity in multilingual semantic matching: *Context knowledge* and *user involvement*. On the one hand, semantic matching can take advantage of context knowledge set of domain-based specific resources to provide extra information and increase matching recall. This feature improves the efficiency of the enrichment of the annotated corpora with multilingual specific knowledge sources, without having to expend energy on irrelevant resources. On the other hand, we designed a hybrid machine–human approach involving users who can contribute to the matching process without losing multiple results that might be relevant. In the case of a neutral result from the algorithm-based similarity calculation, we invited participants to make final judgments through user-friendly interfaces. The decisions were used to train our algorithm and improve its matching accuracy.

This paper is organized as follows. We begin by presenting the background and related works. In Sec. 3, we explain our approach both to context knowledge and user involvement within the framework, and we report our implementation in Sec. 4. Finally, we conclude with a summary of this work and present suggestions for future research.

2. Related Work

With the emergence of the Semantic Web and Linked Open Data, researchers have proposed several techniques and models to handle matching tasks in natural language: For instance, Word2Vec [21], LSA [22] and LDA [23]. However, most of these tools focus on developing monolingual matching algorithms that no longer satisfy the growing number of multilingual resources. For this reason, we need novel algorithms that can match ontologies and share more than one language. While automatic language translation is presently used to match the content in a multilingual environment and is designed to translate resources in other languages into English, several issues, such as regularities and culture [24], are difficult to address with automatic approaches. Therefore, automatic translation approaches suffer from ambiguity and inadequacy in specific domains.

In the context of multilingual resources, matching content has received some attention in recent years. Shvaiko and Euzenat [25] introduced three types of techniques for element- and structure-level matching: Syntactic, external and semantic techniques. Following several clearly stated algorithms, syntactic techniques interpret the input as a function of the sole structures: For example, the iterative fixpoint computation for matching graphs [26]. External techniques exploit domains auxiliary resources and common knowledge to interpret the input, which might be human or a thesaurus expressing the relationships between terms, such as Wordnet [27], Wikipedia [28] and BabelNet [29]. Finally, semantic techniques use some formal semantics e.g. model-theoretic semantics [30] to interpret the input and justify the result. With a semantic-based matching system, exact algorithms are complete because they guarantee the discovery of all possible mappings, while approximate algorithms tend to be incomplete.

Matching can be performed by discovering common context knowledge through ontology [31–33]. This context is represented as a set of multilingual resources, which have been annotated with concepts from an ontology. The resources are used to extract relationships between ontology entities and provide common anchors to a specific domain. Giunchiglia *et al.* [34] proposed an algorithm S-Match to automatically discover and use missing context knowledge during the matching process, which improved the quality of matching via iterations with a heuristic and enabled the newly discovered knowledge to be reused. A study by Sabou *et al.* [35] focused on finding the connecting paths between two concepts using all of the ontologies available on the semantic web as context knowledge. The rationale for their study was that more ontologies yielded better results. However, this can become a complex procedure and mass-generated information negatively affects the matching results. Moreover, Jain *et al.* [36] and Mascardi *et al.* [37], who used upper-level ontologies as context knowledge in the matching process, provided a common starting point for defining the context. Their studies supported broad semantic interoperability among a large number of domain-specific ontologies. Adding context knowledge yields new information and improves recall in matching tasks. This new information may, however, generate matching errors, which decreases precision.

To improve precision and recall results, user involvement approaches have been widely accepted in the development of active learning systems [38]. A review of the literature indicates that user involvement generally has positive effects, especially with fake news detection [39–42], health care services [43, 44] and smart cities [45, 46]. However, few studies have considered how to involve users in semantic matching. Dragisic *et al.* [47] and Falconer and Storey [48] proposed using graphic interfaces for mapping, allowing users to interact with the alignment in multiple ways and to clarify the consequences of accepting or rejecting a specific match. Other researchers focused on the development of technical tools. For example, Conroy *et al.* [49] proposed a tag mapping tool integrated into Firefox, a web browser, that enables users to participate in the matching process. Raffio *et al.* [50] provided a clip tool that allows users to explicitly specify structural transformations using a visual language, in addition to value coupling to be associated with the correspondences. A recent study by Da Silva *et al.* [51] involving domain experts in the matching process not only allows an expert to provide feedback about matchings between ontology entities but also dynamically evaluates the existing mapping results. The main challenge in the above research is to design ways to involve users to help in the matching process without burdening them with information overload or requiring them to expend a lot of energy. The best solution is to design the human–computer interaction naturally, using the users motivation to actively complete the matching tasks.

3. Framework and Approaches

The purpose of this study is to develop an adaptive method to perform the matching in the semantically related entities' ontologies using data-linking techniques. Data linking — the task of finding equivalent resources that represent the same real-world object [52] can be formalized as an operation that takes collections of data as input and produces a set of binary relations between their entities as output. Based on this mechanism, we explored the effectiveness of finding the correspondence between source and target concepts. This correspondence is used for various tasks from merging ontologies to answering queries, from data translation to data enrichment. In addition, we propose two new features for similarity measures context knowledge and user involvement to improve the accuracy of multilingual semantic matching.

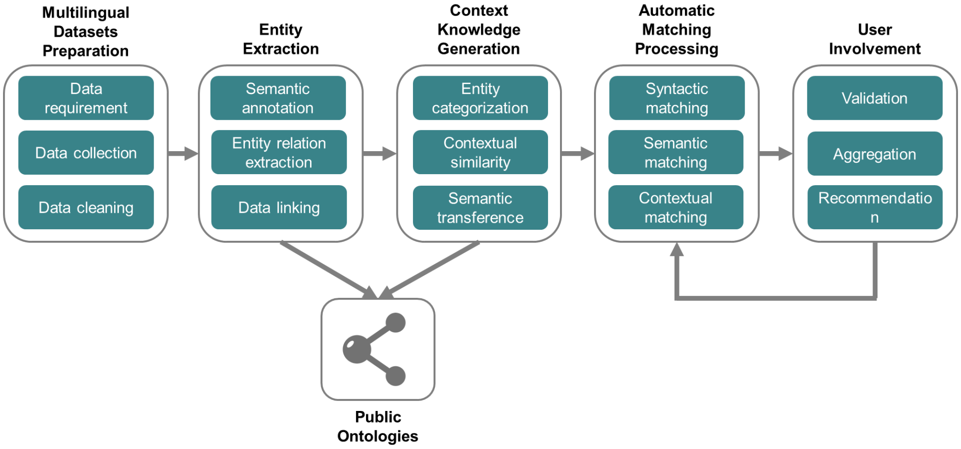


Fig. 1. The framework of the adaptive semantic matching system.

Figure 1 shows the global framework of the matching system, containing six stages with different resources and tools used during the design and implementation of the system.

3.1. *Multilingual datasets preparation*

In the first stage, we gathered and stored the source and target datasets into a relational database. These datasets are usually unstructured, multilingual and identified by the business requirements. In this study, the target dataset usually represents the information that needs to be matched, such as a product's detailed description, an introduction of an activity or profile information of a supplier. The source datasets are often composed of keywords or short phrases to match people, things or events from the system. Because the collected data are not all useful, we cleaned the data, removing incomplete, duplicated, stop words and error datasets. Moreover, we used the language recognition function to detect the language corresponding to each paragraph of text.

3.2. *Public ontologies*

Public ontology, an indispensable component in our process, is a knowledge base representing semantic relations between concepts in a network. Our ontologies refer to an encyclopedic dictionary providing concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations, such as DBpedia [53], Wordnet, [27] BabelNet [29] and Wikidata [54]. The ontology links words to semantic relationships that include hypernyms, hyponyms and synonyms. These are grouped into synsets with short definitions and multilanguage labels. The ontologies are accessible via the APIs or web browser, where they serve to process the named entity extraction and context knowledge generation.

3.3. *Entity extraction*

At this stage, the main tasks are to identify and annotate the entities as key elements from the text. Entity extraction based on semantic technologies disambiguates meaning and understands context, making unstructured data machine-readable and thus available for standard processing actions such as retrieving information, extracting facts and answering questions. For example, the entity orange might mean a fruit, a color or a telecommunications company. Our entity extraction technology can distinguish between all three, by considering every possible meaning and finally suggesting the one that best fits the context.

There are several excellent entity extraction tools, such as DBpedia Spotlight [55], Babelfy [56], NLTK [57] and SpaCy [58]. In this study, we applied DBpedia Spotlight to complete the named entity extraction work, not only because it is an excellent open-source tool with a strong research community but also because it provides comprehensive information in multiple languages based on DBpedia resources ontology, such as definitions, labels, categories and so on. Specifically, we implemented the following operations:

- Semantic annotation: Tagging text with relevant concepts, to identify URIs for resources mentioned within the text and described in the knowledge graph of DBpedia.
- Entity relation extraction: Reveals direct relationships and connections between different entities as well as complex relationships through inferred, indirect connections (e.g. synonyms, hypernyms and hyponyms).
- Data linking: Establish links between knowledge bases and entities, to infer entity similarities by identifying the number and direction of links from different entities.

3.4. *Context knowledge generation*

The term context in our project refers to the information requested that frames and scopes the required knowledge. Hence, context knowledge is a part of external knowledge, a set of domain-based specific resources used to characterize the current tasks situation. Semantic matching takes advantage of such context knowledge to provide extra domain-based information and increase matching recall. This matching is performed by identifying a common context knowledge and using it to generate a set of multilingual resources that have been annotated in the entity extraction stage with the concepts from domain-specific ontologies.

In our approach, we focus first on the task of entity categorization, which allows us to classify different entities into the same category based on the hierarchy from Wikipedia ontology. We used linked data techniques to connect categories coming from two entities, transform them into trees of senses for each concept, and compare the trees to discover hierarchical relations between such concepts. This effort provided similar entities with the same upper-level context to establish an effective

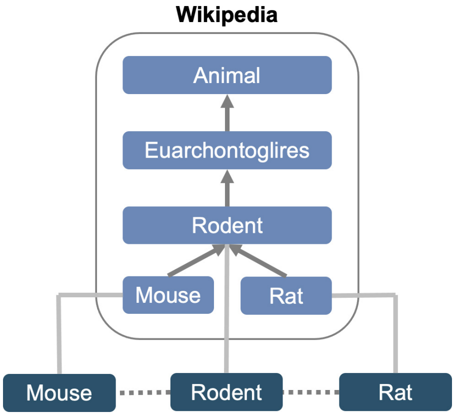


Fig. 2. Context knowledge generation with entity categorization.

knowledge base for subsequent matching tasks. Figure 2 illustrates an example of the generation of context knowledge through entity categorization. When considering the concepts mouse, rodent and rat, the use of a knowledge base such as Wikipedia ontology helps deduce that a mouse is less general than a rodent. In other words, mouse is a hyponym of rodent and situated at the same level as the concept rat. Hence, we can infer that mice, rodents and rats are related and have a matching similarity.

We then used contextual measure and disambiguation to compute the similarity between the source and target concepts based on their category hierarchy trees. Our contextual similarity ranges from 0 to 1, where 0 indicates no similarity and 1 indicates maximum similarity. We computed the similarity score between source concept and target concept to further determine whether these concepts should be matched. In an ideal case, the superclass of the source and target concepts come from the same root. In this situation, the alignment of concepts between source and target is further supported and hence is preferred. Otherwise, the alignment should be penalized. For example, the concept mouse might be aligned to the concept animal, which seems reasonable. However, if the mouse belongs to a superclass such as a computer input device, then the alignment should be penalized because its contextual similarity is low.

In our approach, semantic transference refers to assigning semantics to concepts that have not been described in the target knowledge resource by considering the available information of these concepts in the source knowledge resource. In a multilingual knowledge base such as Wikipedia, the labels describe a concept, which is expressed in different languages. The attribute helps to deal with the transference of semantics across languages, supplement the missing language information in the same context, and match the source concept and target concept between different languages. To illustrate: Assume that a German car accessory supplier opens a new store on an international industrial e-commerce platform. Technically, the e-commerce platform company requires the information sources to be integrated into its multilingual ontology. This task could be done by transferring the semantics of the concepts described in a source knowledge resource to a target knowledge resource, where the target knowledge is in English and the source knowledge is in another language, such as German in this example.

3.5. *Automatic matching processing*

This stage seeks to find the relationship between concepts, or rather to estimate the degree of their similarity by using the generated context knowledge. In this study, we focused on establishing a correspondence between words and texts through three main levels: Syntactic, semantic and contextual.

First, we used the Jaccard method [59] to calculate syntactic similarity, which assumes that the similarity between two texts is proportional to the number of identical words in them. At this level, we focused on comparing the labels related to the concepts. Since a concept has several labels and each label represents one language, this avoids matching results in a multilingual environment. For example, mouse and souris belong to the same concept but exist in English and French, respectively. The Jaccard similarity index is proportional to the number of common unique word roots in the two texts, and the Jaccard measure calculates the similarity value between two words by comparing the weight they are sharing. This weight of measure considers how many different words are associated with a given word in a text. By computing the similarity measure of all word pairs in the corpus, we extracted the list of the most similar words between source and target texts. This process was repeated to reach the best syntactic matching.

Semantic matching is used to complete syntactic methods since the latter are insufficient: It compares only textual concepts and neglects their semantic designations. Based on the results of entity extraction at the early stage, each annotated concept from the text was integrated with the synonyms, hypernyms and hyponyms. This additional information helped to reveal useful linked concepts and expand the scope of matching based on semantic similarity. On the one hand, semantic similarity focuses on the meaning and interpretation-based similarity between the two texts, the analysis used a sophisticated method for extracting meaning-based values between source and target concepts. On the other hand, however, semantic similarity estimated the distance between concepts by using an ontology and offers greater accuracy while keeping irrelevant information to a minimum.

The results were obtained from the semantic matching as input for the initial contextual method. First, we classified the relevant concepts from the same text into different categories. By using the semantic similarity between the source and target categories, we obtained the common topics, which we used to define the context of the content. In other words, in the absence of matching concepts, we could still match the relevant content because of a common context. Second, we used a confidence score that was defined by using the factors such as topical pertinence and contextual ambiguity, to evaluate the trustworthiness of each annotation of the concept. This confidence score is computed by the relative concepts of content in the same context. The high relevant concepts were assigned a higher confidence score. For example, an article about animals contains an introduction to the growth of the mouse. During the matching process, through the analysis of the context information, our system automatically determined that the mouse is a rodent, not a pointing device. Finally, with the enrichment of the matching results from different specific contexts, the results provide useful information to adaptively construct domain-specific knowledge bases for public sharing as linked open data.

3.6. *User involvement*

User involvement relates to the validation of the mapping suggestions generated by the machine-based matching system. In this stage, we focused on developing a hybrid approach to enable natural user involvement and obtain higher quality gains in the semantic matching task. Specifically, our approach contains three sessions: Validation, aggregation and recommendation. The validation session allows a domain expert or normal user to validate a subset of matching suggestions. We analyzed the threshold defined by our system for deciding whether the automatic semantic matching is acceptable. This threshold presents the semantic coherence between source and target concepts by using the semantic similarity coefficients score from 0.0 to 1.0. In the event our matching algorithm produces a neutral result (e.g. less than 0.5), we invited the user to make the final validation.

We applied a majority voting strategy to aggregate the results of the validation from the users. The final result of matching was obtained with the highest number of votes. For example, two users selected rat as the best match for mouse, while only one user selected rodent. The system then recorded the concept rat at the top of the matching list of the mouse concept. Furthermore, our approach supported the use of aggregation results in the computation of mapping suggestions and the recommendation of which matching strategies to use, thereby introducing the domain experts and the users knowledge in the matching generation and recommendation processes. Our hybrid machine–human approach constructed an active learning-based [60] decision-making model that decided whether the results from the automatic matching element are sufficiently good or need a human in the loop.

4. Implementation

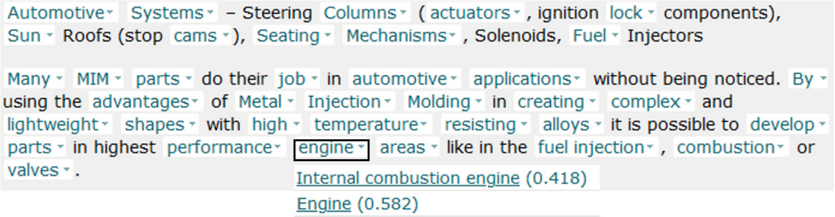
We implemented a prototype based on the framework described above. This prototype was integrated into an online multilingual e-commerce platform containing four steps: content processing, data enrichment with context knowledge, semantic matching and user validation.

4.1. *Content processing*

The content processing function is designed to prepare the original and target datasets before matching. In our case study, data collection involves the source and target datasets separately. The target datasets contain the descriptions of products and the profile of suppliers on an e-commerce platform. Depending on the location of the suppliers, five languages can be used in their descriptions: English, French, German, Italian and Chinese. The source datasets include the users matching request, which can be a sentence or keyword in any of the five languages. Once the data were collected and stored in our database, we focused on three steps in the data cleaning process:

(1) Removed unwanted characters, a primary step in the process of text data cleaning, in which we removed all the tags for the text from HTML sources, for example, HTML format entities, non-alphabets and any other characters that might not be a part of the language. We used regular expression methods to filter out most of the unwanted texts.
(2) Removed stop words. In our case study, keywords are more important than general terms, so removing stop words might increase matching precision. In this step, we used predetermined lists of multilingual keywords from the textual analysis tool quanteda [61] to complete removal.
(3) Stemming and lemmatization. Here, the goal was to reduce inflectional forms of a word to a common base or root form through the natural language toolkit NLTK [57]. For example, the words playing, plays and played share a common root in play, and cars and cars share a common root in car.

The entity extraction task was completed by using DBpedia Spotlight, an open-source tool for automatically annotating concepts from the text. DBpedia Spotlight provides an API to map unstructured information sources to the linked open data cloud through DBpedia ontology. DBpedia Spotlight is a multilingual tool, which identifies URIs for resources mentioned within different textual languages. Moreover, this tool allows configuring the annotations to the specific needs through the DBpedia ontology and quality measures such as prominence, topical pertinence and disambiguation confidence [55]. The confidence parameter was used in our study to measure the precision of concept annotation. The value of the confidence score ranges from 0 to 1, which considers factors such as topical pertinence and contextual ambiguity. During the annotation process, setting a high confidence threshold instructs DBpedia Spotlight to avoid incorrect annotations but risks losing some correct ones.



Automotive Systems – Steering Columns (actuators , ignition lock components), Sun Roofs (stop cams), Seating Mechanisms , Solenoids, Fuel Injectors

Many MIM parts do their job in automotive applications without being noticed. By using the advantages of Metal Injection Molding in creating complex and lightweight shapes with high temperature resisting alloys it is possible to develop parts in highest performance engine areas like in the fuel injection , combustion or valves .

Internal combustion engine (0.418)
Engine (0.582)

Fig. 3. Concepts annotation with confidence parameter.

In some cases where the system is ambiguous, we confirmed the results of annotation through user involvement. Figure 3 shows an example of concept annotation using DBpedia Spotlight. The term engine occurs in two concepts with very close confidence scores, DBpedia: Internal_combustion_engine (0.418) and DBpedia: Engine_motor (0.582). In this situation, we would submit it to an expert for validation to increase the precision of the semantic annotation.

4.2. *Data enrichment with context knowledge*

The goal of data enrichment is to append and enhance the annotated concepts with a relevant context obtained from additional sources in DBpedia ontology, such as synonyms, categories and translations [62, 63]. In this step, we built a context knowledge base in the form of a triple store with the links of each concept to combine the information from external sources. Our triple store was based on the instance of Stardog, a knowledge graph to store the RDF data and model. Stardog is implemented with W3C standards and supports the RDF graph data model and SPARQL query language, which enabled us to enrich the ontology with the extra datasets and relations between the concepts more effectively. The process of data enrichment provides not only additional information for annotated concepts but is also used as the source for the context knowledge ontology construction. Specifically, the

Table 1. Data enrichment properties and examples.

Property name	Description	Example
hasLookupURI	URL of the matching DBPedia concept	https://dbpedia.org/resource/Engine
hasSimilarityScore	The topical relevance of the annotated resource for the given context is measured by the similarity score returned by the content processing step	0.582
hasAnchorText	The word(s) in the text matched to a concept of DBpedia	engine
hasPosStart	The start position of the anchor text in the whole text	411
hasPosEnd	The end position of the anchor text in the whole text	417
type	Type of concept, information coming from the DBpedia ontology	owl:Thing
label	Label of concept, with a language tag	Engine @en; Moteur @fr
abstract	The description of the concept	An engine or motor is a machine designed to convert one form of energy into mechanical energy. @en
subject	The related categories from DBpedia, which is represented using keywords, key phrases or classification codes	Engine technology @en; Moteur @fr; Motor @de
wikiPageRedirects	The synonyms of a concept	Air-breathing_engine; Motor
sameAs	Mapping of concepts with equivalent meaning	http://fr.dbpedia.org/page/Moteur http://de.dbpedia.org/page/Motor

following sources were added in our ontology to enrich the annotated concept as shown in Table 1.

Once the required data are obtained, we stored it in the format of Turtle syntax in the triple store. In principle, a Turtle document allows writing an RDF graph in a compact textual form. An RDF graph is made up of triples consisting of a subject, predicate and object. This structure can be used for semantic matching through SPARQL queries. Table 2 introduces an example of a Turtle document of an annotated concept with additional sources.

Table 2. Example of data enrichment in a turtle document.

<http://datasemlab.ch/smms/100001#34> nlp:hasLookupURI
<http://dbpedia.org/resource/Engine>.
<http://datasemlab.ch/smms/100001#34> nlp:hasAnchorText
"engine"@en.
<http://datasemlab.ch/smms/100001#34> nlp:hasPosStart 411.
<http://datasemlab.ch/smms/100001#34> nlp:hasPosEnd 417.
<http://datasemlab.ch/smms/100001#34 > nlp:hasSimilarityScore 0.582.
<http://dbpedia.org/resource/Engine>
<http://www.w3.org/2000/01/rdf-schema#label> "Engine"@en,
"Moteur"@fr;
<http://purl.org/dc/terms/subject>
<http://dbpedia.org/resource/Category:Engine technology>,
<http://dbpedia.org/resource/Category:Engines>;
<http://dbpedia.org/ontology/abstract> "An engine or motor
is a machine designed
to convert one form of energy into mechanical energy." @en,
"Un moteur est un appareil transformant une nergie
quelconque en nergie canique." @fr.

4.3. *Semantic matching*

The principles of semantic matching can be considered data interlinking, which is particularly beneficial in cross-language matching because the matching resources do not need to use the same natural language. We began to determine the relation between the source and target entities by composing the relations in the path (i.e. sequence of relations) connecting them. The composition method may be functional, order-based or relational. For instance, the relation =, standing for equals to; the relation < standing for subClass by and the relation > standing for superClass to.

Our algorithm first proposed the matching results that hold an equal relation between source and target entities. To avoid matching concepts with the same words but different meanings, the results were based on the computation of disambiguation. In this case, multiple languages that refer to the same concept could be identified and matched in the same way. Second, we used synonyms in the matching task, which helped to identify the nodes in the two structures that semantically relate to one another. For instance, the system identified that a product containing the word

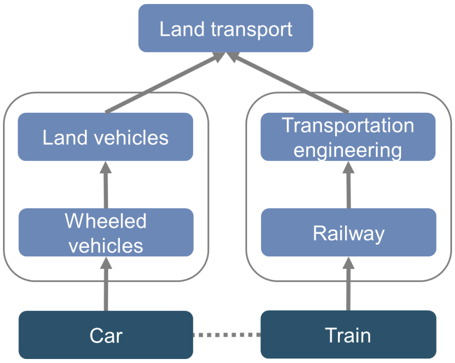
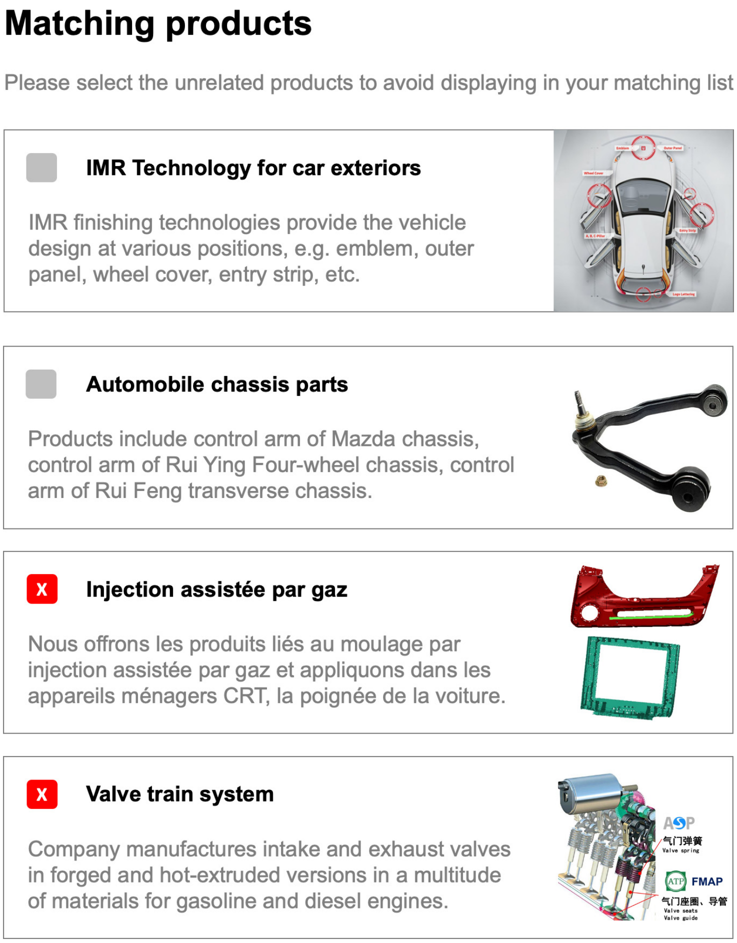


Fig. 4. An example of the paths between the related concepts.



Matching products

Please select the unrelated products to avoid displaying in your matching list

□ **IMR Technology for car exteriors**

IMR finishing technologies provide the vehicle design at various positions, e.g. emblem, outer panel, wheel cover, entry strip, etc.

□ **Automobile chassis parts**

Products include control arm of Mazda chassis, control arm of Rui Ying Four-wheel chassis, control arm of Rui Feng transverse chassis.

☒ **Injection assistée par gaz**

Nous offrons les produits liés au moulage par injection assistée par gaz et appliquons dans les appareils ménagers CRT, la poignée de la voiture.

☒ **Valve train system**

Company manufactures intake and exhaust valves in forged and hot-extruded versions in a multitude of materials for gasoline and diesel engines.

Fig. 5. Matching suggestions and user validation.

car is semantically equivalent to another product automobile in English and voiture in French since they are synonyms. However, because too many synonyms render the matching results biased, we used only synonyms with the best similarity scores [64]. Third, we operated concept mapping through general and specific relations. This operation was based on the extra information from our context knowledge database. The purpose was to find paths between concepts that are not directly related but belong to the same category or subcategory. Figure 4 shows an example of paths found by our system. The concepts car and train can be matched because they belong to a common category, land transport.

4.4. *User validation*

The validation session allows the user to validate matching suggestions. The suggested results come from the semantic matching computation in the multilingual environment, which contains equivalent, similar and context-based relations between source and target concepts. Through the user interface, the system presents matching suggestions (Fig. 5), with product information available on the e-commerce platform, such as product name, short description and photo. The user can reject one or more matching suggestions by selecting the checkbox. Furthermore, the user can review previous decisions and propose different options.

The user's validations are stored in the matching decisions database. We applied the majority voting method to measure the final output of matching results with the highest votes. Matching decisions, including both accepted and rejected results, can be used in the future computation processing stage, as well as to improve our context knowledge information. A validation session is an optional user operation, and with additional user involvement, the matching systems accuracy and efficacy should improve.

5. Conclusion and Future Work

In this paper, we present an adaptive semantic matching framework used to perform matching tasks in a multilingual environment. The development of our framework not only solves the problem of inaccurate matching between multilingual resources but can also be used as a fundamental technique in areas such as resource discovery, data integration, contextual search and schema and ontology merging. We highlight how the Semantic Web and data-linking technologies can help to improve ontology-based natural language processing using rules and queries. To this, we added two new dimensions to the concept matching process, to improve the accuracy of the matching algorithms and eliminate language barriers. Specifically, the generation of context knowledge ontology allows enriching the annotated concepts with additional sources. This approach is well suited to matching multilingual resources since it does not consider the linguistic manifestation of concepts as part of the content. Moreover, our framework permits natural user involvement, an enduring challenge in semantic matching. We show the utility of this approach in our implementation for providing precise matching results through human knowledge and experience.

Looking ahead, we will continue to develop and evaluate matching-based computation strategies and user involvement strategies. Strategies that reuse validation results to reduce the matching processing time or guide computation are especially interesting. In addition, we will integrate debugging strategies into the context knowledge generation process, to improve the results from the entity extraction session. It would also be useful to develop configurable and customizable user interfaces, which the users themselves might help to improve and find solutions that best fit their needs and preferences.

Acknowledgments

The work described in this paper is supported by the Hasler Foundation Switzerland under grant number 21029.

References

[1] S. Melnik, E. Rahm and P. A. Bernstein, Developing metadata-intensive applications with Rondo, *J. Web Semantics* **1** (2003) 47–74.

[2] F. Giunchiglia, M. Yatskevich and P. Shvaiko, Semantic Matching: Algorithms and Implementation, in *Journal on Data Semantics IX*, eds. S. Spaccapietra, P. Atzeni, F. Fages, M.-S. Hacid, M. Kifer, J. Mylopoulos, B. Pernici, P. Shvaiko, J. Trujillo and I. Zaihrayeu, PLecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2007), pp. 1–38.

[3] J. P. McCrae and P. Buitelaar, Linking datasets using semantic textual similarity, *Cybern. Inf. Technol.* **18** (2018) 109–123.

[4] S. D. Cardoso, M. Da Silveira, Y.-C. Lin, V. Christen, E. Rahm, C. Reynaud-Delaître and C. Pruski, Combining Semantic and Lexical Measures to Evaluate Medical Terms Similarity, in *Data Integration in the Life Sciences*, eds. S. Auer and M.-E. Vidal, Lecture Notes in Computer Science (Springer International Publishing, Cham, 2019), pp. 17–32.

[5] H.-H. Do and E. Rahm, COMA — A system for flexible combination of schema matching approaches, in *VLDB '02: Proc. 28th Int. Conf. Very Large Databases*, eds. P. A. Bernstein, Y. E. Ioannidis, R. Ramakrishnan and D. Papadias (Morgan Kaufmann, San Francisco, January 2002), pp. 610–621.

[6] J. Madhavan, P. A. Bernstein and E. Rahm, Generic Schema Matching with Cupid, in *Proc. 27th Int. Conf. Very Large Data Bases VLDB '01* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, September 2001), pp. 49–58.

[7] T. Kant, S. K. Shrivastava, N. K. Tiwary and N. Parween, SoundexHindi: A Phonetic Matching Algorithm for Hindi Written in English, SSRN Scholarly Paper ID 3579322, Social Science Research Network, Rochester, NY (2020).

[8] A. Gal, Why is schema matching tough and what can we do about it? *ACM SIGMOD Record* **35** (2006) 2–5.

[9] D. Holmes and M. McCabe, Improving precision and recall for Soundex retrieval, in *Proc. Int. Conf. Information Technology: Coding and Computing* April 2002, pp. 22–26.

[10] A. Ojugo and A. Eboka, Memetic algorithm for short messaging service spam filter using text normalization and semantic approach, *Int. J. Informatics Commun. Technol.* **9**(1) (2020) 9–18.

[11] E. Chondrogiannis, V. Andronikou, T. Varvarigou and E. Karanastasis, Semantically-enabled context-aware abbreviations expansion in the clinical domain, in *Proc. 9th Int. Conf. Bioinformatics and Biomedical Technology ICBBT'17* (Association for Computing Machinery, New York, NY, USA, 2017), pp. 89–96.

[12] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius and M. Duneld, Synonym extraction and abbreviation expansion with ensembles of semantic spaces, *J. Biomed. Semantics* **5** (2014) 6.

[13] F. Giunchiglia and P. Shvaiko, Semantic matching, *Knowl. Eng. Rev.* **18** (2003) 265–280.

[14] S. Kim, I. Kang and N. Kwak, Semantic sentence matching with densely-connected recurrent and co-attentive information, in *Proc. AAAI Conf. Artificial Intelligence*, Vol. 33 (2019), pp. 6586–6593.

[15] Y. Huang, Q. Wu, C. Song and L. Wang, Learning semantic concepts and order for image and sentence matching, *2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (2018), pp. 6163–6171.

[16] A. Bordes, X. Glorot, J. Weston and Y. Bengio, Joint learning of words and meaning representations for open-text semantic parsing, in *Proc. Fifteenth Int. Conf. Artificial Intelligence and Statistics* (2012), pp. 127–135.

[17] S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang and X. Cheng, A deep architecture for semantic matching with multiple positional sentence representations, in *Proc. AAAI Conf. Artificial Intelligence*, Vol. 30 (2016), pp. 2835–2841.

[18] H. Elmeleegy, M. Ouzzani and A. Elmagarmid, Usage-based schema matching, *2008 IEEE 24th Int. Conf. Data Eng.*, April 2008, pp. 20–29.

[19] A. Nandi and P. A. Bernstein, HAMSTER: Using search clicklogs for schema and taxonomy matching, in *Proc. VLDB Endowment*, Vol. 2, August 2009, pp. 181–192.

[20] A. F. Newell, A. Carmichael, P. Gregor, N. Alm and A. Waller, Information technology for cognitive support, in *The Human-Computer Interaction Handbook*, 2nd edn. (CRC Press, 2007), pp. 464–481.

[21] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, in *Proc. 26th Int. Conf. Neural Information Processing Systems — Volume 2 NIPS'13* (Curran Associates Inc., Red Hook, NY, USA, 2013), pp. 3111–3119.

[22] S. T. Dumais, Latent semantic analysis, *Annu. Rev. Inf. Sci. Technol.* **38**(1) 188–230 (2004).

[23] D. Falush, M. Stephens and J. K. Pritchard, Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies, *Genetics* **164** (2003) 1567–1587.

[24] C. Schäffner, Running before walking?: Designing a translation programme at undergraduate level, in *Developing Translation Competence*, eds. C. Schäffner and B. Adab, Benjamins Translation Library (John Benjamins, Amsterdam (NL), 2000), pp. 143–156.

[25] P. Shvaiko and J. Euzenat, A survey of schema-based matching approaches, in *Journal on Data Semantics IV*, ed. S. Spaccapietra, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2005), pp. 146–171.

[26] S. Melnik, H. Garcia-Molina and E. Rahm, Similarity flooding: A versatile graph matching algorithm and its application to schema matching, in *Proc. 18th Int. Conf. Data Engineering*, February 2002, pp. 117–128.

[27] G. A. Miller, WordNet: A lexical database for English, *Commun. ACM* **38** (1995) 39–41.

[28] E. Gabrilovich and S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis, in *Proc. 20th Int. Joint Conf. Artifical Intelligence IJCAI'07* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007), pp. 1606–1611.

[29] R. Navigli and S. P. Ponzetto, BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artif. Intell.* **193** (2012) 217–250.

[30] B. B. Jackson, Model-theoretic semantics as model-based science, *Synthese* **199** (2021) 3061–3081.

[31] A. Locoro, J. David and J. Euzenat, Context-based matching: Design of a flexible framework and experiment, *J. Data Semantics* **3** (2014) 25–46.

[32] F. Lin, J. Butters, K. Sandkuhl and F. Ciravegna, Context-based ontology matching: Concept and application cases, *2010 10th IEEE Int. Conf. Computer and Information Technology*, June 2010, pp. 1292–1298.

[33] F. Duchateau, Z. Bellahsene and M. Roche, A context-based measure for discovering approximate semantic matching between schema elements, Report, RR-06053 (2006), p. 11.

[34] F. Giunchiglia, P. Shvaiko and M. Yatskevich, Discovering missing background knowledge in ontology matching, in *Proc. 2006 Conf. ECAI 2006: 17th European Conf. Artificial Intelligence August 29 – September 1, 2006* (IOS Press, NLD, 2006), pp. 382–386.

[35] M. Sabou, M. d'Aquin and E. Motta, Exploring the semantic web as background knowledge for ontology matching, in *Journal on Data Semantics XI*, eds. S. Spaccapietra, J. Z. Pan, P. Thiran, T. Halpin, S. Staab, V. Svatek, P. Shvaiko and J. Roddick, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2008), pp. 156–190.

[36] P. Jain, P. Z. Yeh, K. Verma, R. G. Vasquez, M. Damova, P. Hitzler and A. P. Sheth, Contextual ontology alignment of LOD with an upper ontology: A case study with proton, in *The Semantic Web: Research and Applications*, eds. G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer and J. Pan, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2011), pp. 80–92.

[37] V. Mascardi, A. Locoro and P. Rosso, Automatic ontology matching via upper ontologies: A systematic evaluation, *IEEE Trans. Knowl. Data Eng.* **22** (2010) 609–623.

[38] B. Settles, Active learning literature survey, Technical Report, University of Wisconsin-Madison Department of Computer Sciences (2009).

[39] S. Shabani and M. Sokhn, Hybrid machine-crowd approach for fake news detection, *2018 IEEE 4th Int. Conf. Collaboration and Internet Computing (CIC)*, October 2018, pp. 299–306.

[40] Z. Liu, S. Shabani, X. Yu, M. Sokhn and N. Glassey Balet, A collaborative intelligence approach to fighting COVID-19 false news: A Chinese case, in *Human Centred Intelligent Systems: Proc. KES-HCIS 2022 Conf.* (Springer, 2022), pp. 3–12.

[41] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant and A. Krause, Fake news detection in social networks via crowd signals, in *Companion Proc. Web Conf. 2018 WWW'18* (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018), pp. 517–524.

[42] J. Kim, B. Tabibian, A. Oh, B. Schölkopf and M. Gomez-Rodriguez, Leveraging the crowd to detect and reduce the spread of fake news and misinformation, in *Proc. Eleventh ACM Int. Conf. Web Search and Data Mining WSDM '18* (Association for Computing Machinery, New York, NY, USA, February 2018), pp. 324–332.

[43] B. Fischer, A. Peine and B. Östlund, The importance of user involvement: A systematic review of involving older users in technology design, *The Gerontologist* **60** (2020) e513–e523.

[44] D. de Beurs, I. van Bruinessen, J. Noordman, R. Friele and S. van Dulmen, Active involvement of end users when developing web-based mental health interventions, *Front. Psychiatry* **8** (2017) 72.

[45] H. Kopackova and J. Komarkova, Participatory technologies in smart cities: What citizens want and how to ask them, *Telematics Informatics* **47** (2020) 101325.

[46] M. Himanen, The significance of user involvement in smart buildings within smart cities, in *Designing, Developing, and Facilitating Smart Cities* (Springer, Cham, 2017), pp. 265–314.

[47] Z. Dragisic, V. Ivanova, P. Lambrix, D. Faria, E. Jiménez-Ruiz and C. Pesquita, User validation in ontology alignment, *The Semantic Web - ISWC 2016. ISWC 2016* (2016), pp. 200–217.

[48] S. M. Falconer and M.-A. Storey, A cognitive support framework for ontology mapping, *The Semantic Web. ISWC 2007, ASWC 2007*, Lecture Notes in Computer Science, Vol. 4825 (2007), pp. 114–127.

[49] C. Conroy, R. Brennan, D. O'Sullivan and D. Lewis, User evaluation study of a tagging approach to semantic mapping, in *The Semantic Web: Research and Applications*, eds. L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou and E. Simperl, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2009), pp. 623–637.

[50] A. Raffio, D. Braga, S. Ceri, P. Papotti and M. A. Hernandez, Clip: A visual language for explicit schema mappings, *2008 IEEE 24th Int. Conf. Data Engineering*, April 2008, pp. 30–39.

[51] J. D. Silva, K. Revoredo, F. Baião and J. Euzenat, Alin: Improving interactive ontology matching by interactively revising mapping suggestions, *Knowl. Eng. Rev.* **35** (2020) e1.

[52] A. Ferrara, A. Nikolov and F. Scharffe, Data linking for the semantic web, *Int. J. Semantic Web Inf. Syst.* **7** (2011) 46–76.

[53] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann, DBpedia — A crystallization point for the Web of Data, *J. Web Semantics* **7** (2009) 154–165.

[54] C. Müller-Birn, B. Karran, J. Lehmann and M. Luczak-Rösch, Peer-production system or collaborative ontology engineering effort: What is Wikidata? in *Proc. 11th Int. Symp. Open Collaboration OpenSym '15* (Association for Computing Machinery, New York, NY, USA, 2015), pp. 1–10.

[55] P. N. Mendes, M. Jakob, A. García-Silva and C. Bizer, DBpedia spotlight: Shedding light on the web of documents, in *Proc. 7th Int. Conf. Semantic Systems I-Semantics '11* (Association for Computing Machinery, New York, NY, USA, 2011), pp. 1–8.

[56] A. Moro, A. Raganato and R. Navigli, Entity Linking meets Word Sense Disambiguation: A unified approach, *Trans. Assoc. Comput. Linguist.* **2** (2014) 231–244.

[57] E. Loper and S. Bird, NLTK: The Natural Language Toolkit, in *Proc. ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics — Volume 1 ETMTNLP'02* (Association for Computational Linguistics, USA, 2002), pp. 63–70.

[58] M. Honnibal, Introducing SpaCy, Explosion (2015), https://explosion.ai/blog/introducing-spacy.

[59] G. Grefenstette, *Explorations in Automatic Thesaurus Discovery* (Springer Science & Business Media, 2012).

[60] B. Settles, *Active Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning (Morgan & Claypool Publishers, 2012).

[61] K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller and A. Matsuo, quanteda: An R package for the quantitative analysis of textual data, *J. Open Source Softw.* **3** (2018) 774.

[62] Z. Liu, J. Shan, N. Glassey Balet and G. Fang, Semantic social media analysis of chinese tourists in Switzerland, *Inf. Technol. Tourism* **17**(2) (2017) 183–202.

[63] Y. Gutiérrez, S. Vázquez and A. Montoyo, A semantic framework for textual data enrichment, *Expert Syst. Appl.* **57** (2016) 248–269.

[64] P. Ataee, How to Build a Fast “Most-Similar Words” Method in SpaCy, *Towards Data Science* (2020).