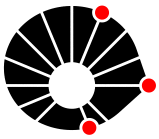

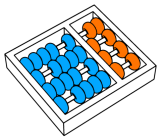

Investigating neighbour concepts for cross-lingual ontology alignment

Gabriel Oliveira dos Santos *Juliana Medeiros Destro*
Julio Cesar dos Reis

Technical Report - IC-19-01 - Relatório Técnico
February - 2019 - Fevereiro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Investigating neighbour concepts for cross-lingual ontology alignment

Gabriel Oliveira dos Santos Juliana Medeiros Destro
Julio Cesar dos Reis
Institute of Computing, University of Campinas, São Paulo, Brazil

February 2019

Abstract

Cross-lingual ontology alignments play a key role for the semantic integration of data described in different languages. The task of automatic cross-lingual ontology matching requires exploring similarities measures. Such measures compute the degree of relatedness between two given terms from ontology's concepts. Although the literature has extensively studied similarity measures for monolingual ontology alignments, the use of similarity measures for the creation of cross-lingual ontology mappings still requires further research. In this work, we define an algorithm for automatic cross-lingual ontology matching based on the analysis of neighbour concepts to improve the effectiveness of the composed similarity approach, a technique to calculate the degree of similarity between concept contents in different languages. Experimental results with OAEI datasets indicate that our novel approach including neighbour concepts for mapping identification has a good effectiveness.

1 Introduction

There is a growing number of ontologies described in different natural languages. The mappings among different ontologies are relevant for the integration of heterogeneous data sources to facilitate the exchange of information between systems. Although automatic monolingual ontology matching has been extensively investigated [30], cross-lingual ontology matching still demands further investigations aiming to automatically identify correspondences between ontologies described in different languages.

In this context, accurate automatic methods are essential for ensuring the quality of the generated mappings. Current ontologies have highly grown in size and

differences between the used alphabets hamper the use of simple string comparison techniques. Similarity measures play a key role to obtain well-defined ontology mappings because they allow calculating the level of lexical and semantic similarity between concepts [24]. Cross-lingual ontology matching approaches in the literature have not yet thoroughly investigated the influence of similarity calculation neither have they analyzed the influence of neighbour concepts in the matching process.

In this work, we propose an original cross-lingual ontology alignment technique based on the analysis of neighbour concepts relying on composed similarity measure [6] by combining both syntactic and semantic similarity techniques. Syntactic similarity computes a score calculated based on string analysis (extracted from labels of concepts), whereas the semantic similarity is computed taking into account background knowledge such as synonyms and the context in which terms appear (*e.g.*, use of external dictionaries and vocabularies). Our investigation explores a *Weighted Overlap* measure [25] relying on the neutral-domain semantic network *BabelNet* [20] and computes a weighted mean of semantic and syntactic similarities.

The proposed technique takes into account those concepts immediately related to a given concept from a source ontology (the neighbours), and those ones also directly linked to a concept from a target ontology. On this basis, the method finds the highest value of similarities among these concepts, in this work we name such value as neighbourhood similarity. The neighbourhood similarity is used to improve the correctness of mappings, so we combine it with the composed similarity. However, the combination is only applied if the initial value of composed similarity is in a doubtful band, that is, between a default and minimum threshold (parameters).

We carried out a series of experiments to empirically investigate the quality of mappings generated by our technique. Our experiments explored conference-domain ontologies, described in English, Portuguese and Spanish, from the *MultiFarm* [1] dataset [18]. *MultiFarm* turns available curated mappings established among multilanguage ontologies. This dataset has been extensively used to assess cross-lingual ontology matching methods. The obtained results indicate that syntactic and semantic similarities may be given the same importance in order to obtain a good accuracy. Our experiments suggest that the threshold and language in which the ontologies are described play an important role in the quality of generated alignments.

The remaining part of this report is organised as follows: Section 2 describes the related work; Section 3 formalizes the fundamental concepts of our proposal; Section 4 reports on our proposed technique; Section 5 describes the experimental results whereas Section 6 discusses our findings; Section 7 provides the conclusion remarks.

[1] https://www.irit.fr/recherches/MELODI/multifarm

2 Related Work

There has been a number of investigations on specific aspects of cross-language for ontology matching. Meilicke *et al.* [2] studied the effectiveness of a set of matching systems based on a dataset defined to evaluate ontology alignment. Their results indicated the difficulties of traditional ontology matching algorithms for carrying out multilingual ontology alignment. Trojahn *et al.* [29] described an extensive survey of matching systems and strategies for accomplishing multilingual and cross-lingual ontology matching.

Several approaches have explored the translation effects and the use of a third language in cross-language ontology alignment. In particular, Fu *et al.* [8] analyzed the impact of automatic translations on multilingual ontology alignment, highlighting the translation's relevance for achieving adequate matching quality. Spohr *et al.* [27] studied the translation of concept labels to a third language for matching two ontologies described in different languages.

Ontology alignment techniques have considered the use of similarity methods, which aim to calculate the degree of relatedness between concepts exploring different sources (e.g. dictionary, thesauri, etc.). Stoutenburg [28] argued that the use of ontologies combined with linguistic resources as background knowledge might enhance ontology matching processes. This appears as an alternative to syntactic similarity measures relying only on string comparison to determinate the similarity value.

The use of multiple similarity measures for the ontology alignment task has been hardly investigated in the literature. Nguyen and Conrad [23] proposed an ontology matching method based on the combination of lexical-based, structure-based, and semantic-based techniques. After obtaining the structural correspondences among the concepts, the method explores a semantic similarity based on WordNet dictionary and the results are combined. Their approach was evaluated with monolingual ontology alignments. Further investigations are necessary to understand whether a combination and use of semantic similarity can be relevant for cross-lingual ontology alignment.

Experimental studies have analyzed the influence of syntactic and semantic similarity methods and the structure of terms denoting concepts in ontologies in the context of cross-language alignment [3]. These studies highlight the potential influence of similarity measures. Though, there is no proposal for combining the techniques for cross-language ontology alignment and experimental evaluations against adequate datasets.

We have thoroughly surveyed existing cross-lingual ontology matching methods. We analyzed five recent approaches: *CroLOM* [15], *SOCOM++* [9], *YAM++* [22], KEPLER [14] and LogMap [12].

The proposal of *CroLOM* is based on natural languages processing techniques (such as lemmatization, stopwords elimination and stemming) to normalize labels

extracted from ontologies. These entities are translated into English, as a pivot language, and the technique computes a Cartesian product among the concepts that compose the ontologies. They apply semantic and syntactic similarity measures in a hybrid way to identify potential mappings. The syntactic similarity is calculated from the *Levenshtein distance* [17], whereas the semantic similarity considers the category of words. At this stage, an initial filter is applied to select candidate correspondences containing the maximum similarity value. Then, a second filter is applied to identify the correspondences that contain similarity value upper than a given threshold.

The *SOCOM++* approach considers several setups with different parameters. In contrast to *CroLOM*, it translates concept labels of the source ontology to the same language of the target ontology, thus no pivot languages are considered. Afterwards, both ontologies are described in the same language and monolingual matching methods are applied. In this process, the context of a given concept is analyzed considering all immediately neighbour concepts to improve the quality of the obtained alignment. This approach was designed to support user's influence on adjustments in the translation of the selected labels, and thus users can analyze the resulting mappings and propose changes.

In YAM++ , concept labels of both ontologies (source and target) are translated into the English language. The concepts are filtered in a stage named candidate filtering. In this stage, heuristic filters are applied to selected candidate correspondences, reducing the search space. In the following stage, the method analyses the neighbourhood of previously selected concepts to discover as many as possible high accurate mappings. Finally, the selected mappings go through a process of semantic verification [21], in which those correspondences considered inconsistent are removed.

The KEPLER's approach relies on divide and conquer proposal, first split up the ontology into small blocks, maximizing the relationship inside the block, and minimizing the relationship between the blocks themselves. On the following step, it translates the ontologies to English as the pivot language, and uses the indexing strategy to reduce the searching space. It considers Candidate Mappings Identification, which queries documents in a vector space that contains a set of ontological entities and their synonyms obtained via WordNet for each Ontology. Finally, the algorithm filters the candidate mappings by using two filters: the first filter eliminates the redundancy between these candidates by eliminating possible duplicates; The second filter eliminates false positives candidates.

LogMap considers a Lexical indexation, which is an inverted index used to store the lexical information contained. It exploits ontology modularisation techniques to reduce the size of problem. The relevant modules in the input ontologies together with (a subset of) the candidate mappings are encoded in LogMap using a Horn propositional representation. This approach extends Dowling-Gallier's algorithm to track all mappings that may be involved in the unsatisfiability of a class and performs a greedy local repair; that is, it repairs unsatisfiabilities on-the-fly and only

looks for the first available repair plan. It considers a Semantic Indexation, which allows to answer many entailment queries as an index lookup operation over the input ontologies and the mappings computed. The semantic index complements the use of the propositional encoding to detect and repair unsatisfiable classes in the input ontologies.

Our approach differentiates from the above-mentioned proposals because we combine both semantic and syntactic similarities by computing the composed similarity. In addition, we define a similarity value to the neighbourhood with the aim of improving the correctness of the generated mappings.

3 Fundamental Concepts

This section formalizes the fundamental concepts in this investigation.

3.1 Ontologies

In the semantic web context [11], ontologies define a common vocabulary in a domain. They are used for semantic representation in computational systems, describing the definition of concepts and the relationship among them.

Definição 3.1 (Ontology) *An ontology $\mathcal{O}$ describes a domain in terms of concepts, attributes and relationships [10]. Formally, an ontology $\mathcal{O} = (\mathcal{C}_\mathcal{O}, \mathcal{R}, \mathcal{A}_\mathcal{O})$ consists in a set of concepts $\mathcal{C}_\mathcal{O}$ interrelated by a set of directed relations $\mathcal{R}$. Each concept $c \in \mathcal{C}_\mathcal{O}$ has an unique identifier and it is associated to a set of attributes $\mathcal{A}_\mathcal{O}(c) = \{a_1, a_2, ..., a_p\}$. Each relation $r(c_1, c_2) \in \mathcal{R}$ can be described as a tuple $(c_1, c_2, r(c_1, c_2))$, where $r(c_1, c_2)$ is a function returning the type of relationship between the concepts (c_1, c_2) (e.g., "$\equiv$", "$\sqsubseteq$", etc.). The symbols "$\equiv$" and "$\sqsubseteq$" represent relationships "equivalence" and "is-a", respectively. Furthermore, the relationships can express domain-related relations. For instance, considering the biomedical domain, the concepts c_1 : "Insulin" and c_2 : "Diabetes" may be related by the following function: $r(c_1, c_2) =$ "Treats".*

Definição 3.2 (Neighbour Concepts) *We define neighbour concepts of a given concept $c \in O$ as all the concepts which is directly related to it. Formally, the neighbourhood of c is the set $nbh = \{cpt | cpt \in O \wedge dist(c, cpt) = 1\}$, where $dist(c, cpt)$ is the distance (in terms of the number of edges) between 'c' and 'cpt'.*

Figure 1 presents an illustrative example of neighbour concepts. The neighbourhood of "*Pancreas*" is composed by "*Endocrine System*", "*Digestive System*", "*Insulin*" and "*Glucagon*", because all of them are directly related to "*Pancreas*".

Once the distance between "*Kidney*" and "*Pancreas*" is equal to two, they are not considered neighbour concept of "*Pancreas*".

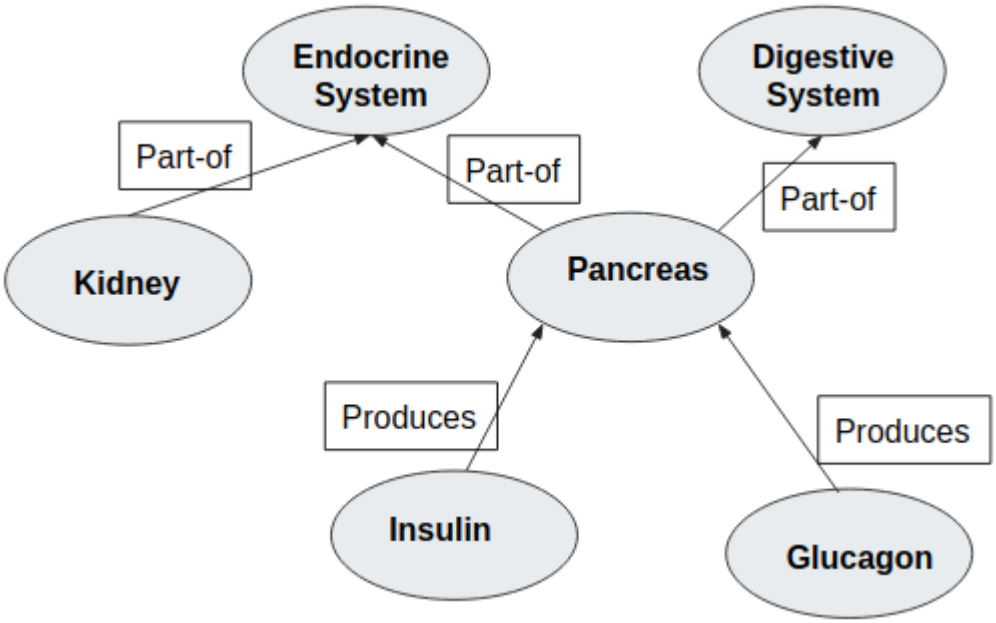

Figure 1: Example of neighbourhood

3.2 Cross-Lingual Ontology Alignment

Ontology alignment refers to the process of identifying mappings among concepts from different ontologies. Formally:

Definição 3.3 (*Ontology Alignment*) *The alignment stands for the process of identifying the relation between concepts from different ontologies. For concepts* $c_i \in \mathcal{C}_{\mathcal{O}_1}$ *and* $c_j \in \mathcal{C}_{\mathcal{O}_2}$*, the alignment is expressed by the tuple* $m_{ci \to cj} = (c_i, c_j, Rel(c_i, c_j))$*, where* $Rel(c_i, c_j) \in \mathcal{R}$ *is the relationship between* c_i *and* c_j*.*

This research addresses the problem of cross-lingual ontology alignment. This problem is formally defined as follows:

Definição 3.4 (Cross-lingual ontology alignment) *Let* O_X *and* O_Y *be ontologies described in different natural language "X" and "Y", respectively; and* $c_i \in \mathcal{C}_{\mathcal{O}_X}$ *and* $c_j \in \mathcal{C}_{\mathcal{O}_Y}$*. The problem relies on automatically identifying the adequate set of tuples* $m_{ci \to cj} = (c_i, c_j, Rel(c_i, c_j))$*, where* $Rel(c_i, c_j) \in \mathcal{R}$ *is the relationship between these concepts. For instance, considering the concepts* $c_1 \in \mathcal{C}_{\mathcal{O}pt}$ *and* $c_2 \in \mathcal{C}_{\mathcal{O}en}$*, from ontologies described in Portuguese and English, respectively, such that* $c_1 =$ *"Cabeça" and* $c_2 =$ *"Head", the alignment between these concepts is* $m_{c1 \to c2} = (c_1, c_2, \equiv)$*.*

Definição 3.5 (Mappings) *The final result of the alignment process is a set containing the mappings found between the concepts from two given ontologies. Formally, the mapping between the ontologies O_1 and O_2 is given by each element of $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}(\lambda) = \{m_{ci \to cj} | c_i \in \mathcal{C}_{\mathcal{O}_1} \wedge c_j \in \mathcal{C}_{\mathcal{O}_2} \wedge sim(c_i, c_j) \geq \lambda\}$, where "$\lambda$" is the threshold (minimum value to consider similar) and $sim(c_i, c_j)$ is the similarity between c_i and c_j (cf. Subsection 3.3).*

3.3 Similarity Measures

Definição 3.6 (Similarity between concepts) *Given two concepts c_i and c_j from an ontology (or from different ontologies), the similarity value between them is defined as the maximum similarity value among the attributes of c_i and c_j. Formally:*

$$sim(c_i, c_j) = \arg\max \ sim(a_{ix}, a_{jy}) \tag{1}$$

where $sim(a_{ix}, a_{jy})$ is the relatedness degree between the pair of attributes a_{ix} and a_{jy} from c_i and c_j, respectively. The similarity may be calculated in different linguistic levels, from string-based methods to semantic techniques [4].

Levenshtein Distance is an algorithm that computes a syntactic similarity, which can be understood as the minimum number of single-character editions (insertions, deletions or substitutions) needed to change a string s into s'. We has chosen this algorithm to compute the syntactic similarity because Levenshtein Distance has been well-studied and has been extensively used to spelling correction, then it is considered a good alternative to syntactic analysis [19].

Semantic Similarity Measure. Semantic similarity between concepts is a metric to evaluate how similar two given concepts are, considering their meanings in a certain context. For instance, the words "lead" and "iron" are much more similar considering the metal context than "lead" and "leader". On the other hand, when we consider the organizational context "lead" and "leader" may be more similar than "lead" and "iron".

There are algorithms to calculate semantic similarity. Usually, these algorithms explore an external resource such as vocabulary, dictionaries and thesauri, which help to compute the similarity. In this work, we use Weighted Overlap applied to NASARI vectors, together with the neutral-domain semantic network *BabelNet* [20].

This choice allows us to understand the influence of semantic similarity in both neutral-domain and specific-domain. NASARI helps us to compute the similarity value in multilingual contexts because it uses vectors based on "*synsets*" (set of synonyms) used by *Babelnet* [1]. The vectors are created in two steps: first, for a given concept, a set of Wikipedia pages where the concept is mentioned are collected. The second step consist in process the collected contextual information using a statistical measure (lexical specificity [16]), aiming at finding the most relevant words and

synsets appearing in the contextual information and assigning to each one of them a weight (based on the statistical measure). Each of the these words and synsets are used as dimensions in the vector-based representation.

Table 1 shows the semantic vector-based representation of two *Babel synsets* (*i.e.*, the identification used in BabelNet to represent a given meaning of a word and all the synonyms expressing that meaning in a range of different languages). On each row of the NASARI vector table (exemplified by two rows in Table 1), the first column is the *Babel synsets* ID and the second column is the textual description of the synset (*e.g.*, the synsetID bn:00000009n represents the synset "*100 (number)*"). The vector dimensions are described from column three onwards, and are represented by a *Babel synset* ID and its correspondent weight (*e.g.*, vector dimension in column *synset1_weight1*, where bn:00058285n is the dimension and 332.33 is the weight). Vectors are truncated to the non-zero dimensions only (*i.e.*, all dimensions present weight above zero). Because vectors present *Babel synset* as their dimensions, they are comparable across languages.

Babel SynsetId	Wikipedia PageTitle	synset1_weight1	...	synset$n$_weight$n$
bn:00000009n	100 (number)	bn:00058285n_332.33	...	bn:00031261n_9.35
bn:00000010n	1000 (number)	bn:00058285n_347.11	...	bn:00024261n_2.11

Table 1: Example of word-based vector representation.

NASARI leverages *Weighted Overlap (WO)* method applied to the semantic vectors representations [13] to calculate the semantic similarity between two elements el_1 and el_2 (*cf.* Equation (2)):

$$sem(el_1, el_2) = WO(v_1, v_2) \tag{2}$$

Weighted Overlap calculates the similarity between the meanings of two given lexical items. Formally:

$$WO(v_1, v_2) = \frac{\sum_{i=1}^{|S|}(r_i^1 + r_i^2)^{-1}}{\sum_{i=1}^{|S|}(2i)^{-1}} \tag{3}$$

At the equation 3, S refers to the set of overlapping dimensions between the two vectors (*i.e.*, dimensions appearing on both vectors; in the example in table 1, dimension bn:00058285n under column *synset1_weight1*). The r_q^j is the rank of dimension q in the vector v_j. Note that the weight is not used on WO equation; it is used only for ranking (*i.e.*, sorting) the dimensions.

Given two word-based vector representation v_1 and v_2 of the string elements el_1 and el_2, respectively. The string elements can be any string looked up for a correspondence on *BabelNet* and matched with a vector in NASARI (*e.g.*,in Table 1, el_1 can be

represented by the string "*100 (number)*", and el_2 by the string "*1000 (number)*"). [25].

Definição 3.7 (Composed Similarity) *We define the composed similarity combining syntactic and semantic measures. Let sem(c_1, c_2) (Equation* (2)*) be the semantic similarity and syn(c_1, c_2) the syntactic one (Equation* (1)*) between the concepts c_1 and c_2, respectively. Formally:*

$$simC(c_1, c_2) = \frac{\alpha syn(c_1, c_2) + \beta sem(c_1, c_2)}{\alpha + \beta} \tag{4}$$

where α and β are constants.

Note that both semantic and syntactic similarities are a particular case of the composed similarity, when α and β are equal to zero, respectively.

We explore the composed similarity together with Neighbourhood Analysis (*cf.* Section 4) in our cross-lingual ontology alignment technique.

4 Neighbourhood Analysis for Cross-Lingual Ontology Alignment

We propose an algorithm which combines composed similarity with neighbourhood analysis for cross-lingual ontology alignment. Algorithm 1 creates a cross-lingual alignment between two distinct ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ expressed in different natural languages. The algorithm considers the following input arguments:

- Input ontologies $\mathcal{O}_1, \mathcal{O}_2$
- $\lambda \in (0, 1]$ - default threshold
- $min_\lambda \in [0, \lambda)$ - minimum threshold
- α - Syntactic weight
- β - Semantic weight
- *pivot* - The pivot language

The algorithm starts with mapping set $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \leftarrow \emptyset$ (line 1) and the similarity variables with zero. It calculates the cartesian product from the set of concepts $\mathcal{C}_{\mathcal{O}_1}$ and $\mathcal{C}_{\mathcal{O}_2}$ from ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively. The algorithm computes the similarity value based on a syntactic measure (line 9). It considers automatic translation

of labels of concepts c_1 and c_2 to a pivot language providing (w_1, w_2). The syntactic similarity is calculated relying on the strings (w_1, w_2). The semantic similarity value is also computed. To this end, for each tuple $(c_1, ling_{c_1}, c_2, ling_{c_2})$, composed by the concepts c_1 and c_2, and their respective natural languages $ling_{c_1}$ and $ling_{c_2}$, the algorithm calls the function $babelnet(c_1, ling_{c_1}, c_2, ling_{c_2})$. Such function is based on *synsets* used by *Babelnet* and by the NASARI semantic vectors (*cf.* Section 3.3) to calculate the *Weighted Overlap* (Equation (3)).

On this basis, the algorithm calculates the weighted average, assigning weights previously defined α and β to the syntactic syn_{sim} and semantic sem_{sim} similarities, respectively. It results on the composed similarity $composed_{sim}$ (lines 14 and 15). At this stage, if the composed similarity is good enough, that is, it is greater than the defined threshold (similarity value higher than λ), then it considers the mapping as correct (line 20). On the other hand, if the $composed_{sim}$ composed similarity value is greater than or equal to min_λ and it is less than or equal to λ, thus assuming that the mapping is doubtful, the algorithm verifies the neighbourhood of the involved concepts in order to ensure the quality of mappings.

Algorithm 1: Cross-lingual ontology alignment based on composed similarity measure considering neighbourhood analysis

Require: $\mathcal{O}_1, \mathcal{O}_2, \lambda, min_\lambda \in [0,1], \alpha, \beta, pivot$

1: $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \leftarrow \emptyset$ {Initialize the mapping as an empty set}
2: $syn_{sim} \leftarrow 0$
3: $sem_{sim} \leftarrow 0$
4: $nbh_{sim} \leftarrow 0$
5: **for all** $c_1 \in \mathcal{C}_{\mathcal{O}_1}$ **do**
6: **for all** $c_2 \in \mathcal{C}_{\mathcal{O}_2}$ **do**
7: **if** $\alpha > 0$ **then**
8: $w_1 \leftarrow translate(c_1, \text{pivot})$, $w_2 \leftarrow translate(c_2, \text{pivot})$
9: $syn_{sim} \leftarrow syntactic_{sim}(w_1, w_2)$
10: **end if**
11: **if** $\beta > 0$ **then**
12: $sem_{sim} \leftarrow semantic_{sim}(c_1, ling_{c_1}, c_2, ling_{c_2})$
13: **end if**
14: $composed_{sim} = \frac{\alpha sim_{sin} + \beta sem_{sim}}{\alpha + \beta}$ {Compute the composed similarity value}
15: $similarity \leftarrow composed_{sim}$
 {If the mapping is doubtful, then analyze the neighbourhood of concepts}
16: **if** $min_\lambda \leq composed_{sim} \leq \lambda$ **then**
17: $nbh_{sem} \leftarrow neighbourhood_{sim}(c_1, c_2)$ {Algorithm 2}
18: $similarity \leftarrow composed_{sim}^{(1-nbh_{sem})}$
19: **end if**
20: **if** $similarity \geq \lambda$ **then**
21: $m_{c1 \to c2} \leftarrow (c_1, c_2, \equiv)$
22: $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \leftarrow \mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \cup \{m_{c1 \to c2}\}$
23: **end if**
24: **end for**
25: **end for**
26: **return** $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}$ {Generated mappings}

The neighbourhood analysis computes the maximum similarity among the neighbours of the considered concepts (source and target). Algorithm 2 concerns computing a similarity among the neighbours of the concepts c_s and c_t (source and target concepts given as input). First, it extracts the neighbourhood of concepts c_s and c_t to nbh_s and nbh_t, respectively (line 1 and 2 in Algorithm 2). The algorithm aims to find the pair of neighbour concepts (one from the source ontology and the other one from the target one) with the maximum similarity value.

Algorithm 2: Neighbourhood analysis

Require: c_s, c_t {Given the concepts c_s and c_t from the source and target ontologies}
{Extract the neighbourhood of concepts c_s and c_t to nbh_s and nbh_t}

```
1: nbh_s ← neighbourhood(c_s)
2: nbh_t ← neighbourhood(c_t)
3: maxSim ← 0
4: for all n_1 ∈ nbh_s do
5:    for all n_2 ∈ nbh_t do
6:       sim ← similarity(n_1, n_2)
7:       if sim > maxSim then
8:          maxSim ← sim
9:       end if
10:   end for
11: end for
12: return maxSim
```

Figure 2 presents an example to illustrate the technique of neighbourhood analysis. We consider two ontologies[2], on the left side the source ontology is described in Portuguese language and on the right the target ontology is described in English language.

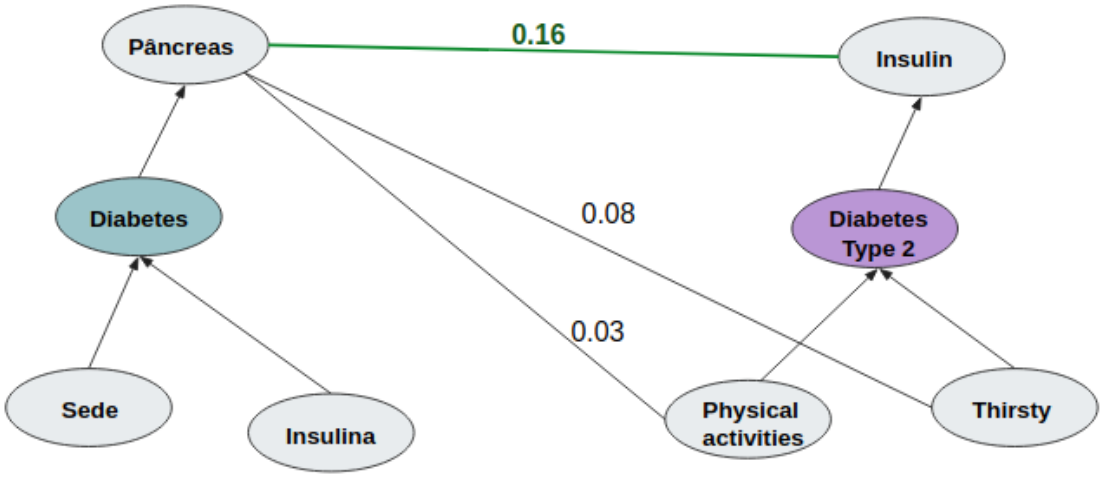

Figure 2: Finding the maximum similarity among the neighbour concepts

In the example of Figure 2, the concepts "*Diabetes*" and "*Diabetes Type* 2" are

[2]These ontologies are considered only for the purpose of this example. They were not extracted from real-world ontologies.

under analysis. Let's suppose the initial similarity score between them was given by 0.80 and the minimum threshold and default threshold equal 0.33 and 0.95, respectively. Thus they must go through the neighbourhood analysis and their neighbours are evaluated to find the maximum similarity value, once the similarity measure is less than the default threshold and greater than minimum threshold. In this illustration, the maximum similarity is related to "*Pâncreas*" is 0.16, which links temporarily "*Pâncreas*" to "*Insule*"

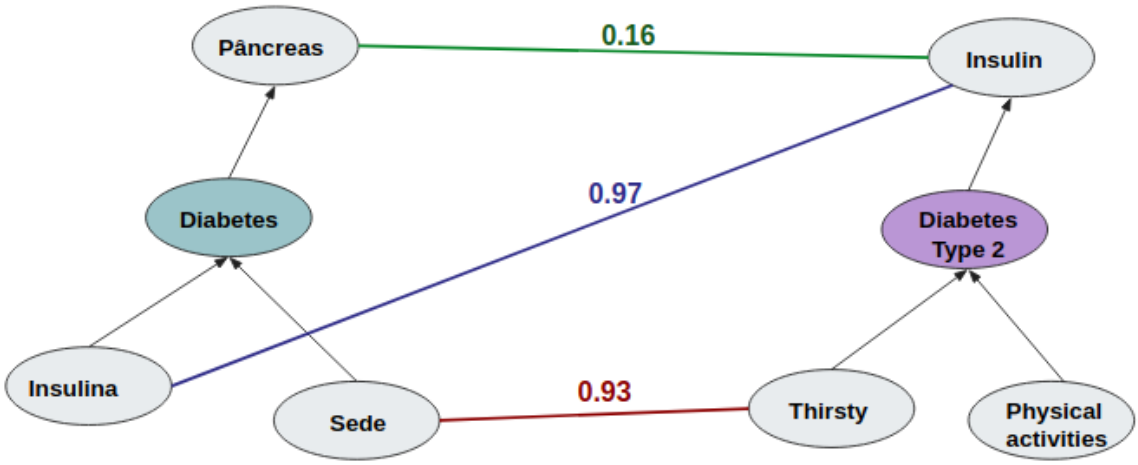

Figure 3: Found the pair of concepts with maximum similarity for each concept from the original ontology

After the analysis of all neighbour concepts, we find the maximum similarity values among them, 0.16 between "*Pâncreas*" and "*Insule*"; 0.97 between "*Insulina*" and "*Insule*" and 0.93 between "*Sede*" and "*Thirsty*" (*cf.* Figure 3). At the end of the process, the neighbourhood similarity is computed as the greatest similarity score. In this example, it is equal to 0.97.

The neighbourhood similarity value returned by Algorithm 2 updates the similarity value considering $composed_{sim}^{(1-nbh_{sem})}$ in Algorithm 1 (line 18). Therefore, after the neighbourhood analysis process the final similarity value equal to $0.80^{(1-0.97)} = 0.80^{0.03} = 0.99$, thus these concepts under analysis become similar, because the final similarity is greater than the default threshold. Note, when the neighbourhood similarity is high, close to 1, the resulting similarity also approaches to 1, therefore it is likely to surpass the default threshold, and then be considered as correct. This method assumes as correct mappings which the neighbour concepts are quite similar even if the pair of concepts under analysis itself is not so similar. Finally, the Algorithm 1 verifies whether the similarity value computed is greater than or equals to a beforehand input threshold λ. If such condition is satisfied, the algorithm inserts the mapping $(c_1, c_2, \equiv)$ into the set $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}$ indicating a cross-lingual correspondence

between the concepts.

5 Experiments

This evaluation aims to analyze the quality of mappings generated by our proposed technique which considers the structure of ontologies to align ontologies. We conducted a series of two experiments relying on a set of curated mappings manually established between ontologies described in different languages. In particular, we assessed our proposal for ontologies originally described in English and Spanish mapped into another ontology described in the Portuguese language.

5.1 Datasets and Procedure

Our experiments are based on ontologies related to the conference domain from the *MultiFarm dataset* version released in 2015. The considered dataset is used in the *OAEI* (Ontology Alignment Evaluation Initiative)[3]. The *MultiFarm* [18] benchmark is a comprehensive dataset for cross-lingual ontology matching. The original *MultiFarm* dataset is composed by a set of 7 ontologies of the Conference domain[4], translated into 8 languages[5] and the corresponding cross-lingual alignments between them. This dataset is based on the *OntoFarm* dataset, which has been successfully used for several years in the OAEI Conference track. The cross-lingual alignments of this dataset were manually curated and may be used as a reference to assess algorithms that build automatic cross-lingual ontology mappings. There are alignments founded in a set of 45 pairs of languages. For instance, the pair PT-ES refers to the case involving Portuguese and Spanish. For each pair, there are 25 alignments covering the mentioned ontologies.

Our experiments built cross-lingual ontology mappings by using English as a pivot language. We initially choose English because a great number of resources such as background knowledge is available. The results obtained by executing the Algorithm 1 in different scenarios were compared with the reference mappings from the *MultiFarm dataset*, then metrics of precision, recall and f-measure [26] were calculated.

We executed the algorithm 1 setting different weights and default thresholds, but considering the minimum threshold just equals to 0.33. We used the reference mappings between the ontologies described in English and Spanish mapped into those concepts in the Portuguese Language. The weights followed the fractions $\{\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}\}$, considering the constraint $\alpha + \beta = 1$. We present results varying

[3]http://oaei.ontologymatching.org.

[4]Cmt, Conference, ConfOf, Iasted, Sigkdd

[5](English) – Chinese (cn), Czech (cz), Dutch (nl), French (fr), German (de), Portuguese (pt), Russian (ru), Spanish (es)

the threshold level to comprehend its role in the studied scenarios. We vary the threshold in $\{0.66, 0.75, 0.80, 0.95\}$, which were selected based on the fractions $\{\frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{19}{20}\}$. The threshold 0.95 was chosen to evaluate the behaviour of the algorithm in contrast to the high level of threshold.

5.2 Experimental Results

Figure 4 and Figure 5 present the obtained results of our experiment. The horizontal axis of the charts represents the syntactic weight; *i.e.*, the weight assigned to syntactic similarity and thus the semantic weight refers to the difference of 1 of the syntactic weight; the vertical axis presents the achieved scores for precision, recall and f-measure.

Figure 4 (mapping English into Portuguese) indicates that results increase in terms of precision as the threshold increases. On the other hand, the recall drops slightly in most cases. When we take into account higher thresholds as 0.95, the recall decreases significantly, by presenting a fall of about 0.11 in comparing the syntactic weights 0.20 and 0.80. The experiments demonstrated that better results are obtained setting threshold close to 0.80. Our analysis shows that the best results were obtained by thresholds equal to 0.75 and 0.80.

Although the results presented in Figure 4 indicate an increase in the precision as the syntactic weight grows, our findings in analyzing Figure 5 (mapping Spanish into Portuguese) show a contrary behaviour. The precision decreases with syntactic weight increase, whereas the recall rises slightly. Furthermore, the recall of the matching from Spanish to Portuguese is greater than the one of matching English to Portuguese. This finding may indicate an influence of the language in the matching process.

A thorough comparison between the two configurations evaluated indicates that when further weight is given to the syntactic similarity measure, and consequently considering less importance to the semantic, the f-measure falls gradually. This reveals that less accurate mappings are generated. Such behaviour may mean that both syntactic and semantic similarities are relevant to produce cross-lingual mappings with quality, because type of similarity complements the other. The best results of f-measure in all scenarios are around the syntactic weight 0.50, i.e., when both similarity measures present the same importance in the composed approach.

6 Discussion

Cross-lingual ontology matching relies on several different approaches to obtain mappings that interrelate ontologies described in distinct languages. Cross-lingual ontology matching requires adequate techniques relying on similarity measures to overcome the matching task barrier. Ontological structure and similarity measures might help

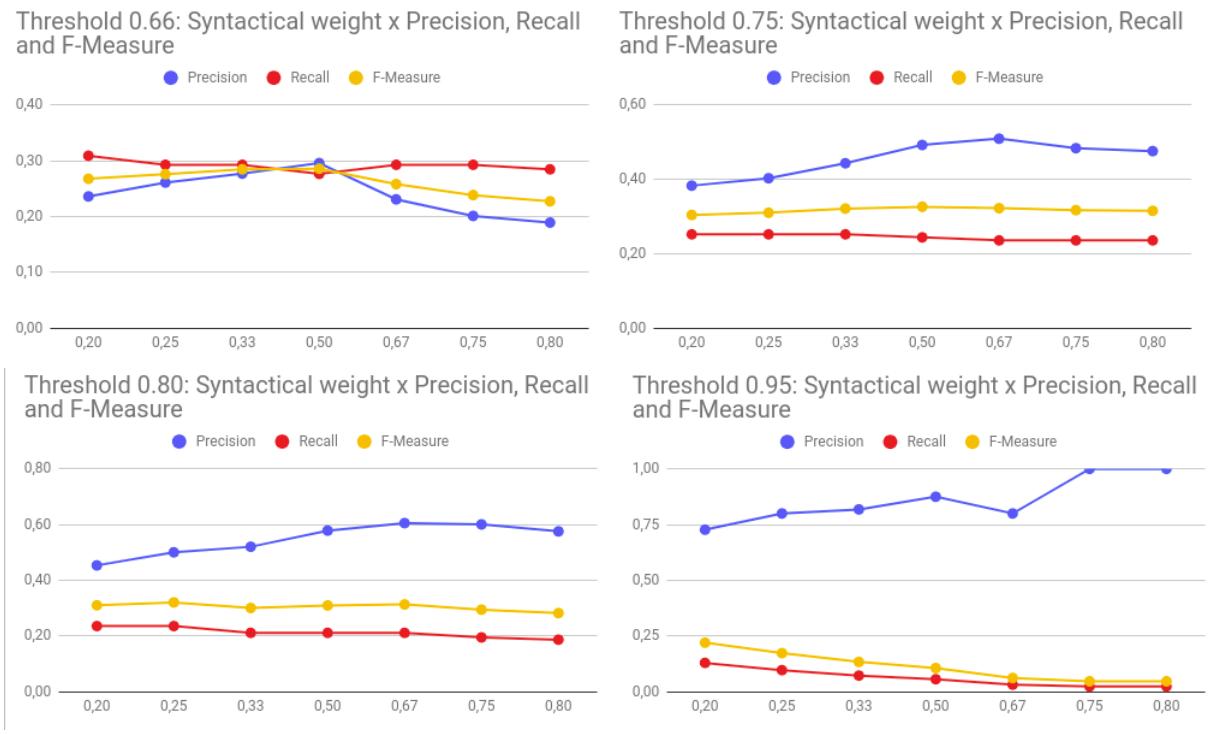

Figure 4: Results carrying out the Experiment. Horizontal axis: Syntactical weight; Vertical axis: precision, recall and f-measure. Ontology alignment from English to Portuguese by using composed similarity considering the neighbourhood analysis. MultiFarm 2015 Ontology: Conference [EN] - Conference [PT]. Pivot Language: English.

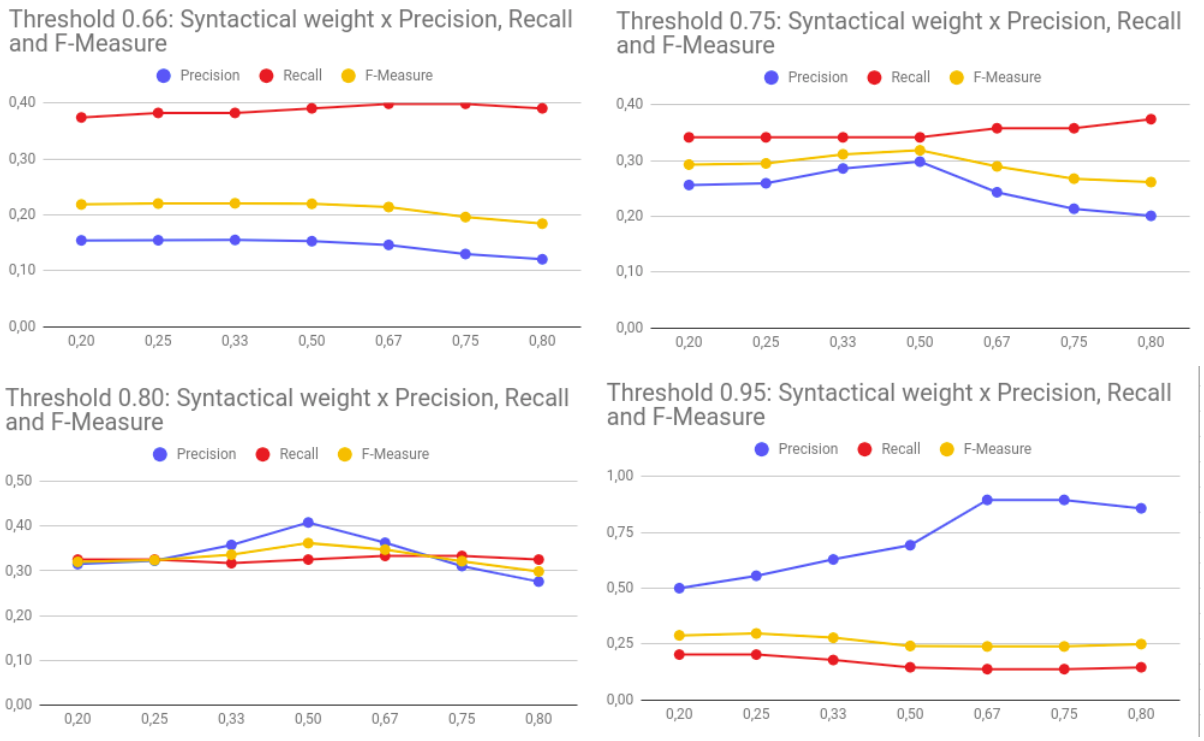

Figure 5: Results carrying out the Experiment. Horizontal axis: Syntactical weight; Vertical axis: precision, recall and f-measure. Ontology alignment from Spanish to Portuguese by using composed similarity considering the neighbourhood analysis. MultiFarm 2015 Ontology: Conference [ES] - Conference [PT]. Pivot Language: English.

in matching algorithms to determinate the adequate mappings. Existing techniques can favor from the understanding of the benefits and limitations of syntactic and semantic similarity approaches to develop a better combination of them.

In this context, this investigation contributed with several experiments to determinate the relevant aspects to be considered in the alignment of ontologies described in different languages. Our experiments were designed to help us understanding the effects of considering the ontological structure in the matching process and the quality of the generated cross-lingual ontology mappings.

Our proposal concerns the influence of the ontological structure and similarity measures on cross-language ontology matching. Our goal was to understand how to combine them aiming to build accurate cross-lingual ontology alignments. To this end, we took into account the weighted average between syntactic and semantic similarities. Our approach considered the neighbour concepts directed related to another specific concept under analysis.

The choice of weights assigned to each similarity measure played an important role in the results. As we showed empirically, semantic and syntactic similarities might have the same relevance, *i.e.*, the same weight. Considering the syntactic weight close to 0.50 generated the best mappings, *i.e.*, it resulted in alignments with the highest f-measure value. Thus, our technique may be understood as a good alternative to syntactic or semantic only methods. It might perform even better taking into account the correct parameters.

We found that the gain of effectiveness may vary according to the language describing the content of the ontologies. Comparing the Figures 4 and 5, we observe that the precision obtained in matching from English to Portuguese is better than those ones from Spanish to Portuguese. On the other hand, the recall obtained in the alignment from Spanish to Portuguese is considerably better.

A possible explanation for this behaviour might be the use of English as a pivot language. The experiments that built mappings from Portuguese into English (only one translation needed) may reduce the influence of automatic translation and improves the similarity precision. When aligning ontologies described in Spanish to the ones described in Portuguese, two automatic translations are needed, which might impact the correctness of the created mappings.

In cases of languages that share the same root and are closer, our experiments considered Portuguese and Spanish, similar words might be translated to the same word in English, which increases the syntactic similarity, once the words is considered the same. On the other hand, when we take into account the mappings from English to Portuguese, just one translation was performed, hence two similar words can still be considered different. For instance, consider the concepts from ontologies described in English, Portuguese and Spanish, labelled "weird", "estranho" and "extraño", respectively. Both labels "estranho" and "extraño" may be translated to "strange" in English as the pivot language, thus the syntactic similarity is given as 1. On the

other hand, when we consider the matching English to Portuguese, "estranho" may be translated to "strange", but it is compared to "weird", therefore the syntactic similarity is less than 1, which reduces the composed score.

The results showed an influence of threshold; as the threshold rises, the precision also increases. It may be explained by considering equivalence of only those concepts with a high level of similarity. However, f-measure reduces as the threshold increases because large values assigned to threshold turns the algorithm disregards concepts that are equivalent, but somehow was assigned a lower level of similarity than expected by the threshold. For instance, the similarity between "strange" and "estranho" equals to 0.89, but the given threshold is 0.95, thus "estranho" is not mapped to "strange". As a result, the recall drops substantially, because many correct correspondences are ignored, and thus f-measure decreases. Empirically, we concluded that the thresholds generating the more accurate mappings were $\lambda = 0.75$ and $\lambda = 0.80$.

Although the composed similarity considering the neighbourhood showed as a good alternative to syntactic or semantic only methods, there might be issues when working with ontologies whose labels are complex sentences. Our approach explored *Babelnet*, which in these scenarios fail in not finding correspondences. *Babelnet* works fine to interpret simple terms. Thus, it might be useful considering semantic algorithms such as stop-words elimination and stemming, etc. to break the complex sentences into simple structures.

Table 2 describes the results obtained by related work (ontology alignment systems) presented in OAEI (the version of 2018) with the same dataset in which our experiments were conducted. By comparing our obtained results to the presented systems, our best f-measure mean of 0.34 (obtained considering the best results in the matching English to Portuguese and Spanish to Portuguese) surpasses two of the four assessed systems, AML [7], whose f-measure is equal to 0.27 and XMAP [5], which presented 0.14 of f-measure. With these results, our algorithm gets closer to the other two better tools, just about 0.15 from KEPLER [14], the best tool placed in the competition in 2018.

Work	Precision	Recall	F-Measure
KEPLER [14]	0.85	0.36	0.49
LogMap [12]	0.95	0.28	0.41
AML [7]	0.96	0.16	0.27
XMAP [5]	0.13	0.19	0.14

Table 2: Results obtained with existing ontology alignment systems on OAEI (Multifarm Track) in 2018 considering the same ontology in different languages.

Our obtained findings support the hypothesis that composing different types of similarity measures and taking into account the neighbour concepts for cross-lingual

ontology mappings with adequate parameter values can reveal satisfactory generated ontology alignments.

7 Conclusion

Alignment of large ontologies described in different natural languages remains an open research challenge. In this work, we proposed an approach based on the weighted mean of syntactic and semantic similarities for this task. Our approach considered the influence of neighbour concepts on the cross-lingual alignment method, combining it with the composed similarity. The defined algorithms were implemented and we carried out a series of experiments to evaluate the effectiveness of this approach. Our findings based on experiments with standard datasets reveled the effectiveness of combining similarity measure and considering the neighbourhood of concepts in the cross-language ontology alignment problem. Future work involves to improve our cross-lingual alignment proposal by considering different combinations of background Knowledge, such as specific-domain thesauri to evaluate the semantic similarity. In addition, we plan to investigate different ways of computing the syntactic and semantic similarities considering additional stages in the preprocessing of concept labels.

Acknowledgments

This work has the financial support of São Paulo Research Foundations (FAPESP) (grants #2017/23522-3 and #2017/02325-5)[6].

References

[1] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*.

[2] Ondrej Šváb-Zamazal Christian Meilicke, Cassia Trojahn and Dominique Ritze. In *Multilingual ontology matching evaluation–a first report on using multifarm. In The Extended Semantic Web Conference(ESWC 2012): Satellite Events*, pages 132–147. Springer, 2012.

[3] Juliana Medeiros Destro, Julio Cesar dos Reis, Ariadne Maria Brito, Rizzoni Carvalho, and Ivan Luiz Marques Ricarte. Influence of semantic similarity measures on ontology cross-language mappings. *Proceedings of the Symposium on Applied Computing*, pages 323–329, 2017.

[6]The opinions expressed in here are not necessarily shared by the financial support agency.

[4] Duy Dinh, Julio Cesar Dos Reis, Cédric Pruski, Marcos Da Silveira, and Chantal Reynaud-Delaître. Identifying relevant concept attributes to support mapping maintenance under ontology evolution. *Web semantics: Science, services and agents on the world wide web*, 29:53–66, 2014.

[5] Warith Eddine Djeddi, Sadok Ben Yahia, and Mohamed Tarek Khadir. XMap : Results for OAEI 2018. *CEUR Workshop Proceedings*, 2288, 2018.

[6] Gabriel Oliveira dos Santos and Julio Cesar dos Reis. Identificação de mapeamentos entre ontologias em diferentes línguas. Technical Report IC-18-09, Institute of Computing, University of Campinas, July 2018.

[7] Daniel Faria, Booma S. Balasubramani, Vivek R. Shivaprabhu, Isabela Mott, Catia Pesquita, Francisco M. Couto, and Isabel F. Cruz. Results of AML participation in OAEI 2018. *CEUR Workshop Proceedings*, 2288, 2018.

[8] Bo Fu, Rob Brennan, and Declan O 'sullivan. Cross-lingual ontology mapping - an investigation of the impact of machine translation. In *Proceedings of the 4th Annual Asian Semantic Web Conference (ASWC 2009)*. Springer, 2009.

[9] Bo Fu, Rob Brennan, and Declan 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*, 15:15–36, 2012.

[10] Thomas R. Gruber. In *Toward principles for the design of ontologies used for knowledge sharing*, volume 43, pages 907–928. International Journal of Human-Computer Studies, 1995.

[11] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928, December 1995.

[12] E. Jimenez-Ruiz, B. Cuenca Grau, and V. Cross. LogMap family participation in the OAEI 2018. *CEUR Workshop Proceedings*, 2288, 2018.

[13] Mohammad Taher Pilehvar José Camacho-Collados and Roberto Navigli. Nasari: a novel approach to a semantically-aware representation of items. In *Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL 2015)*, pages 567–577, Denver,USA, 2015.

[14] Marouen Kachroudi, Gayo Diallo, and Sadok Ben Yahia. KEPLER at oaei 2018. *CEUR Workshop Proceedings*, 2288, 2018.

[15] Abderrahmane Khiat. Crolom results for OAEI 2017: summary of cross-lingual ontology matching systems results at OAEI. *CEUR Workshop Proceedings*, 2032:129–134, 2017.

[16] Pierre Lafon. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165, 1980.

[17] Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, feb 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[18] Christian Meilicke, Raul Garcia-Castro, Fred Freitas, Willem Robert van Hage, Elena Montiel-Ponsoda, Ryan Ribeiro de Azevedo, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, Andrei Tamilin, Cássia Trojahn dos Santos, and Shenghui Wang. Multifarm: A benchmark for multilingual ontology matching. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15:62–68, 2012.

[19] Gonzalo Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, March 2001.

[20] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[21] DuyHoa Ngo and Zohra Bellahsene. Efficient semantic verification of ontology alignment. *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 1:141–148, 2015.

[22] DuyHoa Ngo and Zohra Bellahsene. Overview of YAM++ - (not) yet another matcher for ontology alignment task. *Web Semantics: Science, Services and Agents on the World Wide Web*, 41:30–49, 2016.

[23] Thi Thuy Anh Nguyen and Stefan Conrad. Ontology matching using multiple similarity measures. In *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 01, pages 603–611, 2015.

[24] Catia Pesquita, Daniel Faria, André O. Falcão, Phillip W. Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7), 2009.

[25] Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. Align, disambiguate and walk: A unified approach for measuring semantic similarity. pages 1341–1351, 2013.

[26] David M W Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation. *Journal of Machine Learning Technologies.*, 2:37–63, 2011.

[27] Dennis Spohr, Laura Hollink, and Philipp Cimiano. A machine learning approach to multilingual and cross-lingual ontology matching. In *Proceedings of the 10th International Semantic Web Conference (ISWC 2011)*, pages 665–680. Springer, 2011.

[28] S. K. Stoutenburg. Acquiring advanced properties in ontology mapping. In *Proceedings of the 2nd PhD workshop on Information and knowledge management*, pages 9–16, 2008.

[29] Cássia Trojahn, Bo Fu, Ondřej Zamazal, and Dominique Ritze. State-of-the-art in multilingual and cross-lingual ontology matching. In *Towards the Multilingual Semantic Web*, pages 119–135. Springer, 2014.

[30] Maria Vargas-Vera and Miklos Nagy. State of the art on ontology alignment. *Int. J. Knowl. Soc. Res.*, 6(1):17–42, January 2015.