Coupling of WordNet Entries for Ontology Mapping using Virtual Documents

Frederik C. Schadd and Nico Roos

Department of Knowledge Engineering,
Maastricht University, Maastricht, The Netherlands,
frederik.schadd@maastrichtuniversity.nl,
roos@maastrichtuniversity.nl

Abstract.
Facilitating information exchange is a crucial service for ontology-based knowledge systems. This can be achieved by the mapping of two heterogenous ontologies. Many mapping frameworks utilize language-based knowledge resources such as WordNet. By coupling all ontology concepts to a corresponding entry in WordNet, one can quantify the lexical relatedness of any two ontology concepts. However, coupling the correct entry is a difficult task due to the ambiguous nature of names. Coupling the wrong entries hence yields similarity values that do not correctly express the relatedness of two given concepts, resulting in a poor performance of the overall mapping framework. This paper proposes an approach for the more accurate coupling of ontology concepts with their corresponding WordNet entries. The basis of the proposed approach is the creation of separate virtual documents representing the different ontology concepts and WordNet entries and coupling these according to their document similarities. The extent of improvements using this approach are evaluated using a data set originating from the Ontology Alignment Evaluation Initiative (OAEI). Furthermore, the performance of a framework using our approach is demonstrated using the results of the OAEI 2011 competition.

1 Introduction

Sharing and reusing knowledge is an important aspect in modern information systems. Since multiple decades, researchers have been investigating methods that facilitate knowledge sharing in the corporate domain, allowing for instance the integration of external data into a company's own knowledge system. Ontologies are at the center of this research, allowing the explicit definition of a knowledge domain. With the steady development of ontology languages, such as the current OWL language [11], knowledge domains can be modelled with an increasing amount of detail. Due to the Semantic Web vision [2], information sources on the future World Wide Web will store machine readable information, allowing autonomous agents to collect and interpret information automatically. Just as in current knowledge systems, each information source on the World Wide Web will store its structured content with a publicly available ontology describing the semantics of stored information. Such ontologies are generally developed independently, resulting in many different ontologies describing the same

domain. Thus, agents roaming the Semantic Web need to be able to integrate knowledge of heterogenous sources into their own representation of a specific domain.

Commonly, ontology mapping tools combine a variety of similarity measures using advanced aggregation techniques. The application of an extraction technique on the aggregated similarities can then be used to produce an alignment. The focus of this article lies on similarities measures that utilize lexical ontologies. More specifically, we investigate the automatic identification of corresponding entries in these ontologies through the use of virtual documents and information retrieval techniques, such that the semantic relatedness of any two ontology entities can be accurately specified. This article expands on previous research [17] by applying a formal virtual document model and evaluating the system against state-of-the-art frameworks in the OAEI 2011 campaign.

2 Related Work

Matching heterogenous ontologies has traditionally been done either manually or using semi-automatic tools. However, many research groups have focused their attention on automatic mapping approaches. This has led to the development of ontology mapping frameworks [18] which all utilize different techniques and resources. Many of these include lexical ontologies such as WordNet in their matching procedure. Falcon-AO [15] was the first framework to successfully apply the concept of virtual documents in the ontology mapping process. Here, virtual documents are created where each document represents a different ontology concept, such that a similarity matrix can be computed by applying a document similarity measure on the virtual documents.

Budanitsky et al. [3] evaluated five different measures of expressing the semantic relatedness between WordNet concepts, which subsequently can be applied to approaches that use different lexical ontologies. Buitelaar et al. [4] proposed a linguistic model as a labelling system, such that natural language can be generated using the ontology concepts. Such a model would be useful for situations when the name of a concept has no matching entry in a lexical ontology, allowing the linguistic decomposition of a name such that an appropriate lexical entry might still be mapped.

The techniques applied in this research are related to the field of Word Sense Disambiguation, which can be approached using numerous different techniques [14]. The strongest related technique is the *Lesk* [10] method, however key differences to the proposed approach is that it is limited to the glossary of the concept, omitting other information such as labels and the data of related concepts, and that it does not allow for a weighting of the terms according to a specified document model. The *Extended-Lesk* [1] method does also incorporate glossaries of related concepts, however still lacks the inclusion of non-glossary information and the weighting of terms according to their origin within the ontology, which the proposed approach does provide.

3 Motivation

Lexical ontologies are useful assets for ontology mapping systems. Established research primarily focused on developing frameworks or theoretical models which allow sophisticated reasoning functionalities, provided the ontology concepts are annotated, or

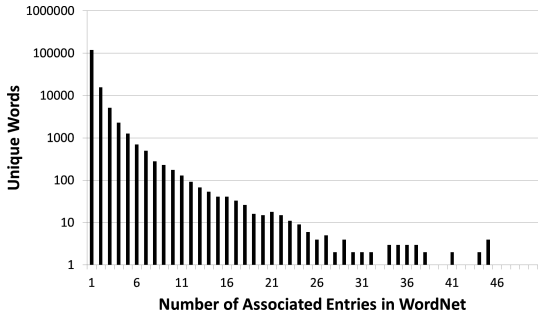

Fig. 1. Frequency of the amounts of possible concepts that can be coupled to a given word. All unique words occurring in WordNet were used as data.

'coupled', using the framework constructs. However, in order to utilize a lexical ontology or appropriate framework it is necessary that the given ontologies actually contain these couplings. Unfortunately, this is rarely the case, meaning that the ontology concepts need to be coupled during the mapping procedure. This is not a straight-forward task since words can have different meaning, such that when looking up the name of a concept in a lexical ontology it can occur that there are multiple entries for that name. Figure 1 indicates the extent of such situations occurring within WordNet [13], by displaying the frequency of the amounts of possible concepts that can be coupled to a given word, using all unique concept labels that occur in WordNet as queries.

From Figure 1 one can see that, while there is a large collection of words that only have one entry in WordNet, a significant proportion of the data leads to multiple entries. This issue becomes increasingly prevalent when the concept names do not directly occur in a lexical ontology, due to the names being composite words or technical terms. Research is required into methods that can automatically couple ontology concepts to entries in lexical ontologies for the situation when such couplings are not specified.

4 Virtual Documents

We will provide a brief introduction to virtual documents and provide a detailed description of the creation of a virtual document representing the meaning of a ontology concept or WordNet entry. The general definition of a virtual document [20] is any document for which no persistent state exists, such that some or all instances of the given document are generated at run-time. A simple example would be creating a template for a document and completing the document using values stored in a database.

In this domain the basic data structure used for the creation of a virtual document is a linked-data model. It consists of different types of binary relations that relate concepts in order to create an exploitable structure, i.e. a graph. RDF [9] is an example of a linked-data model, which can be used to denote an ontology according to the OWL specification [11]. The inherent data model of WordNet has similar capacities, however it stores its data using a database. A key feature of a linked-data model is that it not only

allows the extraction of literal data for a given concept, but also enables the exploration of concepts that are related to that particular concept, such that the information of these neighboring concepts can then be included in the virtual document.

We will provide a generalized description of the creation of a virtual document based on established research [15]. The generalization has the purpose of providing a description that is not only applicable to an OWL/RDF ontology like the description given in Qu et al. [15], but also to the WordNet model. To provide the functions that are used to create a virtual document, the following terminology is used:

Synset: Basic element within WordNet, used to denote a specific meaning using a list of synonyms. Synsets are related to other synsets by different semantic relations, such as hyponymy and holonymy.

Concept: A named entity in the linked-data model. A concept denotes a named class or property given an ontology and a synset when referring to WordNet.

Link: A basic component of a linked-data model for relating elements. A link is directed, originating from a source and pointing towards a target, such that the type of the link indicates what relation holds between the two elements. An example of a link is a triplet in an RDF graph.

sou(s), type(s), tar(s): The source element, type and target element of a link s, respectively. Within the RDF model, these three elements of a link are also known as the subject, predicate and object of a triplet.

Collection of words: A list of unique words where each word has a corresponding weight in the form of a rational number.

+: Operator denoting the merging of two collections of words.

A concept definition within a linked-data model contains different types of literal data, such as a name, different labels, annotations and comments. The RDF model expresses some of these values using the *rdfs:label*, *rdfs:comment* relations. Concept descriptions in WordNet have similar capacities, but the labels of a concepts are referred to as its synonyms and the comments of a concept are linked via the glossary relation.

Definition 1. *Let ω be a concept of a linked-data model, the description of ω is a collection of words defined by (1):*

$$\begin{aligned} Des(e) = \alpha_1 * &\textit{ collection of words in the name of } \omega \\ +\alpha_2 * &\textit{ collection of words in the labels of } \omega \\ +\alpha_3 * &\textit{ collection of words in the comments of } \omega \\ +\alpha_4 * &\textit{ collection of words in the annotations of } \omega \end{aligned} \tag{1}$$

α_1, α_2, α_3 and α_4 are each rational numbers in $[0, 1]$, such that words can be weighed according to their origin.

Next to accumulating information that is directly related to a specific concept, one can also include the descriptions of neighboring concepts that are associated with that concept via a link. Such a link can be a standard relation that is defined in the linked-data model, for instance the specialization relation. However, it can also be a relation that is defined specifically for this ontology, such as an object property in the OWL language. The OWL language supports the inclusion of blank-node concepts which allow complex

logical expressions to be included in concept definitions. However, since not all linked-data models support the blank-node functionality, among which WordNet, these are omitted in our generalization. For more information on how to include blank nodes in the description, consult the work by Qu et al. [15].

To explore neighboring concepts, three neighbor operations are defined. $SON(\omega)$ denotes the set of concepts that occur in any link for which ω is the source of that link. Likewise $TYN(\omega)$ denotes the set of concepts that occur in any link for which ω is the type of that link and $TAN(\omega)$ denotes the set of concepts that occur in any link for which ω is the target. WordNet contains inverse relations, such as hypernym being the inverse of the hyponym relation. When faced with two relations which are the inverse of each other, only one of the two should be used such that descriptions of neighbors are not included twice in the virtual document. The formal definition of the neighbor operators is given below.

Definition 2. *Let ω be a named concept and s be a variable representing an arbitrary link. The set of source neighbors $SON(\omega)$ is defined by (2), the set of type neighbors of ω is defined by (3) and the set of target neighbors of ω is defined by (4).*

$$SON(\omega) = \bigcup_{sou(s)=\omega} \{type(s), tar(s)\} \tag{2}$$

$$TYN(\omega) = \bigcup_{type(s)=\omega} \{sou(s), tar(s)\} \tag{3}$$

$$TAN(\omega) = \bigcup_{tar(s)=\omega} \{sou(s), type(s)\} \tag{4}$$

Given the previous definitions, the definition of a virtual document of a specific concept can be formulated as follows.

Definition 3. *Let ω be a concept of a linked-data model. The virtual document of ω, denoted as $VD(\omega)$, is defined by (5):*

$$\begin{aligned} VD(\omega) =& Des(\omega) + \beta_1 * \sum_{\omega' \in SON(\omega)} Des(\omega') \\ &+ \beta_2 * \sum_{\omega' \in TYN(\omega)} Des(\omega') + \beta_3 * \sum_{\omega' \in TAN(\omega)} Des(\omega') \end{aligned} \tag{5}$$

Here, β_1, β_2 and β_3 are rational numbers in $[0, 1]$. This makes it possible to allocate a different weight to the descriptions of neighboring concepts of ω compared to the description of the concept ω itself.

5 Coupling Synsets

Our proposed approach aims at improving matchers applying lexical ontologies, in this case WordNet. When applying WordNet for ontology mapping, one is presented with the problem of identifying the correct meaning, or synset, for each entity in both ontologies that are to be matched. The goal of our approach is to automatically identify

the correct synsets for each entity of an ontology using information retrieval techniques. Given two ontologies O_1 and O_2 that are to be matched, O_1 contains the sets of entities $E_x^1 = \{e_1^1, e_2^1, ..., e_m^1\}$, where x distinguishes between the set of classes, properties or instances, O_2 contains the sets of entities $E_x^2 = \{e_1^2, e_2^2, ..., e_n^2\}$, and $C(e)$ denotes a collection of synsets representing entity e. The main steps of our approach, performed separately for classes, properties and instances, can be described as follows:

1. For every entity e in E_x^i, compute its corresponding set $C(e)$ by performing the following procedure:
   (a) Assemble the set $C(e)$ with synsets that might denote the meaning of entity e.
   (b) Create a virtual document of e, and a virtual document for every synset in $C(e)$.
   (c) Calculate the document similarities between the virtual document denoting e and the different virtual documents originating from $C(e)$.
   (d) Discard all synsets from $C(e)$ that resulted in a low similarity score with the virtual document of e, using some selection procedure.
2. Compute the WordNet similarity for all combinations of $e^1 \in E_x^1$ and $e^2 \in E_x^2$ using the processed collections $C(e^1)$ and $C(e^2)$.

The essential operation of the approach is the exclusion of synsets from the WordNet similarity calculation. This is determined using the document similarities between the virtual documents originating from the synsets and the virtual document originating from the ontology concepts. Figure 2 illustrates steps 1.b - 2 of our approach for two arbitrary ontology entities e^1 and e^2: Once the similarity matrix, meaning all pairwise similarities between the entities of both ontologies, are computed, the final alignment of the mapping process can be extracted or the matrix can be combined with other similarity matrices.

5.1 Synset Selection and Virtual Document Similarity

The initial step of the approach entails the allocation of synsets that might denote the meaning of a concept. The name of the concept, meaning the fragment of its URI, and alternate labels, when provided, are used for this purpose. While ideally one would prefer synsets which contain an exact match of the concept name or label, precautions must

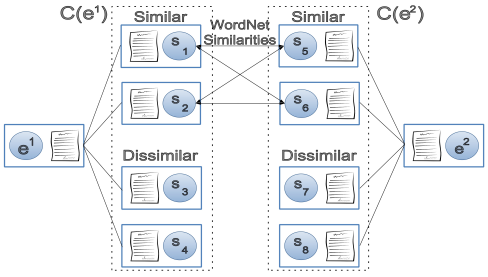

Fig. 2. Visualization of step 1.b-2 of the proposed approach for any entity e^1 from ontology O_1 and any entity e^2 from ontology 2.

be made for the eventually that no exact match can be found. For this research, several pre-processing methods have been applied such as the removal of special characters, stop-word removal and tokenization. It is possible to enhance these precautions further by for instance the application of advanced natural language techniques, however the investigation of such techniques in this context is beyond the scope of this research. When faced with ontologies that do not contain concept names using natural language, for instance by using numeric identifiers instead, and containing no labels, it is unlikely that any pre-processing technique will be able to reliably identify possible synsets, in which case a lexical similarity is ill-suited for that particular matching problem.

In the second step, the virtual document model as described in section 4 is applied to each ontology concept and to each synset that has been gathered in the previous step. The resulting virtual document are represented using the well known vector-space model [16]. In order to compute the similarities between the synset documents and the concept documents, the established cosine-similarity is applied [19].

5.2 Synset Selection

Once the similarities between the entity document and the different synset documents are known, a selection method is applied in order to only couple synsets that resulted in a high similarity value, while discarding the remaining synsets. It is possible to tackle this problem from various angles, ranging from very lenient methods, discarding only the very worst synsets, to strict methods, coupling only the highest scoring synsets. Several selection methods have been investigated for this research, such that both strict and lenient methods are tested. To test lenient selection methods, two methods using the arithmetic (A-MEAN) and geometric mean (G-MEAN) as a threshold have been investigated. Two other methods have been tested in order to investigate whether a more strict approach is more suitable. The first method, annotated as M-STD, consists of subtracting the standard deviation of the similarities from the maximum obtained similarity, and using the resulting value as a threshold. This method has the interesting property that it is more strict when there is a subset of documents that is significantly more similar than the remaining documents, and more lenient when it not as easy to identify the correct correspondences. The second investigated strict method (MAX) consists of only coupling the synset where its corresponding virtual document resulted in the highest similarity value.

5.3 WordNet Distance

After selecting the most appropriate synsets using the document similarities, the similarity between two entities can now be computed using their assigned synsets. This presents the problem of determining the similarity between two sets of synsets, where one can assume that within each of these sets resides one synset that represents the true meaning of its corresponding entity. Thus, if one were to compare two sets of synsets, assuming that the originating entities are semantically related, then one can assume that the resulting similarity between the two synsets that both represent the true meaning of their corresponding entities, should be a high value. Inspecting all pairwise similarities between all combinations of synsets between both sets should yield at least one high

similarity value. When comparing two sets originating from semantically unrelated entities, one can assume that there should be no pairwise similarity of high value present. A reasonable way of computing the similarity of two sets of synsets is to compute the maximum similarity over all pairwise combination between the two sets.

There exist several ways to compute the semantic similarity within WordNet [3] that can be applied, however finding the optimal measure is beyond the scope of this paper. Here, a similarity measure with similar properties as the Leacock-Chodorow similarity [3] has been applied. The similarity $sim(s_1, s_2)$ of two synsets s_1 and s_2 is computed using the distance function $dist(s_1, s_2)$, which determines the distance of two synsets inside the taxonomy, and the over depth D of the taxonomy:

$$sim(s_1, s_2) = \begin{cases} \frac{D - dist(s_1, s_2)}{D} & \text{if } dist(s_1, s_2) \leq D \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

This measure is similar to the Leacock-Chodorow similarity in that it relates the taxonomic distance of two synsets to the depth of the taxonomy. In order to ensure that the resulting similarity values fall within the interval of $[0, 1]$ and thus can be integrated into larger mapping systems, the log-scaling has been omitted in favor of a linear scale.

6 Experiments

In this section, the experiments that have been performed to test the effectiveness of our approach will be presented. Subsection 6.1 details an evaluation on the conference data set, originating from the Alignment Evaluation Initiative 2010 (OAEI 2010) competition [5], which demonstrates to what extent our Synset coupling method can improve a framework using WordNet. Subsections 6.2 and 6.3 compares our matcher, referred to as MaasMatch (MM), against existing frameworks using the results from the OAEI 2011 campaign [6] in which MaasMatch participated. For this research, the weighing parameters for the virtual documents were all given the equal value of 1 such that the vectors resemble document vectors originating from human-written documents, since a sensitivity analysis of these parameters is beyond the scope of this article and will be addressed in future research. The WordNet similarity matrix is combined with the similarity matrix stemming from the Jaro [8] string similarity using the average similarity of each pairwise combination, upon which the Naive descending extraction algorithm [12] is applied to generate a temporary mapping. For the experiment in subsection 6.1 a threshold of 0.7 is used, where for the OAEI 2011 evaluation a threshold of 0.95 has been applied. MaasMatch can be downloaded from the SEALS-platform, which can be accessed at `http://www.seals-project.eu/`.

When evaluating the performance of an ontology mapping procedure, the most common practise is to compare a generated alignment with a reference alignment of the same data set. Measures such as precision and recall [7], can then be computed to express the correctness and completeness of the computed alignment. Given a generated alignment A and reference alignment R, the precision $P(A, R)$ and recall $R(A, R)$ of the generated alignment A are defined as:

$$P(A, R) = \frac{R \cap A}{A} \quad (7) \qquad R(A, R) = \frac{R \cap A}{R} \quad (8)$$

Given the precision and recall of an alignment, a common measure to express the overall quality of the alignment is the F-measure [7]. Given a generated alignment A and a reference alignment R, the F-measure can be computed al follows:

$$\text{F-measure} = \frac{2 * P(A,R) * R(A,R)}{P(A,R) + R(A,R)} \tag{9}$$

The F-measure is the harmonic mean between precision and recall. Given that these measurements require a reference alignment, they are often inconvenient for large-scale evaluations, since reference alignments require an exceeding amount of effort to create. The used data sets, however, do feature reference alignments, such that the performance of a mapping approach can easily be computed and compared.

6.1 Synset Coupling

To investigate to what extent our approach improves a framework using a WordNet similarity, we evaluated our framework using different variations of our approach on the conference data set of the OAEI 2010 competition. This data set consists of real-world ontologies describing the conference domain and contains a reference alignment for each possible combination of ontologies from this data set. Figure 3 displays the results of our approach on the conference data set. Each entry in Figure 3 denotes a different synset selection procedure, which are arranged according to their strictness, such that the most lenient method is located on the far left side and the most strict method is located on the far right. Note that the most lenient method, denoted as 'none', does not discard any synsets based on their document similarities, resulting in the equivalent of a conventional WordNet similarity, which can be used as a basis for comparison. From Figure 3 we can see two notable trends. First and foremost is the observation that the more strict the synset selection procedure is, the higher the overall performance of the matcher is, as indicated by the F-Measure. This is solely due to a steady increase of the precision of the alignments. Secondly, it is notable that the recall of the alignments decreases slightly upon increasing the strictness of the selection procedure. This can be explained by the possibility that during the selection synsets are discarded that better denote the meaning of a given concept than its similarity value indicates.

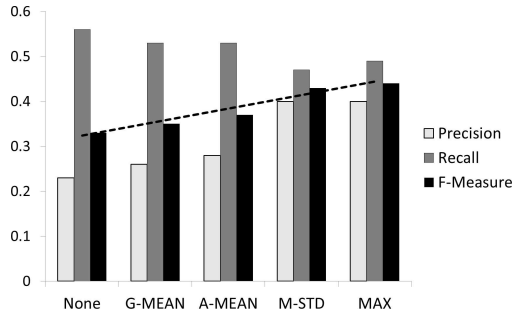

Fig. 3. Evaluation of coupling methods on the OAEI 2010 Conference data set.

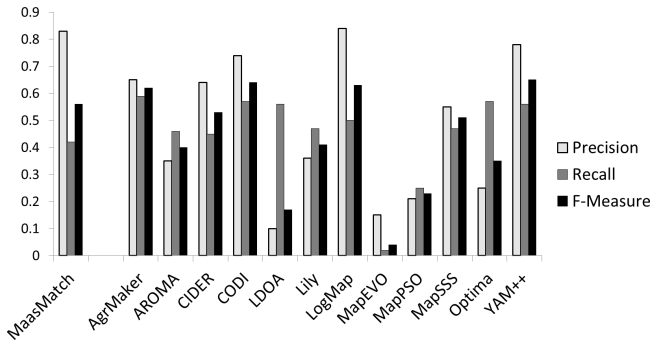

Fig. 4. Results of MaasMatch in the OAEI 2011 competition on the conference data set, compared against the results of the other participants

Overall, the highest performing variation of our coupling technique achieved an f-measure 0.44, which is an increase of 0.11 when compared to our framework without a selective coupling method. These results indicate that our coupling technique improves the computed WordNet similarities to such an extent that the computed alignments exhibit a significant increase in quality, mostly with regard to their precision.

6.2 OAEI 2011: Conference Dataset

Using the best performing synset selection method, as determined in subsection 6.1, our framework has been evaluated in the OAEI 2011 competition. The results of the evaluation on the conference data set can be seen in Figure 4. From Figure 4 one can see that MaasMatch achieved a high precision and moderate recall over the conference data set, resulting in the fifth-highest f-measure among the participants, which is above average. A noteworthy aspect of this result is that this result has been achieved by only applying lexical similarities, which are better suited at resolving naming conflicts as opposed to other conflicts. This in turn also explains the moderate recall value, since it would require a larger, and more importantly a more varied set of similarity values, to deal with the remaining types of heterogeneities as well. Hence, it is encouraging to see these good results when taking into account the moderate complexity of the framework.

6.3 OAEI 2011: Benchmark Dataset

The benchmark data set is a synthetic data set, where a reference ontology is matched with many systematic variations of itself. These variations include many aspects, such as introducing errors or randomizing names, omitting certain types of information or altering the structure of the ontology. Since a base ontology is compared to variations of itself, this data set does not contain a large quantity of naming conflicts, which our approach is targeted at. However, it is interesting to see how our framework performs when faced with every kind of heterogeneity. Figure 5 displays the results of the OAEI 2011 evaluation on the benchmark data set.

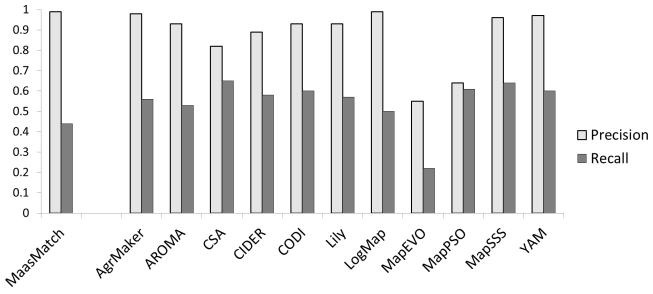

Fig. 5. Results of MaasMatch in the OAEI 2011 competition on the benchmark data set, compared against the results of the other participants

From Figure 5 we can see that the overall performance MaasMatch resulted in a high precision score and relatively low recall score when compared to the competitors. The low recall score can be explained by the fact that the WordNet similarity of our approach relies on collecting synsets using information stored in the names of the ontology concepts. The data set regularly contains ontologies with altered or scrambled names, making it extremely difficult to couple synsets that might denote the meaning of an entity. These alterations also have a negative impact on the quality of the constructed virtual documents, especially if names or annotations are scrambled or completely left out, resulting in MaasMatch performing poorly in benchmarks that contain such alterations. Despite these drawbacks, it was possible to achieve results similar to established matchers that address all types of heterogeneities. Given these results, the performance can be improved if measures are added which tackle other types of heterogeneities, especially if such measures increase the recall without impacting the precision.

7 Conclusion

In this paper, we proposed a method to improve the coupling of ontology concepts with their corresponding WordNet entries. The experiment on the OAEI 2010 data set shows that our approach increases the quality of the computed alignments, mainly with regards to their precision. Furthermore, it is established that strict coupling methods produce better results than lenient coupling methods. The result of the OAEI 2011 evaluation show that a framework using the proposed technique can compete with established frameworks, especially with regards to the conference data set. However, the results of the benchmark data set indicate a reliance on the presence of adequate concept names and descriptions. Future research can be performed on improving the robustness of our approach when given distorted names and descriptions.

The recall of the alignments slightly decreases if our approach is applied, indicating that occasionally the correct meaning of an entity is not established. A possible solution would be the improvement of the representative strength of the virtual documents. This can be achieved by refining the current virtual document model, such that for instance descriptions from different OWL types of relations receive different weights.

References

[1] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th international joint conference on Artificial intelligence*, IJCAI'03, pages 805–810, San Francisco, CA, USA, 2003.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

[3] A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources, second meeting of the NAACL*, 2001.

[4] P. Buitelaar, P. Cimianop, P. Haase, and M. Sintek. Towards linguistically grounded ontologies. In *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 111–125. Springer Berlin / Heidelberg, 2009.

[5] J. Euzenat, A. Ferrara, C. Meilicke, J. Pane, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Svab-Zamazal, V. Svatek, and C. Trojahn. First results of the ontology alignment evaluation initiative 2010. In *Proceedings of ISWC Workshop on OM*, 2010.

[6] J. Euzenat, A. Ferrara, R.W. van Hague, L. Hollink, C. Meilicke, A. Nikolov, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Svab-Zamazal, and C. Trojahn dos Santos. Results of the ontology alignment evaluation initiative 2011. In *Proc. 6th ISWC workshop on ontology matching (OM), Bonn (DE)*.

[7] F. Giunchiglia, M. Yatskevich, P. Avesani, and P. Shvaiko. A large dataset for the evaluation of ontology matching. *Knowl. Eng. Rev.*, 24:137–157, June 2009.

[8] M. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *J. of the American Statistical Association*, 84(406):pp. 414–420, 1989.

[9] O. Lassila, R. R. Swick, and W3C. Resource description framework (rdf) model and syntax specification, 1998.

[10] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, 1986.

[11] D. L. McGuinness and F. van Harmelen. OWL web ontology language overview. W3C recommendation, W3C, February 2004.

[12] C. Meilicke and H. Stuckenschmidt. Analyzing mapping extraction approaches. *The Second International Workshop on Ontology Matching*, 2007.

[13] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38:39–41, November 1995.

[14] R. Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69, February 2009.

[15] Y. Qu, W. Hu, and G. Cheng. Constructing virtual documents for ontology matching. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 23–31, New York, NY, USA, 2006. ACM.

[16] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[17] F. C. Schadd and N. Roos. Improving ontology matchers utilizing linguistic ontologies: an information retrieval approach. In *Proceedings of the BNAIC 2011*, 2011.

[18] P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. In *Journal on Data Semantics IV*, volume 3730, pages 146–171. 2005.

[19] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 1 edition, May 2005.

[20] C. Watters. Information retrieval and the virtual document. *J. Am. Soc. Inf. Sci.*, 50:1028–1029, September 1999.