Exploiting WordNet as Background Knowledge

Chantal Reynaud, Brigitte Safar

LRI, Université Paris-Sud, Bât. G, INRIA Futurs
Parc Club Orsay-Université - 2-4 rue Jacques Monod, F-91893 Orsay, France
{chantal.reynaud, brigitte.safar}@lri.fr

Abstract. A lot of alignment systems providing mappings between the concepts of two ontologies rely on an additional source, called background knowledge, represented most of the time by a third ontology. The objective is to complement others current matching techniques. In this paper, we present the difficulties encountered when using WordNet as background knowledge and we show how the *TaxoMap* system we implemented can avoid those difficulties.

1 Introduction

In order to identify mappings between the concepts of an ontology, called the source ontology (O_{Src}), with concepts of another one, called the target ontology (O_{Tar}), a lot of recent works use additional descriptions, called background knowledge, represented by a third ontology O_{BK} [1,2,9,4,7,8,10]. The common objective is to complement current matching techniques which may fail in some cases. Some works as [1,2,10] assume that ontology alignment can rely on a unique and predefined ontology that covers a priori all the concepts of the ontologies to be matched. Conversely, other works [9] suppose that there does not exist a priori any suitable ontology. Hence, their idea is to dynamically select online available ontologies. In this paper, we present the difficulties encountered when using WordNet as background knowledge, in particular the misinterpretation problem coming from the different senses of a term, and how the *TaxoMap* system we implemented avoids these difficulties. The solution that we propose aims at limiting the meanings of the terms involved in a match. Experimental results are given and the increase of precision obtained with a limitation of the senses of the terms is shown.

2 Use of WordNet

WordNet is an online lexical resource for English language that groups synonym terms into synsets, each expressing a distinct concept. The term associated with a concept is represented in a lexicalized form without any mark of gender or plural. Synsets are related to each other with terminological relations such as hypernym relations. WordNet can be used for ontology matching in several ways. A first way is to extend the label of a concept with the synonyms in WordNet belonging to the synset of each term contained in the label [3]. Another way is to exploit WordNet restricted to a concept hierarchy only composed of hypernym relations. Given two nodes in this hierarchy, equivalence relation can be inferred if their distance is lower than a given threshold [4]. Other works compute similarity measures [5,6,8]. This last approach leads to relevant results when the application domains of the ontologies to be mapped are very close and targeted. Conversely, results are much less satisfactory when application domains are larger. Indeed a term can belong to several synsets. This leads to misinterpretations and false positive mappings.

We illustrate this problem with results coming from experiments performed with TaxoMap [8] on the taxonomies Russia-A (O_{Tar}) and Russia-B (O_{Src}) loaded from the Ontology Matching site [11]. Both taxonomies describe Russia from different view points but Russia-B contains extra information on the means of transport. Fig.1 represents the WordNet subgraph that is exploited in the search of the terms of O_{Tar} (grey circles in Fig.1) the most similar to the terms in O_{Src} (white circles in Fig.1) that denote vehicles. As no term in O_{Tar} correspond to means of transport, all terms in O_{Src} that refer to vehicles will be related to 'Berlin', a term belonging to three synsets respectively corresponding to a city in Germany, a musician and a kind of car.

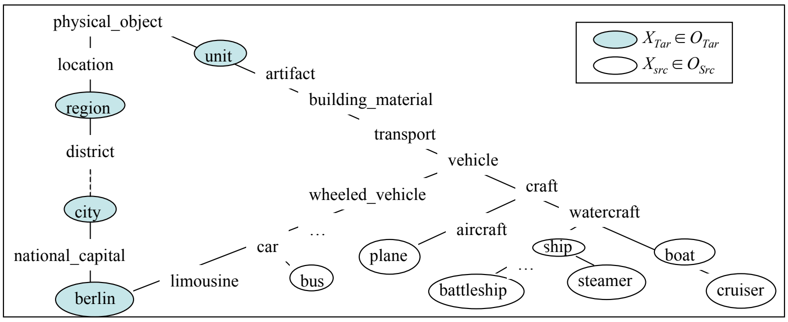

Figure 1. WordNet sub-graph.

To avoid this problem, the *TaxoMap* system relies on a limitation of the senses of the terms in WordNet. It performs a two-step process: a sub-tree is first extracted from WordNet, which only corresponds to the senses assumed to be relevant to the domain of the involved ontologies. Second, mappings are identified in this sub-tree.

2.1 Extraction of a sub-tree relevant to the domain from WordNet

The extraction of a sub-tree starts with a manual phase. If the application domains of the ontologies to be mapped are close and targeted, an expert has to identify the concept, noted $root_A$, that is the most specialized concept in WordNet which generalizes all the concepts of both ontologies. If the target ontology is relative to several distinct application domains**,** the expert has to identify several root nodes in order to cover all the topics. Then, the extraction of the relevant sub-tree needs the search of relations between all the concepts in O_{Tar} and in O_{Src} not yet mapped and $root_A$. Hypernyms of each concept are looked for in the WordNet hierarchy until $root_A$, or one of the WordNet roots, is reached. For example, a search on cantaloupe will result in these two following derivation paths:

Path 1: cantaloupe → sweet melon → melon → gourd → plant → organism → Living

Path 2: cantaloupe → sweet melon → melon → edible fruit → green goods → food

The paths from the invoked terms to the $root_A$ (*food* in the example) will only be selected because they represent the only accurate senses for the application. That way, a sub-graph, denoted T_{WN}, is obtained. It is composed of the union of all the concepts and relations of the selected paths (cf. Fig. 2). The T_{WN}'s root is the concept the most general in the application, $root_A$, leaf nodes correspond to the concepts of the ontologies to be mapped (circles in Fig.2), middle nodes have been extracted from WordNet but possibly belong to one of the two ontologies too.

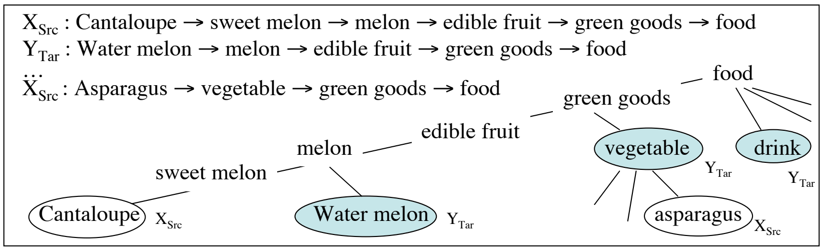
X_{Src} : Cantaloupe → sweet melon → melon → edible fruit → green goods → food
Y_{Tar} : Water melon → melon → edible fruit → green goods → food
...
X_{Src} : Asparagus → vegetable → green goods → food
food
green goods
edible fruit
melon
sweet melon
vegetable
drink
Y_{Tar}
Y_{Tar}
Cantaloupe
X_{Src}
Water melon
Y_{Tar}
asparagus
X_{Src}

Figure 2. An example of sub-graph T_{WN} where the root is food

In the Russia experiment, the chosen roots (*Location, Living Thing, Structure* and *Body of Water*) covering all the topics of O_{Tar} are not hypernyms of terms in O_{Src} relative to vehicles and no derivation path computed from these terms is retained. Missing terms are preferred over misinterpretations. Recall of matching process will be smaller but precision higher.

2.2 Mappings identification

Identification of relevant mappings consists of discovering, for each concept in O_{Src}, the closest concept in O_{Tar} that is its ancestor and that belongs to its derivation path rooted in $root_A$. So, from the sub-graph in Fig.2, the mapping (asparagus *isA* vegetable) can be derived. That process does not allow the discovery of mappings for cantaloupe because none of its ancestor nodes is a concept in O_{Tar}. All are middle nodes coming from WordNet. However, it should be very interesting to map cantaloupe to Watermelon because they are two specializations of melon, so very close semantically. Such a mapping can be derived using a similarity measure between nodes of T_{WN}. There is evidence that correspondences discovered thanks to such measures cannot be used to derive "semantic" mappings (as *isA* or *Eq* relation) which have a clear semantic and which could then be automatically exploited [8]. But there is also evidence that it would be a great pity not to exploit discovered information. So, we propose to retain such relations which will be labelled '*isClose*' as "potential" mappings for which an expert evaluation will be necessary.

Consequently, the choice in *TaxoMap* is to discover, for each concept X_{Src} in O_{Src} not yet mapped, the concept Y_{Sim} in O_{Tar} that is the most similar according to a similarity score. From that correspondence we derive the potential mapping (X_{Src} *isClose* Y_{Sim}). Then we extract, as we mentioned before, the set of semantic mappings in T_{WN}. If a concept Y_{Sim} is linked to the same concept X_{Src} both in a semantic and in a potential mapping, only the semantic mapping is retained. For example, the concept vegetable in O_{Tar} being the concept the most similar to the concept asparagus in O_{Src}, we derive the potential mapping (asparagus *isClose* vegetable). However, the semantic mapping (asparagus *isA* vegetable) can also be derived. We will then consider that this semantic mapping will be the only retained one. On the opposite, as no semantic mapping has been derived for the concept cantaloupe, the potential mapping (cantaloupe *isClose* Watermelon) will be retained.

3 Experiments

Different experiments have been performed in the micro-biology domain and on taxonomies used for tests in the Ontology Matching community [11]. All these experiments showed that if the application domain is too large, we can not use a unique root. Indeed, in that case, the concept in WordNet which is an hypernym of all the concepts to be mapped, is too general (*entity)* and T_{WN} is too big. It is composed of all the nodes in the WordNet hierarchy without any restriction. Several meanings are mixed. This leads to the derivation of non relevant mappings.

We give results obtained with the Russia taxonomies. As we see in Tab. 1, with a single root (Entity) our technique has derived 61 *isA* and 15 *isClose* mappings among 162 terms in Russia-B not yet mapped by others techniques (370 terms were to be mapped at the beginning). As no reference mappings were delivered, the results have been evaluated manually. Only 29 out of 61 *isA* mappings and 8 out of 15 *isClose* mappings were correct. In particular, all mappings relative to vehicle are false (cf. FIG.1). A significant increase of the precision of the found mappings has been obtained when several roots have been specified. In that case, several distinct sub-trees are built in the same time, one per sub-domain. Four roots have been identified: Location, Living Thing, Structure and Body of Water. Then 35 *isA* and 11 *isClose* mappings have been derived. 29 out of 35 *isA* mappings and 9 out of 11 *isClose* mappings were correct. In particular, all geographic mappings relative to towns, countries, regions and rivers were relevant. Even though the same number of correct *isA* mappings (29) appears as the results of the two experiments, these mappings are not all the same. For example, the (alcohol *isA* drink) mapping is not identified in the second experiment because the concept drink of O_{Tar} is not covered by the chosen roots. On the opposite, the (pine *isA* plant) mapping is identified whereas without senses limitation the incorrect (pine *isA* material) mapping was found.

	With a single root (*Entity*)	With several roots
# *isA* mappings found (relevant)	61 (29)	35 (29)
# *isClose* mappings found (relevant)	15 (8)	11 (9)
Total Number of mappings (relevant)	76 (37)	46 (38)
Recall (Precision)	0,23 (0,49)	0,23 (0,83)

Table 1. Number of found mappings among 162 terms in Russia-B not yet mapped

A more precise choice of the roots would very probably increase recall. In our application context, as the identification step of the roots in WordNet can be done just in reference to O_{Tar}, this task is only performed once and the identified roots will be exploited whatever the taxonomies of information sources to be aligned with O_{Tar} might be. Hence it is worthwhile to pay attention to this identification step. Our first results are already promising. Yet we think they could be even better with a more precise choice of the roots.

4 Conclusion

So, in conclusion the use of external background knowledge can be very interesting when the context of interpretation of the involved concepts is precisely known. It allows the obtention of semantic relations and so overcomes the major limitations of syntactic approaches. WordNet is often used as background resource. However, the drawback is that it is difficult to get relevant information if the meaning of the searched terms is not known. The results of our experiments using WordNet indicate that our approach based on the definition of multiple roots is a promising solution when the domain of the background knowledge is too large. Whatever the domain, the sub-tree grouping terms relevant to the application can be extracted from WordNet with our system. We showed how semantic mappings, when they exist, can be found in this sub-tree and how, when they do not exist, meaningful proximity relationships can be found instead.

References

1. Z. Aleksovski, M. Klein, W. Ten Kate, F. Van Harmelen. Matching Unstructured Vocabularies using a Background Ontology, In *Proc. of EKAW'06,* Springer-Verlag, 2006.
2. Z. Aleksovski, M. Klein, W. Ten Kate, F. Van Harmelen. Exploiting the Structure of Background Knowledge used in Ontology Matching». In *Proceedings of ISWC'06 Workshop on Ontology Matching (OM-2006),* Athens, Georgia, USA, November 2006.
3. T. L. Bach, R. Dieng-Kuntz, F. Gandon. On Ontology matching Problems – for Building a Corporate Semantic Web in a Multi-Communities Organization, in *Proc. of ICEIS* (4), 2004.
4. F. Giunchiglia, P. Shvaiko, M. Yatskevich. Discovering Missing Background Knowledge in Ontology Matching, In *Proc. of ECAI 06.*
5. Y. Kalfoglou, B. Hu. Crosi Mapping System (CMS) Result of the 2005 Ontology Alignment Contest. In *Proc of K-Cap'05 Integrating Ontologies WS*, pp. 77-85, Banff, Canada, 2005.
6. T. Pedersen, S. Patwardhan, J. Michelizzi. WordNet::Similarity - Measuring the Relatedness of Concepts, In *Proc. of AAAI-04*, July 25-29, San Jose, CA, 2004.
7. C. Reynaud, B. Safar. When usual structural alignment techniques don't apply. In *Proc. of ISWC'06 Ontology Matching WS, (OM-2006),* Poster, 2006.
8. C. Reynaud, B. Safar. Structural Techniques for Alignment of Taxonomies: Experiments and Evaluation, In TR 1453, LRI, Université Paris-Sud, Juin 2006.
9. M. Sabou, M. D'Aquin, E.Motta. Using the Semantic Web as Background Knowledge for Ontology Mapping, In *Proc. of ISWC'06 Ontology Matching WS (OM-2006),* 2006.
10. S. Zhang, O. Bodenreider. NLM Anatomical Ontology Alignement System. Results of the 2006 Ontology Alignment Contest. In *Proc. of ISWC'06 Ontology Matching* 2006.
11. http://www.atl.external.lmco.com/projects/ontology/ontologies/russia/