Exploiting BabelNet for generating subsumption

Mouna Kamel[1], Daniela Schmidt[2], Cassia Trojahn[1], and Renata Vieira[2]

[1] Institut de Recherche en Informatique de Toulouse, Toulouse, France
{mouna.kamel,cassia.trojahn}@irit.fr
[2] Pontificia Universidade Catolica do Rio Grande do Sul, Porto Alegre
daniela.schmidt@acad.pucrs.br, renata.vieira@pucrs.br

1 Introduction

Whereas the ontology matching field has developed fully in the last decades, most matching approaches are still limited to generating equivalences between entities of different ontologies. However, for many tasks, finding subsumption relations may be useful. Despite the variety of matching approaches in the literature, most of them rely on string-based techniques as an initial estimate of the likelihood that two elements refer to the same real world phenomenon, hence, the found correspondences represent equivalences with terms similarly written rather than subsumptions. This paper presents an approach relying on background knowledge from BabelNet (BN) [3] and on the notion of *context*. The latter has been exploited in different ways in ontology matching [2, 4]. They are used for disambiguating the senses that better express the meaning of ontology concepts when looking for subsumption relations between them in BN.

2 Proposed approach

The matching process is divided in two steps. The first step disambiguates the ontology concept, and the second looks for a subsumption relation between two concepts.

Concept disambiguation. It finds the semantically closer BN synset for a concept. We adopt the notion of *context* as a *bag of words*. For each ontology concept c, from the source s and target t ontologies, the context ctx_c is constructed from the available information about the concept (ID, labels, information on super and sub-concepts, etc.). The context of BN synsets ctx_{bn} is constructed from their sense and main glosses terms. We adapt the word sense disambiguation method of Lesk [1], which relies on the calculation of the word overlap between the sense definitions of two or more target terms. Here, we overlap the context ctx_c and all ctx_{bn}, coming from the synsets retrieved when looking for c in BN. We retrieve the highest overlap. The overlap function is based on the edit distance similarity between words rather than on the exact match.

Subsumption detection. Given c_s and c_t concepts from the source and target ontologies, and their respectively retrieved synsets syn_s and syn_t obtained in the previous step, we look for a subsumption relation between c_s and c_t. For that purpose, we check if syn_t belongs to the set of hypernyms $Hyper(syn_s)$, where $Hyper(syn_s)$= $\bigcup_k Hyper^k(syn_s)$ and k is length of the path from syn_s to one of its hypernym synsets, based on a depth-first search strategy.

3 Experimentation

Material and methods. We used the set of 7 ontologies from the OAEI conference data set that are involved in the 21 available reference alignments. In our experiments, compounds with no entry in BN have been pre-processed by removing the modifiers (e.g. "Invited speaker" is a "Speaker"). We empirically selected k=2 for the path length and 0.8 as edit distance threshold. We used as reference the subsumptions inferred from the available equivalence reference alignments, using Hermit and the Alignment API 4.5. As many concepts do not have any super or sub concepts, we considered 2 settings: contexts as introduced above and the whole ontology as context for each concept. The best results, which are reported here, were obtained with the latter.

Results and discussion. Table 1 shows the results (measures were computed using the Alignment API). Overall, the best results are obtained when considering alignments close to those expected (extended and semantic measures) rather than exact ones. Looking at the results for each pair of ontologies, the best results where obtained for different pairs when using the different measures: *edas-ekaw* (classical), *confOf-edas* (extended) and *conference-sigkdd* (semantic). The overall low results are mainly due to two reasons: a high number of concepts can not be found in BN and using the modifier does help so much in this task; the construction of contexts suffers from the lack of annotations in the ontologies (as well many concepts do not have any super or sub concepts), and hence, contexts are not rich enough for disambiguating the synsets.

Table 1. Results for the 21 pairs (and those discarding empty alignments) and best pair results.

Average (21 pairs)						Best pair results					
Classical		Extended		Semantic		Classical		Extended		Semantic	
Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
.06 (.23)	.02 (.07)	.14 (.16)	.05 (.06)	.02 (.02)	.22 (.22)	.22	.08	.50	.11	.14	.15

4 Conclusions

We presented an approach for generating subsumption correspondences relying on BabelNet. This task is still a gap in the field and the initial results presented here can be improved in different ways. We plan to improve the disambiguation strategy, exploiting word embeddings, to automatically enrich the ontology with annotations, to adopt a hybrid approach combining both lexical and background knowledge, to work on the confidence of the correspondences, and to look for other relations like meronymy.

References

1. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *SIGDOC*, pages 24–26. ACM, 1986.
2. A. Maedche and S. Staab. Measuring Similarity between Ontologies. In *13th Conf. on Knowledge Engineering and Knowledge Management*, pages 251–263, 2002.
3. R. Navigli and S. P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *AI*, 193:217–250, 2012.
4. F. C. Schadd and N. Roos. Coupling of Wordnet Entries for Ontology Mapping Using Virtual Documents. In *7th Workshop on Ontology Matching*, pages 25–36, 2012.