Automatic Lexical Annotation Applied to the SCARLET Ontology Matcher

Laura Po[1] and Sonia Bergamaschi[1]

DII, University of Modena and Reggio Emilia, Italy
name.surname@unimore.it

Abstract. This paper proposes lexical annotation as an effective method to solve the ambiguity problems that affect ontology matchers. Lexical annotation associates to each ontology element a set of meanings belonging to a semantic resource. Performing lexical annotation on the ontologies involved in the matching process allows to detect false positive mappings and to enrich matching results by adding new mappings (i.e. lexical relationships between elements on the basis of the semantic relationships holding among meanings).
The paper will go through the explanation of how to apply lexical annotation on the results obtained by a matcher. In particular, the paper shows an application on the SCARLET matcher.
We adopt an experimental approach on two test cases, where SCARLET was previously tested, to investigate the potential of lexical annotation. Experiments yielded promising results, showing that lexical annotation improves the precision of the matcher.

Key words: ontology matching, lexical annotation, mapping discovery

1 Introduction

Finding correspondences between heterogeneous conceptual structures is inherent to all systems that combine multiple information sources. In information integration and ontology engineering communities, the task of identified matching is a core task.

Many different matching solutions have been proposed in literature [4]. They take advantage of various properties of ontologies, e.g., structures, data instances, semantics, or labels, and use techniques from different fields, e.g., statistics and data analysis, machine learning, automated reasoning, and linguistics. Some approaches have been proposed for validating mappings with respect to the semantics of the involved ontologies. Example of works in this direction are the S-Match system [6] and the theoretical study proposed in [17]. Although the approaches are important for the validation of mappings, they do not discern elements with different meanings. Some other tools incorporate the linguistic features inside the matcher (we find some example in H-MATCH [3], Cupid [9] and Falcon-AO [8]). Differently from these matchers that include a linguistic

component, lexical annotation is able to disambiguate elements, to enable an effective comparison of them from other online ontologies or thesauri.

Until now, we developed automatic lexical annotation techniques that allow us to extract lexical knowledge from structured and semi-structured data sources detecting mappings useful for the data integration process [2]. These techniques have been developed within the MOMIS data integration system [1] and the evaluation performed on real data sets has shown good performance.

In this paper, we apply automatic lexical annotation on the elements involved in the mappings discovered by the a matcher. In particular, we will show the application of lexical annotation on the SCARLET matcher[1], but lexical annotation can be applied in general to the output of different matchers.

The SCARLET matcher has been selected as a candidate matcher because it belongs to a new generation of ontology matchers that focused on exploiting the increasing amount of online semantic data available on the Web. These applications handle the high semantic heterogeneity introduced by the increasing number of available online ontologies (different domains, different points of view, different conceptualisations). These matching algorithms exhibit very good performance, but they rely on merely syntactical techniques to anchor the terms to be matched to those found on the Semantic Web. As a result, their precision can be affected by ambiguous terms. A critical issue is to solve these ambiguity problems by introducing lexical annotation techniques, which validate the mappings by exploring the semantics of the elements involved in the matching process. In addition, lexical annotation allows the discovery of new mappings (derived from the lexical knowledge), thus enriching the results of the matcher.

Automatic lexical annotation is obtained by the application of a set of Word Sense Disambiguation (WSD) algorithms. We make use of the tool ALA (Automatic Lexical Annotator) [2] to detect the correct annotations for each ontology concept. Then, we apply rules to detect the false positive mappings discovered by the matcher and to discover new mapping among concepts.

The evaluation has been done on two test cases and compared with other WSD techniques previously tested on the SCARLET output [7].

The paper is organized as follows: Section 2 describes the SCARLET matcher, section 3 focus on the application of lexical annotation techniques on the matcher output. In section 4 the evaluation of lexical annotation techniques is shown on two different test cases. Conclusion are sketched in section 5.

2 SCARLET matcher

SCARLET[2] [11, 15] is a technique for discovering relationships between two concepts by making use of online available ontologies. Developed in the context of

[1] http://scarlet.open.ac.uk/
[2] http://scarlet.open.ac.uk/

the NeOn[3] and OpenKnowledge[4] projects, SCARLET has been primarily used to support tasks such as ontology matching and enrichment.

SCARLET discovers semantic relationships between concepts by using the entire Semantic Web as a source of background knowledge: by using semantic search engines (Swoogle [5] and WATSON [13]), it finds online ontologies containing concepts with the same names as the candidate concepts and then it derives mappings from the relationships in the online ontologies.

Scarlet is able to identify disjoint relations, subsumption relations, and correspondences [12, 14]. All relations are obtained by using derivation rules which explore not only direct relations but also relations deduced by applying subsumption reasoning within a given ontology.

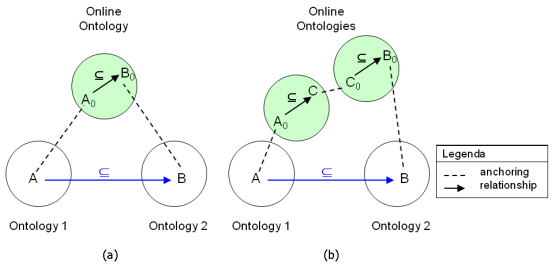

Fig. 1. SCARLET strategy using a single online ontology (a) or more ontologies (b)

Figure 1 illustrates the idea of SCARLET. A and B are the concepts to relate, and the first step is to find online ontologies containing concepts A_0 and B_0 equivalent to A and B. This process is called *anchoring* and A_0 and B_0 are called the *anchor terms* (or *anchor concepts*). Based on the relationships that link A_0 and B_0 in the retrieved ontologies, a mapping is then derived between A and B. In Figure 1 (a), the strategy assumes that a semantic relationship between the candidate concepts can be discovered in a single ontology. However, some relationships could be distributed over several ontologies. Therefore, SCARLET develops a recursive strategy to combine knowledge contained in several ontologies, and thus derives mappings from two (or more) ontologies, as shown in Figure 1 (b).

2.1 Limitations of SCARLET

As depicted in [12], the SCARLET paradigm is feasible. A baseline implementation of the SCARLET technique applied on a large-scale, real life data set has led to a precision value of 70% which correlates with the performance of other background knowledge based matchers. An analysis of the causes of false positive

[3] http://www.neon-project.org/web-content/
[4] http://www.openk.org/

mappings revealed that more than half of them were due to an incorrect anchoring caused by ambiguities: elements of the source ontology have been anchored to online ontologies on the basis of the syntax. Therefore, we can affirm that the major limitation of SCARLET prototype remains its simple, string comparison based anchoring which generated more than half of the false mappings.

SCARLET is not able to take advantage of the ontological context in which a concept appears. Instead, lexical annotation techniques, exploiting the context, can define the meaning for the concept itself. By identifying a meaning (or a set of meanings) for a concept it is possible to, more accurately, compare the concept with the concepts that appear in online ontologies. For example, if we look for a online ontology that contains the term "star" we retrieve 14 results[5]. Some of these ontologies use the word "star" as a famous actor/actress, some other assume the meaning of a celestial body. Because the SCARLET anchoring ground on all the retrieve documents, it can potentially derived false positive relationships.

On the SCARLET matcher some disambiguation techniques have already been applied. In [7] two different techniques of WSD are investigated to improve the SCARLET results, by detecting and solving the ambiguity problems inherent to the use of heterogeneous sources of knowledge. The experiments carried out confirmed that precision can be improved by using the semantic techniques. However, both the techniques proposed (semantic similarity measures) have an important limitation: they need some training set to detect an accurate threshold under which two terms are not consider synonyms. Unlike these techniques, the method proposed in this paper offers a definite answer regarding the detection of synonym relationships.

3 Lexical Annotation applied to the SCARLET matcher

An Annotation is a piece of information added in a book, document, online record, video, or other data. Lexical Annotation is a particular kind of Annotation that refers to a semantic resource. Each lexical annotation has the property to own one or more lexical descriptions. Lexical annotation of an ontology class is the explicit assignment of its meaning w.r.t. a semantic resource (i.e. entries in a thesaurus, dictionary or semantic network). Lexical Annotation differs from the Ontology-based Annotation where the annotation is performed w.r.t. an ontology and it is not mandatory that an ontology class has a lexical description.

Lexical annotation leads to several improvements in the matching process:

- it improves the precision of the matcher by detecting the false positive mappings;
- it enriches the matching results by discovering new mappings based on the lexical relationships among meanings;

[5] Data obtained on 21st June 2009 looking in WATSON for a word match of class entity.

– it is able to identify synonymous and more general classes of a concept, giving the matcher the possibility to widen the search among online ontologies.

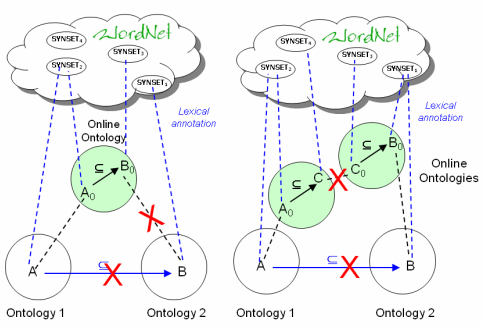

Fig. 2. Lexical annotation improvements: detection of false positive mappings

The third improvement has not been developed yet, it will be the focus of our future work.

Detection of false positive mappings The idea is to apply a combination of WSD algorithms to the source and background ontologies involved in the matching process. After the annotation of these ontologies, we examined the concepts involved in the anchoring. If a concept and its anchoring concept have disregarding meanings (i.e. if they do not have the same list of meanings), the anchoring is discharged. Lexical annotation can thus filter out wrong anchoring (with a good precision) and so, it can improve the efficiency of the matcher. Figure 2 shows how lexical annotation influences the anchoring. Let us focus on the (a) subfigure, after the annotation of all the concepts involved in the anchoring (A, B, A_0, B_0), it is possible to compare the meanings of a concept with the meanings of its anchoring concept. The anchoring between A and A_0 is preserved because the concepts have the same meanings. Instead, the anchoring between B and B_0 is discharged because the concepts have different meanings. As a consequence, the mapping among A and B is detected as false positive mapping. Also anchoring across online ontologies benefit from the lexical annotation, as shown in figure 2 (b).

New mapping discovery Lexical annotation can also enrich the matching results by discovering new mappings. In Figure 3, the process of identifying new mappings among elements is shown. First, the lexical annotation of the source and target elements is performed, then, the WordNet network is explored looking for lexical relationships between the selected meanings. For any relationships found, a mapping is inserted between the corresponding source and target elements.

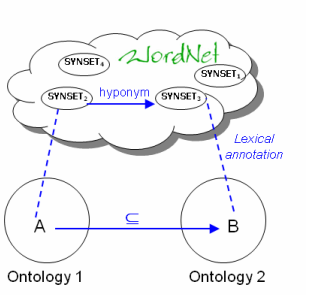

Fig. 3. Lexical annotation improvements: new mapping discovery

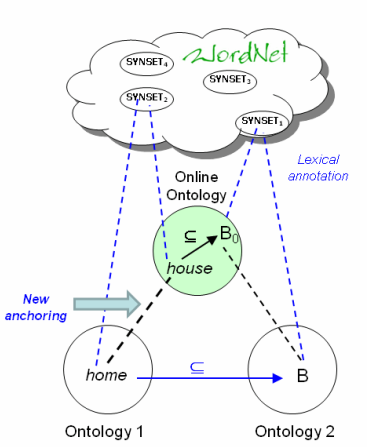

Fig. 4. Lexical annotation improvements: identification of synonymous and more general concepts

Identification of synonymous and more general concepts Lexical annotation of an ontology leads important consequences. Identifying a meaning for a ontology class means that we are able to detect synonymous (terms that share a meaning) and more general concepts of the given class. Synonymous and more general concepts are new terms that can be used by the matcher to widen the search in online ontologies (see Figure 4).

3.1 WSD techniques

To perform lexical annotation we use a combination of WSD algorithms of different natures. Ensemble methods are becoming more and more popular as they allow one to overcome the weaknesses of single approaches [10]. Different strategies can be applied such as majority voting, probability mixture, rank-based combination or maximum entropy combination. We chose to combine algorithms in a sequential composition based on their reliability: each algorithm has been

previously tested on several scenarios and its precision has been calculated; then, this precision value has been used to define the reliability of the algorithm.

We employ the ALA tool [2] to perform lexical annotation of the onologies involved in the matching process (source ontologies and online ontologies). With ALA we combine the output of four WSD algorithms (Structural Disambiguation algorithm, WordNet Domains Disambiguation algorithm, Gloss Similarity algorithm and Iterative Gloss Similarity algorithm) and two heuristic rules (Monosemic heuristic rule and WordNet first sense heuristic rule). We select a sequential composition to apply the WSD algorithms: only the first algorithm is executed on the entire data source, the following algorithms are executed only on the set of concepts that were not disambiguated by the previous ones.

4 Evaluation

The application of lexical annotation techniques on the SCARLET results has been tested on two test cases.

4.1 NALT and AGROVOC false positive mappings evaluation

The first test case was composed of real life thesauri [15]: the United Nations Food and Agriculture Organization (FAO)'s **AGROVOC** thesaurus, and the United States National Agricultural Library (NAL) Agricultural thesaurus **NALT**. On this scenario, a sample of 1000 mappings obtained by SCARLET has been manually validated, resulting in a promising 70% precision. Our evaluation has been performed on the 217 false positive mappings (the detection of them was previously done by a domain expert that knows the ontologies and their characteristics and is able to select correct mappings). After the lexical annotation of concepts involved in the anchoring, we discovered if the meanings are consistently linked or not, then we detected the false positive mappings.

We performed the automatic lexical annotation on each sub-ontology involved in the mapping and then, evaluated the results of the annotation on the anchoring. As previously explained in section 3, it was sufficient that the lexical annotation reveal that a concept has a meaning different from its anchoring, so that the anchoring is discharged and the mapping is revealed not valid. Thanks to the lexical annotation of the concepts, 12 out of 14 mappings have been recognized as false positive.

Unfortunately, not all the background ontologies were still available when we performed the test and some of them were not correctly written. At the end, the test case was not meaningful for the lexical annotation evaluation because it examines only 14 mappings.

4.2 OAEI evaluation

The second test case is based on the OAEI 2006 benchmark; this was the test case where SCARLET and another disambiguation method have been previ-

ously evaluated [15]. The benchmark[6] is bibliographic domain, the bibliographic ontologies we took into account are the reference ontology and the Karlsruhe ontology.

On this test case, we tested the lexical annotation techniques over both correct and incorrect anchoring to evaluate not only which wrong mappings are discharged after lexical annotation, but even, which true negative mappings are lost due to lexical annotation. For each matching found by SCARLET, we compared the meanings of the terms on the source ontologies with the meaning of the correspondent anchoring terms in the background ontologies. If both the couples have converging meanings, the anchoring is confirmed. If one couple has disregarding meanings, the anchoring is discharged. After lexical annotation the obtained results have been compared with the manual evaluation done by an expert on the entire set of matching.

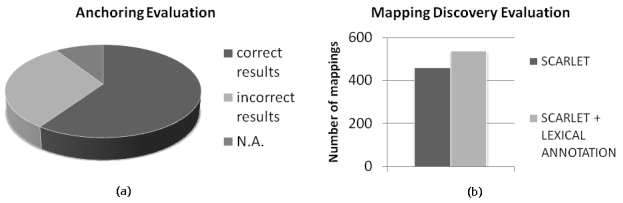

Fig. 5. Lexical annotation evaluation on OAEI 2006 benchmark: detection of incorrect anchoring (a) and new mapping discovery (b)

We examined a set of 109 mappings. The evaluation found agreement with the manual evaluation of the anchoring results in 65 cases (62 true positive anchoring and 3 true negative anchoring). The disagreement with the manual evaluation has been found in 34 cases (in these cases our algorithm retrieved 25 false positive and 9 false negative). 10 cases were impossible to disambiguate. A graphical representation is shown in figure 5 (a).

Moreover, we compared our results with a multiontology disambiguation method [18] that has been applied on the SCARLET matcher [7] and evaluated on the OAEI test case [15]. Because the multiontology disambiguation method retrieves similarity measures, the comparison of these two disambiguation methods permitted to evaluate some possible threshold on the similarity measures (we retrieved a threshold for 0.19).

On the OAEI scenario we also evaluated how lexical annotation can enrich the matcher results proving new relationships. After the lexical annotation of the OAEI sources each concept has one or more meanings associated. Exploring the WordNet network, we computed a mapping between two concepts, if a relationships exists between their meanings in WordNet. Some of these mappings confirmed the relationships found by SCARLET (we retrieved 18 mappings that

[6] available at http://oaei.ontologymatching.org/2006/benchmarks/

confirm the SCARLET results), and some other were new mappings that enrich the matcher results (we retrieved 77 new mappings, with a precision of 0.75%). Figure 5 (b) reports the improvements yielded by the lexical annotation.

5 Conclusion and Future Work

In this paper we described and experimentally investigated the application of automatic lexical annotation techniques in order to solve the ambiguity problems and to improve the results obtained by a matcher. The method has been applied on the SCARLET matcher, a semantic web based matcher which discovers mappings between two concepts by making use of online ontologies. Nevertheless, the method could be coped with any matcher.

We adopted an experimental approach on two test cases where SCARLET was previously tested, to investigate the potentiality of our method. The results confirmed our initial hypothesis (the precision is increased by solving ambiguity problems and new relationships are discovered), thus proving the value of the approach.

As mentioned in section 3, Lexical Annotation is able to identify synonymous and generalization of concepts, giving the matcher the possibility to widen the search among online ontologies, this will be the focus of our future work. This paper constitutes the kernel of an automatic lexical annotator for real world schemata/ontologies. To cope with complex schemata/ontologies, our method needs to be extended by including the treatment of compound terms and abbreviations [16].

6 Acknowledgements

The research reported in the paper was developed during a research period at KMI (Knowledge and Media Institute at Open University, Milton Keynes UK). We thank Marta Sabau, Enrico Motta, Jorge Gracia for their help with the evaluation reported in this paper. This work was partially supported by MUR FIRB Network Peer for Business project (http://www.dbgroup.unimo.it/nep4b) and by the IST FP6 STREP project 2006 STASIS (http://www.dbgroup.unimo.it/stasis).

References

1. D. Beneventano, S. Bergamaschi, F. Guerra, and M. Vincini. Synthesizing an integrated ontology. *IEEE Internet Computing*, pages 42–51, Sep-Oct 2003.
2. S. Bergamaschi, L. Po, S. Sorrentino, and A. Corni. Dealing with uncertainty in lexical annotation. In *ER Demo Sessions*, 2009.
3. S. Castano, A. Ferrara, and S. Montanelli. Matching ontologies in open networked systems: Techniques and applications. pages 25–63, 2006.
4. N. Choi, I.-Y. Song, and H. Han. A survey on ontology mapping. *SIGMOD Record*, 35(3):34–41, 2006.

5. L. Ding, R. Pan, T. W. Finin, A. Joshi, Y. Peng, and P. Kolari. Finding and ranking knowledge on the semantic web. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 156–170. Springer, 2005.
6. F. Giunchiglia, P. Shvaiko, and M. Yatskevich. Semantic schema matching. In R. Meersman, Z. Tari, M.-S. Hacid, J. Mylopoulos, B. Pernici, Ö. Babaoglu, H.-A. Jacobsen, J. P. Loyall, M. Kifer, and S. Spaccapietra, editors, *OTM Conferences (1)*, volume 3760 of *Lecture Notes in Computer Science*, pages 347–365. Springer, 2005.
7. J. Gracia, V. Lopez, M. d'Aquin, M. Sabou, E. Motta, and E. Mena. Solving semantic ambiguity to improve semantic web based ontology matching. In P. Shvaiko, J. Euzenat, F. Giunchiglia, and B. He, editors, *OM*, volume 304 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.
8. N. Jian, W. Hu, G. Cheng, and Y. Qu. Falconao: Aligning ontologies with falcon. In B. Ashpole, M. Ehrig, J. Euzenat, and H. Stuckenschmidt, editors, *Integrating Ontologies*, volume 156 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2005.
9. J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with cupid. In P. M. G. Apers, P. Atzeni, S. Ceri, S. Paraboschi, K. Ramamohanarao, and R. T. Snodgrass, editors, *VLDB*, pages 49–58. Morgan Kaufmann, 2001.
10. R. Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2), 2009.
11. M. Sabou, M. d'Aquin, and E. Motta. Using the semantic web as background knowledge for ontology mapping. In P. Shvaiko, J. Euzenat, N. F. Noy, H. Stuckenschmidt, V. R. Benjamins, and M. Uschold, editors, *Ontology Matching*, volume 225 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2006.
12. M. Sabou, M. d'Aquin, and E. Motta. Exploring the semantic web as background knowledge for ontology matching. *J. Data Semantics*, 11:156–190, 2008.
13. M. Sabou, M. Dzbor, C. Baldassarre, S. Angeletou, and E. Motta. Watson: A gateway for the semantic web. In *Poster session of the European Semantic Web Conference, ESWC*, 2007.
14. M. Sabou and J. Gracia. Spider: bringing non-equivalence mappings to oaei. In *Third International Workshop On Ontology Matching (OM2008)*, October 2008.
15. M. Sabou, J. Gracia, S. Angeletou, M. d'Aquin, and E. Motta. Evaluating the semantic web: A task-based approach. In K. Aberer, K.-S. Choi, N. F. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, editors, *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 423–437. Springer, 2007.
16. S. Sorrentino, S. Bergamaschi, M. Gawinecki, and L. Po. Schema normalization for improving schema matching. In A. H. F. Laender, S. Castano, U. Dayal, F. Casati, and J. P. M. de Oliveira, editors, *ER*, volume 5829 of *Lecture Notes in Computer Science*, pages 280–293. Springer, 2009.
17. H. Stuckenschmidt, L. Serafini, and H. Wache. Reasoning about ontology mappings. In *Technical Report, ITC-IRST, Trento*, 2005.
18. R. Trillo, J. Gracia, M. Espinoza, and E. Mena. Discovering the semantics of user keywords. *J. UCS*, 13(12):1908–1935, 2007.