Exploiting Relation Extraction for Ontology Alignment

Elena Beisswanger

Jena University Language and Information Engineering (JULIE) Lab,
Friedrich-Schiller-Universität Jena,
Fürstengraben 30, 07743 Jena, Germany
elena.beisswanger@uni-jena.de

Abstract. When multiple ontologies are used within one application system, aligning the ontologies is a prerequisite for interoperability and unhampered semantic navigation and search. Various methods have been proposed to compute mappings between elements from different ontologies, the majority of which being based on various kinds of similarity measures. As a major shortcoming of these methods it is difficult to decode the semantics of the results achieved. In addition, in many cases they miss important mappings due to poorly developed ontology structures or dissimilar ontology designs. I propose a complementary approach making massive use of relation extraction techniques applied to broad-coverage text corpora. This approach is able to detect different types of semantic relations, dependent on the extraction techniques used. Furthermore, exploiting external background knowledge, it can detect relations even without clear evidence in the input ontologies themselves.

Keywords: Ontology Alignment, Relation Extraction, Wikipedia

1 Background and Problem Statement

Ontologies specify the major terms and concepts (also called classes) of a domain and their relations in a formal manner. An increasing number of information systems in different application domains rely on ontologies to organize data. While in case of the Semantic Web they are used to define the semantics of (Web) documents, in biomedicine they serve as vocabulary to semantically annotate huge literature collections and factual data stores. In biomedical natural language processing (bio-NLP), in turn, ontologies support (amongst others) information extraction and semantic search applications.

However, especially in the field of biomedicine conceptual knowledge is scattered over various different, often disconnected ontologies. While some of them topically overlap (such as two different anatomy ontologies), others complement each other rather by design (such as ontologies for anatomical structures, cells, proteins, biological processes, drugs and diseases) [19]. Both, extraction patterns and search queries easily transcend the conceptual coverage of a single ontology. As a consequence, missing links between ontologies hamper effective information extraction and search, besides generally limiting data interoperability. The process of linking related ontology elements (*viz.*, classes, relations, and class instances) is called ontology alignment (OA) (cf. [8]).

OA has become an active field of research. Various methods have been proposed, most of them grounded in the intuition that elements with similar features (string-based, structural, extensional or semantics-based ones, cf. [8]) tend to be semantically related. Typically, given a certain similarity measure, similarity values are computed for pairs of elements and a threshold is chosen to decide which value is needed for a pair to be accepted as being "semantically related".

However, besides being sensitive to differing naming conventions and poorly developed or dissimilar ontology structures (cf., e.g., [1]), a major drawback of many similarity-based approaches is that the interpretation of their results is rather difficult. This applies both, to the type of relation existing between elements found to be similar (commonly, an equivalence relation is assumed[1]), and the similarity scores themselves. As a consequence, the lack of clear semantics hampers the incorporation of such alignments in reasoning applications and cross-ontology search. Generally, for the alignment of (complementary) biomedical ontologies other relation types than equivalence are critical, for example, *subClassOf*, *partOf*, and less common ones, such as *locatedIn*, *treats*, or *regulates*.

A completely different approach to OA with the potential to detect many different types of semantic relations is looking for relation evidences not within the given ontologies themselves, but in large-size, broad-coverage text corpora. This requires both, a suitable text corpus and an appropriate relation extraction (RE) machinery. Regarding the latter, in the field of NLP, there is a large body of work that could be exploited, targeting the extraction of various different relation types from text (cf., e.g., [11,9,5]). Concerning the text corpus, Wikipedia excels as a good candidate, for several reasons. First, it is a huge conglomerate of collaboratively assembled encyclopedic knowledge that currently seems to be unmatched in its size, broad coverage and up-to-dateness. Second, besides the free-text parts packed with definition phrases, it comes with a wide range of additional, more structured relation sources that could be used to support and complement results from free-text-based RE. These include the Wikipedia infoboxes, holding a multitude of conceptual relations in terms of implicit subject-predicate-object triples, the category system and articles linked to it, forming a huge concept graph with untyped semantic relations as edges, and the cross-links between articles, representing association-type like relations.

Along these lines the following research questions were derived:

1. How can established RE approaches contribute to the alignment of ontologies? How can they be adapted to the alignment use case? In particular, how can ontology class mentions be detected in text?[2]
2. Given Wikipedia as data source, how can relations extracted from free-text parts be integrated with relations extracted from structured parts of articles?
3. How can corpus-based ontology alignment methods be evaluated, in particular if they target relations other than equivalence, such as *subClassOf*?

[1] Very few systems also detect *subClassOf* relations (cf., e.g., [7]).

[2] Note that in my work I focus on the alignment of ontology classes only.

In the following I will discuss related work, outline expected contributions, present my working plan and conclude with an overview on the current state of my work and the next steps to be taken.

2 Related Work

Automatic ontology alignment is hampered by the fact that in many existing ontologies the meaning of classes and relations is insufficiently specified. To compensate for this shortcoming, alignment approaches have been developed incorporating various kinds of external background knowledge (e.g., [1,16,22]). Structured resources, such as ontologies (e.g., [1]) or WordNet (e.g., [16]) are preferably used, due to their easily accessible semantics. However, their coverage is generally limited and for many domains such resources lack completely. The opposite is the case for unstructured text. It is available in large quantities across many domains. However, relations are hidden in natural language phrases and an appropriate NLP system is required to access them. I am aware of only few alignment approaches exploiting relation extraction from text. One example is the work by van Hage et al., experimenting with basic linguistic methods (Hearst pattern matching on the Web and parsing definition phrases from an online dictionary) to discover *subClassOf* relations in the domain of food [22].

In **Ontology learning** (OL), a neighboring field of OA, relation extraction from text is much more common. OL is concerned with the automatic construction (or extension) of ontologies from given data sets, such as text corpora or databases. A typical text-based OL system extracts relevant terms and term variants, groups them to concepts and subsequently identifies *subClassOf* relations forming the backbone of the ontology (cf., [6]). The population of ontologies with instances is also widespread. For example, the Sofie Framework extracts facts from free-text parts of Wikipedia articles to extend ontologies with instance data [21]. Only few systems go further and extract other relations than *subClassOf* and *instanceOf* (cf., [23]). In the case of ontology extension, as in the case of OA, a major challenge is to recognize the linguistic appearance of known concepts in text.

Concept recognition comprises two (not necessarily separate) steps: candidate detection and candidate disambiguation. While the first step influences recall of the RE procedure, the second one, tackling lexical ambiguity arising from homonymy and polysemy of words, has an impact on precision. Several concept recognition tools have been released, most of them relying on matching concept labels against text. The techniques used range from simple string matching procedures to advanced forms incorporating detailed linguistic analysis and synonym enrichment, as in the case of MetaMap (a system frequently used in the field of bio-NLP) [2]. Some terms in text qualify as mapping target for more than one concept. While simpler systems typically enumerate all candidates, more sophisticated solutions employ word sense disambiguation (WSD) techniques to identify the correct mapping (for a comprehensive survey, cf., [14]).

Recently some new WSD approaches have been proposed exploiting Wikipedia specific information, such as page links and disambiguation pages (cf., e.g., [12]).

Automatic extraction of semantic relations from text is a broad research field in NLP. A plethora of statistical, rule-based, and machine learning-based approaches has been proposed targeting different types of relations, ranging from hypernymy (cf., e.g., [11,20,17]) and meronymy (cf., e.g., [9]), denoting the *subClassOf* and *partOf* relation on the linguistic level, to domain-specific relations (cf., e.g., [5]). The first version of the alignment system I am developing will focus on the detection of *subClassOf* relations between classes. Thus I am particularly interested in work on automatic hypernym extraction. Most common approaches either rely on lexico-syntactic patterns (cf., e.g., [11,20]), or exploit the distributional similarity or co-occurrence of terms (cf., e.g., [17]). As a unique feature, pattern-based approaches detect hypernymy relations explicitly mentioned in text. While Hearst utilizes a small set of hand crafted patterns (such as "term_1 *is a* term_2") [11], Snow et al. achieve a major improvement in recall by automatically deriving a much larger set of patterns from text and using them as features in a machine learning approach [20].

In recent years, **Wikipedia** has become a popular resource for RE and other NLP tasks and applications (for a survey, cf., [12]). So far relation extraction efforts mainly concentrate on structured facets of Wikipedia, such as infoboxes, page links, and the category system. Amongst others, Ponzetto and Strube created a taxonomy based on the Wikipedia category system by refining the previously untyped semantic relations [15]. Bizer et al. built DBpedia, consisting of over 4.5 million RDF triples mainly derived from Wikipedia infobox templates [4]. Recently WikiNet was published, a collection of 3 million concepts and over 36 million relations mainly extracted from Wikipedia infoboxes and the category system [13]. Both, results and extraction machinery of some of these projects have been made publicly available. Fewer efforts target the full-text body of Wikipedia articles, an example is [21].

3 Expected Contributions

1. The main contribution of my work will be an **ontology alignment system** exploiting conceptual relations entangled in unstructured and structured parts of a huge text corpus, *viz.* Wikipedia. While the current version is restricted to the extraction of *subClassOf* relations from free-text, two extensions of the system are scheduled: the detection of other types of semantic relations critical for the alignment of biomedical ontologies and the incorporation of relations extracted from structured parts of Wikipedia.
2. The UIMA-based[3] **NLP tools** I am developing for the analysis of Wikipedia articles are designed to work independently from the alignment system. Thus, they can be deployed in other application scenarios, too.

[3] Unstructured Information Management Architecture (`http://uima.apache.org/`)

3. Amongst the few existing RE-based alignment systems, my system will be distinguished by a proper **concept recognition** step. To assess the state-of-the-art in concept recognition, a thorough investigation of existing approaches will be carried out. Concept recognition also is a key issue in other tasks involving both, ontologies and textual data (such as semantic search, semantic annotation of text, or text-based OL). Thus, the intended study is of potential interest even outside the OA community.
4. Finally, my work will cover **evaluation strategies** for alignments holding relations of other types than equivalence (a first step in this direction was taken in terms of the "Oriented matching" task of the 2009 Campaign of the Ontology Alignment Evaluation Initiative [7]), as well as an **alignment algorithm** that can even cope with large-sized ontologies, avoiding an exhaustive analysis of class and label pairs.

4 Alignment System Design and Development

The alignment system will consist of the following components:

1. **Alignment algorithm**. It decides on which class comparisons to be made, queries the index, filters the query result, invokes the RE module and the relation repository, and integrates results.
2. **Lucene index**. The index contains Wikipedia articles, sentence-wise.
3. **Relation extraction module**. The RE module extracts relations between class pairs from free-text parts of Wikipedia.
4. **Relation repository**. The relation repository contains relations extracted from structured parts of Wikipedia.
5. **Result store**. The result store saves the output of the RE module across different runs of the system, to avoid duplicate work.

Procedure. The alignment of two ontologies will proceed as follows. First, the system imports the ontologies. Next, the alignment algorithm starts selecting pairs of classes to be compared. For each selected class pair that is not yet in the result store label pairs are formed. For each label pair that is not yet in the result store the index is queried for sentences containing normalized forms of both labels. A filtering step eliminates those sentences in which at least one label refers to a wrong word sense. The remaining sentences and the label pair are handed over to RE module, which, in turn, searches the sentences for relation evidences. Sentences with overlapping labels are dealt with separately. Next, the relation repository is queried for additional relation evidences. Based on all evidences found, the alignment algorithm decides whether the class pair is related or not. Results for newly analyzed class and label pairs are saved in the result store. If no more class pairs need to be analyzed, results are integrated and cleaned up and the final alignment is exported.

Relation extraction. The RE module of the alignment system will incorporate a dependency feature-based relation classifier, similar to the one proposed in [20]. In the current version, it predicts *subClassOf* relations only (as in [20]).

However, in principle the classifier could be enabled to detect also other relation types, given that sufficient train and test data is available (see Section 5). As baseline for the *subClassOf* extraction, a second RE module is used, relying on the original Hearst patterns [11] (in this respect, it is similar to [22]).

In the alignment system, relation extraction is preceded by a two-stage **concept recognition** step. First, class labels are detected in text by means of an extended string matching procedure. It involves lower-casing, stop word removal, removal of special characters, stemming, and a filtering step evaluating part-of-speech tags and syntactic information of class labels and text. In the second stage (which is not implemented yet), ambiguities will be resolved considering both, the context of candidate classes in the respective ontology (e.g., the labels of adjacent classes and relations, as in [21]) and Wikipedia specific information. For the latter, existing Wikipedia-based WSD approaches will be evaluated.

Alignment algorithm. The overall alignment process is governed by the alignment algorithm. There are two major tasks to perform: the selection of class pairs to be analyzed, and the decision about mappings between classes. For the first task, a brute force approach is used in the current version of the system. All classes in one ontology are compared to all classes in the other one, considering all possible label pairs. Since for large ontologies this implies high computational costs, the adoption of a new, optimized procedure is scheduled. Inspired by existing work (such as the Anchor-Flood algorithm, looking for mappings between previously defined blocks of similar classes only [18]), for the selection of class pairs it will consider both, the structure of the input ontologies and already computed mappings. The second task is solved based on two input streams: relation evidences from free-text parts of Wikipedia articles delivered by the RE module, and relation evidences originating from structured parts as they will be available from the relation repository. While a first version of the RE module is already in place, the relation repository is still pending.

5 Status and Next Steps

So far I have developed an UIMA-based NLP application for relation extraction from Wikipedia. It comprises two text processing pipelines (one for creating the Wikipedia index required by the OA system, the other for the RE task itself), and a scheduling system that allows to run several pipeline instances simultaneously (a prerequisite to efficiently process a large data collection such as Wikipedia). Besides existing text processing components from JCoRe [10] (sentence splitter, tokenizer, POS tagger, chunker, etc.), the two pipelines include the following newly developed components: a UIMA Collection Reader for Wikipedia [3] (it makes Wikipedia articles accessible for subsequent UIMA analytics by parsing the MediaWiki mark-up and filtering relevant contents), a new indexing component, a UIMA-based Hearst pattern matcher, and a second, more advanced RE module, incorporating a dependency feature-based relation classifier. To build the classifier, I basically parsed sentences extracted from Wikipedia abstracts (the first paragraph, before the table of contents), extracted noun phrases as anchor pairs, labeled them as being hyponym/hypernym pairs in WordNet[4], extracted the dependency paths between all anchor pairs, took “frequent” paths as features, generated feature vectors for the labeled anchor pairs (taking the frequency of occurrence of a path between an anchor pair as feature value), and trained the classifier with all anchor pairs of which the feature vector contained a minimum number of non-zero values. Currently, a manual *subClassOf* annotation project is running that will deliver gold standard data required to evaluate the RE modules.

There are two immediate next steps when the gold standard has been completed. First, the performance of the RE module will be evaluated. Second, the concept recognition step will be refined (which precedes the actual relation extraction) by implementing the scheduled disambiguation stage. The evaluation will be rerun to assess which impact it has on RE results. Thereafter, the next major steps will be to enhance the alignment algorithm and to prepare the RE module for the detection of new relation types. For each relation type, train and test data must be provided to retrain and evaluate the included classifier. Finally, the relation repository will be populated with relations extracted from structured parts of Wikipedia (e.g., incorporating harmonized results of [15,4,13]).

In conclusion, this doctoral project lies at the junction of two different avenues of research, *viz.* NLP-based relation extraction and ontology alignment. The main challenge is to properly integrate these currently almost unrelated approaches, in order to open up the rich reservoir of conceptual relations entangled in natural language texts for OA. Furthermore, it requires to respond to the methodological requirements of aligning concrete, large-sized (bio-)ontologies. Up until now, I have implemented a first simple version of an alignment system working along these lines. Although it still lacks many of the envisaged sophisticated features, it can already discover *subClassOf* relations between ontology classes applying a well established RE approach to the English Wikipedia.

Acknowledgments

I would like to thank my supervisors Udo Hahn and Heiner Stuckenschmidt for their assistance, and the anonymous reviewers for their helpful comments.

References

1. Aleksovski, Z., Klein, M.C.A., ten Kate, W., van Harmelen, F.: Matching unstructured vocabularies using a background ontology. In: Proceedings of the EKAW 2006 conference. pp. 182–197 (2006)
2. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: The METAMAP program. In: Proceedings of the AMIA 2001. pp. 17–21 (2001)
3. Beisswanger, E., Hahn, U.: JULIE Lab’s UIMA Collection Reader for Wikipedia. In: Proceedings of the LREC 2010 workshop on New Challenges for NLP Frameworks. pp. 15–19 (2010)

[4] `http://wordnet.princeton.edu/`

4. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - a crystallization point for the web of data. Web Semantics 7(3), 154–165 (2009)
5. Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I.: Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In: Proceedings of the IJCAI 2005. pp. 659–664 (2005)
6. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag (2006)
7. Euzenat, J., Ferrara, A., Hollink, L., Isaac, A., Joslyn, C., Malaisé, V., Meilicke, C., Nikolov, A., Pane, J., Sabou, M., Scharffe, F., Shvaiko, P., Spiliopoulos, V., Stuckenschmidt, H., Sváb-Zamazal, O., Svátek, V., dos Santos, C.T., Vouros, G.A., Wang, S.: Results of the Ontology Alignment Evaluation Initiative 2009. In: Proceedings of the ISWC 2009 workshop on Ontology Matching (2009)
8. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag (2007)
9. Girju, R., Badulescu, A., Moldovan, D.: Automatic discovery of part-whole relations. Computational Linguistics 32(1), 83–135 (2006)
10. Hahn, U., Buyko, E., Landefeld, R., Mühlhausen, M., Poprat, M., Tomanek, K., Wermter, J.: An overview of JCoRe, the JULIE Lab UIMA component repository. In: Proceedings of the LREC 2008 workshop on UIMA for NLP. pp. 1–7 (2008)
11. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the ACL 1992 conference. pp. 539–545 (1992)
12. Medelyan, O., Milne, D., Legg, C., Witten, I.H.: Mining meaning from Wikipedia. International Journal of Human-Computer Studies 67(9), 716–754 (2009)
13. Nastase, V., Strube, M., Boerschinger, B., Anas, E.: WikiNet: A very large scale multi-lingual concept network. In: Proceedings of the LREC 2010 (2010)
14. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys 41(2), 1–69 (2009)
15. Ponzetto, S.P., Strube, M.: Deriving a large scale taxonomy from Wikipedia. In: Proceedings of the AAAI 2007 conference. pp. 1440–1445 (2007)
16. Reynaud, C., Safar, B.: Exploiting WordNet as background knowledge. In: Proceedings of the ISWC 2007 workshop on Ontology Matching (2007)
17. Sanderson, M., Croft, W.B.: Deriving concept hierarchies from text. Proceedings of the SIGIR 1999 conference pp. 206–212 (1999)
18. Seddiqui, M.H., Aono, M.: An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size. Web Semantics 7(4), 344–356 (2009)
19. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.E.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25(11), 1251–1255 (2007)
20. Snow, R., Jurafsky, D., Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. In: Advances in Neural Information Processing Systems 17. pp. 1297–1304. MIT Press (2005)
21. Suchanek, F.M., Sozio, M., Weikum, G.: SOFIE: A Self-Organizing Framework for Information Extraction. In: Proceedings of the WWW 2009 conference (2009)
22. Van Hage, W.R., Katrenko, S., Schreiber, G.: A method to combine linguistic ontology-mapping techniques. In: Proceedings of ISWC 2005. pp. 732–744 (2005)
23. Völker, J., Haase, P., Hitzler, P.: Learning expressive ontologies. In: Proceedings of the 2008 conference on Ontology Learning and Population. pp. 45–69 (2008)