SAMBO Results for the Ontology Alignment Evaluation Initiative 2007

He Tan and Patrick Lambrix

Department of Computer and Information Science
Linköpings universitet
581 83 Linköping, Sweden

Abstract. This article describes a system for ontology alignment, SAMBO, and presents its results for the benchmark and anatomy tasks in the 2007 Ontology Alignment Evaluation Initiative. For the benchmark task we have used a strategy based on string matching as well as the use of a thesaurus, and obtained good results in many cases. For the anatomy task we have used a combination of string matching and the use of domain knowledge. This combination performed well in former evaluations using other anatomy ontologies.

1 Introduction

Many ontologies have already been developed and many of these ontologies contain overlapping information. Often we would want to be able to use multiple ontologies. For instance, companies may want to use community standard ontologies and use them together with company-specific ontologies. Applications may need to use ontologies from different areas or from different views on one area. Ontology builders may want to use already existing ontologies as the basis for the creation of new ontologies by extending the existing ontologies or by combining knowledge from different smaller ontologies. Further, different data sources in the same domain may have annotated their data with different but similar ontologies. In each of these cases it is important to know the relationships between the terms in the different ontologies. It has been realized that this is a major issue and some organizations have started to deal with it. For instance, regarding anatomy ontologies there is the CARO (http://www.bioontology.org/wiki/-index.php/CARO:Main_Page) effort and earlier the SOFG effort (http://www.sofg.org/).

To deal with this issue we developed and continue developing SAMBO, System for Aligning and Merging Biomedical Ontologies. We use the term 'alignment' for defining the relationships between terms in different ontologies. We use the term 'merging' when we, based on the alignment relationships between ontologies, create a new ontology containing the knowledge included in the source ontologies. In the remainder of the paper we only discuss the alignment component of SAMBO.[1] In section 2 we describe the purpose, the framework on which SAMBO is based, the techniques used, and the adaptations made for OAEI 2007. Section 3 describes the test runs and general comments are given in section 4. The paper concludes in section 5.

[1] SAMBO also merges two source ontologies in OWL syntax with given alignment relationships using a reasoner.

2 Presentation of the system

2.1 State, purpose, general statement

Although several of our methods and techniques are general and applicable to different areas, when developing SAMBO, we have focused on biomedical ontologies. Research in biomedical ontologies is recognized as essential in some of the grand challenges of genomics research [2]. Further, there exist de facto standard ontologies such as GO, and much support is being provided to the community to develop and publish ontologies in the biomedical domain in a principled way through, for instance, the OBO Foundry initiative (http://www.obofoundry.org/). There are also many overlapping ontologies available in the field, many of which are available through OBO. The field has also matured enough to start tackling the problem of overlap in the ontologies and standardization efforts such as SOFG and CARO have started.

Ontologies may contain concepts, relations, instances and axioms. Most biomedical ontologies are controlled vocabularies, taxonomies, or thesauri. This means that they may contain concepts, is-a and part-of relations, and sometimes a limited number of other relationships. Therefore, we have focused on methods that are based on these ontology components. For some approaches we have also used documents about a concept as instances for that concept. We have not dealt with axioms.

2.2 Framework

SAMBO is based on the framework shown in figure 1 [5]. The framework consists of two parts. The first part (*I* in figure 1) computes alignment suggestions. The second part (*II*) interacts with the user to decide on the final alignments. An alignment algorithm receives as input two source ontologies. The algorithm includes one or several matchers, which calculate similarity values between the terms from the different source ontologies. The matchers may use knowledge from different sources. Alignment suggestions are then determined by combining and filtering the results generated by one or more matchers. By using different matchers and combining and filtering the results in different ways we obtain different alignment strategies. The suggestions are then presented to the user who accepts or rejects them. The acceptance and rejection of a suggestion may influence further suggestions. Further, a conflict checker is used to avoid conflicts introduced by the alignment relationships. The output of the alignment algorithm is a set of alignment relationships between terms from the source ontologies.

2.3 Specific techniques used

In this section we describe the matchers, and combination and filtering techniques that are available in SAMBO. These matchers and techniques were previously evaluated using test cases for aligning Gene Ontology and Signal Ontology, and for aligning Medical Subject Headings (MeSH) and the Anatomical Dictionary for the Adult Mouse (MA) [5] using the KitAMO evaluation environment [7].[2] In addition to these techniques we

[2] An introduction to SAMBO and KitAMO can be found in [6].

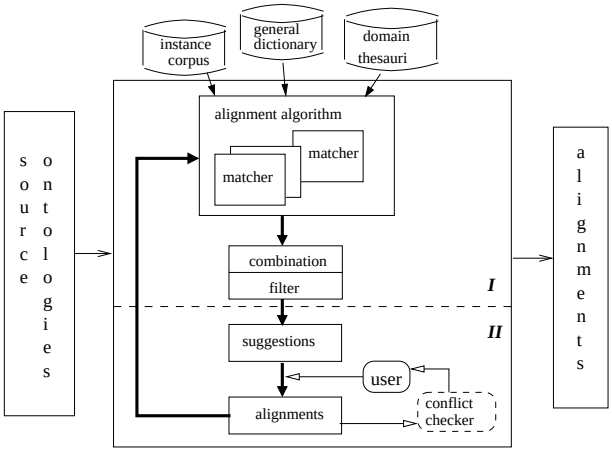

Fig. 1. Alignment framework [5].

have also experimented with other matchers [9, 11] and another filtering technique [1], some of which may be added to SAMBO in the future. We are also working on methods for recommendation of alignment strategies [10] which we intend to integrate into SAMBO in the future.

Matchers SAMBO contains currently five basic matchers: two terminological matchers, a structure-based matcher, a matcher based on domain knowledge, and a learning matcher.

Terminological matchers. The basic terminological matcher, *Term* contains matching algorithms based on the textual descriptions (names and synonyms) of concepts and relations. In the current implementation, the matcher includes two approximate string matching algorithms, n-gram and edit distance, and a linguistic algorithm. An n-gram is a set of n consecutive characters extracted from a string. Similar strings will have a high proportion of n-grams in common. Edit distance is defined as the number of deletions, insertions, or substitutions required to transform one string into the other. The greater the edit distance, the more different the strings are. The linguistic algorithm computes the similarity of the terms by comparing the lists of words of which the terms are composed. Similar terms have a high proportion of words in common in the lists. A Porter stemming algorithm is employed to each word. These algorithms were evaluated in [4] using MeSH anatomy (ca 1400 terms) and MA (ca 2350 terms). Term computes similarity values by combining the results from these three algorithms using a weighted sum. The combination we use in our experiments (weights 0.37, 0.37 and 0.26 for the linguistic algorithm, edit distance and n-gram, respectively) outperformed the individual individual algorithms in our former evaluations [4]. Further, the matcher *TermWN* is based on Term, but uses a general thesaurus, WordNet (http://wordnet.princeton.edu/), to enhance the similarity measure by looking up the hypernym relationships of the pairs of words in WordNet.

Structural matcher. The structural matcher is an iterative algorithm based on the is-a and part-of hierarchies of the ontologies. The algorithm requires as input a list of alignment relationships and similarity values and can therefore not be used in isolation. The intuition behind the algorithm is that if two concepts lie in similar positions with respect to is-a or part-of hierarchies relative to already aligned concepts in the two ontologies, then they are likely to be similar as well. For each pair of concepts (C_1,C_2) in the original list of alignment relationships the structural matcher augments the original similarity value for pairs of concepts (C_1',C_2') such that C_1' and C_2' are equivalent to, are in an is-a relationship with, or participate in a part-of relationship with C_1 and C_2, respectively. The augmentation depends on the relationship and on the distance between the concepts in the is-a and part-of hierarchies. The augmentation diminishes with respect to distance. The new similarity value can also not exceed 1. In our earlier experiments we used a maximal distance of 2 and the effect on ancestors is lower than the effect on descendants.

Use of domain knowledge. Another strategy is to use domain knowledge. Our matcher *UMLSKSearch* uses the Metathesaurus in the Unified Medical Language System (UMLS, http://www.nlm.nih.gov/research/umls/). The similarity of two terms in the source ontologies is determined by their relationship in UMLS. In our experiments we used the UMLS Knowledge Source Server to query the UMLS Metathesaurus with source ontology terms. The querying is based on searching the normalized string index and normalized word index provided by the UMLS Knowledge Source Server. We used version 2007AB of UMLS. As a result we obtain concepts that have the source ontology term as their synonym. We assign a similarity value of 1 if the source ontology terms are synonyms of the same concept and 0 otherwise.[3]

Learning matcher. The matcher makes use of life science literature that is related to the concepts in the ontologies. It is based on the intuition that a similarity measure between concepts in different ontologies can be defined based on the probability that documents about one concept are also about the other concept and vice versa. The strategy contains the following basic steps. (i) For each ontology that we want to align we generate a corpus of PubMed abstracts. In our implementation we generated a corpus of maximally 100 PubMed abstracts per concept using the programming utilities provided by the retrieval system Entrez (http://www.ncbi.nlm.nih.gov/Entrez/). (ii) For each ontology a document classifier is generated. This classifier returns for a given document the concept that is most closely related to the document. To generate a classifier the corpus of abstracts associated to the classifier's ontology is used. In our algorithm we use a naive Bayes classification algorithm (based on the code available at http://www.cs.utexas.edu/users/mooney/ir-course/). (iii) Documents of one ontology are classified by the document classifier of the other ontology and visa versa. (iv) A similarity measure between concepts in the different ontologies is computed by using the results of step (iii). The similarity is computed as

$$lsim(C_1, C_2) = \frac{n_{NBC2}(C_1, C_2) + n_{NBC1}(C_2, C_1)}{n_D(C_1) + n_D(C_2)}$$

[3] Observe that this is slightly different from the version reported in [5] where we used version 2005AA of UMLS and we assigned a similarity value of 1 for two terms with the exact same names, 0.6 if the source ontology terms are synonyms of the same concept, and 0 otherwise.

where $n_D(C)$ is the number of abstracts originally associated with C, and $n_{NBCx}(C_p, C_q)$ is the number of abstracts associated with C_p that are also related to C_q as found by classifier $NBCx$ related to ontology x. More details about this algorithm as well as some extensions can be found in [9].

Combinations The user is given the choice to employ one or several matchers during the alignment process. The similarity values for pairs of concepts can then be determined based on the similarity values computed by one matcher, or as a weighted sum of the similarity values computed by different matchers.

Filtering The current filtering method is threshold filtering. Pairs of concepts with a similarity value higher than or equal to a given threshold value are returned as alignment suggestions to the user.

2.4 Adaptations made for the evaluation

SAMBO is an interactive alignment system. The alignment suggestions calculated by SAMBO are normally presented to the user who accepts or rejects them. Alignment suggestions with the same concept as first item in the pair are shown together to the user. Therefore, SAMBO shows the user the different alternatives for aligning a concept. This is a useful feature, in particular when the system computes similarity values which are close to each other and there is no or only a small preference for one of the suggestions. Further, the acceptance and rejection of a suggestion may influence which suggestions are further shown to the user.

The computation of the alignment suggestions in SAMBO is based on the computation of a similarity value between the concepts. The computation of the similarity values does not take into account what the relationship of the alignment should be. However, when an alignment is accepted, the user can choose whether the alignment relationship should be an equivalence relation or an is-a relation.

As the OAEI evaluation only considers the non-interactive part of the system and the computation of the similarity values does not take the relationship into account, we had to modify the computation of the suggestions. It would not make sense to have alignment suggestions where a concept appears more than once as the user would not be able to make a choice. Therefore, we decided to filter SAMBO's alignment suggestion list such that only suggestions are retained where the similarity between the concepts in the alignment suggestion is higher than or equal to the similarity of these concepts to any other concept according to the alignment suggestion list. (In the case there are different possibilities, one is randomly chosen.)

2.5 Link to the system and parameters file

The SAMBO project page is at http://www.ida.liu.se/∼iislab/projects/SAMBO/.

2.6 Link to the set of provided alignments (in align format)

The suggested alignments are available at
http://www.ida.liu.se/∼iislab/projects/SAMBO/OAEI/2007/.

3 Results

We have provided alignment suggestions for the tasks 'benchmark' and 'anatomy'. Tests were performed on a PC (Pentium(R) D CPU 2.80GHz 2.79GHz, RAM 0.99GB, Windows XP).

3.1 benchmark

The results for the benchmark task were obtained by using TermWN with threshold 0.6. As a preprocessing step we split names based on capital letters occurring within a name. For instance, 'InCollection' was split into 'In Collection'. We did not use the comment field. The results may be improved using also this field.

We assume that ontology builders use a reasonable naming scheme and thus we did not tackle the cases where labels were replaced by a random one. Therefore, the recall for tests 201-202, 248-254, 257-262, 265-266 is low. For these cases we may use other kind of information in the ontology such as the comment field or the structure. We also did not focus on different natural languages (206-207, 210) or subsumption relationships (302).

Regarding the other cases we received high precision and recall except for cases 205 and 209. For 205 and 209 we had expected that using WordNet would be an advantage. Therefore, we compared the results with a run using Term (without WordNet). The differences between the results for Term and TermWN were small for all cases, including cases 205 and 209.

3.2 anatomy

The results for the anatomy task were obtained by first running UMLSKSearch and suggesting the pairs with similarity value 1 and then running Term with threshold 0.6 on the remainder of the pairs. With respect to the computation of the suggestions, this would be similar to having a matcher that returns as similarity value for a pair the maximum of the similarity value for the pair according to UMLSKSearch and the similarity value for the pair according to Term, and then using 0.6 as threshold.

4 General comments

A problem that users face is that often it is not clear how to get the best alignment results given that there are many strategies to choose from. In most systems, including SAMBO) there usually is no strategy for choosing the matchers, combinations and filters in an optimal way. Therefore, we used our experience from previous evaluations [5] to decide which matchers to use for which task. The lack of an optimization strategy is also the reason why we did not provide results for the second and third test for anatomy (optimization with respect to precision and recall, respectively). The results for precision and recall for SAMBO may be influenced by the filtering phase. Intuitively, higher thresholds lead to higher precision and lower recall, while lower thresholds usually lead to higher recall, but lower precision. However, for SAMBO as stand-alone

system, there is no strategy for how to choose the threshold for optimizing precision or recall. In the future, however, this may be possible using recommendation methods for alignment strategies such as proposed in [10] that will be able to recommend matchers, combinations and filters based on the alignment task and evaluation methods.

The OAEI deals with the non-interactive part of the alignment systems. This allows for evaluating how good the alignment suggestions are. However, for some systems, such as SAMBO, the list of alignment suggestions is only an initial list and is updated after each acceptance or rejection of a suggestion.

5 Conclusion

We have briefly described our ontology alignment system SAMBO and some results of running SAMBO on the alignment tasks of OAEI.

For the benchmark task we have used TermWN and obtained good results in many cases. We expect that the results will still improve when we use more information available in the ontology, such as the comment field and the structure. Therefore, we will continue this task using Term and TermWN also on the comment field, as well as using our structural matcher. Further, in earlier tests, also our advanced filtering technique described in [1] usually improves the results of Term and TermWN.

Regarding the anatomy task we have used a combination of UMLSKSearch and Term, which performed best in former evaluations using other anatomy ontologies. We are currently also evaluating instance-based matchers.

A major problem is deciding which algorithms should be used for a given alignment task. This is a problem that users face, and that we have also faced in the evaluation. We expect that recommendation strategies [10, 8, 3] will alleviate this problem.

References

1. B Chen, H Tan, and P Lambrix. Structure-based filtering for ontology alignment. In *Proceedings of the IEEE WETICE Workshop on Semantic Technologies in Collaborative Applications*, pages 364–369, 2006.
2. F Collins, E Green, A Guttmacher, and M Guyer. A vision for the future of genomics research. *Nature*, 422:835–847, 2003.
3. M Ehrig, S Staab, and Y Sure. Bootstrapping ontology alignment methods with apfel. In *Proceedings of the International Semantic Web Conference*, pages 186–200, 2005.
4. P Lambrix and H Tan. Merging daml+oil ontologies. In J Barzdins and A Caplinskas, editors, *Databases and Information Systems - Selected Papers from the Sixth International Baltic Conference on Databases and Information Systems*, pages 249–258. IOS Press, 2005.
5. P Lambrix and H Tan. Sambo - a system for aligning and merging biomedical ontologies. *Journal of Web Semantics, Special issue on Semantic Web for the Life Sciences*, 4(3):196–206, 2006.
6. P Lambrix and H Tan. Ontology alignment and merging. In Burger, Davidson, and Baldock, editors, *Anatomy Ontologies for Bioinformatics: Principles and Practice*. Springer, 2007.
7. P Lambrix and H Tan. A tool for evaluating ontology alignment strategies. *Journal on Data Semantics, LNCS 4380*, VIII:182–202, 2007.
8. M Mochol, A Jentzsch, and J Euzenat. Applying an analytic method for matching approach selection. In *Proceedings of the Workshop on Ontology Matching*, 2006.

9. H Tan, V Jakonienė, P Lambrix, J Aberg, and N Shahmehri. Alignment of biomedical ontologies using life science literature. In *Proceedings of the International Workshop on Knowledge Discovery in Life Science Literature, LNBI 3886*, pages 1–17, 2006.
10. H Tan and P Lambrix. A method for recommending ontology alignment strategies. In *Proceedings of the 6th International Semantic Web Conference*, 2007.
11. T Wächter, H Tan, A Wobst, P Lambrix, and M Schroeder. A corpus-driven approach for design, evolution and alignment of ontologies. In *Proceedings of the Winter Simulation Conference*, pages 1595–1602, 2006. Invited contribution.