BLOOMS on AgreementMaker: results for OAEI 2010

Catia Pesquita[1], Cosmin Stroe [2], Isabel F. Cruz[2], Francisco M. Couto[1]

[1] Faculdade de Ciencias da Universidade de Lisboa, Portugal
cpesquita-at-xldb.di.fc.ul.pt, fcouto-at-di.fc.ul.pt
[2] ADVIS Lab, Department of Computer Science, University of Illionois at Chicago
cstroe1@cs.uic.edu, ifc@cs.uic.edu

Abstract. BLOOMS is an ontology matching method developed as part of an ontology extension system for biomedical ontologies. It combines two lexical similarity measures with similarity propagation. These matchers are applied sequentially, following their precision yield: first lexical similarity based on exact matches, followed by partial matches, and finally these similarities are propagated throughout the ontologies. Partial matches are based on the specificity of words within the ontologies vocabularies. Semantic propagation of similarities is made according to the semantic distance between ontology concepts given by semantic similarity measures. Alignments are extracted after each matcher, to favor precision, since BLOOMS was specifically designed to be as automated as possible. For the participation in OAEI 2010 BLOOMS was integrated into the AgreementMaker system, which provided ontology loading and navigation capabilities. We participated only in the anatomy track, in the tasks #1 and #2 (f-measure and precision), given that BLOOMS was specifically designed for the automated matching of biomedical ontologies. We obtained encouraging results with an f-measure of 0.828 for task #1 and a precision of 0.967 for task #2. Although the current implementation of BLOOMS results in very good precision values, recall is below that of the highest performing systems. This motivates our future work in improving our semantic propagation algorithm and exploiting external resources.

1 Presentation of the system

BLOOMS is an ontology matching method specifically intended for application to biomedical ontologies. The matching of biomedical ontologies has become a focus of interest in recent years due to the increasingly important role that biomedical ontologies are playing in the knowledge revolution that has swept the Life Sciences domain in the last decade. The pressing need for these resources resulted in the parallel development of ontologies by different groups and institutions, giving rise not only to different ontologies covering the same domain, but also to a lack of shared standards and logical links between related ontologies. The alignment of biomedical ontologies is thus crucial to take full advantage of them.

Biomedical ontologies present specific challenges and opportunities for their alignment. One relevant feature of many biomedical ontologies that hinders their alignment is their size, for instance the Gene Ontology contains over 30,000 concepts and ChEBI over 500,000. Many of the systems developed for other domains have

difficulty in handling such large ontologies. On the other hand, most biomedical ontologies support few types of relationships, which can hinder the performance of matchers that explore more complex structures. Also, in most biomedical ontologies edges do not all represent the same semantic distance between concepts, for instance, edges deeper in the ontology usually represent shorter distances than edges closer to the root concept.

Another relevant feature is the rich textual information in the form of concept names, synonyms and definitions that most biomedical ontologies have. This can play a crucial role in matching algorithms that exploit lexical resources but it can also be an obstacle since biomedical terminology has a high degree of ambiguity.

In recent years OAEI has been the major play field for biomedical ontologies alignment, in its anatomy track. One important finding of previous OAEI anatomy tracks is that several matches are rather trivial and can be found by simple string comparison techniques. Based on this notion, the work in [1] has applied a simple string matching algorithm to several ontologies available in the NCBO BioPortal, and reported high levels of precision in most cases. There are several possible explanations for this, including the simple structure of most biomedical ontologies, their high number of synonyms and low language variability. To improve on the results of simple string matching, the most successful systems in previous OAEI editions [2,3] have shown the advantages of two distinct strategies: (1) exploitation of external knowledge and (2) composition of different matchers followed by propagation of similarity. The first strategy uses background knowledge resources such as the UMLS to support lexical matching of concepts [4-6]. The second strategy propagates similarities between ontology concepts throughout the ontology graphs, based on the assumption that a match between two concepts should contribute to the match of their adjacent concepts, according to a propagation factor [7].

BLOOMS was designed to leverage on the success of simple lexical matching methods, while still finding alignments where lexical similarity is low, by using global computation techniques. It couples a lexical matching algorithm based on the specificity of words in the ontology vocabulary, with a novel global similarity computation approach that takes into account the semantic variability of edges.

1.1 State, purpose, general statement

The original purpose of BLOOMS is to provide the ontology matching component of an ontology extension system called Auxesia. This system combines ontology matching and ontology learning techniques to propose new concepts and relations to biomedical ontologies. Consequently, BLOOMS was specifically designed to match biomedical ontologies in a fully automated fashion, favoring precision over recall.

Although BLOOMS was specifically designed to be applied to biomedical ontologies, its current implementation is domain-independent since it can function without external forms of knowledge. To capitalize on the specific characteristics of most biomedical ontologies, BLOOMS joins a lexical matcher to exploit the rich textual component with a global similarity computation technique to handle the cases where

synonyms exist but are not shared between ontologies. Furthermore, BLOOMS can also exploit annotation corpora, which are available for some biomedical ontologies, to improve the propagation of similarity.

1. Specific techniques used

BLOOMS has a sequential architecture composed of three distinct matchers: Exact, Partial and Semantic Broadcast Match. While the first two matchers are based on lexical similarity, the final one is based on the propagation of previously calculated similarities throughout the ontology graph. Figure 1 depicts the general structure of BLOOMS.

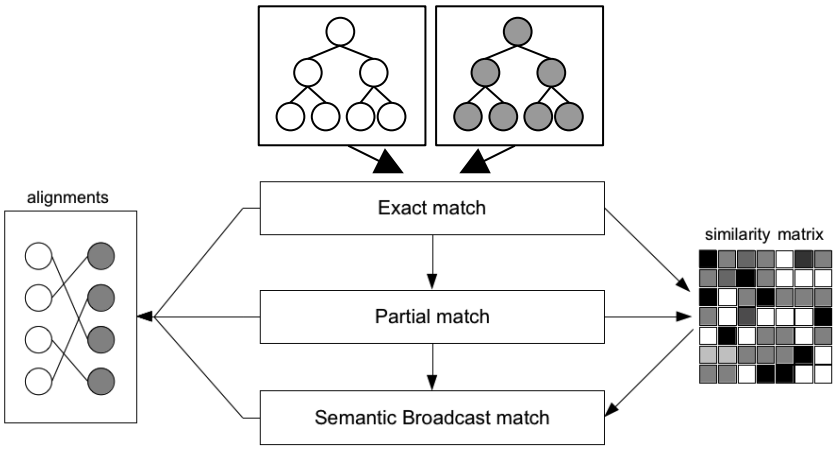

Figure 1. Diagram of BLOOMS architecture. Given two ontologies, BLOOMS first extracts alignments based on Exact matches, then on Partial matches, and finally it propagates the similarities generated by those two strategies using the Semantic Broadcast approach.

1.2 .1 Lexical similarity

Exact and Partial matchers use lexical similarity based on textual descriptions of ontology concepts. Textual descriptors of concepts include their labels, synonyms and definitions. Since ontology concepts usually have several textual descriptors (e.g., name, synonyms, definitions), the similarity between two ontology concepts is given by the maximum similarity between all possible combinations of descriptors.
The first matcher, Exact Match, is run on textual descriptions after normalization and corresponds to a simple exact match, where the score is either 1 or 0.

The second matcher, Partial Match, is applied after processing all concept's labels, synonyms and definitions through tokenizing strings into words, removing stopwords, performing normalization of diacritics and special characters, and finally stemming (Snowball). If the concepts share some of the words in their descriptors, i.e. are partial matches, the final score is given by a Jaccard similarity, which is calculated by the number of words shared by the two concepts, over the number of words they both have. Alternatively, each word can be weighted by its evidence content.
The notion of evidence content (EC) of a word [1] is based on information theory and can be considered a term relevance measure, since it measures the relevance of a word within the vocabulary of an ontology. It is calculated as the negative logarithm of the relative frequency of a word in the ontology vocabulary:

$$EC(word) = -\log freq(word \in V_{\text{ontology}})$$

The ontology vocabulary corresponds to all words in all descriptors of all concepts in the ontology. The final frequency of a word within an ontology corresponds to the number of concepts that contain it in any of their descriptors. This means that a word that appears multiple times in the label, definition or synonyms of a concept is only counted once, preventing bias towards concepts that have many synonyms with very similar word sets. The evidence content of words that are common to both ontologies, is given by the average of their ECs within each ontology.

1.2 .2 Semantic Broadcast

After the lexical similarities are computed, they are used as input for a global similarity computation technique, Semantic Broadcast (SB). This novel approach takes into account that the edges in the ontology graph do not all convey the same semantic distance between concepts.
This strategy is based on the notion that concepts whose relatives are similar should also be similar. A relative of a concept is an ancestor or a descendant whose distance to the concept is smaller than a factor *d*. To the initial similarity between concepts, SB adds the sum of all similarities of the alignments between all relatives weighted by their semantic gap *sG*, to a maximum contribution of a factor *c*. This is given by the following:

$$\text{Sim}_{final}(c_a, c_b) = \text{Sim}_{lex}(c_a, c_b) + c\left(\sum_{|D(r_i, c_a) \wedge D(r_j, c_b) < d \wedge r_i, r_j \in A} \text{Sim}_{lex}(r_i, r_j) \cdot \text{sG}(c_a, r_i, c_b, r_j)\right)$$

where c_a and c_b are concepts from ontologies *a* and *b*, and r_i and r_j are relatives of c_a and c_b at a distance *D* smaller than a factor *d* whose match belongs to the set of extracted alignments *A*.
The semantic gap between two matches corresponds to the inverse of the average semantic similarity between the two concepts from each ontology. Several metrics

can be used to calculate the similarity between ontology concepts, in particular, measures based on information content have been shown to be successful [2].
In BLOOMS we currently implement three information content based similarity measures: Resnik [3], Lin [4] and a simple semantic difference between each concept's ICs. The information content of an ontology concept is a measure of its specificity in a given corpus. Many biomedical ontologies possess annotation corpora that are suited to this application. Nevertheless, semantic similarity can also be given by simpler methods based on edge distance and depth.
Semantic broadcast can also be applied iteratively, with a new run using the similarity matrix provided by the previous.

1.2.3 Alignment Extraction

Alignment extraction in BLOOMS is sequential. After each matcher is run, alignments are extracted according to a predefined threshold of similarity and cardinality of matches, so that the concepts already aligned are not processed for matchers down the line. Each successive matcher has its own predefined threshold.

1.3 Adaptations made for the evaluation

With the purpose of participating in OAEI, BLOOMS was integrated into the AgreementMaker system [5] due to its extensible and modular architecture. We were particularly interested in benefiting from its ontology loading and navigation capabilities, and its layered architecture that allows for serial composition since our approach combines two matching methods that need to be applied sequentially. Furthermore, we also exploited the visual interface during the optimization process of our matching strategy, since although it is not a requirement for our methods, we found it to be extremely useful, it supports a very quick and intuitive evaluation.
Since neither the mouse or the human anatomy ontologies have an annotation corpus, the Semantic Broadcast algorithm used a semantic similarity measure based on edge distance and depth, where similarity decreases with the number of edges between two concepts, and edges further away from the root correspond to higher levels of similarity.

2 Results

BLOOMS was only submitted to the anatomy track, since it is being specifically developed to handle biomedical ontologies. The anatomy track contains 4 tasks: in the first three tasks, matchers should be optimized to favor f-measure, precision and recall, in turn. In the fourth task, an initial set of alignments is given, that can be used to improve the matchers performance. In addition to the classical measures of precision, recall and f-measure, the OAEI initiative also employs recall+, which

measures the recall of non-trivial matches, since in the anatomy track a large proportion of matches can be achieved suing simple string matching techniques.
We only participated in tasks #1 and #2, since BLOOMS is designed to favor precision.

2.1 anatomy

Taking advantage of the SEALS platform we ran several distinct configurations of BLOOMS, testing different parameters and also analyzing the contribution of each matcher to the final alignment.
We found that after the first matcher is run, the alignments produced have a very high precision (0.98), but the recall is somewhat low (0.63). Each of the following matchers increases recall while slightly decreasing precision, which was expected given the increasing laxity they provide.
We also found that weighting the partial match score using word evidence content did not significantly alter results when compared to the simple Jaccard similarity.
For task #1 we used a Partial Match threshold of 0.9 and a final threshold of 0.4. Semantic Broadcast was run to propagate similarities through ancestors and descendants at a maximum distance of 2, and contribution was set to 0.4. Using the SEALS evaluation platform, we obtained 0.954 precision, 0.731 recall, for a final F-measure of 0.828 and a recall+ of 0.315.
For task #2 we used a Partial Match threshold of 0.9 and did not use Semantic Broadcast. With this strategy, we ensured a higher precision, of 0.967. However, recall was not much lower than the one in task #1, 0.725 , which resulted in a final f-measure of 0.829.
We did not participate in other tasks, since BLOOMS was originally intended to yield a high precision, as it is intended to be run in a fully automated fashion as a part of an ontology extension system.

3 General comments

We find that the SEALS platform is a very valuable tool in improving matching strategies. We find however that the 100 minute time limit might be detrimental to strategies that need to process large external resources.

3.1 Comments on the results

BLOOMS was designed to be as fully automated as possible, so it is more geared towards increased precision than recall. Comparing our results for tasks #1 and #2, they clearly indicate that our semantic broadcast strategy does not represent a very heavy contribution to recall, but that we do capture nearly 10% more matches when using both the Exact and Partial Match strategies, than Exact Match alone. Also our

recall+ is not very high, again highlighting the need to expand our strategy to improve recall.
Nevertheless, we find our performance to be comparable to the best systems in 2009, and in 2010 our f-measure in task #1 is 5% lower that the best performing system, whereas in task #2 we are the second best system, with a slight difference of 0.1% in precision. These are encouraging results and we fully intend to participate in future events with an improved version of BLOOMS.

3.2 Discussions on the way to improve the proposed system

We are planning on implementing several strategies for improvement in the near future, some of which were already a part of our initial strategy, but were not yet implemented at the time of OAEI 2010. To improve the lexical similarity matchers, future versions of BLOOMS will take into account spelling variants and mistakes, and we will also investigate the feasibility of using external resources such as UMLS and WordNet to increase the number of synonyms for both terms and words. We feel this would greatly improve the recall of our strategy. Regarding similarity propagation, we will work extensively on improving our semantic broadcast approach, by exploring alternative strategies for the computation of information content independently of an annotation corpus, and thus expand the number of semantic similarity measures that can be used. We will also adapt semantic broadcast to propagate dissimilarity, and decrease the similarity between concepts that might have a high lexical similarity but very distinct neighborhoods.

4 Conclusion

Participating in the anatomy track of OAEI 2010 has given us an opportunity to evaluate a matching algorithm developed with the practical purpose of being used in a semi-automated ontology extension system, Auxesia. Our matching algorithm, BLOOMS, is intended to be as automated as possible, and thus its current implementation favors precision. This was clearly visible in the results we obtained in tasks #1 and #2 of the anatomy track of OAEI 2010, where we obtained high ranking precision values within the top 3, but lower recall.
In future versions of BLOOMS we will implement several strategies designed to improve recall, while minimizing precision loss.
The lessons learned throughout this period will undoubtedly contribute to an improvement of our method.

Acknowledgements

The work performed at University of Lisbon by Catia Pesquita and Francisco M. Couto was supported by the Multiannual Funding Program and the PhD grant SFRH/ BD/42481/2007.
The work performed at UIC by Cosmin Stroe and Isabel F. Cruz has been partially sponsored by the National Science Foundation under Awards IIS-0513553 and IIS-0812258

References

1. Ghazvinian, A., Noy, N. , Musen, M. (2009). Creating mappings for ontologies in biomedicine: Simple methods work. In AMIA Annual Symposium (AMIA 2009), 2. 73
2. Caracciolo, C., Hollink, L., Ichise, R., Meilicke, C., Pane, J. , Shvaiko, P. (2008). Results of the Ontology Alignment Evaluation Initiative 2008. The 7th International Semantic Web Conference. 33, 73
3. Ferrara, A., Hollink, L., Isaac, A., Joslyn, C., Meilicke, C., Nikolov, A., Pane, J., Shvaiko, P., Spiliopoulos, V. , Wang, S. (2009). Results of the Ontology Alignment Evaluation Initiative 2009. Fourth International Workshop on Ontology Matching, Washington, DC , 1. 16, 33, 74
4. Zhang, S., Bodenreider, O. (2007). Lessons Learned from Cross-Validating Alignments between Large Anatomical Ontologies. MedInfo, 12, 822--826. 33
5. Lambrix, P., Tan, H. (2006). SAMBO - system for aligning and merging biomedical ontologies. Web Semantics: Science, Services and Agents on the World Wide Web, 4, 196--206. 31, 33
6. Jean-Mary, Y.R., Shironoshita, E.P. , Kabuka, M.R. (2009). Ontology matching with semantic verification. Web Semantics: Science, Services and Agents on the World Wide Web, 7, 235--251. 33
7. Cruz, I.F., Antonelli, F.P., Stroe, C. (2009). AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. PVLDB, 2, 1586- 1589. 82
8. Couto, F., Silva, M. & Coutinho, P. (2005). Finding genomic ontology terms in text using evidence content. BMC Bioinformatics, 6, S21. 57, 64
9. Pesquita, C., Faria, D., Falcão, A.O., Lord, P., Couto, F.M., Bourne, P.E. (2009). Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology, 5
10. Resnik, P. (1998). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research.
11. Lin D (1998) An information-theoretic definition of similarity. Proc. of the 15th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann. pp. 296–304.