BLOOMS on AgreementMaker: Results for OAEI 2010

Catia Pesquita[1], Cosmin Stroe [2], Isabel 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
ifc@cs.uic.edu

Abstract. BLOOMS is an ontology matching method developed as part of an ontology extension system. It combines lexical similarity measures with similarity propagation based on semantic distance. For the participation in OAEI 2010 BLOOMS was integrated into the Agreement Maker system which has competed in previous years. Although BLOOMS was specifically designed to be as automated as possible, and thus favors precision, results were encouraging.

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.

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. Auxesia combines ontology matching and ontology learning techniques to propose new concepts and relations to bio-ontologies. Consequently, BLOOMS was specifically designed to match bio-ontologies taking into consideration some of their more relevant characteristics: bio-ontologies can have a large number of concepts, and usually provide a large textual component in the form of labels, synonyms and definitions; also, they typically have few types of relations defined between the concepts and little or no axiomatization.

Although BLOOMS was specifically designed to be applied to bio-ontologies, it is a domain-independent strategy since it can function without external forms of knowledge. To capitalize on the specific characteristics of most bio-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 capitalize on annotation corpora, which are a feature of some biomedical ontologies initiatives.

1. Specific techniques used

BLOOMS has a sequential architecture composed of three distinct matchers: Exact Match, Partial Match and Semantic Broadcast. 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 the general structure of BLOOMS.

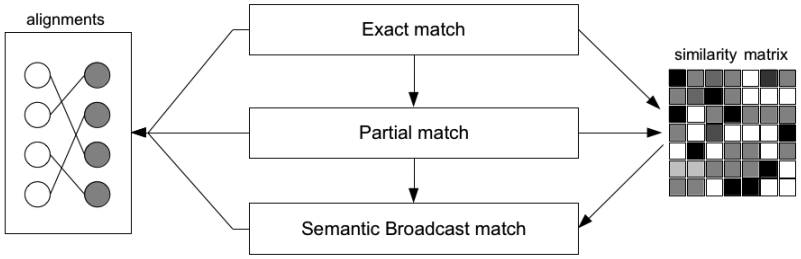

Figure 1. Diagram of BLOOMS architecture.

1.2 .1 Lexical similarity

The first two matchers used in BLOOMS 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 1.0 or 0.0.
The second matcher, Partial Match, is applied after processing of 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.
The ontology vocabulary is corresponds to all words in the all descriptors of all concepts in the ontology. The final frequency of a word 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.

1.2 .2 Semantic Broadcast

After the lexical similarities are computed, they are used as input for a global similarity computation technique, Semantic Broadcast. 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, to a maximum contribution of a factor c.

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 concepts 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.

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, since it supports a very quick and intuitive evaluation.

Since neither the mouse or the human anatomy ontologies have an annotation corpus, we had to adapt the Semantic Broadcast algorithm to use a semantic similarity measure based on edge distance and depth, so that edges further away from the root correspond to higher levels of similarity.

2 Results

BLOOMS was only submitted to the anatomy track.

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 taskk #1 we used a Partial Match threshold of 0.9 and a final threshold of 0.4. Using the SEALS evaluation platform, we obtained 0.954 precision, 0.731 recall, for a final F-measure of 0.828.

For task #2 we used a Partial Match threshold of 0.9 and did not use semantic broadcast.

We did not participate in other tasks, since BLOOMS was originally intended to yield a high precision.

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. Nevertheless, we find our performance to be comparable to the best systems in 2009, and hope to participate in future events with an improved version.

3.2 Discussions on the way to improve the proposed system

We are planning on implementing several strategies for improvement in the near future, namely using semantic broadcast to propagate dissimilarity, and decrease the similarity between concepts that might have a high lexical similarity but very distinct neighborhoods. We would also like to implement different semantic similarity measures, possibly exploring alternative strategies for the computation of information content.

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. The lessons learned throughout this period will undoubtedly contribute to an improvement of our method.

References

1. Couto, F., Silva, M. & Coutinho, P. (2005). Finding genomic ontology terms in text using evidence content. BMC Bioinformatics, 6, S21. 57, 64
2. 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
3. Resnik, P. (1998). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Arti cial Intelligence Research.
4. 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.
5. Cruz, I.F., Antonelli, F.P. & Stroe, C. (2009). AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. PVLDB, 2, 1586- 1589. 82