Simple library thesaurus alignment with SILAS

Roelant Ossewaarde[1]

Linguistics Department
University at Buffalo, the State University of New York
rao3@buffalo.edu

Abstract. This paper describes a system written in C which employs the instance-based approach to ontology alignment of library thesauri. It computes relatedness relations between subsets of a library catalogue. Even a very basic method to calculate confidence in the found relation yields usable alignment results.

1 Presentation of the system

1.1 State, purpose, general statement

The system in this paper (SILAS - Simple Instance-based Library-thesaurus Alignment System) is an example of 'instance-based ontology matching' [3]; it measures the similarity between subsets annotated with words from different ontologies to match up the concepts described by these words.

Definitions An *ontology* is taken to be an organization of concepts C by some set of ontological relations. An *agent* is a human or computer program that employs an ontology. A *concept* is a particular agent's conceptualization of an element of the domain of discourse. A concept is denoted by one or more words, which may be shared between agents. In the specific case of the library task, the goal is to find concepts in one ontology that are equivalent or related to concepts in the other one.

A thesaurus is typically used in a library to create a subset of books - we assume that for each word $w_i \in W$ that denotes a concept $c_i \in C$, there is a subset $B_i \in B$ of books that is related to that concept. For the purpose of ontology alignment, a subset B_i is assumed to represent the meaning of the concept c_i described by word w_i in that database. Terminology: subset B_i represents the *extensional meaning*, or *extension*, of the concept c_i. For example, the concept ROWING may be described by the Dutch word *roeien* and the English word *rowing*. The extension of the concept ROWING is formed by the set of books which have their subject tagged as *roeien* or *rowing*.

Assumptions SILAS relies on the following assumptions about the relationship between subsets and concepts:

1. Assume a concept c_A (denoted by word w_A) in ontology A maps onto a concept c_B (denoted by word w_B) in ontology B. Then the subset of books described by word w_A will generally be the same as the subset of books described by word w_B of B.

For example, if both ontology A and B have a term for the concept *rowing*, a librarian who works with ontology A will generally tag the same books as a librarian who works with ontology B.

2. Converse: consider two partitionings of the set of books, one tagged as about concept c_A and the other tagged as about concept c_B. The more overlap the two subsets show, the stronger the semantic relation between concepts c_A and c_B. If there is enough overlap, the concepts are said to be the same.
   For example, if there is a subset of books tagged with the words from ontology A for the concept *rowing* in that ontology, and there is also a subset of books tagged with the words from ontology B for the concept *rowing*, and the subsets contain the exact same members, then the two concepts are probably equivalent.
3. It is assumed that for every concept in ontology A, there is a somewhat equivalent concept in ontology B.

The two ontologies used for alignment were the ones provided by the organizers, ontology 'GTT' with 32530 concepts and ontology 'Brinkman' with 4845 concepts.

The approach The system identifies overlapping subsets and computes a very simple metric to predict semantic relatedness between the concepts of which the subsets are extensions. The library track was particularly suited for this approach because it provides both a pair of ontologies, roughened and shaped by actual use, and a large set of datapoints actually described by the ontologies.

For this particular task, the collection of the library of the University of Amsterdam was used to provide the datapoints. This library has a collection of about 4 million books, of which about 90% is annotated using either of the two thesauri.

Systems such as AUTOMS [4] and Falcon-AO [2] use WordNet senses as an intermediary between concepts in two ontologies. But external semantic annotation using resources such as WordNet is limited to languages for which such resources exist, and is limited to domains for which such resources are formulated.

SILAS is based on the condition that no semantic knowledge from outside the ontology is available at runtime. The meaning of the ontological concepts is in some way coded through its use; the extension of a concept represents its meaning. For example, if two different librarians with two different thesauri make a subset of all books on rowing and label the subset with their respective terms for that sport, the system should find that the used terms probably describe the same concept, because they both have the same subset as their extension.

1.2 Specific techniques used

Confidence scores The overlap between sets may be relevant, and it may not. In the latter case, it is noise that needs to be filtered out. As an example, the category 'antillianen' (*Antillians*) shows some overlap with the category 'anticonceptie' (*contraceptives*). Perhaps there have been included in the subset publications on the use of contraceptives among Antillians, but that overlap should clearly not be taken as evidence that Antillians and contraceptives are in any way semantically related.

For each overlapping pair of sets, a confidence score is calculated, which expresses the following characteristics:

1. Two sets are more likely to be aligned if they have more elements in common, instead of less.
2. Two sets, one of which large, the other small, are more likely to aligned if their average portion of joint elements is more, instead of less.
3. Two sets are more likely to be aligned if they are identified by the same words. (*lexical booster*)

The confidence score is implemented by ranking properties of the overlapping sets A and B using a variation of the often used Jaccard similarity measure:

1. Let confidence = 0 and $x = |A \cap B|$ (the number of elements in the intersection of A and B);
   (a) if $((x/|A|) + (x/|B|))/2 > 0.05$, confidence +1, if > 0.15, additional +1.
   (b) i. if $x \geq 5$ and $x/|A| \geq 0.3$, confidence +4; otherwise
   ii. if $x/|A| \geq 0.05$, confidence + 1, if $x/|A| \geq 0.2$, additional +1.
   Similarly so for set B.
2. if the name of the concept described by subset A is identical to the name of the concept described by subset B, confidence +5. (*lexical booster*)

This yields a confidence score S between 0 and 15. If $3 \geq S \geq 5$, two concepts are related; if $6 \geq S \geq 15$, two concepts are equivalent. If there are no alignment candidates with a confidence score high enough for equivalence, but there is an alignment candidate with the same name, this candidate is chosen to be the equivalent concept. Related concepts have been assigned the relation `broadMatch`, equivalent relation have been assigned the relation `exactMatch`. The confidence score is then mapped onto a scale $\{0 \ldots 1\}$.

The rationale for the lexical booster is that it is obvious that two concepts with the same name are probably equivalent. However, lexical similarity alone is not enough for concepts to be judged related. Its use is mainly for ranking alignment candidates: if there is a lower-ranked alignment candidate with a similar name, it gets promoted to the top ranking. The numbers used in the algorithm were initially chosen on the basis of intuition and adjusted after hand-checking the results for 5% of all categories. The algorithm becomes more selective if the thresholds for relatedness and equivalence are raised.

Alignment procedure The procedure of alignment is as follows:

1. The words describing ontological concepts in the two ontologies were translated into two sets search terms T.
   A software tool `skostool` was developed in C, on the basis of the open-source RDF-parsing library `libraptor`. The tool enables easy browsing, searching and manipulation of the concepts in the ontologies, and provides a C API-interface to the files provided by the organizers. In particular, it can compute hierarchical measures that we intend to use in subsequent versions of our alignment protocol.

2. Using automated spiders, for each $t_i \in T$, all records tagged with term t_i were retrieved into subset $B_i \in B$.
3. Each subset corresponding to a concept from ontology GTT was then compared to all subsets corresponding to concepts from ontology Brinkman.
   Each comparison yields an overlap score for each pair $\{B_i, B_j\}$, where B_i is a subset according to ontology GTT and B_j is a subset according to ontology Brinkman. If the overlap between the two sets is sufficiently high so that the confidence score is greater than 0, B_j is considered an *alignment candidate* for B_i.
   The overlap is determined by finding the intersection $B_i \cap B_j$, using ISBN as a unique identifier for each $b \in B_i(B_j)$. The comparison function was coded in C. The runtime on a linux PC is about 3 hours.
4. For each B_i, the alignment candidates were scored according to the confidence measurement system. If the confidence exceeds certain thresholds, the concepts of which subsets B_i and B_j are extensions are judged *related* or *equivalent*.

1.3 Link to the system and parameters file

The system and its documentation is available from `http://www.buffalo.edu/~rao3/oaei2007`. The system requires a POSIX-compliant C-compiler to build, and makes use of the open source libraries `libxml` and `libraptor`. Two memory leak bugs were fixed in `libraptor` in order to make it work flawlessly in combination with `libxml`; patches have been submitted to the maintainer of `libraptor`.

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

The alignments, both in human readable form and in XML-format are available from `http://www.buffalo.edu/~rao3/oaei2007`.

2 Results

Because participation was limited to the library track (scored blind), no formal results can be reported yet. Informal review by the author of the system's performance shows that with the current parameters, in most cases both the precision and recall of the algorithm are high enough to be satisfactory. Some of the judgements are subjective (is 'pottery' really equivalent to 'ceramic arts'?); fellow linguists judged more than 90% of the relations as 'acceptible as correct' in an informal evaluation of the first 500 concepts.

3 General comments

3.1 Comments on the results

There are a few benefits of the instance-based approach which make it interesting for further study.

First, the approach is independent of outside semantic knowledge at runtime. No sources such as WordNet or FrameNet are required to generate the mappings. The assumption is that humans, in this case librarians, have already assigned meaning to a term by formulating its extension, ie. by identifying a subset of books on that subject, and it would be a waste not to use that knowledge in mapping.

Second, the system is blind as to the exact terms used to describe concepts. The two ontologies may as well have been given in completely unrelated languages; as long as they are both used to describe the same set of data points, the algorithm will find matches. The advantage of this can easily be seen in a situation of academic libraries in different countries; each library would use its own localized thesaurus to describe the proceedings of this conference, and if all works well, the algorithm would pick up the relevant localized terms and align them.

3.2 Discussions on the way to improve the proposed system

The system presented is in its infant stages. The following are a few possible improvements for SILAS v2.

- The system currently cannot compute hierarchical relations. We'd like to implement that, for example using the MedOnt / MedCount algorithms described in [1] or a similar methodology. For example, given a set of alignment candidates from one ontology, the hierarchy between the different candidates may be reconstructed using the original ontology and factored in while scoring the confidence rating, perhaps using the semantic similarity between nodes of the ontology as computed in the DSSim-system [5].
- The system regards hierarchical relations as non-transitive. In other words, if concept c_1 taxonomically dominates concept c_2 (ex. *science* describes the superset of *linguistics*), a book tagged as subject c_2 would not also occur in the extension of concept c_1. If the relation is considered transitive, as it perhaps should (although [3] suggests a decrease in performance in this particular task once hierarchical information is taken into account), the recall of the alignment improves. However the precision decreases significantly, due to the fact that the confidence ratings used for this task cannot adequately distinguish a higher and a lower hierarchical concepts. Because by definition the extension of the concept SCIENCE includes all scientific concepts, such as BIOLOGY and LINGUISTICS, each of these concepts would show overlap with SCIENCE and would be considered an alignment candidate. Until the algorithm can be restricted to only choose alignment candidates in roughly the same hierarchical area, transitive relations would degrade the quality of the system.
- The lexical booster is very simple. It is currently based on simple pattern matching, and does not detect plural-singular alternations and such. A better way of computing lexical similarity may provide a more fine-grained confidence measure and a better way of using lexical information in ranking alignment candidates.

4 Conclusion

SILAS uses a very naive, simple, number-crunching approach. Official results have not been made available yet, but a glance over the computed alignments shows at least

a satisfactory performance, given the low complexity of the methodology. It makes effective use of a source of information - the annotated database - that is often ignored, and as such may provide a starting point for combinations with other methodologies.

References

1. Alistair E. Campbell and Stuart C. Shapiro. Algorithms for ontological mediation. In S. Harabagiu, editor, *Usage of WordNet in Natural Language Processing Systems: Proceedings of the Workshop*, pages 102–107, New Brunswich, NJ, 1998. COLING-ACL.
2. Wei Hu, Gong Cheng, Dongdong Zheng, Xinyu Zhong, and Yuzhong Qu. The results of Falcon-AO in the OAEI 2006 campaign. In Shvaiko et al. [6].
3. Antoine Isaac, Lourens Van der Meij, Stefan Schlobach, and Shenghui Wang. An empirical study of instance based ontology matching. ISWC 2007 + ASWC 2007, 2007.
4. Konstantinos Kotis, Alexandros Valarakos, and George Vouros. AUTOMS: Automated Ontology Mapping through Synthesis of methods. In Shvaiko et al. [6].
5. Miklos Nagy, Maria Vargas-Vera, and Enrico Motta. Dssim-ontology mapping with uncertainty. In Shvaiko et al. [6].
6. Pavel Shvaiko, Jérôme Euzenat, Natalya Noy, Heiner Stuckenschmidt, Richard Benjamins, and Michael Uschold, editors. *Proceedings of the 1st International Workshop on Ontology Matching (OM-2006)*, 2006.