Hybrid mappings of complex questions over an integrated semantic space

Conference or Workshop Item

Hybrid Mappings of Complex Questions over an Integrated Semantic Space

Gaston Burek, Anne De Roeck and Zdenek Zdrahal
Computing Research Centre
The Open University, Milton Keynes, MK7 6AA, UK
[G.G.Burek, A.DeRoeck, Z.Zdrahal]@open.ac.uk

Abstract

We address the issue of measuring semantic similarity between ontologies and text by means of applying Latent Semantic Analysis. This method allows ranking of vector representations describing semantic relations according to their cosine similarity with a particular query. Our work is expected to make contributions including the introduction of reasoning about uncertainty when mapping between ontologies, an algorithm that can perform automatic mapping between concepts or relations derived from text and concepts or relations belonging to different ontologies, and the capability to infer implicit similarity between concepts or relations.

1. Introduction and motivations

Within the Semantic Web (SW) [6] formal conceptualizations of knowledge in different domains are implemented in the form of ontologies [5]. A reduced set of ontologies can be related semantically to heterogeneous sources to facilitate the management of information online. The main problem with this way of representing semantics is that it is very unlikely that ontology developers will agree on the same ontology for formalizing a single domain [10]. Consequently, there is a need for the development of automatic mapping methods aimed to generate interoperability between ontologies. Today, mappings are coded manually making this process very expensive [9]. Moreover, it is expected that a large number of linguistic resources online will be encoded as Natural Language (NL) [2]. This also makes it imperative to develop interoperability between ontologies and other linguistic resources.

The motivation of our work is to provide integrated access to text documents. We propose to achieve that goal by generating mappings between Natural Language Expressions (NLE; i.e. complex questions) and a set of ontologies and other linguistic resources (i.e. NLE formalized within a document collection) integrated within a semantic space. The space is generated by means of applying Latent Semantic Analysis (LSA) [3] to a Vector Space Model (VSM) [12] containing weighted frequency vector representations of ontologies and text. This approach for formalizing the semantic space is intended to explore the use of term frequencies characterizing meanings as contexts. Also, to solve the problems arising from a) mapping knowledge entities (i.e. classes, instances and relations) that belong to two or more different ontologies and b) mapping NL to the semantic space.

The rest of the paper is organized as follow: section two describes the theoretical background for analyzing our research problem; section three specifies the problem, section four presents an example and section five proposes a mapping method together with the solution for the example described in section four. Sections six, seven and eight describe future work, contributions and acknowledgements respectively.

2. Theoretical background

2.1. Ontologies

Ontologies are used to provide semantics and structure to the data. Ontologies formalize knowledge by organizing concepts within taxonomy of classes. These classes have certain attributes that differentiate them one from each other and can be instantiated by fixing the values for those attributes. The instantiations of classes are organized as a Knowledge Base (KB).

The lack of agreement between ontologies is exposed by the terminology gaps occurring between them. To solve this problem it is necessary to measure semantic similarities between classes and relations within the ontologies. This process is called ontology mapping and transforms instances of a particular

Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05)
1529-4188/05 $20.00 © 2005 IEEE

ontology into instances of other ontology. Ontologies within the SW environment need to be integrated by means of mapping.

2.2. Similarity measures and Latent Semantic Analysis

2.2.1. Similarity measures

Similarity measures are used to measure the relevance of objects within a knowledge base and/or in a document collection (e.g. terms, documents, functions, commands, etc.) for a particular statement formalizing a question (e.g. queries). Those measures [7] combined with a probabilistic knowledge representation framework [4] can be applied to measure semantic similarity between vector representations of structured (i.e. hierarchical) and non-structured (bags of words) information.

2.2.2. Latent Semantic Analysis

LSA extends the VSM, a probability model, implemented as a term-to-context matrix of weighted terms. Within the matrix, rows represent the weighted frequency of the term occurring within different contexts and columns represent documents or other contexts (e.g. sentences, paragraphs, etc). This model induces global knowledge from term frequencies within local contexts belonging to a document collection. The data entry used in the model [8] is first order local associations between stimuli and context in which those stimuli occur. LSA uses Single Value Decomposition (SVD) [1] of the term-to-context matrix to capture higher order associations and to identify the semantic dimensions that are statistically significant to characterize the model used in the generation of language.

2.2.3. Cosine similarity

The cosine similarity measure is used to calculate semantic similarity between two columns/rows within the term-to-context matrix. The cosine of the angle between two vectors is defined as the inner product between the vectors divided by the product of their length.

$$Cos\theta = \frac{v.w}{\| v \| . \| w \|}$$

2.3. Question Answering

Question Answering (QA) systems generate direct answers to user questions consulting information stored in one single source (e.g. document collection, database, etc) or a set of sources (e.g. World Wide Web). In particular, each system uses different procedures to generate the answer according to the type of source used (e.g. data bases, ontologies knowledge bases, text documents, etc.). Most QA systems are composed of four components (Question Analysis, Document Retrieval, Passage Retrieval, and Answer Extraction) [13]. To find correspondence between a question and a set of possible answers systems measure semantic similarity between the queries and the information contained within the sources.

Once a set of answers is selected from each available source all the answers need to be ranked to select the best one. To do so systems need to understand how similar or dissimilar those answers are compared to each other. In addition, the answers may be specified differently (e.g. text, logical predicates, etc) depending on the knowledge representation used by the particular source from which they are extracted (i.e. text or ontology).

3. Problem specifications

Integrating ontologies and text within the SW requires making sense of the different terminology used within the various sources. Also, it will expose all the problems related to the 'terminology gap': the fact that concepts and their semantic relations can be expressed in different ways within a particular ontology, and that there is not robust and scalable method for relating text to meanings formalized by ontologies. Our research explores term-concept dimension for solving the problem of mapping between semantic relations formalized as attributes of classes with the ontologies, NLE and queries. Mapping queries and NLE expressions to semantic relations cannot be compared only by their name because they may share the same meaning in a particular context but the same mining may be described using a different terminology (i.e. synonymy). Also the same vocabulary may be used to describe semantically different relations (i.e. homonymy).

To compare those semantic relations it is necessary to have a way of calculating a degree of similarity between them and to reason about uncertainty in the similarity. By using LSA a degree of similarity between sets of related concepts can be calculated by means of finding co-occurrence of terms in definitions of concepts. In the case of ontologies, classes' names and attributes can be used for this purpose.

4. Problem example

The following problem example arises from mapping a complex question over heterogeneous data sources (i.e. text and ontologies) and illustrates the difficulties of measuring similarity between different semantic relations defined within ontologies or text documents.

Rule 1	([NNS]) <:VBP:>IN ([NNP]) → VBP (NNS, NNP)
Rule 2	([NNP NN]) <:VBD:>IN ([JJ NN]) →VBD (NNP, JJ NN)
Rule 3	([NNP NNP]) <: VBZ VBN :> IN ([DT JJ NN]) → VBZ VBN (NNP NNP, JJ NN)
Rule 4	([NNP NNP]) <: VBZ VBN :> IN ([DT JJ _NN]) → VBZ VBN (NNP NNP, JJ NN)

Table 1. Syntactic Rules

Given the complex question "What researchers work in the ALPHA project financed by the Argentine government?" we can decompose it in two queries by using a shallow parser[1] and syntactic rules to build semantic relations.

Applying Rule 1 and Rule 2 (See Table 1) to the shallow parser output (see Figure 1) for the complex question we can derive respectively the semantic relations Q1 and Q2 (See Table 2). In addition, we assume that two ontologies (i.e. O1 and O2) are available. O1 contains instances describing university staff in South America and defines the relation R1. O2 contains instances of projects funded by governments in South America and defines the relations R2 and R3 (See Table 2). Further more, we make also the assumption that the sentences "South Foundation is sponsored by the Argentine government" and "South Foundation is honored by the Brazilian government" are NLE semantically related to O1 and O2. Applying Rule 3 and Rule 4 (See Table 1) to the shallow parser output for both sentences (See Figure 2) we derived the semantic relations R4 and R5 respectively (See Table 2).

Finally, to answer the complex question we need to find within the five available relations the answers for Q1 and Q2. Q1 requires a list of one or more answers and Q2 requires a yes/no answer. The problem yet to be solved is to confirm that Q2 is similar to R2 and more similar to R4 than to R5. We also need to confirm that Q1 is similar to R1 and R3 (See Table 4).

```
     ([ What_WDT researchers_NNS])
            <: work_VBP :>
in_IN([the_DT ALPHA_NNP project_NN ])
          <: financed_VBD :>
     by_IN ([the_DT Argentine_JJ
           government_NN ])
```

Figure 1. Shallow parser results for Q1 and Q2

```
([South_NNP Foundation_NNP ])
 <: is_VBZ sponsored_VBN :>
by_IN ([ the_DT argentine_JJ
      government_NN ])

([South_NNP Foundation_NNP ])
  <: is_VBZ honoured_VBN :>
by_IN ([ the_DT brazilian_JJ
      government_NN ])
```

Figure 2. Shallow parser results for R4 and R5

5. A proposed method for hybrid mappings

The mapping method proposed here uses a VSM to represent ontologies, semantically related linguistics resources and queries. Then it applies LSA and the cosine similarity to measure similarity between the relations and queries.

5.1. Vector representations for semantic relations

Our method proposes the representation of semantic relations formalized by ontologies and derived from

[1] The shallow parser used in the example is NLProcessor Copyright © 2000 2001 Infogistics Ltd.

NLE (See Table 2) by means of a term-to-context matrix, where the columns of the matrix represent the semantic relations defined as a bag of words and the rows represent the processed (i.e. stemmed) version of the terms used to define those relations. The matrix entries are the frequency of the processed words for each bag representing semantic relations. A semantic relation derived from an NLE is represented by all the terms used within the relation. A semantic relation within ontology is represented as a vector that contains the frequency of terms (names and properties) associated to all classes related by the semantic relation.

Q1	work in(researchers, ALPHA project)
Q2	financed by(ALPHA project, argentine government)
R1	works in(Martin Smith, University of Buenos Aires)
R2	funded by(ALPHA project, South Foundation)
R3	collaborate with(University of Buenos Aires, ALPHA project)
R4	sponsored by(South Foundation, Argentine government)
R5	honored by(South Foundation, Brazilian government)

Table 2. Queries and semantic relations

The conformed matrix represents an integrated semantic space that defines all semantic relations defined by ontology in addition of all the relations derived from other sources (i.e. text documents). Once the documents are built all the terms are preprocessed (i.e. stemming) and frequencies are weighted by means of applying any of the exiting frameworks for term weighting [11].

Vectors representing queries contain the frequencies of the stemmed terms appearing as arguments within the relation. For instance, given Q1 and Q2, derived from the complex question described in the example of the previous section, the vector representing Q1 contains a frequency combination of one in the rows corresponding to the stemmed terms "work", "research", "alpha" and "project". The vector representing Q2 will contain a frequency combination of one in the rows corresponding to the stemmed terms "financ", "alpha", "project", "argentin" and "govern".

	R1	**R2**	**R3**	**R4**	**R5**
Q1	0.81	0.79	0.93	0.36	0.36
Q2	0.29	0.99	0.59	0.86	0.83

Table 3. Cosine similarity results for Q1 and Q2

5.2. Query mapping

We apply LSA to the matrix defining the integrated semantic space and then calculate the cosine similarity between the relation and each of the queries. In this way we obtain a ranking of similarity for each relation.

5.3. Example solution

To solve the problem described in section three we calculate the cosine similarity between Q1 and Q2 (See Table 3) and each of the relations (R1, R2, R3, R4, R5). Once we have chosen the two higher values for the cosine similarity, the results confirm that Q1 is similar to R1 and R3 and that Q2 is similar to R2 and R4. A similar conclusion can be drawn by analyzing the arguments of the semantic relations for the reason that in the presented example the different semantic relations share vocabulary. However, this will not be true in most of the cases when using different ontologies and online documents. Although common vocabulary is not shared among all the online sources, LSA seems to be a method capable of describing imprecise mappings, due to its probabilistic representation of information. Those imprecise mappings capture uncertainty that arises for different reasons: either the mapping language may be restricted to express mappings with complete certainty, or the concepts in the two models simply do not match up precisely [9].

6. Future work

We are currently building a narrow domain QA system to run a series of experiments using ontologies and news articles describing driving rules, transport vehicles, and road accidents. Evaluating the system is a particularly complex task given the fact that we are dealing with artificially created knowledge representations (i.e. ontologies and vector representations). It is unlikely that a human being will be able to determine if an example of such representations describes the right answer for a particular query. For that reason our evaluation will involve the use of Frequently Asked Questions (FAQ) from online repositories. We expect that within the semantic space queries derived from a question will map close to one of the semantic relations derived from the answer.

7. Contributions

This work makes several research contributions. In particular, the use of probabilistic methods such as LSA to measure semantic similarity between structured information sources (i.e. ontologies) and no-structured (i.e. text documents) ones. We analyze the information represented by semantically related concepts within the term/concept dimension. Semantic relations held between concepts within ontology are evaluated as contexts containing particular concepts. The second contribution is the introduction of uncertainty when mapping those concepts and relations by means of using a probabilistic method to measure semantic similarity that takes into account frequencies of terms used in a particular context (e.g. terms used to name properties of a class). The third contribution is an algorithm that can perform automatic mapping between concepts or relations belonging to different ontologies. A final contribution is the possibility of capturing implicit similarity between concepts or relations. This capability is provided by the SVD method incorporated within the LSA by means of capturing similarity between concepts or relations even if their descriptions do not use common terms.

8. Acknowledgement

This research was supported by the Advanced Knowledge Technologies (AKT) project and the Knowledge Media Institute at The Open University. AKT is an Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01.

9. References

[1] M. Berry, *Large-scale singular value computations*, Volume 6-1, pages 13-49, 1992.

[2] P. Buitelaar, T. Declerck, N. Calzolari, A. Lenci. "Language Resources and the Semantic Web", *Proceedings of the ELSNET/ENABLER workshop*, Paris, France, August 28th/29th 2003

[3] S.,C.,Deerwester, S.,T.,Dumais, T. K Landauer,., G. W., Furnas, R.A. Harshman "Indexing by Latent Semantic Analysis" *JASIS*, 1990, 41(6): 391-407.

[4] D. Florescu, D. Koller and A. Levy, "Using Probabilistic Information in Data Integration", *Proceedings of the 23rd VLDB Conference*, Athens, Greece, 2003.

[5] T. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", *International Journal of Human and Computer Studies*. Special Issue: Formal Ontology, Conceptual Analysis and Knowledge Representation, 2005.

[6] J. Hendler, T. Berners-Lee, and E. Miller, "Integrating Applications on the Semantic Web". *Journal of the Institute of Electronic Engineers of Japan*, 2002, 122(10):676-680.

[7] W. P Jones, and G. W Furnas, "Pictures of relevance: A Geometric Analysis of similarity measures, *Journal of the American society for information science*, 1987, 38(6) 420-442

[8] T. K. Landauer, and , S. T Dumais "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge", *Psychological Review*,1997, 104(2) , 211-240.

[9] J. Madhavan, P. Bernstein, P. Domingos and A. Halevy, "Representing and Reasoning about Mappings between Domain Models", *Proceedings of the AAAI Eighteenth National Conference on Artificial Intelligence*, 2002.

[10] N.F. Noy "What do we need for ontology integration on the Semantic Web", *Proceedings of the Semantic Integration Workshop*, Collocated with the Second International Semantic Web Conference (ISWC03), 2003 .

[11] G. Salton, G. and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing and Management*, 1988, 24 (5):513-523.

[12] G. Salton, A. Wong and C. Yang, "A Vector Space Model for Automatic Indexing" Communications of the ACM, 1971, 18(11):613-620.

[13] S, Tellex, S., B. Katz, J. Lin, G. Marton, and A, Fernandes, "Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering", *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR), Toronto, Canada, 2003.

1529-4188/05 $20.00 © 2005 **IEEE**