Measuring Concept Relatedness Using Language Models

Dolf Trieschnigg[1], Edgar Meij[2], Maarten de Rijke[2] and Wessel Kraaij[3]
[1] HMI, University of Twente, Enschede, The Netherlands
[2] ISLA, University of Amsterdam, The Netherlands
[3] TNO ICT, Delft, The Netherlands
trieschn@ewi.utwente.nl, {emeij,mdr}@science.uva.nl, kraaijw@acm.org

ABSTRACT

Over the years, the notion of concept relatedness has attracted considerable attention. A variety of approaches, based on ontology structure, information content, association, or context have been proposed to indicate the relatedness of abstract ideas. We propose a method based on the cross entropy reduction between language models of concepts which are estimated based on document-concept assignments. The approach shows improved or competitive results compared to state-of-the-art methods on two test sets in the biomedical domain.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods*

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Language Models, Semantic Relatedness

1. INTRODUCTION

Humans prefer to think and reason in terms of concepts rather than words. For computer-aided reasoning the relationships between concepts are often explicitly modeled in an ontology. In this context it is important to have a measure which indicates the *semantic relatedness* of concepts. In IR this measure can, for instance, be used for expanding a query with related concepts.

In this work we assume to have an ontology consisting of concepts and relationships as well as a document collection in which to each document one or more concepts have been assigned. In the literature, four categories of concept relatedness measures can be distinguished: based on structure, information content, association, or context.

Firstly, concept relatedness can be based on ontology structure. Concepts close to each other in the structure are assumed to be strongly related. Path length is a typical indicator of this kind of relatedness [1]. More sophisticated methods take into account the depth of the concepts in the structure or look at the *lowest common subsumer* [5].

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

Secondly, methods based on information theory have been proposed. These methods take into account the Information Content (IC) of the concepts. The IC indicates the specificity of a concept and can be related to the ratio of documents assigned to a concept. Resnik [7] proposed to measure the relatedness of concepts by looking at the IC of the lowest common subsumer. Lin [3] extended this by also taking into account the IC of individual concepts.

Thirdly, different association-based methods can be used to determine the relatedness of concepts, such as Dice, Jaccard and Overlap coefficients [8]. The co-occurrence of instances of two concepts—which in our current case means concepts assigned to the same document(s)—serves as an relatedness indicator. Other, collocation-based measures such as Pointwise Mutual Information (PMI) and Log Likelihood Ratio (LLR) can be used for this purpose as well [4].

Finally, relatedness of concepts has been estimated by considering the context of concepts, where the context of a concept consists of text discussing it. Pedersen et al. [6] present an approach in which the relatedness of concepts is defined as the cosine of the angle between two *context vectors*.

In this paper we present a novel context based measure of concept relatedness, based on cross entropy reduction. After introducing our method, we compare it to the methods introduced earlier, by comparing the results with relatedness judgments provided by human assessors.

2. PROPOSED RELATEDNESS MEASURE

As a concept distance measure we propose to use a symmetrical version of the Cross Entropy Reduction (CER) between two concept language models. A concept language model θ_c is defined as a distribution over words based on a concatenation of all documents annotated with concept c. The concept language models are smoothed with a background language model, based on the collection (Jelinek-Mercer smoothing, $\lambda = 0.7$ background).

The rationale behind our CER-based notion of concept relatedness is that related concepts are surrounded by similar language. The CER quantifies how much better a certain language model is in modeling a certain observed text in comparison with modeling by a collection model. CER has already been successfully applied to ad hoc retrieval and topic detection and tracking [2]. The CER is defined as follows:

$$\begin{aligned} \mathrm{CER}(\theta_c; M, \theta_{c'}) &= H(\theta_{c'}, M) - H(\theta_{c'}, \theta_c) \\ &= \sum_t P(t|\theta_{c'}) \log \frac{P(t|\theta_c)}{P(t|M)} \end{aligned}$$

where θ_c is the language model of a concept c, M is a background language model and $H(\theta_1, \theta_2)$ is the cross entropy between two language models. The incorporation of $H(\theta_{c'}, M)$ is essential since it makes the resulting scores comparable across different concept pairs. The relationship with KL divergence is as follows: $\text{CER}(\theta_c; M, \theta_{c'}) = \text{KL}(\theta_c||M) - \text{KL}(\theta_c||\theta_{c'})$. A symmetrical version of CER is used as a concept distance [9]:

$$d_{\text{CER}}(c, c') = \frac{\text{CER}(\theta_c; M, \theta_{c'}) + \text{CER}(\theta_{c'}; M, \theta_c)}{2}$$

3. EXPERIMENTAL SETUP

To assess the quality of our relation estimation method, we look at correlations with *semantic* relatedness as indicated by human assessors. Performance is measured by looking at the level of agreement between the two gold standard sets and each method, using both Kendall's tau rank correlation coefficient and Pearson's correlation coefficient. As ontology we use the Medical Subject Headings (MeSH) controlled vocabulary thesaurus maintained by the National Library of Medicine (NLM). The 2007 PubMed baseline distribution—consisting of around 16 million biomedical abstracts— is used for the creation of the concept language models. This bibliographic database has been manually indexed using MeSH concepts by curators from the NLM.

Caviedes and Cimino [1] kindly provided us with a test set of 55 MeSH concept pairs (11 unique concepts), judged on relatedness by three physicians on a 1 to 10 scale. As a second test set we use a set of 24 concept pairs (47 unique concepts) derived from Pedersen et al. [6], judged by experts on a 1 to 4 scale. This test set was developed for evaluating relatedness measures on the SNOMED-CT ontology.

We compare our approach to two structure-based methods (*path* and *Nguyen* [5]), one information content approach (*Lin* [3]), three association-based approaches (*Dice*, *LLR* and *PMI*), and one context-based approach (*context* [6]).

4. RESULTS & DISCUSSION

Table 1 shows the correlation between the different methods and the humanly assessed concept pairs. Several things are worth noting. First, our CER-based measure of concept relatedness performs best on test set 1 and second best on test set 2. Indeed, our CER-based measure shows a very strong correlation with the judgments of test set 1; the correlation with the judgments of set 2 is much smaller, but compared to other methods CER still performs very well. It is remarkable that a simple association-based method such as PMI performs well on both test sets. The context-based approach proposed by Pedersen et al. performs poor in this setting; it seems that the terms describing MeSH concepts do not lead to effective context vectors.

Next, it seems that test set 2 is 'more difficult' than test set 1, since all measures show a drop in correlation. The structure-based measures show an especially sharp drop in performance, perhaps because the MeSH structure does not describe the concept relations as completely as the SNOMED CT ontology.

Our method can be interpreted as a contextual extension to association-based measures: it intrinsically biases concept pairs which have been assigned to the same documents, but it is not limited to this co-occurrence information. Moreover, the lack of dependence on structure shows to be beneficial, especially on the second test set.

	Test set 1		Test set 2	
	K	P	K	P
path	0.566 ‡	0.798 ‡	0.319 †	0.376
Nguyen	0.573 ‡	0.812 ‡	0.290	0.422 †
Lin	0.585 ‡	0.839 ‡	0.285	0.431 †
Dice	0.698 ‡	0.650 ‡	0.486 ‡	0.617 ‡
LLR	0.642 ‡	0.552 ‡	0.469 ‡	0.503 †
PMI	0.711 ‡	0.855 ‡	**0.570** ‡	0.530 ‡
context	0.404 ‡	0.651 ‡	0.364 †	0.486 †
CER	**0.781** ‡	**0.953** ‡	0.537 †	**0.618** ‡

Table 1: Absolute correlation (K = Kendall τ coefficient, P = Pearson's correlation) between metrics and ground truth (best scores are marked in boldface). †/‡: significant correlation at 0.05/0.01 level.

5. CONCLUSION AND FUTURE WORK

We have described a novel concept relatedness measure based on the cross entropy reduction between concept language models. We have shown that the measure performs well on an ontology and document collection from the biomedical domain and is able to outperform other relatedness measures. Future work should point out whether the approach is also useful in more general domains and how robust the method is with respect to the ontology and document collection used.

6. ACKNOWLEDGEMENTS

This work was carried out in the context of the BioRange programme of the Netherlands Bioinformatics Centre, supported by a BSIK grant through the Netherlands Genomics Initiative. Edgar Meij was supported by the Virtual Laboratory for e-Science project. Maarten de Rijke was supported by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 017.001.-190, 640.001.501, 640.002.501, STE-07-012 and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

7. REFERENCES

[1] J. E. Caviedes and J. J. Cimino. Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics*, 37(2):77–85, April 2004.
[2] W. Kraaij. *Variations on Language Modeling for Information Retrieval*. PhD thesis, University of Twente, June 2004.
[3] D. Lin. An information-theoretic definition of similarity. In *ICML '98*, pages 296–304, 1998.
[4] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.
[5] H. Nguyen and H. Al-Mubaid. New ontology-based semantic similarity measure for the biomedical domain. In *2006 IEEE Int. Conf. on Granular Computing*, pages 623–628, 2006.
[6] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40 (3):288–299, 2007.
[7] P. Resnik. Semantic similarity in a taxonomy. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.
[8] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.
[9] D. Trieschnigg and W. Kraaij Hierarchical topic detection in large digital news archives: Exploring a sample based approach. *Journal of Digital Information Management*, 3(1): 21–26, 2005.