# Omiotis: A Thesaurus-Based Measure of Text Relatedness

George Tsatsaronis[1], Iraklis Varlamis[2], Michalis Vazirgiannis[3], and Kjetil Nørvåg[1]

[1] Department of Computer and Information Science,
Norwegian University of Science and Technology
[2] Department of Informatics and Telematics, Harokopio University
[3] Department of Informatics, Athens University of Economics and Business

**Abstract.** In this paper we present a new approach for measuring the relatedness between text segments, based on implicit semantic links between their words, as offered by a word thesaurus, namely WordNet. The approach does not require any type of training, since it exploits only WordNet to devise the implicit semantic links between text words. The paper presents a prototype on-line demo of the measure, that can provide word-to-word relatedness values, even for words of different part of speech. In addition the demo allows for the computation of relatedness between text segments.

## 1 Introduction

Text-relatedness can be perceived in several different ways. Primarily, as lexical relatedness or similarity between texts, based on a vectorial representation of texts and a standard similarity measure (e.g. Cosine). Secondly, by capturing the latent semantic relations between dimensions (words) in the constructed vector space model, by using techniques of latent semantic analysis [1]. Another aspect of text relatedness, probably of equal importance, is the semantic relatedness between two text segments. For example, the sentences ”*The shares of the company dropped 14 cents*” and ”*The business institution’s stock slumped 14 cents*” have an obvious semantic relatedness, which traditional measures of text similarity fail to recognize. In this paper we present an on-line demo of a new measure of semantic relatedness between words (SR) and its expansion (Omiotis, from the Greek word for relatedness or similarity) to measure semantic relatedness between text segments. Our measure of relatedness lies in the use of WordNet, and all of its available semantic information. The contribution of this demo is twofold: (a) it can measure the semantic relatedness between words of any part of speech, and consequently between sentences, and (b) it computes relatedness values very fast, based on an index of all the pair-wise relatedness values between WordNet synsets (11 billion combinations). This is the first time, to the best of our knowledge, that such an index has been created.

## 2 Measuring Semantic Relatedness

The expansion of WordNet with semantic relations that cross parts of speech has widened the possibilities of semantic network construction from text. Recent approaches in

W. Buntine et al. (Eds.): ECML PKDD 2009, Part II, LNAI 5782, pp. 742–745, 2009.

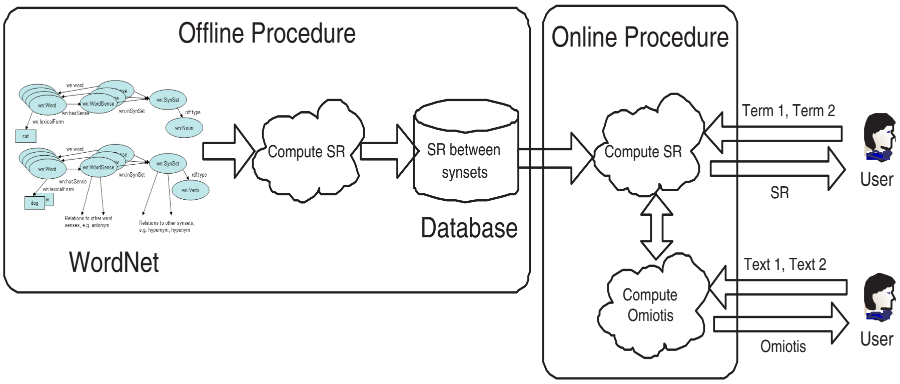


**Fig. 1.** System architecture

semantic network construction from word thesauri, e.g. from Navigli and Velardi [2], utilize all of the semantic relations in WordNet. The applications of these methods in areas like Word Sense Disambiguation highlight the advantages of using the full range of semantic relations that WordNet offers. In this work, we find all the semantic paths that connect two sense nodes in WordNet using all of the provided semantic relations by WordNet. To the best of our knowledge it is the first time that a measure of semantic relatedness combines three factors in tandem: (a) path length connecting concepts; (b) concepts' depth in the used thesaurus, and (c) thesaurus' edges importance. An analysis of state of the art measures that use semantic information from word thesauri can be found in [3]. Figure 1 shows the architecture of the developed system. For the computation of the semantic relatedness between two terms (**SR**), we first compute the semantic relatedness values for all pairwise combinations of their senses. The semantic relatedness between a pair of senses considers the path length, captured by compactness, and the path depth, captured by semantic path elaboration as explained in details in [4]. The values of SR range in [0,1]. In the case when only one of the terms exists in WordNet, the semantic relatedness between them is 0. If both terms are the same, and this term exists in WordNet, then it is considered 1.

Following the work we presented in [5], the measure of semantic relatedness between text segments (**Omiotis**) perceives texts as sets of terms (*bag of words*) in a vector space, and uses TF-IDF for term weighting. Omiotis value $O_{A,B}$ for a pair of texts $A$ and $B$ is defined as:

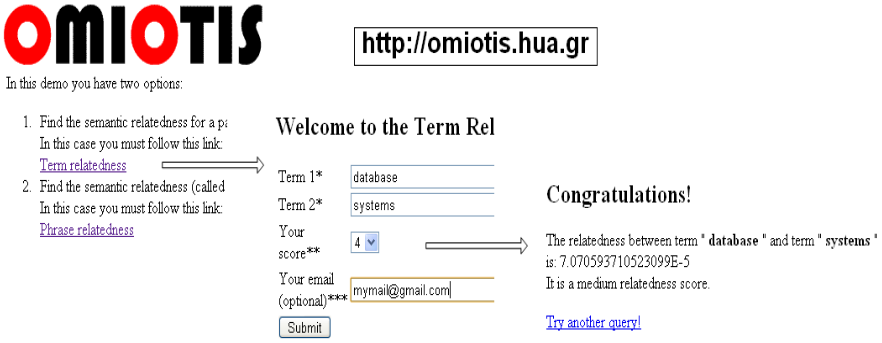
$$O_{A,B} = \frac{1}{2}\left[\frac{1}{|A|}\left(\sum_{i=1}^{|A|}\lambda_{i,x(i)} \cdot SR(k_i, h_{x(i)})\right) + \frac{1}{|B|}\left(\sum_{j=1}^{|B|}\lambda_{y(j),j} \cdot SR(k_{y(j)}, h_j)\right)\right] \quad (1)$$

In the above equation, $SR(t_i, t_j)$ is considered as the semantic relatedness between terms $t_i$ and $t_j$, as defined previously, $\lambda_{t_i,t_j}$ is the sum of the terms' TF-IDF values, $k_i$ the $i$th term of $A$, $h_i$ the $i$th term of $B$ and $x(i)$ and $y(j)$ are defined respectively:

$$x(i) = \underset{j \in [1,|B|]}{\arg\max}(\lambda_{i,j} \cdot SR(k_i, h_j)) \; and \; y(j) = \underset{i \in [1,|A|]}{\arg\max}(\lambda_{i,j} \cdot SR(k_i, h_j)) \quad (2)$$

## 3 On-line Demo

The computation of Omiotis entails a series of steps, the complexity of which is strongly related to the $SR$ measure. In order to improve the system's scalability, we have pre-computed and stored all $SR$ values between every possible pair of synsets in a RDBMS. This is a one-time computation cost, which dramatically decreases the computational complexity of Omiotis. The database schema has three entities, namely *Node*, *Edge* and *Paths*. *Node* contains all WordNet synsets. *Edge* indexes all edges of the WordNet graph adding weight information for each edge computed using the SR measure. Finally, *Paths* contains all pairs of WordNet synsets that are directly or indirectly connected in the WordNet graph and the computed relatedness. These pairs were found by running a Breadth First Search (BFS) starting from all WordNet roots for all POS. The RDBMS, which exceeds 220 Gbytes in size, has indexed in total $155.327$ unique synsets, whereas the number of processed edges is $324.268$. In total, the number of processed synset pairs exceeds the $11$ billion combinations. The current implementation takes advantage of the database structures (indices, stored procedures etc) in order to decrease the running time of Omiotis. Because we have pre-computed the relatedness values for all synset pairs, the time required for processing $100$ pairs of terms is $\simeq 1$ sec, which makes the computation of Omiotis feasible and scalable. As a proof of concept, we have developed an on-line demo version of the SR and the Omiotis measures (publicly available at http://omiotis.hua.gr), where the user can test the term-to-term and sentence-to-sentence semantic relatedness measures. Figure 2 shows a walk-through example of the demo for computing relatedness between *database* and *systems*. Similarly, the user can compute text-to-text relatedness. Once the relatedness computation takes place for the desired input, a screen showing the relatedness value appears to the user's browser.



OMIOTIS

http://omiotis.hua.gr

In this demo you have two options:

1. Find the semantic relatedness for a pa
   In this case you must follow this link:
   Term relatedness
2. Find the semantic relatedness (called
   In this case you must follow this link:
   Phrase relatedness

Welcome to the Term Rel

Term 1* database
Term 2* systems
Your score** 4
Your email (optional)*** mymail@gmail.com

Congratulations!

The relatedness between term " **database** " and term " **systems** " is: 7.070593710523099E-5
It is a medium relatedness score.

Try another query!

**Fig. 2.** Omiotis on-line demo

## 4 Applications

SR and Omiotis have been evaluated in several different text related tasks. Initially, SR has been evaluated as a measure for Word Sense Disambiguation in [4]. The results showed that the measure produces state of the art precision in the Senseval competitions (*all English words* task). Furthermore, SR has been evaluated in measuring word-to-word relatedness, in three widely used data sets [4], where experiments showed that it produces the highest Spearman rank order correlation coefficient compared to the human judgements, than any other dictionary-based measure [3]. In addition, Omiotis has been embedded in the text retrieval task [6], by using the implementation of the measures described in the next section, inside the TERRIER retrieval platform. Experiments in three TREC collections show that Omiotis can boost retrieval performance by even up to 2% compared to traditional retrieval models that do not take into account semantic information from text. Finally, measures of semantic relatedness have been embedded in the past in many different linguistic exercises, like for example the SAT analogy tests and the TOEFL synonym questions (consult http://www.aclweb.org/aclwiki/), as well as in paraphrase recognition [7]. From the aforementioned, it is induced that both SR and Omiotis can be embedded in a variety of applications, mainly due to the fact that they can measure relatedness for words of all parts of speech. Regarding related systems, the offered functionality of computing word-to-word relatedness by SR is also offered by the *WordNet::Similarity* software package [8], which implements a wide range of measures, but as our experimental analysis in [4] shows, SR outperforms all of these measures in capturing semantic relatedness compared to the human judgements in three data sets. The functionality offered by Omiotis, in computing text relatedness between text segments, taking into account semantic information from WordNet, is not offered, to the best of our knowledge, by another on-line system. Finally, as far as WordNet synset relatedness values is concerned, currently there is no other on-line system that has indexed all the possible WordNet synset combination relatedness values.

## References

1. Landauer, T., Foltz, P., Laham, D.: Introduction to latent semantc analysis. Discourse Processes 25, 259–284 (1998)
2. Navigli, R., Velardi, P.: Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. IEEE Trans. Pattern Anal. Mach. Intell. 27(7), 1075–1086 (2005)
3. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics 32(1), 13–47 (2006)
4. Tsatsaronis, G., Varlamis, I., Vazirgiannis, M.: Word sense disambiguation with semantic networks. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2008. LNCS (LNAI), vol. 5246, pp. 219–226. Springer, Heidelberg (2008)
5. Varlamis, I., Vazirgiannis, M., Halkidi, M., Nguyen, B.: Thesus: Effective thematic selection and organization of web document collections based on link semantics. IEEE TKDE Journal 16(6), 585–600 (2004)
6. Tsatsaronis, G., Panagiotopoulou, V.: A generalized vector space model for text retrieval based on semantic relatedness. In: Proc. of the 12th EACL (SRW session), pp. 70–78 (2009)
7. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: Proc. of the 21st AAAI (2006)
8. Patwardhan, S., Banerjee, S., Pedersen, T.: Senserelate:targetword - a generalized framework for word sense disambiguation. In: Proc. of AAAI (demo session), pp. 1692–1693 (2005)