HAL
open science

A semantic similarity framework exploiting multiple parts-of-speech

Giuseppe Pirrò, Jérôme Euzenat

To cite this version:

Giuseppe Pirrò, Jérôme Euzenat. A semantic similarity framework exploiting multiple parts-of-speech. Proc. 9th international conference on ontologies, databases, and applications of semantics (ODBASE), Oct 2010, Heraklion, Greece. pp.1118-1125, 10.1007/978-3-642-16949-6_33 . hal-00793282

HAL Id: hal-00793282

https://inria.hal.science/hal-00793282v1

Submitted on 22 Feb 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Semantic Similarity Framework Exploiting Multiple Parts-of Speech

Giuseppe Pirró*, Jérôme Euzenat

INRIA Grenoble Rhône-Alpes & LIG, Montbonnot, France
{*Giuseppe.Pirro*, *Jerome.Euzenat*}@*inrialpes.fr*

Abstract. Semantic similarity aims at establishing resemblance by interpreting the meaning of the objects being compared. The Semantic Web can benefit from semantic similarity in several ways: ontology alignment and merging, automatic ontology construction, semantic-search, to cite a few. Current approaches mostly focus on computing similarity between nouns. The aim of this paper is to define a framework to compute semantic similarity even for other grammar categories such as verbs, adverbs and adjectives. The framework has been implemented on top of WordNet. Extensive experiments confirmed the suitability of this approach in the task of solving English tests.

Key words: Semantic Similarity, Feature Based Similarity, Ontologies, Synonymy detection

1 Introduction

Similarity gives an estimation of to what extent two or more objects are alike. It is especially useful when there is only a partial knowledge between the objects being compared and is one of the pillar of important processes such as memory, categorization, decision making, problem solving, and reasoning [17]. The origin of similarity studies has to be found in psychology and cognitive science where different models have been postulated. Similarity found its way different different areas ranging from databases [6] to distributed systems [3]. The Semantic Web is one of the most active community in which similarity has been extensively used. For instance, similarity helps to compute mappings between different ontologies [13], repair ontology mappings [10] or compute similarity between ontologies. In information retrieval, similarity is used to complement the vector-space model [21] while in natural language processing it is useful, for instance, in word sense disambiguation [7]. In artificial intelligence, there have been defined several ways of computing similarity. However, two main branches can be identified. On one hand, knowledge-base methods exploit some semantic artefacts (e.g., WordNet) encoding human knowledge; here similarity is computed by investigating how the two entities being compared are arranged in the considered structure. A striking observation is that existing distance or similarity measures are only applicable to the hierarchical relations, which makes them only applicable to some syntactic

* This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

categories (e.g., nouns and verbs). On the other hand, statistic-based methods (see for instance [19, 2]) compute similarity by analyzing the co-occurrences of the two words being compared in large corpora (e.g., documents indexed by a search engine).

The main contribution of this paper is the definition of a general framework for computing semantic similarity between words, which takes into account relations both within and across different parts-of-speech. As will be explained later, a feature model to represent objects, that is, nouns, verbs, adjectives and adverbs is presented, which stems from our previous work [12]. Upon this model a framework to compute semantic similarity is presented. By using this framework all the existing similarity measures can be augmented to work with multiple parts-of-speech as well. To have an insight, if one were calculating the similarity between *democratic* and *liberal*, using existing methods it would not be possible. However, if we observe that *democratic* and *liberal* are related to *democracy* and *liberty* in the *noun* taxonomy we can compute their similarity. We adopt WordNet [11] as reference ontology since it it provides the most comprehensive representation of lexical knowledge ontologically encoded. The applications of this work can be several, among which word similarity, synonymy recognition, document summarization and clustering, ontology mapping and automatic thesauri construction.

The remainder of this paper is organized as follows. Section 2 provides some background on WordNet. Section 3 surveys on popular similarity measures both knowledge-based and statistic-based. Section 4 presents the similarity framework and the logical path toward its definition. Section 5 discusses the evaluation of the framework while Section 6 concludes the paper.

2 Background on WordNet

The WordNet ontology O is a graph, where nodes represent concepts and edges encode relations between concepts. There are several semantic relations that connect nodes, referred to as *synsets* i.e., sets of similar entities. WordNet gives synset definitions of four different parts-of-speech, that is, *nouns*, *verbs*, *adjectives* and *adverbs*. However, only nouns and verbs are arranged in a taxonomic structure; adjectives and adverbs are defined both in terms of relations with the same part of speech and with the noun and verb taxonomies. Each synset (S) definition, in WordNet has the following form:

$$S_x =< id, W, R, g > \tag{1}$$

where id is a unique identifier, the set W contains pairs of the form $W =< w, n >$, where w is a word in the synset and n is the *sense number* for this word. Moreover, the set R contains pairs of the form $r =< s_r, id_r, pos >$ where s_r is the type of semantic relation that relates the given synset with the target synset id_r and pos is the part of speech of id_r. Finally, g is the gloss for the synset, which is a description in natural language. The set R is different depending on the part-of speech considered.

2.1 Features

To better understand the reasoning that motivates the present work, the notion of feature has to be introduced. An object *feature* (a word in our case) can be seen as a property of the object. In the case of *nouns* and *verbs*, words in the hierarchy inherit all the feature of their superordinate even if they can have their own specific features. As an example, in the WordNet noun taxonomy, since *car* and *bicycle* both serve to transport people or objects, in other words they are both types of vehicles, they share all features pertaining to *vehicle*. However, each word has also its specific features as *steering wheel* for *car* and *pedal* for *bicycle*. Again, in the verb taxonomy, the verb *dress* inherits all the features of its superordinate i.e., *wrap up*.

Since this work aims at exploiting the relations between parts-of-speech, the notion of feature has to be extended to also encompass relations across parts-of-speech. In this setting, the features of the adjective *active*, for instance, include its relations in the noun taxonomy with *action* as well. We will discuss in more detail such reasoning later in Section 4. For the time being, the main intuition is that in this work the feature-based model postulated by Tversky, will be projected in the information-theoretical model introduced by Resnik for the purpose of computing similarity taking into account multiple parts-of speech.

3 Related work

This section describes some well-known similarity measures along with two of the most prominent statistic-based approaches, that is, PMI-IR, and Normalized Google Distance (NGD).

Information Theoretic Approaches. Information theoretic approaches to semantic similarity employ the notion of Information Content (IC), which quantifies the informativeness of concepts. IC values are obtained by associating probabilities to each concept in an ontology on the basis of its occurrences in large text corpora. In the specific case of hierarchical ontologies, these probabilities are cumulative as we travel up from specific concepts to more abstract ones. This means that every occurrence of a concept in a given corpus is also counted as an occurrence of each class containing it. Resnik [15] was the first to leverage IC for the purpose of semantic similarity. Resnik's formula to compute similarity states that similarity depends on the amount of information two concepts share, which is given by the Most Specific Common Abstraction (*msca*), that is, the concept that subsumes the two concepts being compared. Starting from Resnik's work two other similarity measures were proposed. The first, by Jiang and Conrath [5] and the second by Lin [9]. Both measures leverage IC-values calculated in the same manner as proposed by Resnik. The improvement with these measures is that they correct some problems with Resnik's similarity measure by considering the IC of the two concepts as well.

Ontology based approaches. As for ontology based approaches, the work by Rada et al. [14] is similar to the Resnik measure since it also computes the

msca between two concepts, but instead of considering the IC as the value of similarity, it considers the number of links that were needed to attain the *msca*. Obviously, the less the number of links separating the concepts the more similar they are. The work by Hirst et al. is similar to the previous one but it uses a wider set of relations in the ontology (e.g., part-of) coupled with rules restricting the way concepts are transversed [4].

Hybrid approaches. Hybrid approaches usually combine multiple information sources. Li et al. [8] proposed to combine structural semantic information in a nonlinear model. The authors empirically defined a similarity measure that uses shortest path length, depth and local density in a taxonomy. In [22] the *OSS* semantic distance function, combining *a-priori* scores of concepts with concept distance, is proposed.

3.1 Statistic-based approaches.

PMI-IR [19] is a unsupervised learning algorithm for recognizing synonyms, based on statistical data acquired by querying a Web search engine. PMI-IR uses Pointwise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words. NGD [1] distance is a measure of semantic relatedness, which is computed by considering the number of hits, for a set of keywords, returned by the Google search engine. The intuition is that words with the same or similar meanings are close in terms of Google distance whereas words not related in meaning are distant.

4 A general framework for computing similarity

This section describes the framework devised to compute semantic similarity by exploiting multiple parts-of speech. The main idea is to consider each synset definition and to complement it with information that can be inferred following its relations with the same or other parts-of-speech.

4.1 The general similarity framework

We can see a WordNet synset description along with the relations with other synsets as the definition of its features (see Section 2.1). For instance, the WordNet definition of the adjective *democratic* includes both relations with other adjectives (with same part of speech) and a relation with the noun *democracy*. The framework relies on the feature based model proposed by Tversky [20], which states that similarity depends from the presence/absence of certain qualitative features. According to the feature-based model, the similarity of a concept c_1 to a concept c_2 is a function of the features common to c_1 and c_2, those in c_1 but not in c_2 and those in c_2 but not in c_1. In our case, each definition, in terms of synset, has to take into account features that can be derived both from the same and related parts-of speech. Thus, in our previous example, the features of the adjective *democratic*, should also encompass the features derived from the noun *democracy*. This consideration is particularly useful if one were computing the semantic similarity with another adjective e.g., *liberal* since *liberal*, on its turn,

is related to the noun definition of *liberality*. Without this enrichment, existing approaches would fail in computing semantic similarity, since adjectives are not arranged in a taxonomy differently from names or verbs.

In more detail, the cornerstone of the proposed similarity framework is the *msca* in the IC domain, which reflects the information shared by two concepts c_1 and c_2 and that in a feature-based formulation of similarity can be seen as the intersection of features from c_1 and c_2. Starting from this assumption, it is immediate to infer that the set of features specific to c_1 (resp. c_2) is given by $IC(c_1) - IC(msca)$ (resp. $IC(c_2) - IC(msca)$) in the information content formulation. A more comprehensive discussion about the mapping between features and IC is provided in our previous paper [12]. Once the IC of the two concepts and that of the msca are available, one can exploit existing IC-based similarity measures to compute similarity. At this point, what is needed is a method to compute the IC of words by taking into account relations both with the the same and other parts of speech.

4.2 Information Content mapping to features

Similarity measures based on IC, usually obtain IC-values by parsing large text corpora and counting occurrences of words as discussed in Section 3. This has two main advantages; on one hand it requires time and on the other hand it may be corpus dependent. In [18] a new way of obtaining IC-values directly from a taxonomic structure, called intrinsic Information Content *iIC* is discussed. We extend the idea of *iIC* to adjectives and adverbs by taking into account their relations with nouns and verbs. As discussed before, adjectives and adverbs are related to nouns and verbs by semantic relations enabling to assess features of each synset, in terms of IC, that can be exploited to compute semantic similarity. In particular, for each adjective and adverb synset, the multi part of speech IC (IC_m) for each synset S is defined as follows:

$$IC_m(S) = \sum_{j=1}^{m} \frac{\sum_{k=1}^{n} iIC(c_k \in C_{R_j})}{|C_{R_j}|} \ . \tag{2}$$

This formula takes into account all the m kinds of relations that connect a given adjective or adverb synset S with nouns and verbs. In particular, for all the synsets at the other end of a particular relation (i.e., each $c_k \in C_{R_j}$) the average iIC is computed. This enables to take into account the expressiveness of an adjective or adverb in terms of its relations with nouns and verbs.

At this point, each IC-based similarity measure can be rewritten by using the IC_m definition described in equation (2). It is important to point out that the similarity measures considered in our evaluation were originally formulated to work only with noun definitions apart from the Resnik measure, which has been evaluated on verbs as well [16] even if IC values were obtained in the classical manner, that is, by word counting.

5 Evaluation

This section discusses the evaluation of the proposed framework to compute semantic similarity using different parts-of-speech. In particular, three existing similarity measures (described in Section 3) based in IC have been rewritten according to the proposed framework. For the evaluation, we implemented the Similarity Based English Test Solver (SB-ETS), which is useful in the task of meaning recognition and synonymy detection. Given a base word and four choices, SB-ETS returns the most similar word. For each of the four considered datasets, the percentage of correct answers has been calculated.

5.1 English vocabulary test evaluation

The similarity measures have been evaluated against PMI-IR and NGD for which we considered two different search engines i.e., Google (G) and Yahoo (Y). This can give an insight of how much these approaches depends on the search algorithm implemented by the search engine an the amount of data the search engine indexes. For PMI-IR we considered the best results obtained by using three different types of score described in [19]. We performed evaluations by also adopting tagging, stemming and elimination of stopwords in the case of sentences. As a tagger we used the JMontyTagger [1] whereas a basic stemmer has been implemented. The results obtained for each of the considered similarity measures along with the time elapsed for each evaluation are reported in Table 1 (with tagging) and Table 2 (without tagging). In the column *Na* it is indicated the number of tests for which it has not been computed the result since the words were not found in WordNet. Table 3 reports the results for statistic-based approaches.

Table 1. Results for similarity measures with tagging.

	VOA			**TOEFL**			**Sat**			**GRE**			**GMAT**			**D5**		
	P	**Na**	**t(s)**	**P**	**Na**	**t(s)**	**P**	**Na**	**t(s)**	**P**	**F**	**t(s)**	**P**	**Na**	**t(s)**	**P**	**Na**	**t(s)**
Res	0.6	0	18	0.6	1	5	0.9	0	11	0.5	2	4	0.6	2	5	0.5	11	161
J&C	0.6	0	18	0.5	1	3	0.8	0	8	0.4	2	3	0.6	2	3	0.5	11	181
Lin	0.6	0	18	0.6	1	4	0.9	0	8	0.5	2	3	0.6	2	3	0.5	11	244

As can be observed, the evaluations performed after tagging the words being compared are poorer as compared to those in which tagging was not performed apart from the VOA dataset. This result may depend on the performance of the tagger used. In the case of not tagging, all the parts-of-speech of a given word have been considered. Performance of similarity measures are different depending on the considered test. For instance, the VOA test seems to be the most difficult; here the precision of similarity measures range from 0.5 for the Resnik measure to 0.6 for the J&C measure in the case of not tagging whereas it is 0.6

[1] http://web.media.mit.edu/∼hugo/montylingua/

Table 2. Results for similarity measures without tagging.

	VOA			TOEFL			Sat			GRE			GMAT			D5		
	P	**Na**	**t(s)**	**P**	**Na**	**t(s)**	**P**	**Na**	**t(s)**	**P**	**F**	**t(s)**	**P**	**Na**	**t(s)**	**P**	**Na**	**t(s)**
Res	0.5	0	43	0.8	0	9	0.8	0	12	0.9	0	5	0.9	0	5	0.6	24	184
J&C	0.6	0	41	0.6	0	7	0.8	0	11	0.8	0	6	0.9	0	5	0.7	24	186
Lin	0.5	0	47	0.8	0	7	0.8	0	11	0.9	0	6	0.9	0	5	0.7	24	208

Table 3. Results for statistic-based methods.

	VOA			TOEFL			Sat			GRE			GMAT			D5		
	P	**Na**	**t(s)**	**P**	**Na**	**t(s)**	**P**	**Na**	**t(s)**	**P**	**F**	**t(s)**	**P**	**Na**	**t(s)**	**P**	**Na**	**t(s)**
PMI-IR-G	0.5	0	82	0.6	0	80	0.7	0	68	0.6	0	95	0.8	0	64	0.5	0	3024
NGD-G	0.6	0	85	0.5	0	85	0.3	0	70	0.6	0	61	0.5	0	70	0.4	0	2134
PMI-IR-Y	0.4	0	287	0.5	0	243	0.6	0	254	0.7	0	262	0.7	0	290	0.5	0	4778
NGD-Y	0.4	0	380	0.5	0	373	0.4	0	289	0.5	0	248	0.5	0	261	0.4	0	2754

for all measures in the case of tagging. Similarity measures perform better in the GMAT test where the value of 0.9 is reached. In the D5 dataset, which includes 300 tests, the Lin measure performs better than the other. Statistic based approaches are overcame by similarity measure in all tests. PMI-IR exploiting Google performs better than the other statistic approaches. The advantage of all these approaches is the wider coverage in terminology; in fact the number of not answered tests is equal to 0 in each case whereas the number of tests not answered by similarity measures in D5 is 24. Another comparison between similarity measures and statistic-based approaches can be done in terms of time elapsed. Similarity measures clearly are faster than statistic-based approaches. In the D5 test, which is the largest, the time elapsed for all the 300 tests ranges from 184 secs for the Resnik measure to 4788 secs for the NGD exploiting Yahoo as search engine. This indicates that in several applications such as document clustering or text similarity, statistic-based approaches result to be unusable.

6 Concluding Remarks and Future Work

This paper described a framework to compute similarity between words belonging to different parts-of-speech. To have an insight of how the framework performs we considered automatic scoring of English tests, such as the well known TOEFL. An extensive evaluation followed by a comparison with statistic-based approaches showed the suitability of the framework. As future work, the main direction is that of investigating similarity measures between words belonging to different parts of speech as for instance the noun *car* and the verb *run*. Besides, other interesting applications could be text summarization or plagiarism detection.

References

1. R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE TKDE*, 19(3):370–383, 2007.
2. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
3. C. Hai, J. Hanhua. Semrex: Efficient search in semantic overlay for literature retrieval. *FGCS*, 24(6):475–488, 2008.
4. G. Hirst and D. St-Onge. *in WordNet: An Electronic Lexical Database*, chapter Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. MIT Press, 1998.
5. J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. ROCLING X*, 1997.
6. V. Kashyap and A. Sheth. Schematic and semantic similarities between database objects: A context-based approach. *VLDB Journal*, 5:276–304, 1996.
7. C. Leacock, M. Chodorow, and G. A. Miller. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.
8. Y. Li, A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE TKDE*, 15(4):871–882, 2003.
9. D. Lin. An information-theoretic definition of similarity. In *Proc. of Conf. on Machine Learning*, pages 296–304, 1998.
10. C. Meilicke, H. Stuckenschmidt, and A. Tamilin. Repairing ontology mappings. In *AAAI*, pages 1408–1413, 2007.
11. G. Miller. Wordnet an on-line lexical database. *International Journal of Lexicography*, 3(4):235–312, 1990.
12. G. Pirrò. A semantic similarity metric combining features and intrinsic information content. *Data Knowl. Eng.*, 68(11):1289–1308, 2009.
13. G. Pirrò, M. Ruffolo, and D. Talia. Secco: On building semantic links in peer to peer networks. *Journal on Data Semantics*, XII:1–36, 2008.
14. R. Rada, H. Mili, and M. Bicknell, E. andBlettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:17–30, 1989.
15. P. Resnik. Information content to evaluate semantic similarity in a taxonomy. In *Proc. of IJCAI 1995*, pages 448–453, 1995.
16. P. Resnik and M. Diab. Measuring verb similarity. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 399–404, 2000.
17. B. Schaeffer and R. Wallace. Semantic similarity and the comparison of word meanings. *J. Experiential Psychology*, 82:343–346, 1969.
18. N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *Proc. of ECAI 2004*, pages 1089–1090, 2004.
19. P. D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *ECML*, pages 491–502, 2001.
20. A. Tversky. Features of similarity. *Psychological Review*, 84(2):327–352, 1977.
21. G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis, and E. E. Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *WIDM*, pages 10–16. ACM, 2005.
22. V. S. Zuber and B. Faltings. Oss: A semantic similarity function based on hierarchical ontologies. In *IJCAI*, pages 551–556, 2007.