Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 96 (2016) 365 – 374
Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Improving semantic relatedness assessments: ontologies meet textual corpora

Montserrat Batet[a]*, David Sánchez[b]

[a]*Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss, 5, Parc Mediterrani de la Tecnologia, 08860 Castelldefels, Catalonia (Spain)*
[b]*UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Catalonia (Spain)*

Abstract

Even though the calculation of the semantic similarity between textual entities has received a lot of attention by the research community, the more general notion of semantic relatedness (which considers both taxonomic and non-taxonomic knowledge) has been significantly less studied and, in general, stays one step behind in terms of accuracy. In this paper, we improve semantic relatedness assessments by aggregating the highly-accurate ontology-based estimation of semantic similarity with the distributional resemblance of textual terms computed from large textual corpora. As a result, our approach is able to improve the accuracy of related works on a standard benchmark.

Peer-review under responsibility of KES International

Keywords: knowledge, semantics, semantic relatednes; semantic similarity; ontologies; textual information distribution.

1. Introduction

To automatically manage, classify and exploit textual data, researchers require from computerized mechanisms able to *understand* textual resources. In this context, the estimation of the semantic resemblance between textual terms is a fundamental issue. The research community distinguishes two main notions. On the one hand, *semantic*

* Corresponding author. Tel.: +34 932535714; fax: +34 977559710.
E-mail address: mbatetsa@uoc.edu

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.149

similarity is understood as the degree of taxonomic resemblance between two textual terms or concepts [1] (e.g. *horse* and *zebra* are semantically similar because both are *equines*); being an essential metric for data classification, many researchers have proposed metrics to automatically compute semantic similarities [2-5]. On the other hand, we find the more general notion of *semantic relatedness*, which is the focus of this paper and that captures the semantic resemblance between terms as a function of both their taxonomic and non-taxonomic (e.g., meronymy, functionality, cause-effect, etc.) relationships [1]; for example, *horse* is semantically related to *stable*, even though they are not semantically similar.

To estimate semantic similarity/relatedness, researchers extract and exploit the explicit or implicit knowledge available in different kinds of (linguistic) sources. On the one hand, ontology-based measures rely on the structured domain knowledge offered by ontologies (i.e., concepts and semantic relationships), which may be complemented with the distribution of linguistic data in textual corpora. On the other hand, distributional approaches exclusively rely on textual corpora and assume that textual terms with similar distributions (i.e., those that tend to co-occur) have similar meanings [6].

By reviewing the current literature [1] (see also Section 2), we can conclude that semantic similarity measures mostly rely on ontologies, because the latter provide the unambiguous and well-defined taxonomic structures/knowledge needed to assist similarity assessments. On the contrary, distributional measures naturally capture the more general notion of semantic relatedness because they do not usually distinguish the type of semantic relationship underlying term co-occurrence; for example, *flu* and *cold* usually co-occur because both are taxonomically similar *diseases of the respiratory system*, but *flu* and *fever* (being taxonomically distant) also usually co-occur because the latter is a symptom of the former. Semantic relatedness has been less studied than semantic similarity, both because of its higher complexity (since different kinds of semantic relationships should be considered in aggregate), and because non-taxonomic knowledge/relationships are rarely modeled in ontologies [7] (due to the burden that the manual modeling of the potentially large number of non-taxonomic relationships implies). Moreover, because distributional measures rely on large electronic textual corpora (such as the Web) to capture the information distribution at a social scale [8], they are hampered by the ambiguity (i.e., polysemy or synonymy) underlying plain natural language text. Due to these reasons, semantic relatedness measures stay, generally, one step behind of similarity measures in terms of accuracy [9-13].

In this paper, we tackle the limitations of (distributional) semantic relatedness measures by relying on the following intuitive idea: because semantic relatedness is a function of both the taxonomic and non-taxonomic resemblance between terms, we propose computing their semantic relatedness as the aggregation of their (strictly taxonomic) similarity, which we estimate from the structured knowledge provided by an ontology, and their (either taxonomic and non-taxonomic) distributional resemblance computed from a large textual corpus. With a semantically and numerically coherent aggregation of both dimensions, we aim at improving semantic relatedness assessments by taking advantage of the higher accuracy that characterize ontology-based similarity measures and compensating (some of) the limitations of strictly distributional relatedness measures. Our approach has been evaluated and compared with related works on a standard benchmark achieving state of the art accuracy.

The rest of the paper is organized as follows. Section 2 surveys and discusses related works on semantic similarity/relatedness assessment. Section 3 presents our proposal, which combines the high accuracy of ontology-based similarity assessment and the more general notion of relatedness captured by data distributions. Section 4 evaluates and compares the accuracy of our approach on a standard benchmark. The final section contains the conclusions and depicts some lines of future research.

2. Related work

Being semantic similarity the most studied paradigm in the literature, a plethora of similarity measures have been proposed in the last years. These can be classified into different calculation paradigms according to the kind of semantic evidences and knowledge sources they use to perform the similarity assessment.

Edge-counting measures evaluate similarity as the inverse to the number of *is-a* edges separating two terms modeled as concepts in an ontology [2, 14-16]. These measures are intuitive and easy to implement, but their accuracy is limited by the fact that *just* the shortest taxonomic path between the two concepts is considered; this

omits other taxonomic knowledge (e.g., taxonomic ancestors in paths other than the shortest one) that are usually available in current ontologies [3].

Feature-based measures overcome this limitation by quantifying similarity according to the number of taxonomic ancestors that the two concepts have or not have in common [17, 18], which exploit much more taxonomic knowledge than edge-counting measures.

Finally, *Information Content-based measures* quantify the similarity according to the amount of information (i.e., Information Content (IC)) that the concepts to be compared have in common, which is represented by the IC of the most concrete concept that is a taxonomic ancestor to both concepts, that is, their Least Common Subsumer (LCS) [19]; this value is usually normalized by the informativeness of the individual concepts [20, 21].

The informativeness of concepts has been usually computed as a function of the probability of occurrence of these concepts in textual corpora [19-21]. However, since corpora contain terms that can be ambiguous (e.g., polysemous), term occurrences should be properly disambiguated in order to accurately compute concept probabilities. Moreover, corpora-based IC calculus requires from large and heterogeneous enough tagged corpora to avoid data sparseness, which may not be available for all domains of knowledge. To overcome these problems, in recent years, several authors have proposed estimating concept probabilities from the number of taxonomic descendants (and/or ancestors) modeled in an ontology [11, 22, 23]. The underlying idea is that concepts with a large number of hyponyms (respectively, a small number of ancestors) appear more frequently in textual corpora, because they are general and can be referred by means of any of their hyponyms. This approach overcomes most of the limitations of corpus-based IC calculation (specifically, the need of a tagged corpus and data sparseness and language ambiguity issues). Moreover, thanks to the large amount of taxonomic knowledge that this calculation exploits (i.e., taxonomic ancestors and specializations), ontology-based IC-based similarity measures tend to provide state of the art accuracy, as it has been shown in recent empirical studies [4, 22, 23].

Semantic relatedness, on the other hand, usually relies on the information distribution in textual corpora to infer the degree of semantic relationship (of any kind) between two textual entities. These *distributional* approaches can be classified into *first order co-occurrence measures*, which compute relatedness as a function of the probability of explicit co-occurrence of terms [8, 10, 24] and *second order co-occurrence measures*, which estimate relatedness as a function of the co-occurrence of the words appearing in the linguistic contexts of the terms to compare [12, 25, 26]; by relying on second order co-occurrences, the latter measures aim at minimizing language ambiguity (because the context of a term helps to semantically disambiguate it) and tackle the issue that, sometimes, semantically related words (such as synonyms or antonyms) do not directly co-occur [27].

Ontology-based relatedness measures, which consider not only taxonomic subsumption but any other semantic relationship modeled in the ontology (e.g., meronymy, holonymy, cause, effect, etc.) are much scarcer. As a matter of fact, in addition to the complexity of integrating taxonomic and non-taxonomic knowledge, available ontologies rarely model a non-taxonomic relationships [7] and, when they do, they tend to offer much less detail than for taxonomic knowledge (e.g., in WordNet, around 90% of all the semantic relationships are taxonomical [28]. Again, we can find measures based on counting semantic edges (in this case, of any kind) [29], those based on aggregating semantic features of different kinds available in the ontology (e.g., ancestors, synonyms, meronyms, glosses, etc.) [30, 31], and IC-based measures that estimate the IC of concepts not only from their set of hyponyms, but also from the set of concepts that are non-taxonomically related with them [32]; as above, in the latter case, the idea is that concepts subsuming a large number of hyponyms (taxonomic IC) and/or (non-taxonomically) semantically related with concepts with large sets of hyponyms (non-taxonomic IC) have a high probability of appearance in general corpora. To integrate taxonomic and non-taxonomic informativeness, the authors use weights that tune the contribution of each dimension to the final relatedness assessment.

3. A new semantic relatedness measure

From the review carried out in the previous section, we can extract the following conclusions. First, relatedness measures have a complexity higher than similarity measures because taxonomic and non-taxonomic knowledge/resemblance should be properly integrated. Moreover, because distributional relatedness measures [8, 10, 24] rely on the linguistic information distribution in large electronic textual corpora (such as the Web), they are usually affected by the language ambiguity (e.g, polysemy and synonymy) that appears when the distribution of

concepts is estimated from the occurrence of (non-disambiguated) words; specifically, polysemous words (that may refer to several concepts) overestimate the probability of occurrence of individual concepts, whereas the fact that a concept may be referred with different synonyms would underestimate the concept occurrence. Finally, ontology-based relatedness measures are hampered by the fact that, in general, ontologies do not model non-taxonomic relationships with enough detail (or, at least with the same detail as for taxonomic relationships) [7, 28]. In consequence, the accuracy reached by relatedness measures is usually lower than that achieved by state of the art similarity measures [1, 9-13].

In order to minimize the above problems, we propose capturing the taxonomic and non-taxonomic dimensions of semantic relatedness in different ways and aggregating them. Our idea is to estimate the semantic relatedness of two terms as the aggregation of their (taxonomic) semantic similarity, which will be computed from the unambiguous and well-defined taxonomic knowledge modeled in ontologies, and their (taxonomic and non-taxonomic) distributional resemblance, which will be computed from the information distribution in a large textual corpus. Our aim is to take advantage of the higher accuracy of ontology-based similarity measures while minimizing the limitations of distributional approaches. To ensure a coherent result, the aggregation of both assessments should be semantically and numerically coherent. Regarding semantic coherence, both assessments should interpret and quantify semantics in a similar way; from a numerical perspective, the values resulting from the two assessments should be in the same numerical scale/range in order to avoid one dominating over the other.

In our approach, the strictly taxonomic resemblance between the terms to compare is assessed according to an ontology-based semantic similarity. As discussed in section 2, from the different similarity calculation paradigms available in the literature, Information Content-based measures that compute ICs intrinsically from the taxonomic structure of the ontology (i.e. ontology-based IC-based similarity measures) achieve the best accuracy [4, 22].

Specifically, we use the well-established IC-based similarity measure proposed by Lin [21] with the Seco et al. [11] approach to intrinsically compute the IC, a combination that, as shown in [22] offered one of the highest accuracies ever reported for a set of evaluation benchmarks.

As discussed in section 2, the informativeness of concepts has been traditionally computed as the inverse of their probability of occurrence in textual corpora [19-21]:

$$IC(a) = -\log p(a) \tag{1}$$

To overcome the limitations of corpora-based IC calculations, Seco et al. [11] proposed estimating concept probabilities for IC calculation (see Eq. (1)) from to the number of hyponyms of the concept in an ontology, as follows:

$$IC_{Seco\_et\_al}(a) = \frac{\log\left(\dfrac{hypo(a)+1}{total\_concepts}\right)}{\log\left(\dfrac{1}{total\_concepts}\right)} = 1 - \frac{\log\left(hypo\left(a\right)+1\right)}{\log\left(total\_concepts\right)}, \tag{2}$$

where $hypo(a)$ is the number of hyponyms in the taxonomic tree strictly below the concept a. a +1 is added to this number in the numerator in order to avoid $log(0)$ when a is a leaf concept of the taxonomy. This value is normalized by the maximum number of hyponyms a concept may have, which corresponds to the hyponyms of the root node of the taxonomy (including itself), that is, the *total_concepts* in the taxonomy.

On the other hand, Lin measures the similarity between concepts a and b according to the ratio between the informativeness of their Least Common Subsumer (LCS(a,b)) (i.e. the most specific concept that subsumes a and b in the taxonomy) and the informativeness of each individual concept. The similarity results are thus in the [0..1] range. By applying the Seco et al.'s IC calculation to Lin's measure, we compute the taxonomic component of our relatedness measure ($rel_{Taxonomic}$) as follows:

$$rel_{Taxomic}(a,b) = sim_{Lin}(a,b) = \frac{2 \times IC_{Seco\_et\_al}(LCS(a,b))}{(IC_{Seco\_et\_al}(a) + IC_{Seco\_et\_al}(b))} \quad (3)$$

Being the taxonomic part of our measure based on IC, the non-taxonomic part, which is based on a distributional measure, should also capture semantic evidences of relatedness in an information theoretic way; this will ensure the semantic coherence between the two dimensions of the relatedness assessment. To the best of our knowledge, the only distributional measure relying on the information theory is the Pointwise Mutual Information (PMI), which quantifies the semantic resemblance between two textual entities *a* and *b* according to their relative probability of co-occurrence in a corpus. Specifically PMI compares the probability of observing *a* and *b* together and observing them independently.

$$PMI(a,b) = \log \frac{p(a,b)}{p(a) \times p(b)} \quad (4)$$

As a matter of fact, PMI can be expressed in terms of IC, as follows:

$$PMI(a,b) = \log \frac{p(a,b)}{p(a) \times p(b)} = \log p(a,b) - \log p(a) - \log p(b) = IC(a) + IC(b) - IC(a,b) \quad (5)$$

In its standard form, PMI yields the following numerical bounds [-∞ … *min*(-log *p*(*a*),-log *p*(*b*))]. In order to be numerically coherent with the similarity values provided by the Lin's measure, which are bounded in the [0..1] range, we used the PMI's normalized form (NPMI) [33], which uses *IC*(*a*,*b*) as normalizing factor. Thus, we compute the non-taxonomic component of our relatedness measure ($rel_{Non\text{-}taxonomic}$) as follows:

$$rel_{Non-Taxonomic}(a,b) = NPMI(a,b) = \frac{PMI(a,b)}{IC(a,b)} \quad (6)$$

Strictly speaking, NPMI provides results in the [-1..+1] range, where 0 means that *a* and *b* are independent (i.e., they co-occur by chance), 1 means that *a* and *b* are perfectly associated (i.e., they always co-occur) and -1 means that *a* and *b* are exclusive (i.e., they never co-occur). In practice, however, because words appearing in a textual context do not co-occur by chance and all words are related up to some degree [27] (i.e., the notion of mutually exclusive words does not exist and, if they do not co-occur in a corpus is because of data sparseness), one can expect that the actual values provided by NPMI for two textual entities are within the positive range.

To compute term probabilities that are representative of the underlying semantics as understood by human beings, we require a textual corpus that is representative of the actual information distribution in society. To do so, we rely on the largest general-purpose electronic textual corpus currently available: the Web. In this respect, former studies have highlighted the suitability of the Web as a faithful representation of the information distribution at a social scale [8]. Moreover, the page count provided by Web Search Engines (WSEs) when querying specific terms can be used as an efficient proxy of the terms' information distribution at a Web scale. In this respect, PMI was the first distributional measure that was adapted to exploit WSEs' hit counts to measure term probabilities [24]. Specifically, the probabilities we require to compute the $rel_{Non\text{-}taxonomic}$ component of our measure are measured as follows:

$$p(a) = \frac{page\_count("a")}{total\_webs}, \quad (7)$$

where $page\_count(a)$ is the number of web sites containing a according to a WSE and *total_webs* is the number of web sites indexed by the WSE. Likewise, to measure the probability of co-occurrence of a and b we use the AND operator of the WSE:

$$p(a,b) = \frac{page\_count("a" AND "b")}{total\_webs} \tag{8}$$

Finally, by aggregating the two components, we define our relatedness measure as follows:

$$relatedness(a,b) = \alpha \times rel_{Taxonomic}(a,b) + (1-\alpha) \times rel_{Non-Taxonomic}(a,b), \tag{9}$$

where α balances the contribution of each dimension to the final relatedness assessment.

4. Evaluation

In this section, we evaluate the accuracy of our measure and compare it against related works on a well-known similarity/relatedness benchmark.

4.1. Benchmark and knowledge sources

In order to conduct objective and reproducible evaluations of similarity/relatedness measures, researchers rely on benchmarks consisting of sets of word pairs whose similarity/relatedness have been agreed among human experts (who act as gold standard). Then, the accuracy of the computerized measures is quantified by calculating the correlation between the results they provide for the word pairs in the benchmark and the ratings provided by the human experts on the same pairs. Specifically, the Pearson's correlation coefficient is commonly used in the literature [3, 11, 31]: the higher the correlation (i.e., the closer to 1), the better the computerized measure mimics human judgments of similarity/relatedness.

In our experiments we have used the WordSim353 benchmark [34, 35], which consists of 353 English word pairs, each associated with a semantic relatedness value resulting from averaging 13 to 16 human ratings. Specifically, human experts were asked to rate the semantic relatedness between the word pairs in the benchmark in a scale from 0 (for totally unrelated words) to 10 (for highly related words or synonyms). It is important to note that the size of this benchmark is one order of magnitude larger than the classic benchmarks used to evaluate similarity (which encompass from 30 to 65 word pairs) [36, 37]; this will make the differences observed among the different measures more statistically significant.

On the other hand, as done in most related works [4, 5, 21, 32], we use WordNet [38] as the ontology to assess the taxonomic dimension of our measure (Eq. (3)). WordNet is a general purpose and domain-independent thesaurus that describes more than 100,000 general concepts, which are structured by means of semantic relations. Although WordNet models some non-taxonomic relations (e.g., meronymy/part-of), the backbone of the semantic structure is the subsumption hierarchy which accounts around 90% of all the semantic relationships [28].

Finally, in order to assess the term probabilities we need for the non-taxonomic dimension of our measure (Eq. (6)), we use Bing WSE (http://www.bing.com), and set the *total_webs* constant (Eqs. (7) and (8)) to 15 billions (according to http://www.worldwidewebsize.com/).

4.2. Results

Our measure has been compared with several representative similarity/relatedness measures relying on different calculation paradigms (see Section 2). Table 1 depicts the Pearson correlation achieved for each measure on the evaluation benchmark. In all cases, WordNet was used as the ontology to guide the assessments.

Rows 1 to 3 show the correlation achieved by three ontology-based similarity measures: Wu & Palmer [2] and Leacock & Chodorow [15], which are based on counting taxonomic edges, and the IC-based measure by Lin [21] with the Seco et al.'s [11] intrinsic IC calculation. Note that the latter corresponds to the taxonomic component of our relatedness measure (Eq. (3)).

Rows 4 to 10 depict the correlation of seven semantic relatedness measures. The first one is the Web-based NPMI, which is a distributional first order co-occurrence measure and corresponds to the non-taxonomic dimension of our measure (Eq. (6)). The second one is the Gloss Overlap measure [25], which is a second order co-occurrence distributional measure that uses WordNet glosses instead of a plain textual corpora to compute relatedness. The remaining ones are IC-based measures (Resnik's [19], Jiang & Conrath's [20], Lin's [21], P&S [39] and FaITH [32]), which were adapted by Pirró & Euzenat in [32] to measure ontology-based relatedness. This was done by extending the intrinsic IC calculus in which the measure rely, not only to hyponyms but also to non-taxonomically related concepts (see Section 2). To integrate the taxonomic IC estimation (based on the number of hyponyms) and the non-taxonomic ones (based on the average intrinsic informativeness of non-taxonomically related concepts), the authors in [32] gave a weight of 0.4 to the taxonomic part and 0.6 to the non-taxonomic one. The results for all these relatedness measures except for the NPMI are those reported in [32] for the WordSim353 benchmark.

The correlation achieved by our measure is depicted in the last row of Table 1. To enable a fair comparison with the former relatedness measures, which represent our closest related works, we set α=0.4 in Eq. (9) so that, as in [32], we set a weight of 0.4 for the taxonomic part and of 0.6 for the non-taxonomic one. Moreover, and also as in [32], this value of α results in the best results for our method.

Table 1. Pearson correlation for several similarity/relatedness measures for the WordSim353 benchmark.

Measure	Pearson correlation
Wu & Palmer	0.32
Leacock & Chodorow	0.36
Lin (with Seco et al. IC calculus)	0.36
NPMI	0.34
Gloss Overlap	0.21
Resnik (with Pirró & Euzenat extended IC)	0.40
Jiang & Conrath (with Pirró & Euzenat extended IC)	0.40
Lin (with Pirró & Euzenat extended IC)	0.40
P&S (with Pirró & Euzenat extended IC)	0.41
FaITH (with Pirró & Euzenat extended IC)	0.43
Our measure (Eq. (9))	0.45

To test the statistical significance of the correlation values we also computed the *p-value* of the correlation, which states the probability that the observed correlation occurred by chance. In all cases, the *p-values* were less than 0.000035 (the *p-value* for the measure with the lower correlation, i.e. Gloss Overlap). A *p-value* below 0.001 (0.1% chance) is a proof of statistical significance under the strictest standards [40].

Analyzing the results, we observe that ontology-based similarity measures are, in general, less accurate than relatedness ones. The low accuracy (0.32 to 0.36) achieved by these measures was expected because they estimate similarity (i.e., only the taxonomic relationships modeled in the ontology were considered), whilst the benchmark associates word pairs with relatedness ratings.

Regarding distributional approaches to semantic relatedness, we see that the first order co-occurrence measure (NPMI), which relies on direct term co-occurrences estimated from the WSE's page counts, also offers a limited accuracy (0.34). This shows that page-counts alone are not accurate enough to estimate a reliable resemblance between terms, mainly because of the ambiguity of word occurrences in plain textual corpora such as the Web. The second order co-occurrence measure (Gloss Overlap) performed even worse than NPMI. Even though second order measures are able to estimate the resemblance between semantically related terms that do not directly co-occur, in

this case, it seems that the WordNet glosses in which the measure relies were not detailed/large enough (in comparison with the large linguistic corpus provided by the Web) to enable accurate estimations.

IC-based measures using the extended IC calculation proposed in [32] provide better accuracy than both ontology-based similarity measures and distributional relatedness ones. On the one hand, we see that the classical IC-based similarity measures by Resnik, Lin and Jiang & Conrath performed quite well for relatedness (0.40 to 0.41) when the intrinsic IC calculation in which they rely also considered the non-taxonomic relationships modeled in the ontology. On the other hand, the FaITH measure, which was specifically designed to exploit the notion of extended IC proposed in [32] was able the obtain the best accuracy so far (0.43). Certainly, the fact that these measures exploit more ontological knowledge (i.e., both taxonomically and non-taxonomically related concepts) gives them an advantage over ontology-based similarity measures.

In any case, our measure was able to provide an even higher accuracy (0.45). Compared with pure distributional measures, our proposal was able to overcome the limitations imposed by the language ambiguity inherent to the estimation of term probabilities from plain textual corpora, thanks to the use of highly reliable ontology-based similarity assessments. Compared with ontology-based relatedness measures, the improvement shown by our measure suggests that the amount of non-taxonomic knowledge currently available in most ontologies, even though semantically accurate, are not detailed enough to replace the (implicit) non-taxonomic knowledge inherent to large textual corpora (such as the Web). In this respect and, as stated above, the coverage of WordNet and other ontologies of non-taxonomic relationships is quite marginal [7, 28].

5. Conclusions

In this paper we presented a semantic relatedness calculation method that, by basing the assessment of the taxonomic resemblance on an accurate ontology-based similarity measure and the assessment of the non-taxonomic resemblance on the distribution of terms in large textual corpora, was able to improve the accuracy of related works in a standard evaluation benchmark.

Special care has been put to coherently integrate the taxonomic and non-taxonomic dimensions of the relatedness assessment, both semantically and numerically. In comparison with ontology-based relatedness measures, our method can be applied to ontologies with poor or even null coverage of non-taxonomic relationships. Compared to distributional measures, our approach is less affected by the language ambiguity inherent to the estimation of term probabilities from plain textual corpora.

Regarding this last aspect, as future work, we plan to minimize the effect of language ambiguity by contextualizing potentially ambiguous queries with additional terms (e.g., taxonomic ancestors) extracted from the underlying ontology [41]. With this, we expect to obtain more accurate probabilities that, in turn, will help to improve the accuracy of the relatedness assessment. Moreover, we also plan combining several ontologies in order to further increase the taxonomical accuracy [42-45]. Finally, we also plan to test our method with other semantic benchmarks and/or in specific domains (e.g. medicine) and applications (e.g., semantic annotation, document classification, semantic disambiguation, etc.).

Acknowledgements

This work was partly supported by the European Commission under the H2020 project “CLARUS”, by the Spanish Government through projects TIN2014-57364-C2-R “SmartGlacis”, TIN2011-27076-C03-01 “Co-Privacy” and TIN2015-70054-REDC “Red de excelencia Consolider ARES” and by the Government of Catalonia under grant 2014 SGR 537.

The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of UNESCO. M. Batet is supported by a Postdoctoral grant from Ministry of Economy and Competitiveness (MINECO) (FPDI-2013-16589).

References

1 . M. Batet, D. Sánchez, A review on semantic similarity, in: Encyclopedia of Information Science and Technology, Third Edition, IGI Global, 2014, pp. 7575-7583.
2 . Z. Wu, M. Palmer, Verb semantics and lexical selection, in: 32nd annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Las Cruces, New Mexico, 1994, pp. 133 -138.
3 . M. Batet, D. Sánchez, A. Valls, An ontology-based measure to compute semantic similarity in biomedicine, Journal of Biomedical Informatics, 44 (2011) 118-125.
4 . D. Sánchez, M. Batet, D. Isern, Ontology-based Information Content computation, Knowledge-based Systems, 24 (2011) 297-303.
5 . P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, Journal of Artificial Intelligence Research, 11 (1999) 95-130.
6 . U. Waltinger, I. Cramer, TonioWandmacher, From Social Networks To Distributional Properties: A Comparative Study On Computing Semantic Relatedness, in: N. Taatgen, H. van Rijn (Eds.) Thirty-First Annual meeting of the Cognitive Science Society, CogSci 2009, Cognitive Science Society, Amsterdam, Netherlands, 2009, pp. 3016-3021.
7 . L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V. Doshi, J. Sachs, Swoogle: A Search and Metadata Engine for the Semantic Web, in: thirteenth ACM international conference on Information and knowledge management, CIKM 2004, ACM Press, Washington, D.C., USA, 2004, pp. 652-659.
8 . R.L. Cilibrasi, P.M.B. Vitányi, The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering, 19 (2006) 370-383.
9 . D. Bollegala, Y. Matsuo, M. Ishizuka, Measuring Semantic Similarity between Words Using Web Search Engines, in: C. Williamson, M.E. Zurko (Eds.) 16th international conference on World Wide Web, WWW 2007, ACM Press, Banff, Alberta, Canada, 2007, pp. 757-766.
10 . D. Bollegala, Y. Matsuo, M. Ishizuka, A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web, in: P. Koehn, R. Mihalcea (Eds.) Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, ACL and AFNLP, Singapore, Republic of Singapore, 2009, pp. 803–812.
11 . N. Seco, T. Veale, J. Hayes, An Intrinsic Information Content Metric for Semantic Similarity in WordNet, in: R. López de Mántaras, L. Saitta (Eds.) 16th European Conference on Artificial Intelligence, ECAI 2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, IOS Press, Valencia, Spain, 2004, pp. 1089-1090.
12 . S. Patwardhan, T. Pedersen, Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts, in: EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together, Trento, Italy, 2006, pp. 1-8.
13 . M. Batet, A study on semantic similarity and ots application to clustering, VDM Verlag Dr. Müller, 2011.
14 . R. Rada, H. Mili, E. Bichnell, M. Blettner, Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man, and Cybernetics, 9 (1989) 17-30.
15 . C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, in: WordNet: An electronic lexical database, MIT Press, 1998, pp. 265-283.
16 . Y. Li, Z. Bandar, D. McLean, An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources., IEEE Transactions on Knowledge and Data Engineering, 15 (2003) 871-882.
17 . D. Sánchez, M. Batet, D. Isern, A. Valls, Ontology-based semantic similarity: A new feature-based approach, Expert Systems with Applications, 39 (2012) 7718-7728.
18 . A. Maedche, S. Staab, Measuring similarity between ontologies, in: 13th International Conference on Knowledge Engineering and Knowledge Management, 2002, pp. 251-263.
19 . P. Resnik, Using Information Content to Evalutate Semantic Similarity in a Taxonomy, in: C.S. Mellish (Ed.) 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, 1995, pp. 448-453.
20 . J.J. Jiang, D.W. Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, in: International Conference on Research in Computational Linguistics, ROCLING X, Taipei, Taiwan, 1997, pp. 19-33.
21 . D. Lin, An Information-Theoretic Definition of Similarity, in: J. Shavlik (Ed.) Fifteenth International Conference on Machine Learning, ICML 1998, Morgan Kaufmann, Madison, Wisconsin, USA, 1998, pp. 296-304.
22 . D. Sánchez, M. Batet, A New Model to Compute the Information Content of Concepts from Taxonomic Knowledge, International Journal on Semantic Web and Information Systems, 8 (2012) 34-50.
23 . D. Sánchez, M. Batet, Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective Journal of Biomedical Informatics, 44 (2011) 749-759.
24 . P.D. Turney, Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, in: L. De Raedt, P. Flach (Eds.) 12th European Conference on Machine Learning, ECML 2001, Springer-Verlag, Freiburg, Germany, 2001, pp. 491-502.
25 . S. Banerjee, T. Pedersen, Extended Gloss Overlaps as a Measure of Semantic Relatedness, in: G. Gottlob, T. Walsh (Eds.) 18th International Joint Conference on Artificial Intelligence, IJCAI 2003, Morgan Kaufmann, Acapulco, Mexico, 2003, pp. 805-810.
26 . S. Wan, R.A. Angryk, Measuring Semantic Similarity Using WordNet-based Context Vectors, in: M. El-Hawary (Ed.) IEEE International Conference on Systems, Man and Cybernetics, SMC 2007, IEEE Computer Society, Montreal, Quebec, Canada, 2007, pp. 908 - 913.
27 . B. Lemaire, G. Denhière, Effects of High-Order Co-occurrences on Word Semantic Similarities, Current Psychology Letters - Behaviour, Brain and Cognition, 18 (2006) 1.
28 . A. Devitt, C. Vogel, The topology of WordNet: Some Metrics, in: P. Sojka, K. Pala, P. Smrz, C. Fellbaum, P. Vossen (Eds.) 2nd Global Wordnet Conference, GWC 2004, Masaryk University, Brno, Czech Republic, 2004, pp. 106-111.

29 . G. Hirst, D. St-Onge, Lexical chains as representations of context for the detection and correction of malapropisms, in: C. Fellbaum (Ed.) WordNet: An Electronic Lexical Database, MIT Press, 1998, pp. 305–332.
30 . M.A. Rodríguez, M.J. Egenhofer, Determining semantic similarity among entity classes from different ontologies, IEEE Transactions on Knowledge and Data Engineering, 15 (2003) 442–456.
31 . E.G.M. Petrakis, G. Varelas, A. Hliaoutakis, P. Raftopoulou, X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies, Journal of Digital Information Management, 4 (2006) 233-237.
32 . G. Pirrò, J. Euzenat, A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness, in: International Semantic Web Conference, 2010, pp. 615-630.
33 . G. Bouma, Normalized (Pointwise) Mutual Information in Collocation Extraction, in: E.d.C.S. Chiarcos (Ed.) Biennial GSCL Conference 2009, Gunter Narr Verlag, Tübingen, Germany, 2009, pp. 31–40.
34 . L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, Placing Search in Context: The Concept Revisited, ACM Transactions on Information Systems, 20 (2002) 116-131.
35 . E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, A. Soroa, A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches, in: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, 2009, pp. 19-27.
36 . H. Rubenstein, J. Goodenough, Contextual correlates of synonymy, Communications of the ACM, 8 (1965) 627-633.
37 . G.A. Miller, W.G. Charles, Contextual correlates of semantic similarity, Language and Cognitive Processes, 6 (1991) 1-28.
38 . C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Massachusetts, 1998.
39 . G. Pirró, A semantic similarity metric combining features and intrinsic information content, Data & Knowledge Engineering, 68 (2009) 1289-1308.
40 . V.E. Johnson, Revised standards for statistical evidence, in: Proceedings of the National Academy of Sciences of the United States of America, 2013.
41 . D. Sánchez, M. Batet, A. Valls, K. Gibert, Ontology-driven web-based semantic similarity, Journal of Intelligent Information Systems, 35 (2010) 383-413.
42 . M. Batet, D. Sanchez, A. Valls, K. Gibert, Semantic similarity estimation from multiple ontologies, Applied Intelligence, 38 (2013) 29-44.
43 . D. Sánchez, A. Solé-Ribalta, M. Batet, F. Serratosa, Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain, Journal of Biomedical Informatics, 45 (2012) 141-155
44 . D. Sánchez, M. Batet, A semantic similarity method based on information content exploiting multiple ontologies, Expert Systems with Applications, 40 (2013) 1393-1399.
45 . M. Batet, S. Harispe, S. Ranwez, D. Sánchez, V. Ranwez, An information theoretic approach to improve semantic similarity assessments across multiple ontologies, Information Sciences, 283 (2014) 197-210.