A review on semantic similarity

Montserrat Batet
Universitat Rovira i Virgili, Catalonia, Spain
David Sánchez
Universitat Rovira i Virgili, Catalonia, Spain

INTRODUCTION

In recent years, due to the enormous development of the Information Society and the Word Wide Web, the automated understanding of electronic textual resources has acquired a great interest in the research community. The counter stone of textual understanding is the assessment of the *semantic similarity* between textual entities (e.g., words, sentences or documents). In general, semantic similarity aims at assessing a score that reflects the resemblance between the meanings of the compared entities, so that algorithms (e.g., classification, clustering, etc.) can seamlessly manage textual information from a numerical perspective. To do so, similarity measures exploit one or several information or knowledge sources and rely on different theoretical principles. Semantic similarity estimation has received a lot of attention in the last decade, also becoming a hot topic in many application areas such as natural language processing, information management and retrieval, textual data analysis and classification, and the Semantic Web.

Considering the plethora and heterogeneity of semantic similarity approaches available in the literature, this chapter offers researchers and practitioners aiming to develop or to exploit similarity measures: i) a description on the main notions involved in the assessment of semantic similarity, ii) a classification of the usual approaches proposed in the literature according to the theoretical principles on which they rely, iii) a comparative study of these approaches, highlighting their advantages and shortcomings under the dimensions of accuracy, applicability and dependency on external resources, and iv) a list of research challenges.

BACKGROUND

According to the kind of knowledge source exploited to extract semantic evidences on which semantic similarity estimation relies, measures can be grouped in several families.

Ontology-based measures estimate the similarity of two concepts according to the structured knowledge representation offered by ontologies. They can be classified into:

i) *Edge-based measures*, which estimate similarity according to the number of semantic links separating the two concepts in the ontology (Al-Mubaid & Nguyen, 2006; Leacock & Chodorow, 1998; Li, *et al.*, 2003; Rada, *et al.*, 1989; Wu & Palmer, 1994).

ii) *Feature-based measures*, which rely on the amount of overlapping ontological features (e.g., taxonomic ancestors, concept descriptions, etc.) between the compared concepts (Petrakis, *et al.*, 2006; Rodríguez & Egenhofer, 2003; Sánchez, *et al.*, 2012a).

iii) *Measures based on quantifying the amount of information* (i.e., Information Content (IC)) that concepts have in common (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995). In these approaches, commonalties are extracted from the taxonomical structure offered by ontologies, whereas the informativeness of concepts can be computed either extrinsically from the concept occurrences in a corpus (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995) or intrinsically, according to the number of taxonomical descendants and/or ancestors modeled in the ontology (Pirró, 2009; Sánchez & Batet, 2012; Seco, *et al.*, 2004).

Distributional approaches only use textual corpora to infer semantics. They are based on the assumption that words with similar distributional properties have similar meanings (Waltinger, *et al.*, 2009). Thus, they assess term similarities according to their co-occurrence in corpora. As words may co-occur due to different kinds of relationships (i.e., taxonomic and non-taxonomic), distributional measures capture the more general notion of semantic *relatedness* in contrast to *similarity*, which is understood strictly as taxonomical resemblance. Distributional approaches can be classified into:

i) *First order co-occurrence measures*, which assume that terms with similar meanings have a tendency to co-occur and measure relatedness directly from the explicit co-occurrence of terms (Bollegala, *et al.*, 2009; Chen, *et al.*, 2006; Cilibrasi & Vitányi, 2006; Turney, 2001).

ii) *Second order co-occurrence measures*, which estimate relatedness as a function of the co-occurrence of words appearing in the context in which the compared terms occur (Banerjee & Pedersen, 2003; Deerwester, *et al.*, 2000; Pedersen, *et al.*, 2007; Sahami & Heilman, 2006; Wan & Angryk, 2007).

DISCUSSION

In this section, the advantages and drawbacks of the different families of measures introduced in the previous section are pointed out, under the dimensions of expected accuracy, applicability and dependency on external resources.

In general, edge-counting measures are able to provide reasonably accurate results when a detailed and taxonomically homogenous ontology is used. They have a low computational cost (compared to approaches relying on textual corpora) and they are easily implementable and applicable. However, they just evaluate the shortest taxonomical path between concept pairs as evidences of distance (i.e., the opposite to similarity). This is a drawback, because many ontologies (e.g., WordNet, SNOMED-CT or MeSH) incorporate multiple taxonomical inheritance that would result in several taxonomical paths connecting concept pairs. Those paths represent explicit knowledge that is omitted by edge-counting methods (Batet, *et al.*, 2011). Compared to other ontology-based paradigms (i.e., feature-based and IC-based approaches) they usually provide the lowest accuracy (Batet, *et al.*, 2011).

Feature-based measures exploit more semantic evidences than edge-counting approaches, evaluating both commonalties and differences of compared concepts (e.g., common and different taxonomical ancestors). Since the additional knowledge helps to better differentiate concept pairs, they tend to be more accurate than edge-based measures. However, since some feature-based approaches rely on semantic features other than taxonomical, like non-taxonomic relationships (e.g., meronymy), or concept descriptions (e.g., synsets, glosses, etc.), these measures can only be applied to a subset of the available ontologies, in which this information is available. In fact, domain ontologies very occasionally model other semantic features apart from taxonomical relationships (Ding, *et al.*, 2004). Another issue that hampers the applicability of feature-based measures as general purpose solutions is the fact that many of them (Petrakis, *et al.*, 2006; Rodríguez & Egenhofer, 2003) incorporate ad-hoc weighting parameters that balance the contribution of each semantic feature. To minimize these issues, some approaches rely on a coherent integration of just taxonomic features, which avoids the need of weighting parameters while retaining a high accuracy (Sánchez, *et al.*, 2012a).

IC-based measures assume that the similarity between the meanings of a pair of concepts is represented by the informativeness (i.e., IC) of their Least Common Subsumer (LCS) found in the background ontology. Thus, the way in which the IC is computed is crucial. Classic approaches (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995) obtain the IC of a concept by computing the inverse of its probability of appearance in a corpus. Thus, infrequent terms are considered more informative than common or general ones. However, textual corpora contain terms and ontologies model concepts, and, hence, a proper disambiguation and conceptual annotation of each word found in the corpus is required in

order to accurately compute concept appearance probabilities. Moreover, corpora should be large and heterogeneous enough to provide robust probabilities and avoid data sparseness (i.e., the fact that there are not enough data to extract valid conclusions about the distribution of terms). IC-based measures also require that the probability of appearance of terms monotonically increases as one moves up in the taxonomy; so, a concept is coherently considered more informative than any of its taxonomical ancestors and less informative than its descendants. This requirement is fulfilled by computing the probability of appearance of a term as the probability of encountering the term and any of its specializations in the given corpus. Thus, the ontology must be as complete as possible (i.e., it should include most of the specializations of each concept) in order to provide reliable results. If either the ontology or the corpus changes, probability re-computations need to be recursively executed for the affected concepts. Scalability problems due to the need of manual annotation of corpora contents and corpora dependency and availability hamper the applicability and accuracy of these measures (Sánchez, *et al.*, 2009).

To overcome these limitations, some authors intrinsically derive IC values from just ontological knowledge (Pirró, 2009; Sánchez & Batet, 2011; 2012; Seco, *et al.*, 2004). Those works rely on the assumption that the taxonomic structure of ontologies is organized in a meaningful way, according to the principle of cognitive saliency (Blank, 2003): concepts are specialized when they need to be differentiated from existing ones. In comparison to corpora-based IC computation models, intrinsic IC computation models consider that abstract ontological concepts with many hyponyms are more likely to appear in a corpus because they can be implicitly referred in text by means of all their specializations. So, concepts with many specializations will provide less information than concepts at the leaves of the hierarchy, which provide maximum information as they are not further differentiated. Intrinsic IC-based approaches overcome most of the problems observed for corpus-based IC approaches (specifically, the need of corpus processing and data-sparseness). Moreover, these approaches achieve a similar, or even better accuracy than corpus-based IC calculation when applied over detailed and fine grained ontologies (e.g., WordNet (Fellbaum, 1998)). However, for small or very specialized ontologies with a limited taxonomical depth and low branching factor, the resulting IC values between concepts would be too homogenous to enable a proper differentiation.

The main drawback of all the above ontology-based similarity measures is their complete dependency on the degree of coverage and detail of the input ontology (Cimiano, 2006). Hence, ontology-based measures require a unique, rich and consistent knowledge source with a relatively homogeneous distribution of semantic links and good inter-domain coverage to work properly (Pirró, 2009). This is sometimes hard to achieve since, in many domains, knowledge is spread through different ontologies. To tackle this problem, some authors are working on exploiting multiple ontologies in similarity assessment. The use of multiple ontologies provides additional knowledge (Al-Mubaid & Nguyen, 2009) that helps to improve the similarity estimation and to solve cases in which terms are not contained in a unique ontology. Ontology-based measures relying on multiple ontologies are a challenge because different ontologies may present significant differences in their levels of detail, granularity and semantic structures, making the comparison and integration of similarities computed from different sources difficult (Al-Mubaid & Nguyen, 2009). In the multi-ontology setting, for edge-counting and feature-based similarities, the challenge is to find terminologically or semantic equivalent concepts between ontologies (Al-Mubaid & Nguyen, 2009; Batet, *et al.*, 2013; Rodríguez & Egenhofer, 2003; Sánchez, *et al.*, 2012b). On the other hand, IC-based measures depend on the ability to retrieve an appropriate LCS that subsumes the meaning of concepts belonging to different ontologies, and the ability to coherently integrate IC values computed from different ontologies (Sánchez & Batet, 2013).

First order co-occurrence measures usually rely on large raw corpora like the Web, from which term co-occurrence and, thus, similarity are estimated from the page-count provided by a Web Search Engine when querying both terms. Compared to IC-based measures relying on tagged corpora, distributional measures require neither any knowledge source nor manual annotation to support the

assessment. Thanks to the coverage offered by a corpus as large and heterogeneous as the Web, distributional measures can be applied to terms that are not typically considered in ontologies such as named entities or recently minted terms. However, their reliance on raw textual corpora is also a drawback. First, term co-occurrences estimated by page-counts omit the type of semantic relationship that motivated the co-occurrence. Words may co-occur because they are taxonomically related, but also because they are antonyms or by pure chance. So, page counts give a rough estimation of statistical dependency. Second, page counts deal with words rather than concepts. Due to the ambiguity of language and the omission of word context analysis, polysemy and synonymy may negatively affect the estimation of similarity/relatedness. Polysemic words associated to a concept cause their page counts to contain a combination of all their senses. The presence of synonyms for a given concept produces that page counts underestimate the occurrence of concepts. Finally, page counts may not necessarily be equal to word frequencies because queried words might appear several times in a Web resource. Due to these reasons, some authors have questioned the usefulness of page counts alone as a measure of relatedness(Bollegala, *et al.*, 2009). Other authors (Lemaire & Denhière, 2006) also questioned the effectiveness of relying just on first order co-occurrences because studies on large corpora gave examples of strongly associated words that never co-occur (Bollegala, *et al.*, 2009; Lund & Burgess, 1996). This situation is caused, in many cases, by the fact that both words tend to co-occur with a third one. In fact, empirical studies have shown that first order co-occurrence measures exploiting raw corpora present a limited accuracy (Sánchez, et al., 2009).

Second-order co-occurrence measures were developed to tackle the situation in which closely related words do not necessarily co-occur (Lemaire & Denhière, 2006). In such approaches, two words are considered similar or related to the extent that their contexts are similar (Harris, 1985). Some approaches build contexts of words from a corpus of textual documents or the Web (Sahami & Heilman, 2006). However, some authors have criticized their usefulness to compute relatedness because, while semantic relatedness is inherently a relation on concepts, Web-based approaches measure a relation on words (Budanitsky & Hirst, 2006). Indeed, big-enough sense-tagged corpora are needed to obtain reliable concept distributions from word senses. Moreover, commercial bias, spam and noise are problems that may affect distributional measures when using the Web as corpus. To minimize these problems, some authors exploited concept descriptions/glosses offered by structured thesaurus like WordNet. They argue that words appearing in a gloss are likely to be more relevant for the concept's meaning than text drawn from a generic corpus and, in consequence, glosses may represent a more reliable context. That approach is based on the principle that, if two terms have similar glosses (i.e., their textual descriptions overlap), they are likely to have similar meanings (Banerjee & Pedersen, 2003; Patwardhan & Pedersen, 2006; Wan & Angryk, 2007). The use of glosses instead of the Web as corpora, results in a significant improvement of accuracy. However, those measures are hampered by their computational complexity, since the creation of context vectors in such a big dimensional space is difficult. Moreover, by relying on WordNet glosses, those measures are hardly applicable to other ontologies in which glosses or textual descriptions are typically omitted (Ding, et al., 2004).

FUTURE RESEARCH DIRECTIONS

Even though a large amount of contributions on semantic similarity are available, there is still room for improvement. In general, most measures rely on the estimation of the degree of commonalty (and, optionally, differences) between the compared concepts. By improving these estimators a large amount of specific measures would be likely improved. For example, in IC-based measures, the assessment of the commonality from the IC of the LCS could consider and integrate the informativeness of all the common ancestors, and not just the most specific one. In this manner, additional semantic evidences will be exploited. Regarding the IC calculus, corpora-based approaches may overcome data sparseness by using large electronic resources such as the Web. In such case, language ambiguity of resulting page counts may be minimized by contextualizing (i.e., disambiguating) web queries (Sánchez,

et al., 2009). The same approach can be also considered in distributional approaches dealing with unstructured textual corpora. Regarding intrinsic IC calculus, new strategies can be envisioned to measure the concept's concreteness/generality from ontological knowledge. Moreover, non-taxonomic ontological features can be also considered to measure relatedness instead of similarity (Pirró & Euzenat, 2010).

As stated in the previous section, the main limitation of ontology-based measures is that they depend on the completeness of the input ontology. This may be overcome by combining the knowledge provided by several ontologies. New strategies should be developed to coherently integrate ontological features without necessarily requiring to merge all the ontologies, which is costly and prone to inconsistences (Rodríguez & Egenhofer, 2003).

Most ontology-based measures focus on similarity rather than relatedness. Even though many ontologies omit the non-taxonomic relationships that need to be considered during the relatedness assessment, there exist some knowledge sources that incorporate relationships such as meronymy and antonymy (e.g.,WordNet). New methods to integrate these heterogeneous semantic features without necessarily requiring from tuning parameters would be desirable.

Finally, semantic similarity has been successfully applied to many natural language related tasks such as word sense disambiguation (Patwardhan, *et al.*, 2003), synonym detection (Lin, 1998) or automatic language translation(Cilibrasi & Vitányi, 2006). However, they have been scarcely considered in other research areas also dealing with textual resources, such as data clustering (Batet, 2011) or privacy protection (Martínez, *et al.*, 2012). In such cases, incorporating semantic analysis to already existing algorithms would likely improve the coherency and accuracy of the results.

CONCLUSION

This chapter provided a description and discussion of the main approaches for semantic similarity/relatedness assessment.

Among the different approaches, the ontology-based approach has proven to be one of the most effective. Within this family of measures, feature-based approaches and those relying on intrinsic IC calculus usually produce the most accurate results. Indeed, they exploit more semantic evidences than approaches based on edge-counting and they are less affected by data sparseness than methods relying on corpora. On the other hand, distributional approaches, even though usually less accurate, can be applied to terms that are not usually modeled in ontologies, such as named entities.

In any case, even by considering the large number of proposals, there is still room for future research to improve the accuracy and applicability of similarity/relatedness measures.

REFERENCES

Al-Mubaid, H., & Nguyen, H. A. (2006). A cluster-based approach for semantic similarity in the biomedical domain. *Proc. of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2006* (pp. 2713–2717). New York, USA.

Al-Mubaid, H., & Nguyen, H. A. (2009). Measuring Semantic Similarity between Biomedical Concepts within Multiple Ontologies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 39*(4), 389-398.

Banerjee, S., & Pedersen, T. (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. In G. Gottlob & T. Walsh (Eds.), *Proc. of the 18th International Joint Conference on Artificial Intelligence, IJCAI 2003* (pp. 805-810). Acapulco, Mexico.

Batet, M. (2011). Ontology based semantic clustering. *AI Communications, 24*(3), 291-292.

Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics, 44*(1), 118-125.

Batet, M., Sánchez, D., Valls, A., & Gibert, K. (2013). Semantic similarity estimation from multiple ontologies. *Applied Intelligence, 38*(1), 29-44.

Blank, A. (2003). Words and Concepts in Time: Towards Diachronic Cognitive Onomasiology. In R. Eckardt, K. von Heusinger & C. Schwarze (Eds.), *Words and Concepts in Time: towards Diachronic Cognitive Onomasiology* (pp. 37-66). Berlin, Germany: Mouton de Gruyter.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2009). A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web. In P. Koehn & R. Mihalcea (Eds.), *Proc. of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2009* (pp. 803–812). Singapore, Republic of Singapore.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of semantic distance. *Computational Linguistics, 32*(1), 13-47.

Chen, H.-H., Lin, M.-S., & Wei, Y.-C. (2006). Novel association measures using web search with double checking. *Proc. of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, COLING-ACL 2006* (pp. 1009-1016). Sydney, Australia.

Cilibrasi, R. L., & Vitányi, P. M. B. (2006). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 370-383.

Cimiano, P. (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*: Springer-Verlag.

Deerwester, S., Dumais, S. T., & Harshman, R. (2000). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V., & Sachs, J. (2004). Swoogle: A Search and Metadata Engine for the Semantic Web. *Proc. of the thirteenth ACM international conference on Information and knowledge management, CIKM 2004* (pp. 652-659). Washington, D.C., USA.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts: MIT Press.

Harris, Z. (1985). Distributional structure. In J. J. Katz (Ed.), *The Philosophy of Linguistics* (pp. 26–47). New York: Oxford University Press.

Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proc. of the International Conference on Research in Computational Linguistics, ROCLING X* (pp. 19-33). Taipei, Taiwan.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification *WordNet: An electronic lexical database* (pp. 265-283): MIT Press.

Lemaire, B., & Denhière, G. (2006). Effects of High-Order Co-occurrences on Word Semantic Similarities. *Current Psychology Letters - Behaviour, Brain and Cognition, 18*(1), 1.

Li, Y., Bandar, Z., & McLean, D. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering, 15*(4), 871-882.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In J. Shavlik (Ed.), *Proc. of the Fifteenth International Conference on Machine Learning, ICML 1998* (pp. 296-304). Madison, Wisconsin, USA.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research, Methods, Instruments and Computers, 28*(2), 203-208.

Martínez, S., Sánchez, D., Valls, A., & Batet, M. (2012). Privacy protection of textual attributes through a semantic-based masking method. *Information Fusion, 13*(4), 304-314.

Patwardhan, S., Banerjee, S., & Pedersen, T. (2003). Using Measures of Semantic Relatedness for Word Sense Disambiguation. In A. F. Gelbukh (Ed.), *Proc. of the 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, CICLing 2003* (pp. 241-257). Mexico City, Mexico.

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. *Proc. of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together* (pp. 1-8). Trento, Italy.

Pedersen, T., Pakhomov, S., Patwardhan, S., & Chute, C. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics, 40*(3), 288-299.

Petrakis, E. G. M., Varelas, G., Hliaoutakis, A., & Raftopoulou, P. (2006). X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies. *Journal of Digital Information Management, 4*, 233-237.

Pirró, G. (2009). A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering, 68*(11), 1289-1308.

Pirró, G., & Euzenat, J. (2010). A feature and information theoretic framework for semantic similarity and relatedness. *Proc. of the International Semantic Web Conference* (pp. 615-630).

Rada, R., Mili, H., Bichnell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics, 9*(1), 17-30.

Resnik, P. (1995). Using Information Content to Evalutate Semantic Similarity in a Taxonomy. In C. S. Mellish (Ed.), *Proc. of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995* (pp. 448-453). Montreal, Quebec, Canada.

Rodríguez, M. A., & Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering, 15*(2), 442–456.

Sahami, M., & Heilman, T. D. (2006). A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In L. Carr, D. De Roure & A. Iyengar (Eds.), *Proc. of the 15th International World Wide Web Conference, WWW 2006* (pp. 377 - 386). Edinburgh, Scotland

Sánchez, D., Batet, M., Valls, A., & Gibert, K. (2009). Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems, 35*(3), 383-413.

Sánchez, D., & Batet, M. (2011). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective *Journal of Biomedical Informatics, 44*(5), 749-759.

Sánchez, D., & Batet, M. (2012). A New Model to Compute the Information Content of Concepts from Taxonomic Knowledge. *International Journal on Semantic Web and Information Systems, 8*(2), 34-50.

Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012a). Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications, 39*(9), 7718-7728.

Sánchez, D., Solé-Ribalta, A., Batet, M., & Serratosa, F. (2012b). Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. *Journal of Biomedical Informatics, 45*(1), 141-155

Sánchez, D., & Batet, M. (2013). A Semantic Similarity Method Based on Information Content Exploiting Multiple Ontologies. *Expert Systems with Applications, 40*(4), 1393–1399.

Seco, N., Veale, T., & Hayes, J. (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In R. López de Mántaras & L. Saitta (Eds.), *Proc. of the 16th European Conference on Artificial Intelligence, ECAI 2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004* (pp. 1089-1090). Valencia, Spain.

Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt & P. Flach (Eds.), *Proc. of the 12th European Conference on Machine Learning, ECML 2001* (pp. 491-502). Freiburg, Germany.

Waltinger, U., Cramer, I., & TonioWandmacher (2009). From Social Networks To Distributional Properties: A Comparative Study On Computing Semantic Relatedness. In N. Taatgen & H. van Rijn (Eds.), *Proc. of the Thirty-First Annual meeting of the Cognitive Science Society, CogSci 2009* (pp. 3016-3021). Amsterdam, Netherlands.

Wan, S., & Angryk, R. A. (2007). Measuring Semantic Similarity Using WordNet-based Context Vectors. In M. El-Hawary (Ed.), *Proc. of the IEEE International Conference on Systems, Man and Cybernetics, SMC 2007* (pp. 908 - 913). Montreal, Quebec, Canada.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *Proc. of the 32nd annual Meeting of the Association for Computational Linguistics* (pp. 133 -138). Las Cruces, New Mexico.

ADDITIONAL READINGS

Batet, M., Valls, A., Gibert, K., & Sánchez, D. (2010). Semantic clustering using multiple ontologies. *Proc. of the 13th International Conference on the Catalan Association for Artificial Intelligence* (pp. 207-216).

Batet, M., Gibert, K., & Valls, A. (2011). Semantic clustering based on ontologies: an application to the study of visitors in a Natural Reserve. *Proc. of the 3rd International Conference on Agents and Artificial Intelligence, ICAART 2011* (pp. 283-289). Rome, Italy.

Batet, M. (2012). *A study on semantic similarity and its application to clustering: Enabling the classification of textual data*: VDM Verlag Dr. Müller.

Batet, M., Erola, A., Sánchez, D., & Castellà-Roca, J. (2013). Utility preserving query log anonymization via semantic microaggregation. *Information Sciences, 242*, 49-63

Bichindaritz, I., & Akkineni, S. (2006). Concept mining for indexing medical literature. *Engineering Applications of Artificial Intelligence, 19*(4), 411-417.

Brown corpus. http://icame.uib.no/brown/bcm.html,

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Proc. of the Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 10-15). Pittsburgh, USA.

Caviedes, J. E., & Cimino, J. J. (2004). Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics, 37*(2), 77-85.

Domingo-Ferrer, J., Sánchez, D., & Rufian-Torrell, G. (2013). Anonymization of nominal data based on semantic marginality. *Information Sciences, 242*, 35-48.

Euzenat, J., & Shvaiko, P. (2007). *Ontology Matching*: Springer Verlag.

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological Engineering* (2nd ed.): Springer-Verlag.

Jimeno-Yepes, A., McInnes, B. T., & Aronson, A. R. (2011). Collocation analysis for UMLS knowledge-based word sense disambiguation. *BMC Bioinformatics 12*(S-3), S4.

Martínez, S., Sánchez, D., & Valls, A. (2012a). Semantic adaptive microaggregation of categorical microdata. *Computers & Security, 31*(5), 653-672

Martínez, S., Valls, A., & Sánchez, D. (2012b). Semantically-grounded construction of centroids for datasets with textual attributes. *Knowledge-Based Systems, 35*, 160-172.

Martínez, S., Sánchez, D., & Valls, A. (2013). A semantic framework to protect the privacy of electronic health records with non-numerical attributes. *Journal of Biomedical Informatics, 46*(2), 294–303.

Mesh. http://www.nlm.nih.gov/mesh/,

Pedersen, T., Pakhomov, S., Patwardhan, S., & Chute, C. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics, 40*(3), 288-299.

Sánchez, D. (2008). *Domain Ontology Learning from the Web*: VDM Verlag.

Sánchez, D., & Isern, D. (2011). Automatic extraction of acronym definitions from the Web. *Applied Intelligence, 34*(2), 311-327.

Sánchez, D., Isern, D., & Millán, M. (2011). Content Annotation for the Semantic Web: an Automatic Web-based Approach. *Knowledge and Information Systems, 27*(3), 393-418.

Sánchez, D., Moreno, A., & Vasto, L. D. (2012). Learning relation axioms from text: An automatic Web-based approach. *Expert Systems with Applications, 39*(5), 5792-5805.

Sánchez, D., Castellà-Roca, J., & Viejo, A. (2013). Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines. *Information Sciences, 218*, 17-30.

Stevenson, M., & Greenwood, M. A. (2005). A semantic approach to IE pattern induction. In K. Knight (Ed.), *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, COLING-ACL 2005* (pp. 379–386). Ann Arbor, Michigan, USA.

Vicient, C., Sánchez, D., & Moreno, A. (2013). An automatic approach for ontology-based feature extraction from heterogeneous textual resources. *Engineering Applications of Artificial Intelligence, 26*(3), 1092-1106.

WordNet. http://wordnet.princeton.edu/

KEY TERMS AND DEFINITIONS

Semantic similarity: It estimates the taxonomic resemblance of two terms, based on the evaluation of the common semantic evidences extracted from one or several knowledge sources (e.g., textual corpus, thesaurus, taxonomies/ontologies, etc.)). For example, *moose* and *reindeer* are similar because they are ruminant mammals of the family Cervidae.

Semantic dissimilarity or distance: It is the inverse to similarity. It is computed from the disjoint semantic evidences extracted from one or several knowledge sources. Dissimilarity or distance measure can be converted to similarities through a linear transformation.

Semantic relatedness: It estimates the closeness of the semantic relationship between two terms considering both taxonomic and non-taxonomic (e.g., meronymy, antonymy, functionality, cause-effect, etc.) knowledge. For example, *reindeer* is related to *antlers* or *radiotherapy* is related to *cancer*.

Ontology: It is a structured knowledge source that explicitly and consensually represents the concepts and the semantic interrelations of a domain, with the purpose of sharing and re-using knowledge. Ontologies should, at least, contain a taxonomical backbone. Non-taxonomic relationships, properties stating concept attributes, logical axioms defining restrictions or instances representing specific objects can be also included.

Information content (IC): It measures the informativeness of an entity as the inverse of its probability of occurrence. Thus, general entities are assumed to provide less information than more specialized ones, since the former are more likely to appear in a discourse.

Least Common Subsumer (LCS): It is the most specific common ancestor of two concepts found in a given ontology. Semantically, it represents the commonality of the pair of concepts. For example, the LCS of *moose* and *kangaroo* in WordNet is *mammal.*

Tagged corpora: It is a set of textual resources in which words and phrases have been annotated (i.e., disambiguated) according to their conceptualizations modeled in one or several ontologies.