Semantic Measures

Yoan Chabot
University of Burgundy, France

Christophe Nicolle
University of Burgundy, France

INTRODUCTION

Significant advances in terms of syntactic, structural and schematic heterogeneity have been achieved by adopting conventions and standards. The IT community is now trying to solve the problem of semantic heterogeneity (particularly in the Semantic Web field). To reach this objective, it is necessary to enable machines to understand the semantics of terms.

Semantics, as opposed to syntax, defines the mental representation of concepts corresponding to the symbols used in texts or images. When a person reads a text, he uses a semantization process which enables him to associate an interpretation to each sign identified. This operation uses a number of underlying processes such as measuring semantic distance between the meanings of several terms. Reasoning about the semantic proximity of terms is trivial for a human. However, this task is very complex for machines, and requires access to a large number of definitions of specific field terms.

This article aims to present the semantic measures and outlines the various techniques used to compute these measures. Three criteria are commonly used in literature to classify semantic measures: the type of measures, the source of knowledge used, and the type of approach. The first section of this article presents the three types of measures. In the next section, the different kinds of knowledge resources which can be used to compute measures will be presented. The approaches used to compute semantic measures and the works related to each of these approaches will be introduced in the third section. Several techniques to assess the accuracy of semantic measures will be given in the last section of this article.

BACKGROUND

Similarity and Semantic Relatedness

This section discusses the three different types of semantic measures. Depending on the applications and the developers' needs, proposals may be semantic relatedness measures, semantic similarity measures or semantic distance. The semantic measure, which allows to quantify the distance between the meanings of two concepts, is a generic term covering several concepts (Budanitsky & Hirst, 2006) (Gracia & Mena, 2008):

- Semantic relatedness covers all possible semantic relationships. This is a broader measure than the measure of semantic similarity. Indeed, the terms which do not share a common meaning can be considered semantically close, as they can be linked by a meronym or antonym relationship. They can also be linked by a functional relationship or frequent association relationship (e.g. "car" and "road," "lion" and "Africa").
- Semantic similarity is a special case of semantic relatedness. This distance uses only synonymy, hyponymy and hyperonymy relationships to determine whether two words share common characteristics.
- Semantic distance is often viewed as the inverse of semantic relatedness or semantic similarity. If the proximity increases, the semantic distance decreases. In most cases, the two "visions" of the term "distance" are compatible.

DOI: 10.4018/978-1-4666-5888-2.ch460

However, there are exceptions. For example, antonym concepts are semantically dissimilar but still very close, due to the antonymous relationship. Generally, it is accepted that semantic distance is the inverse of semantic relatedness.

Knowledge Resources

This section talks about the second criterion used for classifying the measures: the source of knowledge used to compute semantic measures. Among the most common sources, there are dictionaries, thesaurus, Wikipedia, DBPedia and the Web. Some proposals do not use knowledge sources, including some statistical methods. In this section, a summary of strengths and drawbacks of each source is proposed.

Dictionary and Thesauri

Dictionaries and thesauri have the same goal, that of providing information on the meaning of words. Dictionaries, however, are more focused on information about grammar, etymology and the pronunciation of words. Meanwhile, thesauri provide information about the relationships between words (synonymy, antonymy...) but also between concepts (the hierarchical relationship, relationship association...). A large majority of measures use this type of knowledge source and more specifically the thesaurus WordNet (Fellbaum & others, 2005) (other proposals use the Roget's thesaurus (Roget, 1911) or the Macquarie thesaurus (Bernard, 1984)).

The use of dictionaries or thesauri has three major drawbacks. The first limitation is the low coverage of those knowledge sources. Indeed, thesauri such as WorldNet contain few proper names ("Genghis Khan," "François 1er" etc.) and specialized terms ("potassium nitrate," "TCP-IP") (Gracia & Mena, 2008). The second disadvantage is that it is necessary to request experts to supply the knowledge base, which makes the expansion process long and tedious. Finally, dictionaries and thesauri contain more information about the terms themselves ("car" is a synonym of "automobile") than about knowledge in general (e.g. a relation between the word "lion" and the word "Africa").

Wikipedia

The advent of Web 2.0 has enabled online communities to work together to create lexical resources like Wikipedia or Wiktionary. The collaborative nature of Wikipedia enables it to grow quickly and have high reactivity on world events. This last point enables this encyclopedia to have updated content and recent topics (Wikipedia: About, 2002).

Wikipedia is now considered to be one of the most significant multilingual knowledge bases (Gabrilovich & Markovitch, 2007). In addition, it provides a more structured knowledge base than search engines, and with a wider coverage than WordNet (Strube & Ponzetto, 2006). The use of this support keeps the advantages of the techniques based on the thesaurus, while providing better coverage. In addition, Wikipedia provides information on proper names or specialized terms. However, Wikipedia is not similar to the entire web with regard to the discovery and evaluation of semantic relations implied (Gracia & Mena, 2008). For example, the term "stomach ache" and "aspirin" are not mentioned together in a Wikipedia article. However, thousands of pages containing both terms together exist on the Internet.

DBPedia

DBPedia is a project with the objective to extract information from Wikipedia and to provide a structured format (using RDF which is a Semantic Web technology) on the Web. The data structure combined with a very large amount of data (over a billion RDF triples so far) enables DBPedia to be a source of knowledge with a strong potential for many applications, including semantic measurements (Bizer, et al., 2009).

Web

The property of maximum coverage (presented later in this state of the art) has encouraged the authors to work with the Web as a source of knowledge. The Web is a potentially endless source of information. Nevertheless, it is important to note that the proportion of domain experts is small compared to the total number of Internet users. Therefore, the Web cannot be considered as a corpus of high quality (Gracia & Mena, 2008). However, with the large data volume, it is easy to remove noise and to identify relevant knowledge using statistical methods. Using the Web as a corpus has become an increasingly widespread idea.

Due to the difficulty in handling the Web, some of the measures using it as knowledge sources use search engines to extract relevant information (Gracia & Mena, 2008). This has led to the creation of semantic measures such as Normalized Google Distance (Cilibrasi & Vitanyi, 2007) which uses information on the number of results returned by a query on the Google search engine to evaluate the semantic distance between two terms.

Summary

Table 1 provides an overview of the characteristics of each source of knowledge.

The line "Size" reflects the amount of information contained in each source of knowledge. The "Coverage" is based on the presence of specialized terms and proper names in the database. The line "Structure" shows the capacity of each source to provide structured information. Finally, the line "Growth" is used to define the growth rate of each source. This level is directly dependent on the ease of updating the database, and whether or not the assistance of specific agents (experts) is needed.

This comparison study between knowledge sources enables us to draw some conclusions. Firstly, DBPedia seems to be a good alternative to the dictionary and thesauri. Indeed, the synergy between knowledge of Wikipedia and the structuration provided by Semantic Web technologies enables DBpedia to have a large amount of well-structured knowledge. On the other hand, the Web is also a promising source of knowledge. However, processes to reduce noise and extract unstructured knowledge are required.

Table 1. Overview of the knowledge sources

	Dictionary and Thesauri	**Wikipedia**	**DBPedia**	**Web**
Size	+	++	++	+++
Coverage	+	++	++	+++
Structure	+++	++	+++	+
Growth	+	++	++	+++

APPROACHES AND METHODS OF SEMANTIC MEASURES

The last criterion to classify semantic measures is the approach used. In this section, we present those different approaches and the associated proposals of semantic measures identified in the literature.

Graph-Based Approach

This kind of approach uses a semantic network and sees it as a graph. Thus, the concepts of the network are represented by nodes, and relationships between concepts are represented by edges. The vast majority of graph approaches consider only the subsumption hierarchy of the network. They are therefore often similarity measures. Measurements are made using the path length between the entities to compare. The shorter the path between them, the more similar the concepts are (Resnik, 1995).

(Rada, Mili, Bicknell, & Blettner, 1989) introduced a metric called "Distance" which measures the length of the shortest path between two terms (represented by nodes) in a graph. This approach, using the medical metadata system MeSH (Medical Subject Headings) and the principles outlined by Resnik, is principally designed to compute the conceptual distance between two concepts of a concept hierarchy. Thanks to this simplification, Rada was able to obtain good results. In the context of semantic networks, the path length between two nodes of the graph is not sufficient to correctly measure the semantic distance because the properties composing the graph are not necessarily equally significant. However, when properties are reduced to subsumption links, the path length can measure the semantic distance quite effectively.

Other proposals have attempted to propose variations of the basic idea "the shorter a path is, the more similar the related concepts are." In his proposal (Sussna, 1993) introduced the idea of measuring the semantic relatedness, considering other relationships in addition to subsumption. Sussna proposes to associate weights to each arc composing the path between two nodes involved in the distance computation process using two techniques:

- The TSF (type-specific fan-out) factor: the numbers of arcs of the same type leaving a given node affects the weight of those arcs. An increasing number of arcs of the same type lead to a decrease of the weights of those arcs.
- A depth-relative scaling process: several proposals use the length of the paths in the hierarchy to measure the semantic relatedness of two elements. However, in the case of WordNet in particular, the path length depends on the hierarchical level of the terms studied. Thus, a length equal to one between two concepts located at the top of the hierarchy is a greater distance than the same length between two concepts located at the bottom of the hierarchy.

In conclusion, several proposals have tried to improve the basic idea of "the shorter the path between them is, the more similar the concepts are," of the graph-based approach. Among the proposed improvements, the depth-relative scaling introduced by (Sussna, 1993) enables to take into account the depth of the studied nodes.

Information Content

This approach, introduced by Resnik in (Resnik, 1993) (Resnik, 1995) (Resnik, 2011) uses the subsumption relations between concepts (using a semantic network), and the information contained in a corpus. In his approach, the author uses two assumptions:

- The more general the least common parent of two concepts of a hierarchy is, the lower the similarity between them is. The author states that the similarity between two concepts is based on the amount of information they have in common. In a subsumption hierarchy, the amount of shared information may be given by the position of the last common parent.
- The more specific (inverse of general) a concept is, the lesser its probability of occurrence is. To evaluate the specificity of a concept, the informational content formula introduced by Shannon in information theory (Shannon, 2001) is used. The information content is used to evaluate the specificity of a concept by measuring the probability of encountering an instance of this concept (the lower a concept is in the hierarchy, the less likely it is to be instantiated) and is given by the formula -log *p* (concept).

This measure of specificity is then used on the least common subsumer of the studied concepts:

$$sim(X,Y) = -\log p\Big(lcs(X,Y)\Big)$$

where *lcs* is the least common subsumer of *X* and *Y*.

The lower use of network links to determine the similarity enables Resnik to partially solve the problem of distance variation between the links (scaling problem). The network is only used here to identify the common parent of a pair of concepts and numerical values derived entirely from the corpus used. However, Resnik's approach also has disadvantages. The proposed method makes all pairs of concepts sharing a common parent have an equal similarity. Thus, certain pairs of concepts with different values in graph-based approaches (because the number of links between them is different) have equal distances in Resnik's approach (Budanitsky & Hirst, 2006).

Distributional Approach

The proposals of this group are based on the distributional hypothesis (Rubenstein & Goodenough, 1965) which says that semantically closed terms tend to appear in similar contexts.

In his proposal (Hindle, 1990), introduced a measure of similarity between two terms using this assumption: "In each language, there are restrictions on what words can appear together in the same construction" (e.g. "car" with "drive," "build" or "sell"). Hindle argues that each word can therefore be represented by the context where it occurs. If two words share a similar context, it means that they are similar. The measure proposed by Hindle gives good results which support the distributional hypothesis. However, these results are obtained under certain conditions highlighted by the author such as the use of a large text.

Information Overlap

This approach is based on the assumption that semantically similar words share common content. The measurement of relatedness is based on the overlap (similarities) of the descriptions of the two terms studied (Lesk, 1986). The more important the overlap is, the more similar the terms are (because intuitively, the same words are used to express the same ideas).

In his work, Lesk introduces a relatedness measure used in an application to find the correct meaning of a term in a given context. To disambiguate terms, a comparison between short glosses of the ambiguous term and contextual terms is performed. The meaning of the ambiguous word which matches the best with contextual glosses is chosen at the end of the process. One of the main limitations of the Lesk proposition is that the glosses used are extracted from WordNet, and are therefore too short to obtain accurate results.

In (Banerjee & Pedersen, 2003) (Patwardhan, Banerjee, & Pedersen, 2003), the authors propose to enhance the Lesk measure by increasing the size of the glosses. To do this, the authors introduced the extended gloss overlap measure which extends glosses with the definitions of terms related to the studied terms. The related terms are identified by using Wordnet relations but also implicit relations identified with the overlap measure (e.g. vehicle and car are explicitly connected by a hypernym relation while car and tyre are implicitly connected).

An improvement of the scoring mechanism is also proposed for the extended gloss overlap measure. To compute the Lesk measure, the number of words shared by two terms is taken into account. Banerjee and Pedersen argue that "this mechanism does not differentiate between single word and phrasal overlaps." To make this distinction, the authors suggest giving a more significant score to words which share longer sequence of consecutive words.

The final formula of the extended gloss overlap measure is given in Box 1 where *rel*(*X*,*Y*) is a function which computes the relatedness between two terms, overlap_score is a function that returns the overlapping score of two glosses and $r_1(X)$ and $r_2(Y)$ are functions which return the gloss of terms respectively related to X and Y by the relations r_1 and r_2 which are from the sets of WordNet relations take into account by this approach.

Statistical Approach

The approach based on statistics differs from other approaches, by providing a more mathematical vision of the problem of semantic measures. Proposals using this approach rely on vector or probabilistic models to represent words and concepts, and then use formulas like the cosine similarity measure (Singhal, 2001), Euclidean distance or the Jaccard similarity measure (Jaccard, 1901) to compute the distance between these elements.

In (Gabrilovich & Markovitch, 2007), the ESA (Explicit Semantic Analysis) method to compute the semantic relatedness between two terms is proposed. The main idea of this proposition is to compute the relatedness between the semantics of the Wikipedia articles related to the terms studied. To do this, a large space of concepts derived from Wikipedia is used to represent the meanings of texts using vector representations. Thus, the ESA method enables an explicit semantic analysis, because the concepts used are obvious for humans (Wikipedia concepts). The ESA method is therefore opposed to the LSA method (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), which uses latent concepts.

To achieve their goals, the authors use two components:

Box 1.

$$\boxed{rel(X,Y) = \sum overlap\_score\left(r_1(X), r_2(Y)\right)\ \forall(r_1, r_2)\ \in RELATIONS}$$

- A semantic interpreter which converts Wikipedia articles into weighted vector of Wikipedia concepts. The concepts are ordered according to their relevance with respect to the given text. This phase of classification is performed by a classifier based on the centroid vectors and gives results called «interpretation vectors"
- A vector comparison module which computes the semantic relatedness between two texts with conventional methods such as the cosine similarity measure.

In addition to the use of concepts more understandable by humans, empirical evaluation of the ESA approach confirms that the results of this latter are better than the LSA method.

Conclusion

A large majority of works proposed in this article are directly based on assumptions consistent with human intuition, such as the information overlap hypothesis or distributional hypothesis. This allows their authors to obtain good results because "if the assumptions are deemed reasonable, the similarity measure necessarily follows" (Lin, 1998). However, to say what the best measure is for a given application is still a difficult task that requires evaluation techniques.

Although the empirical evaluation of the measures is beyond the scope of this article, we will introduce some criterions which enable to assess the relevance of a semantic measure in the next section.

QUALIFICATION OF MEASURES

This section gives some elements to allow the reader to know how to qualify a semantic measure. By "qualify a semantic measure," we mean checking if the measure has expected qualities and provides accurate results. For this last point, we present three ways to objectively evaluate a given measure.

Qualitative Properties

To enable comparison of the different proposals and determine which method of measurement is more relevant and effective in meeting a need, it is important to give criteria of comparison. In (Gracia & Mena, 2008), Jorge Garcia gives several qualities expected for a semantic measure:

- **Maximum coverage:** A measure should not be limited to a particular source of knowledge such as WordNet, for example, or an ontology in particular. If the maximum coverage is not observed, the scope of application of the measure produced is limited. The more sources of knowledge that are taken into account, the greater the coverage is.
- **Universality:** Semantic measures must be flexible and generic enough to be used without specific lexical resources and without specific language of knowledge representation. The universality is an indicator of the ability of the measure to adjust to any knowledge sources.
- **Independence from the field:** The number of knowledge bases and by extension, the number of represented fields and subjects is significantly increasing. It is therefore necessary to propose semantic measures which can deal with this heterogeneity. For this, no assumptions should be made about the area and the nature of the knowledge base that will be used.

In Lin (1998), Dekang Lin highlights the fact that the vast majority of proposed semantic measures are related to one application or specific model. This is particularly the case for the cosine similarity measure which requires working with a vector model. A second problem highlighted by this author is that the similarity measures are based on assumptions that have not been explicitly defined. Without knowing the underlying assumptions of each of the proposals, it is difficult to compare these measures on the theoretical level. Thus, all comparative studies adopt an empirical standpoint by using such benchmarks. From this, we can introduce a fourth and final desirable property for a semantic measure:

- **Theoretical foundations:** To enable a more rigorous comparison of the measures, it is essential to provide a theoretical basis consisting of a set of assumptions for any semantic measure.

Although these properties seem relevant, it is difficult to evaluate the ability of a given measure to fulfil these criterion and the latter are subjective. In addition, it is therefore important to study a measure in a more objective way with one of the evaluation methods given below.

Evaluation Methods

A number of methods to evaluate measures objectively have been proposed. There are three main methods of assessment in the literature, (Budanitsky & Hirst, 2006):

- **Theoretical Examination:** This type of evaluation is to verify that the measures fulfil certain mathematical properties (Lin, 1998). It may be, for example, to verify that the measure is metric, or observe singularities in it. This type of evaluation method can be used to perform an initial filtering of measurement methods.
- **Comparison with human judgment:** Human judgment is the result that we want to reach. It is therefore an effective way to determine the relevance of a measure by comparing the results given by it with results given by humans for the same pairs of terms. However, a major drawback is that it is difficult to obtain a reliable and consistent set of human judgments.
- **Evaluation by the application:** The last type of method is the evaluation of measurements in applications. It can be, for example, to compare the effectiveness of various measures to solve the problem of natural language processing, or to assess the adequacy of a measure within a recommender system.

FUTURE RESEARCH DIRECTIONS

The need to measure semantic distance between concepts is a ubiquitous problem in many areas. There are therefore many applications for semantic measures and needs will probably increase in future years. Among the main consumers of semantic measures are fields of Natural Language Processing (disambiguation of words, detection of spelling mistakes, etc.), Information Retrieval (improvement of query accuracy, query extension, etc.) and Knowledge Engineering (ontology matching, etc.).

To date, few methods have all the qualities defined by (Gracia & Mena, 2008) and (Lin, 1998). This is mainly due to the fact that the proposed measures are often ad-hoc proposals, whose properties of universality or independence from the field are rarely sought.

CONCLUSION

In this article, a classification of existing semantic measures has been proposed. We saw that a semantic measure was characterized by several criteria such as the measurement type (semantic relatedness, semantic similarity...), the type of knowledge resources used, and finally the approach used. This last criterion was used here to classify measures. The strengths and weaknesses of various proposals were highlighted for each of these approaches.

REFERENCES

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *International Joint Conference on Artificial Intelligence, 18*, 805--810.

Bernard, J. B. (1984). *The Macquarie Thesaurus: The Book of Words*. Macquarie Library.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia-A crystallization point for the Web of Data. *Web Semantics: Science. Services and Agents on the World Wide Web*, *7*(3), 154–165. doi:10.1016/j.websem.2009.07.002

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13–47. doi:10.1162/coli.2006.32.1.13

Cilibrasi, R., & Vitanyi, P. (2007). The google similarity distance. *Knowledge and Data Engineering. IEEE Transactions on*, *19*(3), 370–383.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science American Society for Information Science*, *41*(6), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Fellbaum, C. (2005). WordNet and wordnets. Encyclopedia of language and linguistics, 2, 665-670.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *Proceedings of the 20th international joint conference on artificial intelligence, 6*, 12.

Gracia, J., & Mena, E. (2008). *Web-based measure of semantic relatedness*. Springer. doi:10.4018/978-1-60566-066-0

Hindle, D. (1990). Noun classification from predicate-argument structures. *Proceedings of the 28th annual meeting on Association for Computational Linguistics* (pp. 268-275).

Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database, 305*, 305-332.

Jaccard, P. (1901). *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation* (pp. 24-26).

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on Machine Learning* (pp. 296-304).

Patwardhan, S., Banerjee, S., & Pedersen, T. (2003). *Using measures of semantic relatedness for word sense disambiguation*. Springer. doi:10.1007/3-540-36456-0_24

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(1), 17–30. doi:10.1109/21.24528

Resnik, P. (1993). Selection and information: a class-based approach to lexical relationships. *IRCS Technical Reports Series*, 200.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 448-453).*

Resnik, P. (2011). 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*, 9–130.

Roget, P. (1911). *Roget's Thesaurus of English Words and Phrases.* TY Crowell Company.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633. doi:10.1145/365628.365657

Shannon, C. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3–55. doi:10.1145/584091.584093

Singhal, A. (2001). Modern information retrieval: A brief overview. *A Quarterly Bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, *24*(4), 35–43.

Strube, M., & Ponzetto, S. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. *Proceedings of the National Conference on Artificial Intelligence, 21*, 1419.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. *In Proceedings of the second international conference on Information and knowledge management* (pp. 67-74).

Wikipedia: About. (2002). In *Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/Wikipedia:About

ADDITIONAL READING

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, 2.*

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Corby, O., Dieng-Kuntz, R., & Faron-Zucker, C. (2004). Querying the semantic web with corese search engine. *ECAI*, *16*, 705.

Euzenat, J., & Shvaiko, P. (2007). *Ontology matching (18)*. Springer Heidelberg.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131. doi:10.1145/503104.503110

Gandon, F., Corby, O., Dieng-Kuntz, R., & Giboin, A. (2005). Proximité conceptuelle et distances de graphes. Proc. Raisonner le Web Sémantique avec des Graphes, Nice, Journée thématique de la plate-forme AFIA, Nice.

Jarmasz, M., & Szpakowicz, S. (2003). S.: Roget's thesaurus and semantic similarity. *In: Proceedings of the RANLP-2003.*

Kozima, H., & Furugori, T. (1993). Similarity between words computed by spreading activation on an English dictionary. *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics* (pp. 232--239).

Laukkanen, M., & Helin, H. (2005). Competence management within and between organizations. *Proc. of 2nd Interop-EMOI Workshop on Enterprise A Similarity in an Ontology, 11.*

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database, 49*(2), 265-283.

Lord, P., Stevens, R., Brass, A., & Goble, C. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics (Oxford, England)*, *19*(10), 1275–1283. doi:10.1093/bioinformatics/btg153 PMID:12835272

Miller, G., & Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28. doi:10.1080/01690969108406936

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, *17*(1), 21–48.

Thieu, M., Steichen, O., Zapletal, E., Jaulent, M., & Bozec, C. (2004). Mesures de similarité pour l'aide au consensus en anatomie pathologique. *Ingénierie des Connaissances (IC).*

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 133-138).

KEY TERMS AND DEFINITIONS

Information Retrieval: Field of information technology whose aim is to provide techniques to process queries for extracting information from corpus.

Knowledge Engineering: Field of information technology whose aim is to provide techniques to store and manipulate knowledge.

Natural Language Processing: Field of information technology to provide methods and algorithms for processing human language. Automatic translation tools, spelling checkers, or even speech recognition software are among the most popular applications.

Ontology: Model of knowledge representation used especially in the areas of Semantic Web and artificial intelligence. Ontologies are used to represent domain knowledge using concepts, relations and axioms.

Semantic Relatedness: This is a broader measure than the measure of semantic similarity. Indeed, the terms which do not share a common meaning can be considered semantically close, as they can be linked by a meronym or antonym relationship.

Semantic Similarity: A semantic measure which is a special case of semantic relatedness. This distance uses only synonymy, hyponymy and hyperonymy relationships, and determines whether two words share common characteristics.

Semantic: As opposed to syntax, semantic defines the mental representation of concepts corresponding to the symbols used in texts or images.