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