Chapter 21
Evaluating Semantic Metrics on Tasks of Concept Similarity

Hansen A. Schwartz
University of Pennsylvania, USA

Fernando Gomez
University of Central Florida, USA

ABSTRACT

In this study, first, concept similarity measures are evaluated over human judgments by using existing sets of word similarity pairs that we annotated with word senses. Next, an application-oriented study is presented to evaluate semantic metrics based on integration into an algorithm, first focused on the task of concept similarity then on the task of concept relatedness. The results found no single measure to be most significantly correlated with human judgments, while an information content-based measure clearly lead to the best results in the application-oriented task of concept similarity. Reinforcing the difference between tasks of concept similarity and concept relatedness, the best measure for an application-oriented task of concept relatedness was a gloss-based relatedness measure rather than a similarity measure. A major conclusion of this work is that similarity measures may perform differently if embedded in specific applications than if they are compared with human judgments.

INTRODUCTION

This chapter presents an evaluation of WordNet-based semantic similarity and relatedness measures in tasks focused on concept similarity. Concept similarity is studied and is used in many disciplines. This evaluation focuses on the application to Natural Language Processing itself. Assuming similarity as distinct from relatedness, this work fills a gap within the current body of work in the evaluation of similarity and relatedness measures. Past studies have either focused entirely on relatedness or only evaluated judgments over words rather than concepts.

Semantic similarity and relatedness has a substantial history in computational linguistics
and natural language processing signifying its importance to the fields. However, an extensive evaluation of similarity and relatedness measures for the task of concept similarity has yet to be carried out. Such an evaluation could benefit applications such as word sense disambiguation or query expansion for information retrieval. This study seeks to address this gap in the current body of work by providing results on the performance of various WordNet-based measures for tasks utilizing similarity judgments among concepts (word senses).

Two distinctions are important within this chapter: that between words and concepts, and that between relatedness and similarity. Although many measures are designed for comparison of concepts (word senses), past comparisons of similarity and relatedness measures with human judgments have looked only into similarity between words themselves. For example, while one would likely agree that ‘bat’ as in ‘a club used for hitting a ball’ is similar to ‘stick’, one would be hard-pressed to agree that ‘bat’ as in ‘nocturnal mouselike mammal with forelimbs modified to form membranous wings’ is also similar to ‘stick’ (definitions from WordNet (Miller et al. 1993)). On the other hand, while application-oriented studies have applied measures to concepts, we have yet to see an evaluation utilizing an application calling for similarity judgments. This paper views similarity as a specific type of relatedness characterized by the relationships: synonymy, antonymy, and hyponymy. As an example, we would say a ‘wooden stick’ is similar and related to a ‘baseball bat’, while a ‘baseball player’ is related but not similar to a ‘baseball bat’. Although this similarity distinction has been noted previously (Resnik 1999; Patwardhan, Banerjee, and Pedersen 2003; Agirre et al. 2009), we believe this paper presents the first evaluation of measures for tasks of concept similarity.

After a review of similarity and relatedness measures, we present a summary of past evaluations. Our approach to evaluate measures is broken into two types of experiments. One type of experiment is based on existing human judgments of similarity which we annotated with senses. As a secondary contribution of this paper, we have made the sense annotated datasets available. The other experiment is application-oriented, integrating measures within a word sense disambiguation (WSD) algorithm that requires in one case similarity judgments and in another case relatedness judgments among concepts. Finally, the results are presented to demonstrate the effectiveness of each measure for tasks of concept similarity, and the difference in results when the task is concept relatedness.

BACKGROUND

Semantic Similarity and Relatedness

Semantic similarity and relatedness metrics measure the strength of relationships between concepts. This chapter studies metrics which take input as concepts defined in WordNet (Miller et al. 1993). It is important that we distinguish the measures we evaluate in this chapter from those that take words as input. Word similarity and relatedness measures, such as those that use co-occurrence information or latent semantic analysis, differ from concept similarity and relatedness measures in the types of tasks they can be used for. Additionally, the notion of similarity in this work is applied over two single concepts; Other works have applied similarity over different terms, such as comparing two pairs of words when measuring analogy (Turney 2006).

Measures of similarity and relatedness are often broken down in to three types: path based, information content, and gloss based (Pedersen et al., 2004). Path based approaches rely entirely on graphs of relationships, using the idea that concepts closer to each other, according to the length of path between the two, are more similar. The other types of measures may take advantage of graphs, but