Joint Distance and Information Content Word Similarity Measure

Issa Atoum and Chih How Bong

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak,
94300 Kota Samarahan,Sarawak,Malaysia
Issa.Atoum@gmail.com , chbong@fit.unimas.my

Abstract. Measuring semantic similarity between words is very important to many applications related to information retrieval and natural language processing. In the paper, we have discovered that word similarity metrics suffer from the drawback of obtaining equal similarities of two words, if they have the same path and depth values in WordNet. Likewise information content methods which depend on word probability of a corpus tend to posture the same drawback. This paper proposes a new hybrid semantic similarity to overcome the drawbacks by exploiting advantages of Li and Lin methods. On a benchmark set of human judgments on Miller Charles and Rubenstein Goodenough data sets, the proposed approach outperforms existing methods in distance and information content based methods.

Keywords: semantic similarity; similarity measures; edge counting; information content; word similarity; WordNet

1 Introduction

Semantic similarity encompasses the semantic likeness between compared words. For example, *fork* and *food* are related but they are not similar, whereas *food* and *salad* are more similar in semantics. Resnik illustrated that word similarity is a subcase of word relatedness [1]. However, many existing word similarity measures do not clearly distinguish between similarity and relatedness but instead they use a score on a scale between 0 and 1 to indicate the degree of semantic relatedness. Semantic similarity measures have been seen widely used in different applications such as: word sense disambiguation, information retrieval, question answering, summarization, machine translation and automatic essay grading[2].

Although many semantic similarity measures exist in past literatures, most of them are either distance or Information Content (IC) based. The distance based (also called edge counting method) usually uses a thesaurus (such as WordNet) to find words' pair shortest path or Least Common Subsumer (LCS) depth length (shortly referring to it depth) to derive a semantic score. Although it seems the methods are reported to work well, one intuitive problem is that words having the same path and same depth will yield an identical score, even though they postulate semantic differences. On the other hand, the fine-grained IC methods not only relying on the structure of thesaurus, but it also depend on the probability of words used in a dictionary. The significant fact is that both methods depend on the LCS, which is the shared ancestor of the two concept words to determine semantic similarity. For example, the LCS of *school bus* and *train* is *public transport*, which is depicted in Fig 1.

Fig. 1 illustrates a fragment of WordNet hierarchy. The numbers indicate the IC values of a node extracted from Brown Dictionary. Noted here that using the distance based methods, the word pairs *wheeled vehicle–rocket* and *bus–train* yield identical score as they have the same shortest path to each other and the same depth in the tree, which is (2,8)[1] respectively. Similarly, the words *car– public transport* will have the same similarity as *rocket– train* where intuitively the later should yield lower similarity score.

By obtaining the IC, we will get that the words pairs *bus* and *train* are more similar than the word pairs *rocket* and *wheeled vehicle*. Likewise, using a distance based method we will get a similarity for *car* and *scooter* rather than getting zero using an information content approach due to unavailability of information for *scooter* in Brown Dictionary.

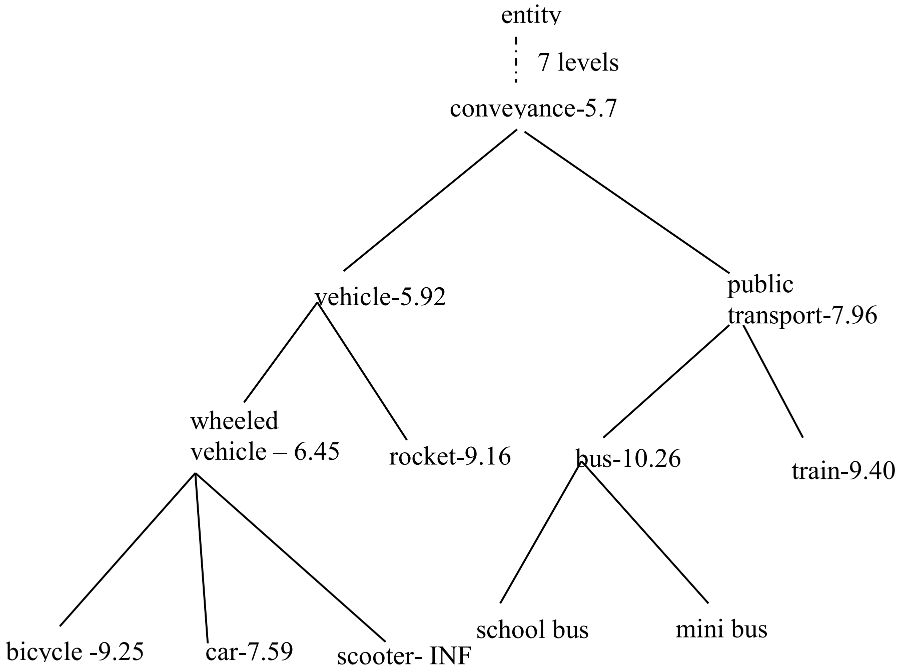

Fig.1. Fragment of WordNet showing nodes in an "*is-a"* relationship and the information content of each node. The figure shows that word pairs such as *train - wheeled vehicle* and *bus-rocket* can have same similarity value using a distance based method because they have the same path and depth values (4,7).

[1] (path, depth)