A Comprehensive Comparative Study of Word and Sentence Similarity Measures

Issa Atoum
Faculty of Information Technology
The World Islamic Sciences & Education University
11947 Amman, Jordan

Ahmed Otoom
Royal Jordanian Air forces
11134 Amman, Jordan

Narayanan Kulathuramaiyer
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan, Sarawak, Malaysia

ABSTRACT
Sentence similarity is considered the basis of many natural language tasks such as information retrieval, question answering and text summarization. The semantic meaning between compared text fragments is based on the words' semantic features and their relationships. This article reviews a set of word and sentence similarity measures and compares them on benchmark datasets. On the studied datasets, results showed that hybrid semantic measures perform better than both knowledge and corpus based measures.

General Terms
Semantic Similarity, Natural Language Processing, Computational Linguistics, Text Similarity

Keywords
Word Similarity, Sentence Similarity, Corpus Measures, Knowledge Measures, Hybrid Measures, Text Similarity

1. INTRODUCTION
Semantic similarity finds a resemblance between the related textual terms. Words are considered semantically similar or related if they have common relationships. For example, *food* and *salad* are semantically similar; obviously *salad* is a type of *food*. Also, *fork* and *food* are related; undoubtedly a *fork* is used to take *food*. Resnik illustrated that word similarity is a subcase of word relatedness[1].

The word similarity is the foundation of the sentence similarity measures. A Sentence similarity method measures the semantics of group of terms in the text fragments. It has an important role in many applications such as machine translation [2], information retrieval [3]–[5], word sense disambiguation [6], spell checking [7], thesauri generation [8], synonymy detection [9], and question answering [10]. Furthermore, semantic similarity is also used in other domains; in medical domain to extract protein functions from biomedical literature [11] and in software quality[12]–[14] to find common software attributes.

Generally, sentence similarity methods can be classified as corpus based, knowledge based and hybrid methods. Corpus based methods depend on building word frequencies from specific corpus. In this category, Latent Semantic Analysis (LSA) [10], [15], [16], and Latent Dirichlet Allocation (LDA) [3], [17], [18] have shown positive outcomes, however they are rather domain dependent [19], [20]. In other words, if the model (i.e. corpus model) was built for news text, it usually performs poorly in another domain such computer science text.

The knowledge based methods usually employ dictionary information such as path and/or depth lengths between compared words to signify relatedness. These methods suffer from the limited number of general dictionary words that might not suit specific domains. Most knowledge based measures depend on WordNet[21]. WordNet is a hand crafted lexical knowledge of English that contains more than 155,000 words organized into a taxonomic ontology of related terms known as synsets. Each synset (i.e. a concept) is linked to different synsets via a defined relationship between concepts. The most common relationships in WordNet are the 'is-a' and 'part –of' relationships.

Hybrid methods combine the corpus based methods with knowledge based methods and they generally perform better.

To the best of authors knowledge, there are a few works that compares sentences [22] [10]. This article compares state of the art word and sentence measures on benchmark datasets. It is found that hybrid measures are generally better than knowledge and corpus based measures.

2. RELATED WORK
2.1 Word Similarity Methods
2.1.1 Corpus based Methods
These methods depend on word features extracted from a corpus. The first category of these methods is based on the information content (IC) of the least common subsumer (LCS) of compared term synsets [23]–[25]. The second category, a group known as distributional methods, depends on distribution of words within a text context. Words co-occurrences are represented as vectors of grammatical dependencies. The distributional method, LSA similarity [16], [26] transforms text to low dimensional matrix and it finds the most common words that can appear together in the processed text. Corpus based methods are domain dependent because they are limited to their base corpora.

2.1.2 Knowledge based Methods
Knowledge based methods use information from dictionaries (such as WordNet) to get similarity scores. Classical knowledge based methods use the shortest path measure [27] , while others extend the path measure with depth of the LCS of compared words [28], [29] . Leacock Chodorow [30] proposed a similarity measure based on number of nodes in a taxonomy and shortest path between compared terms. Hirst and St-Onge [31] considered all types of WordNet relations; the path length and its change in direction. Some methods [23]–[25] have the ability to use intrinsic information rather than information content. Knowledge based methods suffer