Chapter 41
Semantic Similarity Measurement Using Knowledge-Augmented Multiple-prototype Distributed Word Vector

Wei Lu
Beijing Jiaotong University, China

Yuanyuan Cai
Beijing Jiaotong University, China

Kailun Shi
Beijing Jiaotong University, China

Xiaoping Che
Beijing Jiaotong University, China

ABSTRACT

Recent years, textual semantic similarity measurements play an important role in Natural Language Processing. The semantic similarity between concepts or terms can be measured by various resources like corpora, ontologies, taxonomies, etc. With the development of deep learning, distributed vector models are constructed for extracting the latent semantic information from corpora. Most of existing models create a single prototype vector to represent the meaning of a word such as CBOW. However, due to lexical ambiguity, encoding word meaning with a single vector is problematic. In this work, the authors propose a knowledge-augmented multiple-prototype model by using corpora and ontologies. Based on the distributed word vector learned by the CBOW model, the authors append the concept definition and the relational knowledge vector into the target word vector to enrich the semantic information of the word. Finally, the authors perform the experiments on well-known datasets to verify the efficiency of the authors' approach.

DOI: 10.4018/978-1-7998-0414-7.ch041

1. INTRODUCTION

Semantic similarity or relatedness, a basis for textual analysis in natural language processing (NLP) and information retrieval (IR), is commonly used to quantify the degree of the likeness of semantic content and lexical meaning. Textual semantic similarity measurements have been widely applied to a variety of applications, such as web service discovery (Paliwal et al., 2012), word sense disambiguation (Resnik, 1999), text clustering (Song et al., 2009), question answering (Ramprasath and Hariharan, 2012), as well as detection and correction of malapropisms (Hirst and StOnge, 1998). In addition to linguistics, the semantic similarity computing also emerges in other research fields, such as Biomedicine (Pedersen et al., 2007) and Geoinformatics (Schwering and Raubal, 2005). Some studies use the notion of semantic similarity and semantic relatedness interchangeably. Actually, the connotation of semantic relatedness is more general than semantic similarity, which represents a special case of relatedness. Two semantically dissimilar concepts may be related to each other in certain contexts. For example, "bank" and "interest", whose meanings are apparently dissimilar, have some semantic relatedness since generally co-occur in a financial article. In terms of lexical resources, existing semantic similarity measurements are generally classified into knowledge-based methods and corpus-based methods. Knowledge-based measures rely on inherent structure and information content of priori knowledge bases or semantic lexicons such as WordNet (Miller, 1995) and Gene ontology, however, it is limited by the size of knowledge base; While corpus-based measures employ distributional properties of occurrence of words in a given corpus, such as British National Corpus, Wikipedia, and web search. Knowledge-based measures are considered more useful for evaluating the semantic similarity with predefined semantic relationships between words, on the contrary, corpus-based measures are more helpful for assessing the semantic relatedness using the co-occur statistics, but it is hard to show the relationships between words.

Nevertheless, most of existing works employ either knowledge base or corpora, which may suffer from the insufficiency of semantic information and the ambiguity contained in raw corpora. Therefore, some works combine knowledge bases with corpora for extracting the accurate semantic of words and measuring the semantic similarity. In the article by (Jiang and Conrath, 1997) extended WordNet-based method by adding the information content from corpus, and indicated the structural information sources from the lexical taxonomy WordNet, which consists of edge, depth, density and link type. The work by (Li, 2003) replaced the local semantic density in WordNet by the information content derived from corpus for measuring semantic similarity with distinct lexical sources. Moreover, due to polysemy and homonymy, there is other works conduct multiple sense-specific vector representation per word in distributional vector space or distributed vector space. In these multi-prototype vector representations, semantic similarity between two words is then computed as the max similarity or the average similarity of all pairs of prototype vectors.

Most of existing works employ either knowledge base or corpora, which may suffer from the insufficiency of semantic information and the ambiguity contained in raw corpora, to solve this problem, in this work, the authors proposed a knowledge-augmented multiple-prototype vector representation model which represents each word by a set of distinct vectors. The semantic similarity between two words is therefore computed as the max among the similarity values of multiple vector pairs. Inspired by the emerging distributed word representation in deep learning, the authors primarily leverage the corpora to train distributed word vector via neural network language model. Then the authors use WordNet as additional knowledge with unambiguous semantics to augment the semantic information of corpora and conduct multiple prototype vectors per word in low-dimensional vector space. This work contributes to