A Review of Recent Trends: Text Mining of Taxonomy UsingWordNet 3.1 for the Solution and Problems of Ambiguity in Social Media

Ali Muttaleb Hasan, Taha Hussein Rassem, Noorhuzaimi Mohd Noor,
and Ahmed Muttaleb Hasan
Faculty of Computing (Fkom), University Malaysia Pahang, DarulMakmur, Gambang, 26300
Kuantan, Pahang, Malaysia

ABSTRACT

Text processing has been playing a great role in information retrieval to solve the problem of ambiguity in natural language processing, e.g., internet search, data mining, and social media. In semantic similarity, it will be used to analyze the relationships between Word-Pairs on social media. Organizing a huge number of unstructured text documents into a small number of concepts of word sense disambiguation is essential so that the lexical source could incorporate the features for capturing more semantic evidence. Text mining involves the pre-processing of documents collections, text categorization and classification, and extracting information and terms from golden standard data sets. This work proposed the lexical sourced from the semantic representation. The paper contained an evaluation of the advanced measures, which include shortest path, depth, and information content measures. In this paper, we used the same set of measures as previous studies, but different methods such as taxonomy on social media by semantic similarities, such as Synonymy (https://github.com/alimuttaleb/Ali-Muttaleb/blob/master/Synonym.txt), Non-taxonomy, Hypernym, and Glosses. This paper has focused to address the synonymy and ambiguity by incorporating the knowledge in the lexical resources. Thus, each word in a document is linked to its corresponding concept in the lexical resources. To build the semantic representation, these approaches can be classified into two main approaches: knowledge-based and statistical approaches. The knowledge-based approaches depend on structured information that is normally available in forms of dictionaries, thesaurus, lexicons,WordNet 3.1, and ontologies The statistical approaches are based on finding the semantic relations among words using the frequencies of words in a given corpus.

DOI: https://doi.org/10.1007/978-981-15-3284-9_17

KEYWORDS: Semantic similarity, Shortest path measures, Depth measures, Information content measures, WordNet 3.1, Text mining, Knowledge-based Approach

ACKNOWLEDGEMENTS

This work is supported by the University Malaysia Pahang (UMP) via Research Grant UMP (RDU1803141, PGRS190398).