Chapter 2
A New Similarity Measure for Automatic Construction of the Unknown Word Lexical Dictionary

Myunggwon Hwang
Chosun University, South Korea

Pankoo Kim
Chosun University, South Korea

ABSTRACT
This paper deals with research that automatically constructs a lexical dictionary of unknown words as an automatic lexical dictionary expansion. The lexical dictionary has been usefully applied to various fields for semantic information processing. It has limitations in which it only processes terms defined in the dictionary. Under this circumstance, the concept of “Unknown Word (UW)” is defined. UW is considered a word, not defined in WordNet, that is an existing representative lexical dictionary. Here is where a new method to construct UW lexical dictionary through inputting various document collections that are scattered on the Web is proposed. The authors grasp related terms of UW and measure semantic relatedness (similarity) between an UW and a related term(s). The relatedness is obtained by calculating both probabilistic relationship and semantic relationship. This research can extend UW lexical dictionary with an abundant number of UW. It is also possible to prepare a foundation for semantic retrieval by simultaneously using the UW lexical dictionary and WordNet.

INTRODUCTION
Extensive research has been carried out on semantic information processing based on a lexical dictionary. Research on semantic document indexing (Hemayati et al., 2007, Kiryakov et al., 2003), on semantic metadata creation (Jovanovic et al., 2006, Handschuh et al., 2004, Alani et al., 2003), on generating semantic Web content using natural language processing (Java et al., 2007), and on detecting document topics (Kong et al., 2006) has been done. In addition, a lexical dictionary
can become a basis for query expansion (Liu et al., 2004), ontology extension (Navigli et al., 2006, Velardi et al., 2007), information retrieval by semantic similarity (Hliaoutakis et al., 2006), and knowledge integration (Missikoff et al., 2003). While the lexical dictionary is gradually becoming more important and useful for semantic information processing, the most vulnerable aspect of semantic information processing is that it cannot handle concepts that are not defined in the dictionary. This is also true for WordNet, the most representative English lexical dictionary. Even though WordNet has been expanded with manual effort for over ten years, it cannot completely cover all the new words created by changes in real life such as new social phenomena, trends, techniques, product names, and famous persons’ name (athletes, entertainers, and politicians). Manually defining such new words requires a lot of time, cost, labor, and controversy (Velardi et al., 2001, Navigli et al., 2003). This hindrance led to many research endeavors to reduce such factors in constructing or expanding knowledge bases using Web documents and data warehouses (Missikoff et al., 2003, Navigli et al., 2004, Navigli et al., 2006, Velardi et al., 2007). These methods have good result but they fail to not deal with relationships and similarity of terms.

WordNet contains semantic relations between concepts. Many research projects have quantified such relations using various methods and have applied the results for semantic information processing. Its limits are apparent since documents frequently contain words not defined in the dictionary. These undefined words are called “Unknown Word (UW).” For true semantic information processing, we search UWs and terms which relate to the UW. The probabilistic weight based on Bayesian probability and the semantic weight based on WordNet will be calculated to ultimately find the semantic relatedness between an UW and a related term(s). In addition since the dictionary handles concepts rather than words, we apply our newly designed word sense disambiguation (WSD) method, which enables the dictionary to have an accurate synset for related terms. As a result, the dictionary can be compatibly used with WordNet. This research is based on the relation structure of WordNet and uses the synset, therefore it can be regarded as a type of automatic WordNet expansion.

The following paper explains the preprocessing step that covers noun extraction from a document group, UW determination, and the new WSD method. In using these results, terms that relate to the UW are measured for similarity with additional processes. Experimental results are evaluated using various methods in order to validate this approach. Following the conclusion, additional opportunities are suggested for future research.

### PREPROCESSING

WordNet\(^1\), which is used as a knowledge base in this research, defines almost all English terms (Miller et al., 1990, Fellbaum et al., 1998). In this section, the methods for extraction of noun and the WSD method will be described.

#### Extraction of an Unknown Word

In order to extract noun terms from a Web document, pos (part-of-speech) tagging is necessary. We chose pos-tagger version 2006-05-21, developed and provided by The Stanford Natural Language Processing Group\(^2\) (Toutanova et al., 2000, Toutanova et al., 2003). Inserting text into the pos-tagger allows it to assign a POS tag for each word. Among the tags, NN (single noun), NNS (compound noun), NNP (proper noun), and NNPS (compound proper noun) are the types of nouns. Words tagged with NN or NNS are used to create a noun term candidate list (NL\(_c\)) and NNP or NNPS are used to create an unknown word candidate list (UL\(_c\)).

Let us demonstrate with one sentence. “Zinedine Yazid Zidane, popularly nicknamed Zizou,