Detection of new words and their senses in Twitter data using Wikipedia

Apichai Chan-Udom*, Chan Karman†, Yoshimi Suzuki*

*University of Yamanashi
4-3-11, Takeda , Kofu, Yamanashi
{G15DM001, ysuzuki}@yamanashi.ac.jp

†IIJ Innovation Institute
Fujimi, Chiyoda-ku, Tokyo, Japan
chan@iij-ii.co.jp

Abstract

With the growth of real-time information sharing capability, social media site such as Twitter, Facebook becomes one of the most widely-used. Many semantic-oriented applications such as Question Answering, Text Categorization, Sentiment Analysis and Text Summarization by using twitter data need to recognize which words may substitute each other in a meaning preserving manner ((Lin and Pantel, 2001),(Barzilay and McKeown, 2001),(Quirk et al., 2004) and (Metzler et al., 2011)). Fine-grained and large-scale semantic knowledge like WordNet (Fellbaum, 1998), CIMLEX, EDR dictionary, and Bunrui-Goi-Gyo (BGH) is indispensable for these applications. However, such a resource makes it nearly impossible to cover large and fast-changing linguistic knowledge required for these application systems, depending on text-type and subject domain. For example, noun words in WordNet 3.0 consists of 117,798 words. These words are organized into fine-grained classification, 82,115 synsets, each consists of synonyms. The average number of words in a synset is very few, 1.24 per a synset Therefore, considering this resource scarcity problem, semantic tagging of words which do not appear in the WordNet resource but appear in Twitter data has been an interest since the earliest days when a number of large scale corpora have become available. In this paper, we present a method for detecting new words in twitter data which do not appear in the WordNet resource, and identify their senses in order to classify into synsets in the WordNet. We utilized abstract information of Wikipedia, and distributional word representation obtained by Word2vec (Mikolov et al., 2013). Wikipedia forms a rich semantic network connecting entities and concepts, enabling it as a valuable source for knowledge harvesting. Word2vec describes a word vector embedding that is quite commonly used in NLP research. We used these and identified new words and their related synsets.

Keyword: Twitter, distributional word representation, WordNet

1. References

Barzilay, R. and K.R. McKeown, 2001. Extracting paraphrases from a parallel corpus. In *In Proc. of 39th Annual Meeting of the Association for Computational Linguistics*.

Fellbaum, C., 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Lin, D. and P. Pantel, 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7, No.4:343–360.

Metzler, D., E. Hovy, and C. Zhang, 2011. An empirical evaluation of data-driven paraphrase generation techniques. In *In Proc. of 49th Annual Meeting of the Association for Computational Linguistics*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean, 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26th, Curran Associates*.

Quirk, Chris, Chris Brockett, and William Dolan, 2004. Monolingual machine translation for paraphrase generation. In *In Proc. of the 2004 Conference on Empilical Methods in Natural Language Processing*.