Introduction

Topics in a text corpus include features and information. Analyzing these topics can improve a user’s understanding of the corpus. These topics can be divided into two types: those whose meaning can be described in one word and those whose meaning is expressed through a recurring combination of words, also known as multiword expressions (MWE). Out of context, the MWE ‘she sets the bar high’ is ambiguous between a literal and a metaphorical reading. Ambiguity resolution is needed to extract accurate topics. Several well-known techniques have been proposed for topic extraction: TF*PDF (Khoo Khyou Bun et al., 2002), Topic Detection and Tracking (Kuan-Yu Chen et al., 2007), LDA (T. L. Griffiths and M. Steyvers, 2004), inter alia. However, most of these techniques target single words, not MWEs. In this paper, we propose a system that extracts MWE-based topics accurately. Our algorithm breaks down into six steps: Recognition, Pre-Processing, Processing, Candidate Extraction, Topic Validation, and Storing. We benchmark the Evaluation step using ambiguous sentences. Results show that our algorithm identifies MWEs faster and more accurately. This is because it detects problematic expressions, parses them in the light of a repository of resolved MWEs, and manages to provide a correct interpretation. Compiling a repository of MWEs that are correctly parsed and interpreted is time consuming. We show how this can be solved in the near future.

Case study

We present a data processing architecture for extracting MWEs from the corpus (Figure 1). In the processing, candidate words were extracted from a combination of tokenized words with N-grams and reference relations of words with Dependency structure. Dependency is the notion that linguistic units, e.g. words, are connected to each other by directed links (Melčuk, Igor A., 2012). Figure 2 illustrates how we can extract multiword candidates with dependency structure from the example sentence: ‘Shall I wake him up?’

Figure 1: Data processing structure. Framework for topic acquisition from corpus data.

Figure 2: Words candidates from Dependency structure

Conclusion

Often, MWEs cause problems; Google translation, Stanford CoreNLP, etc. Those problems can be solved if the algorithm can extract and recognize MWEs correctly. For this reason, we made a parsing algorithm to extract MWEs from the corpus. The case study shows how to extract MWEs. In the near future, we will create a web application based on our algorithm to visualize the results and integrate the user’s input on MWEs.