HOW CAN WE CAPTURE MULTIWORD EXPRESSION?

Seongmin Mun[1], Guillaume Desagulier[2], Kyungwon Lee[3]
UMR 7114 MoDyCo - CNRS, University Paris Nanterre, France[1]
Life Media Interdisciplinary Program, South Korea[1]
UMR 7114 MoDyCo - University Paris 8, CNRS, University Paris Nanterre, France[2]
Department of Digital Media, South Korea[3]

ABSTRACT

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 in 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 the 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.

KEYWORDS

Multiword expression, Topics extraction, Lexeme, Syntax analysis, Semantic analysis, Parsing

REFERENCES

Journal

Thomas L. Griffiths and Mark Steyvers, 2004, Finding scientific topics, Proceedings of the National Academy of Sciences, Vol. 101, pp 5228-5235.

Kuan-Yu Chen. et al., 2007. Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 19, No. 8, pp 1016-1025.

Conference paper or contributed volume

Khoo Khyou Bun and M. Ishizuka, 2002, Topic extraction from news archive using TF*PDF algorithm, WISE '02 Proceedings of the 3rd International Conference on Web Information Systems Engineering, pp 73-82.