PARSEME Short term Scientific Mission at the Center for the Study of Language and Information (CSLI) at Stanford University

Meghdad Farahmand

Host: Dr. Dan Flickinger

October-December 2015

1 Introduction

MWEs have application in many different NLP areas. For instance they can improve the quality of topic models [5]. [4] discusses the important role that MWE play in corpus linguistics. Efficient extraction of MWEs can improve the performance of applications such as information extraction [1], language generation [2, 6], parsing [7], statistical machine translation [8], and statistical language models [3].
Noun compounds are one of the most frequent and productive types of MWEs in English. They are identified with two main types of idiosyncrasy namely statistical and semantic idiosyncrasy. Our research is focused on noun compounds. Efficient identification of noun compounds is an important task because firstly it targets one of the most frequent classes of MWEs, and secondly it deals with the two mentioned types of idiosyncrasy that are relevant for other types of MWEs as well.

2 Description of the work accomplished during the STSM

During this visit we studied different aspects of MWEs ranging from their definition to their distinguishing properties. We focused our work on English noun compounds due to their high frequency and the fact that they represent most prominent types of idiosyncrasy, i.e. semantic and statistical idiosyncrasy.
We discussed the ambiguity of MWEs and looked for ways to tackle this problem. We studied phenomena such as non-substitutability that can be used to

model the statistical idiosyncrasy and consequently help identify MWEs. Non-substitutability is a property of MWEs and it means that the components of a MWE can not be replaced with their near synonyms. We defined a number of features based on word embedding representation of the components of a MWE that model different aspects of non-substitutability. We particularly focused on non-substitutability partly because it is relevant for all classes of MWEs and partly due to the little amount of research that has been done on this property and different ways of modeling it.
We trained a maxent classifier to identify MWEs starting with one statistical feature and moving on to other features that are inspired by non-substitutability. We analyzed the parameters of the maxent classifier to understand which aspect of non-substitutability is more relevant for identification of MWEs. In addition to development of a model that can be used to efficiently extract MWEs, we hope that our research have provided some insights into non-substitutability and shed some lights to some of the aspects of this property that have been seemingly ambiguous.
Additionally, we studied verb particle constructions to some extent and tried to apply non-substitutability to this type of MWEs. We studied different cases where identification of verb particles can be useful in other NLP tasks.
We also discussed different ways of modeling human intuition on understanding statistical idiosyncrasy and identifying MWEs. We defined a set of guidelines and developed tools and resources for human judges to help them make a better use of their intuition in identifying MWEs.

3 Outcomes of the STSM

We described the work that has been done during this visit in a short paper that was submitted to a major CL conference. Therein we present the detail of this research and its evaluation. Moreover, we are finalizing the annotation of a dataset that was created during this visit. This dataset comprises a set of English noun compounds that are annotated with their level of collocational strength and a decision about their semantic non-compositionality. This dataset will soon become freely available.

4 Confirmation by the host institution of the successful execution of the STSM

Dan Flickinger, CSLI, Stanford University I hereby confirm that the work which Meghdad describes in this report was indeed completed during his research visit at CSLI in 2015. I am also happy to note that I found the collaboration with Meghdad on noun-noun collocations in English to be especially productive, and I look forward to continued work with him as we refine our results for publication. My thanks to the PARSEME consortium for making his visit possible.

References

[1] Timothy Baldwin and Su Nam Kim. Multiword expressions. *Handbook of Natural Language Processing, second edition. Morgan and Claypool*, 2010.

[2] Stefan Evert. *The statistics of word cooccurrences*. PhD thesis, Dissertation, Stuttgart University, 2005.

[3] Stefan Evert. Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2:223–233, 2008.

[4] Stefan Th Gries. 50-something years of work on collocations: what is or should be next.... *International Journal of Corpus Linguistics*, 18(1):137–166, 2013.

[5] Jey Han Lau, Timothy Baldwin, and David Newman. On collocations and topic models. *ACM Trans. Speech Lang. Process.*, 10(3):10:1–10:14, July 2013.

[6] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[7] Joakim Nivre and Jens Nilsson. Multiword units in syntactic parsing. In *In Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*, 2004.

[8] Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 47–54. Association for Computational Linguistics, 2009.