SHORT TERM SCIENTIFIC MISSION

Sentence simplifications for Multi-word Expressions

Shashi Narayan

May 27, 2014

1 Short Term Scientific Mission (STSM)

STSM Topic: Sentence simplifications for Multi-word Expressions

STSM Applicant: Shashi Narayan (shashi.narayan@loria.fr), Université de Lorraine, LORIA, Nancy France

STSM Host: Mirella Lapata (mlap@inf.ed.ac.uk), School of Informatics, University of Edinburgh, UK

STSM Period: May 11th, 2014 to May 23rd, 2014

COST Action: IC1207 PARSEME

COST MC Chair: Dr. Agata Savary (agata.savary@univ-tours.fr)

STSM Reference Number: COST-STSM-ECOST-STSM-IC1207-300314-044347

STSM Type: Regular (from France to United Kingdom)

2 Purpose of the STSM

Sentence simplification plays an important role in various NLP tasks such as parsing and machine translations [Chandrasekar et al., 1996], summarisation [Knight and Marcu, 2000], sentence fusion [Filippova and Strube, 2008], semantic role labelling [Vickrey and Koller, 2008] etc. In this short term scientific mission (STSM), we study the importance of sentence simplification in wide ranging societal application such as a reading aid for people with aphasia [Carroll et al., 1999], for low literacy readers [Watanabe et al., 2009] and for non native speakers [Siddharthan, 2002]. We believe that sentences with multi-word expressions create a major challenge for such communities and therefore, needs a crucial attention during the sentence simplification.

The sentence simplification literature either overlooks multi-word expressions or does not consider them at all. The supervised systems claims to achieve multi-word simplification with phrase substitution but their strength is limited with the small sized parallel corpora of complex/simple sentence pairs. This mission aimed at discovering resources and techniques that could be used for multi-word expression simplification.

3 Description of the work carried out during the STSM

Over the two weeks period, I had several productive meetings with the STSM host Mirella Lapata. We looked and discussed the problem in hand. In particular, we explored the paraphrase database (PPDB, [Ganitkevitch et al., 2013]) for various kinds of multi-word expressions and discussed the ways it could be used for simplification. I also had very fruitful discussions with Mark Steedman, Kristian Woodsend, Siva Reddy, Mike Lewis, yannis konstas, Alexandra Birch and few other researchers at ILCC.

During the second week, I got an opportunity to give a seminar on "Hybrid Simplification using Deep Semantics and Machine Translation" [Narayan and Gardent, 2014] at the weekly meeting of Probabilistic Models of Language Group. We got very good response and the use of semantics in simplification was much appreciated.

4 Description of the main results obtained

We explored PPDB for paraphrases consisting of various kinds of multi-word expressions such as fixed expressions, semi-fixed expressions and syntactically-flexible expressions. Some of the example paraphrases, we found, are shown below:

- "*by and large*" → "*generally*" or "*overall*" or "*broadly*" (Fixed expressions)
- "*in short*" → "*in summary*" (Fixed expressions)
- "*ad hoc*" → "*special*" (Fixed expressions)
- "*take action [ADVP]*" → "*to act [ADVP]*" (Light verbs, syntactically-flexible expressions)

During our study of PPDB, we also found that PPDB also covers dative shifting, passivization, relative clauses and possessive rephrasing.

We also discussed on the ways of developing an unified simplification model combining semantics and the rules from PPDB.

5 Future collaboration with host institution

One major question is that rephrasing does not always mean simplification. In future, we plan to collaborate over the development of an unsupervised unified simplification model which exploit semantics for splitting and deletion [Narayan and Gardent, 2014] and the rephrasing rules from PPDB tuned for simplification [Ganitkevitch et al., 2011].

6 Foreseen publications/articles resulting or to result from the STSM

We plan to submit the outcome of our ongoing research to a conference or a journal soon.

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

I (Mirella Lapata) gladly confirm the successful visit of Shashi to Edinburgh. I had multiple productive meeting with him over the last two weeks of his visit, during which we discussed over our approaches to sentence simplification and problems addressing the issues of multiword expressions. During his second week, he gave an interesting seminar on his simplification approach during our weekly meeting. We plan to continue our collaboration on the ongoing research.

References

[Carroll et al., 1999] Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, volume 99, pages 269–270. Citeseer.

[Chandrasekar et al., 1996] Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th International conference on Computational linguistics (COLING)*, pages 1041–1044. Association for Computational Linguistics.

[Filippova and Strube, 2008] Filippova, K. and Strube, M. (2008). Dependency tree based sentence compression. In *Proceedings of the Fifth International Natural Language Generation Conference (INLG)*, pages 25–32. Association for Computational Linguistics.

[Ganitkevitch et al., 2011] Ganitkevitch, J., Callison-Burch, C., Napoles, C., and Van Durme, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1168–1179, Edinburgh, Scotland, UK. Association for Computational Linguistics.

[Ganitkevitch et al., 2013] Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

[Knight and Marcu, 2000] Knight, K. and Marcu, D. (2000). Statistics-based summarization-step one: Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI) and Twelfth Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pages 703–710. AAAI Press.

[Narayan and Gardent, 2014] Narayan, S. and Gardent, C. (2014). Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL) on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics.

[Siddharthan, 2002] Siddharthan, A. (2002). An architecture for a text simplification system. In *Proceedings of the Language Engineering Conference (LEC)*, pages 64–71. IEEE Computer Society.

[Vickrey and Koller, 2008] Vickrey, D. and Koller, D. (2008). Sentence simplification for semantic role labeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL) and the Human Language Technology Conference (HLT)*, pages 344–352.

[Watanabe et al., 2009] Watanabe, W. M., Junior, A. C., Uzêda, V. R., Fortes, R. P. d. M., Pardo, T. A. S., and Aluísio, S. M. (2009). Facilita: reading assistance for low-literacy readers. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 29–36. ACM.