Addressing Cross-Linguistic Divergences for Machine Translation Using Cognitive Annotation: A French-English Case Study

Elior Sulem[1], Omri Abend[2], and Ari Rappoport[3]
[1]Institute of Computer Science, The Hebrew University of Jerusalem (eliors@cs.huji.ac.il)
[2]School of Informatics, University of Edinburgh (oabend@inf.ed.ac.uk)
[3]Institute of Computer Science, The Hebrew University of Jerusalem (arir@cs.huji.ac.il)

The integration of linguistic information into Machine Translation (MT), that was usually required before the development of statistical models for translation [Brown et al., 1993], has been the object of a new interest during these last few years. Indeed, syntax-based models for SMT [Chiang, 2005, Liu et al., 2006, Mi et al., 2008] take into account the hierarchical structure of the language and capture reordering at the global level. However, syntactic structures change across languages and these cross-linguistic divergences, characteristic of the integration of linguistic information in MT [Dorr, 1994], are still a main problem when the integration is done in statistical models [Ding and Palmer, 2004, Zhang et al., 2008].

Looking for a formalism which conserves structures across languages, a promising option is the use of semantics. However, semantic schemes often follow the syntax of a specific language, as in the case of Propbank [Palmer et al., 2005], and can also be affected by many structural divergences as shown in the English-Czech comparison of AMR (Abstract Meaning Representation) [Xue et al., 2014]. In this context, we propose using UCCA (Universal Cognitive Conceptual Annotation) [Abend and Rappoport, 2013], a structural annotation scheme which can be consistently applied to different languages and which is relatively robust to translation divergences.

In this work, we focus on English and French, and demonstrate UCCA's applicability to French through a systematic theoretical analysis of the major grammatical phenomena. We further construct an annotated parallel corpus of 583 sentences, which we intend to make publicly available. We perform a quantitative and qualitative analysis of the divergences between the languages in the corpus, showing that UCCA abstracts away from many of the divergences encountered by existing annotation schemes.

We in particular show that UCCA provides a better structure conservation than syntax. We also point out a correspondence of 93.6% between UCCA's Scenes in the two languages. Qualitative analysis of the divergences indicates how to avoid many of them, confirms UCCA's bilingual stability and leads to a methodology which can be applied to the analysis of other structural annotations for MT.

The present work therefore proposes an improved structural representation, which addresses cross-linguistic divergences, one of the main challenges in MT, and sets the foundations for integrating it into a full structure-based MT system.

References

Omri Abend and Ari Rappoport. Universal Conceptual Cognitive Annotation (UCCA). *Proc. of ACL-13*, pages 228–238, 2013.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19:263–311, 1993.

David Chiang. A hierarchical phrase-based model for statistical machine translation. *Proc.43rd Annual Meeting of the ACL*, pages 263–270, 2005.

Yuan Ding and Martha Palmer. Synchronous dependency insertion grammars: a grammar formalism for syntax based statistical MT. *Workshop on recent advances in dependency grammars, COLING-04*, 2004.

Bonnie J. Dorr. Machine translation divergences: a formal description and proposed solution. *Computational linguistics*, 20(4):597–635, 1994.

Yang Liu, Qun Liu, and Shouxun Lin. Tree-to-string alignment template for statistical machine translation. *Proc. of COLING-ACL-06*, pages 609–616, 2006.

Haitao Mi, Liang Huang, and Qun Liu. Forest-based translation. *Proc. of ACL-08 HLT*, pages 192–199, 2008.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

Nianwen Xue, Odrej Bojar, Jan Hajic, Martha Palmer, Zdenka Uresova, and Xiuhong Zhang. Not an intelingua, but close: comparison of English AMRs to Chinese and Czech. *Proc. of LREC-14*, pages 1765–1772, 2014.

Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Seng Li. A tree sequence alignment-based tree-to-tree translation model. *Proc. of ACL-08*, pages 559–567, 2008.