Combining, Adapting and Reusing Bi-texts between Related Languages: Application to Statistical Machine Translation (invited talk)

Preslav Nakov
Qatar Computing Research Institute
Tornado Tower, floor 10, PO box 5825
Doha, Qatar
pnakov@qf.org.qa

1 Abstract

Bilingual sentence-aligned parallel corpora, or *bi-texts*, are a useful resource for solving many computational linguistics problems including part-of-speech tagging, syntactic parsing, named entity recognition, word sense disambiguation, sentiment analysis, etc.; they are also a critical resource for some real-world applications such as statistical machine translation (SMT) and cross-language information retrieval. Unfortunately, building large bi-texts is hard, and thus most of the 6,500+ world languages remain resource-poor in bi-texts. However, many resource-poor languages are related to some resource-rich language, with whom they overlap in vocabulary and share cognates, which offers opportunities for using their bi-texts.

We explore various options for bi-text reuse: (i) direct combination of bi-texts, (ii) combination of models trained on such bi-texts, and (iii) a sophisticated combination of (i) and (ii).

We further explore the idea of generating bi-texts for a resource-poor language by adapting a bi-text for a resource-rich language. We build a lattice of adaptation options for each word and phrase, and we then decode it using a language model for the resource-poor language. We compare word- and phrase-level adaptation, and we further make use of cross-language morphology. For the adaptation, we experiment with (a) a standard phrase-based SMT decoder, and (b) a specialized beam-search adaptation decoder.

Finally, we observe that for closely-related languages, many of the differences are at the subword level. Thus, we explore the idea of reducing translation to character-level transliteration. We further demonstrate the potential of combining word- and character-level models.

2 Author's Biography

Dr. Preslav Nakov is a Scientist in the Arabic Language Technologies group at the Qatar Computing Research Institute (QCRI), Qatar Foundation. His research interests include computational linguistics, machine translation, lexical semantics, Web as a corpus, and biomedical text processing. His current research focus is on Arabic language processing, with an emphasis on statistical machine translation to/from Arabic.

Before joining QCRI, Dr. Nakov was at the National University of Singapore, where he worked on text and spoken language machine translation for Asian languages, including Chinese, Malay and Indonesian. Prior to that, he was at the Bulgarian Academy of Sciences and the Sofia University, where he was an honorary lecturer. He received his Ph.D. in Computer Science from the University of California at Berkeley in 2007, supported by a Fulbright grant and a Berkeley fellowship.

Dr. Nakov authored three books, one book chapter, and many research papers at conferences such as ACL, HLT-NAACL, EMNLP, ICML, CoNLL, COLING, EACL, ECAI, and RANLP, and journals such as JAIR, TSLP, NLE and LRE. He received the Young Researcher Award at RANLP'2011. He was also the first to receive the Bulgarian President's John Atanasoff annual award for achievements in the development of the information society (December 2003); the award is named after an American of Bulgarian ancestry who co-invented the first automatic electronic digital computer, the Atanasoff-Berry computer.

Dr. Nakov has served on the program committee of the major conferences and workshops in computational linguistics, including as a co-organizer and an area/publication/tutorial chair.

Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants, page 1,
Hissar, Bulgaria, 13 September 2013.