Reusing Parallel Corpora between Related Languages (invited talk)

Preslav Nakov
National University of Singapore
nakov@comp.nus.edu.sg

Abstract

Recent developments in statistical machine translation (SMT), *e.g.*, the availability of efficient implementations of integrated open-source toolkits like Moses, have made it possible to build a prototype system with decent translation quality for any language pair in a few days or even hours. This is so in theory. In practice, doing so requires having a large set of parallel sentence-aligned bilingual texts (a *bi-text*) for that language pair, which is often unavailable. Large high-quality bi-texts are rare; except for Arabic, Chinese, and some official languages of the European Union (EU), most of the 6,500+ world languages remain resource-poor from an SMT viewpoint. This number is even more striking if we consider language *pairs* instead of individual languages, *e.g.*, while Arabic and Chinese are among the most resource-rich languages for SMT, the Arabic-Chinese language pair is quite resource-poor. Moreover, even resource-rich language pairs could be poor in bi-texts for a specific domain, *e.g.*, biomedical text, conversational text, *etc.*

Due to the increasing volume of EU parliament debates and the ever-growing European legislation, the official languages of the EU are especially privileged from an SMT perspective. While this includes "classic SMT languages" such as English and French (which were already resource-rich), and some important international ones like Spanish and Portuguese, many of the rest have a limited number of speakers and were resource-poor until a few years ago. Thus, becoming an official language of the EU has turned out to be an easy recipe for getting resource-rich in bi-texts quickly.

Our aim is to tap the potential of the EU resources so that they can be used by other non-EU languages that are closely related to one or more official languages of the EU.

We propose to use bi-texts for resource-rich language pairs to build better SMT systems for resource-poor pairs by exploiting the similarity between a resource-poor language and a resource-rich one.

We are motivated by the observation that related languages tend to have (1) similar word order and syntax, and, more importantly, (2) overlapping vocabulary, *e.g.*, *casa* (house) is used in both Spanish and Portuguese; they also have (3) similar spelling. This vocabulary overlap means that the resource-rich auxiliary language can be used as a source of translation options for words that cannot be translated with the resources available for the resource-poor language. In actual text, the vocabulary overlap might extend from individual words to short phrases (especially if the resource-rich languages has been transliterated to look like the resource-poor one), which means that translations of whole phrases could potentially be reused between related languages. Moreover, the vocabulary overlap and the similarity in word order can be used to improve the word alignments for the resource-poor language by biasing the word alignment process with additional sentence pairs from the resource-rich language. We take advantage of all these opportunities: (1) we improve the word alignments for the resource-poor language, (2) we further augment it with additional translation options, and (3) we take care of potential spelling differences through appropriate transliteration.

Speaker's Bio

Dr. Preslav Nakov is a Research Fellow at the National University of Singapore. He received his PhD in Computer Science from the University of California at Berkeley in 2007. Dr. Nakov's research interestes are in the areas of Web as a corpus, lexical semantics, machine translation, information extraction, and bioinformatics.

Proceedings of Recent Advances in Natural Language Processing, page 1,
Hissar, Bulgaria, 12-14 September 2011.