Phrase-based SMT and Solutions to 'Out of Order' Problem

Alok Parlikar

Document of Intent for 11-734; Spring 2008

Abstract

Each language has its own word order of rules. Some languages the verb in the final, some other language, the verb must be requested in the second position. Machine translation is clearly not just a word or phrase mapping a language to another language. We need to get the sequence right way, so that it is consistent with some that will have a native language of it. The problem reordering is very special of Phrase-based SMT in order to make the output sentence more grammatical. In this review, I would like to address this importance of issue. I would briefly review the phrase and POS based approaches then focus mainly on syntax driven approaches to word of reordering.

Note: This abstract has been translated into English from a non-european foreign language by using a current MT system.

Scope and Objectives

When translating between a language pair such as Chinese-English, the average phrase length that phrase based systems typically find is less than two words. This has a problem: the output sentence may not be grammatically correct, especially if the word orders in source and target languages are not consistent. There are several approaches to getting the correct order of words in the output. In this paper, I would like to look at few of these approaches. I intend to briefly look at the principles of phrase-based and POS-based reordering, and then concentrate in the rest of the part on various syntax-based approaches that people have reported to (or to not) work.

I am thinking of the following as an outline for the content matter. The context in each of these is a phrase based SMT system.

- (05%) Need for reordering and complexity among different language pairs. Is any language pair currently of particularly high interest? Why?
- (15%) General principles behind phrase-based reordering approach.
- (15%) General principles behind POS-based reordering approach.
- (50%) Different approaches to syntax-based reordering. (Reordering strategies using different data structures: c-structure/dependency trees; rule-based reordering: linguistic and data-driven approaches, etc)
- (15%) Future scope of syntax-based techniques.

Boot-up Readings

- Kanthak et al. (2005) Novel Reordering Approaches in Phrase-Based Statistical Machine Translation
- Crego et al. (2006) Integration of POStag-based source reordering into SMT decoding by an extended search graph
- Collins et al. (2005) Clause restructuring for statistical machine translation)
- Wang et al. (2007) Chinese Syntactic Reordering for Statistical Machine Translation