Statistical Structure in Language Processing
Project 3 research proposal

Cristina Gârbacea
10407936
cr1st1na.garbacea@gmail.com

Sara Veldhoen
10545298
sara.veldhoen@student.uva.nl

1 Introduction

This proposal continues to build on the phrase based machine translation approach. In principle, the set of possible phrase pairs is far too huge to be computationally manageable. Especially if the size of the corpus grows, the Therefore, we need to restrict this set in some way, to keep only useful/ meaningful phrase pairs. This reduction might be the main challenge in phrase based machine translation.

In the baseline approach we took in the previous assignment, this reduction was based on Giza++ output: only those phrase pairs that were consistent with symmetrized word alignments were added to the phrase table. This yields a large reduction and proves useful for translation, obtaining for instance a BLEU score of 24.68 in our previous assignment (1). But the MT community has moved forward, and can do better now.

One possibility that we aim to investigate in this project, is to use multi-parallel corpora. These data consist of documents in more than two languages with aligned sentences. The process of incorporating evidence from multilingual data in a single system is called *triangulation*.

2 Related work

Two methods are presented in In (2) to filter the phrase table for a language pair which we will dubb *source-target*, based on an intermediate third *bridge* language. Both methods assume an existing *source-target* phrase table, based on Giza++, and filter its entries with evidence from phrase tables *source-bridge* and *bridge-target* where *bridge* is an intermediate language.

In method 1, for each phrase pair $\langle s, t\rangle \in source - target$, it is kept if there is an entire phrase b in the bridge language such that $\langle s, b\rangle \in source-bridge$ and also $\langle b, t\rangle \in bridge-target$.

Method 2 is somewhat more lenient, in that it looks at the words occurring in the phrases instead of an exact match of the entire phrase. An overlap score is assigned to each phrase pair, based on the intersection of the vocabularies in each phrase. The filtering is done by placing a threshold on this score.

The authors of (3) focus less on reducing the phrase tables, but use triangulation to obtain high quality phrase tables from multilingual data, even if they are not from the same corpus. They use a summation over the several intermediate languages to form a probability estimate: $p(s|t) = \sum_i p(s|i)p(i|t)$. Interestingly, the triangulated phrase table is trained separate from the standard phrase table, so that it can be used as a distinct feature in decoding.

3 Goals

In this project we are planning to build a phrase pair extraction tool for phrases of up to length 7 that would use evidence from multilingual aligned corpora. To achieve our goal we aim to use Europarl data for Dutch, English, Romanian, French, German. We are planning to investigate how we can use the evidence from word alignments between several languages to extract phrase pairs and estimate their conditional and joint probabilities. We also want to focus on determining if we can achieve better symmetrization compared to the case when all phrase pairs are consistent with a symmetrized word alignment.

References

[1] Cristina Gârbacea and Sara Veldhoen, 2014. *Phrase based models*, Statistical Structure in Language Processing assignment

[2] Yu Chen, Andreas Eisele, Martin Kay, 2008. *Improving Statistical Machine Translation Efficiency by Triangulation*, LREC

[3] Trevor Cohn and Mirella Lapata, 2007. *Machine translation by triangulation: Making effective use of multi-parallel corpora*, ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS