Statistical Structure in Language Processing
Using multi-parallel data for phrase-table improvement

Cristina Gârbacea
10407936
cr1st1na.garbacea@gmail.com

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

1 Introduction

In principle, the set of possible phrase pairs is far too huge to be computationally manageable, especially if the size of the corpus grows. 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 approach we took in the previous assignment, the reduction was based on Giza++ output: only those phrase pairs that were consistent with symmetrized word alignments were added to the phrase table.

The 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 (2) to filter the phrase table for a language pair, 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*.

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 the phrases. 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 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 Word-alignments

Phrase pair extraction is based on symmetrized word alignments.

The bidirectional word alignments are obtained with `Giza++`, which is a combination of IBM-models. Although the quality of these alignments is not very high on itself, they have proven to be a useful guide in the extraction of phrase pairs.

Because of the design of IBM models, the unidirectional word-alignments are one to many. For symmetrized word alignment, where many-to-many alignments are possible, both directions are used and their alignment points combined. Each word-alignment can be viewed as a matrix with words in both languages along the axes, and binary values in the cells: either the words are aligned, or not. If we combine the two directions, we introduce new values for the cells: the alignment point is either present in both directions ($\mathcal{B}$), in the for-tar alignment ($\mathcal{F}$), in the tar-for alignment ($\mathcal{T}$), or in none (empty).

The Moses alignment symmetrization is aimed to recast such a matrix back into a binary matrix. Union and intersection are the extreme choices, several heuristics are available in Moses that compromise them:

- `union` contains all non-empty cells. This heuristic has a high recall, but might have many false positives.
- `intersection` keeps only the $\mathcal{B}$ cells. These are the high-confidence points, thus

improving precision but (generally) dropping recall.

- `grow` starts from the intersection and adds block-neighboring non-empty cells.
- `grow-diag` is like `grow`, but it also includes diagonally neighboring non-empty cells.
- `final` is used after either heuristic, to add as many as possible unaligned words that are not neighbored.

In these heuristics, Moses does not distinguish between $\mathcal{F}$ and $\mathcal{B}$ cells. Since we use multi-parallel data that all translates to the same target (English), our extraction is not entirely symmetrical and this distinction might be of importance.

Our base-line

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