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

Abstract

In this paper we continue to extend upon our previous assignment by building a phrase pair extraction tool that would use evidence from aligned parallel corpora. Unlike in our previous work where we would only use parallel corpora consisting of two languages, this time we use evidence from multilingual parallel aligned corpora made available by Europarl. We investigate how word alignments between several languages can be of use in extracting consistent phrase pairs. We present an algorithm to do so, and experimented with different symmetrization heuristics to be used as an input. Furthermore, we investigated how the choice of reference languages influences improvement. We compare our results to a baseline system, and to a different approach to phrase table filtering, based on significance testing.

1 Introduction

Phrase-based statistical machine translation relies on estimates of translation probabilities for pairs of phrases, stored in a so-called phrase table. 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. This is a standard approach, that yields a large reduction and proves useful for translation, obtaining for instance a BLEU score of 24.68 in our previous assignment (1).

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*, and presents the advantage of using a wider range of parallel corpora for training.

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.

In (4) the bulk of phrase table is reduced based on the significance testing of phrase pair co-occurence in the parallel corpus. The authors present two methods of testing the significance of

associations in two by two contigency tables departing from calculating the probability that the observed table can occur by chance assuming a model of independence. For this purpose they use Chi-squared and Fisher exact test and experiment with different pruning threshold values ranging between 14 and 25. They outline that while the savings in terms of number of phrases discarded are considerable, the translation quality as measured by the BLEU score is preserved, and even more surprisingly it can even increase as compared to the case when all phrase pairs are kept. This makes their approach a valuable contribution to the field of machine translation.

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 Giza word-alignments are one to many. For symmetrized word alignment, where many-to-many alignments are possible, Giza alignments for 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 source-target alignment ($\mathcal{S}$), in the target-source 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 $\mathcal{S}$ and $\mathcal{T}$ 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

The symmetrized alignments are used for reducing the space of phrase pairs. That means that having less alignment points, as in the intersection heuristic, is actually the more lenient choice for the phrase extraction step: you do not discard phrase pairs based on weak evidence.

4 Experiments

4.1 Data

We use the Europarl version 7 parallel data for English to respectively French, German, Danish, Italian, Dutch, Spanish, and Portuguese. This choice was based on the fact that parallel data exists for parliament proceedings of the same time interval for these languages, so that they are in a sense parallel beyond language level. We excluded Greek because of the following remark: "Some recent Greek data has only parts of transcripts in the files." Moreover, the Greek data was not in a readable format, because of the deviant alphabet.

We use a script that comes with the Moses installation for initial preprocessing. It discards empty lines and lines with sentences that exceed a length threshold. Furthermore, all text is lowercased.

The Europarl recommends to set aside the Q4/2000 portion of the data for testing. However, the clean parallel corpora don't have time annotation, so there is no easy way to make this split. Therefore we decided to just create a split ourselves. Each 50th sentence is removed from the corpus and added to a separate test file. The ratio is based on the fact that the Q4/2000 portion would be 3/(16*12-4)=0.016 of the data.

Sentence alignment The parallel data is sentence aligned, which means that each line in the source text corresponds to a line in the target text.

In order for our phrase extraction to work, we want one single English (target) text file to be aligned to all languages.

Our implementation is quite ad hoc, in that it compares the English sides and just discards those lines that are different. This means we discard part of the data, because in the original parallel data sometimes multiple sentences are concatenated in the sentence alignment. This approach could probably be refined to lose less data, but due to time constraints and memory efficiency reasons we preferred to keep only those lines that are common in all English files.

In what follows we describe our sentence alignment approach. From all the English files in the list of parallel data that we have for the 7 language pairs, we pick one randomly (we chose the English file from the English - Danish corpus in this case) and we retain only those sentences that are present in the other English files belonging to the other corpora. In this way we aim to identify common English sentences in all English files from the parallel corpora that we have. For each such sentence found we preserve the index of the occurrence so that we can retrieve its corresponding translations in the foreign languages we use. Usually these indices will denote the position of the corresponding translated sentence inside the foreign files. However, we are aware that this will not always be the case. For computation efficiency reasons, instead of looping though all the lines in a foreign file in searching for the translated version of the English sentence, we prefer to keep track of the last retrieved position i_l from the English file and count the number of lines n_s skipped since then to look up the corresponding foreign phrase within a window ranging between i_l and $i_l + n_s * 3$. We chose 3 as a safe margin leaving from the premises that we should include in our search all the sentences located in the neighborhood of the expected index. The output of our preprocessing algorithm consists in the English file having only the common sentences and the 7 foreign files for each of the different languages we use containing translated versions of the English sentences which are now properly aligned with our English sentences.

4.2

We run Moses steps 1 and 2 with default setting on all language pairs, to obtain Giza word alignments.

5 Results

6 Conclusion

A major drawback of our approach is the constraint that the multi-parallel corpus has to be aligned, so that the target-sides coincide. Apart from the possibility to improve the tool for extracting such data from a corpus like Europarl, it may be interesting to investigate the possibility to abandon this constraint at all. For instance, by using an existing (bilingually trained) system to provide translations for the English sentences in other languages. This would however severely aggravate the training of the model, so the question is whether the reduction in phrase table expected from our method is worth it.

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

[4] J Howard Johnson, Joel Martin, George Foster and Roland Kuhn, 2007. *Improving Translation Quality by Discarding Most of the Phrasetable*, In Proceedings of EMNLP-CoNLL