The Effects of Pivot Language on Word Alignment Quality

Averie So
Final Project Paper for Computational Linguistics

Abstract

Previous research shows that using a third "pivot" language when training word alignment models can be beneficial to the translation accuracy of phrase-based SMT systems. The current project extends the pivot model from Kumar et al. (2007) by symmetrization and analyses the resulting phrases from these alignment models. With a multi-parallel corpus, a word alignment model from English to French is trained with Spanish as the pivot language. A pivot model is shown to contributes to the diversity of phrase hypotheses. We conclude that pivoted alignment models can be beneficial to phrase-based SMT despite the lack of improvements in widely used alignment quality measures.

1 Introduction

Word alignment is an important step within the pipeline in creating a phrase-based SMT. Phrases of a parallel corpus are directly extracted from word alignment models, and translation probabilities for phrases are then calculated based on these possible candidate phrases. This is the case in major SMT decoders such as Moses (Koehn et al., 2007). With evidence supporting the benefits of high quality word alignments on translation quality (Och and Ney, 2003), improving word alignment generation is therefore an essential task for phrase-baaed SMT research.

With the availability of multi-parallel corpora (parallel corpora between more than two languages), such as the United Nations parallel corpus (Ziemski et al., 2016) and European Parliament (Koehn, 2005), Kumar et al. (2007) explores the benefits of using pivot languages in the generation of word alignment models. One rationale behind this idea is that errors made in a direct word alignment model can potentially be resolved with a third alignment source. Secondly, by having multiple alignment models, more diverse phrases can potentially be extracted. In an Arabic-to-English translation task, they showed that bridged word alignment models with Spanish, French, Chinese and Russian can improve the BLEU scores of their SMT system.

The outcome of Kumar et al. (2007) provides a further insight which concerns the evaluation method of word alignments, which is commonly reported in terms of Precision, Recall and Alignment Error Rate (AER). While bridged alignment models improved translation quality, the AER were higher compared to the baseline model. This suggests that the surface quality of word alignments measured by AER may not fully reflect alignment quality, especially for the purpose of phrase-based SMT.

This project aims to replicate the bridged alignment model by Kumar et al. (2007) and extend it in two directions: (1) explore an alternative way in combining multiple word alignment models with symmetrization heuristics and (2) extract phrases from alignment models as a way of evaluation beyond AER.

The rest of this paper is organised as follows: Sections 1.1, 1.2 and 1.3 explain in technical terms the three main components of this project, which are constructing bridged word alignment models, the symmetrization heuristics and extracting phrases respectively. Section 2 describes in detail the experimental setup. Section 3 reports the alignment performance of the various models. Section 4 presents a discussion of this project.

1.1 Constructing word alignments with a bridge language

According to Kumar et al. (2007), the alignment probabilities between a sentence-pair in source language E and target language F, can be obtained using a matrix multiplication of the alignment probabilities between the source language and the pivot language G, and that between the pivot language and the target language. This approach was developed with the intention of being easily adapted from IBM models.

For each parallel sentence in English $\mathbf{e} = i_1^I$, French $\mathbf{f} = j_1^J$ and Spanish $\mathbf{g} = k_1^K$, a FE probability table is generated by multiplying the corresponding GE and FG probability tables. Consider the alignment $a_j^{FE} = i$ of French word *f* in position *j* to the English word *e* in position *i*. *j* can possibly be aligned to various Spanish words in positions $k \in \{0, 1, 2, \ldots, K\}$; equally, the same set of possible Spanish words can possibly be aligned to *i*. The alignment probability $P(a_j^{FE} = i|\mathbf{e},\mathbf{f})$ is therefore the sum of product of all the probabilities with various Spanish words in positions $k \in \{0, 1, 2, \ldots, K\}$ as pivot. This is captured by the expression below (from Kumar et al., 2007, Equation 4):

$$P(a_j^{FE} = i|\mathbf{e},\mathbf{f}) = \sum_{k=0}^{K} P(a_j^{FG} = k|\mathbf{g}, \mathbf{f})P(a_k^{GE} = i|\mathbf{g}, \mathbf{e})$$

Note that in order for this expression to be valid, some assumptions were made. This includes assuming that there is exactly one translation **g** in language G that correspond to the sentence pair **f** and **e** (Kumar et al., 2007).

1.2 Symmetrization heuristics

As experiments in Kumar et al. (2007) show, bridged word alignments do not show direct improvements through AER scores. It has also been pointed out that using bridge languages can introduce noise into word alignment models (Zadeh, 2009). The current project explores an alternative way of utilising bridged alignment models with symmetrization, a set of heuristics which takes into account alignments of both translation directions EF and FE (Och and Ney, 2003).

Symmetrization was originally motivated by the fact that real word alignments have many-to-many mappings, while word alignment models such as IBM models create many-to-one mappings (Koehn, 2016). One such set of heuristics is called the grow-diag-final method Koehn et al. (2005), it can be understood in terms of three steps (See Appendix A.1 for the pseudo-code):

(1) intersection: the first step preserves only the intersected alignment pairs between the two alignment models EF and FE. For example, an alignment point is preserved if EF contains 0-1 and FE contains 1-0. These intersected points obtains the highest precision since they are confirmed by both alignment directions.

(2) grow-diag: next, more alignment points are added from the union of the EF and FE alignments. Alignment points are added if an unaligned word neighbours an established alignment point. Neighbouring refers to any point directly next to an alignment point, including top, bottom, right, left and the diagonals.

(3) final: in the final step, points that do not neighbour other aligned points are added, if at least the E word or the F word is unaligned.

Considering that alignment symmetrization successfully incorporates information from multiple word alignment sources, the current project adapts this method with the bridged alignment model as the EF direction and the direct alignment model as the FE direction.

1.3 Extracting phrases

In phrase-based SMT, phrase extraction is the next step and a direct outcome of word alignment models. Following Koehn et al. (2005), we extract any phrase pair that is consistent with the word alignment. Consistent means that the words in the phrase pair are aligned to each other and not to any words outside. That is, given a hypothetical phrase pair defined by $(e_1 \ldots e_n)$ and $(f_1 \ldots f_m)$, the phrase is consistent if all $(e_1 \ldots e_n)$ that are aligned only correspond to members of $(f_1 \ldots f_m)$, and vice versa. If a word is unaligned, they can either be included or excluded in the current phrase, resulting in two target/source phrases corresponding to the same source/target phrase.

For our experiments, phrase extraction is intended as a way to reflect alignment quality beyond AER. Since there is no data for gold phrases to be extracted, it would not be fair to evaluate alignment quality by phrase quality. Given that the goal of bridged alignments is to increase the diversity of phrases, we conduct an analysis on the unique contributions made by the bridged model in addition to the baseline model.

2 Methodology

This section presents the experimental setup and functions in the script utilised to obtain the various models in our experiments.

2.1 Data

According to the approach in Kumar et al. (2007), a multi-parallel corpus is required, which means a parallel corpus between at least three languages, where each sentence has translations in the two other languages. The current project aims to align French (FR) and English (EN) words with Spanish (ES) as a pivot language.

In the training of alignment models, part of the European Parliament parallel corpus (Koehn, 2005) was used. A subset of this data is downloadable from the Shared Task of the ACL 2005 Workshop on Building and Using Parallel Texts [1], where gold word alignments between French and English are provided. We do not know how the gold alignments are obtained, except that it is from an automatic tool[2].

The data originally consists of bilingual parallel corpora between EN-FR and EN-ES. A multi-parallel corpus is created by merging the EN-FR and EN-ES corpora by matching identical English sentences, and discarding those that do not exist in either of the bilingual corpora or are repeated sentences. In the experiments, 100k sentences from the multi-parallel corpus are used in training.

Similar to Assignment 5, the test set consists of the first 100 sentences of the training dataset, meaning the test set is not blind.

Before generating the models, the `preprocess` function is applied on the corpus to tokenize and lowercase each sentence.

2.2 Alignment Model training

2.2.1 Baseline model: IBM model 1

The baseline model is a direct alignment model of the direction EN → FR, based on IBM 1 and trained for 20 iterations. After the probability table is generated, the `align` function performs the decoding part as in IBM model 1.

2.2.2 Bridged alignment model: matrix multiplication of posterior probability

To obtain a bridged alignment model, two separate models of the directions EN → ES and ES → FR are trained. The parameters, including the sentences, are exactly the same as the baseline model except for different language pairs.

Given the two alignment models, a EN → FR model is obtained with a `pivot` function, which implements the equation in Section 1.1.

Additionally, the current project follows the treatment in Kumar et al. (2007) for $k = 0$, where the alignment is NULL as a result of the IBM 1 model, such that:

$$P(a_k^{GE} = i|k = 0) = \begin{cases} \epsilon & i = 0 \\ \frac{1-\epsilon}{I} & i \in \{1, 2, \ldots, I\} \end{cases}$$

We follow the original paper in setting $\epsilon = 0.5$.

As such, the probability table of EN → FR is re-generated for every sentence and only contains word candidates that are present in the current sentence. The same decoding function is applied as in the baseline model.

2.2.3 Symmetrization: the grow-diag-final heuristic

To incorporate symmetrization to the bridged alignment model, a reversed model in the direction of FR → EN and a bridged model EN → ES → FR (same model as Section 2.2.2) are considered as the two sets of alignment points. They are of opposite directions following the intention of the grow-diag-final method, and this method performed better than two sets of alignments with the same direction in preliminary experiments. For each sentence to be aligned, the procedure is the same as the bridged alignment model, until after the `align` function. The `grow_diag_final` function is applied to symmetrize between the two sets of alignments and outputs a new alignment. For the pseudo-code of the grow-diag-final heuristic see Appendix A.1 (from Koehn, 2016, p.45).

2.3 Evaluation

2.3.1 AER (Alignment Error Rate)

The quantitative evaluation procedure is the same as in Assignment 5, using the `score-alignments` script from http://github.com/xutaima/jhu-mt-hw/tree/master/hw2. The gold alignment files (*hansards.e, hansards.f, hansards.a*) are modified to be the first 100 lines of the multi-parallel corpus. Note that while gold word alignments are available from the Shared Task, the alignments only contain *Sure* links, and do not contain *Possible* links as in Assignment 5.

[1] http://www.statmt.org/wpt05/mt-shared-task/

[2] According to the website, it is likely to have been obtained from a higher IBM model / Giza++.

2.3.2 Extracted phrases

The `phrase-extract` function follows the phrase extraction algorithm from Koehn (2009, p.133), which is also part of the standard SMT pipeline in Moses. The function extracts phrases for up to 7 tokens by default. While there is no gold phrases, we assume that the phrases extracted from the gold alignments to be the more optimal outcome. However, the main purpose of this evaluation procedure is to determine whether the bridged model can extract additional phrases compared to the baseline model, and this can be also done with human judgement.

3 Results

Table 1 reports the alignment performance in terms of Precision (Prec), Recall (Rec) and Alignment Error Rate (AER). All models are trained with the same size corpus and same number of iterations of IBM 1. The bridged alignment models did not manage to outperform the direct alignment model. This is not surprising, as the bridged models from Kumar et al. (2007) also did not outperform the model without any bridge languages.

model	Prec	Rec	AER
IBM 1	0.621	0.575	0.403
IBM 1 + sym	0.587	0.667	0.375
Bridged	0.510	0.472	0.510
Bridged + sym	0.554	0.646	0.403

Table 1: Alignment performance of various models. In both symmetrized models, the alignment in the reversed direction comes from a direct IBM 1 model in the FR → EN direction.

Symmetrization improves both the baseline and bridged model. We observe that the bridged model benefits more significantly than the baseline model, with a drop of 0.1 compared to 0.03 in AER. Symmetrization improves both Precision and Recall in the bridged language, while reducing Precision for the baseline model. However, such a significant improvement for the bridged model may also be attributed to the additional information given by the direct model in the reversed direction. This may suggest that the noise introduced by a pivot language can partly be alleviated by combining it with a non-bridged model. The AER suggests that a bridged, symmetrized model performs equally well as the baseline model.

Next, we present an analysis of the phrases extracted from the alignment models. Table 2 shows the number of phrases extracted from each model. We make a comparison with the phrases extracted from the gold alignments ("gold phrases"). If we assume that the intersection with "gold phrases" reflects the alignment quality, it is clear that phrase extraction does not agree with the evaluation of the AER. While symmetrized models improved alignments according to AER, they extract significantly fewer phrases. On the other hand, we may compare the number of phrases extracted and the phrases intersecting with the gold phrases, and find that symmetrized models have a much higher proportion of 'correct' phrases out of all extracted phrases; while their non-symmetrized counterparts extract a lot of phrases but only a minority overlaps with the gold phrases.

model	# of phrases extracted	# of phrases in gold alignments
IBM 1	1852	500
IBM 1 + sym	460	338
Bridged	2049	329
Bridged + sym	394	285
gold alignments	3305	-

Table 2: Set of phrases extracted from the corpus by the models. The phrases from gold alignments do not necessarily mean that they are the optimal set of phrases to be extracted.

model	# of unique phrases	# of unique phrases in gold alignments
IBM 1	995	224
Bridged	1234	109
Both	948	304
Total	3177	657

Table 3: Set of unique phrases extracted from the baseline models versus the bridged models. For a clear comparison between the baseline vs bridged model, the symmetrized models are merged into the respective non-symmetrized model.

Table 3 shows the number of unique phrases extracted by the models, and is intended to show the contribution of various models towards the final set of extracted phrases. For the first two rows of Table 3, unique means that the phrase is only extracted by either of the models, but not both. The third row shows the phrases extracted by both baseline and bridged models. Since the symmetrized models extracted significantly fewer phrases, they are merged to the respective non-symmetrized models for a clearer comparison between the baseline model and the bridged model.

As we hypothesised, bridged models do contribute towards the diversity of phrases extracted in addition to the baseline model. If one considers the "gold phrases" as gold phrases, column 2 shows that the bridged models can contribute around 50% in addition to what the baseline models extracted. If one solely considers the diversity of phrases, column 1 shows that the bridged model contributes even more phrases than the baseline model.

4 Discussion

In this paper, a pivoted alignment model is implemented where word alignments probabilities are obtained via a third "pivot" language. This is extended in two ways: 1. to improve word alignments with symmetrization and 2. to evaluate word alignments diversity by extracting phrases.

Consistent with previous research (Kumar et al., 2007), it was confirmed that pivoted word alignments do not outperform baseline IBM models 1. However, we found that a pivoted word alignment model can benefit a SMT system in terms of adding diversity to the set of phrase hypotheses. Further, pivoted word alignments benefit significantly from symmetrization.

Related research have suggested that introducing a pivot language during word alignments may introduce noise in constructing the model, thus resulting in more alignment errors (Zadeh, 2009). Other approaches in pivoting word alignments have also emphasised on methods in reducing such noise. For example, Wang et al. (2006) considered cross-language word similarity in order to disambiguate homonyms, which prevents words from being translated into a wrong context in the pivot language. Alternatively, Levinboim and Chiang (2015) introduced priors to jointly train multiple pivoted alignment models. These studies commonly attempt to solve the translation problem between low resource languages with a high resource language as pivot, which is different from the current project, as our pivoted alignment models require multi-parallel corpora, which is usually only available for high resource languages.

By analysing phrase extraction from alignment models, the current project demonstrates that bridged alignment models do add diversity to phrase hypotheses, which has been shown to benefit translation quality in phrase-based SMT, even though without directly improving alignment quality (Kumar et al., 2007). Further, our analysis sheds light on the limitations of evaluating word alignments solely via Alignment Error Rate (AER). While our symmetrized models perform better in terms of AER, they performed less well when compared to non-symmetrized models in terms of extracting a diverse number of phrases.

However, one main limitation of the current project is the lack of a comprehensive analysis of phrase quality. Given that our eventual goal is to improve phrase-based SMT systems, high quality phrases are needed to improve translation quality. A diverse set of phrases does not necessarily mean that they are of high quality. Moreover, in the SMT pipeline, phrase quality does not only depend on diversity but also on the frequency of these phrases being extracted, which would result in phrase probabilities.

The most obvious way to demonstrate alignment or phrase quality is by completing the SMT pipeline with translation quality measurements such as BLEU scores. Furthermore, the current project could benefit from experimenting with various language pairs and pivot languages. In particular, research disagree regarding whether the choice of a pivot language has an impact on alignment outcomes, and how to best choose an appropriate pivot language (Kumar et al., 2007; Wang et al., 2006). Such research would be beneficial to wider research related to the use of pivot languages in NLP.

References

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn. 2016. Lecture slides: Advanced Alignment Models. http://mt-class.org/jhu/slides/lecture-advanced-alignment-models.pdf.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In

Proceedings of the Second International Workshop on Spoken Language Translation.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Shankar Kumar, Franz Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. *Proceedings of the 2007 Joint EMNLP-CoNLL Conference*, pages 42–50.

Tomer Levinboim and David Chiang. 2015. Multi-task word alignment triangulation for low-resource languages. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1221–1226.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Haifeng Wang, Hua Wu, and Zhanyi Liu. 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 874–881.

Reza Bosagh Zadeh. 2009. Building strong multilingual aligned corpora. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Appendix

A.1 The grow-diag-final heuristic

Algorithm 1 The grow-diag-final heuristic

```
Require: e2f, f2e
  procedure GROW-DIAG()
      while new points added do
          for all English word $e$, French word $f$
           do
              for all neighbours$(e_{new}, f_{new})$ do
                  if $(e_{new}$ unaligned OR
                  $f_{new}$ unaligned$)$ AND
                  $(e_{new}, f_{new}) \in$ union(e2f, f2e)
                   then
                      add$(e_{new}, f_{new})$ to $A$
                  end if
              end for
          end for
      end while
  end procedure

  procedure FINAL()
      for all English word $e$, French word $f$ do
          if $(e$ unaligned OR $f$ unaligned $)$
          AND $(e, f) \in$ union(e2f, f2e)
           then
              add$(e, f)$ to $A$
          end if
      end for
  end procedure
```