Learning Core Word Alignments for Statistical Machine Translation

Frances Yung **Kevin Duh** **Yuji Matsumoto**
Nara Institute of Science and Technology
{pikyufrances-y,kevinduh,matsu}@is.naist.jp

1 Introduction

In this paper, we will discuss our linguistic intuition to treat inserted and deleted words separately from the word alignment model. *Core* word alignments can serve as better constraints for the construction of translation rules. The claims of our hypothesis are illustrated by a preliminary experiment.

2 Proposal

2.1 Background

A statistical machine translation system basically consists of a language model and a translation model, where the former generates output text in the target language and the later deduces the translation from the source text to the target text. Typically, the translation model is trained on sentence-aligned bilingual texts and the model outputs a series of translation rules coupled with probabilities. Different levels of linguistic knowledge can be extracted from the bilingual texts and incorporated in the translation rules, yet in all cases word-level correspondences in the source and target are critical. Word alignment is thus the basis of higher translation models.

For almost 20 years, automatic word alignment has been dominantly performed by the IBM Models, which infer word-to-word correspondences by the EM algorithm. A 'NULL' word token is added to each source and target sentence such that unaligned words are represented by alignments with the 'NULL' token. To capture many-to-one alignments, a fertility model is added since IBM Model 3 to explicitly model how many source words are to be aligned to each target word. The IBM Models are implemented by GIZA++[7] , a freely available SMT toolkit that further symmetrizes and grows the many-to-one alignments bi-directionally, yielding many-to-many word alignments. GIZA++ is widely used by many SMT systems at the earliest word alignment step.

Unsupervised word alignment by the IBM Models practically has the shortcoming, among others, of biasing to function words in the alignments. Although function words usually do not have correspondence in the other language, they are often included in automatic many-to-many word alignments as a result of their high frequencies, and thus reducing the chance for correct content-to-content word alignments. According to this observation, the NULL alignments in the IBM Models may not be strong enough to capture word insertion and deletion in the translation process.

2.2 Word Additions in Translation

In the area of translation studies, it is generally accepted that absolute equivalence between two languages is impossible and thus exact word alignments never exist. The word alignment task can thus be seen as linking the 'closest corresponding' words in the bilingual sentences, which can include one-to-one or many-to-many alignments. Nonetheless, human translation is not solely based on aligning source words to target words. Word addition and omission are deliberate techniques applied by translators to achieve *functional equivalence*[6] . For example, the Chinese term *'tai yang'* literally means *'sun'*, yet functionally its equivalence should be *'the sun'* in English. However, there is not any concepts of definite articles in Chinese, nor is it incorporated in the morphology of *'tai yang'*, so additional meaning is conveyed by the term *'the sun'*. Metaphorically speaking, translations of *'tai yang'* to *'sun'* and *'the sun'* are like 1-to-0.8 and 1-to-1.2 alignments respectively, realized as 1-to-1 and 1-to-2 word alignments. The function word *'the'* is added to satisfy the grammatical need of English for a determiner and the semantics of *'tai yang'* provides knowledge to choose 'the' as the proper determiner, whereas the core alignment is between *'tai yang'* and *'sun'*.

Function words added in the translation, as illustrated in the above example, are mostly inferred locally. Addition techniques in translation yet do not limit to local grammatical gaps, but can be inferred more globally, such as from discourse and pragmatic contexts. Complementary words are also added due to the difference in cultural backgrounds of the source and target readers, or for rhetorical purpose. In other cases, grammatical words have to be added when the source sentence is restructured due to the difference in the nature of the two languages. In general, additions inferred from a more global background are more optional, and the choice varies among translators. On the other hand, function words that are not required in target language are omitted in the translation. Content words can be omitted as well when they are redundant, as judged optionally by the translator.

Word addition is extensively applied in translation, especially between distant languages. The translated text is thus the result of two distinctive processes: source-to-target word rendering and word addition (and omission). From another point of view, when a word is translated to two target words, contribution of the two words to the link is likely to be uneven – one word is the core translation, while the other is an attachment inferred locally or globally – even though the two target words may always occur together with the source word statistically.

2.3 'Unaligning' Additional Words

Basing on the translation theory on word addition and omission, we propose to treat inserted words separately in the translation model for SMT. In this work, we apply the principle to word alignment. As the basis for extracting higher level translation rules, word alignment serves to couple corresponding units in the translation sentences. We argue that only the core translation should be aligned since the inserted words in a multiword alignment are not inferred directly from the source word. The linkage, if exists, should be acquired above lexical level. Over-aligned word correspondences may lead to undesirable constraints to the translation rules.

In a many-to-many alignment inferred by the IBM Models, each word contributes equally as long as they co-occur often enough with the source word. Treating an uneven alignment as an even one does little harm if the alignment is supported by many counts. This is the case for a grammatical word addition inferred locally, as in the above *'sun'* example. Translation rules built on such alignments can be applied independently of context. Nonetheless, if the core word is of low frequency, the inserted words may be misleadingly inferred as essential for the translation. In cases where the inserted word is a content word inferred optionally from a larger context, the algorithm may wrongly align only the inserted words. It would be favorable if likely additional words are removed before running the alignment algorithm.

We hypothesize that removing inserted words from the training data can improve words alignments by the IBM Models. After automatic alignment of the remaining core words, the inserted words are to be restored to the training as null alignments. The phrase-based translation model extract translation phrases according to word alignments and phrases including and excluding the unaligned words are both extracted. In this paradigm, the additional word acts as a soft constraint rather than a hard requirement for the many-to-many alignment.

Algorithm to identify inserted words in translation is not within the scope of this work. We evaluated our hypothesis basing on manually annotated word addition. An *inserted word* is defined, based on the translation theory of word addition, as target words that do not have directly corresponding source words. For example, *'the'* is an addition in the translation *'tai yang-the sun'*, but not in the translation *'na tai yang-the sun'*, where *'na'* is a demonstrative determiner in Chinese. While *'friends'* can be translated to the Chinese phrases *'peng you'* or *'peng you men'*, *'men'*, as an optional plural marker, is not an addition since the source word is marked plural. On the other hand, a *core* word is defined as a translated word with a corresponding word in the source, even there is little overlap between their semantics. As in the above examples, *'tai yang-sun'*, *'na-the'*, *'peng you-friends'*, and *'peng you men-friends'* are all links between core words. These include words that are *substituted* due to global context requirement. As long as there is lexical correspondence in the source, the link can be incorporated into the translation model framework.

2.4 Related studies

Past studies also pointed out that links between words are not always straightforward. [1] and [7] suggest that evaluation of automatic word alignments by merging their correspondence with manual *sure* and *possible* links serves as better reference for MT performance. [5] and [2] apply insertion and deletion models as features to extract phrases in phrase-based and hierarchical translation models respectively. We would like to apply this at the word alignment stage so as to avoid wrong alignments between function words and content words and to exclude contextually inferred additional words.

3 Experiment

The purpose of our experiment is to evaluate two claims of our hypothesis. First, we aim to prove that removing additional words in translation is beneficial for the IBM Models to align words. Secondly, we would like to illustrate that restoring additional words after word alignments can improve SMT results.

3.1 Data and Settings

As mentioned, our experiment is based on manually annotated data. We made use of the lately released GALE Chinese-English Word Alignment and Tagging Training Data Part 1-3, a corpus of about 12000 Chinese sentences and their faithful English translation. The source and target sentences are manually aligned to the token level, basing on the *minimum match principle* and the *attachment approach*[4] .

The word alignments are further enriched with annotation of the type of alignment, such as *'Semantic Link'* and *'Functional Link'*. On top of that, to our interest, additional words without direct corresponding source words are annotated according to their functions in the alignment, such as *'Measure Word'*, and *'Context Obligatory Marker'*. Words without a specific tag serve as the core translation in the alignment.

We made use of the 90% of the corpus as training data. This consists of 11,973 pairs of sentences[1], half collected from *web blog* and half from *news wire* articles. It contains 420,777 characters[2] of Mandarin Chinese, translated to 350,753 English words. The translation was organized for MT research purpose, based on a principle to reserve the structure and semantics of the source as much as possible. There are 285,205 alignments in the training set, where each can be expanded to 2.21 one-to-one links on average. 46,510 Chinese characters and 85,503 English words are marked as *inserted words*.

We performed our experiment using the open source statistical machine translation system MOSES [3], in which GIZA++ is incorporated. Results trained on two versions of the corpus are compared - the original corpus and one with all *inserted words* removed. GIZA++ learns 1-to-1 and many-to-1 word alignments in the source-to-target and target-to-source directions respectively and symmetrizes both outputs. We used the standard option of *align-grow-diag-and* for symmetrization. These 3 sets of alignment output were compared with the manual *'gold'* alignments occurring in the corresponding training data. For comparison, many-to-many alignments were expanded to 1-to-1 links between each bilingual word pairs in the alignments. We used *precision*, *recall*, and *F-measure* to evaluate the accuracies.

We continued to uses these word alignments to train translation models using MOSES. Settings for a baseline phrase-based model were used. Good Turing discounting was applied to smooth the phrase translation probabilities. The reordering model was set as *'hier-mslr-bidirectional-fe-allff'*. Since the data set was small, we did not sacrifice a portion for MERT tuning.

We compared the performance of translation model trained by 1) the original corpus; 2) the corpus with inserted words removed; 3) the corpus with inserted words removed but restored, after word alignment and before phrase extraction, as unaligned words[3]. Besides, we compared phrase translation models built from 4) manual many-to-many alignments; and 5) manual many-to-many alignments with *inserted words* modified as unaligned words. We also compare the results using either the *news wire* or *web blog* half of the corpus. In all cases, the language models are trained from the original corpus of the same genre. The MT results are evaluated by BLEU [8], using 10% held out data from the original corpus of the same genre as test data.

3.2 Results and Discussion

	Original corpus No. of gold alignments =631,510		
	Precision	Recall	F-measure
Chi-to-Eng	0.555	0.488	0.561
Eng-to-Chi	0.374	0.322	0.346
Symmetrized	0.548	0.477	0.510
	No. of output= 549,283 No. of true positives =300,981		

	Corpus without inserted words No. of gold alignments =554,568		
	Precision	Recall	F-measure
Chi-to-Eng	**0.663**	**0.529**	**0.589**
Eng-to-Chi	**0.545**	**0.435**	**0.484**
Symmetrized	**0.715**	**0.561**	**0.629**
	No. of output = 435,613 No. of true positives =**311,295**		

Table 1: Alignment accuracies of GIZA++ outputs

Table 1 lists the accuracies of the alignments output by GIZA++. In all three sets of alignments, the alignments trained by the corpus without *inserted words* are more accurate. The English-to-Chinese alignments trained by the original corpus are of particularly low accuracy, but they are much improved when trained by the corpus without inserted words. This leads to a significant rise in the accuracy of the symmetrized alignments. The absolute count of true positives is also higher.

Removing *inserted words* means shortening each pair of aligned sentences while reducing the number of target alignments. Higher accuracy rates are expected. Yet the increase in the absolute number of correct links supports our hypothesis that removing additional words in translation improves word alignment by the IBM Models.

Figure 1 shows the BLEU scores of the MT systems built on various word alignments. Generally, the results are not comparable to state-of-the-art performance, since our data set is very small.

[1] Alignment was rejected by the annotators for a small number of sentences. These are excluded from the training and test data.

[2] To cope with the minimum match principle of word alignment, the Chinese side of the corpus is tokenized to character level. We did not modify it and train the translation models and language models at character level.

[3] The results presented here are based on restoration after on the symmetrized alignments, which are slightly better than those based on restoration before symmetrization

The graph shows that the MT performance greatly dropped when *inserted words* are bluntly removed from the training data. Although more correct core alignments are learnt, all the removed words become unknown in the translation model. Restoring them after word alignment allows the phrase translation table to include them and an increase of about 0.3 BLEU point over the baseline is seen.

Similar trends are found in the *news wore* and *web blog* genres, as well as the data as a whole. Comparing the results obtained by manual alignments, we find that MT is significantly improved by removing the *inserted words* from the alignments and restoring them as unaligned words. In fact, it is an unexpected observation that manual alignments do not always outperform automatic alignments[4]. A possible explanation is that the manual alignments are linguistically oriented, including many implicit links between word elements scattered in different parts of the sentences. These subtle links become noise in the translation model training and unaligning them prevent errors. These findings support our claim that core word alignments are better basis for translation rules, while *inserted words* should be acquired flexibly after word alignments.

4 Conclusion

In this work, we proposed to model directly translated words and inserted words differentially for a statistical machine translation system. Preliminary experiments show that removing inserted words improves word alignment by the IBM Models. Training the translation model after restoring the removed words to the training data improves the overall machine translation performance, even using a small set of manually annotated data. In fact the *word addition* phenomenon is suppressed in our data set strictly translated to preserve the source text structure. Our future direction is to develop a method to identify likely additional words in the translated text, so as to perform translation model training on huge *naturally occurring* translation data and to compare with stronger baselines. We plan to use syntactical information to deduce unalignable grammatical words and translation lexicons to deduce globally inferred word additions. We also plan to evaluate the MT results basing on hierarchical translation models, which should be more sensitive to word alignments than phrase-based models.

References

[1] Alexander Fraser and Daniel Marcu. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, Vol. 33, No. 3, pp. 293–303, September 2007.

[2] Matthias Huck and Hermann Ney. Insertion and deletion models for statistical machine trans. *Proceedings of NAACL*, pp. 347–351, 2012.

[3] Phillip Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. Proceedings of HLT-NAACL, pp. 127–133, May-June 2003.

[4] Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie M. Strassel, and Kazuaki Maeda. Enriching word alignment with linguistic tags. *Proceedings of International Conference on Language Resources and Evaluation*, May 2010.

[5] Arne Mauser, Richard Zens, Evgeny Matusov, Sasa Hasan, and Hermann Ney. The rwth statistical machine translation system for the iwslt 2006 evaluation. *Proceeding of IWSLT*, Vol. 1, pp. 103–110, November 2006.

[6] Eugene A Nida. *Toward a Science of Translating: with Special Reference to Principles and Procedures Involved in Bible Translating.* BRILL, 1964.

[7] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, March 2003.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *Proceedings of ACL*, pp. 311–318, July 2002.

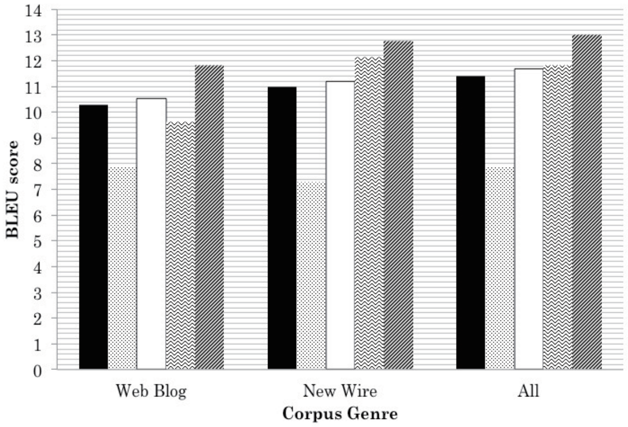

Figure 1: BLEU scores of MT results from various word alignments. From left to right: 1)GIZA++ output trained on original corpus; 2)GIZA++ output trained on corpus with inserted words removed; 3) GIZA++ output trained on corpus with inserted words removed and restored; 4) Manual many-to-many word alignments; 5) Manual many-to-many word alignments with inserted words unaligned

[4] According to the experiment result, automatic alignments are better for *web blog* texts, but manual alignments are better for *news wire* texts.