Analysis and Prediction of Unalignable Words in Parallel Text

Frances Yung    Kevin Duh    Yuji Matsumoto
Nara Institute of Science and Technology 8916-5 Takayama, Ikoma, Nara, 630-0192 Japan
pikyufrances-y|kevinduh|matsu@is.naist.jp

Abstract
Professional human translators usually do not employ the concept of word alignments, producing translations ‘sense-for-sense’ instead of ‘word-for-word’. This suggests that unalignable words may be prevalent in the parallel text used for machine translation (MT). We analyze this phenomenon in-depth for Chinese-English translation. We further propose a simple and effective method to improve automatic word alignment by pre-removing unalignable words, and show improvements on hierarchical MT systems in both translation directions.

1 Motivation
It is generally acknowledged that absolute equivalence between two languages is impossible, since concept lexicalization varies across languages. Major translation theories thus argue that texts should be translated ‘sense-for-sense’ instead of ‘word-for-word’ (Nida, 1964). This suggests that unalignable words may be an issue for the parallel text used to train current statistical machine translation (SMT) systems. Although existing automatic word alignment methods have some mechanism to handle the lack of exact word-for-word alignment (e.g. null probabilities, fertility in the IBM models (Brown et al., 1993)), they may be too coarse-grained to model the ‘sense-for-sense’ translations created by professional human translators.

For example, the Chinese term ‘tai-yang’ literally means ‘sun’, yet the concept it represents is equivalent to the English term ‘the sun’. Since the concept of a definite article is not incorporated in the morphology of ‘tai yang’, the added ‘the’ is not aligned to any Chinese word. Yet in another context like ‘the man’, ‘the’ can be the translation of the Chinese demonstrative pronoun ‘na’, literally means ‘that’. A potential misunderstanding is that unalignable words are simply function words; but from the above example, we see that whether a word is alignable depends very much on the concept and the linguistic context.

As the quantity and quality of professionally-created parallel text increase, we believe there is a need to examine the question of unalignable words in-depth. Our goal is to gain a better understanding of what makes a fluent human translation and use this insight to build better word aligners and MT systems. Our contributions are two-fold:

1) We analyze 13000 sentences of manually word-aligned Chinese-English parallel text, quantifying the characteristics of unalignable words.

2) We propose a simple and effective way to improve automatic word alignment, based on predicting unalignable words and temporarily removing them during the alignment training procedure.

2 Analysis of Unalignable Words
Our manually-aligned data, which we call ORACLE data, is a Chinese-to-English corpus released by the LDC (Li et al., 2010)1. It consists of ~13000 Chinese sentences from news and blog domains and their English translation. English words are manually aligned with the Chinese characters. Characters without an exact counterpart are annotated with categories that state the functions of the words. These characters are either aligned to ‘NULL’, or attached to their dependency heads, if any, and aligned together to form a multi-word alignment. For example, ‘the’ is annotated as [DET], for ‘determiner’, and aligned to ‘tai-yang’ together with ‘sun’.

In this work, any English word or Chinese character without an exact counterpart are called unalignable words, since they are not core to the
multi-word alignment. All other English words or Chinese characters are referred to as core words.

2.1 What kind of words are unalignable?

Analyzing the hand aligned corpus, we find that words annotated as unalignable do not come from a distinct list. Table 1 reveals that 88% of the word types are unambiguously core words. Yet these word types, including singletons, account for only 17% of the word tokens. On the other hand, another 17% of the total word tokens are annotated as unalignable. So, most word types are possibly unalignable but only in a small portion of their occurrence, such as the following examples:

(1a) Chi: yi ge di fang
   Eng: one (measure word) place

(1b) Chi: ge ren
   Eng: personal

(2a) Chi: ming tian zhong wu
   Eng: tomorrow (midday)

(2b) Chi: zai jia
   Eng: at/in/on home

In example (1a), ‘ge’ is a measure word that is exclusive in Chinese, but in (1b), it is part of the multiword unit ‘ge-ren’ for ‘personal’. Similarly, prepositions, such as ‘at’, can either be omitted or translated depending on context.

Nonetheless, unalignable words are by no means evenly distributed among word types. Table 2 shows that the top 100 most frequent unalignable word types already covers 78% and 94% of all Chinese and English unalignable instances, respectively. Word type is thus an important clue.

Intuitively, words with POS defined only in one of the languages are likely to be unalignable. To examine this, we automatically tagged the ORACLE data using the Standford Tagger (Toutanova et al., 2003). We find that the unalignable words include all POS categories of either language, though indeed some POS are more frequent. Table 3 lists the top 5 POS categories that most unalignable words belong to and the percentage they are annotated as unalignable. Some POS categories like DEG are mostly unalignable regardless of context, but other POS tags such as DT and IN depend on context.

Note also that many Chinese unalignable words are nouns (NN) and verbs (VV). Clearly we cannot indiscriminately consider all nouns as unalignable. Some examples of unalignable content words in Chinese are:

(3) Chi: can jia hui jian huo dong
   Eng: participate in the meeting

(4) Chi: hui yi de yuan man ju xing
   Eng: success of the meeting

English verbs and adjectives are often nominalized to abstract nouns (such as ‘meeting’ from ‘meet’, or ‘success’ from ‘succeed’), but such derivation is rare in Chinese morphology. Since POS is not morphologically marked in Chinese, ‘meeting’ and ‘meet’ are the same word. To reduce the processing ambiguity and produce more natural translation, extra content words are added to mark the nominalization of abstract concepts. For example, ‘hui jian’ is originally ‘to meet’. Adding ‘huo dong’ (activity) transforms it to a noun phrase...
(example 3), similar to the the addition of ‘ju
sing’(take place) to the adjective ‘yuan man’(ex-
ample 4). These unalignable words are not lex-
ically dependent but are inferred from the context,
and thus do not align to any source words.

To summarize, a small number of word types
cover 17% of word tokens that are unalignable,
but whether these words are unalignable depends
significantly on context. Although there is no list
of ‘always unalignable’ words types or POS cat-
egories, our analysis shows there are regularities
that may be exploited by an automatic classifier.

3 Improved Automatic Word Alignment

We first propose a classifier for predicting whether
a word is unalignable. Let \((e_i^f, f_J^K)\) be a pair of
sentence with length \(J\) and \(K\). For each word in
\((e_i^f, f_J^K)\) that belongs to a predefined list\(^2\) of
potentially unalignable words, we run a binary clas-
sifier. A separate classifier is built for each word
type in the list, and an additional classifier for all
the remaining words in each language.

We train an SVM classifier based on the fol-
lowing features: **Local context:** Unigrams and
POS in window sizes of 1, 3, 5, 7 around the
word in question. **Top token-POS pairs:** This
feature is defined by whether the token in ques-
tion and its POS tag is within the top \(n\) frequent
token-POS pairs annotated as unalignable like in
Tables 2 and 3. Four features are defined with \(n =
10, 30, 50, 100\). Since the top frequent unalignable
words cover most of the counts as shown in the
previous analysis, being in the top \(n\) list is a strong
positive features. **Number of likely unalignable
words per sentence:** We hypothesize that the
translator will not add too many tokens to the
translation and delete too many from the source
sentence. In the ORACLE data, 68% sentences
have more than 2 unalignable words. We approx-
imate the number of likely unalignable words in
the sentence by counting the number of words
within the top 100 token-POS pairs annotated as
unalignable. **Sentence length and ratio:** Longer
sentences are more likely to contain unalignable
words than shorter sentences. Also sentence ra-
tios that deviate significantly from the mean are
likely to contain unalignable words. **Presence of
alignment candidate:** This is a negative feature
defined by whether there is an alignment candi-
deate in the target sentence for the source word in
question, or vice versa. The candidates are ex-
tracted from the top \(n\) frequent words aligned to
a particular word according to the manual align-
ments of the ORACLE data. Five features are de-
fined with \(n = 5, 10, 20, 50, 100\) and one ‘without
limit’, such that a more possible candidate will be
detected by more features.

Next, we propose a simple yet effective mod-
fication to the word alignment training pipeline:
1. Predict unalignable words by the classifier
2. Remove these words from the training corpus
3. Train word alignment model (e.g. GIZA++)\(^3\)
4. Combine the word alignments in both direc-
tions with heuristics (grow-diag-final-and)
5. Restore unaligned words to original position
6. Continue with rule extraction and the rest of
the MT pipeline.

The idea is to reduce the difficulty for the word
alignment model by removing unaligned words.

4 End-to-End Translation Experiments

In our experiments, we first show that removing
manually-annotated unaligned words in ORACLE
data leads to improvements in MT of both trans-
lation directions. Next, we show how a classifier
trained on ORACLE data can be used to improve
MT in another large-scale un-annotated dataset.\(^4\)

4.1 Experiments on ORACLE data

We first performed an ORACLE experiment using
gold standard unaligned word labels. Follow-
ing the training pipeline in Section 3, we removed
gold unalignable words before running GIZA++
and restore them afterwards. 90% of the data is
used for alignment and MT training, while 10% of
the data is reserved for testing.

The upper half of Table 4 list the alignment
precision, recall and F1 of the resulting align-
ments, and quality of the final MT outputs. **Base-
line** is the standard MT training pipeline with-
out removal of unaligned words. Our **Proposed
approach** performs better in alignment, phrase-
based (PBMT) and hierarchical (Hiero) systems.
The results, evaluated by BLEU, METEOR and
TER, support our hypothesis that removing gold
unaligned words helps improve word alignment
and the resulting SMT.

\(^2\)We can suppress the NULL probabilities of the model.
\(^3\)All experiments are done using standard settings for
Moses PBMT and Hiero with 4-gram LM and mslr-
bidirectional-fe reordering (Koehn et al., 2007). The clas-
sifier is trained using LIBSVM (Chang and Lin, 2011).
Table 4: MT results of ORACLE and REAL experiments. Highest score per metric is bolded. \{+ / −\} indicates statistically significant improvement/degradation, \(p < 0.05\). (P: precision; R: recall; B: BLEU; M: METEOR; T:TER)

For comparison, a naive classifier that labels all top-30 token-POS combinations as unalignable performs poorly as expected (PBMT BLEU: 9.87 in C-E direction). We also evaluated our proposed classifier on this task: the accuracy is 92% and it achieves BLEU of 11.55 for PBMT and 10.84 for Hiero in C-E direction, which is between the results of gold-unalign and baseline.

### 4.2 Experiments on large-scale REAL data

We next performed a more realistic experiment: the classifier trained on ORACLE data is used to automatically label a large data, which is then used to train a MT system. This REAL data consists of parallel text from the NIST OpenMT2008.\(^5\) MT experiments are performed in both directions.

The lower half of Table 4 shows the performance of the resulting MT systems. We observe that our proposed approach is still able to improve over the baseline. In particular, Hiero achieved statistical significant improvements in BLEU and METEOR.\(^6\) Comparing to the results of PBMT, this suggests our method may be most effective in improving systems where rule extraction is sensitive to the underlying alignments, such as Hiero and Syntax-based MT. Table 5 shows the lexical translations for some rare Chinese words: the baseline tends to incorrectly align these to function words (garbage collection), while the proposed method’s translations are more reasonable.

To evaluate how much annotation is needed for the classifier, we repeat experiments using different proportions of the ORACLE data. Figure 1 shows training by 20% of the data (2600 sents.) already leads to significant improvements \((p < 0.05)\), which is a reasonable annotation effort.

### 5 Conclusion

We analyzed in-depth the phenomenon of unalignable words in parallel text, and show that what is unalignable depends on the word’s concept and context. We argue that this is not a trivial problem, but with an unalignable word classifier and a simple modified MT training pipeline, we can achieve small but significant gains in end-to-end translation. In related work, the issue of dropped pronouns (Chung and Gildea, 2010) and function words (Setiawan et al., 2010; Nakazawa and Kurohashi, 2012) have been found important in word alignment, and (Fossum et al., 2008) showed that syntax features are helpful for fixing alignments. An interesting avenue of future work is to integrate these ideas with ours, in particular by exploiting syntax and viewing unalignable words as aligned at a structure above the lexical level.

---

\(^{5}\)We use the standard MT08 test sets; the training data includes LDC2004T08, 2005E47, 2005T06, 2007T23, 2008T06, 2008T08, 2008T18, 2009T02, 2009T06, 2009T15, and 2010T03 (34M English words and 1.1M sentences). Since we do not have access to all OpenMT data, e.g. FBIS, our results may not be directly comparable to other systems in the evaluation.

\(^{6}\)Interestingly, PBMT did better than Hiero in this setup.

Table 5: Examples of translations exclusively found in the top 15 lexical translation.
References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27).

Tagyoung Chung and Daniel Gildea. 2010. Effects of empty categories on machine translation. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Victoria Fossum, Kevin Knight, and Steven Abney. 2008. Using syntax to improve word alignment precision for syntax-based machine translation. *Proceedings of the Workshop on Statistical Machine Translation*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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

Toshiaki Nakazawa and Sado Kurohashi. 2012. Alignment by bilingual generation and monolingual derivation. *Proceedings of the International Conference on Computational Linguistics*.

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

Hendra Setiawan, Chris Dyer, and Philip Resnik. 2010. Discriminative word alignment with a function word reordering model. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.