How Much Can We Gain from Supervised Word Alignment?

Jinxi Xu and Jinying Chen
Raytheon BBN Technologies
10 Moulton Street, Cambridge, MA 02138, USA
{jxu,jchen}@bbn.com

Abstract

Word alignment is a central problem in statistical machine translation (SMT). In recent years, supervised alignment algorithms, which improve alignment accuracy by mimicking human alignment, have attracted a great deal of attention. The objective of this work is to explore the performance limit of supervised alignment under the current SMT paradigm. Our experiments used a manually aligned Chinese-English corpus with 280K words recently released by the Linguistic Data Consortium (LDC). We treated the human alignment as the oracle of supervised alignment. The result is surprising: the gain of human alignment over a state of the art unsupervised method (GIZA++) is less than 1 point in BLEU. Furthermore, we showed the benefit of improved alignment becomes smaller with more training data, implying the above limit also holds for large training conditions.

1 Introduction

Word alignment is a central problem in statistical machine translation (SMT). A recent trend in this area of research is to exploit supervised learning to improve alignment accuracy by mimicking human alignment. Studies in this line of work include Haghighi *et al.*, 2009; DeNero and Klein, 2010; Setiawan *et al.*, 2010, just to name a few.

The objective of this work is to explore the performance limit of supervised word alignment. More specifically, we would like to know what magnitude of gain in MT performance we can expect from supervised alignment over the state of the art unsupervised alignment if we have access to a large amount of parallel data. Since alignment errors have been assumed to be a major hindrance to good MT, an answer to such a question might help us find new directions in MT research.

Our method is to use human alignment as the oracle of supervised learning and compare its performance against that of GIZA++ (Och and Ney 2003), a state of the art unsupervised aligner. Our study was based on a manually aligned Chinese-English corpus (Li, 2009) with 280K word tokens. Such a study has been previously impossible due to the lack of a hand-aligned corpus of sufficient size.

To our surprise, the gain in MT performance using human alignment is very small, less than 1 point in BLEU. Furthermore, our diagnostic experiments indicate that the result is not an artifact of small training size since alignment errors are less harmful with more data.

We would like to stress that our result does not mean we should discontinue research in improving word alignment. Rather it shows that current translation models, of which the string-to-tree model (Shen *et al*., 2008) used in this work is an example, cannot fully utilize super-accurate word alignment. In order to significantly improve MT quality we need to improve both word alignment and the translation model. In fact, we found that some of the information in the LDC hand-aligned corpus that might be useful for resolving certain translation ambiguities (e.g. verb tense, pronoun co-references and modifier-head relations) is even harmful to the system used in this work.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 165–169,
Portland, Oregon, June 19-24, 2011.

2 Experimental Setup

2.1 Description of MT System

We used a state of the art hierarchical decoder in our experiments. The system exploits a string to tree translation model, as described by Shen *et al*. (2008). It uses a small set of linguistic and contextual features, such as word translation probabilities, rule translation probabilities, language model scores, and target side dependency scores, to rank translation hypotheses. In addition, it uses a large number of discriminatively tuned features, which were inspired by Chiang *et al.* (2009) and implemented in a way described in (Devlin 2009). Some of the features, e.g. context dependent word translation probabilities and discriminative word pairs, are motivated in part to discount bad translation rules caused by noisy word alignment. The system used a 3-gram language model (LM) for decoding and a 5-gram LM for rescoring. Both LMs were trained on about 9 billion words of English text.

We tuned the system on a set of 4,171 sentences and tested on a set of 4,060 sentences. Both sets were drawn from the Chinese newswire development data for the DARPA GALE program. On average, each sentence has around 1.7 reference translations for both sets. The tuning metric was BLEU, but we reported results in BLEU (Papineni *et al.*, 2002) and TER (Snover *et al.*, 2006).

2.2 Hand Aligned Corpus

The hand aligned corpus we used is LDC2010E63, which has around 280K words (English side). This corpus was annotated with alignment links between Chinese characters and English words. Since the MT system used in this work is word-based, we converted the character-based alignment to word-based alignment. We aligned Chinese word *s* to English word *t* if and only if *s* contains a character *c* that was aligned to *t* in the LDC annotation.

A unique feature of the LDC annotation is that it contains information beyond simple word correspondences. Some links, called special links in this work, provide contextual information to resolve ambiguities in tense, pronoun co-reference, modifier-head relation and so forth. The special links are similar to the so-called possible links described in other studies (Och and Ney, 2003; Fraser and Marcu, 2007), but are not identical. While such links are useful for making high level inferences, they cannot be effectively exploited by the translation model used in this work. Worse, they can hurt its performance by hampering rule extraction. Since the special links were marked with special tags to distinguish them from regular links, we can selectively remove them and check the impact on MT performance.

Figure 1 shows an example sentence with human alignment. Solid lines indicate regular word correspondences while dashed lines indicate special links. Tags inside [] indicate additional information about the function of the words connected by special links.

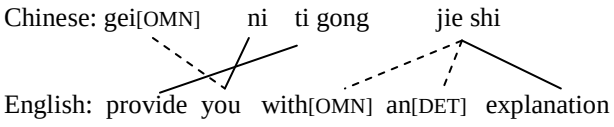

Figure 1: An example sentence pair with human alignment

2.3 Parallel Corpora and Alignment Schemes

Our experiments used two parallel training corpora, aligned by alternative schemes, from which translation rules were extracted.

The corpora are:

- Small: the 280K word hand-aligned corpus, with human alignment removed
- Large: a 31M word corpus of Chinese-English text, comprising a number of component corpora, one of which is the small corpus[1]

The alignment schemes are:

- giza-weak: Subdivide the large corpus into 110 chunks of equal size and run GIZA++ separately on each chunk. One of the chunks is the small corpus mentioned above. This produced low quality unsupervised alignment.

[1] Other data items included are LDC{2002E18,2002L27, 2005E83,2005T06,2005T10,2005T34,2006E24,2006E34, 2006E85,2006E92,2006G05,2007E06,2007E101,2007E46, 2007E87,2008E40,2009E16,2008E56}

- giza-strong: Run GIZA++ on the large corpus in one large chunk. Alignment for the small corpus was extracted for experiments involving the small corpus. This produced high quality unsupervised alignment.
- gold-original: human alignment, including special links
- gold-clean: human alignment, excluding special links

Needless to say, gold alignment schemes do not apply to the large corpus.

3 Results

3.1 Results on Small Corpus

The results are shown in Table 2. The special links in the human alignment hurt MT (Table 2, gold-original vs. gold-clean). In fact, with such links, human alignment is worse than unsupervised alignment (Table 2, gold-original vs. giza-strong). After removing such links, human alignment is better than unsupervised alignment, but the gain is small, 0.72 point in BLEU (Table 2, gold-clean vs. giza-strong). As expected, having access to more training data increases the quality of unsupervised alignment (Table 1) and as a result the MT performance (Table 2, giza-strong vs. giza-weak).

Alignment	Precision	Recall	F
gold-clean	1.00	1.00	1.00
giza-strong	0.81	0.72	0.76
giza-weak	0.65	0.58	0.61

Table 1: Precision, recall and F score of different alignment schemes. F score is the harmonic mean of precision and recall.

Alignment	BLEU	TER
giza-weak	18.73	70.50
giza-strong	21.94	66.70
gold-original	20.81	67.50
gold-clean	22.66	65.92

Table 2: MT results (lower case) on small corpus

It is interesting to note that from giza-weak to giza-strong, alignment accuracy improves by 15% and the BLEU score improves by 3.2 points. In comparison, from giza-strong to gold-clean, alignment accuracy improves by 24% but BLEU score only improves by 0.72 point. This anomaly can be partly explained by the inherent ambiguity of word alignment. For example, Melamed (1998) reported inter annotator agreement for human alignments in the 80% range. The LDC corpus used in this work has a higher agreement, about 90% (Li *et al*., 2010). That means much of the disagreement between giza-strong and gold alignments is probably due to arbitrariness in the gold alignment.

3.2 Results on Large Corpus

As discussed before, the gain using human alignment over GIZA++ is small on the small corpus. One may wonder whether the small magnitude of the improvement is an artifact of the small size of the training corpus.

To dispel the above concern, we ran diagnostic experiments on the large corpus to show that with more training data, the benefit from improved alignment is less critical. The results are shown in Table 3. On the large corpus, the difference between good and poor unsupervised alignments is 2.37 points in BLEU (Table 3, giza-strong vs. giza-weak). In contrast, the difference between the two schemes is larger on the small corpus, 3.21 points in BLEU (Table 2, giza-strong vs. giza-weak). Since the quality of alignment of each scheme does not change with corpus size, the results indicate that alignment errors are less harmful with more training data. We can therefore conclude the small magnitude of the gain using human alignment is not an artifact of small training.

Comparing giza-strong of Table 3 with giza-strong of Table 2, we can see the difference in MT performance is about 8 points in BLEU (20.94 vs. 30.21). This result is reasonable since the small corpus is two orders of magnitude smaller than the large corpus.

Alignment	BLEU	TER
giza-weak	27.84	59.38
giza-strong	30.21	56.62

Table 3: MT results (lower case) on large corpus

3.3 Discussions

Some studies on supervised alignment (e.g. Haghighi *et al*., 2009; DeNero and Klein, 2010) reported improvements greater than the limit we established using an oracle aligner. This seemingly inconsistency can be explained by a number of factors. First, we used more data (31M) to train GIZA++, which improved the quality of unsupervised alignment. Second, some of the features in the MT system used in this work, such as context dependent word translation probabilities and discriminatively trained penalties for certain word pairs, are designed to discount incorrect translation rules caused by alignment errors. Third, the large language model (trained with 9 billion words) in our experiments further alleviated the impact of incorrect translation rules. Fourth, the GALE test set has fewer reference translations than the NIST test sets typically used by other researchers (1.7 references for GALE, 4 references for NIST). It is well known that BLEU is very sensitive to the number of references used for scoring. Had we used a test set with more references, the improvement in BLEU score would probably be higher. An area for future work is to examine the impact of each factor on BLEU score. While these factors can affect the numerical value of our result, they do not affect our main conclusion: Improving word alignment alone will not produce a breakthrough in MT quality.

DeNero and Klein (2010) described a technique to exploit possible links, which are similar to special links in the LDC hand aligned data, to improve rule coverage. They extracted rules with and without possible links and used the union of the extracted rules in decoding. We applied the technique on the LDC hand aligned data but got no gain in MT performance.

Our work assumes that unsupervised aligners have access to a large amount of training data. For language pairs with limited training, unsupervised methods do not work well. In such cases, supervised methods can make a bigger difference.

4 Related Work

The study of the relation between alignment quality and MT performance can be traced as far as to Och and Ney, 2003. A more recent study in this area is Fraser and Marcu, 2007. Unlike our work, both studies did not report MT results using oracle alignment.

Recent work in supervised alignment include Haghighi *et al*., 2009; DeNero and Klein, 2010; Setiawan *et al*., 2010, just to name a few. Fossum *et al*. (2008) used a heuristic based method to delete problematic alignment links and improve MT.

Li (2009) described the annotation guideline of the hand aligned corpus (LDC2010E63) used in this work. This corpus is at least an order of magnitude larger than similar corpora. Without it this work would not be possible.

5 Conclusions

Our experiments showed that even with human alignment, further improvement in MT quality will be small with the current SMT paradigm. Our experiments also showed that certain alignment information suitable for making complex inferences can even hamper current SMT models. A future direction for SMT is to develop translation models that can effectively employ such information.

Acknowledgments

This work was supported by DARPA/IPTO Contract No. HR0011-06-C-0022 under the GALE program[2] (Approved for Public Release, Distribution Unlimited). The authors are grateful to Michael Kayser for suggestions to improve the presentation of this paper.

References

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL,* pages 218–226.

John DeNero and Dan Klein. 2010. Discriminative Modeling of Extraction Sets for Machine Translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1453–1463.

[2] The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

Jacob Devlin. 2009. *Lexical features for statistical machine translation*. Master's thesis, University of Maryland.

Victoria Fossum, Kevin Knight and Steven Abney. 2008. Using Syntax to Improve Word Alignment Precision for Syntax-Based Machine Translation, In *Proceedings of the third Workshop on Statistical MT, ACL*, pages 44-52.

Alexander Fraser and Daniel Marcu. 2007. Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics*. 33(3): 293-303.

Aria Haghighi, John Blitzer, John DeNero and Dan Klein. 2009. Better word alignments with supervised ITG models, In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 923-931.

Xuansong Li. 2009. *Guidelines for Chinese-English Word Alignment*, Version 4.0, April 16, 2009, http://ww.ldc.upenn.edu/Project/GALE.

Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie M. Strassel and Kazuaki Maeda. 2010. Enriching Word Alignment with Linguistic Tags. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta.

Dan Melamed. 1998. Manual annotation of translational equivalence: The blinker project. Technical Report 98-07, Institute for Research in Cognitive Science, Philadelphia.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Hendra Setiawan, Chris Dyer, and Philip Resnik. 2010. Discriminative Word Alignment with a Function Word Reordering Model. In *Proceedings of 2010 Conference on Empirical Methods in Natural Language Processing*, pages 534–544.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08: HLT*, pages 577–585.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223-231.