Combining Unsupervised and Supervised Alignments for MT: An Empirical Study

Jinxi Xu and **Antti-Veikko I. Rosti**
Raytheon BBN Technologies, 10 Moulton Street, Cambridge, MA 02138, USA
{jxu,arosti}@bbn.com

Abstract

Word alignment plays a central role in statistical MT (SMT) since almost all SMT systems extract translation rules from word aligned parallel training data. While most SMT systems use unsupervised algorithms (e.g. GIZA++) for training word alignment, supervised methods, which exploit a small amount of human-aligned data, have become increasingly popular recently. This work empirically studies the performance of these two classes of alignment algorithms and explores strategies to combine them to improve overall system performance. We used two unsupervised aligners, GIZA++ and HMM, and one supervised aligner, ITG, in this study. To avoid language and genre specific conclusions, we ran experiments on test sets consisting of two language pairs (Chinese-to-English and Arabic-to-English) and two genres (newswire and weblog). Results show that the two classes of algorithms achieve the same level of MT performance. Modest improvements were achieved by taking the union of the translation grammars extracted from different alignments. Significant improvements (around 1.0 in BLEU) were achieved by combining outputs of different systems trained with different alignments. The improvements are consistent across languages and genres.

1 Introduction

Word alignment plays a central role in training statistical machine translation (SMT) systems since almost all SMT systems extract translation rules from word aligned parallel training data. Until recently, most SMT systems used GIZA++ (Och and Ney, 2003), an unsupervised algorithm, for aligning parallel training data. In recent years, with the availability of human aligned training data, supervised methods (e.g. the ITG aligner (Haghighi et al., 2009)) have become increasingly popular.

The main objective of this work is to show the two classes (unsupervised and supervised) of algorithms are complementary and combining them will improve overall system performance. The use of human aligned training data allows supervised methods such as ITG to more accurately align frequent words, such as the alignments of Chinese particles (e.g. "bei", "de", etc) to their English equivalents (e.g. "is/are/was/..", "of", etc). On the other hand, supervised methods can be affected by suboptimal alignments in hand-aligned data. For example, the hand-aligned data used in our experiments contain some coarse-grained alignments (e.g. "lianhe guo" to "United Nations") although fine-grained alignments ("lian-he" to "United" and "guo" to "Nations") are usually more appropriate for SMT. Unsupervised methods are less likely to be affected by this problem. We used two well studied unsupervised aligners, GIZA++ (Och and Ney, 2003) and HMM (Liang et al., 2006) and one supervised aligner, ITG (Haghighi et al., 2009) as representatives in this work.

We explored two techniques to combine different alignment algorithms. One is to take the union of the translation rules extracted from alignments produced by different aligners. This is motivated by studies that showed that the coverage of translation rules is critical to SMT (DeNeefe et al., 2007). The

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 667–673,
MIT, Massachusetts, USA, 9-11 October 2010.

other method is to combine the outputs of different MT systems trained using different aligners. Assuming different systems make independent errors, system combination can generate a better translation than those of individual systems through voting (Rosti et al., 2007).

Our work differs from previous work in two ways. Past studies of combining alternative alignments focused on minimizing alignment errors, usually by merging alternative alignments for a sentence pair into a single alignment with the fewest number of incorrect alignment links (Ayan and Dorr, 2006). In contrast, our work is based on the assumption that perfect word alignment is impossible due to the intrinsic difficulty of the problem, and it is more effective to resolve translation ambiguities at later stages of the MT pipeline. A main focus of much previous work on word alignments is on theoretical aspects of the proposed algorithms. In contrast, the nature of this work is purely empirical. Our system was trained on a large amount of training data and evaluated on multiple languages (Chinese-to-English and Arabic-to-English) and multiple genres (newswire and weblog). Furthermore, we used a state of the art string-to-tree decoder (Shen et al., 2008) to establish the strongest possible baseline. In comparison, experiments in previous studies typically used one language pair and one genre (usually newswire), a reduced amount of training data and a phrase based decoder.

This paper is organized as follows. Section 2 describes the three alignment algorithms. Section 3 describes the two methods used to combine these aligners to improve MT. The experimental setup used to compare these methods is presented in Section 4. Section 5 shows the results including a discussion. Section 6 discusses related work. Section 7 concludes the paper.

2 Alignment Algorithms

We used three aligners in this work: GIZA++ (Och and Ney, 2003), jointly trained HMM (Liang et al., 2006), and ITG (Haghighi et al., 2009). GIZA++ is an unsupervised method based on models 1-5 of Brown et al. (1993). Given a sentence pair $e - f$, it seeks the alignment a that maximizes the probability $P(f, a|e)$. As in most previous studies using GIZA++, we ran GIZA++ in both directions, from e to f and from f to e, and symmetrized the bidirectional alignments into one, using a method similar to the grow-diagonal-final method described in Och and Ney (2003). We ran GIZA++ up to model 4.

The jointly trained HMM aligner, or HMM for short, is also unsupervised but it uses a small amount of hand-aligned data to tweak a few high level parameters. Low level parameters are estimated in an unsupervised manner like GIZA++.

The ITG aligner is a supervised method whose parameters are tuned to optimize alignment accuracy on hand-aligned data. It uses the inversion transduction grammar (ITG) (Wu, 1997) to narrow the space of possible alignments. Since the ITG aligner uses features extracted from HMM alignments, HMM was run as a prepossessing step in our experiments. Both the HMM and ITG aligners are publicly available[1].

3 Methods of Combining Alternative Alignments for MT

We explored two methods of combining alternative alignments for MT. One is to extract translation rules from the three alternative alignments and take the union of the three sets of rules as the single translation grammar. Procedurally, this is done by concatenating the alignment files before extracting translation rules. We call this method *unioned grammar*. This method greatly increases the coverage of the rules, as the unioned translation grammar has about 80% more rules than the ones extracted from the individual alignment in our experiments. As such, decoding is also slower.

The other is to use system combination to combine outputs of systems trained using different aligners. Due to differences in the alignment algorithms, these systems would produce different hypotheses with independent errors. Combining a diverse set of hypotheses could improve overall system performance. While system combination is a well-known technique, to our knowledge this work is the first to apply it to explicitly exploit complementary alignment algorithms on a large scale.

Since system combination is an established technique, here we only briefly discuss our system com-

[1] http://code.google.com/p/berkeleyaligner/

bination setup. The basic algorithm was described in Rosti et al. (2007). In this work, we use incremental hypothesis alignment with flexible matching (Rosti et al., 2009) to produce the confusion networks. 10-best lists from all systems are collected first. All 1-best hypotheses for each segment are used as confusion network skeletons, the remaining hypotheses are aligned to the confusion networks, and the resulting networks are connected in parallel into a joint lattice with skeleton specific prior probabilities estimated from the alignment statistics on the initial arcs. This lattice is expanded with an unpruned bigram language model and the system combination weights are tuned directly to maximize the BLEU score of the 1-best decoding outputs. Given the tuned system combination weights, a 300-best list is extracted from the lattice, the hypotheses are rescored using an unpruned 5-gram language model, and a second set of system combination weights is tuned to maximize the BLEU score of the 1-best hypothesis of the re-scored 300-best list. The same rescoring step is also applied to the outputs of individual systems.

4 Experiment Setup

To establish strong baselines, we used a string-to-tree SMT system (Shen et al., 2008), one of the top performing systems in the NIST 2009 MT evaluation, and trained it with very large amounts of parallel and language model data. The system used large sets of discriminatively tuned features (up to 55,000 on Arabic) inspired by the work of Chiang et al. (2009). To avoid drawing language, genre, and metric specific conclusions, we experimented with two language pairs, Arabic-English and Chinese-English, and two genres, newswire and weblog, and report both BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) scores. Systems were tuned to maximize BLEU on the tuning set using a procedure described in Devlin (2009).

The sizes of the parallel training corpora are 238M words (target side) for Arabic-English MT and 265M words for Chinese-English. While the majority of the data is publicly available from the Linguistic Data Consortium (LDC), some of the data is available under the DARPA GALE program. Due to the size of the parallel corpora, we divided them into five chunks and aligned them in parallel to save time. Due to its running complexity, we ran ITG only on sentences with 60 or fewer words. For longer sentences, we used HMM alignments instead, which were conveniently generated in the preprocessing step of ITG aligner. For language model training, we used about 9 billion words of English text, most of which are from English Gigaword corpus and GoogleNews. Each system used a 3-gram LM for decoding and a 5-gram LM for re-scoring. The same 5-gram LM was also used for re-scoring system combination results.

For each combination of language pair and genre, we used three development sets:

- `Tune`, which was used to tune parameters of individual MT systems. Each system was tuned ten iterations based on BLEU.
- `SysCombTune`, which was used to tune parameters of system combination. A subset of it was also used as validation for determining the best iteration in tuning individual systems.
- `Test`, which was the blind test corpus for measuring performances of both individual systems and system combination.

Test materials were drawn from two sources: NIST MT evaluations 2004 to 2008, and development and evaluation data for the DARPA GALE program. Due to the mixing of different data sources, some test sentences have four reference translations while the rest have only one. The average number of references per test sentence varies across test sets. For this reason, MT scores are not comparable across test sets. Table 1 shows the size and the average number of references per sentence of the test sets.

Two hand-aligned corpora were used to train the ITG aligner: LDC2009E82 (Arabic-English) and LDC2009E83 (Chinese-English). We re-tokenized the corpora using our tokenizers and projected the LDC alignments to our tokenization heuristically. The projection was not perfect and sometimes created very coarse-grained alignments. We used a set of filters to remove such problematic data. We ended up with 3,667 Arabic-English and 879 Chinese-English hand-aligned sentence pairs with sufficient quality for training automatic aligners.

language and genre	Tune	SysCombTune	Test
Arabic newswire	2963 (2.9)	3223 (2.7)	2242 (2.7)
Arabic web	4597 (1.5)	4526 (1.4)	2703 (2.7)
Chinese newswire	3085 (2.6)	3001 (2.7)	2055 (1.4)
Chinese web	4221 (1.3)	4285 (1.3)	3092 (1.2)

Table 1: Numbers of sentences and average number of references (in parentheses) of test sets

5 Results

Three baseline systems were trained using the three different aligners. Case insensitive BLEU and TER scores for Arabic newswire, Arabic weblog, Chinese newswire, and Chinese weblog are shown in Tables 2, 3, 4, and 5, respectively[2]. The BLEU scores on the `Test` set are fairly similar but the ordering between different alignment algorithms is mixed between different languages and genres. To compare the two alignment combination strategies, we trained a system using the union of the rules extracted from the alternative alignments (`union` in the tables) and a combination of the three baseline system outputs (`3 syscomb` in the tables). The system with the unioned grammar was also added as an additional system in the combination marked by `4 syscomb`.

As seen in the tables, unioned grammar and system combination improve MT on both languages (Arabic and Chinese) and both genres (newswire and weblog). While there are improvements on both `SysCombTune` and `Test`, the results on `SysCombTune` are not totally fair since it was used for tuning system combination weights and as validation for optimizing weights of the MT systems. Therefore our discussion will focus on results on `Test`. (We did not show scores on `Tune` because systems were directly tuned on it.) Statistical significance is determined at 95% confidence level using the bootstrap method described in Koehn (2004), and is only applied on results obtained on the blind `Test` set.

For unioned grammar, the overall improvement in BLEU is modest, ranging from 0.1 to 0.6 point compared with the best baseline system, with little change in TER score. The improvements in BLEU score are statistically significant for Arabic (both genres), but not for Chinese. The improvements in TER are not significant for either language.

System combination produces bigger improvements in performance. Compared with the best baseline system, the improvement in BLEU ranges from 0.8 to 1.6 point. There are also noticeable improvements in TER, around 1.0 point. The TER improvements are mostly explained by the hypothesis alignment algorithm which is closely related to TER scoring (Rosti et al., 2009). The results are interesting because all three baseline systems (GIZA++, HMM and ITG) are identical except for the word alignments used in rule extraction. The results confirm that the aligners are indeed complementary, as we conjectured earlier. Also, the four-system combination yields consistent gains over the three-system combination, suggesting that the system using the unioned grammar is somewhat complementary to the three baseline systems. The statistical test indicates that both the three and four system combinations are significantly better than the single best alignment system for all languages and genres in BLEU and TER. In most cases, they are also significantly better than unioned grammar.

Somewhat surprisingly, the GIZA++ trained system is slightly better than the ITG trained system on all genres but Chinese weblog. However, we should point out that such a comparison is not entirely fair. First, we only ran ITG on short sentences. (For long sentences, we had to settle for HMM alignments for computing reasons.) Second, the hand-aligned data used for ITG training are not very clean, as we said before. The ITG results could be improved if these problems were not present.

[2]Dagger (†) indicates statistically better results than the best individual alignment system. Double dagger (‡) indicates statistically better results than both best individual alignment and unioned grammar. Bold indicates best Test set performance among individual alignment systems.

	SysCombTune		Test	
System	BLEU	TER	BLEU	TER
GIZA++	51.31	38.01	**50.96**	**38.38**
HMM	50.87	38.49	50.84	38.87
ITG	51.04	38.44	50.69	38.94
union	51.55	37.93	51.53†	38.32
3 syscomb	52.66	37.20	52.43‡	37.69‡
4 syscomb	52.80	37.05	52.55‡	37.46‡

Table 2: MT results on Arabic newswire (see footnote 2).

	SysCombTune		Test	
System	BLEU	TER	BLEU	TER
GIZA++	27.49	55.00	**38.00**	**49.55**
HMM	27.42	55.53	37.81	50.12
ITG	27.19	55.32	37.77	49.94
union	27.66	54.82	38.43†	49.43
3 syscomb	27.65	53.89	38.70†	48.72‡
4 syscomb	27.83	53.68	38.82‡	48.53‡

Table 3: MT results on Arabic weblog (see footnote 2).

	SysCombTune		Test	
System	BLEU	TER	BLEU	TER
GIZA++	36.42	54.21	**26.77**	57.67
HMM	36.12	54.50	26.17	58.22
ITG	36.23	54.11	26.53	**57.40**
union	36.57	54.07	26.83	57.37
3 syscomb	37.60	53.19	27.46‡	56.88‡
4 syscomb	37.77	53.11	27.57‡	56.57‡

Table 4: MT results on Chinese newswire (see footnote 2).

	SysCombTune		Test	
System	BLEU	TER	BLEU	TER
GIZA++	18.71	64.10	16.94	63.46
HMM	18.35	64.66	16.66	64.02
ITG	18.76	63.67	**16.97**	**63.29**
union	18.97	63.86	17.22	63.20
3 syscomb	19.66	63.40	17.98‡	62.47‡
4 syscomb	19.80	63.32	18.05‡	62.36‡

Table 5: MT results on Chinese weblog (see footnote 2).

5.1 Discussion

Inter-aligner agreements provide additional evidence about the differences between the aligners. Suppose on a common data set, the sets of alignment links produced by two aligners are A and B, we compute their agreement as $(|A \bigcap B|/|A| + |A \bigcap B|/|B|)/2$. (This is the average of recall and precision of one set by treating the other set as reference.) The agreement between GIZA++ and ITG is around 78% on a subset of the Arabic-English parallel data. The agreements between GIZA++ and HMM, and between HMM and ITG are slightly higher, around 83%. Since ITG could not align long sentences, we only used short sentences (at most 60 words in length) in our calculation.

Due to the large differences between the aligners, significantly more rules were extracted with the unioned grammar method in our experiments. On average, the size of the grammar (number of rules) was increased by about 80% compared with the baseline systems. The larger grammar results in more combinations of partial theories in decoding. However, for computing reasons, we kept the beam size of the decoder constant despite the increase in grammar size, potentially pruning out good theories. Performance could be improved further if larger beam sizes were used. We will leave this to future work.

6 Related Work

Ayan and Dorr (2006) described a method to minimize alignment errors by combining alternative alignments into a single alignment for each sentence pair. Deng and Zhou (2009) used the number of extractable translation pairs as the objective function for alignment combination. Och and Ney (2003) and Koehn et al. (2003) used heuristics to merge the bidirectional GIZA++ alignments into a single alignment. Despite differences in algorithms and objective functions in these studies, they all attempted to produce a single final alignment for each sentence pair. In comparison, all alternative alignments are directly used by the translation system in this work.

The unioned grammar method in this work is very similar to Giménez and Màrquez (2005), which combined phrase pairs extracted from different alignments into a single phrase table. The difference

from that work is that our focus is to leverage complementary alignment algorithms, while theirs was to leverage alignments of different lexical units produced by the same aligner.

Some studies leveraged other types of differences between systems to improve MT. For example, de Gispert et al. (2009) combined systems trained with different tokenizations.

The theory behind the GIZA++ aligner was due to Brown et al. (1993). The theory of Inversion Transduction Grammars (ITG) was due to Wu (1997). The ITG aligner (Haghighi et al., 2009) used in this work extended the original ITG to handle blocks of words in addition to single words. The use of HMM for word alignment can be traced as far back as to Vogel et al. (1996). The HMM aligner used in this work was due to Liang et al. (2006). It refined the original HMM alignment algorithm by jointly training two HMMs, one in each direction. Furthermore, it used a small amount of supervised data to tweak some high level parameters, although it did not directly use the supervised data in training.

7 Conclusions

We explored two methods to exploit complementary alignment algorithms. One is to extract translation rules from all alternative alignments. The other is to combine outputs of different MT systems trained using different aligners. Experiments on two language pairs and two genres show consistent improvements over the baseline systems.

Acknowledgments

This work was supported by DARPA/IPTO Contract No. HR0011-06-C-0022 under the GALE program[3] (Approved for Public Release, Distribution Unlimited). The authors are grateful to John DeNero and John Blitzer for their help with the Berkeley HMM and ITG aligners.

[3]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.

References

Necip Fazil Ayan and Bonnie J. Dorr. 2006. A maximum entropy approach to combining word alignments. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 96–103.

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):263–311.

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.

Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. 2009. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 73–76.

Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. 2007. What can syntax-based MT learn from phrase-based MT? In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 755–763.

Yonggang Deng and Bowen Zhou. 2009. Optimizing word alignment combination for phrase table training. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 229–232.

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

Jesús Giménez and Lluís Màrquez. 2005. Combining linguistic data views for phrase-based SMT. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 145–148.

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 IJCNLP of the AFNLP*, pages 923–931.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the ACL*, pages 48–54.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 104–111.

Franz J. 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.

Antti-Veikko I. Rosti, Bing Xiang, Spyros Matsoukas, Richard Schwartz, Necip Fazil Ayan, and Bonnie J. Dorr. 2007. Combining outputs from multiple machine translation systems. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the ACL*, pages 228–235.

Antti-Veikko I. Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2009. Incremental hypothesis alignment with flexible matching for building confusion networks: BBN system description for WMT09 system combination task. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 61–65.

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, Linnea Micciula, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Stephan Vogel, Hermann Ney, and Christoph Tillman. 1996. HMM-based word alignment in statistical translation. In *The 16th International Conference on Computational Linguistics*, pages 836–841.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3).