

An Empirical Study of Diversity of Word Alignment and its Symmetrization Techniques for System Combination

Thoudam Doren Singh[†]

Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
thoudam.doren@gmail.com

Abstract

The present work reports system combination task for the Chinese-English statistical machine translation systems. We focus on the strategy to build the candidate systems to enhance the gain of BLEU score by introducing diversity at the early stage of the system combination. One of the most effective strategies is to carry out system combination of the various systems with different word alignment algorithm. Our approach differs from previous work in one important aspect that we report on the diversity of the alignment refinement heuristics of word alignment techniques that are complementary to each other for the system combination. This approach could harness several word alignment possibilities and proved to be beneficial in generating consensus translation where the acting backbone which determines the word order is permitted to switch after each word. We carried out experiments on candidate systems of phrasal and hierarchical paradigms and system combination of both the paradigms as well. To our surprise, the combo systems using the various word alignments with various symmetrization techniques of both the MT paradigms show gain of 0.8 to 2.07 absolute BLEU score against the best candidates of the respective test sets.

1 Introduction

Machine translation outputs system combination based on confusion network decoding carried out has shown significant improvement in performance. The system combination task focused on core aspects of MT systems, i.e., decoding algorithm or alignment algorithm. The most prominent technique used in the system combination is consensus translation based on confusion network by rescoring using TER or METEOR. An

incremental alignment method to build confusion networks based on the TER algorithm (Rosti et al., 2008) yields significant improvement in BLEU score on the GALE test sets. Multi-objective optimization framework which supports heterogeneous information sources to improve alignment in machine translation system combination techniques has proven to be useful approach in Chinese-English data sets (Xia et al., 2013). We explore two approaches of system combination (a) using the word alignment and its different symmetrization method for both phrase based and hierarchical SMT paradigms to build candidate systems if they are complementary to each other for the system combination by sampling the outputs (b) using jointly trained HMM based SMT candidate systems together with MGIZA++ based best candidate systems for system combination for both phrase based and hierarchical SMT paradigms.

2 Related Work

A multiple string alignment algorithm is used to compute a single confusion network to generate a consensus hypothesis through majority voting by (Bangalore et al., 2001). There are other system combination techniques that uses TER (Snover et al., 2006) or ITG(Karakos et al., 2008) to align system outputs. The confusion network (Rosti et al., 2008) based system combination approach could improve the performance of the combined output. In this approach, one of the candidate system output which is acting as backbone determines the word order. The other candidate system outputs vote for insertion, deletion and substitution operations. Och and Ney (2003) and Koehn et al. (2003) used heuristics to merge the bidirectional GIZA++ alignments into a single alignment. Macherey and Och (2007) gave evidence that the systems to be combined should be of similar quality and need to be almost uncorrelated in order to be beneficial for system combination. The generation of diverse hypotheses from a single MT systems using traits

[†] Currently at the Department of Computer Science, University of Houston, Texas, USA. This work was carried out while the author was at NUS, Singapore.
124
D. S. Sharma, R. S. Singh and S. Shekhar, Proc. of the 12th Intl. Conference on Natural Language Processing, pages 124–129, Trivandrum, India. December 2015. ©2015 NLP Association of India (NLP AI)

such as n-gram frequency, rule frequency, average output length and average rule length without making any algorithmic change have shown gain in BLEU score (Devlin and Matsoukas, 2012) using standard system combination technique. Xu and Rosti (2010) reported the combination of unsupervised (GIZA++ and HMM based aligner) and supervised (Inversion Transduction Grammar –ITG) classes of alignment algorithms and strategy to combine them to improve overall system performance. The ITG aligner (Haghighi et al., 2009) which is an extended work of original ITG Wu (1997) to handle block of words in addition to single words is used in their work.

3 Candidate Systems

In order to establish diverse baseline systems, we start our experiment by building various baseline systems to sort out the best candidate systems using the symmetrization method for Chinese-English language pair on FBIS domain. In total, we trained, tuned and tested 20 primary candidate systems. The Chinese sentences are segmented using the Chinese segmentation module of (Low et al., 2005). Out of the 20 primary candidates, 18 of them are built using MGIZA++. Two unsupervised Chinese-English SMT systems – one phrase based SMT (PBSMT) and one hierarchical phrase based SMT (HPBSMT) are also trained based on jointly trained HMM alignment. In the jointly trained HMM aligner (Liang et al., 2006), which is an extension of classical HMM based alignment (Vogel et al., 1996), two IBM Model 1s are trained jointly, one in each direction, for 2 iterations and these parameters are used to train two HMMs jointly for 2 iterations using Berkeley Aligner¹.

3.1 Word Alignment

In the present work, we focus on the alignment symmetrization heuristic rather than the individual alignments technique such as the MGIZA++, HMM or ITG. The baseline alignment model does not allow a source word to be aligned with more than one target word which is a birth defect of IBM models. To solve this problem, training from both directions (source-to-target and target-to-source) is performed and symmetrized the two alignments to increase the quality into an alignment matrix using various combination methods (Och and Ney, 2003; Koehn et al., 2003). Exper-

iments to find out the best word alignment method are conducted for the IWSLT’05 task by Koehn et al. (2005) using some of the symmetrization techniques. The various word alignment models and its different symmetrization methods that include *source-to-target*(srctotgt), *target-to-source*(tgttosrc), *union*, *intersection*, *grow*, *grow-diag*, *grow-diag-final*, *grow-diag-final-and*, *grow-final* of MGIZA++ are used to build the baseline candidate systems for the system combination task between Chinese-English language pair. These alignment heuristics are used in two different machine translation paradigms – phrase based and hierarchical. The MGIZA++ (Gao and Vogel, 2008), a multi-threaded and drop-in replacement version of GIZA++ (Och and Ney, 2003) which is a toolkit of the original IBM alignment models (Brown et al., 1993) implementation is used to build the above models. In the case of jointly trained HMM model, an EM-like algorithm to maximize the intuitive objective function that incorporates both data likelihood and a measure of agreement between models is derived to make the predictions of the models that agree at test time but also encourage agreement during training. Thus, the symmetrization is improved by explicitly modeling the agreement between the two alignments and optimizing it during the EM training (Liang et al., 2006) and implemented in Berkeley aligner.

3.2 Phrase Based SMT systems

In all the processes of training the PBSMT systems, all the parameters of the MT systems are kept the same except the parameter that generates phrase table from a specific alignment. We trained nine different Chinese-English PBSMT systems on the FBIS corpus using 220205 sentences up to length 85, which includes 8247326 English words. These translation models are built using statistical alignment toolkits, viz. MGIZA++ and Berkeley Aligner. The Berkeley aligner is used for jointly trained HMM alignment model (Liang et al., 2006). Standard training regimen is used - 5 iterations of model 1, 5 iterations of HMM, 3 iterations of Model 3, and 3 iterations of Model 4 for the MGIZA++ based models. All these systems are tuned using MTC13 data set of 1928 sentences. Hierarchical lexicalized reordering model (Galley and Manning, 2008) is used to improve non-local reordering and MBR (Kumar and Byrne, 2004) decoding is carried out to minimize expected loss of translation errors under loss functions to select a

¹ <https://code.google.com/p/berkeleyaligner/>

consensus translation on NIST Chinese-English test sets. Maximum phrase length is set to 7 for MGIZA++ based models and 5 for jointly trained HMM based models. A 5-gram interpolated with Knesyer Ney smoothing (Chen and Goodman, 1998) language model built using SRILM (Stolcke, 2002) is used. In our experiment, the Chinese-English SMT systems are decoded using Moses SMT toolkit² (Koehn et al., 2007). All the systems are evaluated on NIST 2002 /2003 /2004 /2005/ 2006 /2008 test sets using automatic scoring toolkit *mteval-v11b.pl* giving BLEU with four reference translations. Table 1 shows the BLEU scores of PBSMT and HPBSMT candidate systems.

3.3 Hierarchical Phrase Based SMT systems

Nine different baseline hierarchical phrase based statistical machine translation (HPBSMT) systems are trained, tuned and tested. In the process of building the HPBSMT systems; all the parameters and components of the MT systems are kept the same except the one that generates specific alignment. The hierarchical phrase based models (Chiang, 2005) of Chinese-English language pair are built using the Moses SMT toolkit (Koehn et al., 2007) choosing the different alignment models using MGIZA++ and Berkeley aligner. To reduce the number of lexical items in the grammar, the maximum phrase length is set to 5 in the process of candidate MT systems training. Good Turing discounting is used for rule scoring to reduce actual counts. We carried out training for all the Chinese-English hierarchical phrase based SMT systems on the FBIS corpus of 220205 sentences up to length 85. The individual baseline candidate systems are tuned using MTC13 data set of 1928 sentences using minimum error rate training (Och, 2003) to optimize the feature weights and tested on NIST 2002 /2003 /2004 /2005 /2006 /2008 test sets with four reference translations. A 5-gram interpolated with Knesyer Ney smoothing language model built using SRILM (Stolcke, 2002) is used. The hierarchical phrase-based models are trained without the reordering model since most of the reordering behavior between the source and the target languages are expected to be captured by the synchronous grammar.

4 System combination of the Chinese-English SMT systems

The system combination of the candidate systems is carried out using MEMT3 (Heafield and Lavie, 2010). Alignment of outputs at the token level is carried out using a variant of METEOR (Denkowski and Lavie, 2010) aligner. This approach uses case-insensitive exact, stem, synonym and unigram paraphrase matched data for the alignment. Thereafter, a search space is defined on top of these alignments and a hypothesis starts with the first word of some sentences. The duplication in the output is prevented by ensuring that a hypothesis contains at most one word from each group of aligned words. Tuning and decoding is carried out using interpolated 5-gram language model with Knesyer-ney smoothing built using SRILM (Stolcke, 2002). A beam search with recombination is used since the search space is exponential in the sentence length. The language model log probability is used as a feature to bias translations towards fluency. Hypotheses to recombine are detected by hashing the search space state, feature state and hypothesis history up to a length requested by the features. The Best n+n combination (n+n combo hereafter) picks up the best n candidate system from PBSMT and best n candidate from HPBSMT systems. The outputs of the candidate systems are sampled and carried out the experiments choosing the best 3/4/5/6/7/8/9 candidate sets from each paradigm i.e., 3 from phrase based and 3 from hierarchical phrase based, 4 from phrase based and 4 from hierarchical phrase based and so on for NIST 2003, NIST 2004, NIST 2005, NIST 2006 and NIST 2008 test sets. The combo systems are tuned on NIST 2002 test set using Z-MERT (Zaidan, 2009). The result of the combo systems of each MT paradigm is shown by Table 2 for PBSMT and Table 3 for HPBSMT systems. All the baseline candidate systems and combo systems are evaluated in terms of BLEU metric (Papineni et al, 2002) with four reference translations on NIST test sets using standard MT evaluation toolkit *mteval-v11b.pl*. Table 6 shows BLEU scores of best candidates and the combo system of Berkeley Aligner and MGIZA++ based systems and gain in BLEU scores.

² <http://www.statmt.org/moses/>

³ <http://kheafield.com/code/memt/>

Symmetrization technique	NIST2002		NIST2003		NIST2004		NIST2005		NIST2006		NIST2008	
	BLEU		BLEU		BLEU		BLEU		BLEU		BLEU	
	PB	HPB										
grow-diag-final-and	31.56	30.75	29.56	28.30	32.14	31.51	25.82	25.02	28.10	27.02	21.65	20.87
intersection	29.81	29.80	28.11	28.07	30.66	30.70	26.83	26.76	28.19	27.71	22.16	20.68
union	27.56	27.52	25.69	25.13	28.29	28.42	25.26	24.69	25.56	25.51	21.04	20.40
grow-diag-final	30.07	28.65	28.30	26.66	30.67	29.45	25.13	23.03	26.58	25.02	21.72	20.20
grow-diag	31.44	31.17	29.83	28.89	32.06	31.84	28.66	27.76	29.82	28.76	22.47	21.83
grow	30.68	30.17	28.89	28.10	31.50	31.23	27.81	27.09	29.12	28.62	22.53	21.48
srctotgt	29.67	28.09	27.58	25.77	29.80	28.82	26.83	25.44	27.22	25.67	21.89	19.79
tggtosrc	31.26	30.14	29.11	28.02	31.90	31.32	25.61	24.49	27.75	26.27	21.65	20.76
grow-final	30.11	28.50	28.42	26.56	30.60	29.28	24.89	23.12	26.66	25.01	21.20	20.49

Table 1: BLEU scores of different PB-SMT(PB) and HPB-SMT(HPB) systems keeping unknown words

Combo Systems	NIST2003 BLEU		NIST2004 BLEU		NIST2005 BLEU		NIST2006 BLEU		NIST2008 BLEU	
Best Candidate	29.83	Gain	32.14	Gain	28.66	Gain	29.82	Gain	22.53	Gain
Best 3 combo	29.75	-0.08	32.43	0.29	28.88*	0.22	29.75	-0.07	23.38	0.85
Best 4 combo	30.05	0.22	32.55	0.41	28.78	0.12	29.88	0.06	22.65	0.12
Best 5 combo	30.12	0.29	32.60	0.46	28.85	0.19	29.62	-0.2	23.83	1.3
Best 6 combo	30.43	0.6	32.69	0.55	29.08	0.42	30.25	0.43	24.24*	1.71
Best 7 combo	30.05	0.22	32.57	0.43	29.01	0.35	30.46	0.64	23.48	0.95
Best 8 combo	30.50	0.67	32.83	0.69	28.70	0.04	29.61	-0.21	23.37	0.84
All-systems-combo	30.33	0.5	32.76	0.62	29.01	0.35	30.09	0.27	24.43	1.9

Table 2: BLEU scores of System combination of different alignment PBSMT models

Combo systems	NIST 2003		NIST 2004		NIST 2005		NIST 2006		NIST 2008	
	BLEU	Gain	BLEU	Gain	BLEU	Gain	BLEU	Gain	BLEU	Gain
Best Candidate	28.89	-	31.84	-	27.76	-	28.76	-	21.83	-
Best 3 combo	28.80	-0.09	31.76	-0.08	27.63	-0.13	28.76	0.0	22.21	0.38
Best 4 combo	29.18	0.29	31.91	0.07	27.94	0.18	29.35	0.59	21.91	0.08
Best 5 combo	29.28	0.39	32.43	0.59	27.98	0.22	29.37	0.61	22.93	1.10
Best 6 combo	29.45	0.56	32.38	0.54	27.90	0.14	29.45	0.69	22.93	1.10
Best 7 combo	29.31	0.42	32.38	0.54	28.09	0.33	29.65	0.89	23.57	1.74
Best 8 combo	29.33	0.44	32.18	0.34	27.55	-0.21	29.20	0.44	23.24	1.41
All-systems-combo	29.31	0.42	32.27	0.43	27.85	0.09	28.99	0.23	23.64	1.81

Table 3: System combination of different alignment models of HPBSMT and gain in BLEU

Alignment Model	NIST2002	NIST2003	NIST2004	NIST2005	NIST2006	NIST2008
	BLEU	BLEU	BLEU	BLEU	BLEU	BLEU
PBSMT- HMM	32.35	29.88	32.24	28.14	29.64	22.68
HPBSMT -HMM	31.73	29.53	31.82	27.91	29.03	21.91

Table 4: BLEU scores of jointly trained HMM candidate systems using Berkeley Aligner

Combo Systems	NIST2003 BLEU score		NIST2004 BLEU score		NIST2005 BLEU score		NIST2006 BLEU score		NIST2008 BLEU score	
Best Candidate	29.83	Gain	32.14	Gain	28.66	Gain	29.82	Gain	22.53	Gain
Best 3+3 combo	30.36	0.53	32.65	0.51	29.29	0.63	30.65	0.83	23.59	1.06
Best 4+4 combo	30.29	0.46	33.01	0.87	29.46	0.80	30.14	0.32	24.26	1.73
Best 5+5 combo	30.61	0.78	32.84	0.70	29.34	0.68	30.56	0.74	24.19	1.66
Best 6+6 combo	30.72	0.89	33.41	1.27	29.18	0.52	30.95	1.13	24.60	2.07
Best 7+7 combo	30.53	0.70	33.40	1.26	29.31	0.65	30.28	0.46	24.08	1.55
Best 8+8 combo	30.18	0.35	32.90	0.76	29.06	0.40	30.20	0.38	24.35	1.82
All-9+9-systems-combo	30.29	0.46	33.23	1.09	29.17	0.51	30.07	0.25	24.56	2.03

Table 5: System combination of MGIZA++ alignment models of PBSMT and Hierarchical PBSMT

Alignment Models	NIST02	NIST03	NIST04	NIST05	NIST06	NIST08
PBSMT- HMM	32.35	29.88	32.24	28.14	29.64	22.68
HPBSMT- HMM	31.73	29.53	31.82	27.91	29.03	21.91
Best-PBSMT-MGIZA++	31.56	29.83	32.14	28.66	29.82	22.53
Best-HPBSMT-MGIZA++	31.17	28.89	31.84	27.76	28.76	21.83
Combo of Four Systems	-	30.96	32.92	29.35	30.57	23.80
Gain in BLEU	-	1.08	0.68	0.69	0.75	1.12

Table 6: BLEU scores of best candidates and the combo system of Berkeley Aligner and MGIZA++

5 Discussion

We compare the best candidate systems performance against the combo system performance on NIST test sets and find out the gain in BLEU score. To our knowledge, our work is the first to report the extensive exploitation of the various alignment symmetrization heuristics for system combination. We carried out extensive experiments on Chinese-English language pair on specific domain and found that certain set of systems could improve output of the system combination. It is difficult to identify the right set of combinations of the candidate systems that shows the maximum gain in the BLEU score of the combined outputs unless we carry out all the possible system combination. The combination of all candidate systems does not always give the highest gain in BLEU score. The best 6+6 combo gives maximum gain in BLEU score as shown in Table no 5 against the combo systems of each paradigm for all the test sets except for NIST 2005 test set. The best 4+4 combo gives the maximum gain in BLEU score for NIST 2005 test set. The *grow-diag* alignment symmetrization method works best for HPBMST systems in terms of BLEU score of all the NIST test sets. The system combination of the PBSMT systems shows interesting characteristics of diversity. For example, there are two sets of NIST 2005 Best-3-system, viz. system combination of (*grow-diag*, *grow*, *intersection*) gives BLEU score 28.35 and system combination of (*grow-diag*, *grow*, *srctotgt*) results BLEU score of 28.88. This shows that *srctotgt* based system show higher degree of complementary behavior than *intersection* based system to *grow-diag* and *diag* based models though *intersection* and *srctotgt* systems give same BLEU scores.

6 Conclusion

The combo systems of PBSMT based candidate systems show gain in the range of 0.42 to 1.9 absolute BLEU score and combo systems of HPBSMT candidates show gain of 0.33 to 1.81 absolute BLEU score from the respective best

candidates of the corresponding test sets. The overall gain of combo systems of both paradigms in the BLEU score from the best candidate systems of respective test set are in the range of 0.8 to 2.07 absolute. So, the outputs of the PBSMT and HPBSMT are complementary in system combination task of FBIS Chinese-English parallel corpora. The jointly trained HMM based PBSMT systems outperform the other candidates of MGIZA++ for NIST 2002, NIST 2003, NIST 2004 and NIST 2008 test sets. The combo systems of the four candidate systems using two aligner viz. MGIZA++ and Berkeley aligner and two MT paradigms show gain of BLEU score in the range of 0.68 to 1.12 from the best candidates of the respective test sets.

Acknowledgement

This work was supported by the grant under project ‘‘Research Collaboration between NUS and CSIDM: Phase 2 [R-252-100-372-490]’’. I thank Prof. Hwee Tou Ng of Dept of Computer Science, National University of Singapore.

References

- Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904.
- Antti-Veikko I Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2008. Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In *Proc. of WSMT*.
- 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, Suntec, Singapore.
- Damianos Karakos, Jason Eisner, Sanjeev Khudanpur, Markus Dreyer. 2008. Machine Translation System Combination using ITG-based Alignments, In *Proceedings of ACL-08: HLT*, Short Papers (Companion Volume), pages 81–84, Columbus, Ohio, USA.
- David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation, *Pro-*

- ceedings of the 43rd Annual Meeting of the ACL, pages 263–270, Ann Arbor.
- Dekai Wu, 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3).
- Franz Josef Och, Hermann Ney. 2003 A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, vol. 29, pages 19–51.
- Franz Josef Och, Minimum error rate training in statistical machine translation, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, p.160-167, July 07-12, 2003, Sapporo, Japan.
- Jacob Devlin and Spyros Matsoukas. 2012. Trait-Based Hypothesis Selection For Machine Translation, *In proceeding of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 528–532, Montréal, Canada.
- Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. *In Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*.
- Jinxi Xu and Antti-Veikko I. Rosti. 2010. Combining Unsupervised and Supervised Alignments for MT: An Empirical Study, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 667–673, MIT, Massachusetts, USA, 9-11.
- Kenneth Heafield, Alon Lavie. 2010. Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme. *The Prague Bulletin of Mathematical Linguistics*, No. 93, pages 27–36, ISBN 978-80-904175-4-0. doi: 10.2478/v10108-010-0008-4..
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea 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*.
- Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model, *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Honolulu.
- Omar Zaidan. 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *Prague Bulletin of Mathematical Linguistics*, 91:79-88.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA.
- 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, New York.
- Peter F. Brown, Vincent J. Della Pietra, Stephen, A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Philipp Koehn, Franz Josef Och, Daniel Marcu, 2003. Statistical phrase based translation. *In proceedings of the Joint conference on HLT-NAACL*, Volume 1, Pages 48-54. Edmonton, Canada.
- Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation
- 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 Proc. of ACL: Poster*, pages 177–180.
- Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool, *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49-57, Columbus, Ohio, USA.
- Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation, *In proceeding of HLT-NAACL*, pages 169-176, Boston, Massachusetts, USA.
- Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. *In Automatic Speech Recognition and Understanding*.
- Tian Xia, Zongcheng Ji, Shaodan Zhai, Yidong Chen, Qun Liu, Shaojun Wang. 2013. Improving Alignment of System Combination by Using Multi-objective Optimization, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 535–544, Seattle, Washington, USA.
- Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University Center for Research in Computing Technology.
- Stephen Vogel, Hermann Ney and Christoph Tillmann. 1996. HMM-based Word Alignment in Statistical Translation. *In COLING '96: The 16th International Conference on Computational Linguistics*, pages 836-841, Copenhagen, Denmark.
- Wolfgang Macherey, Franz Josef Och, 2007. An empirical study on computing consensus translation from multiple machine translation systems, *Proceedings of the 2007 joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 986-995, Prague.