The NICT Translation System for IWSLT 2011

Andrew Finch

Multilingual Translation Group
MASTAR Project
National Institute of Information
and Communications Technology
Kyoto, Japan

Chooi-Ling Goh

Multilingual Translation Group
MASTAR Project
National Institute of Information
and Communications Technology
Kyoto, Japan

Graham Neubig

Graduate School of Informatics
Kyoto University
Yoshida Honmachi
Sakyo-ku
Kyoto, Japan

Eiichiro Sumita

Multilingual Translation Group
MASTAR Project
National Institute of Information
and Communications Technology
Kyoto, Japan

Abstract

This paper describes NICT's participation in the IWSLT 2011 evaluation campaign for the TED speech translation Chinese-English shared-task. Our approach was based on a phrase-based statistical machine translation system that was augmented in two ways.

Firstly we introduced rule-based re-ordering constraints on the decoding. This consisted of a set of rules that were used to segment the input utterances into segments that could be decoded almost independently. This idea here being that constraining the decoding process in this manner would greatly reduce the search space of the decoder, and cut out many possibilities for error while at the same time allowing for a correct output to be generated. The rules we used exploit punctuation and spacing in the input utterances, and we use these positions to delimit our segments. Not all punctuation/spacing positions were used as segment boundaries, and the set of used positions were determined by a set of linguistically-based heuristics.

Secondly we used two heterogeneous methods to build the translation model, and lexical reordering model for our systems. The first method employed the popular method of using GIZA++ for alignment in combination with phrase-extraction heuristics. The second method used a recently-developed Bayesian alignment technique that is able to perform both phrase-to-phrase alignment and phrase pair extraction within a single unsupervised process. The models produced by this type of alignment technique are typically very compact whilst at the same time maintaining a high level of translation quality. We evaluated both of these methods of translation model construction in isolation, and our results show their performance is comparable. We also integrated both models by linear interpolation to obtain a model that outperforms either component. Finally, we added an indicator feature into the log-linear model to indicate those phrases that were in the intersection of the two translation models. The addition of this feature was also able to provide a small improvement in performance.

1. Introduction

In the IWSLT 2011 evaluation campaign, the NICT team participated in TED speech translation shared-task for Chinese-English. This paper describes the machine translation approach adopted for this campaign.

Our overall approach was to take a phrase-based statistical machine translation decoder and improve its performance by means of two strategies. The first strategy was to constrain the re-ordering process during decoding so that the input is decoded in chunks rather than as a single larger unit. In this manner we aimed to improve the translation accuracy by preventing reordering errors that mix phrases from differ-

Case/Punctuation	BLEU	NIST	WER	PER	GTM	METEOR	TER
Case and punct	0.1190	4.6929	0.7177	0.5746	0.4844	0.4847	67.1430
No case and no punct	0.1106	4.7142	0.7523	0.5977	0.4620	0.4503	71.8380

Table 1: The official results for the NICT system in terms of a variety of automatic evaluation metrics.

ent sections of the sentence that could more effectively be translated separately. Since the input utterances are punctuated and contain spaces that indicate word boundaries we exploited this punctuation and space information as cues to determine the likely positions to delimit segmentation boundaries. In previous work [1] it has been shown that constraints of this type can be useful in managing the decoding of longer sentences. Constraining the search in the right way leaves a simpler problem for the machine translation decoder to solve, and one that can be performed considerably more efficiently than unconstrained decoding over the full search space..

The second strategy was to build the translation model for the system using two heterogeneous methods. The translation model is a key component in any phrase-based SMT system, and building this model using two different techniques can potentially bring benefits in two ways. Firstly, by integrating the two tables we can extend the coverage of our translation model. Secondly, by identifying the set of phrases that are shared between both models we identify those phrases that have support from both processes and these phrase-pairs we hypothesize are more reliable/useful to the system. We therefore added an additional feature into the log-linear model that is intended to bias the system towards choosing phrase-pairs from the overlap of the two translation models.

For reference, the official scores for the NICT system with respect to several of the automatic metrics used for the official evaluation are given in Table 1.

The overall layout of our paper is as follows. In the next section we describe the underlying phrase-based statistical machine translation system that forms the basis of all of the systems reported in this paper. In the following section we describe the rule-based re-ordering constraints based on punctuation and spacing cues present in the input text that we used to constrain and thereby simplify the decoding process. The next section, presents the Bayesian alignment technique that we used to independently build a second translation model, along with techniques for integrating this model into the base system. Finally we conclude and offer some directions for future research.

2. The Base System

2.1. Decoder

The decoder used in these experiments is an in-house phrase-based statistical machine translation decoder OCTAVIAN than can operate in a similar manner to the publicly available MOSES decoder [2]. The base decoder used a fairly standard set of features that were integrated into a log-linear model using independent exponential weights for each feature.

These features were:

1. Two language models each with independent log-linear weights
   - An in-domain language model built on the TED speech training data
   - A language model built on the larger out-of-domain Europarl corpus.
2. Five translation model features
   - Probability of the source phrase given the target
   - A similar feature but based on IBM model 1 [3]
   - Probability of the target phrase given the source
   - A similar feature but based on IBM model 1
   - A phrase-pair insertion penalty feature
3. A simple distance-based distortion feature
4. Six lexical distortion model features
   - Monotone (current phrase-pair)
   - Discontinuous (current phrase-pair)
   - Swap (current phrase-pair)
   - Monotone (previous phrase-pair)
   - Discontinuous (previous phrase-pair)
   - Swap (previous phrase-pair)
5. A word insertion penalty feature

Based on a set of pilot experiments we decoded with no limit on the distances phrases could be moved in the reordering process during decoding. The base model above was augmented with an additional translation model feature intended to be an indicator of the quality/reliability of each phrase-pair; this feature will be explained later in Section 4.2.

2.2. Pre-processing

The Chinese data supplied for this task was not segmented into words. We used the Stanford Chinese word segmentation tool [4, 5] with the Peking University (PKU) model to word-segment this data. The English data was tokenized by applying a number of regular expressions to separate punctuation, and split contractions such as “it’s” and “hasn’t” into two separate tokens. We also removed all case information from the English text to help to minimize issues of data sparseness in the models of the translation system. All punctuation was left in both source and target. We took the decision to generate target punctuation directly using the process of translation, rather than as a punctuation restoration step in post processing based on experiments carried out for last year’s IWSLT shared evaluation [6].

2.3. Post-processing

The output of the translation system was subject to the following post-processing steps which were carried out in the order in which that are listed.

1. Chinese characters were removed. Out of vocabulary words (OOVs) were passed through the translation process unchanged, some of these OOVs were Chinese and some English. We took the decision to keep the English OOVs in the output, but remove the Chinese characters. This seemed like a reasonable strategy that would benefit the system in the human evaluation, but was based on no empirical evidence. In the example given later in the paper (in Figure 3) the OOV ‘Skillman’ would have been (correctly in this case) left in the target world sequence.

2. The output was de-tokenized using a set of heuristics implemented as regular expressions designed to undo the process of English tokenization. Punctuation was attached to neighboring words and tokens that form split contractions were combined into a single token.

3. The output was re-cased using the re-casing tool supplied with the MOSES [2] toolkit. We trained the re-casing tool on untokenized text from the TED talk training data combined with the larger Europarl corpus.

2.4. Training

2.4.1. Language Models

The two language models were both built in the same manner using the SRI language modeling toolkit [7]. 5-gram models were built for decoding the development and test data for evaluation, and 3-gram models were built for decoding during the parameter tuning process to speed up decoding. The language models were smoothed using modified Knesser-Ney smoothing.

2.4.2. Translation Model

The translation model for the base system was built in the the standard manner using a 2-step process. First the training data was word-aligned using GIZA++. Second, the grow-diag-final-and phrase-extraction heuristics from the MOSES [2, 8] machine translation toolkit were used to extract a set of bilingual phrase-pairs using the alignment produced by GIZA++. In our system we also use a second translation model that is created using a Bayesian alignment technique, we will describe the both the technique and the manner in which the translation model is created from it in Section 4.

2.4.3. Parameter Tuning

To tune the values for the log-linear weights in our system, we use the standard minimum error-rate training procedure (MERT) [9]. The weights for the models were tuned using the development data supplied for the task. To perform the MERT tuning we used the publicly available ZMERT framework [10], and this allowed us to easily add and tune additional features into our models. The models were tuned with respect the BLEU metric [11]: ‘BLEU4 Closest’ that is built into the tool.

3. Rule-based Decoding Constraints

3.1. Motivation

Translating long and complex sentences has been a critical problem in machine translation. A standard phrase-based statistical machine translation system cannot solve the problem of word reordering in the target when the source sentence has a complex structure. A syntax-based machine translation system could solve the problem by running a parser on the source sentence in order to get the syntactic structure, but when a sentence is long and complex, the parser may fail to give a correct parse tree. Klein and Manning [12] have shown that the accuracy of parsing decreases as sentence length increases, and the parsing time increases dramatically. However, in this research, we found that even when a sentence is long and complex, it is possible to split a sentence into smaller units which can be translated separately with minor consideration of the context. The main problem here is locating the best locations for the split. We use linguistic information such part-of-speech (POS) tags and commas as clues to determine the split positions. After splitting a sentence into small clauses, the clauses are translated almost independently. This means that word reordering can only be done within a clause, not between clauses. This constraint can be specified using “wall” tag in MOSES (as in Koehn and Haddow [13]), and we implemented the same scheme in the OCTAVIAN decoder.

3.2. Methodology

A large body of previous research has shown that punctuation is very useful when parsing a text [14, 15, 16, 17]. The comma is one such useful mark. Basically, a comma has two roles: as a delimiter to separate different syntactic types, or as a separator to separate the elements of the same category type [18]. However, this information alone is not enough to distinguish whether the comma is suitable to be a split position for machine translation. A comma and the information around the comma could help to find a proper place for a split. Whether or not it is a proper place for a split depends upon if the information on the left and right sides of the comma are able to be translated independently. Punctuation can be very useful in written texts for aiding in comprehension. According to Murata et al. [19], there are more than 8 uses for commas in Japanese written text, and 36.32% of commas are used when the context before and after are independent of each other. This indicates to us that a Japanese comma can be used as a clue for a split positions. However, while a comma is usually used in Japanese to improve readability if a sentence is long and complicated, its use is not compulsory and there are no strict rules on usage, so research is being done on inserting missing punctuation into the text [19, 20]. Similar to Kim and Ehara [21], we employ a rule-based approach to split a sentence into multiple clauses. First, the sentence is part-of-speech (POS) tagged using the Stanford Chinese part-of-speech tagger [22, 23]. In many cases, if there is a comma, the context before and after the comma may be independent and can be translated separately, making a comma a very important clue for locating splitting position candidates. However, not all commas are suitable to be used as split boundaries. We therefore combine the POS tags and commas as clues to determine the split position for long sentences. Table 2 shows some of the POS tags that have been used for splitting Chinese text. These POS tags were analyzed and found to be good markers for splitting position candidates, as the clauses before and after they occur may be independent of each other, and thus able to be translated independently. Two simple rules that were used are:

1. If a POS tag in the head position is found after a comma, then the head will be a split position.
2. If a POS tag in the tail position is found before a comma, then the word after the comma will be a split position.

Examining the segmenation points that were inserted into the development data we observed that our spitting heuristics were applicable to approximately 1/3 of all sentences and furthermore that delimiters were inserted after about half of all commas in the data.

3.3. Examples

A standard phrase-based statistical machine translation system does not work well for translating long sentences. This is because the longer the sentence, the larger the search space for reordering becomes. Therefore, the word order in the translation may not be arranged in the correct order as in the source. Figure 3 shows an example of a long sentence translation with and without constraints in the decoding process. In this example the word order of the translation with no constraints does not follow the source sentence and as a consequence the translation is not satisfactory. However, if we can split the sentences into smaller clauses (delimited in the figure by the <wall /> token), each clause can be translated with a better word order, and the overall translation improves. In this example the first phrase, “几 年 前 ， 在 TED 大会 上 ， ” translates correctly as “a few years ago , at the TED conference ,”, but the unconstrained decoding process has separated the translation of “TED 大会” (TED conference) and placed at the end of the target sentence. This has led to errors in the translation. In the constrained decoding process, the decoder is forced to translate this together with the start of the sentence, this in turn has resulted in a simpler, more monotonic translation process that better matches the structure of the source sentence, which has given rise to fewer errors.

Head Position	
POS tag	Description
AD	adverb
CC	coordinating conjunction
CS	subordinating conjunction
P	preposition
DT	determiner
PN	pronoun
Tail Position	
POS tag	Description
LC	localizer

Table 2: POS tags used a cues for splitting in Chinese

3.4. Experiments

We evaluated the effectiveness of this approach using the the supplied training, development and test corpora for the task. The training procedure was the same in both experiments, the only difference was whether or not constraints were applied to the decoding process. We limited the evaluation to those sentences (approximately 30% of the corpus) to which splitting was applied. The results from this experiment are given in Table 4. The results show that using these constraints on the decoding process gives rise to an improvement in machine translation quality. This is in line with previous results reported on different data sets [1]. The improvement is modest but the approach can be expected to be more effective when the sentences are longer.

Source	几 年 前 ， 在 TED 大会 上 ， <wall /> Peter Skillman 介绍 了 一个 设计 挑战 叫做 “ 棉花糖 挑战 ”
Unconstrained	a few years ago , in a design challenge is called ” the marshmallow , peter |Skillman at ted talk . ”
Constrained	a few years ago , in the ted conference , peter |Skillman introduced a design challenge is called ” the marshmallow . ”

Table 3: An example of the use of split points (the split position is marked here using the <wall /> token) to constrain the decoding process (Skillman here is an out-of-vocabulary word and is marked with a ‘|’ symbol.).

Decoding constraints	**BLEU score**
Unconstrained decoding	10.84
Constrained decoding	11.16

Table 4: The effect of re-ordering constraints on translation quality.

4. Bayesian Alignment

4.1. Motivation

In a standard phrase-based statistical machine translation system (and in the base system we used in this shared evaluation), a two-step alignment and extraction process is commonly used. In the first step, word-level alignment is performed both from source-to-target and from target-to-source using the publicly available GIZA++ [24] tool. In the second step, these two word-level alignments are combined and by means of a set of heuristics, a large set of bilingual phrase pairs that are consistent with these alignments are extracted. This approach although inelegant has proven itself to be highly effective in practice, and this is the reason for its pervasiveness. However, other approaches are possible. DeNero and Klein [25] point out that this two step approach results in word alignments that are not optimal for the final task of generating phrase tables that are used in translation. As a solution to this, they proposed a supervised discriminative model that performs joint word alignment and phrase extraction, and found that joint estimation of word alignments and extraction sets improves both word alignment accuracy and translation results.

In our system we employ a related technique that is able to perform direct phrase-to-phrase alignment and extraction in a single unified framework in a fully unsupervised manner [26]. The technique is based on a Pitman-Yor process model. Bayesian models of this form have recently proved themselves useful in the field of natural language processing, as they typically offer benefits over more traditional techniques based on maximum likelihood. In particular, they model the data according to a power law distribution that is often observed in linguistic data. Moreover, by encouraging the re-use of parameters in the model during training, Bayesian models of this type will prefer to build very compact models with few parameters that have a tendency not to over-fit the data. In many-to-many word alignment this over-fitting manifests itself as a tendency for the models to simply memorize long bilingual sequence pairs rather than explain them with shorter units. In their experimental evaluation, Neubig et al. [26] found that using their Bayesian technique to build translation models for a phrase-based statistical machine translation system resulted in far smaller translation models that were able to give approximately the same translation performance as the larger models produced by GIZA++ and grow-diag-final-and. These results (on Japanese-English) are shown in Figure 1.

4.2. Methodology

In the system used for the shared evaluation, two translation models were combined (together with their corresponding lexical distortion models) by linear interpolation: a translation model created in a standard manner using GIZA++ and grow-diag-final-and phrase-pair extraction heuristics, and a translation model built using the Bayesian aligner. In addition we added a feature into the log-linear model that contributed a constant value to phrase-pairs that occurred in both translation models. The linear interpolation weight and the weight for this extra feature were both tuned in the MERT process. The next section provides some experiments to evaluate the effectiveness of the model combination, and indicator feature designed to bias to the model towards more reliable phrase-pairs discovered by both alignment methods.

4.3. Experiments

In order to evaluate the effectiveness of our approach we carried out experiments on the supplied corpora for the task. The systems evaluated in these experiments were built in an identical manner using the same training procedure. The systems were:

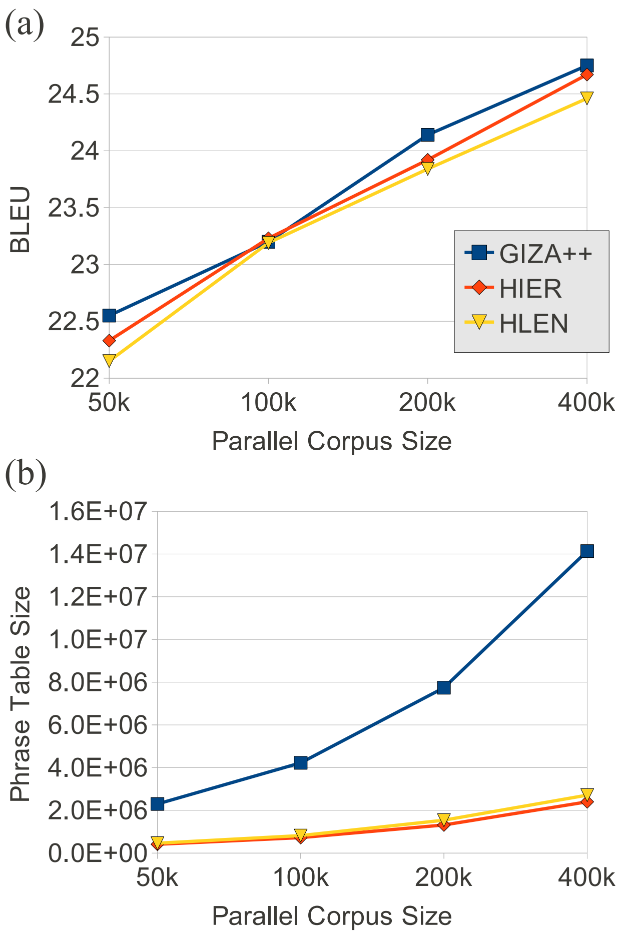

Figure 1: The effect of corpus size on machine translation performance and translation model size, comparing Bayesian alignment to GIZA++/grow-diag-final-and. a) in the figure shows the effect of corpus size on the BLEU score. b) shows the growth in model size as the corpus size is increased.

- A baseline system, that used the standard GIZA++ alignment and grow-diag-final-and phrase extraction heuristics.
- A system that used the unified Bayesian alignment and phrase extraction technique.
- A system that used linearly interpolated models for both the translation model and the lexical re-ordering model.
- A system that used linearly interpolated models for both the translation model and the lexical re-ordering model together with an additional indicator feature in the translation model to indicate those phrases in the intersection of the phrase tables.

Some statistics for the models built using these techniques are shown in Table 5. It is clear from the table that the model built using the Bayesian technique is considerably more compact than that produced in the usual manner using GIZA++ with extraction heuristics, being only around 18% of the size.

Alignment Model	**Total phrase pairs**
GIZA++	5439172
Bayesian	957201
GIZA++ and Bayesian	6059251

Table 5: The translation model size with the various techniques used to build this model.

Table 6 shows the results of our evaluation of these systems. Our results show that the Bayesian model is able to achieve a comparable level of performance to the system trained using GIZA++/extraction heuristics in spite of the huge decrease in model size. This result corroborates the findings of Neubig [26]. Interpolating the models together gives a small improvement in overall performance. Adding the indicator feature gave only a small gain, but the system performance was over the baseline level, so we used this system as out primary submission for the shared evaluation. We are currently conducting an analysis of the differences between these two phrase tables, and these results will be presented in a future publication, but it seems reasonable to assume given the high performance possible using the phrase pairs from the Bayesian alignment, that this approach is extracting only the most useful/reliable phrase pairs in the data. An alternative approach to incorporating the models from the Bayesian alignment would be to introduce the Bayesian models as independent models into the log-linear model. Time constraints ruled this approach out for this year's system, but we would like to run experiments to determine it's effectiveness in future work.

Translation Model	**BLEU score**
GIZA++	11.77
Bayesian	11.53
Combination of GIZA++ and Bayesian	11.84
Combination with indicator feature	11.93

Table 6: Translation quality of systems built with different translation models.

5. Conclusions

This paper described NICT's system for the IWSLT 2011 evaluation campaign for the TED speech translation Chinese-English shared-task. Our approach was based on a fairly typical phrase-based statistical machine translation system that was augmented firstly by introducing rule-based re-ordering constraints on the decoding. This consisted of a simple set of rules to segment the input utterances into segments that could be decoded almost independently. Our experimental results showed that this approach was quite effective in improving system performance. This result is consistent with other results using the technique reported elsewhere [1].

Secondly we used two heterogeneous methods to build the translation model and lexical reordering model for our system. The first method employed the popular method of using GIZA++ for alignment in combination with phrase-extraction heuristics. The second method used a Bayesian alignment technique. We integrated these two systems by means of phrase-table interpolation and our results show that a modest gain in performance can be obtained by doing so. Furthermore, in line with experimental results on other data sets, when trained on the shared task data we found that the translation model arising from the Bayesian alignment/extraction process was considerably more compact than that obtained by using GIZA++ with extraction heuristics in the usual manner. The model size was around 18% of the size, and this compact model gave a translation performance similar to the GIZA++-based technique in spite of its small size, a major advantage of this technique. We also added an indicator function to this model to indicate those phrases that occurred in the intersection of the two models, hopefully an indicator for the most reliable phrases in the model. We found this gave a small improvement in BLEU score.

In future work would like to explore other ways of integrating the models built from Bayesian alignment/extraction with the standard models built using GIZA++ together with extraction heuristics. We are also currently actively researching improvements to the Bayesian alignment technique, including investigating its application to hierarchical phrase-based translation [27].

6. References

[1] C. Goh, T. Onishi, and E. Sumita, "Rule-based reordering constraints for phrase-based smt," in *In Proc. of the 15th International Conference of the European Association for Machine Translation*, 2011, pp. 113–120.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowa, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," in *ACL 2007: proceedings of demo and poster sessions*, Prague, Czeck Republic, June 2007, pp. 177–180.

[3] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[4] H. Tseng, "A conditional random field word segmenter," in *In Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

[5] P.-C. Chang, M. Galley, and C. D. Manning, "Optimizing chinese word segmentation for machine translation performance," in *Proceedings of the Third Workshop on Statistical Machine Translation*, ser. StatMT '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 224–232. [Online]. Available: http://dl.acm.org/citation.cfm?id=1626394.1626430

[6] C.-L. Goh, T. Watanabe, M. Paul, A. Finch, and E. Sumita, "The NICT Translation System for IWSLT 2010," in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, M. Federico, I. Lane, M. Paul, and F. Yvon, Eds., 2010, pp. 139–146.

[7] A. Stolcke, "Srilm - an extensible language model toolkit," 1999. [Online]. Available: http://www.speech.sri.com/projects/srilm

[8] P. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models," in *Machine translation: from real users to research: 6th conference of AMTA*, Washington, DC, 2004, pp. 115–124.

[9] F. J. Och, "Minimum error rate training for statistical machine translation," in *Proceedings of the ACL*, 2003.

[10] O. F. Zaidan, "Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems," *The Prague Bulletin of Mathematical Linguistics*, vol. 91, pp. 79–88, 2009.

[11] K. Papineni, S. Roukos, T. Ward, and W. Zhu, *Bleu: a Method for Automatic Evaluation of Machine Translation*. Thomas J. Watson Research Center: IBM Research Report rc22176 (w0109022), 2001.

[12] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of ACL*, 2003, pp. 423–430.

[13] P. Koehn and B. Haddow, "Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 2009, pp. 160–164.

[14] B. Jones, "Exploring the Role of Punctuation in Parsing Natural Text," in *Proceedings of the COLING*, 1994, pp. 421–425.

[15] T. Briscoe and J. Carroll, "Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels," in *Proceedings of the 4th ACL/SIGPARSE International Workshop on Parsing Technologies*, 1995, pp. 48–58.

[16] M. J. Collins, “A New Statistical Parser Based on Bigram Lexical Dependencies,” in *Proceedings of the ACL*, 1996, pp. 184–191.

[17] M. Jin, M.-Y. Kim, D.-I. Kim, and J.-H. Lee, “Segmentation of Chinese Long Sentences Using Commas,” in *Proceedings of SIGHAN Workshop On Chinese Language Processing*, 2004, pp. 1–8.

[18] G. Nunberg, *The Linguistics of Punctuation*. CSLI lecture notes: no. 18, 1990.

[19] M. Murata, T. Ohno, and S. Matsubara, “Automatic Comma Insertion for Japanese Text Generation,” in *Proceedings of the EMNLP*, 2010, pp. 892–901.

[20] Y. Guo, H. Wang, and J. van Genabith, “A Linguistically Inspired Statistical Model for Chinese Punctuation Generation,” *ACL Transactions on Asian Language Information Processing*, vol. 9, no. 2, pp. 6:1–6:27, 2010.

[21] Y.-B. Kim and T. Ehara, “A Method for Partitioning of Long Japanese Sentences with Subject Resolution in J/E Machine Translation,” in *Proceedings of International Conference on Computer Processing of Oriental Languages*, 1994, pp. 467–473.

[22] K. Toutanova and C. D. Manning, “Enriching the knowledge sources used in a maximum entropy part-of-speech tagger,” in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, ser. EMNLP ’00. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 63–70. [Online]. Available: http://dx.doi.org/10.3115/1117794.1117802

[23] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, ser. NAACL ’03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 173–180. [Online]. Available: http://dx.doi.org/10.3115/1073445.1073478

[24] F. J. Och and H. Ney, “Improved statistical alignment models,” in *ACL00*, Hong Kong, China, 2000, pp. 440–447.

[25] J. DeNero and D. Klein, “Discriminative modeling of extraction sets for machine translation,” in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ser. ACL ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1453–1463. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858681.1858828

[26] G. Neubig, T. Watanabe, E. Sumita, S. Mori, and T. Kawahara, “An unsupervised model for joint phrase alignment and extraction,” in *ACL*, 2011, pp. 632–641.

[27] D. Chiang, “Hierarchical phrase-based translation,” *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.