International Journal of Knowledge *www.ijklp.org*
and Language Processing KLP International ©2011 ISSN 2191-2734
Volume 2, Number 1, January **2011** pp.**1--18**

Two Decades of Statistical Machine Translation

Yongzeng Xue
Department of New Media and Arts
Harbin Institute of Technology
P. O. Box 321, No. 92 West Dazhi Street
Harbin, Heilongjiang, 150001, China
Yongzeng.xue@gmail.com

Received August 2010; revised January 2011

ABSTRACT. *Statistical Machine Translation is the general term for the various data-driven methods that apply statistical models as the core mechanism to automatically translating from one language to another. Since the first practical approach was proposed in 1990, many attempts have been made to improve the state of the art. We review them here, make a clear clue of the achievements over the past twenty years, and point out a few promising directions.*

Keywords: Natural Language Processing, Machine Translation, Statistical Machine Translation

1. **Introduction.** Statistical Machine Translation (SMT) is the general term for the various data-driven methods that apply statistical models as the core mechanism to automatically translating from one language to another. The idea of SMT can be traced back to 1949, when Warren Weaver attacked the problem of machine translation with the idea of cryptography from information theory: “When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’”[1]Therefore, the process of machine translation is called “decoding” in SMT.

The design of a statistical machine translation system involves modeling, training and decoding. Modeling is the problem of how to develop the stochastic process that simulates the behavior of translation, which results in a statistical model, i.e. one or more formulae for computing the probability or score of any candidate translation of a given text. Training is the problem of how to estimate the parameters in the statistical model, usually based on the statistics derived from a large amount of texts that consist of two or more languages. These texts are called corpus. If a corpus consists of both original texts and corresponding translations, it is called parallel corpus; otherwise, if it consists of similar texts in more than one language, it is called comparable corpus. Decoding is the problem of how to generate the best translation of a text given a statistical model and the estimated values of its parameters. “The best” usually means a maximum value in probability or score.

In theory, the translation unit that we concern can be word, phrase, sentences paragraph or even essay. But in practice, we usually perform full machine translation by sentence. As a tradition derived from the first practical system of SMT, which was designed to translate French sentences into English, the language of input (i.e. the source sentence) for translation is often referred to as French, while that of output (i.e. the target sentence) as English. In [2], the probability of a translation from French sentence *F* to English sentence *E* is computed via the Bayes' theorem:

$$P(E \mid F) = \frac{P(F \mid E)P(E)}{P(F)}\,. \tag{1}$$

In Equation (1), P(F|E) is called the translation probability and P(E) is called the language model probability. Since P(F) only depends on F, it is always a fixed number for the input sentence. So we just need to search for the best candidate translation with the greatest value given by

$$\hat{E} = \arg\max_{\hat{E}} P(E \mid F) = \arg\max_{\hat{E}} P(F \mid E)P(E)\,. \tag{2}$$

The process of searching for $\hat{E}$ is called decoding, as it can be viewed as decrypting the sentence of strange symbols, F, into a sentence in familiar language, E.

The translation model P(F|E) in Equation (1) are usually viewed as the instance of a general architecture called the source-channel model, in which "an English string is statistically generated (source), then statistically transformed into French (channel)"[3]. To establish the translation model, we need to give a random process that simulates the process of translation, in which many operations may be applied, such as insertion, deletion, replacement and reordering of linguistic units. Insertion can be viewed as generating words from a NULL word that we presume has already existed in the source sentence. In a similar way, deletion can be viewed as translating words into a NULL word in the target sentence. Replacement, or substitution, is the operation of replacing the words in the source sentence with the translation in the target sentence. Of course, translation is not always a monotone process, thus the operation of reordering is introduced. [3] shows that due to reordering, the decoding problem of even the simplest word-replacement translation model (e.g. IBM Model 1, see [4]) is NP-complete in computational complexity. To illustrate this, reductions were given from two famous NP-complete problems: Hamilton Circuit Problem and Minimum Set Cover Problem, respectively. Please Note that not all reordering processes in SMT result in NP-complete problem.

Another popular paradigm is the log-linear model [5], in which the probability P(E|F) is approximately calculated from a collection of feature functions:

$$P(E \mid F) \approx p_{\lambda_1 \cdots \lambda_M}(E \mid F) = \frac{\exp\{\sum_{m=1}^{M} \lambda_m h_m(E,F)\}}{\sum_{E'} \exp\{\sum_{m=1}^{M} \lambda_m h_m(E',F)\}}\,. \tag{3}$$

The feature function hm(E,F) is usually designed as the logarithm of probability, e.g. hm(.)=logp(.), though it can be designed in any form that renders a real number value. In fact, such feature functions exist in many statistical translation models. The log-linear model can be viewed as a generalization for the source-channel model, as we can get an equivalent result by defining two feature functions:

$$h_1(E,F) = \log p(F \mid E) \tag{4}$$

and

$$h_2(E,F) = \log p(E)\,, \tag{5}$$

and set $\lambda_1=\lambda_2=1$. As the log-linear model is more flexible than the source-channel model, for example, we can add more than one translation model features or language model features, or both, it has been widely used as the framework to compute direct translation probability, while the latter is usually used to estimate the value of the translation feature functions. For the log-linear model, we can obtain the most probable sentence by

$$\hat{E} = \arg\max_{\hat{E}} \{ \sum_{\mathrm{m}=1}^{\mathrm{M}} \lambda_m h_{\mathrm{m}}(\mathrm{E},\mathrm{F}) \}\,. \tag{6}$$

In what follows, section 2 introduces the statistical models that are widely used in SMT. Section 3 overviews the training methods that estimate the model parameters, for both translation models and λs. Section 4 looks into the problem of decoding. Section 5 concludes this paper and lists some promising directions.

2. **Modeling.** In general, one can design as many models as possible to attack different problems on simulating the process of translation. Amongst all those possible models, there are three major types that are widely used to model the behavior of translation: the translation models, the language models and the reordering models, and they are usually designed in the form of statistical models. In this paper we only focus on the translation and reordering models and view the reordering models as a part of statistical translation models, please refer to [6] if you are interested in the language models.

2.1. **Statistical translation models.** The objective of designing a statistical translation model is to establish the relationship of correspondence between the concerned languages. In some early studies, a statistical translation model took into account anything of translation but the correctness of an output sentence, i.e. whether the word sequence is natural and grammatical in one language, which was usually left for the language model. In the simplest case where only the word-for-word translation is involved, one only needs to consider word selection. But in practice the problem becomes more complexity than that. In the past 20 years, we have witnessed a historical process from word-based models, phrase-based models to syntax-based models.

2.1.1. **Word-based translation models.** As a pioneer work at the very beginning of SMT, five word-based statistical translation models were presented by IBM T.J. Watson Research Center, referred as IBM Models 1-5 [4], which can be divided into two kinds: alignment-based models (Models 1,2), in which the process of translation is viewed as first generating word positions in the target sentence that are related to those in the source, and then filling them by probable translation of each source word, and fertility-based models (Models 3-5), in which the process of translation is viewed as first reproducing each source word one or more times, then replacing each reproduced word with its probable translation, and last reordering the word sequence in proper order. They are also word alignment models, as they are able to give word-to-word correspondence of bilingual sentences. A slightly modification on IBM Model 2 results in a homogeneous HMM-based alignment model[7-9], which generates much better experimental results than Model 2[10]. [10] also presented Model 6, a log-linear combination of HMM and Model 4.

There are also many statistical models are applied to the task of word alignment, a set-apart application from machine translation (e.g. [11-14]). Since those models are not directly applied to SMT, we don't plan to discuss them in detail in this paper. Another issue that this paper doesn't discuss is word clustering[15], which is usually applied to avoid data sparseness problem when estimating probabilities.

In that period, many efforts were paid on how to appropriately model the insertion, deletion and replacement of words in translation. Since it is difficult in establishing an enough sophisticated stochastic process that meets the actual translation behavior of human, as well as in estimating a precise value of the probabilities in those models, word-based models are prone to model errors. (I.e. the best translation doesn't have the highest probability according to the model.) Nowadays, the word-based models usually serve to generate word alignments for phrase or syntax-based models.

2.1.2. **Phrase-based translation models.** Phrase-based translation models aims at encoding contextual features into statistical translation models. In higher rank IBM models, e.g. IBM Model 4 or 5, the cohesion of phrases is considered and modeled by a stochastic process of finding the dependence between the head word and each of the rest words in a phrase[4]. This leads to very complex models, especially when attempting to remove the deficiency (i.e. a stealing of probabilities), also with a difficulty in parameter estimating for probabilities. Aiming at a better model for phrasal cohesion, [16,17] presented a structure-based alignment model, which first aligns phrases with rough alignment and then aligns words within each phrase with detailed alignment. The alignment is IBM-style, i.e. many-to-one alignment. In a similar way that divide the translation process into phrase level alignment followed by word level alignment, [18,19] introduced an alignment template model. Alignment template is a matrix of local word links, which enables modeling many-to-many mapping of words. In those early studies, the phrase translation probability was computed from word-to-word probabilities, thus the translation itself was still a word-for-word translation. A different model that introduces phrase directly into SMT was presented in [20], in which the translation unit is non-linguistic phrase, not word. A log-linear framework version of this model is presented in [21]. With a set of rich features, the model showed better performance against IBM Model 4. That indicates that substituting word for phrase makes it convenient to model the process of translation, and gives fairly better results.

The major problem in phrase-based models is how to extract phrases and how to give them scores (similar to probabilities, but may not be a value between 0 and 1). [20,22] found that the non-linguistic phrase, especially the phrase like "is a", plays a very important role in improving the translation results, as it can serve as a "glue" of two linguistic phrases, such as "he is" and "a man". During that period, there have mushroomed many phrase-based models and the effect brought by them continues. In this paper, we classify those models by whether they need word alignment. Similar to [22], many phrase-based models need to extract bilingual phrase pairs from word alignments, which are usually a combination of IBM-style word alignments from two sides[23]. [25,26] selected potential translation pairs by expanding from alignment points, and [27] attempted to translate with a combination of several methods that handle phrase alignments and overlapping phrases. Since alignment errors may affect the precision of phrase extraction, [28] estimated phrase scores from word translation probabilities that cover multiple word alignments. Another trend is to encode more linguistic features into phrase-based models, for example, [29,30] extended word to an integration of several aspects of linguistic information, e.g. morphological, syntactic, or semantic information, thus resulted in a factored phrase-based model, where phrase is composed not only by words, but by more linguistic annotations along with words. Phrase-based models can also established on linguistic phrase, such as chunk[31]. Other models have no need of bootstrapping from word alignment, but extracting phrase pairs and estimating phrase scores directly from statistics of words. For example, [24] presented a joint probability model instead of conditional probability model, [32] applied pair-wise mutual information to estimate the center point for phrase extraction. Benefited from introducing syntactical information, both from constituency trees and dependency trees, syntax-directed phrase extraction methods also show competitive results[33,34].

2.1.3. **Syntax-based translation models.** Syntax-based models can be traced back to as early as the phrase-based ones. [35,36] applied the motivation of producing the translation by deciding whether to exchange two components or not, referred to as Inversion Transduction Grammar (ITG), or a simplified version, Bracketing Transduction Grammar (BTG)[37]. [38,39] presented a syntax-based model of modeling the translation process from the source parse tree to the target sentence. [40] improved this model by allowing loose clone of sub-trees. In those models, the translation unit was still word, not phrase or sub-tree, so the syntax played a role no more than a reordering or re-ranking feature. Though many efforts were made to incorporate syntax features, it seemed that the syntax did not contribute to the overall improvement much[41,42], especially compared with phrase-based models. However, situation changed when the sub-tree structure was taken into account as a whole rather than the single word.

For syntax-based models, there is a distinction between formally syntax-based and linguistically syntax-based models. Formally but non-linguistically syntax-based models are those which introduce a synchronous context-free grammar but do not induce grammars of linguistic annotations. Besides the above-mentioned ITG and BTG models, such models also include the hierarchical phrase-based model[43-46], etc. Formally and linguistically syntax-based models usually rely on a parse tree or packed parse trees and can be distinguished by on which side the parse tree is introduced. [47-49] incorporated syntactic trees on the target side, while [50,51] incorporated them on the source side. [52,53] viewed translation of both source and target parse trees as a general framework of parsing. The major problem of such translation lies in that many syntactic structures between two different languages are naturally non-isomorphic. To attack this problem, Synchronous Tree Adjoining Grammar (STAG) [54-56] and Synchronous Tree Substitution Grammar

(STSG) [57-59] were introduced to syntax-based SMT. To avoid parsing errors brought by 1-best parse tree, [60-62] accepted a collection of possible parse trees as the input.

Except for parse trees, many models were proposed to base themselves on dependency trees[57,63-65]. [66] indicated that it may be more robust to phrase cohesion by using dependency trees. [67] combined the merits of both constituency and dependency trees.

Transducers, which was first introduced to word-for-word translation[68] and viewed as an alternative to grammars, has recently arisen as another approach of knowledge representation for modeling transformation on trees[56,69-71].

2.2. **Reordering models.** At the early time of SMT, reordering (or distortion) served as an operation embedded in the stochastic process that simulates the behavior of translation[4,17,36,37]. [72] showed that the ITG constrains achieve higher flexibility than the IBM model constrains. With the emergence of log-linear models and development of phrase-based SMT, the reordering probability has come to be computed separately from translation model probability. At the beginning, phrase-based SMT adopted relatively simple reordering models that pose limitation on arbitrary long jump[22]. [73] considered the distortions brought by different alignment patterns. Then the reordering pattern of adjacent blocks (i.e. phrase pairs) were considered[74-76], followed by solutions to disadjacent blocks[77]. A further step from those flat reordering models is to incorporate hierarchical reordering into phrase-based SMT, which can be based on BTG[78-81], constituency trees[82] or dependency trees[83].

3. **Training.** Training is a process to estimate parameters of a statistical model through statistics from a corpus. Usually, most of parameters of a statistical translation model are probabilities. There are three widely applied strategies of training: for models that have no hidden variables, the maximum likelihood estimation (MLE) is usually performed, for models that have hidden variables, the expectation-maximization (EM) algorithm is usually applied, and for getting adequate weights of a log-linear model, we usually perform discriminative training.

In word-based statistical translation models, and some other models of word-for-word translation, the word alignment is usually not given in the corpus, and thus becomes a hidden variable. To estimate probabilities under such a case, the EM algorithm[84] was introduced. [4] induced the parameter reestimation formulae for IBM Model 1-5, among which the formulae for Model 1, 2 are not deficient, while those for Model 3-5 are, for the counts are only summed over a collection of probable alignments according to Model 2. [4] showed that Model 1 has a unique local maximum and concaves to it. For transferring from the lower-rank model to the higher-rank model, viterbi training is usually applied[4,10]. Similar EM-based training algorithms were proposed for many translation models of word-for-word translation, with different reestimation formulae[7,17-19,37]. The early studies adopted perplexity to show the effect of training[4,7]. For the task of word alignment, precision, recall and Alignment Error Rate (AER) were also used[10]. [68,85] proposed algorithms of training a collection of finite-state transducers instead of computing probabilities.

For phrase or syntax-based statistical translation models that extract translation equivalences from word alignments, the solution becomes simple because no hidden variables are involved, thus the parameter estimation is usually implemented via maximum likelihood estimation, in which the phrase translation probability distribution is estimated by relative frequency[20-22,29-31,47-50]. Strictly speaking, some models do introduce hidden variables, but they fall into the algebraic models, and the EM training differs slightly from MLE by how to collect counts. And various training methods are presented to improve the translation performance[19,24,86]. Recently, tree kernels have been applied to explore structured features of parse trees[87,88].

For log-linear models, after estimating the probabilities in the feature functions, an extra training should be performed for estimating weights (i.e. λs), also called tuning. The tuning is usually a discriminative training performed on a held-out set the format of which is similar to the test set, called the development set, or dev-set for short. At first, discriminative training was conducted to maximize the direct translation probability, using Generalized Iterative Scaling (GIS)[89] to deal with real-valued feature functions[5]. Soon, Minimum Error Rate Training (MERT) was proposed to directly optimize translation quality by defining a loss function with respect to the automatic evaluation metric[90]. Therefore, maximizing the probability yields to minimizing the evaluation errors. [91] proposed a perceptron-style discriminative algorithm which acts in a similar way of static re-ranking systems, but needs no baseline system. [92] proposed a generalized algorithm which does not use feature functions, n-best list of outputs and dev-set. [93] proposed a semi-supervised approach that combines EM with discriminative training (i.e. EMD) for word alignment task. [94,95] applied log-linear model to refine phrase pair extraction and scoring. [96] argued that log-linear models are not expressive enough to handle few features. They presented BoostedMERT to learn more complex re-rankers than the standard MERT. [97] extended phrase-based MERT to deal with SCFG models.

To summarize, many efforts have been made to refine the basic three training strategies to explore effective usage of more linguistic knowledge and avoid overfitting to data. At the same time, more and more machine learning approaches have been introduced and a new trend that tends to combine temporary outputs of decoding directly into training has arisen.

4. **Decoding.** In SMT, decoding is the process of machine translation, i.e. the process that rewrites the encrypted messages in normal language. The objective of a decoding is to generate the most probable translation for a given sentence, usually by scoring among candidates. As decoding is often an NP-hard problem by itself, approximate algorithms are usually adopted so as to implement decoding in polynomial time. As a result, the decoding algorithm (decoder) may give a suboptimal result, which is called a search error, opposite to the model error (I.e. the translation that has the highest score according to the statistical translation model is not the best one). In other words, a search error occurs when the decoder renders a translation $\hat{E}$ for a given input sentence F but there exists a sentence E' that satisfies

$$P(E' \mid F) > P(\hat{E} \mid F). \tag{7}$$

During decoding, the candidate translation unit is first selected from the knowledge base and combined with each other, forming a temporary string or structure, which is called the partial hypothesis (or translation options) because not all parts of the input sentence are translated. Then the relevant scores are computed and all possible partial hypotheses are re-ranked. That process continues until every part of the input sentence is processed, the corresponding result is called the complete hypothesis.

The first statistical machine translation system, which was based on IBM models and outperformed a popular rule-based system, is Candide[98]. Candide adopted an analysis-transfer-synthesis paradigm, in which a stack-based decoding algorithm was presented[99]. In decoding, the hypotheses are generated by selecting word for each position from the beginning to the end of output sentence. In fact, the hypotheses are not stored in one stack. Multiple stacks are used to store hypotheses of different word length, so as to avoid longer hypothesis from being pruned off due to more production of probabilities. The search problem of IBM Models can be defined as finding both optimal translation $\hat{E}$ and most probable alignment $\hat{A}$ for a given input sentence F:

$$\langle \hat{E}, \hat{A} \rangle = \arg\max_{E,A} P(A, F \mid E) P(E). \tag{8}$$

Many efforts were made for more efficient decoding for IBM models: beam search and dynamic programming (DP) algorithm was proposed for monotone decoding[100,101] and later for limited word reordering[102,103]; multi-pass A* search was proposed with admissible heuristic function and empirical heuristic function[104]; greedy or perturbation search was also suggested to do much faster decoding[17,105-107]; similar iterative approaches were proposed that employ dynamic programming to generate final hypothesis by improving an initial hypothesis[108,109]. [105,106] took Integer Programming (IP) as an optimal search algorithm for Model 4 and compared it with the stack and greedy decoding algorithms. It was observed that the stack decoder is not much inferior to the IP decoder, and search errors only cover a small portion of the total errors. They also found search errors do not have a significant effect on the measures of translation quality, i.e. BLEU scores. [110] proposed another optimal decoder using Cutting-Plane Integer Linear Programming (ILP), which is an exact solution and more practical than traditional IP.

Beam search and stack-based decoding algorithms are widely applied to phrase-based and informally syntax-based models[18-22,29-31,50,60], which can be classified into the framework of weighted finite-state transducers (FSTs). Formally syntax-based models usually apply CKY-style algorithms to decoding[39,52-59,61-63,77,78,111], as well as tree transducers[69-71]. [52] illustrated the relationship between parsing, synchronization and translation: suppose D is the dimensionality of grammar and I is the dimensionality of input, a generalized parser that can parse in a cross-language manner becomes a synchronous parser if D=I (esp. an ordinary parser if D=I=1), a translator if D≥I, and a synchronizer if D≤I. Similar to training, discriminative approaches have also been introduced into decoding recently. To do that, consensus statistics are used to rank and combine the decoding outputs.

The Minimum Bayes Risk (MBR) decoding was presented as a sentence-level consensus-based decoding which tries to rank the partial hypotheses by loss function rather than probability[112,113]. Various phrase or word-level decoding approaches were proposed, such as consensus voting[114,115] and word-based system combination[116,117]. System combination, or multi-engine machine translation, is a strategy that combines translation outputs from several machine translation systems to generate better results. Rather than re-ranking the complete hypotheses of the candidate systems, focus has recently been posed on the capacity of generating final results from an ensemble of candidate partial hypotheses[118-120], as well as incorporating a large scale of linguistic features or rules directly into the process of decoding[121,122]. Partly because of that, [122,123] proposed a way towards extensible and general-purpose decoders.

To speed up decoding of phrase and syntax-based models, two ways have been followed in recent years. One is to employ better techniques to refine the inherent properties of the existing algorithms, such as different directional generation[124,125], different extension of hypothesis[126], different hill-climbing algorithm[127], effective pruning[128], Monte Carlo simulation[129], etc. Another is to pose linguistic constraints, such as cohesive phrases[130,131], constituent boundaries[132], translation boundaries[133] and the selective use of syntactic constraints[134], on beam search.

The problem of decoding is how to find the optimal result within a limited time and search space. To tackle this problem, many efforts have been made by introducing sophisticated heuristics or ruling out redundant hypotheses with well-designed constraints.

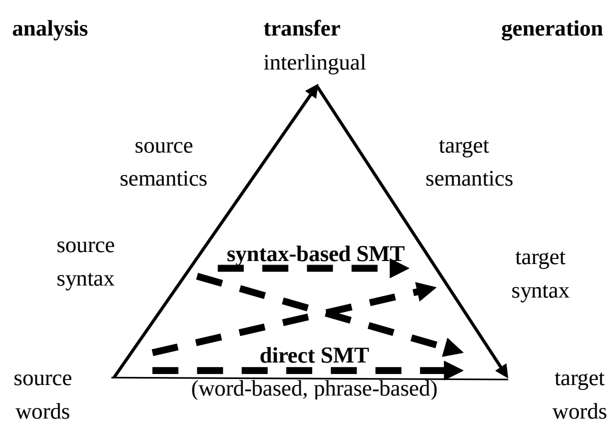

FIGURE 1. The machine translation pyramid.

5. **Conclusions.** As a paradigm of machine translation, SMT is going upwards the machine translation pyramid, where direct, syntax-based, semantic and interlingua-based translation approaches come in turn from bottom to top (Figure 1). Now semantic-based approaches have come under consideration[135,136].

More generally, researchers have held a belief for decades that linguistic information can help in improving the quality of machine translation. The open question of SMT is not whether to introduce linguistic features, but how to introduce them. As can be seen from the zigzag course of incorporating syntax features into SMT, as well as the use of word sense disambiguation (WSD)[137,138], the unsatisfying results got at the beginning time are not due to the ineffectiveness of introduced features, but due to misuse of them. We also appeal that in order to explore substantial power of linguistic knowledge, theoretical research is of the same importance as empirical methods.

Acknowledgment. This work was supported by the project of National Natural Science Foundation of China (No.60903063). We apologize for not introducing relevant issues such as transliteration, MT evaluation, WSD, and many others due to our limited knowledge. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] W. Weaver, *Machine Translation of Languages: Fourteen Essays*, Technology Press, Cambridge, MA, 1955.

[2] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer and P. S. Roossin, A statistical approach to machine translation, *Computational Linguistics*, vol.16, no.2, pp.1-24, 1990.

[3] K. Knight, Decoding complexity in word-replacement translation models, *Computational Linguistics*, vol.25, no.4, pp. 607-615, 1999.

[4] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer, The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics*, vol.19, no.2, pp. 263-311, 1993.

[5] F. J. Och and H. Ney, Discriminative training and maximum entropy models for statistical machine translation, *Proc. of the Association for Computational Linguistics*, Philadelphia, PA, pp.295-302, 2002.

[6] Ronald Rosenfeld, Two decades of statistical language modeling: where do we go from here? *Proc. of IEEE*, vol.88, no.8, pp.1270-1278, 2000.

[7] S. Vogel, H. Ney and C. Tillmann, HMM-based word alignment in statistical translation, *Proc. of the 16th Int. Conf. on Computational Linguistics*, Copenhagen, Denmark, pp.836-841, 1996.

[8] F. J. Och and H. Ney, A comparison of alignment models for statistical machine translation, *Proc. of the 18th Int. Conf. on Computational Linguistics*, Saarbrücken, Germany, pp.315-320, 2000.

[9] K. Toutanova, H. T. Ilhan and C. Manning, Extensions to HMM-based statistical word alignment models. *Proc. of the 2002 Conf. on Empirical Methods in Natural Language Processing*, Philadelphia, USA, pp.87-94, 2002.

[10] F. J. Och and H. Ney, A systematic comparison of various statistical alignment models, *Computational Linguistics*, vol.29, no.1, pp.19-51, 2003.

[11] W. Gale and K. Church, Identifying word correspondences in parallel texts, *Proc. of the 4th DARPA workshop on Speech and Natual Language*, Pacific Grove, California, pp.152-157, 1991

[12] I. D. Melamed, Models of translational equivalence among words, *Computational Linguistics*, vol.26, no.2, pp.221-249, 2000.

[13] R. C. Moore, W.-T. Yih and A. Bode, Improved discriminative bilingual word alignment, *Proc. of the*

21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, pp.513-520, 2006.

[14] Y. Liu, Q. Liu and S. Lin, Discriminative word alignment by linear modeling, *Computational Linguistics*, vol.36, no.3, pp.303-339, 2010.

[15] F. J. Och, An efficient method for determining bilingual word classes, *Proc. of the 9th Conf. of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, 1999.

[16] Y.-Y. Wang and A. Waibel, Modeling with structures in statistical machine translation, *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics, Volume 2*, Montreal, Quebec, Canada, pp.1357-1363, 1998.

[17] Y.-Y. Wang, *Grammar Inference and Statistical Machine Translation*, Ph.D. Thesis, Carnegie Mellon University, 1998.

[18] F. J. Och and H. Ney, The alignment template approach to statistical machine translation, *Computational Linguistics*, vol.30, no.4, pp.417-449, 2004.

[19] F. J. Och, *Statistical Machine Translation: From Single-Word Models to Alignment Templates*, Ph.D. Thesis, RWTH-Aachen, 2002.

[20] P. Koehn, F. Och and D. Marcu, Statistical phrase-based translation, *Proc. of the Human Language Technology Conf. 2003 (HLT-NAACL 2003)*, Edmonton, Canada, pp.127-133, 2003.

[21] P. Koehn and K. Knight, Feature-rich statistical translation of noun phrases, *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, pp.311-318, 2003.

[22] P. Koehn, *Noun Phrase Translation*, Ph.D. Thesis, University of Southern California, 2003.

[23] F. J. Och and H. Ney, Improved statistical alignment models, *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, China, pp.440-447, 2000.

[24] D. Marcu and D. Wong, A phrase-based, joint probability model for statistical machine translation, *Proc. of the 2002 Conf. on Empirical Methods in Natural Language Processing*, Philadelphia, PA, pp.133-139, 2002.

[25] A. Venugopal, S. Vogel and A. Waibel, Effective phrase translation extraction from alignment models, *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, Sopporo, Japan, pp.319-326, 2003.

[26] C. Tillmann, A projection extension algorithm for statistical machine translation, *Proc. of the 2003 Conf. on Empirical Methods in Natural Language Processing*, Sapporo, Japan, pp.1-8, 2003.

[27] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao and A. Waibel, The cmu statistical machine translation system, *Proc. of MT Summit IX*, New Orleans, Louisiana, pp.110-117, 2003.

[28] Y. Liu, T. Xia, X.Y. Xiao and Q. Liu, Weighted alignment matrices for statistical machine translation, *Proc. of EMNLP 2009*, Singapore, pp.1017-1026, 2009.

[29] P. Koehn and H. Hoang, Factored translation models, *Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech, pp. 868-876, 2007.

[30] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, Moses: open source toolkit for statistical machine translation, *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proc. of the Demo and Poster Sessions*, Prague, Czech Republic, pp.177-180, 2007.

[31] T. Watanabe, E. Sumita and H. G. Okuno, Chunk-based statistical translation, *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp.303-310, 2003.

[32] Y. Zhang, S. Vogel and A. Waibel, An integrated phrase segmentation and alignment algorithm for statistical machine translation, *Proc. of 2003 Int. Conf. on Natural Language Processing and Knowledge Engineering* , Beijing, China, pp.567-573, 2003.

[33] H. Hassan, K. Sima'an and A. Way, Supertagged phrase-based statistical machine translation, *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp.288-295, 2007.

[34] A. Zollmann, A. Venugopal, F. Och and J. Ponte, A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT, *Proc. of the 22nd Int. Conf. on Computational Linguistics*, Manchester, UK, pp.1145-1152, 2008.

[35] D. Wu, Stochastic inversion transduction grammars and bilingual parsing of parallel corpora, *Computational Linguistics*, vol.23, no.3, pp.377-403, 1997.

[36] D. Wu and H. Wong, Machine translation with a stochastic grammatical channel, *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics, Volume 2*, Montreal, Quebec, Canada, pp.1408-1415, 1998.

[37] D. Wu, A polynomial-time algorithm for statistical machine translation, *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, USA, pp.152-158, 1996.

[38] K. Yamada and K. Knight, A syntax-based statistical translation model, *Proc. of 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, pp.523-530, 2001.

[39] K. Yamada, *A Syntax-based Statistical Translation Model*, Ph.D. Thesis, University of Southern California, 2002.

[40] D. Gildea, Loosely tree-based alignment for machine translation, *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp.80-87, 2003.

[41] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin and D. Radev, Syntax for statistical machine translation, Technical Report, *Johns Hopkins University 2003 Summer Workshop on Language Engineering*, Center for Language and Speech Processing, Baltimore, MD, 2003.

[42] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L.B. Shen, D. Smith, K. Eng, V. Jain, Z. Jin and D. Radev, A smorgasbord of features for statistical machine translation, *In HLT/NAACL 2004*, Boston, MA, pp.160-167, 2004.

[43] D. Chiang, A hierarchical phrase-based model for statistical machine translation, *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, pp.263-270, 2005.

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

[45] H. Setiawan, M. Y. Kan, H. Li and P. Resnik, Topological ordering of function words in hierarchical phrase-based translation, *Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int. Joint Conf. on Natural Language Processing of the AFNLP*, Suntec, Singapore, pp.324-332, 2009.

[46] L. Cui, D. Zhang, M. Li, M. Zhou and T. Zhao, A joint rule selection model for hierarchical phrase-based translation, *Proc. of the ACL 2010 Conf. Short Papers*, Uppsala, Sweden, pp.6-11, 2010.

[47] M. Galley, M. Hopkins, K. Knight and D. Marcu, What's in a translation rule? *Proc. of HLT-NAACL 2004*, Boston, Massachusetts, pp.273-280, 2004.

[48] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang and I. Thayer, Scalable inference and training of context-rich syntactic translation models, *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.961-968, 2006.

[49] D. Marcu, W. Wang, A. Echihabi and K. Knight, SPMT: statistical machine translation with syntactified target language phrases, *Proc. of EMNLP 2006*, Sydney, Australia, pp.44-52, 2006.

[50] Y. Liu, Q. Liu and S. Lin, Tree-to-string alignment template for statistical machine translation, *Proc. of*

the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, pp.609-616, 2006.
[51] L. huang, K. Knight and A. Joshi, Statistical syntax-directed translation with extended domain of locality, *Proc. of the 7th AMTA*, Boston, MA, pp.66-73, 2006.
[52] I. D. Melamed, Statistical machine translation by parsing, *Proc. of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp.653-660, 2004.
[53] I. D. Melamed, G. Satta and B. Wellington, Generalized multitext grammars, *Proc. of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp.661-668, 2004.
[54] S. Shieber, Probabilistic synchronous tree-adjoining grammars for machine translation: the argument from bilingual dictionaries, *Proc. of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, Rochester, New York, pp.88-95, 2007.
[55] S. DeNeefe and K. Knight, Synchronous tree adjoining machine translation, *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing*, Singapore, pp.727-736, 2009.
[56] A. Maletti, A tree transducer model for synchronous tree-adjoining grammars, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.1067-1076, 2010.
[57] Y. Ding and M. Palmer, Machine translation using probabilistic synchronous dependency insertion grammars, *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, pp.541-548, 2005.
[58] M. Zhang, H. Jiang, A. T. Aw, J. Sun, S. Li and C. L. Tan, A tree-to-tree alignment-based model for statistical machine translation. *MT Summit XI*, Copenhagen, Denmark, pp.535-542, 2007.
[59] D. Chiang, Learning to translate with source and target syntax, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.1443-1452, 2010.
[60] Y. Liu, Y. Huang, Q. Liu and S. Lin, Forest-to-string statistical translation rules, *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp.704-711, 2007.
[61] H. Zhang, M. Zhang, H. Li, A. Aw and C. L. Tan, Forest-based tree sequence to string translation model, *Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int. Joint Conf. on Natural Language Processing of the AFNLP*, Suntec, Singapore, pp.172-180, 2009.
[62] H. Mi, L. Huang and Q. Liu, Forest-based translation, *Proc. of ACL-08: HLT*, Columbus, Ohio, pp.192-199, 2008.
[63] C. Quirk and C. Cherry, Dependency treelet translation: syntactically informed phrasal SMT, *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, pp.271-279, 2005.
[64] Z. Zabokrtsky and M. Popel, Hidden markov tree model in dependency-based machine translation, *Proc. of the ACL-IJCNLP 2009 Conf. Short Papers*, Suntec, Singapore, pp.145-148, 2009.
[65] D. Yuan, *Machine Translation Using Probabilistic Synchronous Dependency Insertion Grammars*, Ph.D. Thesis, University of Pennsylvania, 2006.
[66] H. J. Fox, Phrasal cohesion and statistical machine translation, *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, Philadelphia, PA, pp.304-311, 2002.
[67] H. Mi and Q. Liu, Constituency to dependency translation with forests, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.1433-1442, 2010.
[68] H. Alshawi, S. Bangalore and S. Douglas, Learning dependency translation models as collections of

finite-state head transducers, *Computational Linguistics*, vol.26, no.1, pp.45-60, 2000.

[69] J. Graehl, K. Knight and J. May, Training tree transducers, *Computational Linguistics*, vol.34, no.3, pp.391-427, 2008.

[70] G. Iglesias, A. de Gispert, E. R. Banga and W. Byrne, Hierarchical phrase-based translation with weighted finite state transducers, *Proc. of Human Language Technologies: The 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, pp.433-441, 2009.

[71] J. May, K. Knight and H. Vogler, Efficient inference through cascades of weighted tree transducers, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.1058-1066, 2010.

[72] R. Zens and H. Ney, A comparative study on reordering constraints in statistical machine translation, *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp.144-151, 2003.

[73] Y. Al-Onaizan and K. Papineni, Distortion models for statistical machine translation, *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.529-536, 2006.

[74] C. Tillman and T. Zhang, A localized prediction model for statistical machine translation, *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, pp.557-564, 2005.

[75] C. Tillmann, A unigram orientation model for statistical machine translation, *Proc. of HLT-NAACL 2004: Short Papers*, Boston, Massachusetts, pp.101-104, 2004.

[76] S. Kumar and W. Byrne, Local phrase reordering models for statistical machine translation, *Proc. of HLT-EMNLP 2005*, Vancouver, Canada, pp.161-168, 2005.

[77] M. Nagata, K. Saito, K. Yamamoto and K. Ohashi, A clustered global phrase reordering model for statistical machine translation, *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.713-720, 2006.

[78] D. Xiong, Q. Liu and S. Lin, Maximum entropy based phrase reordering model for statistical machine translation, *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.521-528, 2006.

[79] H. Setiawan, M.-Y. Kan and H. Li, Ordering phrases with function words, *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp.712-719, 2007.

[80] D. Xiong, M. Zhang, A. Aw and H. Li, A linguistically annotated reordering model for BTG-based statistical machine translation, *Proc. of ACL-08: HLT, Short Papers*, Columbus, Ohio, pp.149-152, 2008.

[81] Y. Ni, C. Saunders, S. Szedmak and M. Niranjan, Handling phrase reorderings for machine translation, *Proc. of the ACL-IJCNLP 2009 Conf. Short Papers*, Suntec, Singapore, pp.241-244, 2009.

[82] C.-H. Li, M. Li, D. Zhang, M. Li, M. Zhou and Y. Guan, A probabilistic approach to syntax-based reordering for statistical machine translation, *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp.720-727, 2007.

[83] P.-C. Chang and K. Toutanova, A discriminative syntactic word order model for machine translation, *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp.9-16, 2007.

[84] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete Data via the EM algorithm, *Journal of the Royal Statistical Society-Part B*, vol.39, no.1, pp.1–38, 1977.

[85] K. Knight and Y. Al-Onaizan, Translation with finite-state devices, *Proc. of the 3rd Conf. of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, Langhorne, PA, pp.421-437, 1998.

[86] J. Wuebker, A. Mauser and H. Ney, Training phrase translation models with leaving-one-out, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.475-484, 2010.

[87] M. Zhang and H. Li, Tree kernel-based SVM with structured syntactic knowledge for BTG-based phrase reordering, *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing*, Singapore, pp.698-707, 2009.

[88] M. Zhang, H. Zhang and H. Li, Convolution kernel over packed parse forest, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.875-885, 2010.

[89] J. N. Darroch and D. Ratcliff, Generalized iterative scaling for long-linear models, *Annals of Mathematical Statistics*, vol.43, no.5, pp.1470-1480, 1972.

[90] F. J. Och, Minimum error rate training in statistical machine translation, *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp.160-167, 2003.

[91] P. Liang, A. Bouchard-Côté, D. Klein and B. Taskar, An end-to-end discriminative approach to machine translation, *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.761~768, 2006.

[92] C. Tillmann and T. Zhang, A discriminative global training algorithm for statistical MT, *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.721-728, 2006.

[93] A. Fraser and D. Marcu, Semi-supervised training for statistical word alignment, *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.769-776, 2006.

[94] B. Zhao, S. Vogel, M. Eck and A. Waibel, Phrase pair rescoring with term weighting for statistical machine translation, Proc. of EMNLP, Barcelona, Spain, pp.206-213, 2004.

[95] Y. Deng, J. Xu and Y. Gao, Phrase table training for precision and recall: what makes a good phrase and a good phrase pair? *Proc. of ACL-08: HLT*, Columbus, Ohio, pp.81-88, 2008.

[96] K. Duh and K. Kirchhoff, Beyond log-linear models: boosted minimum error rate training for n-best re-ranking, *Proc. of ACL-08: HLT, Short Papers*, Columbus, Ohio, pp.37-40, 2008.

[97] S. Kumar, W. Macherey, C. Dyer and F. Och, Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices, *Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int. Joint Conf. on Natural Language Processing of the AFNLP*, Suntec, Singapore, pp.163-171, 2009.

[98] A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. R. Gillett, J. D. Lafferty, R. L. Mercer, H. Printz, L. Ureš, The Candide system for machine translation, *Proc. Of ARPA Workshop on Human Language Technology*, Plainsboro, NJ, pp.157-162, 1994.

[99] A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, A. S. Kehler and R. L. Mercer, Language translation apparatus and method using context-based translation models, U.S. Patent, No. 5510981, 1996.

[100] C. Tillmann, S. Vogel, H. Ney and A. Zubiaga, A DP-based search using monotone alignments in statistical translation, Proc. of the 35th Annual Conf. of the Association for Computational Linguistics, Madrid, Spain, pp.289-296, 1997.

[101] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga and H. Sawaf, Accelerated DP-based search for statistical translation, *European Conf. on Speech Communication and Technology*, Rhodes, Greece, pp.2667-2670, 1997.

[102] C. Tillmann and H. Ney, Word re-ordering and DP-based search in statistical machine translation, *Proc. of the 18th Int. Conf. on Computational Linguistics, Saarbrucken, Germany*, pp.850-856, 2000.

[103] C. Tillmann, *Word Re-Ordering and Dynamic Programming based Search Algorithms for Statistical Machine Translation*, Ph.D. thesis, RWTH Aachen, 2001.

[104] F. J. Och, N. Ueffing and H. Ney, An efficient A* search algorithm for statistical machine translation, *Data-Driven Machine Translation Workshop*, Toulouse, France, pp.55-62, 2001.

[105] U. Germann, M. Jahr, K. Knight, D. Marcu and K. Yamada, Fast decoding and optimal decoding for machine translation, *Proc. of 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, pp.228-235, 2001.

[106] U. Germann, M. Jahr, K. Knight, D. Marcu and K. Yamada, Fast and optimal decoding for machine translation, *Artificial Intelligence*, vol.154, issue 1-2, pp.127-143, 2004.

[107] U. Germann, Greedy decoding for statistical machine translation in almost linear time, *Proc. of HLT-NAACL 2003*, Edmonton, Canada, 2003.

[108] I. García-Varea, F. Casacuberta and H. Ney, An iterative, DP-based search algorithm for statisticalmachine translation, *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, pp. 1235-1238, 1998.

[109] I. García-Varea and F. Casacuberta, Search algorithms for statistical machine translation based on dynamic programming and pruning techniques, *Proc. of Machine Translation Summit VIII*, Santiago de Compostela, Spain, pp.115-120, 2001.

[110] S. Riedel, J. Clarke, Revisiting optimal decoding for machine translation IBM model 4, *Proc. of the 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics, Companion Volume*, Boulder, Colorado, pp.5-8, 2009.

[111] K. Yamada and K. Knight, A decoder for syntax-based statistical MT, *Proc. of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, pp.303-310, 2002.

[112] S. Kumar and W. Byrne, Minimum bayes-risk decoding for statistical machine translation, *Proc. of HLT-NAACL 2004*, Boston, Massachusetts, pp.169-176, 2004.

[113] R. W. Tromble, S. Kumar, F. Och and W. Macherey, Lattice minimum bayes-risk decoding for statistical machine translation, *Proc. of the 2008 Conf. on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, pp.620-629, 2008.

[114] S. Bangalore, G. Bordel and G. Riccardi, Computing consensus translation from multiple machine translation systems, *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy, pp.351-354, 2001.

[115] E. Matusov, N. Ueffing and H. Ney, Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment, *Proc. of the 11th Conf. of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, pp.33-40, 2006.

[116] A.-V. Rosti, N. F. Ayan, B. Xiang, S. Matsoukas, R. Schwartz and B. Dorr, Combining outputs from multiple machine translation systems, *Human Language Technologies 2007: The Conf. of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, pp.228-235, 2007.

[117] K. C. Sim, W. J. Byrne, M. J. F. Gales, H. Sahbi and P. C. Woodland, Consensus network decoding for statistical machine translation system combination, *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Honolulu, Hawaii, pp.105-108, 2007.

[118] Y. Chen, A. Eisele, C. Federmann, E. Hasler, M. Jellinghaus and S. Theison, Multi-engine machine translation with an open-source SMT decoder, *Proc. of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp.193-196, 2007.

[119] M. Li, N. Duan, D. Zhang, C.-H. Li and M. Zhou, Collaborative decoding: partial hypothesis re-ranking using translation consensus between decoders, *Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int. Joint Conf. on Natural Language Processing of the AFNLP*, Suntec, Singapore, pp.585-592, 2009.

[120] T. Xiao, J. Zhu, M. Zhu and H. Wang, Boosting-based system combination for machine translation, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.739-748, 2010.

[121] C. Tillmann, A rule-driven dynamic programming decoder for statistical MT, *Proc. of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, Columbus, Ohio, pp.37-45, 2008.

[122] C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman and P. Resnik, Cdec: a decoder, alignment, and learning framework for finite-state and context-free translation models, *Proc. of the ACL 2010 System Demonstrations*, Uppsala, Sweden, pp.7-12, 2010.

[123] Z. Li and S. Khudanpur, A scalable decoder for parsing-based machine translation with equivalent language model state maintenance, *Proc. of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, Columbus, Ohio, pp.10-18, 2008.

[124] T. Watanabe and E. Sumita, Bidirectional decoding for statistical machine translation, *Proc. of the 19th Int. Conf. on Computational Linguistics*, Taipei, Taiwan, pp.1-7, 2002.

[125] T. Watanabe, H. Tsukada and H. Isozaki, Left-to-right target generation for hierarchical phrase-based translation, *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.777-784, 2006.

[126] B. Zhao and S. Chen, A simplex armijo downhill algorithm for optimizing statistical machine translation decoding parameters, *Proc. of Human Language Technologies: The 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Boulder, Colorado, pp.21-24, 2009.

[127] P. Deshpande, R. Barzilay and D. Karger, Randomized decoding for selection-and-ordering problems, *Human Language Technologies 2007: The Conf. of the North American Chapter of the Association for Computational Linguistics; Proc. of the Main Conf.*, Rochester, New York, pp.444-451, 2007.

[128] M. Pust and K. Knight, Faster MT decoding through pervasive laziness, *Proc. of Human Language Technologies: The 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, pp.141-144, 2009.

[129] A. Arun, C. Dyer, B. Haddow, P. Blunsom, A. Lopez and P. Koehn, Monte Carlo inference and maximization for phrase-based translation, *Proc. of the Thirteenth Conf. on Computational Natural Language Learning*, Boulder, Colorado, pp.102-110, 2009.

[130] N. Bach, S. Vogel and C. Cherry, Cohesive constraints in a beam search phrase-based decoder, *Proc. of Human Language Technologies: The 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, pp.1-4, 2009.

[131] C. Cherry, Cohesive phrase-based decoding for statistical machine translation, *Proc. of ACL-08: HLT*, Columbus, Ohio, pp.72-80, 2008.

[132] Y. Marton and P. Resnik, Soft syntactic constraints for hierarchical phrased-based translation, *Proc. of ACL-08: HLT*, Columbus, Ohio, pp.1003-1011, 2008.

[133] D. Xiong, M. Zhang and H. Li, Learning translation boundaries for phrase-based decoding, *Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, pp.136-144, 2010.

[134] H. Cao and E. Sumita, Filtering syntactic constraints for statistical machine translation, *Proc. of the ACL 2010 Conf. Short Papers*, Uppsala, Sweden, pp.17-21, 2010.

[135] K. Y. Su and J. S. Chang, A customizable, self-learnable parameterized MT system: the next generation, MT SUMMIT VII, Singapore, pp.182-188, 1999.

[136] B. Chen, G. Foster and R. Kuhn, Bilingual sense similarity for statistical machine translation, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.834-843, 2010.

[137] M. Carpuat and D. Wu, Word sense disambiguation vs. statistical machine translation, *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan, pp.387-394, 2005.

[138] Y. S. Chan, H. T. Ng and D. Chiang, Word sense disambiguation improves statistical machine translation, *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp.33-40, 2007.