Meta-Structure Transformation Model for Statistical Machine Translation

Jiadong Sun, Tiejun, Zhao and **Huashen Liang**
MOE-MS Key Lab of National Language Processing and speech
Harbin Institute of Technology
No. 92, West Da-zhi Street ,Harbin Heilongjiang ,150001 ,China
jiadongsun@hit.edu.cn
{tjzhao, hsliang }@mtlab.hit.edu.cn

Abstract

We propose a novel syntax-based model for statistical machine translation in which meta-structure (**MS**) and meta-structure sequence (**SMS**) of a parse tree are defined. In this framework, a parse tree is decomposed into **SMS** to deal with the structure divergence and the alignment can be reconstructed at different levels of recombination of **MS** (**RM**). **RM** pairs extracted can perform the mapping between the substructures across languages. As a result, we have got not only the translation for the target language, but an **SMS** of its parse tree at the same time. Experiments with BLEU metric show that the model significantly outperforms Pharaoh, a state-art-the-art phrase-based system.

1 Introduction

The statistical approach has been widely used in machine translation, which use the noisy-channel-based model. A joint probability model, proposed by Marcu and Wong (2002), is a kind of phrase-based one. Och and Ney (2004) gave a framework of alignment templates for this kind of models. All of the phrase-based models outperformed the word-based models, by automatically learning word and phrase equivalents from bilingual corpus and reordering at the phrase level. But it has been found that phrases longer than three words have little improvement in the performance (Koehn, 2003). Above the phrase level, these models have a simple distortion model that reorders phrases independently, without consideration of their contents and syntactic information.

In recent years, applying different statistical learning methods to structured data has attracted various researchers. Syntax-based MT approaches began with Wu (1997), who introduced the Inversion Transduction Grammars. Utilizing syntactic structure as the channel input was introduced into MT by Yamada (2001). Syntax-based models have been presented in different grammar formalisms. The model based on Head-transducer was presented by Alshawi (2000). Daniel Gildea (2003) dealt with the problem of the parse tree isomorphism with a cloning operation to either tree-to-string or tree-to-tree alignment models. Ding and Palmer (2005) introduced a version of probabilistic extension of Synchronous Dependency Insertion Grammars (**SDIG)** to deal with the pervasive structure divergence. All these approaches don't model the translation process, but formalize a model that generates two languages at the same time, which can be considered as some kind of tree transducers. Graehl and Knight (2004) described the use of tree transducers for natural language processing and addressed the training problems for this kind of transducers.

In this paper, we define a model based on the **MS** decomposition of the parse trees for statistical machine translation, which can capture structural variations and has a proven generation capacity. During the translation process of our model, the parse tree of the source language is decomposed into different levels of **MS** and then transformed into the ones of the target language in the form of **RM**. The source language can be reordered according to the structure transformation. At last, the target translation string is generated in the scopes of **RM**. In the framework of this model,

Proceedings of the Second Workshop on Statistical Machine Translation, pages 64–71,
Prague, June 2007. ©2007 Association for Computational Linguistics

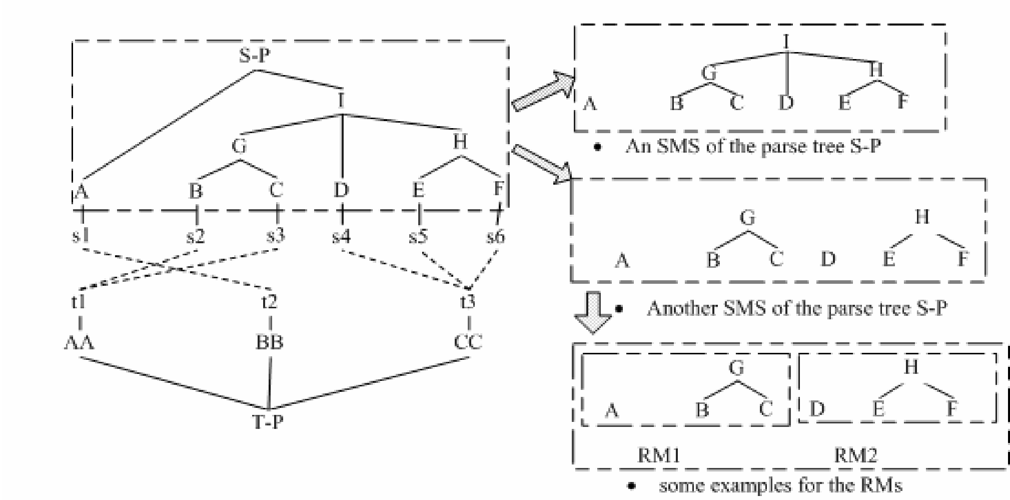

Figure 1: MS and the SMS and RM for a given parser tree

the **RM** transformation can be regarded as production rules and be extracted automatically from the bilingual corpus. The overall translation probability is thus decomposed.

In the rest of this paper, we first give the definitions for **MS**, **SMS**, **RM** and the decomposition of the parse tree in section 2.1, we give a detailed description of our model in section 2.2, section 3 describes the training details and section 4 describes the decoding algorithms, and then the experiment (section 5) proves that our model can outperform the baseline model, pharaoh, under the same condition.

2 The model

2.1 MS for a parse tree

A source language sentence (s1 s2 s3 s4 s5 s6), and its parse tree S-P, are given in Figure 1.We also give the translation of the sentence, which is illustrated as (t1 t2 t3).Its parse tree is T-P.

Definition 1

MS of a parse tree

We call a sub-tree a **MS** of a parse tree, if it satisfies the following constraints:

1. An **MS** should be a sub-tree of a parse tree
2. Its direct sons of the leaf nodes in the sub-tree are the words or punctuations of the sentence

For example, each of the sub-trees in the right-hand of Figure 1 is an **MS** for the parse tree of S-P. The sub-tree of [I [G, D, H]] of S-P is not an **MS**, because the direct sons of the leaf nodes, G, D, H, are not words in the sentence of (s1 s2 s3 s4 s5 s6).

Definition 2 SMS and RM

A sequence of **MS** is called a **meta-structure sequence (SMS)** of a parse tree if and only if,

1. Its elements are **MS** of the parse tree
2. The parse tree can be reconstructed with the elements in the same order as in the sequence.

It is denoted as **SMS [T(S)]**.[1] Two examples for the concept of **SMS** can be found in Figure1.

RM(recombination of MS) is a sub-sequence of **SMS**. We can express an **SMS** as different $RM_1^k[T(S)]$ .The parse tree of S-P in Figure1 is decomposed into **SMS** and expressed in the framework of **RM**. The two RM, $RM_1^2[S-P]$, are used to express its parse tree in Figure1.It is noted that there is structure divergence between the two parse trees in Figure1. The corresponding node of Node **I** in the tree S-P cannot be found in the tree T-P. But under the conception of **RM**, the structure alignments can be achieved at the level of **RM**, which is illustrated in Figure2.

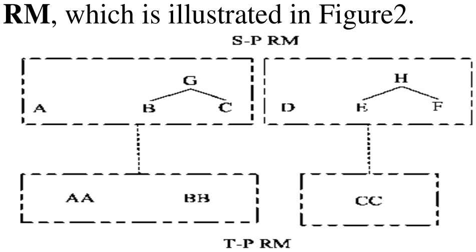

Figure2.The RM alignments for S-P and T-P

[1] **T[S]** denotes the parse tree of a given sentence
f and **e** denote the foreign and target sentences

In Figure2, both of the parse trees are decomposed and reconstructed in the forms of RM. The alignments based on RM are illustrated at the same time.

2.2 Description of the model

In the framework of Statistical machine translation, the task is to find the sentence **e** for the given foreign language **f**, which can be described in the following formulation.

$$\tilde{e} = \arg\max_{e}\{\ P(e \mid f)\} \qquad (1)$$

To make the model have the ability to model the structure transformation, some hidden variables are introduced into the probability equation. To make the equations simple to read, we take some denotations different from the above definitions. **SMS[T(S)]** is denoted as SM[T(S)].

The first variable is the **SM[T(S)]**, we induce the equation as follows,

$$\begin{aligned} P(e \mid f) &= \sum_{SM\ (T(f))} P(e, SM\ [T(f)] \mid f) \\ &= \sum_{SM[T(f)]} P(SM[T(f)] \mid f) P(e \mid SM[T(f)], f) \end{aligned} \qquad (2)$$

$$\begin{aligned} &P(e \mid SM[T(f)], f) = \\ &\sum_{\mathrm{SM}[T(e)]} P(e,\ \mathrm{SM}[T(e)] \mid SM[T(f)], f) \\ &= \sum_{\mathrm{SM}[T(e)]} P(\mathrm{SM}[T(e)] \mid SM[T(f)], f) \times \\ &\qquad P(e \mid \mathrm{SM}[T(e)], SM[T(f)], f) \end{aligned} \qquad (3)$$

In order to simplify this model we have two assumptions:

An assumption is that the generation of **SMS** [T (e)] is only related with SMS[T(f)]:

$$\begin{aligned} &P(SM[T(e)] \mid SM[T(f)], f) \\ &\quad \equiv P(SM\ [T(e)] \mid SM\ [T(f)]) \end{aligned} \qquad (4)$$

Here we do all segmentations for any **SMS** of **[T (f)]** to get different $RM_1^k[T(f)]$.

$$\begin{aligned} &P(SM[T(e)] \mid SM[T(f)]) = \\ &\sum_{RMT(f)]} \prod_{i=1}^{k} P(RM_i[T(e)] \mid RM_i[T(f)]) \end{aligned} \qquad (5)$$

The use of RM is to decompose bi-lingual parse trees and get the alignments in different hierarchical levels of the structure.

Now we have another assumption that all $P(SM[T(f)] \mid f)$ should have the same probability α. A simplified form for this model is derived:

$$\begin{aligned} &P(e \mid f) = \\ &\sum_{SM(T(f))} \sum_{SM(T(e))} \alpha \times \\ &\sum_{RM\ [T(f)]} \prod_{i=1}^{k} P(RM_i[T(e)] \mid RM_i[T(f)]) \\ &\times P(e \mid RM_i[T(e)], RM_i[T(f)], f) \end{aligned} \qquad (6)$$

, Where $P(e \mid RM_i[T(e)], RM_i[T(f)], f)$ can be regarded as a lexical transformation process, which will be further decomposed.

In order to model the direct translation process better by extending the feature functions, the direct translation probability is obtained in the framework of maximum entropy model:

$$\begin{aligned} &P(e \mid f) \\ &= \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(e, SM[T(e)], SM[T(f)], f)]}{\sum_{e, SM\ [T(e)], SM\ [T(f)]} \exp[\sum_{m=1}^{M} \lambda_m h_m(e, SM[T(e)], SM[T(f)], f)]} \end{aligned} \qquad (7)$$

We can achieve the translation according to the function below:

$$\tilde{e} = \operatorname{argmax}\{\exp[\sum_{m=1}^{M} \lambda_m h_m(e, SM[T(e)], SM[T(f)], f)\} \qquad (8)$$

A detailed list of the feature functions for the model and some explanations are given as below:

- Just as the derivation in the model, we take into consideration of the structure transformation when selecting the features. The **MS** are combined in the forms of RM and transformed as a whole structure.

$$h_1(e, f) = \log \prod_{i=1}^{k} P(RM_i[T(e)] \mid RM_i[T(f)]) \qquad (9)$$

$$h_2(e, f) = \log \prod_{i=1}^{k} P(RM_i[T(f)] \mid RM_i[T(e)]) \qquad (10)$$

- Features to model lexical transformation processes, and its inverted version, where the symbol **L (RM$_i$ [T(S)])** denotes the words belonging to this sub-structure in the sentence. In Figure1, **L (RM_1)** denotes the words, s1 s2 s3, in the source language. This part of transformation happens in the scope of each RM, which means that all the words in any RM can be transformed into the target language words just in the way of phrase-based model, serving as another reordering factor at a different level:

$$h_3(e,f) = \log\prod_{i=1}^{k} P(L(RM_i[T(e)])) \mid L(RM_i[T(f)]) \tag{11}$$

$$h_4(e,f) = \log\prod_{i=1}^{k} P(L(RM_i[T(f)])) \mid L(RM_i[T(e)]) \tag{12}$$

- We define a 3-gram model for the RM of the target language, which is called a structure model according to the function of it in this model.

$$h_5(e,f) = \log\prod_{i=1}^{k} P(RM_i[T(e)] \mid RM_{i-2}[T(e)], RM_{i-1}[T(e)]) \tag{13}$$

This feature can model the recombination of the parse structure of the target sentences. For example in Figure3, $P(CC \mid AA, BB)$ is used to describe the probability of the RM sequence, (AA, BB) should be followed by RM (CC) in the translation process. This function can ensure that a more reasonable sub-tree can be generated for the target language. That would be explained further in section 3.

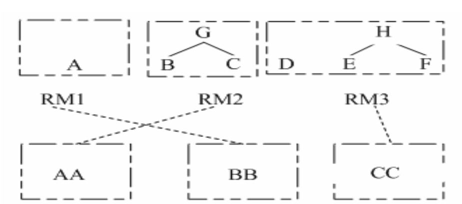

Figure3. The 3-gram structure model

- The 3-gram language model is also used

$$h_6(e,f) = \log P(e) \tag{14}$$

The phrase-based model (Koehn, 2003) is a special case of this framework, if we take the whole structure of the parse tree as the only MS of the parse tree of the sentence, and set some special feature weights to zero.

From the description above, we know the framework of this model. When transformed to target languages, the source language is reordered at the RM level first. In this process, only the knowledge of the structure is taken into consideration. It is obvious that a lot of sentences in the source language can have the same RM. So this model has better generative ability. At the same time, RM is a subsequence of **SMS**, which consists of different hierarchical MS. So RM is a structure, which can model the structure mapping across the sub-tree structure. By decomposing the source parse tree, the isomorphic between the parse trees can be obtained, at the level of RM.

When reordering at the RM level, this model just takes an RM as a symbol, and it can perform a long distance reordering job according to the knowledge of RM alignments.

3 Training

For training the model, a parallel tree corpus is needed. The methods and details are described as follows:

3.1 Decomposition of the parse tree

To reduce the amount of **MS** used in decoding and training, we take some constrains for the **MS**.

（1）.The height of the sub-tree shouldn't be greater than a fixed value α ;

（2）. $\dfrac{N(Leaf-nodes)}{N(height)} \geq \beta$

Given a parse tree, we get the initial **SMS** in such a top -down and left- to –right way.

Any node is deleted if the sub-tree can't satisfy the constrains (1), (2).

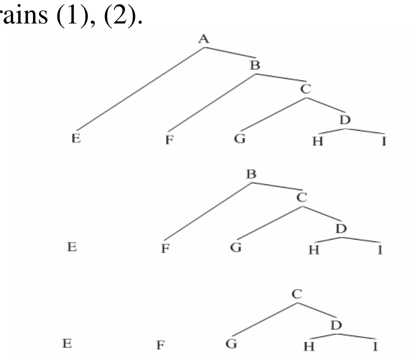

Figure3. Decomposition of a parse tree

RMS for Ch-Parse Tree	RMS for EN-Parse Tree	Pro for transformation
AP[AP[AP[a-a]-usde]-m]	NPB [DT-JJ-NN-PUNC.]	0.000155497
AP[AP[AP[r-a]-usde]-m]	NPB[PDT-DT-JJ-NN]	0.0151515
AP[AP[BMP[m-q]-a]-usde] wj	ADVP [RB-RB-PUNC.]	0.00344828
AP[AP[BMP[m-q]-a]-usde] wj	DT CD JJ NNS PUNC	0.0833333
AP[AP[BMP[m-q]-a]-usde] wj	DT JJ NN NNS PUNC.	0.015625

Table 1 some examples of the RM transformation

RM1	RM2	RM3	P(RM3|RM1,RM2)
IN	NP-A[NPB[PRP-NN]	IN	0.2479237
NPB	NP-A[NPB[PRP-NN]	VBZ	0.2479235
IN	NP-A[NPB[PRP-NN]	MD	0.6458637
<s>	NP-A[NPB[PRP-NN]	VBD	0.904308

Table 2 Examples for the 3-gram structure model of RM

Generate all of the **SMS** by deleting a node in any **Ms** to generate new **SMS**, applying the same operation to any **SMS**

3.2 Parallel SMS and Estimation of the parameters for RM transformations

We can get bi-lingual **SMS** by recombining all the possible **SMS** obtained from the parallel parse trees. $m*n$ Parallel **SMS** can be obtained if m is the number of **SMS** for a parse tree in the source language, n for the target one.

The alignments of the parallel **MS** and extraction can be performed in such a simple way. Given the parallel tree corpus, we first get the alignments based on the level of words, for which we used GIZA++ in both of the directions. According to the knowledge of the word alignments, we derived the alignments of leave nodes of the given parse trees, which are the direct root nodes of the words. Then all the knowledge of the words is discarded for the RM extraction. The next step for the extraction of the RM is based on the popular phrase-extraction algorithm of the phrase-based statistical machine translation model. The present alignment and phrase extraction methods can be applied to the extraction of the MS and RM [T(S)].

$$P(RM_{EI} \mid RM_{Fi}) = \frac{Count(RM_{Fi}, RM_{EI})}{\sum_{RM_{Ei}} Count(RM_{Fi}, RM_{Ei})}$$

$Count(A,B)$ is the expected number of times A is aligned with B in the training corpus.Table1 shows some parameters for this part in the model.

Training n-gram model for the monolingual structure model is based on the English RM of each parse tree, selected from the parallel tree corpus. The 3-gram structure model is defined as follows:

$$P(RM_I[T(e)] \mid RM_{i-2}[T(e)], RM_{i-1}[T(e)]) = \frac{Count(RM_{I-2}, RM_{I-1}, RM_I)}{\sum_j Count(RM_{I-2}, RM_{I-1}, RM_j)}$$

$Count(A,B,C)$ is the times of the situation, in which the RM is consecutive sub-trees of the parse trees in the training set. Some 3-gram parameters in the training task are given in Table2.

We didn't meet with the serious data sparseness problem in this part of work, because most of the MS structures have occurred enough times for parameters estimation. But we still set some fixed value for the unseen parameters in the training set.

4 Decoding

A beam search algorithm is applied to this model for decoding, which is based on the frame of the beam search for phrase-based statistical machine translation (Koehn et al, 03).

Here the process of the hypothesis generation is presented. Given a sentence and its parse tree, all the possible candidate **RM** are collected, which can cover a part of the parse tree at the bottom. With the candidates, the hypotheses can be formed and extended.

For example, all the parse tree's leaf nodes of a Chinese sentence in Figure4, are covered by **[r]**, **[pron]** and **VP[vg-BNP[pron-n]]** in the order of choosing candidate RM{ (1), (2), (3)}.

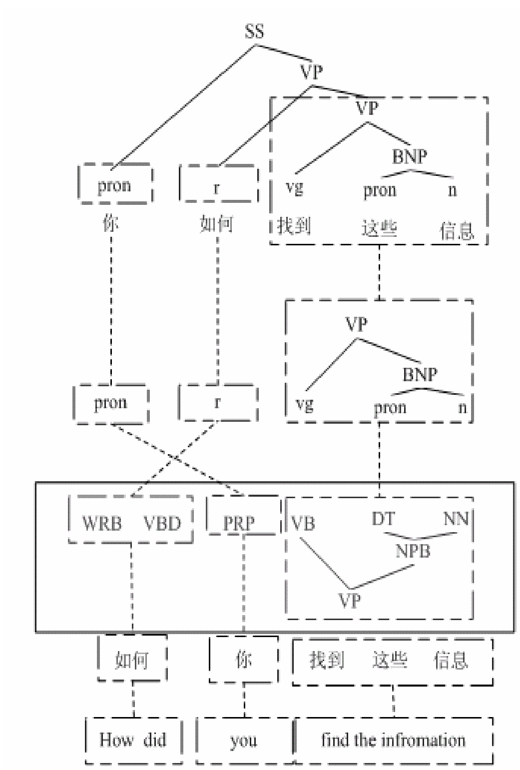

Figure4. Process of translation based on RM

$$(r, WRB \quad VBD) \tag{1}$$

如何 $\rightarrow$ how did

$$(pron, PRP) \tag{2}$$

你 $\rightarrow$ you

$$\begin{aligned}(&VP[vg - BNP[pron - n]],\\ &VP[VB - NPB[DT - NN]])\end{aligned} \tag{3}$$

得到 这些 信息 $\rightarrow$ find the information

Before the next expansion of a hypothesis, the words in the scope of the present RM are translated into the target language and the corresponding $RM_i\,[T(e)]$ is generated. For example, when $(r, WRB \quad VBD)$, is used to expand the hypothesis , the words in the sub-tree are translated into the target language, 如何 $\rightarrow$ how did.

We also need to calculate the cost for the hypotheses according to the parameters in the model to perform the beam search. The task for the beam search is to find the hypothesis with the least cost. When the expansion of a hypothesis comes to the final state, the target language is generated. All of the leave nodes of the parse tree for the source language are covered. The parser for the target language isn't used for decoding. But a target **SMS** is generated during the process of decoding to achieve better reordering performance.

5 Experiments

The experiment was conducted for the task of Chinese-to-English translation. A corpus, which consists of 602,701 sentence pairs, was used as the training set. We took CLDC 863 test set as our test set (http://www.chineseldc.org/resourse.asp), which consists of 467 sentences with an average length of 14.287 Chinese words and 4 references. To evaluate the result of the translation, the BLEU metric (Papineni et al. 2002) was used.

5.1 The baseline

System used for comparison was Pharaoh (Koehn et al., 2003; Koehn, 2004), which uses a beam search algorithm for decoding. In its model, it takes the following features: language model, phrase translation probability in the two directions, distortion model, word penalty and phrase penalty, all of which can be achieved with the training toolkits distributed by Koehn. The training set and development set mentioned above were used to perform the training task and to tune the feature weights by the minimum error training algorithm. All the other settings were the same as the default ones. SRI Language Modeling Toolkit was used to train a 3-gram language model. After training, 164 MB language model were obtained.

5.2 Our model

All the common features shared with Pharaoh were trained with the same toolkits and the same corpus. Besides those features, we need to train the structure transformation model and the monolingual structure model for our model. First, 10,000 sentence pairs were selected to achieve the

System	BLEU-n	n-gram precisions							
	4	1	2	3	4	5	6	7	8
Pharaoh	0.2053	0.6449	0.4270	0.2919	0.2053	0.1480	0.1061	0.0752	0.0534
Ms system	0.2232	0.6917	0.4605	0.3160	0.2232	0.1615	0.1163	0.0826	0.0587

Table3. Comparison of Pharaoh and our system

		Features						
System	$P_{lm}(e)$	P(RT)	P(IRT)	P_w(f|e)	P_w(e|f)	Word	Phr	Ph(RM)
Pharaoh	0.151	----	------	0.08	0.14	-0.29	0.26	-----
MS system	0.157	0.16	0.23	0.06	0.11	-0.20	0.22	0.36

Table4.Feature weights obtained by minimum error rate training on development set

training set for this part of task. The Collins parser and a Chinese parser of our own lab were used. After processing this corpus, we get a parallel tree corpus. SRI Language Modeling Toolkits were used again to train this part of parameters. In this experiment, we set $\alpha = 3$,and $\beta = 1.5$. 149MB RMS $[T(s)]$ pairs and a 25 MB 3-gram monolingual structure model were obtained.

6. Conclusion and Future work

A framework for statistical machine translation is created in this paper. The results of the experiments show that this model gives better performance, compared with the baseline system.

This model can incorporate the syntactic information into the process of translation and model the sub-structure projections across the parallel parse trees.

The advantage of this frame work lies in that the reordering operations can be performed at the different levels according to the hierarchical RM of the parse tree.

But we should notice that some independent assumptions were made in the decomposition of the parse tree. In the future, a proper method should be introduced into this model to achieve the most possible decomposition of the parse tree. In fact, we can incorporate some other feature functions into the model to model the structure transformation more effectively.

Acknowledgement

Thanks to the reviewers for their reviews and comments on improving our presentation of this paper.

References

A.P.Dempster,N.M.Laird, and D.B.Rubin 1977.Maximum likelihood from imcomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(Ser B):1-38.

Christoph Tillman. *A projection extension algorithm for statistical machine translation*. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, June 30-July 4, 2003, 1-8.

Daniel Gildea.2003.*Loosely tree based alignment for machine translation*. In Proceedings of ACL-03

Daniel Marcu, William Wong. *A phrase-based, joint probability model for statistical machine translation*. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, USA, July 11-13, 2002, 133-139.

Dekai Wu. 1997. *Stochastic inversion transduction grammars and bilingual parsing of parallel corpora.* Computational Linguistics, 23(3):3-403.

F.Casacuberta, E. Vidal: *Machine Translation with Inferred Stochastic Finite-state Transducers.* Computational Linguistics, Vol. 30, No. 2, pp. 205-225, June 2004

Franz J. Och, C. Tillmann, Hermann Ney. *Improved alignment models for statistical machine translation*. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP), College Park, MD, USA, June 21-22, 1999, 20-28.

Franz J. Och, Hermann Ney.2002 *Discriminative training and maximum entropy models*. In Proceedings of ACL-00, pages 440-447, Hong Kong, Octorber.

Hiyan Alshawi, Srinvas Bangalore, and Shona Douglas. 2000. *Learning dependency translation models as* *collections of finite state head transducers* Computational Linguistics, 26(1):45-60.

Ilya D. Melamed. *Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons.* Proceedings of the Third Workshop on Very Large Corpora, Boston, USA, July 30, 1995, 197-211.

Jonathan Graehl Kevin Knight *Training Tree Transducers* In Proceedings of NAACL-HLT 2004, pages 105-112.

Kenji Yamada and Kevin Knight 2001. *A Syntax-based statistical translation model*. In Proceedings of the 39th Annual Meeting of the association for computational Linguists(ACL 01), Toulouse, France, July 6-11

Michael John Collins. 1999. *Head-driven statistical Models for Natural Language Parsing*. Ph.D. thesis,University of Pennsyvania,Philadelphia.

P. Koehn, Franz Josef Och, Daniel Marcu. *Statistical phrase-based translation*. Proceedings of the Conference on Human Language Technology, Edmonton, Canada, May 27-June 1, 2003, 127-133.

P. Koehn: Pharaoh: *a Beam Search Decoder for Phrase-based Statistical Machine Translation Models* . Meeting of the American Association for machine translation(AMTA), Washington DC, pp. 115-124 Sep./Oct. 2004

Peter F. Brown ,Stephen *A. Della Pietra,Vincent J.Della Pietra, and Robert Merrcer.1993. The mathematics of statistical machine translation:Parameter estimation.*Computational Linguistics,19(2).:263-311.

Quirk, Chris, Arul Menezes, and Colin Cherry. *Dependency Tree Translation*. Microsoft Research Technical Report: MSR-TR-2004-113.

Regina Barzilay and Lillian Lee. 2003. *Learning to paraphrase: An supervised approach using multiple-sequence alignment.* In Proceedings of HLT/NAACL

S. Nie β en , H. Ney: Statistical Machine Translation with Scarce Resources using Morpho-syntactic Information. Computational Linguistics, Vol. 30 No. 2, pp. 181-204, June 20

Yuan Ding and Martha Palmer. 2005. *Machine translation using probabilistic synchronous dependency insert grammars.* In Proceedings of 43rd Annual Meeting of the NAACL-HLT2004, pages 273-280..