POS-based Reordering Models for Statistical Machine Translation

Deepa Gupta, Mauro Cettolo, Marcello Federico

FBK-irst, Centro per la Ricerca Scientifica e Tecnologica
via Sommarive 18, 38050 Povo di Trento, Italy
{gupta,federico,cettolo}@itc.it
http://hermes.itc.it

Abstract

We present a novel word reordering model for phrase-based statistical machine translation suited to cope with long-span word movements. In particular, reordering of nouns, verbs and adjectives is modeled by taking into account target-to-source word alignments and the distances between source as well as target words. The proposed model was applied as a set of additional feature functions to re-score N-best translation candidates generated by a statistical machine translation system featuring state-of-the-art lexicalized reordering models. Experiments showed relative BLEU score improvement up to 7.3% on the BTEC Japanese-to-English task, and up to 1.1% on the Europarl German-to-English task.

1. Introduction

In machine translation (MT), one of the main problems to handle is word reordering. A word is "reordered" when it and its translation occupy different positions within the corresponding sentence. In Statistical MT (SMT) (Brown et al., 1993), word reordering is faced from two points of view: constraints and modeling. If arbitrary word-reorderings are permitted, the exact decoding problem is NP-hard (Knight, 1999); it can be made polynomial-time by introducing proper constraints, such as IBM constraints (Berger et al., 1996a) and Inversion Transduction Grammars (ITG) constraints (Wu, 1997). Among all the allowed word-reorderings, it is expected that some are more likely than others. The aim of reordering models, known also as distortion models, is that of providing a measure of the plausibility of word movements. Most of the distortion models developed so far are unable to exploit linguistic context to score reorderings: they just predict target positions on the basis of other (source and target) positions.

A few years ago SMT moved from words to phrases as basic units of translation. Phrases are sequences of words, not necessarily with a syntactic meaning, that allow to model local reorderings, short idioms, insertions and deletions that are sensitive to local context. They are a simple mechanism but powerful enough to really improve performance (Koehn et al., 2003; Och and Ney, 2004). Nevertheless, they are able to capture only local phenomena. In (Chiang, 2005) an interesting extension toward hierarchical phrases was proposed, which allows one to predict long-span reordering phenomena, too.

In this work we present a novel word reordering model. In particular, our goal is to model reorderings concerning three major part-of-speech (POS) classes, namely nouns, verbs and adjectives. Relevant statistics are collected from word-aligned parallel texts regarding the distance between target words and the distance between the corresponding source words. The model was applied as a set of additional feature functions for re-scoring N-best lists generated by a phrase-based SMT system.

The paper is organized as follows. Section 2 highlights some relevant and typical reordering phenomena occurring between German and English, two languages which often show significant word movements. Section 3 encompasses an overview of major approaches to the problem of word reordering. Section 4 briefly introduces our phrase-based SMT system. Section 5 presents our novel reordering model. Then, in Section 6 experiments on the BTEC Japanese-to-English task and on the Europarl German-to-English task are described and results are discussed. Finally, some conclusion are drawn in Section 7.

original German sentence:
in wien gab es eine große konferenz .

literal English translation :
in vienna was held a major conference .

reordered English sentence:
a major conference was held in vienna

Figure 1: German to English translation example.

2. Example of Word Reordering

In many cases, German and English show very different word orders. Consider the example reported in Figure 1. If the original German sentence (first entry) is translated word by word into English, the result is the string of the second entry. Some word movements (underlined) are required to get the syntactically correct version of the English sentence (see third row). In particular, a swap of the position of the constituents "in vienna" and "a major conference" is observed.

The phenomenon occurring here is due to the fact that in English the verb follows the subject, while in German the case is the opposite. This is only a simple example, but the characteristics of the two languages often yield long-distance word movements.

In order to capture such aspects of the translation in a general manner, a phrase-based system should be enhanced by means of effective distortion models. In the following section, a brief overview of the most significant previous attempts of attacking the reordering problem is given, together with a discussion of the advantages our approach should have over them.

3. Related Work

One of the main research areas in SMT is word/phrase reordering models. Many reordering models have recently been proposed in the literature. The simplest but effective way to capture movements of target phrases is the use of a relative distortion probability distribution $d(a_i, b_{i-1})$, where a_i denotes the start position of the source phrase that is translated into the i-th target phrase, while b_{i-1} denotes the end position of the source phrase translated into the $i-1$-th target phrase. Systems described in (Och and Ney, 2004; Koehn et al., 2003; Federico and Bertoldi, 2005), and many others, adopt this strategy.

In (Och et al., 2004; Tillmann, 2004; Tillmann and Zhang, 2005), reordering models work on the concept of block, which is a pair of source and target phrases. Each block is associated with an orientation with respect to its predecessor block. During decoding, the probability of a sequence of blocks with the corresponding orientations is computed.

Many recent papers on reordering models are inspired by the block orientation idea introduced by Tillman, like (Kumar and Byrne, 2005; Zens and Ney, 2006; Xiong et al., 2006; Nagata et al., 2006; Al-Onaizan and Papineni, 2006).

In (Kumar and Byrne, 2005) the block orientation is implemented through weighted finite state transducers. Unfortunately, that model cannot capture all possible phrase movements.

Discriminative lexicalized reordering models are presented in (Zens and Ney, 2006). Several types of features are tested: word-based, word class-based, POS-based and based on local context.

Also (Xiong et al., 2006) exploit a discriminative model to predict reordering of consecutive blocks. Two kinds of reorderings are considered: straight and inverted. Any block reordering is allowed, no matter whether it was observed in training or not.

A global reordering model is presented in (Nagata et al., 2006) that explicitly models long distance reordering. It predicts four types of reordering patterns: monotone adjacent, monotone gap, reverse adjacent and reverse gap. By collapsing into the same neutral class monotone gaps and reverse gaps, it models only three possible events similarly to local reordering models (Tillmann and Zhang, 2005).

The distortion model proposed in (Al-Onaizan and Papineni, 2006) assigns a probability distribution over possible relative jumps conditioned on source words. It consists of three components: outbound, inbound and pair distortion. The model's parameters are directly estimated from word alignments.

In (Lee and Roukos, 2004) and (Lee, 2006), the aim is to capture particular syntactic phenomena occurring in the source language which are not preserved by the target language. POS rules are applied for preprocessing the source side both in translation model training and in decoding.

All models referred to above were tested on different language pairs, including Arabic, Chinese, English, German and Japanese languages.

Apart Chinese, which is typologically inconsistent (Newmeyer, 2004), each one of other languages has its own grammatical properties which are peculiar but nevertheless comparable. Hence, the reordering model we propose in this work tries to exploit the "grammatical compatibility" between source and target languages. In fact, we try to model the movements of three major part of speech classes (verbs, nouns and adjectives), looking at where the words translated so far are located. Our model considers the reorderings from the target language point of view, namely English. Moreover, differently from what can happen in lexicalized models, our model does not suffer from data sparseness, since statistics are collected for POS classes instead of plain words.

4. The Phrase-based SMT System

Given a string $\mathbf{f}$ in the source language, the goal of SMT is to select the string $\mathbf{e}$ in the target language which maximizes the posterior distribution $\Pr(\mathbf{e} \mid \mathbf{f})$. In phrase-based translation, words are no longer the only units of translation, but they are complemented by strings of consecutive words, the phrases. By assuming a log-linear model (Berger et al., 1996b; Och and Ney, 2002), the optimal translation can be searched for by exploiting a set of *feature functions*, designed to model different aspects of the translation process.

Our translation system works in two steps. In the first stage, the beam search decoder available in `Moses` (Koehn et al., 2007),[1] computes an N-best list of translations. `Moses` is an open source toolkit for statistical machine translation which includes, besides the decoder, tools for training translation and lexicalized reordering models, and a minimum error training procedure for estimating optimal interpolation weights.

In the second stage, the N-best translations are re-scored by applying additional feature functions and re-ranked: the top-ranked translation is finally output. The log-linear models used in both steps have interpolation parameters which are estimated from a development set by applying a minimum error training procedure (Och, 2003).

The reordering model presented in the following section is the only additional feature function applied for re-scoring the N-best lists.

5. The POS-based Reordering Model

We assume that we have a parallel training corpus provided with inverted word alignments, that is alignments from target to source positions. Let $(\mathbf{f}, \mathbf{e})$ be a source-target sentence pair, and let $\mathbf{a}$ be an inverted alignment which maps target positions i into source positions $a_i = j$.

For any target position i, we look for its predecessor i^* that is aligned to the rightmost source position. Our interest is indeed in the difference between the two positions, denoted by Δ_i. Formally:

$$\Delta_i = \begin{cases} a_i - a_{i^*} & \text{if } i > 1 \\ 1, & \text{if } i = 1 \end{cases} \qquad i^* = \arg\max_{w<k<i} a_k$$

where w denotes the window size. By setting w to zero, i^* is searched among all the positions covered so far.

Intuitively, Δ_i is negative when some word reordering occurred: namely when some source position following a_i has

[1] http://www.statmt.org/moses/

i	1	2	**3**	4	5	6	*7*
e_i	a\DT	major\JJ	**conference\NN**	was\VBD	held\VBN	in\IN	*vienna\NN*
$j = a_i$	5	6	**7**	3	3	1	*2*
f_j	eine	große	konferenz	gab	gab	in	wien
	original German sentence :	in wien gab es eine große konferenz					

Figure 2: Example of English-to-German word alignment.

i	1	2	3	**4**	5	6	7	*8*
e_i	we\PRP	have\VBP	not\RB	**done\VBN**	enough\RB	in\IN	that\DT	*sector\NN*
$j = a_i$	5	4	6	**8**	7	1	2	*3*
f_j	wir	haben	nicht	getan	genug	in	diesem	bereich
	original German sentence :	in diesem bereich haben wir nicht genug getan						

Figure 3: Example of English-to-German word alignment.

been already covered. The value corresponds to the amount of movement relative to a_i. When Δ_i is positive, then the source word covered by e_i was not anticipated by any of its following words. The value corresponds to the distance between a_i and its closest covered position.

In this work we focused our attention on the behavior of target words belonging to one of three major POS classes: verb (V), noun (N) and adjective (A). Reordering statistics of POS classes were obtained by POS tagging the target (English) side of the aligned corpus. Table 1 provides for each class the corresponding tags used by the POS tagger.[2]

Part of Speech	POS Tag
Verb(V)	MD, VB, VBD, VBG VBN, VBP, VB
Noun(N)	NN, NNS, NNP
Adjective(A)	JJ, JJR, JJS

Table 1: Working POS tag set.

Consider again the example introduced in Figure 1. Figure 2 details both the alignment and the tagging of the target side. The English word *vienna\NN* at position 7, tagged as noun, is aligned to the second word of the German sentence. Assuming $w = 0$, the highest alignment before $vienna$ is 7, which corresponds to the word *conference*. Hence, Δ_7=2-7=-5. This indicates that the position covered by $wien$ was anticipated by a higher position at distance 5.

Examples of Δ_i distributions for the considered POS classes of e_i are shown in Figure 5. Statistics were computed on a parallel Japanese-to-English corpus.

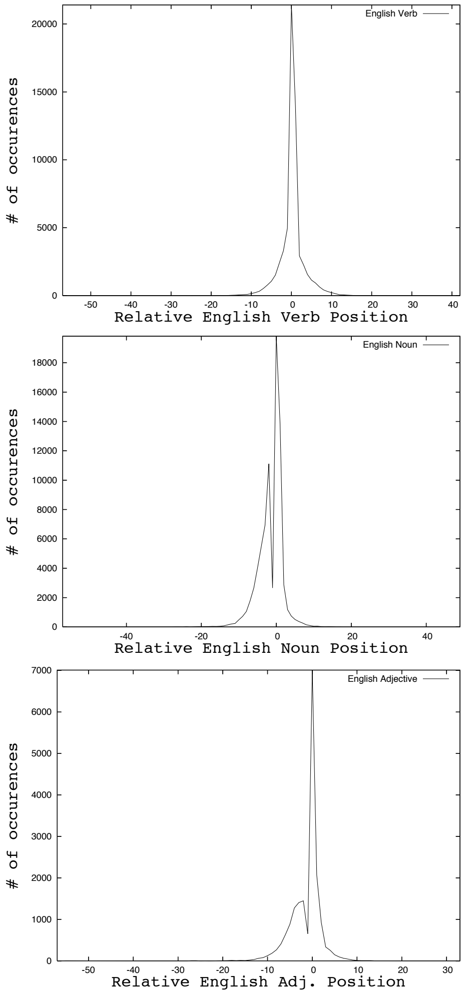

Figure 5: Δ distributions of English verb/noun/adj.

The statistics discussed so far just depend on the class of e_i. A more detailed model can be obtained by also taking into account the POS class of e_{i^*}. As an example, consider in Figure 3 the English word *sector\NN* at position 8, and in Figure 4 the English word *president\NN* at position 7. Both words are tagged as NN (noun). According to the proposed reordering model definition, Δ_i's for these two positions have the same value, namely -5. Hence, in order to distinguish the observations, the tag information corresponding to i^* is also used. In addition, the distance $d_i = i - i^*$ between the two target positions is also considered. Notice that while the POS class for i is restricted to nouns, verbs and adjectives, any of the possible 32 POS tags provided by

[2] http://www.lsi.upc.es/∼nlp/SVMTool/

i	1	2	3	**4**	5	6	*7*
e_i	i\FW	prefer\VBP	to\TO	**wait\VB**	,\,	mr\NN	*president\NN*
$j = a_i$	4	6	0	**7**	0	1	*2*
f_j	ich	lieber		warten		herr	präsident
	original German sentence:	herr präsident , ich würde lieber warten					

Figure 4: Example of English to German word alignment.

our tagger is considered for target position i^*.

Statistics on Δ_i are hence collected by taking into account the target POS classes of the target words at positions i and i^*, and their distance, in shorthand g_i, g_i^*, and d_i. We will also use the notation Δ, g, g^*, d when the index i is not specified.

5.1. Model Definition

According to the plots of Figure 5,[3] Δ's are assumed to have a Normal distribution, as a first approximation. Then, for every distance d and pair of classes g and g^*, sample mean and variance of the Δ variable are computed on the aligned corpus as follows:

$$\hat{\mu}(g, g^*, d) = \frac{\sum_{\mathbf{f},\mathbf{e}} \sum_{i=1}^{|\mathbf{e}|} \Delta_i \delta(g_i, g)\delta(g_i^*, g^*)\delta(d_i, d)}{\sum_{\mathbf{f},\mathbf{e}} \sum_{i=1}^{|\mathbf{e}|} \delta(g_i, g)\delta(g_i^*, g^*)\delta(d_i, d)}$$

$$\hat{\sigma}(g, g^*, d) = \frac{\sum_{\mathbf{f},\mathbf{e}} \sum_{i=1}^{|\mathbf{e}|} (\Delta_i - \hat{\mu})^2 \delta(g_i, g)\delta(g_i^*, g^*)\delta(d_i, d)}{\sum_{\mathbf{f},\mathbf{e}} \sum_{i=1}^{|\mathbf{e}|} \delta(g_i, g)\delta(g_i^*, g^*)\delta(d_i, d)}$$

where $\delta(x, y) = 1$ if $x = y$ and 0 otherwise. Hence, once POS classes g, g^* and distance d are determined, a normalized value of Δ can be computed:

$$\Delta(g, g^*, d) = \frac{\Delta - \hat{\mu}(g, g^*, d)}{\hat{\sigma}(g, g^*, d)}$$

that is assumed to follow the standard normal distribution $\mathcal{N}(x; 0, 1)$.

Finally, distortion models for each of the three POS classes considered for g are computed through suitable feature functions. For instance the feature function for verbs is defined as follows:

$$h_V(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \frac{\sum_{i=1}^{l} \delta(g_i, V)\mathcal{N}(\Delta(g_i, g_i^*, d_i); 0, 1)}{\sum_{i=1}^{l} \delta(g_i, V)} \qquad (1)$$

The feature functions for the classes N and A are computed similarly. In equation 1, the score is normalized with respect to the number of occurrences of the considered POS tag. In fact, different entries of a given N-best list can contain a different number of words tagged with the same POS. Finally, as back-off score for never observed events, the density value of the lower limit of the .95 quantile of the standard Normal distribution is taken.

In order to test the proposed model, we have employed adjective, noun and verb models as additional features in the re-scoring stage of our SMT system. In order to compute model scores, word alignments are needed for each N-best entry. While the decoder returns alignment information at the phrase-level, word-level alignments were computed by refining such phrase-alignment via IBM Model 1 (Brown et al., 1993).

6. Experiment Settings and Results

6.1. Translation Tasks and Setup

Experiments were carried out on the Basic Traveling Expression Corpus (BETC) (Takezawa et al., 2002) and the Europarl task (Koehn, 2005). Details about the employed training, development and test sets are reported in Tables 2 and 3. BTEC is a multilingual corpus which contains tourism-related sentences similar to those that are found in phrase books. We worked on the Japanese-to-English translation direction. Experiments were performed on several evaluation sets, made available by the International Workshop of Spoken Language Translation (IWSLT). In particular, for each source sentence of those sets, 16 references are available with the exception of devset06 sources for which only 7 references are available.

Europarl data were used for testing our models on the German-to-English direction. The four available evaluation sets played the role of development and test sets.[4] Only one reference translation is available for each of them. The two test sets denoted as test06-in and test06-out in Table3 are the official evaluation sets of the 2006 NAACL shared task, namely the in-domain and out-of-domain evaluation sets, respectively.

Translation performance is reported in terms of case-insensitive BLEU% score and word error rate (WER). The latter is expected to capture well the quality of translations in terms of word reorderings.

The `Moses` decoder was run with the maximum reordering distance set to 6 and, among other models, a lexicalized reordering model trained specifying the option "orientation-bidirectional-fe" (Koehn et al., 2005).

In re-scoring experiments, for each Japanese sentence at most 1000-best (English) translation candidates were extracted, while for each German sentence at most 5000-best (English) translations were generated. The model weights of the log-linear interpolation were estimated on the corresponding development sets by optimizing a combination of BLEU and NIST scores.

[3] Actually, Δ distributions shown in the figure just depend on the class of the current target position i. Nevertheless, similar shapes are observed even if Δ's are made dependent on the POS class of the word at i^* and on the distance d_i=$i - i^*$.

[4] please refer the website of NAACL/HLT shared task 2006 for further details on data sets related to this task.

training set	#sentences	language	#words	dictionary size
BTEC	39,954	Jpn	472,702	12,667
		Eng	443,853	9851
Europarl	751,088	Ger	16,760,047	195,292
		Eng	17,554,825	65,889

Table 2: Statistics of training sets.

task	type	lang.	#sentences	#words	dictionary size
CSTAR03	dev	Jpn	506	5091	929
IWSLT04	test		500	5046	955
IWSLT05	test		506	5153	958
devset06	test		489	6818	1202
dev2006	dev	Ger	2000	55136	8790
devtest06	test		2000	54247	8660
test06-in	test		2000	55533	8807
test06-out	test		1064	26818	6303

Table 3: Statistics of development/test sets.

6.2. Results and discussion

Translation performance on development and test sets for Japanese-to-English and German-to-English tasks are provided in Tables 4 and 5, respectively. Experiments were carried out by setting the window size w to different values; best scores were obtained with window size 2 and 4 for the Japanese-to-English and German-to-English tasks, respectively.

set	system	BLEU	WER
CSTAR03	1-best	56.52	35.21
	re-scored	58.67	34.51
IWSLT04	1-best	50.83	38.83
	re-scored	51.29	38.12
IWSLT05	1-best	51.59	36.76
	re-scored	51.95	36.30
devset06	1-best	15.13	79.37
	re-scored	16.24	78.38

Table 4: Results for the Japanese-to-English task.

Rows “1-best” provide performance of the decoder. Rows “re-scored” refer to scores measured on the best translations found after N-best lists are re-scored using as additional features the verb, noun, and adjective reordering models.

The use of the proposed reordering models consistently improved the performance of the state-of-the-art SMT system which already exploits in decoding the really effective lexicalized reordering model called “orientation-bidirectional-fe” (Koehn et al., 2005).

In the Japanese-to-English task, absolute improvements of 0.46%, 0.36% and 1.11% BLEU scores were observed on the IWSLT04, IWSLT05 and devset06 test sets, respectively. On the German-to-English task, BLEU increased by 0.12% and 0.19% absolute on devtest06 and test06-out sets. There is a small degradation of BLEU on test06-in set, but a significant reduction of WER (67.42% to 66.99%). It is worth noticing that WER improved in all experiments. It is well known that translation improvements in word-reordering do not necessarily reflect on BLEU score improvements. In particular, the BLEU score is especially insensitive to word order changes as long as there are few matches of long n-grams between output and references. This seems to be especially true for our German-to-English task, for which BLEU score increments are quite limited or not observed at all. On the contrary, the WER measure is more sensitive to word movements, given that the match is computed by aligning the whole output string with each reference translation.

set	system	BLEU	WER
dev06	1-best	26.47	66.37
	re-scored	26.52	66.01
devtest06	1-best	25.74	67.21
	re-scored	25.86	66.80
test06-in	1-best	26.06	67.42
	re-scored	25.96	66.99
test06-out	1-best	17.61	75.34
	re-scored	17.80	74.64

Table 5: Results for the German-to-English task.

In conclusion, the fact that our method yields only small score improvements should not be too surprising. First, there is a lack of sensitivity of some metrics, as explained above; then, there is the fact that we are trying to improve over an already well performing distortion model. In fact, in previous experiments (not reported here) we obtained significantly better improvements by re-scoring N-best lists generated by a decoder with a plain distance-based distortion model (Koehn et al., 2003).[5] However, those improvements were also significantly smaller than those achieved by applying the lexicalized distortion model (available with the `Moses` decoder). Hence, to our view, the only correct way to proceed was to challenge the strongest available baseline.

6.3. Examples

Figure 6 compares some automatic Japanese-to-English translations generated by the decoder and re-scoring module. Interestingly, some reordering phenomena missed in decoding, even if the decoder exploits a really effective lexicalized reordering model, are properly captured by our model. Similarly, Figure 7 shows some examples taken from the German-to-English task, together with the gold reference translation. It can be noticed that the re-scoring stage outputs more fluent translations.

7. Conclusions

We have presented a novel POS-based reordering model, which regards three major classes, namely nouns, verbs and adjectives. Observed events involve the distance between target phrases and the distance between the corresponding source phrases; statistics are collected by exploiting target-to-source alignments.

[5] by the way the only one available in the `Pharaoh` decoder.

1-best	is on the third floor restaurant .
re-scored	the restaurant on the third floor .
1-best	is this the french wine very much
re-scored	this is is very famous french wine .
1-best	the money i already paid .
re-scored	i already paid the money .
1-best	a bottle of two bottles of whisky and brandy
re-scored	two bottles of whisky and one bottle of brandy
1-best	okay . see you pick up tomorrow , please .
re-scored	yes . please come and pick up again tomorrow .
1-best	can i have dinner ? in my room .
re-scored	can i have my meal in my room ?
1-best	which track it
re-scored	what track does it leave from ?
1-best	is better , to go by car .
re-scored	it's better to go by car .
1-best	do you have a friend of mine injured .
re-scored	my friend is injured .
1-best	what is the name this street ?
re-scored	what street is this ?
1-best	the tomorrow twenty-one me a birthday .
re-scored	tomorrow for my twenty-one birthday .

Figure 6: Reordering phenomena: examples of Japanese-to-English translations before and after re-scoring.

The model has been employed as additional feature function in the re-scoring stage of a SMT system. Experiments were reported on the BTEC corpus for the Japanese-to-English task and on the Europarl corpus for the German-to-English task. Results showed that the proposed reordering model is able to further improve performance of a decoder which already exploits a state-of-the-art lexicalized reordering model.

Acknowledgments

This work has been funded by the European Union under the integrated project TC-STAR- Technology and Corpora for Speech to Speech Translation-(IST-2002-FP6-506738. http://www.tc-star.org)

8. References

Y. Al-Onaizan and K. Papineni. 2006. Distortion Models for Statistical Machine Translation. In *Proc. of ACL*, Sydney, Australia.

A.L. Berger, P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, A.S. Kehler, and R.L. Mercer. 1996a. Language Translation Apparatus and Method Using Context-Based Translation Models. U.S. Patent 5,510,981.

A.L. Berger, S.A. Della Pietra, and V.J. Della Pietra. 1996b. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1).

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2).

D. Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of ACL*, Ann Arbor, MI.

M. Federico and N. Bertoldi. 2005. A Word-to-Phrase Statistical Translation Model. *ACM Transactions on Speech and Language Processing*, 2(2).

K. Knight. 1999. Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics*, 25(4).

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of HLT/NAACL*, Edmonton, Canada.

P. Koehn, A. Axelrod, A. Birch Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proc. of IWSLT*, Pittsburgh, PA.

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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL*, Prague, Czech Republic.

P. Koehn. 2005. A Parallel Corpus for Statistical Machine Translation. In *Proc. of MT Summit*, Phuket, Thailand.

S. Kumar and W. Byrne. 2005. Local Phrase Reordering Models for Statistical Machine Translation. In *Proc. of HLT-EMNLP*, Vancouver, Canada.

Y.-S. Lee and S. Roukos. 2004. IBM Spoken Language Translation System Evaluation. In *Proc. of IWSLT*, Kyoto, Japan.

Y.-S. Lee. 2006. Morpho-Syntax in Statistical Machine Translation. In *OpenLab*, Trento, Italy. http://tc-star.itc.it/openlab2006/.

M. Nagata, K. Saito, K. Yamamoto, and K. Ohashi. 2006. A Clustered Global Phrase Reordering Model for Statistical Machine Translation. In *Proc. of ACL*, Sydney, Australia.

F. Newmeyer. 2004. Word Order and Parameterized Grammars: A Critical Look. In *Johns Hopkins IGERT Workshop*, Baltimore, MD.

F.J. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of ACL*, Philadelphia, PA.

F.J. Och and H. Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4).

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. 2004. A Smorgasbord of Features for Statistical Machine Translation. In *Proc. of HLT/NAACL*, Boston, MA.

F.J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of ACL*, Sapporo, Japan.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In *Proc. of LREC*, Las Palmas, Spain.

C. Tillmann and T. Zhang. 2005. A Localized Prediction

1-best	in venezuela is a dangerous . halt
re-scored	venezuela is in a dangerous halt .
ref	venezuela is mired in a dangerous stalemate .
1-best	consolidation . reform is not , however ,
re-scored	consolidation , however , is not a reform .
ref	consolidation , however , is not reform .
1-best	new proposal is now before us . a green paper
re-scored	new proposal before us now is a green paper .
ref	the new proposal before us is for a green paper .
1-best	conflicts arising now rather than within the member states . between them
re-scored	conflicts arise within the member states now rather than between them .
ref	conflicts are more likely to arise within rather than between states .
1-best	cooperation will , i hope , on foreign policy . extend
re-scored	cooperation will hopefully also extend to the foreign policy .
ref	we are hoping that the cooperation will extend to external policy .
1-best	after the current estimates complaints every third inhabitants in europe . on noise
re-scored	after the current estimates every third inhabitants complaints about noise in europe .
ref	the commission now estimates that one in every three europeans complains about noise .
1-best	in both cases , the situation at the moment by the commission . monitored
re-scored	in both cases , the situation is currently monitored by the commission .
ref	the commission is currently monitoring the situation in both cases .
1-best	seems to me to be the concept of ivoritt quite justified . to be
re-scored	the concept of ivoritt seems to me to be totally justified .
ref	the concept of ivorian nationality would appear to me to be perfectly well founded .
1-best	issues with which we are concerned . technically complex and often
re-scored	the issues we deal with which are often complicated and technical .
ref	it is true that the subjects we are dealing with are sometimes complex and technical .

Figure 7: Reordering phenomena: examples of German-to-English translations before and after re-scoring.

Model for Statistical Machine Translation. In *Proc. of ACL*, Ann Arbor, MI.

C. Tillmann. 2004. A Unigram Orientation Model for Statistical Machine Translation. In *Companion Vol. of the Joint HLT and NAACL Conference*, Boston, MA.

D. Wu. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3).

D. Xiong, Q. Liu, and S. Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In *Proc. of ACL*, Sydney, Australia.

R. Zens and H. Ney. 2006. Discriminative Reordering Models for Statistical Machine Translation. In *Proc. of HLT-NAACL Workshop on SMT*, New York, NY.