Refinements in BTG-based Statistical Machine Translation

Deyi Xiong, Min Zhang, Aiti Aw
Human Language Technology
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
{dyxiong, mzhang, aaiti}@i2r.a-star.edu.sg

Haitao Mi, Qun Liu and Shouxun Lin
Key Lab of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
Beijing China, 100080
{htmi, liuqun, sxlin}@ict.ac.cn

Abstract

Bracketing Transduction Grammar (BTG) has been well studied and used in statistical machine translation (SMT) with promising results. However, there are two major issues for BTG-based SMT. First, there is no effective mechanism available for predicting orders between neighboring blocks in the original BTG. Second, the computational cost is high. In this paper, we introduce two refinements for BTG-based SMT to achieve better reordering and higher-speed decoding, which include (1) reordering heuristics to prevent incorrect swapping and reduce search space, and (2) special phrases with tags to indicate sentence beginning and ending. The two refinements are integrated into a well-established BTG-based Chinese-to-English SMT system that is trained on large-scale parallel data. Experimental results on the NIST MT-05 task show that the proposed refinements contribute significant improvement of 2% in BLEU score over the baseline system.

1 Introduction

Bracket transduction grammar was proposed by Wu (1995) and firstly employed in statistical machine translation in (Wu, 1996). Because of its good trade-off between efficiency and expressiveness, BTG restriction is widely used for reordering in SMT (Zens et al., 2004). However, BTG restriction does not provide a mechanism to predict final orders between two neighboring blocks.

To solve this problem, Xiong et al. (2006) proposed an enhanced BTG with a maximum entropy (MaxEnt) based reordering model (MEBTG). MEBTG uses boundary words of bilingual phrases as features to predict their orders. Xiong et al. (2006) reported significant performance improvement on Chinese-English translation tasks in two different domains when compared with both Pharaoh (Koehn, 2004) and the original BTG using flat reordering. However, error analysis of the translation output of Xiong et al. (2006) reveals that boundary words predict wrong swapping, especially for long phrases although the MaxEnt-based reordering model shows better performance than baseline reordering models.

Another big problem with BTG-based SMT is the high computational cost. Huang et al. (2005) reported that the time complexity of BTG decoding with m-gram language model is $O(n^{3+4(m-1)})$. If a 4-gram language model is used (common in many current SMT systems), the time complexity is as high as $O(n^{15})$. Therefore with this time complexity translating long sentences is time-consuming even with highly stringent pruning strategy.

To speed up BTG decoding, Huang et al. (2005) adapted the hook trick which changes the time complexity from $O(n^{3+4(m-1)})$ to $O(n^{3+3(m-1)})$. However, the implementation of the hook trick with pruning is quite complicated. Another method to increase decoding speed is cube pruning proposed by Chiang (2007) which reduces search space significantly.

In this paper, we propose two refinements to address the two issues, including (1) reordering heuris-

tics to prevent incorrect swapping and reduce search space using swapping window and punctuation restriction, and (2) phrases with special tags to indicate beginning and ending of sentence. Experimental results show that both refinements improve the BLEU score significantly on large-scale data.

The above refinements can be easily implemented and integrated into a baseline BTG-based SMT system. However, they are not specially designed for BTG-based SMT and can also be easily integrated into other systems with different underlying translation strategies, such as the state-of-the-art phrase-based system (Koehn et al., 2007), syntax-based systems (Chiang et al., 2005; Marcu et al., 2006; Liu et al., 2006).

The rest of the paper is organized as follows. In section 2, we review briefly the core elements of the baseline system. In section 3 we describe our proposed refinements in detail. Section 4 presents the evaluation results on Chinese-to-English translation based on these refinements as well as results obtained in the NIST MT-06 evaluation exercise. Finally, we conclude our work in section 5.

2 The Baseline System

In this paper, we use Xiong et al. (2006)'s system Bruin as our baseline system. Their system has three essential elements which are (1) a stochastic BTG, whose rules are weighted using different features in log-linear form, (2) a MaxEnt-based reordering model with features automatically learned from bilingual training data, (3) a CKY-style decoder using beam search similar to that of Wu (1996). We describe the first two components briefly below.

2.1 Model

The translation process is modeled using BTG rules which are listed as follows

$$A \to [A^1, A^2] \tag{1}$$

$$A \to \langle A^1, A^2 \rangle \tag{2}$$

$$A \to x/y \tag{3}$$

The lexical rule (3) is used to translate source phrase x into target phrase y and generate a block A. The two rules (1) and (2) are used to merge two consecutive blocks into a single larger block in a straight or inverted order.

To construct a stochastic BTG, we calculate rule probabilities using the log-linear model (Och and Ney, 2002). For the two merging rules (1) and (2), the assigned probability $Pr^m(A)$ is defined as follows

$$Pr^m(A) = \Omega^{\lambda_\Omega} \cdot \triangle_{p_{LM}(A^1,A^2)}^{\lambda_{LM}} \tag{4}$$

where Ω, the reordering score of block A^1 and A^2, is calculated using the MaxEnt-based reordering model (Xiong et al., 2006) described in the next section, λ_Ω is the weight of Ω, and $\triangle_{p_{LM}(A^1,A^2)}$ is the increment of language model score of the two blocks according to their final order, λ_{LM} is its weight.

For the lexical rule (3), it is applied with a probability $Pr^l(A)$

$$
\begin{aligned}
Pr^l(A) = \; & p(x|y)^{\lambda_1} \cdot p(y|x)^{\lambda_2} \cdot p_{lex}(x|y)^{\lambda_3} \\
& \cdot p_{lex}(y|x)^{\lambda_4} \cdot exp(1)^{\lambda_5} \cdot exp(|y|)^{\lambda_6} \\
& \cdot p_{LM}^{\lambda_{LM}}(y)
\end{aligned} \tag{5}
$$

where $p(\cdot)$ are the phrase translation probabilities in both directions, $p_{lex}(\cdot)$ are the lexical translation probabilities in both directions, $exp(1)$ and $exp(|y|)$ are the phrase penalty and word penalty, respectively and λs are weights of features. These features are commonly used in the state-of-the-art systems (Koehn et al., 2005; Chiang et al., 2005).

2.2 MaxEnt-based Reordering Model

The MaxEnt-based reordering model is defined on two consecutive blocks A^1 and A^2 together with their order $o \in \{straight, inverted\}$ according to the maximum entropy framework.

$$\Omega = p_\theta(o|A^1, A^2) = \frac{exp(\sum_i \theta_i h_i(o, A^1, A^2))}{\sum_o exp(\sum_i \theta_i h_i(o, A^1, A^2))} \tag{6}$$

where the functions $h_i \in \{0, 1\}$ are model features and θ_i are weights of the model features trained automatically (Malouf, 2002).

There are three steps to train a MaxEnt-based reordering model. First, we need to extract reordering examples from unannotated bilingual data, then generate features from these examples and finally estimate feature weights.

For extracting reordering examples, there are two points worth mentioning:

1. In the extraction of useful reordering examples, there is no length limitation over blocks compared with extracting bilingual phrases.

2. When enumerating all combinations of neighboring blocks, a good way to keep the number of reordering examples acceptable is to extract smallest blocks with the *straight* order while largest blocks with the *inverted* order .

3 Refinements

In this section we describe two refinements mentioned above in detail. First, we present fine-grained reordering heuristics using swapping window and punctuation restriction. Secondly, we integrate special bilingual phrases with sentence beginning/ending tags.

3.1 Reordering Heuristics

We conduct error analysis of the translation output of the baseline system and observe that Bruin sometimes incorrectly swaps two large neighboring blocks on the target side. This happens frequently when inverted order successfully challenges straight order by the incorrect but strong support from the language model and the MaxEnt-based reordering model. The reason is that only boundary words are used as evidences by both language model and MaxEnt-based reordering model when the decoder selects which merging rule (straight or inverted) to be used [1]. However, statistics show that boundary words are not reliable for predicting the right order between two larger neighboring blocks. Al-Onaizan and Papineni (2006) also proved that language model is insufficient to address long-distance word reordering. If a wrong inverted order is selected for two large consecutive blocks, incorrect long-distance swapping happens.

Yet another finding is that many incorrect swappings are related to punctuation marks. First, the source sequence within a pair of balanced punctuation marks (quotes and parentheses) should be kept

---

[1]In (Xiong et al., 2006), the language model uses the leftmost/rightmost words on the target side as evidences while the MaxEnt-based reordering model uses the boundary words on both sides.

Chinese: 他 说：「这 是 个 非常 严重 的 情 况 ， 我们 只能 希望， 能 有 <u>加快 行动</u> 的 可 能性 。」
Bruin: <u>urgent action</u> , he said : "This is a very serious <u>situation</u> , we can only hope that there will be a possibility ."
Bruin+RH: he said : "This is a very serious situation , we can only hope that there will be the possibility to <u>expedite action</u> ."
Ref: He said: "This is a very serious situation. We can only hope that it is possible to speed up the operation."

Figure 1: An example of incorrect long-distance swap. The underlined Chinese words are incorrectly swapped to the beginning of the sentence by the original Bruin. RH means reordering heuristics.

within the punctuation after translation. However, it is not always true when reordering is involved. Sometime the punctuation marks are distorted with the enclosed words sequences being moved out. Secondly, it is found that a series of words is frequently reordered from one side of a structural mark, such as commas, semi-colons and colons, to the other side of the mark for long sentences containing such marks. Generally speaking, on Chinese-to-English translation, source words are translated monotonously relative to their adjacent punctuation marks, which means their order relative to punctuation marks will not be changed. In summary, punctuation marks place a strong constraint on word order around them.

For example, in Figure 1, Chinese words "加快行动" are reordered to sentence beginning. That is an incorrect long-distance swapping, which makes the reordered words moved out from the balanced punctuation marks "「" and "」", and incorrectly precede their previous mark "，".

These incorrect swappings definitely jeopardize the quality of translation. Here we propose two straightforward but effective heuristics to control and adjust the reordering, namely swapping window and punctuation restriction.

Swapping Window (SW): It constrains block swapping in the following way

ACTIVATE $A \rightarrow \langle A^1, A^2 \rangle$ **IF** $|A_s^1| + |A_s^2| < sws$

where $|A_s^i|$ denotes the number of words on the source side A_s^i of block A^i, sws is a pre-defined swapping window size. Any inverted reordering beyond the pre-defined swapping window size is prohibited.

Punctuation Restriction (PR): If two neighboring blocks include any of the punctuation marks $p \in \{ , 、 ： ； 「 」 《 》 （ ） " " \}$, the two blocks will be merged with straight order.

Punctuation marks were already used in parsing (Christine Doran, 2000) and statistical machine translation (Och et al., 2003). In (Och et al., 2003), three kinds of features are defined, all related to punctuation marks like quotes, parentheses and commas. Unfortunately, no statistically significant improvement on the BLEU score was reported in (Och et al., 2003). In this paper, we consider this problem from a different perspective. We emphasize that words around punctuation marks are reordered ungrammatically and therefore we positively use punctuation marks as a hard decision to restrict such reordering around punctuations. This is straightforward but yet results in significant improvement on translation quality.

The two heuristics described above can be used together. If the following conditions are satisfied, we can activate the inverted rule:

$$|A_s^1| + |A_s^2| < sws \,\&\&\, P \bigcap (A_s^1 \bigcup A_s^2) = \emptyset$$

where P is the set of punctuation marks mentioned above.

The two heuristics can also speed up decoding because decoding will be monotone within those spans which are not in accordance with both heuristics. For a sentence with n words, the total number of spans is $O(n^2)$. If we set $sws = m$ ($m < n$), then the number of spans with monotone search is $O((n-m)^2)$. With punctuation restriction, the non-monotone search space will reduce further.

3.2 Phrases with Sentence Beginning/Ending Tags

We observe that in a sentence some phrases are more likely to be located at the beginning, while other phrases are more likely to be at the end. This kind of location information with regard to the phrase position could be used for reordering. A straightforward way to use this information is to mark the beginning and ending of word-aligned sentences with $\langle s \rangle$ and $\langle /s \rangle$ respectively. This idea is borrowed from language modeling (Stolcke, 2002). The corresponding tags at the source and target sentences are aligned to each other, i.e, the beginning tag of source sentences is aligned to the beginning tag of target sentences, similarly for the ending tag. Figure 2 shows a word-aligned sentence pair annotated with the sentence beginning and ending tag.

During training, the sentence beginning and ending tags ($\langle s \rangle$ and $\langle /s \rangle$) are treated as words. Therefore the phrase extraction and MaxEnt-based reordering training algorithm need not to be modified. Phrases with the sentence beginning/ending tag will be extracted and MaxEnt-based reordering features with such tags will also be generated. For example, from the word-aligned sentence pair in Figure 2, we can extract tagged phrases like

$\langle s \rangle$ 西藏 ||| $\langle s \rangle$ Tibet 's

成绩 $\langle /s \rangle$ ||| achievements $\langle /s \rangle$

and generate MaxEnt-based reordering features with tags like

$$h_i(o, b^1, b^2) = \begin{cases} 1, & b^2.t_1 = \langle /s \rangle, o = s \\ 0, & otherwise \end{cases}$$

where b^1, b^2 are blocks, t_1 denotes the last source word, $o = s$ means the order between two blocks is straight. To avoid wrong alignments, we remove tagged phrases where only the beginning/ending tag is extracted on either side of the phrases, such as

$\langle s \rangle$ ||| $\langle s \rangle$ Those 。

$\langle /s \rangle$ ||| $\langle /s \rangle$

During decoding, we first annotate source sentences with the beiginning/ending tags, then translate them as what Bruin does. Note that phrases with sentence beginning/ending tags will be used in the same way as ordinary phrases without such tags during decoding. With the additional support of language model and MaxEnt-based reordering model, we observe that phrases with such tags are always moved to the beginning or ending of sentences correctly.

⟨s⟩	西藏	金融	工作	取得	显著	成绩	⟨/s⟩
⟨s⟩	Tibet 's	financial	work	has gained	remarkable	achievements	⟨/s⟩

Figure 2: A word-aligned sentence pair annotated with the sentence beginning and ending tag.

4 Evaluation

In this section, we report the performance of the enhanced Bruin on the NIST MT-05 and NIST MT-06 Chinese-to-English translation tasks. We describe the corpus, model training, and experiments related to the refinements described above.

4.1 Corpus

The bilingual training data is derived from the following various sources: the FBIS (LDC2003E14), Hong Kong Parallel Text (Hong Kong News and Hong Kong Hansards, LDC2004T08), Xinhua News (LDC2002E18), Chinese News Translation Text Part1 (LDC2005T06), Translations from the Chinese Treebank (LDC2003E07), Chinese English News Magazine (LDC2005E47). It contains 2.4M sentence pairs in total (68.1M Chinese words and 73.8M English words).

For the efficiency of minimum-error-rate training, we built our development set using sentences not exceeding 50 characters from the NIST MT-02 evaluation test data (580 sentences).

4.2 Training

We use exactly the same way and configuration described in (He et al., 2006) to preprocess the training data, align words and extract phrases.

We built two four-gram language models using Xinhua section of the English Gigaword corpus (181.1M words) and the English side of the bilingual training data described above respectively. We applied modified Kneser-Ney smoothing as implemented in the SRILM toolkit (Stolcke, 2002).

The MaxEnt-based reordering model is trained using the way of (Xiong et al., 2006). The difference is that we only use lexical features generated by tail words of blocks, instead of head words, removing features generated by the combination of two boundary words.

	Bleu(%)		Secs/sent	
Bruin	29.96		54.3	
sws	RH^1	RH^1_2	RH^1	RH^1_2
5	29.65	29.95	42.6	41.2
10	30.55	31.27	46.2	41.8
15	30.26	31.40	48.0	42.2
20	30.19	31.42	49.1	43.2

Table 1: Effect of reordering heuristics. RH^1 denotes swapping window while RH^1_2 denotes swapping window with the addition of punctuation restriction.

4.3 Translation Results

Table 1 compares the BLEU scores [2] and the speed in seconds/sentence of the baseline system Bruin and the enhanced system with reordering heuristics applied. The second row gives the BLEU score and the average decoding time of Bruin. The rows below row 3 show the BLEU scores and speed of the enhanced Bruin with different combinations of reordering heuristics. We can clearly see that the reordering heuristics proposed by us have a two-fold effect on the performance: improving the BLEU score and decreasing the average decoding time. The example in Figure 1 shows how reordering heuristics prevent incorrect long-distance swapping which is not in accordance with the punctuation restriction.

Table 1 also shows that a 15-word swapping window is an inflexion point with the best tradeoff between the decoding time and the BLEU score. We speculate that in our corpus most reorderings happen within a 15-word window. We use the FBIS corpus to testify this hypothesis. In this corpus, we extract all reordering examples using the algorithm of Xiong et al. (2006). Figure 3 shows the reordering length distribution curve in this corpus. Accord-

---

[2]In this paper, all BLEU scores are case-sensitive and evaluated on the NIST MT-05 Chinese-to-English translation task if there is no special note.
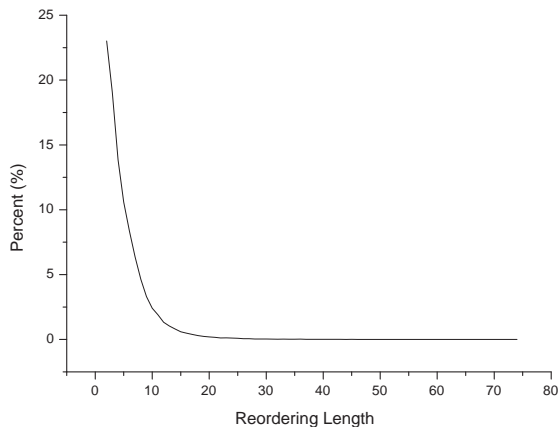
Figure 3: Reordering length distribution. The horizontal axis (reordering length) indicates the number of words on the source side of two neighboring blocks which are to be swapped. The vertical axis represents what proportion of reorderings with a certain length is likely to be in all reordering examples with an inverted order.

	Bleu(%)
Without Special Phrases	31.40
With Special Phrases	32.01

Table 2: Effect of integrating special phrases with the sentence beginning/ending tag.

ing to our statistics, reorderings within a window not exceeding 15 words have a very high proportion, 97.29%. Therefore we set $sws = 15$ for later experiments.

Table 2 shows the effect of integrating special phrases with sentence beginning/ending tags into Bruin. As special phrases accounts for only 1.95% of the total phrases used, an improvement of 0.6% in BLEU score is well worthwhile. Further, the improvement is statistically significant at the 99% confidence level according to Zhang's significant tester (Zhang et al., 2004). Figure 4 shows several examples translated with special phrases integrated. We can see that phrases with sentence beginning/ending tags are correctly selected and located at the right place.

Table 3 shows the performance of two systems on the NIST MT-05 Chinese test data, which are (1)

System	Refine	MT-05	MT-06
Bruin	-	29.96	-
EBruin	RH	31.40	30.22
EBruin	RH+SP	32.01	-

Table 3: Results of different systems. The refinements RH, SP represent reordering heuristics and special phrases with the sentence beginning/ending tag, respectively.

Bruin, trained on the large data described above; and (2) enhanced Bruin (EBruin) with different refinements trained on the same data set. This table also shows the evaluation result of the enhanced Bruin with reordering heuristics, obtained in the NIST MT-06 evaluation exercise. [3]

5 Conclusions

We have described in detail two refinements for BTG-based SMT which include reordering heuristics and special phrases with tags. The refinements were integrated into a well-established BTG-based system Bruin introduced by Xiong et al. (2006). Reordering heuristics proposed here achieve a twofold improvement: better reordering and higher-speed decoding. To our best knowledge, we are the first to integrate special phrases with the sentence beginning/ending tag into SMT. Experimental results show that the above refinements improve the baseline system significantly.

For further improvements, we will investigate possible extensions to the BTG grammars, e.g. learning useful nonterminals using unsupervised learning algorithm.

---

[3]Full results are available at http://www.nist.gov/speech/tests/mt/doc/mt06eval_official_results.html.

With Special Phrases	Without Special Phrases
⟨s⟩ **Japan had** already pledged to provide 30 million US dollars of aid due to the tsunami victims of the country . ⟨/s⟩	**originally has** pledged to provide 30 million US dollars of aid from Japan tsunami victimized countries .
⟨s⟩ **the results of the survey is based on** the results of the chiefs of the Ukrainian National 50.96% cast by chiefs . ⟨/s⟩	**is based on** the survey findings Ukraine 50.96% cast by the chiefs of the chiefs of the country .
⟨s⟩ and at the same time , the focus of the world have been transferred to **other areas . ⟨/s⟩**	and at the same time , **the global focus has shifted he**.

Figure 4: Examples translated with special phrases integrated. The bold underlined words are special phrases with the sentence beginning/ending tag.

References

Yaser Al-Onaizan, Kishore Papineni. 2006. Distortion Models for Statistical Machine Translation. In *Proceedings of ACL-COLING 2006*.

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, Michael Subotin. 2005. The Hiero Machine Translation System: Extensions, Evaluation, and Analysis. In *Proceedings of HLT/EMNLP*, pages 779 – 786, Vancouver, October 2005.

David Chiang. 2007. Hierarchical Phrase-based Translation. In *computational linguistics*, 33(2).

Christine Doran. 2000. Punctuation in a Lexicalized Grammar. In *Proceedings of Workshop TAG+5*, Paris.

Zhongjun He, Yang Liu, Deyi Xiong, Hongxu Hou, Qun Liu. 2006. ICT System Description for the 2006 TC-STAR Run #2 SLT Evaluation. In *Proceedings of TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain.

Liang Huang, Hao Zhang and Daniel Gildea. 2005. Machine Translation as Lexicalized Parsing with Hooks. In *Proceedings of the 9th International Workshop on Parsing Technologies (IWPT-05)*, Vancouver, BC, Canada, October 2005.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas*, pages 115 – 124.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *International Workshop on Spoken Language Translation*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL 2007, demonstration session, Prague, Czech Republic, June 2007.

Yang Liu, Qun Liu, Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. In *Proceedings of ACL-COLING 2006*.

Robert Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of CoNLL-2002*.

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical Machine Translation with Syntactified Target Language Phraases. In *Proceedings of EMNLP*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL 2002*, pages 295 – 302.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, Dragomir Radev. 2003. Final Report of Johns Hopkins 2003 Summer Workshop on Syntax for Statistical Machine Translation.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 901-904.

Dekai Wu. 1995. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *Proceedings of IJCAL 1995*, pages 1328-1334, Montreal, August.

Dekai Wu. 1996. A Polynomial-Time Algorithm for Statistical Machine Translation. In *Proceedings of ACL 1996*.

Deyi Xiong, Qun Liu and Shouxun Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In *Proceedings of ACL-COLING 2006*, pages 521–528.

R. Zens, H. Ney, T. Watanabe, and E. Sumita. 2004. Reordering Constraints for Phrase-Based Statistical Machine Translation. In *Proceedings of CoLing 2004*, Geneva, Switzerland, pp. 205-211.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of LREC 2004*, pages 2051–2054.