Simpler Is Better: Re-evaluation of Default Word Alignment Models in Statistical MT

Mark Fishel

Department of Computer Science, University of Tartu,
Liivi 2 - 307, Tartu 50606, Estonia
fishel@ut.ee

Abstract. Although several recent studies have shown that alignment quality is a poor indicator of the resulting translation quality, the word alignment models currently considered to be default (the so-called IBM models and HMM-based alignment) have been evaluated using the alignment error rate. We argue that from a machine translation perspective it makes sense to use simpler alignment models. Here we show that not only do the sequential models result in the same or better translation quality, but even from the set of sequential alignment models simpler ones can match the performance of the HMM-based model, whereas using computationally less expensive and faster algorithms to train and align new sentence pairs. Empirical evaluation is performed on a phrase-based and a parsing-based translation system.

Keywords: Word alignment, machine translation, relative reordering

1 Introduction

A majority of state-of-the-art statistical machine translation (SMT) systems operate with multi-word units but still use word alignment as an intermediate step for learning the translation models. Such is the case for two wide-spread machine translation frameworks: phrase-based SMT of (Koehn *et al.*, 2003) and hierarchical phrase-based SMT of (Chiang, 2005). In phrase-based systems word alignment is used to construct phrase tables and in hierarchical phrase-based systems – to extract the synchronous grammar rules.

The word alignment models that are currently considered as default are the so-called IBM models 1 to 5 (Brown *et al.*, 1993) and the HMM-based alignment model (Vogel *et al.*, 1996). The main work evaluating them is (Och and Ney, 2003) where they are compared in the context of word alignment only (i.e. based on the alignment error rate). Specifically, the default setup of the well known implementation of the models, GIZA++, is derived from (Och and Ney, 2003) and involves training the following models in sequence: IBM model 1, HMM-based model, IBM model 3 and IBM model 4.

Although Och and Ney (2003) mention briefly that "improved alignment quality yields an improved subjective quality of the statistical machine translation system as well", a number of recent studies suggest otherwise – namely, that the correlation between the alignment quality and the quality of the resulting translation is rather weak (see the following section on related work). This suggests that the best default word alignment models are not necessarily optimal in terms of the resulting translation quality.

Some recent works (e.g. (Liang *et al.*, 2006) or (DeNero and Klein, 2007)) are already based on the sequential HMM word alignment model, rather than the fertility-based models 3 or 4 . The former is computationally less complex and at the same time still includes the essential parts of the word alignment (i.e. lexical correspondence and changes in word order).

In this work we show that a simpler alignment model introduced together with HMM-based alignment in (Vogel *et al.*, 1996), but discarded due to worse alignment error rate, results in essentially the same translation quality like HMM-based alignment in almost all cases – i.e. the relative-distortion IBM model 2. At the same time, it does not include a first-order dependency of the alignment, which means that it is much simpler to implement and train it. The experiment results also support the common knowledge that HMM-based alignment works just as well or sometimes better than IBM model 4, usually used by default.

In the following section we review the related work on the link between the word alignment quality and translation quality. Section 3 consists of a theoretical overview of different aspects of the word alignment task in the models in question. In section 4 we present the experiments with simpler default and alternative models on translations from Chinese, Czech, Estonian, Finnish, German and Korean into English and back.

2 Related Work

Several papers point out that the scores designed for word alignment (alignment error-rate, F-score) and translation (BLEU, NIST), are not heavily correlated. In particular (Fraser and Marcu, 2007) and (Ayan and Dorr, 2006) distinctly state that alignment error rate is a poor indicator of translation quality.

(Lopez and Resnik, 2006) artificially degrade the alignment quality in order to show that it does not cause a significant drop in translation quality. They further show that with careful feature engineering the flaws of the underlying word alignment can be compensated.

(Vilar *et al.*, 2006) give two examples of word alignment modifications which cause worse alignment quality and nevertheless better translation quality. This is achieved by adapting the alignments to the specific requirements of translation.

(Guzman *et al.*, 2009) inspect word alignments and their characteristics, especially the number of unaligned words, and their influence on phrase pair extraction. They show that an increased number of unaligned words causes degraded translation quality. Analyzing manually evaluated phrase pairs they come up with translation model features that account for the number of unaligned words and improve the translation quality. (Lambert *et al.*, 2009) tune alignment for the F-score and the BLEU score. They show that the two objectives are not the same and produce different translation models. (Ganchev *et al.*, 2008) use agreement-driven training of alignment models and replace Viterbi decoding with posterior decoding. This results in improvements both in the alignment quality as well as translation quality.

A brief comparison of the IBM models in SMT context is performed in (Koehn *et al.*, 2003). The comparison is based on the BLEU scores and covers IBM models 1 to 4. The given brief conclusions are that using different alignment models does not cause significant changes in translation quality. Model 1 is noted for lower scores and models 2 and 4 are said to produce similar results. Our results suggest the contrary for the latter point.

(He, 2007) introduce word dependent HMM-based word alignment. They apply fully lexicalized transition modeling by additionally conditioning the first-order dependency of the alignment on the corresponding output word. They show that this modified alignment model can match the performance of IBM model 4. As it will be shown later in this work, usual HMM-based alignment models also achieve the same result.

3 Word Alignment

The task of word alignment is to find matching words in a pair of sentences that have the same meaning in two different languages. Although in reality sometimes whole phrases are translated with no direct correspondence in meaning between the used words (e.g. idioms), the classic approach is to align single words.

Although the task is essentially symmetrical, the IBM and HMM-based models focus on aligning every word in one sentence to at most one word in the other one. For simplicity's sake we will use the standard notation of $\mathbf{f}$ and $\mathbf{e}$ for the two sentences.

The aim is therefore to find an alignment $\mathbf{a}$, which is a vector of indexes indicating which words in $\mathbf{e}$ the words in $\mathbf{f}$ are aligned to; in other words, the word f_j is aligned to the word e_i if $a_j = i$. Also any a_j can be equal to 0, in which case the word f_j is said to be unaligned.

Here we will focus on the IBM models (Brown *et al.*, 1993), the HMM-based alignment model (Vogel *et al.*, 1996) and a modification of the original IBM model 2 introduced in (Och and Ney, 2000) and referred to as *diagonal-oriented* model 2.

3.1 Lexical Correspondence

Lexical (or translational) correspondence of the single words in the two sentences is perhaps the main aspect of word alignment and is present in all of the described models. In all the models considered here lexical correspondence is treated as independent of the word positions in the sentences or any context of either words; it is modeled via a probability distribution $p(f|e)$.

Although lexical correspondence is a very important aspect, constricting an alignment model to it (as it is the case with model 1) results in serious model flaws: in case of any lexical ambiguity the model will select the most probable word pair in all cases, and the model would not be able to resolve a conflict between repeated items, like punctuation marks or same words.

3.2 Distortion

Modeling the different word order in the two sentences is also referred to as distortion (Brown *et al.*, 1993). Just like the lexical correspondence aspect, in the IBM models it is considered to be independent of the words themselves or their context. The models 2 and 3 include a distortion component based only on the absolute word positions and the sentence lengths: $p(a_j|j, J, I)$, where $J = |\mathbf{f}|$ and $I = |\mathbf{e}|$. The problem with absolute word positions is that, simply put, same words can occur at different positions in sentences of different length. This means that a separate parameter subset models each different position and sentence length which, in addition to unnecessary treating of the same words differently, can easily suffer from the sparse data effect.

A modification of the original model 2, introduced originally in (Vogel *et al.*, 1996) and developed further in (Och and Ney, 2000), is instead based on the distance between a_j and a scaled j: $d = a_j - \left\lfloor \frac{j \cdot I}{J} \right\rfloor$.

Finally, (Och and Ney, 2000) introduce lexicalized reordering, which is additionally conditioned on some general classes of the words ($C(f_j)$, $C(e_{a_j})$).

3.3 Distortion First-order Dependency

(Vogel *et al.*, 1996) make a step further from independent single pair distortions. They treat alignment as a Markov process with the source words as the observed and the alignments – as hidden variables. With first-order Markov dependency assumption the alignment pairs are not any more independent. This makes the training/aligning algorithms more complicated. Still with the help of dynamic programming these can be solved with no approximations.

First-order dependency is a simple way to take into consideration the context of the word pair. If a neighboring (unambiguous) word pair is aligned with an atypical relative distortion, an HMM-based model is capable of deciding to align the current word similarly.

3.4 Fertility

The aspect of fertility aims at modeling the fact that a single word can either be unconnected with the other sentence or be aligned with more than one word. In models 3 and 4 this is modeled explicitly: $p(\phi|e)$, where ϕ denotes the number of words in $\mathbf{f}$ that e is connected to. Since this aspect influences the first-order dependencies as well as distortion, it makes learning and applying of such models even more complicated and is solved with approximations in practice.

Table 1: The size of the training parts of the used parallel corpora

Corpus	Number of sentence pairs	Number of words (English)	Number of words (Foreign)
OPUS KDE4 (Korean-English)	$64.1 \cdot 10^3$	$0.32 \cdot 10^6$	$0.33 \cdot 10^6$
OPUS KDE4 (Chinese-English)	$103.7 \cdot 10^3$	$0.57 \cdot 10^6$	$0.78 \cdot 10^6$
CzEng, tech. docs (Czech-English)	$0.97 \cdot 10^6$	$7.27 \cdot 10^6$	$6.59 \cdot 10^6$
JRC-Acquis (Estonian-English)	$1.09 \cdot 10^6$	$27.91 \cdot 10^6$	$20.18 \cdot 10^6$
Europarl (German-English)	$1.52 \cdot 10^6$	$41.98 \cdot 10^6$	$39.81 \cdot 10^6$
Europarl (Finnish-English)	$1.59 \cdot 10^6$	$43.94 \cdot 10^6$	$31.58 \cdot 10^6$

4 Experiments

In this section we compare the influence of the word alignments on the resulting translation quality. The main aim is to test, whether it is necessary to use the complex default models or not – i.e., whether simpler models can match their performance.

The default way of training the IBM models, as proposed by (Och and Ney, 2003), is IBM model 1, HMM-based model, IBM model 3 and finally model 4, whereas the resulting parameters of the simpler models are used as the initial values of the more elaborate models. We first evaluate, how stopping at an earlier stage of word alignment influences translation quality. The hypothesis here, dictated by common knowledge, is that HMM-based alignment can perform just as well or even better than IBM model 4.

As a next step we compare the default alignment model training sequence to an alternative, which uses different variants of IBM model 2 as a final step. Our aim is to see whether some variant can match the performance of the HMM-based model and IBM model 4.

4.1 Experiment Setup

We evaluate the influence of word alignment on two translation systems – a phrase-based system trained with Moses (Koehn *et al.*, 2007) and a hierarchical phrase-based system trained with Joshua (Li *et al.*, 2009). In both cases we used 5-gram language models from SRI LM (Stolcke, 2002) and minimum error rate training included in the toolkits.

Word alignment was done with GIZA++ (Och and Ney, 2003) for both systems. We modified its implementation to support three kinds of IBM2-based models: the absolute reordering-based model already included in GIZA++ (abbreviated as IBM2), the relative reordering-based model (abbreviated as IBM2(r)) and the latter, augmented with lexicalization, as suggested by (Och and Ney, 2000) (abbreviated as IBM2(r-l)).

We performed all of the experiments on the following language pairs and corpora: the Chinese-English and Korean-English parts of the OPUS KDE4 corpus (Tiedemann, 2009), Czech-English technical documentations from CzEng (Bojar and Žabokrtský, 2009), the Estonian-English part of the JRC-Acquis (Steinberger *et al.*, 2006) and Finnish-English and German-English parts of Europarl (Koehn, 2005); all experiments included both translation directions.

Two independent held-out sets, each 2500 sentence pairs, were reserved for minimum error-rate training and validation; the resulting sizes of the training parts after preprocessing and separating the dev and test sets are summarized in table 1.

4.2 Results

We used the BLEU (Papieni *et al.*, 2001) and NIST (NIST, 2002) scores for evaluation and paired bootstrap resampling (Riezler and Maxwell, 2005) for significance testing.

The results of the first part of the experiments, which is stopping midway in training the word alignment models, are presented on table 1. The scores of translations based on IBM model 1 are

noticeably lower than all other models; also after the HMM-based model there is a noticeable drop at the IBM model 3 for almost every language pair, after which the IBM4 model scores rise to the level of the HMM model again.

Significance testing reveals that only in case of Korean-English translation the NIST score of the HMM model is significantly lower (p-value 0.009) than the score of the IBM model 4; in all other cases both scores of the HMM model is either insignificantly different or significantly higher.

The IBM2 models were trained in the same way as the HMM-based model: starting with the IBM model 1. The resulting scores of the IBM2 models are compared to the HMM-based model in table 2. As expected, the absolute reordering-based IBM2 model results are considerably lower than all other models in all experiments.

The only case where an IBM2-based model outperforms an HMM-based model is English-Czech translation with IBM2(r) (based on the BLEU score). Also Czech-English translation results are essentially the same for IBM2(r) based on both scores with Joshua, and English-Czech, English-Korean and English-Chinese results for IBM2(r-l) – with Moses.

Most other examples of IBM2-based models having insignificantly different scores from the HMM-based models are IBM2(r) with Joshua (German-English and English-German NIST, Chinese-English and English-Chinese BLEU) and IBM2(r-l) with Moses (Czech-English BLEU, German-English NIST, Chinese-English NIST).

In all other cases HMM-based alignment outperforms IBM2-based alignments significantly. However the score differences are relatively small (0.5-0.6 BLEU and 0.04-0.06 NIST points) in many cases. The main translation directions where the difference between the HMM-based alignment and the best IBM2-based model is high with both phrase-based and parsing-based translation are Estonian-English, Korean-English, Finnish-English and English-Estonian.

To conclude, it seems that the IBM4 model can be safely replaced with the HMM-based model. Also, although the IBM2-based models did not outperform HMM entirely, for many language pairs the difference was estimated as insignificant and for some others – significant but relatively small. Thus the relative distortion-based IBM2 models can serve as an efficient trade-off between efficiency and quality.

References

Ayan, Necip Fazil and Bonnie J. Dorr. 2006. Going beyond AER: An extensive analysis of word alignments and their impact on mt. In *Proceedings of ACL/COLING'06*.

Bojar, Ondřej and Zdeněk Žabokrtský. 2009. CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92. in print.

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

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL'05*, pp. 263–270, Ann Arbor, USA.

DeNero, John and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of ACL'07*, p. 17, Prague, Czech Republic.

Fraser, Alexander and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3), 293–303.

Ganchev, Kuzman, João V. Graça, and Ben Taskar. 2008. Better alignments = better translations? In *Proceedings of ACL'08*, pp. 986–993, Columbus, USA.

Guzman, Francisco, Qin Gao, and Stephan Vogel. 2009. Reassessment of the role of phrase extraction in PBSMT. In *Proceedings of MT Summit XII*, Ottawa, Canada.

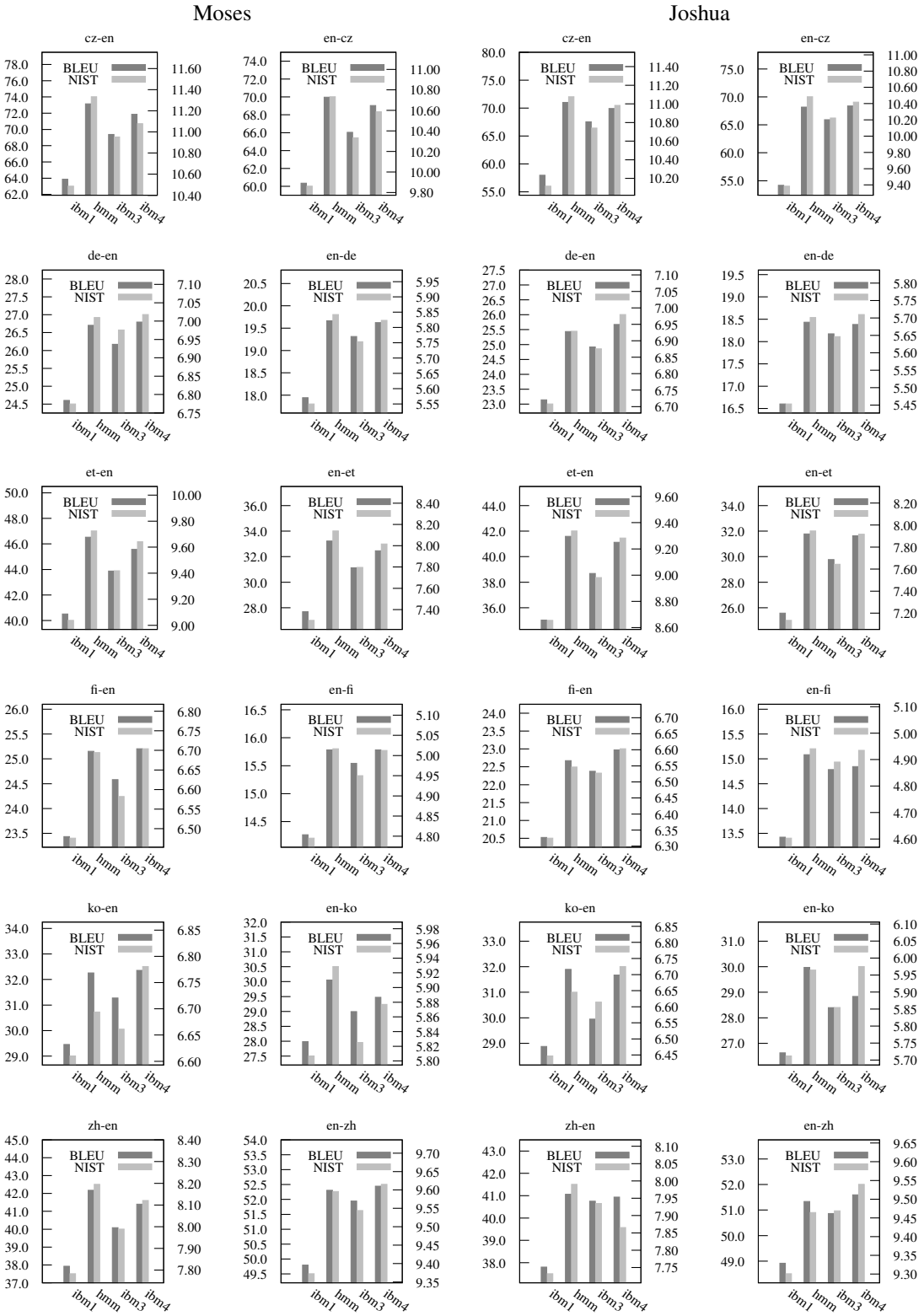

Figure 1: Results of early-stopping experiments: comparison of the influence of word alignment models IBM1, HMM-based, IBM3 and IBM4 on the resulting translation quality.

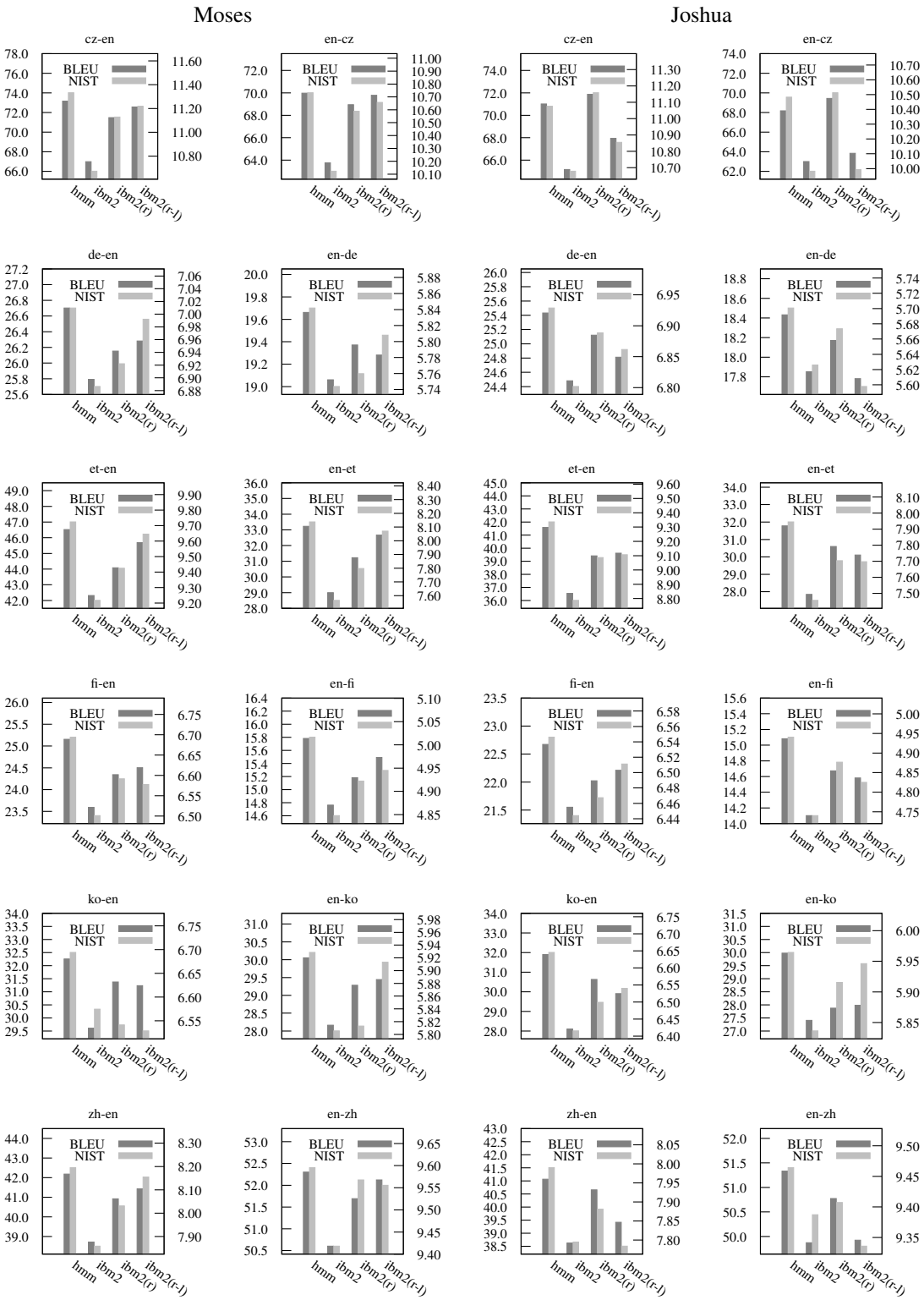

Figure 2: Results of comparing the HMM-based model to IBM2-based models.

He, Xiaodong. 2007. Using word-dependent transition models in HMM-based word alignment for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 80–87, Prague, Czech Republic.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, Phuket, Thailand.

Koehn, Philipp, 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, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL'07*, pp. 177–180, Prague, Czech Republic.

Koehn, Philipp, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL-HLT'03*, pp. 48–54, Edmonton, Canada.

Lambert, Patrik, Yanjun Ma, Sylwia Ozdowska, and Andy Way. 2009. Tracking relevant alignment characteristics for machine translation. In *Proceedings of MT Summit XII*, pp. 268–275, Ottawa, Canada.

Li, Zhifei, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 135–139, Athens, Greece.

Liang, Percy, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of NAACL-HLT'06*, pp. 104–111, New York, USA.

Lopez, Adam and Philip Resnik. 2006. Word-based alignment, phrase-based translation: what's the link? In *Proceedings of AMTA'06*, pp. 90–99.

NIST. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Technical report.

Och, Franz J. and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19–51.

Och, Franz Joseph and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of COLING'2000*, pp. 1086–1090, Saarbrücken, Germany.

Papieni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL'01*, pp. 311–318, Philadelphia, PA, USA.

Riezler, Stefan and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, pp. 57–64, Ann Arbor, USA.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC'06*, pp. 2142–2147, Genoa, Italy.

Stolcke, Andreas. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP'02*, volume 2, pp. 901–904, Denver, Colorado, USA.

Tiedemann, Jörg. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of RANLP'09*, pp. 237–248, Borovets, Bulgaria.

Vilar, David, Maja Popovic, and Hermann Ney. 2006. AER: Do we need to "improve" our alignments. In *Proceedings of IWSLT'06*, pp. 205–212.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of COLING'96*, pp. 836–841, Copenhagen, Denmark.