Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT?

Nadir Durrani
University of Edinburgh
dnadir@inf.ed.ac.uk

Alexander Fraser **Helmut Schmid**
Ludwig Maximilian University Munich
fraser,schmid@cis.uni-muenchen.de

Hieu Hoang **Philipp Koehn**
University of Edinburgh
hieu.hoang,pkoehn@inf.ed.ac.uk

Abstract

The phrase-based and N-gram-based SMT frameworks complement each other. While the former is better able to memorize, the latter provides a more principled model that captures dependencies across phrasal boundaries. Some work has been done to combine insights from these two frameworks. A recent successful attempt showed the advantage of using phrase-based search on top of an N-gram-based model. We probe this question in the reverse direction by investigating whether integrating N-gram-based translation and reordering models into a phrase-based decoder helps overcome the problematic phrasal independence assumption. A large scale evaluation over 8 language pairs shows that performance does significantly improve.

1 Introduction

Phrase-based models (Koehn et al., 2003; Och and Ney, 2004) learn local dependencies such as reorderings, idiomatic collocations, deletions and insertions by memorization. A fundamental drawback is that phrases are translated and reordered independently of each other and contextual information outside of phrasal boundaries is ignored. The monolingual language model somewhat reduces this problem. However i) often the language model cannot overcome the dispreference of the translation model for nonlocal dependencies, ii) source-side contextual dependencies are still ignored and iii) generation of lexical translations and reordering is separated.

The N-gram-based SMT framework addresses these problems by learning Markov chains over sequences of minimal translation units (MTUs) also known as tuples (Mariño et al., 2006) or over operations coupling lexical generation and reordering (Durrani et al., 2011). Because the models condition the MTU probabilities on the previous MTUs, they capture non-local dependencies and both source and target contextual information across phrasal boundaries.

In this paper we study the effect of integrating tuple-based N-gram models (TSM) and operation-based N-gram models (OSM) into the phrase-based model in Moses, a state-of-the-art phrase-based system. Rather than using POS-based rewrite rules (Crego and Mariño, 2006) to form a search graph, we use the ability of the phrase-based system to memorize larger translation units to replicate the effect of source linearization as done in the TSM model.

We also show that using phrase-based search with MTU N-gram translation models helps to address some of the search problems that are non-trivial to handle when decoding with minimal translation units. An important limitation of the OSM N-gram model is that it does not handle unaligned or discontinuous target MTUs and requires post-processing of the alignment to remove these. Using phrases during search enabled us to make novel changes to the OSM generative story (also applicable to the TSM model) to handle unaligned target words and to use target linearization to deal with discontinuous target MTUs.

We performed an extensive evaluation, carrying out translation experiments from French, Spanish, Czech and Russian to English and in the opposite direction. Our integration of the OSM model into Moses and our modification of the OSM model to deal with unaligned and discontinuous target tokens consistently improves BLEU scores over the

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 399–405,
Sofia, Bulgaria, August 4-9 2013.

baseline system, and shows statistically significant improvements in seven out of eight cases.

2 Previous Work

Several researchers have tried to combine the ideas of phrase-based and N-gram-based SMT. Costa-jussà et al. (2007) proposed a method for combining the two approaches by applying sentence level reranking. Feng et al. (2010) added a linearized source-side language model in a phrase-based system. Crego and Yvon (2010) modified the phrase-based lexical reordering model of Tillman (2004) for an N-gram-based system. Niehues et al. (2011) integrated a bilingual language model based on surface word forms and POS tags into a phrase-based system. Zhang et al. (2013) explored multiple decomposition structures for generating MTUs in the task of lexical selection, and to rerank the N-best candidate translations in the output of a phrase-based. A drawback of the TSM model is the assumption that source and target information is generated monotonically. The process of reordering is disconnected from lexical generation which restricts the search to a small set of precomputed reorderings. Durrani et al. (2011) addressed this problem by coupling lexical generation and reordering information into a single generative process and enriching the N-gram models to learn lexical reordering triggers. Durrani et al. (2013) showed that using larger phrasal units during decoding is superior to MTU-based decoding in an N-gram-based system. However, they do not use phrase-based models in their work, relying only on the OSM model. This paper combines insights from these recent pieces of work and show that phrase-based search combined with N-gram-based and phrase-based models in decoding is the overall best way to go. We integrate the two N-gram-based models, TSM and OSM, into phrase-based Moses and show that the translation quality is improved by taking both translation and reordering context into account. Other approaches that explored such models in syntax-based systems used MTUs for sentence level reranking (Khalilov and Fonollosa, 2009), in dependency translation models (Quirk and Menezes, 2006) and in target language syntax systems (Vaswani et al., 2011).

3 Integration of N-gram Models

We now describe our integration of TSM and OSM N-gram models into the phrase-based system. Given a bilingual sentence pair (F, E) and its alignment (A), we first identify minimal translation units (MTUs) from it. An MTU is defined as a translation rule that cannot be broken down any further. The MTUs extracted from Figure 1(a) are $A \rightarrow a, B \rightarrow b, C \ldots H \rightarrow c$[1] and $D \rightarrow d$. These units are then generated left-to-right in two different ways, as we will describe next.

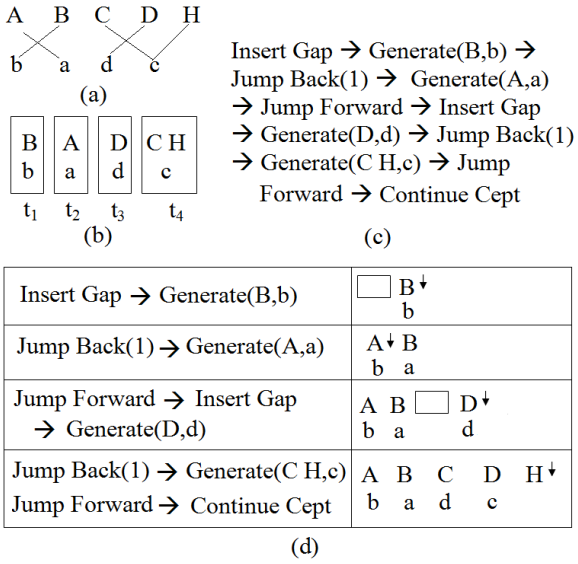

Figure 1: Example (a) Word Alignments (b) Unfolded MTU Sequence (c) Operation Sequence (d) Step-wise Generation

3.1 Tuple Sequence Model (TSM)

The TSM translation model assumes that MTUs are generated monotonically. To achieve this effect, we enumerate the MTUs in the target left-to-right order. This process is also called source linearization or tuple unfolding. The resulting sequence of monotonic MTUs is shown in Figure 1(b). We then define a TSM model over this sequence $(t_1, t_2, \ldots, t_J)$ as:

$$p_{tsm}(F, E, A) = \prod_{j=1}^{J} p(t_j | t_{j-n+1}, ..., t_{j-1})$$

where n indicates the amount of context used. A 4-gram Kneser-Ney smoothed language model is trained with SRILM (Stolcke, 2002).

Search: In previous work, the search graph in TSM N-gram SMT was not built dynamically like in the phrase-based system, but instead constructed as a preprocessing step using POS-based rewrite rules (learned when linearizing the source side). We do not adopt this framework. We use

[1] We use ... to denote discontinuous MTUs.

phrase-based search which builds up the decoding graph dynamically and searches through all possible reorderings within a fixed window. During decoding we use the phrase-internal alignments to perform source linearization. For example, if during decoding we would like to apply the phrase pair "C D H – d c", a combination of t_3 and t_4 in Figure 1(b), then we extract the MTUs from this phrase-pair and linearize the source to be in the order of the target. We then compute the TSM probability given the $n-1$ previous MTUs (including MTUs occurring in the previous source phrases). The idea is to replicate rewrite rules with phrase-pairs to linearize the source. Previous work on N-gram-based models restricted the length of the rewrite rules to be 7 or less POS tags. We use phrases of length 6 and less.

3.2 Operation Sequence Model (OSM)

The OSM model represents a bilingual sentence pair and its alignment through a sequence of operations that generate the aligned sentence pair. An operation either generates source and target words or it performs reordering by inserting gaps and jumping forward and backward. The MTUs are generated in the target left-to-right order just as in the TSM model. However rather than linearizing the source-side, reordering operations (gaps and jumps) are used to handle crossing alignments. During training, each bilingual sentence pair is deterministically converted to a unique sequence of operations.[2] The example in Figure 1(a) is converted to the sequence of operations shown in Figure 1(c). A step-wise generation of MTUs along with reordering operations is shown in Figure 1(d). We learn a Markov model over a sequence of operations $(o_1, o_2, \ldots, o_J)$ that encapsulate MTUs and reordering information which is defined as follows:

$$p_{osm}(F, E, A) = \prod_{j=1}^{J} p(o_j | o_{j-n+1}, ..., o_{j-1})$$

A 9-gram Kneser-Ney smoothed language model is trained with SRILM.[3] By coupling reordering with lexical generation, each (translation or reordering) decision conditions on $n-1$ previous (translation and reordering) decisions spanning across phrasal boundaries. The reordering decisions therefore influence lexical selection and vice versa. A heterogeneous mixture of translation and reordering operations enables the OSM model to memorize reordering patterns and lexicalized triggers unlike the TSM model where translation and reordering are modeled separately.

Search: We integrated the generative story of the OSM model into the hypothesis extension process of the phrase-based decoder. Each hypothesis maintains the position of the source word covered by the last generated MTU, the right-most source word generated so far, the number of open gaps and their relative indexes, etc. This information is required to generate the operation sequence for the MTUs in the hypothesized phrase-pair. After the operation sequence is generated, we compute its probability given the previous operations. We define the main OSM feature, and borrow 4 supportive features, the *Gap*, *Open Gap*, *Gap-width* and *Deletion* penalties (Durrani et al., 2011).

3.3 Problem: Target Discontinuity and Unaligned Words

Two issues that we have ignored so far are the handling of MTUs which have discontinuous targets, and the handling of unaligned target words. Both TSM and OSM N-gram models generate MTUs linearly in left-to-right order. This assumption becomes problematic in the cases of MTUs that have target-side discontinuities (See Figure 2(a)). The MTU $A \rightarrow g \ldots a$ can not be generated because of the intervening MTUs $B \rightarrow b, C \ldots H \rightarrow c$ and $D \rightarrow d$. In the original TSM model, such cases are dealt with by merging all the intervening MTUs to form a bigger unit t'_1 in Figure 2(c). A solution that uses split-rules is proposed by Crego and Yvon (2009) but has not been adopted in Ncode (Crego et al., 2011), the state-of-the-art TSM N-gram system. Durrani et al. (2011) dealt with this problem by applying a post-processing (PP) heuristic that modifies the alignments to remove such cases. When a source word is aligned to a discontinuous target-cept, first the link to the least frequent target word is identified, and the group of links containing this word is retained while the others are deleted. The alignment in Figure 2(a), for example, is transformed to that in Figure 2(b). This allows OSM to extract the intervening MTUs $t_2 \ldots t_5$ (Figure 2(c)). Note that this problem does not exist when dealing with source-side discontinuities: the TSM model linearizes discontinuous source-side MTUs such as $C \ldots H \rightarrow c$. The

[2] Please refer to Durrani et al. (2011) for a list of operations and the conversion algorithm.

[3] We also tried a 5-gram model, the performance decreased slightly in some cases.

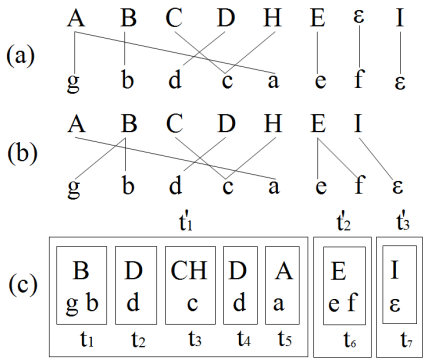

Figure 2: Example (a) Original Alignments (b) Post-Processed Alignments (c) Extracted MTUs – $t'_1 \ldots t'_3$ (from (a)) and $t_1 \ldots t_7$ (from (b))

OSM model deals with such cases through *Insert Gap* and *Continue Cept* operations.

The second problem is the unaligned target-side MTUs such as $\varepsilon \rightarrow f$ in Figure 2(a). Inserting target-side words "spuriously" during decoding is a non-trival problem because there is no evidence of when to hypothesize such words. These cases are dealt with in N-gram-based SMT by merging such MTUs to the MTU on the left or right based on attachment counts (Durrani et al., 2011), lexical probabilities obtained from IBM Model 1 (Mariño et al., 2006), or POS entropy (Gispert and Mariño, 2006). Notice how $\varepsilon \rightarrow f$ (Figure 2(a)) is merged with the neighboring MTU $E \rightarrow e$ to form a new MTU $E \rightarrow ef$ (Figure 2 (c)). We initially used the post-editing heuristic (PP) as defined by Durrani et al. (2011) for both TSM and OSM N-gram models, but found that it lowers the translation quality (See Row 2 in Table 2) in some language pairs.

3.4 Solution: Insertion and Linearization

To deal with these problems, we made novel modifications to the generative story of the OSM model. Rather than merging the unaligned target MTU such as $\varepsilon - f$, to its right or left MTU, we generate it through a new *Generate Target Only (f)* operation. Orthogonal to its counterpart *Generate Source Only (I)* operation (as used for MTU t_7 in Figure 2 (c)), this operation is generated as soon as the MTU containing its previous target word is generated. In Figure 2(a), $\varepsilon - f$ is generated immediately after MTU $E - e$ is generated. In a sequence of unaligned source and target MTUs, unaligned source MTUs are generated before the unaligned target MTUs. We do not modify the decoder to arbitrarily generate unaligned MTUs but hypothesize these only when they appear within an extracted phrase-pair. The constraint provided by the phrase-based search makes the *Generate Target Only* operation tractable. Using phrase-based search therefore helps addressing some of the problems that exist in the decoding framework of N-gram SMT.

The remaining problem is the discontinuous target MTUs such as $A \rightarrow g \ldots a$ in Figure 2(a). We handle this with target linearization similar to the TSM source linearization. We collapse the target words g and a in the MTU $A \rightarrow g \ldots a$ to occur consecutively when generating the operation sequence. The conversion algorithm that generates the operations thinks that g and a occurred adjacently. During decoding we use the phrasal alignments to linearize such MTUs within a phrasal unit. This linearization is done only to compute the OSM feature. Other features in the phrase-based system (e.g., language model) work with the target string in its original order. Notice again how memorizing larger translation units using phrases helps us reproduce such patterns. This is achieved in the tuple N-gram model by using POS-based split and rewrite rules.

4 Evaluation

Corpus: We ran experiments with data made available for the translation task of the *Eighth Workshop on Statistical Machine Translation*. The sizes of bitext used for the estimation of translation and monolingual language models are reported in Table 1. All data is true-cased.

Pair	Parallel	Monolingual	Lang
fr–en	≈39 M	≈91 M	fr
cs–en	≈15.6 M	≈43.4 M	cs
es–en	≈15.2 M	≈65.7 M	es
ru–en	≈2 M	≈21.7 M	ru
		≈287.3 M	en

Table 1: Number of Sentences (in Millions) used for Training

We follow the approach of Schwenk and Koehn (2008) and trained domain-specific language models separately and then linearly interpolated them using SRILM with weights optimized on the held-out dev-set. We concatenated the news-test sets from four years (2008-2011) to obtain a large dev-setin order to obtain more stable weights (Koehn and Haddow, 2012). For Russian-English and English-Russian language pairs, we divided the tuning-set news-test 2012 into two halves and used

No.	System	fr-en	es-en	cs-en	ru-en	en-fr	en-es	en-cs	en-ru
1.	Baseline	31.89	35.07	23.88	33.45	29.89	35.03	16.22	23.88
2.	1+pp	31.87	35.09	23.64	33.04	29.70	35.00	16.17	24.05
3.	1+pp+tsm	31.94	35.25	23.85	32.97	29.98	35.06	16.30	23.96
4.	1+pp+osm	32.17	**35.50**	24.14	33.21	**30.35**	**35.34**	16.49	**24.22**
5.	1+osm*	32.13	**35.65**	**24.23**	**33.91**	**30.54**	**35.49**	**16.62**	**24.25**

Table 2: Translating into and from English. Bold: Statistically Significant (Koehn, 2004) w.r.t Baseline

the first half for tuning and second for test. We test our systems on news-test 2012. We tune with the k-best batch MIRA algorithm (Cherry and Foster, 2012).

Moses Baseline: We trained a Moses system (Koehn et al., 2007) with the following settings: maximum sentence length 80, grow-diag-final-and symmetrization of GIZA++ alignments, an interpolated Kneser-Ney smoothed 5-gram language model with KenLM (Heafield, 2011) used at runtime, msd-bidirectional-fe lexicalized reordering, sparse lexical and domain features (Hasler et al., 2012), distortion limit of 6, 100-best translation options, minimum bayes-risk decoding (Kumar and Byrne, 2004), cube-pruning (Huang and Chiang, 2007) and the no-reordering-over-punctuation heuristic.

Results: Table 2 shows uncased BLEU scores (Papineni et al., 2002) on the test set. Row 2 (+pp) shows that the post-editing of alignments to remove unaligned and discontinuous target MTUs decreases the performance in the case of ru-en, cs-en and en-fr. Row 3 (+pp+tsm) shows that our integration of the TSM model slightly improves the BLEU scores for en-fr, and es-en. Results drop in ru-en and en-ru. Row 4 (+pp+osm) shows that the OSM model consistently improves the BLEU scores over the Baseline systems (Row 1) giving significant improvements in half the cases. The only result that is lower than the baseline system is that of the ru-en experiment, because OSM is built with PP alignments which particularly hurt the performance for ru-en. Finally Row 5 (+osm*) shows that our modifications to the OSM model (Section 3.4) give the best result ranging from $[0.24-0.65]$ with statistically significant improvements in seven out of eight cases. It also shows improvements over Row 4 (+pp+osm) even in some cases where the PP heuristic doesn't hurt. The largest gains are obtained in the ru-en translation task (where the PP heuristic inflicted maximum damage).

5 Conclusion and Future Work

We have addressed the problem of the independence assumption in PBSMT by integrating N-gram-based models inside a phrase-based system using a log-linear framework. We try to replicate the effect of rewrite and split rules as used in the TSM model through phrasal alignments. We presented a novel extension of the OSM model to handle unaligned and discontinuous target MTUs in the OSM model. Phrase-based search helps us to address these problems that are non-trivial to handle in the decoding frameworks of the N-gram-based models. We tested our extentions and modifications by evaluating against a competitive baseline system over 8 language pairs. Our integration of TSM shows small improvements in a few cases. The OSM model which takes both reordering and lexical context into consideration consistently improves the performance of the baseline system. Our modification to the OSM model produces the best results giving significant improvements in most cases. Although our modifications to the OSM model enables discontinuous MTUs, we did not fully utilize these during decoding, as Moses only uses continous phrases. The discontinuous MTUs that span beyond a phrasal length of 6 words are therefore never hypothesized. We would like to explore this further by extending the search to use discontinuous phrases (Galley and Manning, 2010).

Acknowledgments

We would like to thank the anonymous reviewers for their helpful feedback and suggestions. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n ° 287658. Alexander Fraser was funded by Deutsche Forschungsgemeinschaft grant Models of Morphosyntax for Statistical Machine Translation. Helmut Schmid was supported by Deutsche Forschungsgemeinschaft grant SFB 732. This publication only reflects the authors views.

References

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June. Association for Computational Linguistics.

Marta R. Costa-jussà, Josep M. Crego, David Vilar, José A.R. Fonollosa, José B. Mariño, and Hermann Ney. 2007. Analysis and System Combination of Phrase- and N-Gram-Based Statistical Machine Translation Systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 137–140, Rochester, New York, April.

Josep M. Crego and José B. Mariño. 2006. Improving Statistical MT by Coupling Reordering and Decoding. *Machine Translation*, 20(3):199–215.

Josep M. Crego and François Yvon. 2009. Gappy Translation Units under Left-to-Right SMT Decoding. In *Proceedings of the Meeting of the European Association for Machine Translation (EAMT)*, pages 66–73, Barcelona, Spain.

Josep M. Crego and François Yvon. 2010. Improving Reordering with Linguistically Informed Bilingual N-Grams. In *Coling 2010: Posters*, pages 197–205, Beijing, China, August. Coling 2010 Organizing Committee.

Josep M. Crego, François Yvon, and José B. Mariño. 2011. Ncode: an Open Source Bilingual N-gram SMT Toolkit. *The Prague Bulletin of Mathematical Linguistics*, 96:49–58.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A Joint Sequence Translation Model with Integrated Reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA, June.

Nadir Durrani, Alexander Fraser, and Helmut Schmid. 2013. Model With Minimal Translation Units, But Decode With Phrases. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Minwei Feng, Arne Mauser, and Hermann Ney. 2010. A Source-side Decoding Sequence Model for Statistical Machine Translation. In *Conference of the Association for Machine Translation in the Americas 2010*, Denver, Colorado, USA, October.

Michel Galley and Christopher D. Manning. 2010. Accurate Non-Hierarchical Phrase-Based Translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 966–974, Los Angeles, California, June. Association for Computational Linguistics.

Adrià Gispert and José B. Mariño. 2006. Linguistic Tuple Segmentation in N-Gram-Based Statistical Machine Translation. In *INTERSPEECH*.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2012. Sparse Lexicalised Features and Topic Adaptation for SMT. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 268–275.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, 7.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic, June. Association for Computational Linguistics.

Maxim Khalilov and José A. R. Fonollosa. 2009. N-Gram-Based Statistical Machine Translation Versus Syntax Augmented Machine Translation: Comparison and System Combination. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 424–432, Athens, Greece, March. Association for Computational Linguistics.

Philipp Koehn and Barry Haddow. 2012. Towards Effective Use of Training Data in Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 317–321, Montréal, Canada, June. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL*, pages 127–133, Edmonton, Canada.

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, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007 Demonstrations*, Prague, Czech Republic.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July.

Shankar Kumar and William J. Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *HLT-NAACL*, pages 169–176.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram-Based Machine Translation. *Computational Linguistics*, 32(4):527–549.

Jan Niehues, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider Context by Using Bilingual Language Models in Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 198–206, Edinburgh, Scotland, July. Association for Computational Linguistics.

Franz J. Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(1):417–449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Morristown, NJ, USA.

Christopher Quirk and Arul Menezes. 2006. Do We Need Phrases? Challenging the Conventional Wisdom in Statistical Machine Translation. In *HLT-NAACL*.

Holger Schwenk and Philipp Koehn. 2008. Large and Diverse Language Models for Statistical Machine Translation. In *International Joint Conference on Natural Language Processing*, pages 661–666, January 2008.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Intl. Conf. Spoken Language Processing*, Denver, Colorado.

Christoph Tillman. 2004. A Unigram Orientation Model for Statistical Machine Translation. In *HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts.

Ashish Vaswani, Haitao Mi, Liang Huang, and David Chiang. 2011. Rule Markov Models for Fast Tree-to-String Translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 856–864, Portland, Oregon, USA, June.

Hui Zhang, Kristina Toutanova, Chris Quirk, and Jianfeng Gao. 2013. Beyond Left-to-Right: Multiple Decomposition Structures for SMT. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, June. Association for Computational Linguistics.