Offline Extraction of Overlapping Phrases for Hierarchical Phrase-Based Translation

Sariya Karimova, Patrick Simianer, Stefan Riezler

Computational Linguistics, Heidelberg University
69120 Heidelberg, Germany
{karimova,simianer,riezler}@cl.uni-heidelberg.de

Abstract

Standard SMT decoders operate by translating disjoint spans of input words, thus discarding information in form of overlapping phrases that is present at phrase extraction time. The use of overlapping phrases in translation may enhance fluency in positions that would otherwise be phrase boundaries, they may provide additional statistical support for long and rare phrases, and they may generate new phrases that have never been seen in the training data. We show how to extract overlapping phrases offline for hierarchical phrase-based SMT, and how to extract features and tune weights for the new phrases. We find gains of $0.3 - 0.6$ BLEU points over discriminatively trained hierarchical phrase-based SMT systems on two datasets for German-to-English translation.

1. Introduction

Decoding in SMT amounts to searching for the most probable (Viterbi) derivation of a target string given the source string. Standard SMT decoders perform at the same time a search for the optimal segmentation of the source sentence into disjoint spans of words, which are translated by rules encoding bi-phrases. This means that irrespective of whether phrases are contiguous [1], non-contiguous [2, 3], or hierarchical [4], the application of phrase rules at decoding time disallows overlapping words. However, the use of overlapping phrases might have several advantages: First, they may enhance fluency in positions that would otherwise be phrase boundaries. Second, overlapping phrases may provide additional statistical support for long and rare phrases extracted from the training data. Finally, and most importantly, overlapping phrases may constitute new phrases that have never been seen in the training data but may be applicable to the test data.

The few approaches that did attempt to integrate overlapping phrases into SMT decoding in the past [5, 6, 7] were handicapped mostly by the additional decoding complexity. The need to counterbalance exponential growth of the search space with very restrictive reordering constraints prevented these approaches to be competitive with state-of-the-art phrase-based SMT. The exception is Tribble et al. [8] who reported significant gains for using overlapping phrases over their own baseline. The key idea in this approach is to circumvent decoder integration and instead to generate overlapping phrases *offline*, by merging existing contiguous phrases into longer bi-phrases that have overlapping words in both source and target.

In this work, we will revive this approach, and extend it to hierarchical phrases. We show how to merge and filter overlapping phrases created from hierarchical and non-hierarchical phrases, and how to extract and tune features for the new phrases. An experimental comparison with a state-of-the-art hierarchical phrase-based decoder [9] shows gains of $0.3 - 0.6$ BLEU points on two datasets for German-to-English translation.

2. Related Work

The potential of overlapping phrases to improve fluency and to smooth prediction of long and rare phrases has been discovered independently in a few instances in prior work. The crux of most of these approaches is an efficient integration of overlapping phrases into decoding. For example, the exponential number of translation hypotheses arising from overlapping phrases has been managed in beam search decoding frameworks by reordering constraints that allow only adjacent non-overlapping phrases to be swapped [5, 7]. This reordering constraint seems to be too restrictive since it impacts translation quality in comparison to state-of-the-art phrase-based SMT.

Alternatively, sampling-based approaches [6] or graph-search techniques [10, 11, 12] have been used for decoding with overlapping phrases. These approaches suffer from search errors due to necessary abstractions in sampling or due to necessary approximations in adaptation of graph search algorithms to SMT decoding.

The work related closest to our approach is that of Tribble et al. [8, 13]. Their key idea is to circumvent decoder integration and instead to generate overlapping phrases in an offline manner. In contrast to our work, their approach is restricted to merging contiguous phrases. Furthermore, they extract a single feature (based on phrase-internal word alignments) for new phrases and do not learn discriminative weights. A similar idea has also been presented for Example-Based MT

[14, 15] where the focus is on combining given overlapping phrases by a new search algorithm.

An alternative to enriching the repository of phrases with overlapping phrase rules is the design of context-sensitive features for discriminative training. Target context is clearly exploited by large language models. Word-sense disambiguation inspired features [16] allow to exploit source context, and recent approaches successfully merged source and target context into a powerful decoding feature [17]. However, these approaches are orthogonal to our work.

3. Generating Overlapping Phrases with and without Variables

Hierarchical phrases can be formalized as rules of a synchronous CFG [4]. We denote terminals consisting of contiguous phrases by T, and the single non-terminal variable by NT. The key idea is to merge base rules into new rules by pivoting on overlapping words. We apply this idea to base rules consisting of terminals only (T rules) and to base rules including non-terminals (NT rules).

As a first step, we apply the technique of [18] to extract rules for German-to-English translation from the News Commentary and TED data (see Section 5.1). Tables 1 and 2 show the token counts of rule shapes for the extracted grammars.

We see that base rules consisting of terminals only (rule shape T-T) are quite frequent in the extracted grammars for both datasets. To these rules, the ideas of [8], namely merging all base rules that have overlapping words on both source and target can be applied directly. For base rules including non-terminals (rule shape including NT), merging of rules can be done at word overlaps in terminals at the head of one rule with terminals at the tail of another rule.

Because of the huge number of potential new rules, we apply several filtering steps to the merging process. For T rules, we firstly restrict our attention to base rules with more than one terminal on source and target side. Secondly, we apply count cutoffs of less than 5, 8, and 11 occurrences of base rules in the training set. Lastly, given the test set, we only store merged rules whose source sides are in principle applicable to the test set. For rules including NTs, we restrict our attention to base rules with exactly one NT on source and target. Furthermore, we consider only base rules that are seen at least 17, 20, or 23 times in the training set. Lastly, a pre-filtering based on applicability of merged rules to test set sources is done. Tables 3 and 4 show the counts of base rules and merged rules before and after filtering on the News Commentary and TED datasets.

Overall, these filtering steps resulted in a considerable number of new rules, i.e., rules that are unseen in the training set. Table 5 shows the percentages of overlapping phrase rules that are applicable to the test data, but are unseen in the training data, together with their actual use in the 1-best translation of the test data. We find that new rules are composed at a considerably higher percentage from base T rules than from base NT rules, resulting in a similar usage pattern

	News Commentary testset		TED testset	
	new	used	new	used
T 5	65.3	25.3	63.5	54.5
T 8	54.8	18.7	49.6	43.3
T 11	47.1	10.6	40.1	36.7
NT 17	21.7	2.5	37.8	10.8
NT 20	17.7	4.5	35.2	8.6
NT 23	15.3	5.5	32.7	6.9
T + NT	24.7	16.4	38.03	23.0

Table 5: Percentages of overlapping phrase rules composed from base rules and unseen in training ("new"), out of rules of the same form applicable to the test set, together with their usage in translating the test set ("used"), out of rules of the same form used to translate the test set.

(1) $X \rightarrow$ ⟨ es stellt sich heraus ||| it turns out ⟩

(2) $X \rightarrow$ ⟨ stellt sich heraus , dass ||| turns out that ⟩

(3) $X \rightarrow$ ⟨ es stellt sich heraus , dass ||| it turns out that ⟩

Figure 1: T rule (3) merged from rules (1) and (2).

of more T rules than NT rules used in 1-best translations of both test sets. As expected, these percentages are decreasing the more restrictive the count cutoffs are set. A combination of T and NT rules shows a pattern of composition and usage in between T rules and NT rules.

Across all extracted rules, the average number of words in merged rules is as little as 0.1 tokens higher than in base rules for News Commentary, and increases on average up to more than 1 token for the TED data set. For the majority of cases, the overlap is 1 token in source and target. In $1 - 2\%$ of the cases, the overlap is 2 tokens, and only 0.1% of the new phrases overlap in 3 or 4 tokens.

An example for a merger of two T rules (1) and (2) into a new rule (3), with an overlap of 3 source tokens and 2 target tokens, is given in Fig. 1. A merger of two rules including NTs is given in Fig. 2. Here, the overlap in target and source is 2 tokens.

4. Feature Extraction and Tuning

[8] use IBM model 1 word-level alignments of the merged phrases to directly assign probabilities to the new phrases. In this work, we use the SMT decoder `cdec` [9] that combines features into a log-linear model and offers several learners for discriminative tuning of weights.

We compare four feature configurations. First, we use

Count	Shape
359,406	T NT T - T NT T
270,813	NT T NT T - NT T NT T
267,528	T NT T NT - T NT T NT
155,250	T NT T NT T - T NT T NT T
129,400	T - T
109,447	T NT - T NT
104,924	NT T - NT T
99,615	NT T NT - NT T NT
58,824	T NT T NT - T NT NT
50,253	NT T NT T - NT NT T
35,015	T NT T NT T - T NT NT T
28,496	NT T NT T - T NT NT T
24,523	T NT T - T NT
23,821	T NT T NT - T NT NT T
22,705	NT T - T NT
20,658	NT T NT - T NT NT
20,639	NT T NT T - NT T NT
20,498	NT T NT - NT NT T
20,455	T NT - NT T

Count	Shape
20,003	NT T - T NT T
17,480	NT T NT - T NT T NT
17,276	T NT T - NT T
16,967	NT T NT T - T NT T NT T
16,559	T NT T NT - NT T NT
16,465	T NT T NT - T NT T NT T
15,965	NT T NT - NT T NT T
15,366	T NT - T NT T
11,736	T NT T NT T - T NT T NT
11,378	T NT T NT T - NT T NT T
10,625	NT T NT T - T NT T NT
8,691	T NT T NT - NT T NT T
2,693	NT T NT T - T NT NT
1,948	NT T NT - T NT NT T
1,525	T NT T NT - NT NT T
848	NT T NT - T NT T NT T
576	T NT T NT T - NT T NT
459	T NT T NT T - T NT NT
303	T NT T NT T - NT NT T

Table 1: Rule shapes in the grammar extracted from News Commentary.

(1) $X \rightarrow \langle$ ist wirklich X_1 , aber ||| is really X_1 , but $\rangle$

(2) $X \rightarrow \langle$, aber man X_1||| , but you $X_1 \rangle$

(3) $X \rightarrow \langle$ ist wirklich X_1 , aber man X_2|||
is really X_1 , but you $X_2 \rangle$

Figure 2: NT rule (3) merged from rules (1) and (2).

`cdec`'s implementation of lexical phrase probabilities for source words f and target words e:

$$MaxLexFgivenE = -\sum_i \log_{10} p_{max}(f_i|e) \quad (1)$$

and

$$MaxLexEgivenF = -\sum_i \log_{10} p_{max}(e_i|f). \quad (2)$$

Second, we add a new feature that indicates whether a rule is created by merging as follows:

$$NewRule = \begin{cases} 1 & \text{if the rule is new,} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Third, we calculate the following standard statistics among new rules that were merged from base rules extracted for the test set:

$$EgivenFCoherent = -\log_{10}(count\_EF/count\_F) \quad (4)$$

$$SampleCountF = \log_{10}(1 + count\_F) \quad (5)$$

$$CountEF = \log_{10}(1 + count\_EF) \quad (6)$$

$$IsSingletonF = \begin{cases} 1 & \text{if } count\_F = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

$$IsSingletonFE = \begin{cases} 1 & \text{if } count\_EF = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Last, we take inspiration from [19]'s adaptive features that combine counts from a lookup in post-editing data with counts from the suffix array sample extracted for the test set. In our case, this corresponds to combining count statistics for new rules only (denoted by subscript $\mathcal{L}$) with count statistics for base rules extracted for the test set (denoted by subscript $\mathcal{S}$):

$$EgivenFCoherent = -\log_{10}((count\_EF_{\mathcal{S}} + count\_EF_{\mathcal{L}})/ (count\_F_{\mathcal{S}} + count\_F_{\mathcal{L}})) \quad (9)$$

$$SampleCountF = \log_{10}(1 + count\_F_{\mathcal{S}} + count\_F_{\mathcal{L}}) \quad (10)$$

Count	Shape
373,500	T NT T - T NT T
284,364	T NT T NT - T NT T NT
277,682	NT T NT T - NT T NT T
204,562	T NT T NT T - T NT T NT T
97,485	T - T
92,133	T NT - T NT
86,469	NT T - NT T
85,518	NT T NT - NT T NT
47,617	T NT T NT - T NT NT
43,403	T NT T NT T - T NT NT T
38,121	NT T NT T - NT NT T
29,213	NT T NT T - T NT NT T
25,302	T NT T NT - T NT NT T
20,839	T NT T - T NT
20,173	NT T NT T - T NT T NT T
17,559	NT T NT T - NT T NT
17,328	NT T - T NT T
16,404	NT T NT - T NT T NT
16,087	T NT T NT - T NT T NT T
14,166	T NT T NT T - T NT T NT
14,039	NT T NT - NT T NT T
13,836	T NT T - NT T
13,476	T NT - T NT T
13,078	NT T NT - NT NT T
12,907	T NT T NT - NT T NT
12,893	NT T - T NT
12,658	T NT - NT T
12,376	NT T NT - T NT NT
10,454	T NT T NT T - NT T NT T
7,159	NT T NT T - T NT T NT
5,170	T NT T NT - NT T NT T
1,566	NT T NT - T NT NT T
1,368	NT T NT T - T NT NT
836	T NT T NT - NT NT T
813	NT T NT - T NT T NT T
496	T NT T NT T - NT T NT
343	T NT T NT T - T NT NT
234	T NT T NT T - NT NT T

Table 2: Rule shapes in the grammar extracted from TED talks.

$$CountEF = \log_{10}(1 + count\_EF_{\mathcal{S}} + count\_EF_{\mathcal{L}}) \quad (11)$$

$$MaxLexFgivenE = p_{max}(\tilde{f}|\tilde{e}) = -\sum_i \log_{10} p_{max}(f_i|e) \quad (12)$$

$$MaxLexEgivenF = p_{max}(\tilde{e}|\tilde{f}) = -\sum_i \log_{10} p_{max}(e_i|f) \quad (13)$$

$$IsSingletonF = \begin{cases} 1 & \text{if } count\_F_{\mathcal{S}} + count\_F_{\mathcal{L}} = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

$$IsSingletonFE = \begin{cases} 1 & \text{if } count\_EF_{\mathcal{S}} + count\_EF_{\mathcal{L}} = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (15)$$

$$NewRule = \begin{cases} 1 & \text{if the rule is new,} \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

Discriminative tuning is performed on the respective tuning sets of the News Commentary and TED data. We use the pairwise ranking learner of [20] for this purpose. In addition to the standard handful of dense feature, sparse features for rule shapes, rule identifiers, and bigrams in rule source and target are extracted from grammar rules.

NC	train	train-lm	tune	test
Sentences	136,227	180,657	1,057	1,064
Words de	3,005,252		26,205	23,593
Words en	2,909,346	3,797,500	25,660	22,518
TED	train	train-lm	tune	test
Sentences	139,563	158,641	1,172	746
Words de	2,195,030		21,270	11,831
Words en	2,332,370	2,715,777	21,679	12,734

Table 6: News Commentary and TED de-en parallel data.

5. Translation Experiments

5.1. Systems and Data

The data used in our experiments are the German-English parallel data provided in the News Commentary and TED releases of WMT 2007[1] and IWSLT 2013[2], respectively. Table 6 gives the basic data statistics for News Commentary (NC) and TED data.

The bilingual SMT system used in our experiments is the state-of-the-art SCFG decoder `cdec` [9][3]. We built grammars using its implementation of the suffix array extraction method described in [18]. Word alignments are built from all parallel data using fast_align [21]. SCFG models use the same settings as described in [4]. For language modeling, we built a modified Kneser-Ney smoothed 5-gram language

[1] `http://statmt.org/wmt07/shared-task.html`
[2] `http://www.iwslt2013.org/`
[3] `http://www.cdec-decoder.org`

News Commentary	Base rules	Merged rules	Unique	Applicable in test	Unique
all	129,400				
>1 token	72,322				
T 5	6,823	364,642	352,171	6,311	5,739
T 8	4,434	171,715	167,125	3,414	3,165
T 11	3,286	100,513	98,268	2,203	2,054
TED	Base rules	Merged rules	Unique	Applicable in test	Unique
all	97,485				
>1 token	62,671				
T 5	6,073	370,611	363,789	8,823	7,637
T 8	4,088	181,227	178,010	4,828	4,235
T 11	3,115	105,657	103,906	3,203	2,855

Table 3: Counts of base rules and merged rules with terminals only before and after filtering.

News Commentary	Base rules	Merged rules	Unique	Applicable in test	Unique
all	694,105				
NT 17	14,107	563,980	556,476	18,588	14,919
NT 20	11,592	324,790	319,919	13,794	11,039
NT 23	9,774	198,447	194,880	10,915	8,690
TED	Base rules	Merged rules	Unique	Applicable in test	Unique
all	643,132				
NT 17	14,684	1,980,618	1,940,402	34,696	28,293
NT 20	12,256	1,345,298	1,316,680	26,856	21,750
NT 23	10,334	908,066	887,474	21,118	16,938

Table 4: Counts of base rules and merged rules with nonterminals before and after filtering.

model [22, 23].

All data were normalized, tokenized and lowercased; German compounds were split. For tokenization, lowercasing and other preprocessing steps we used the scripts distributed with the Moses SMT toolkit [24]. For compound splitting in German texts a standard empirical approach of [25] was employed.

5.2. Experimental Results

Table 7 shows BLEU [26] results for MERT [27] optimization of dense feature weights, and for pairwise ranking [20] optimization of sparse feature weights. MERT runs were repeated three times to account for optimizer instability [28]. The pairwise ranking technique was stable in this respect. Statistical significance is measured using Approximate Randomization [29, 30] where result differences with a p-value smaller than 0.05 are considered significant.

In order to investigate a possible correspondence of the patterns of composition and usage shown in Table 5, we evaluate overlapping phrases merged from base T rules and base NT rules separately. Table 8 shows BLEU results for different frequency cutoffs for base rules (see Section 3) and different feature sets (see Section 4) on the News Commentary data for German-to-English translation. All results are nominal improvements over the PairRank baseline in Table 7, with several statistically significant result differences. Best results, namely an improvement of 1.3 BLEU points over the MERT baseline, and a gain of 0.6 BLEU points over the pairwise-ranking baseline are obtained for merging overlapping rules from base T rules, using all adaptive features. Best results for merging rules from NT rules are slightly lower.

Table 9 evaluates the same configurations of base rule cutoffs and features on the TED talk data. Here the best result is a nominal improvement of 0.3 BLEU points over the baseline, obtained by merging rules from base T rules. Again, this result is slightly better than merging rules from base NT rules. However, in case of the TED data, no result difference

	News Commentary	TED
MERT	24.95	25.94
PairRank	25.69†	25.90

Table 7: Baseline results for News Commentary and TED talks German-to-English translation. Statistically significant differences to MERT are denoted with †.

Cutoff	Features (1)-(2)	(1)-(3)	(1)-(8)	(9)-(16)
T 5	25.83	25.82	25.83	*25.86*
T 8	25.99	25.99	26.02	**26.24**†
T 11	25.93	26.08†	*26.12*†	25.75
NT 17	25.76	*26.13*†	26.01	25.84
NT 20	*26.14*†	25.70	25.89	25.97
NT 23	25.76	25.90	*26.22*†	25.82

Table 8: Results for News Commentary, German-English translation. Best results for a certain feature set in *italics*, best result overall in **bold**. Significant differences compared to the PairRank baseline of Table 7 are denoted with †.

Cutoff	Features (1)-(2)	(1)-(3)	(1)-(8)	(9)-(16)
T 5	25.89	25.94	25.93	*26.00*
T 8	25.96	25.95	**26.23**	26.01
T 11	25.84	*26.13*	25.79	25.98
NT 17	25.71	25.93	*26.04*	25.82
NT 20	25.50	26.03	26.02	*26.10*
NT 23	25.57	*26.04*	26.01	25.78

Table 9: Results for TED, German-English translation. Best results for a certain feature set in *italics*, best result overall in **bold**. Significant differences compared to the PairRank baseline of Table 7 are denoted with †.

is statistically significant compared to the PairRank baseline.

Table 10 shows an evaluation for a combination of overlapping phrase rules merged from base T rules and base NT rules. Combining the best configurations for generating overlapping phrases from T-only and NT base rules yields results that are about 0.1 BLEU point lower than the best results in Tables 8 and 9. Result differences are statistically significant for News Commentary, but not for TED experiments.

Overall, we find a correspondence of BLEU improvements shown in Tables 8, 9, 10 with the pattern of composition and usage shown in Table 5, with higher gains and higher usage for T rules compared to NT rules.

6. Conclusion

We presented an application of the idea of offline merging of bi-phrases into longer phrases with overlapping words to the framework of hierarchical phrase-based translation. The advantages of overlapping phrases in translation are enhanced fluency in positions that would otherwise be phrase boundaries. Furthermore, a large number of new phrases can be generated that have never been seen in the training data but are applicable to the test data. Our approach maintains all the benefits of using overlapping phrases at translation time, without the pain of having to modify the decoder to deal with overlapping phrases.

Our experimental results on two datasets for German-to-English translation show gains of $0.3-0.6$ BLEU points over a baseline system that implements discriminatively trained hierarchical phrase-based SMT. We conjecture that improved quality at translation time might be worth the overhead of building overlapping rules at phrase extraction time.

	News Commentary	TED
T + NT	**26.15**†	**26.10**

Table 10: Best results for combination of NT and T overlapping phrases on TED and News Commentary, German-English translation. Significant differences compared to the PairRank baseline of Table 7 are denoted with †.

7. References

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the Human Language Technology Conference and the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'03)*, Edmonton, Cananda, 2003.

[2] M. Simard, N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, and K. Yamada, "Translating with non-contiguous phrases," in *Proceedings of the Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, BC, Canada, 2005.

[3] M. Galley and C. D. Manning, "Accurate non-hierarchical phrase-based translation," in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (HLT-NAACL)*, Los Angeles, CA, 2010.

[4] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, 2007.

[5] M. Kääriäinen, "Sinuhe - statistical machine translation using a globally trained conditional exponential family translation model," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Singapore, 2009.

[6] B. Roth, A. McCallum, M. Dymetman, and N. Cancedda, "Machine translation using overlapping alignments and samplerank," in *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, CO, 2010.

[7] Z. Wang and J. Shawe-Taylor, "A kernel regression framework for SMT," *Machine Translation*, vol. 24, pp. 87–102, 2010.

[8] A. Tribble, S. Vogel, and A. Waibel, "Overlapping phrase-level translation rules in an SMT engine,"

in *Proceedings of the International Conference on NLP and Knowledge Engineering (NLP-KE)*, Beijing, China, 2003.

[9] C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik, "cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models," in *Proceedings of the ACL 2010 System Demonstrations*, Uppsala, Sweden, 2010.

[10] C. Cortes, M. Mohri, and J. Weston, "A general regression framework for learning string-to-string mappings," in *Predicting Structured Data*, G. Bakhir, T. Hofmann, and B. Schölkopf, Eds. Cambridge, MA: The MIT Press, 2007, pp. 143–168.

[11] N. Serrano, J. Andrés-Ferrer, and F. Casacuberta, "On a kernel regression approach to machine translation," in *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis*, Póvoa de Varzim, Portugal, 2009.

[12] E. Biçici and D. Yuret, "Regmt system for machine translation, system combination, and evaluation," in *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT)*, Edinburgh, Scotland, UK, 2011.

[13] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao, and A. Waibel, "The CMU statistical machine translation system," in *Proceedings of MT Summit IX*, New Orleans, LA, 2003.

[14] R. D. Brown, R. Hutchinson, P. N. Bennett, J. G. Carbonell, and P. Jansen, "Reducing boundary friction using translation-fragment overlap," in *Proceedings of MT Summit IX*, New Orleans, LA, 2003.

[15] R. Hutchinson, P. N. Bennett, J. G. Carbonell, P. Jansen, and R. D. Brown, "Maximal lattice overlap in example-based machine translation," Computer Science Department, Carnegie Mellon University, Paper 324, Tech. Rep., 2003.

[16] Y. S. Chang, H. T. Ng, and D. Chiang, "Word sense disambiguation improves statistial machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, 2007.

[17] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, "Fast and robust neural network joint models for statistical machine translation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, MD, 2014.

[18] A. Lopez, "Hierarchical phrase-based translation with suffix arrays," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, Czech Republic, 2007.

[19] M. Denkowski, C. Dyer, and A. Lavie, "Learning from post-editing: Online model adaptation for statistical machine translation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*, Gothenburg, Sweden, 2014.

[20] P. Simianer, S. Riezler, and C. Dyer, "Joint feature selection in distributed stochastic learning for large-scale discriminative training in SMT," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Korea, 2012.

[21] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of ibm model 2," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, Georgia, 2013.

[22] K. Heafield, "KenLM: faster and smaller language model queries," in *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation (WMT'11)*, Edinburgh, UK, 2011.

[23] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified kneser-ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, Sofia, Bulgaria, 2013.

[24] P. Koehn, H. Hoang, A. Birch, C. Callison-Birch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic, 2007.

[25] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, Hungary, 2003.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," Yorktown Heights, N.Y., Tech. Rep. IBM Research Division Technical Report, RC22176 (W0190-022), 2001.

[27] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the Human Language Technology Conference and the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'03)*, Edmonton, Cananda, 2003.

[28] J. Clark, C. Dyer, A. Lavie, and N. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, OR, 2011.

[29] E. W. Noreen, *Computer Intensive Methods for Testing Hypotheses. An Introduction*. New York: Wiley, 1989.

[30] S. Riezler and J. Maxwell, "On some pitfalls in automatic evaluation and significance testing for MT," in *Proceedings of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI, 2005.