Hierarchical Hybrid Translation between English and German

Yu Chen **Andreas Eisele**
German Research Center for Artificial Intelligence (DFKI) GmbH
Saarbrücken, Germany
{Yu.Chen, Andreas.Eisele}@dfki.de

Abstract

We present new results from a hybrid combination of rule-based machine translation (RBMT) with a variant of statistical machine translation (SMT) that supports hierarchical structures and is therefore able to preserve more of the linguistic structures obtained from the RBMT system than versions of SMT that operate on flat phrases alone. Having shown in (Chen and Eisele, 2010) for the first time that a tighter integration of hierachical MT systems from different paradigms leads to consistent improvements for translation from German to English in various experimental settings, the current paper generalizes the approach to translation from English to German, where we observe similar improvements.

These findings indicate that hybrid combinations of MT paradigms can benefit from structural similarities in the underlying models, which makes us expect even stronger benefits from a tight integration of different approaches.

.

1 Introduction

Research in machine translation has made significant progress in recent years. Statistical machine translation (SMT) systems became especially dominant in this area, motivated by the ability to create new systems from existing training data without much effort, but also encouraged by the fact that many evaluation measures that solely rely on string matching allow to implement incremental improvements without having to solve the really hard issues. However, human assessments (Callison-Burch et al., 2009) show that rule-based systems can still translate better than SMT systems in many cases. The errors produced by different types of systems are somehow complementary (Thurmair, 2004). In addition to improving MT techniques for certain kinds of systems, another stream of research in MT aims at combining existing methods, that is, to build hybrid approaches.

One way to integrate SMT and RBMT is to apply a variant of standard statistical methods to induce information from translations made by RBMT systems and to incorporate the information into the core of a SMT system (Eisele et al., 2008). Most RBMT systems benefit from the large accurate lexicons and complex grammars that took enormous human efforts through decades. By the hybrid combination, the linguistic information supplied in RBMT systems are expected to fill gaps in lexical knowledge of the SMT system, which is particularly lacking when translating texts in domains different from the training data.

The hybrid framework in (Eisele et al., 2008) outperforms the original SMT system that acts as the core, however the improvements over the RBMT systems were not consistent when the RBMT systems actually created better translations compared to the SMT system. It is mostly because this hybrid combination method is unable to make use of well-formedness in RBMT, which, in fact, is one of the most significant advantages of systems based on linguistic knowledge. On the contrary, the correct syntactic structures are decomposed into small pieces that are no longer connected to each other any more. Similar problems also exist for post-editing approaches (Dugast et al., 2007).

Furthermore, similar to a general system combination approach, the improvement of such a hybrid system greatly depends on the number and the diversity of the systems. Most results on this track report improvements only with more than 2 systems. Excluding RBMT systems in system combination tasks may degrade the overall performance by 6% (Leusch et al., 2009). However, it is rather unrealistic in practice to use 6 RBMT engines in addition to the SMT core as described in (Eisele et al., 2008) as most high quality customized RBMT systems are not freely available. Thus, we restrict ourselves to hybrid architectures involving only one RBMT system and one SMT decoder here.

This paper substitutes the core SMT system with hierarchical phrase-based SMT system inspired by (Chiang, 2007) in the hope of preserving more syntactic structures while introducing additional lexical information to the SMT system. For our experiments, we use Joshua as the decoder (Li et al., 2009). The experiments in translation from German to English documented in (Chen and Eisele, 2010) showed that the hybrid system was able to outperform both its SMT and RBMT components significantly. We also compared our system to a setup that follows (Eisele et al., 2008) and achieve much more reliable improvements over both in-domain and out-of-domain tasks in terms of BLEU score (Papineni et al., 2001). These results motivated us to extend the approach by inverting the language pair, which is the main focus of the current paper.

2 Previous Work

There have been various approaches proposed for combining MT systems into multi-engine architectures since (Frederking and Nirenburg, 1994). The most straightforward method is to attempt to select the best output from a number of systems so as to form a multi-engine system from the group of independent systems. Individual hypotheses in such setups remain as is (Tidhar and Küssner, 2000; Akiba et al., 2001; Callison-Burch and Flournoy, 2001; Akiba et al., 2002; Nomoto, 2004; Eisele, 2005). More sophisticated combinations aim at recombining the best pieces available from multiple hypotheses into a new utterance that can be better than the best of the given candidates, as described in (Rayner and Carter, 1997; Hogan and Frederking, 1998; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Matusov et al., 2006; Rosti et al., 2007).

When recombining multiple translations, it is essential to find the correspondences between alternative renderings of a source-language expression proposed by different MT systems. Due to the distinct errors and word ordering in hypotheses from different systems, it is generally difficult to identify the alignment between the source and the outputs or among the MT outputs. Therefore, a central component of a multi-engine system is a specialized module for word alignment.

Another key to a competitive recombination system is how to select the most proper combination of alternative building blocks. It is not only necessary to consider the plausibility of each individual building block but also crucial to take into account the relation between the building blocks. Although many method determine the word order by selecting a skeleton before recombination, recent work in system combination allows flexible word orders determined by various features (He and Toutanova, 2009; Zhao and He, 2009). Such a optimization process is almost identical to the search in a SMT decoder that seeks naturally sounding combinations of highly probable partial translations.

3 Architecture

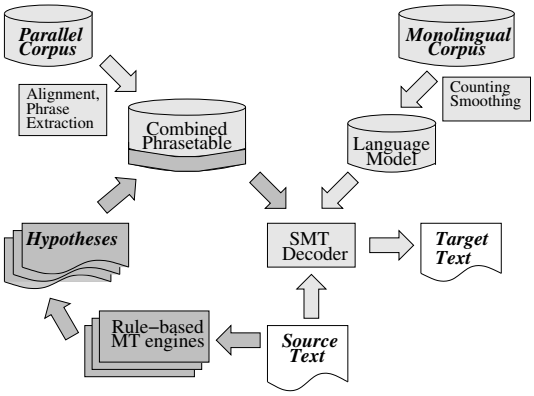

Figure 1: Hybrid architecture of the system

The system as shown in Figure 1 uses a SMT system with a modified translation model as the central element. This modification is performed by augmenting the standard phrase table with entries extracted from translations produced by a rule-based MT system. Given such additional knowledge from the RBMT, the SMT decoder makes decision for the final translation using the standard algorithm by recombining the building blocks that have been contributed by both the RBMT system

and the original SMT system.

It differs from the design proposed in (Eisele et al., 2008) mainly in two aspects: the formalism of translation models and the method of combining translation models from different sources. Accordingly, we adapt the corresponding decoder and the training procedure. The following sections will give more details.

3.1 Components

A phrase-based SMT decoder, as implemented in the Moses Toolkit (Koehn et al., 2007), works on correspondences between contiguous sequences of words from both languages. Because of this limitation, it is fairly difficult for such models to reflect global syntactic structures from the training data. The correct translations are frequently outside of the search space of the SMT decoder. Since many syntactic constructions in RBMT do not necessarily involve complete sequences, the grammaticality of the translations can be easily destroyed during the phrase extraction phase and the damage is unlikely to be recovered in later stages, which explains why the previous hybrid system (Eisele et al., 2008) does not always outperform the RBMT systems.

Instead of a phrase-based SMT decoder, we employ a parsing-based decoder, which assumes a probabilistic synchronous context-free grammar (SCFG) (Venugopal et al., 2007) comprising a set of source-language terminal symbols, a set of target-language terminal symbols, a shared set of non-terminal symbols and a set of rules. The composite weight of a translation is a linear combination of feature function weights and feature function values. Given a source sentence, the decoder uses a chart-parsing algorithm with beam search to construct a target sentence that has the best composite weight among all possible derivations. This paper only considers hierarchical rules, that is, representations of translation correspondence in rather flat structures (Chiang, 2007). An example of such hierarchical rules for German-to-English translation, ignoring the feature values, is

$$X \rightarrow \langle \textit{den}\ X_1\ \textit{habe ich}\ X_2, \textit{I have}\ X_2\ X_1 \rangle.$$

This rule generalizes the correspondence between phrase pairs such as "den *Termin* habe ich *vergessen* — I have *forgotten* the *appointment*". It is much harder to achieve this kind of generalization with the same training data by pure phrase-based models. Syntax-based translation models generalize even better than hierarchical models, but usually require additional resources for parsing texts in at least one of the languages. Hierarchical decoding is relatively less demanding from this aspect. Meanwhile, RBMT translations built upon carefully crafted grammatical rules involve much less sophisticated syntactic structures than human translations. Hierarchical models should be able to capture such knowledge to a certain extent.

The rule-based system we use here is *Lucy* (Alonso and Thurmair, 2003), a commercial translation engine with its roots in the transfer-based METAL system that has evolved for years. The system contains various linguistic components, including: the lexicons (mono- and bilingual), analysis grammars, transfer rules, a generation module and other linguistic knowledge sources. According to human assessments carried out in recent WMT shared tasks (Callison-Burch et al., 2008; Callison-Burch et al., 2009), *Lucy* was among the best systems for German-to-English and English-to-German translation, superior to most SMT systems for tasks in the news domain. As a commercial MT system, *Lucy* do no supply any intermediate analysis. The RBMT system is used as a black box in our hybrid system.

3.2 Phrase table from RBMT outputs

The major component of this architecture is the phrase table generated with the RBMT system. We apply the general approach proposed in (Eisele et al., 2008) except that our phrase table also contains phrases with gaps, modeled as a SCFG.

Given no explicit word alignment information from the RBMT system, construction of the phrase table consists of two steps: word alignment and phrase extraction. We construct this RBMT phrase table using a bilingual corpus (RBMT corpus) that consists of given input texts and corresponding translation produced by the RBMT system.

This bilingual RBMT corpus is usually far smaller than any commonly used training corpora. The standard statistical word alignment algorithm is most likely unable to acquire reliable alignment models for such a small corpus. Therefore, we create the alignment between the input and the RBMT outputs based on existing alignment models that were generated with sufficiently large corpora. There are two alternative choices for the base alignment models: alignment models specifically trained for RBMT translations and the align-

ment models trained for the SMT core system that already exist in the hybrid system. The alignment procedure starts with mapping the vocabularies in the RBMT corpus to the vocabularies corresponding to the base model and initiating estimates with the base model, thereby building a more precise model for the RBMT corpus. The final word alignment is determined on the resulting alignment models with *grow-diag-final-and* heuristic to include diagonal neighboring words to existing aligned points for higher coverage.

Given the word alignment between the input and RBMT translation, hierarchical rules, phrasal translations and their associated model parameters, are extracted for the complete RBMT corpus using suffix arrays (Lopez, 2008). The extraction setup needs to be identical to translation model construction of the SMT core system, that is, both allow the same number of gaps, the same maximal length of phrases, etc. Still, the model parameters in this model are not directly comparable to the parameters of the core model, since the estimation of the hierarchical model is based on the RBMT corpus, which is much smaller than that used for estimating the model of the core system. The estimates in the RBMT model tend to have higher values.

3.3 Combined Phrase-table

The *union* of the two models, the RBMT model and the statistical model constructed from the training corpus, forms the combined phrase table to be used for final translations. We simply add the entries that only appear in RBMT translations to the original phrase table. The proposal in this paper differs from the system in (Eisele et al., 2008) mainly in handling the features from both models.

The previous method extends phrase tables by adding one binary feature for each individual system, including the SMT core. For a phrase pair, the value of such a feature indicates whether the corresponding system produced this phrase pair.

In the current setup presented in this paper, we retain all features in both translation models, one set from training data and the other set from RBMT translations. In other words, the standard hierarchical grammar with 3 features leads to a hierarchical grammar (equivalent to phrase tables in Moses) in our hybrid system that has 6 features in total. Figure 2 illustrates a few entries from the combined phrase table we used in our experiments. No more than 2 gaps are allowed in the rule extraction. All 6 feature values are listed when a phrase pair exist in both models, such as the first three entries in the list. Phrase pairs in the rest of list do not appear in both models. In this case, the missing feature values are set to 1.0, which yields to 0.0 in log-linear models.

We hope minimum error rate training (MERT) (Och, 2003) is able to balance between features on different bases. As for the hybrid setup with pure phrase-based models, this method would produce 10 columns in the combined phrase table provided 5 features in a standard setup. We are aware of the risk that this combination method may introduce too many features and hence too many opportunities for over-learning at the MERT optimization step, but as this is our first attempt to exploit a new variant of hybrid systems, this approach should be regarded as simple way to explore the feasibility of a new setup.

4 Experiments

4.1 Data and configurations

The experiments we conducted involve both in-domain and out-of-domain tasks. We used release v4 of the Europarl corpus (Koehn, 2005)[1] as the training corpus. Accordingly, the in-domain test inputs are texts from Europarl and the out-of-domain tests are news texts. We tested the hybrid system with two test sets from the WMT 2008 Shared Task[2]. Our development data also consists of the two corresponding sets from the test data of the WMT 2007 Shared Task[3].

We built the core SMT system with the open-source software package *Joshua* (Li et al., 2009). The hierarchical models are trained on sentences with less than 80 tokens. The statistical system also includes a 5-gram language model that was constructed on the target side of the parallel corpus using SRILM toolkit (Stolcke, 2002). Then, we extracted a relatively small hierarchical model from *Lucy*'s translation of the development set and merged it into the large one.

When using the *Joshua* decoder, it is straightforward to use Z-MERT (Zaidan, 2009) rather than the other implementation for minimum error rate training (MERT) (Och, 2003). As a standalone open source tool, Z-MERT is highly op-

[1] http://www.statmt.org/wmt09/translation-task.html
[2] http://www.statmt.org/wmt08/
[3] http://www.statmt.org/wmt07/

source	target	SMT features			RBMT features		
zum	at the	1.9800	1.8958	2.4356	1.9542	1.8255	2.1297
der X_1 , die	the X_1 which	1.2552	1.7833	1.6795	1.0543	1.4845	1.4218
der X_1 der X_2	of the X_1 of the X_2	1.3979	1.1264	1.8677	1.58546	1.0686	1.5023
landesgrenzen	boundaries	1.1563	1.7584	1.1139	1.0	1.0	1.0
X_1 abgeschlossen sein	X_1 be finalised	1.8450	1.7077	1.8586	1.0	1.0	1.0
fakten X_1 der X_2	facts X_1 against the X_2	1.0413	1.0455	3.613	1.0	1.0	1.0
nach den	after that	1.0	1.0	1.0	1.1139	2.1035	2.129
auf der X_1	on which X_1	1.0	1.0	1.0	1.3617	1.4243	2.1300
die X_1 von X_2	who X_1 of X_2	1.0	1.0	1.0	1.3802	1.2750	1.9222

Figure 2: Example entries from combined phrase table

timized for time and space efficiency and apparently faster than Moses' C++ MERT implementation. Z-MERT also works with Moses. However, it is unclear how the performance of both approaches compares, which needs further investigation. The feature weights for the enlarged model are determined by Z-MERT on the respective development sets with the aim to maximize BLEU score.

Similar to the choice of MERT implementation, we used the Berkeley Aligner (DeNero and Klein, 2007) to align our training data. As an alternative to GIZA++ (Och and Ney, 2003), the Berkeley Aligner combines the innovations of recent work in unsupervised word alignment. The joint training of IBM models was able to reduce alignment error rate by 32% relative to GIZA++. When aligning RBMT translations with corresponding source texts based on an alignment model constructed with the complete training data set, we use an existing adaption of GIZA++.

As for testing, we translated the test sets with *Lucy* and constructed corresponding hierarchical models. For each translation task, we integrate the *Lucy* model into the original. The feature weights obtained with the development set are used for translations with the corresponding combined model. For comparison, we built another hybrid system with phrase-based SMT core using Moses Toolkit with a very similar setup: the same data sets (training, tuning and testing), identical word alignments, the same language model and the identical MERT program.

4.2 Results

We evaluated all the translations with BLEU. The results are shown in Table 1. The scores indicate that the hybrid system combining *Joshua* and *Lucy* is able to consistently produce translations better than both systems in isolation. It is obvious

	de-en		de-en	
	EP	NC	EP	NC
Lucy	16.40	17.02	11.23	13.01
Moses	27.27	16.66	19.42	10.27
+Lucy	27.26	16.06	19.19	12.35
Joshua	27.51	16.24	20.69	10.48
+Lucy	**27.52**	**17.69**	**20.89**	**13.21**

Table 1: BLEU scores from both in-domain and out-of-domain experiments

that *Joshua* produce better translations (over 10 BLEU points) than *Lucy* for in-domain tests, however the hybrid system built upon *Joshua* manage to achieve performance close to the SMT system although translations produced by *Lucy* are also consider alongside the human translations in the training corpus. On the other hand, The improvement the hybrid system made was more significant for out-of-domain tests. The difference between the hybrid system and the SMT core increased to nearly 1.5 BLEU. In other words, the hierarchical approach is able to capture the unseen information when RBMT system delivers it even when it is only represented vaguely in the translations.

Figure 3 are example translations produced by all 5 systems in the experiments, including both in-domain and out-of-domain tests. Compared to the stand-alone *Joshua*, our hybrid system clearly benefited from integration with *Lucy*. The system not only made better selection of phrase translations provided by *Lucy* but also adjust the translations with more well-formed overall syntactic structures close to the RBMT translation. In the first example, the SMT systems did not consider the appropriate translation correspondence between the words "*unter*" and "*among*" as translated by the RBMT system. It was translated in a more com-

In-domain

Source	Ich möchte Sie daran erinnern, dass sich unter unseren Verbündeten entschiedene Befürworter dieser Steuer befinden.
Reference	Let me remind you that our allies include fervent supporters of this tax.
Lucy	I would like to remind you of there being decisive proponents of this tax among our allies.
Moses	I would like to remind you that under our allies are strong supporters of this tax.
+Lucy	I would like to remind you that there are among our allies in favour of this tax.
Joshua	I would like to remind you that , under our allies are strong supporters of this tax.
+Lucy	I would like to remind you that there are strong supporters of this tax among our allies.

Out-of-domain

Source	So kooperieren die Hochschulen schon aus Tradition mit den Nachbarländern.
Reference	The university-level institutions' cooperation with the neighboring countries, for instance, is part of a tradition.
Lucy	So the colleges co-operate already from tradition with the neighbor countries closely.
Moses	So the universities from tradition cooperate closely with the neighbouring countries.
+Lucy	So the colleges co-operate closely with the neighbouring already from tradition.
Joshua	So cooperate closely with the neighbouring the universities from tradition.
+Lucy	So the universities, already from tradition, co-operate closely with the neighbouring countries.

Figure 3: Translation examples

Base alignment model	$\emptyset$	Europarl
Moses+Lucy (EP)	19.37	19.19
Moses+Lucy (NC)	12.50	12.38
Joshua+Lucy (EP)	20.83	20.89
Joshua+Lucy (NC)	13.17	13.21

Table 2: BLEU scores of English-German translations with/without base model for aligning RBMT outputs

mon way into "*under*" instead. Both hybrid systems successfully included this translation pair in their phrase tables, however only the system with hierarchical core reallocated the preposition phrase after the head "*stronger supporter*" of the noun phrase. The other hybrid system dropped the head phrase, which leads to an inadequate and non-fluent translation. This is more obvious for out-of-domain tests as illustrated in the second example. The subject of this sentence was missing in the translation given by the original hierarchical system but recovered in the hybrid setup. The phrase-based hybrid system was not able to achieve similar improvement and some key nouns such as "*countries*" are neglected in translation.

Contradictory to the results reported in (Eisele et al., 2008), we were not able to observe clear improvements with the combined system built on Moses even for out-of-domain tests. One possible cause may be the many additional features in the translation model after integration, which makes it more difficult for MERT to reach an optimal feature weight set. In fact, the tuning process of the hybrid system took much longer time compared to its SMT core. More importantly, unlike the previous approach including 6 RBMT systems, our system only consists of only one, which appeared to produce extremely distinct translations compared to the core SMT system. A smaller number of RBMT systems also implies much less linguistic and lexical knowledge that can be derived from the RBMT translations.

4.3 Alignment from RBMT outputs to inputs

Our hybrid setup includes a large-scale base model that was constructed from the Europarl corpus by aligning the translations produced by *Lucy* back to the original input texts. To understand the effect of the base model, we conducted an additional set of experiments for English-German translation with alignments that were built without any base models. Whereas using the base model leads to different alignment results for up to 90% of the sentences, Table 2 shows no significant difference in translation quality in the hybrid outputs.

The base model supposedly provides more evidence on the correspondence between words in the alignment process so that the resulting alignments should be more precise and more consistent with the base model. In other words, correspondences that occur in both the large parallel texts and the RBMT translations are considered more plausible. Since these alignment results are used to generate the RBMT models and eventually combined with the original translation model, the alignment points appearing in both data sets would always lead to phrase pairs with higher overall feature weights given the design of our hybrid system. Therefore,

we reckon that the key factor of better combination is the grammar extraction step rather than the word alignment. This requires further investigations.

5 Conclusion and Future Work

This paper has introduced a novel approach to combine machine translation systems from different schemes. We integrate a commercial RBMT system with hierarchical SMT system by extracting SCFG rules from RBMT translations. The hybrid system inherits the lexicons from both subsystems as well as other merits of each system, including local syntactic constructions defined in RBMT system and the high fluency thanks to the statistical language model.

In order to understand the potential of this hybrid setup, we conducted a series of experiments for German-English and English-German translation. The variation to the previous approach leads to significant improvement over both individual sub-systems and hybrid system built with previous approaches. The improvement for out-of-domain tests was almost 1.5 BLEU points. In addition, we also investigate the translations manually. This evaluation provides strong evidence that we are going into a highly promising direction.

The results reported in this paper are still somewhat preliminary in the sense that many possible (including some desirable) variants of the setup could be tried in the future. For instance, a large language model trained on out-of-domain data should help our approach to achieve bigger improvements. Since hierarchical models have given us clear advantages over pure phrase-based models for learning from RBMT translation, we reckon a tighter integration of SMT and RBMT will eventually lead to significant progress. Such a hybrid system requires more insight into the RBMT system and more careful tackling of the SMT system.

Acknowledgements

This work was supported by European Community through the EuroMatrix Plus project (ICT-231720) funded under the Seventh Framework Programme for Research and Technological Development.

References

Akiba, Yasuhiro, Kenji Imamura, and Eiichiro Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain.

Akiba, Yasuhiro, Taro Watanabe, and Eiichiro Sumita. 2002. Using language and translation models to select the best among outputs from multiple mt systems. In *COLING*.

Alonso, Juan A. and Gregor Thurmair. 2003. The Comprendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA.

Bangalore, Srinivas, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *ASRU*, Italy.

Callison-Burch, Chris and Raymond S. Flournoy. 2001. A program for automatically selecting the best output from multiple machine translation engines. In *Proc. of MT Summit VIII*, Santiago de Compostela, Spain.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Chen, Yu and Andreas Eisele. 2010. Integrating a rule-based with a hierarchical translation system. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*.

Chiang, David. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

DeNero, John and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

Dugast, Loïc, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223, Prague, Czech Republic, June. Association for Computational Linguistics.

Eisele, Andreas, Christian Federmann, Hervé Saint Amand, Michael Jellinghaus, Teresa Herrmann, and Yu Chen. 2008. Using moses to integrate multiple rule-based machine translation engines into a hybrid system. In *the Third Workshop on Statistical Machine Translation*, pages 179–182, Columbus, Ohio, June. Association for Computational Linguistics.

Eisele, Andreas. 2005. First steps towards multi-engine machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 155–158, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Frederking, Robert E. and Sergei Nirenburg. 1994. Three heads are better than one. In *ANLP*, pages 95–100.

He, Xiaodong and Kristina Toutanova. 2009. Joint optimization for machine translation system combination. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1202–1211, Morristown, NJ, USA. Association for Computational Linguistics.

Hogan, Christopher and Robert E. Frederking. 1998. An evaluation of the multi-engine MT architecture. In *Proceedings of AMTA*, pages 113–123.

Jayaraman, Shyamsundar and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proc. of EAMT*, Budapest, Hungary.

Koehn, Philipp, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of Annual meeting of the Association for Computation Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech, June.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit 2005*.

Leusch, Gregor, Evgeny Matusov, and Hermann Ney. 2009. The RWTH system combination system for WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 51–55, Athens, Greece, March. Association for Computational Linguistics.

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*, pages 135–139, Athens, Greece, March. Association for Computational Linguistics.

Lopez, Adam David. 2008. *Machine translation by pattern matching*. Ph.D. thesis, College Park, MD, USA. Adviser-Resnik, Philip S.

Matusov, Evgeny, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *In Proc. EACL*, pages 33–40.

Nomoto, Tadashi. 2004. Multi-engine machine translation with voted language model. In *Proc. of ACL*.

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

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Rayner, Manny and David M. Carter. 1997. Hybrid language processing in the spoken language translator. In *Proc. ICASSP '97*, pages 107–110, Munich, Germany.

Rosti, Antti-Veikko, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr. 2007. Combining translations from multiple machine translation systems. In *Proceedings of HLT-NAACL*, pages 228–235, Rochester, NY, April 22-27.

Stolcke, Andreas. 2002. SRILM - an extensible language modeling toolkit. In *the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, Denver, Colorado.

Thurmair, Gregor. 2004. Comparing rule-based and statistical MT output. In *LREC*.

Tidhar, Dan and Uwe Küssner. 2000. Learning to select a good translation. In *COLING*, pages 843–849.

Venugopal, Ashish, Andreas Zollmann, and Vogel Stephan. 2007. An efficient two-pass approach to synchronous-cfg driven statistical mt. In Sidner, Candace L., Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, *HLT-NAACL*, pages 500–507. The Association for Computational Linguistics.

Zaidan, Omar F. 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.

Zhao, Yong and Xiaodong He. 2009. Using n-gram based features for machine translation system combination. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 205–208, Morristown, NJ, USA. Association for Computational Linguistics.