Combining Multi-Engine Machine Translation and Online Learning through Dynamic Phrase Tables

Rico Sennrich
Insitute of Computational Linguistics
University of Zurich
Binzmühlestr. 14
CH-8050 Zürich
sennrich@cl.uzh.ch

Abstract

Extending phrase-based Statistical Machine Translation systems with a second, dynamic phrase table has been done for multiple purposes. Promising results have been reported for hybrid or multi-engine machine translation, i.e. building a phrase table from the knowledge of external MT systems, and for online learning. We argue that, in prior research, dynamic phrase tables are not scored optimally because they may be of small size, which makes the Maximum Likelihood Estimation of translation probabilities unreliable. We propose basing the scores on frequencies from both the dynamic corpus and the primary corpus instead, and show that this modification significantly increases performance. We also explore the combination of multi-engine MT and online learning.

1 Introduction

Two recent trends in Machine Translation are multi-engine MT, and online learning. In multi-engine MT, the aim is to combine the strengths of different MT systems to perform better than any single system. Online learning is of high interest in the field of interactive MT; In order to increase translation performance and user satisfaction, it is beneficial to consider previous post-edits made by the user of the system.

Both approaches can be implemented within the phrase-based SMT framework by adding a second, dynamic phrase table. This architecture was first described in (Chen et al., 2007), who built a dynamic phrase table trained on translation hypotheses by external MT systems. The online learning system described in (Hardt and Elming, 2010) uses a similar architecture, with the difference that, rather than translations by external systems, previous translations, post-edited by the user, constitute the dynamic corpus.

The aim of this study is to: a) evaluate both approaches in an independent reimplementation and on a different corpus; b) implement and evaluate an alternative scoring procedure that promises further performance gains; c) show the feasibility of combining multi-engine MT and online learning in a single framework.

2 Related Work

System combination for Machine Translation is an active research field. The last two Workshops on Machine Translation (WMT) included a system combination task; an overview is given in (Callison-Burch et al., 2009; Callison-Burch et al., 2010).

The effectiveness of system combination strongly depends on the relative performance of the systems being combined. In the 2009 WMT, (Callison-Burch et al., 2009) conclude that "In general, system combinations performed as well as the best individual systems, but not statistically significantly better than them." A possible reason for this failure to improve on individual systems is given in the following year: "This year we excluded Google translations from the systems used in system combination. In last year's evaluation, the large margin between Google and many of the other systems meant that it was hard to improve on when combining systems. This year, the system combinations perform better than their component systems more often than last year." (Callison-Burch et al., 2010)

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 89–96
Leuven, Belgium, May 2011

We implemented a system combination architecture similar to that described in (Chen et al., 2007). While some approaches treat all systems as black boxes, needing only the 1-best output from each system and a language model, (e.g. (Barrault, 2010; Heafield and Lavie, 2010)), the combined system described by (Chen et al., 2007) is an extension of an existing SMT system. The combination is achieved by taking a vanilla SMT system and adding a second, dynamic phrase table to the existing primary one. (Chen et al., 2007) propose that the dynamic phrase table be trained online on the translation output of several rule-based translation systems. (Chen et al., 2009) expand on this concept by allowing for the inclusion of arbitrary translation systems. We think this distinction into a primary system and several secondary ones is attractive for our translation scenario, as will be explained in section 3.1.

(Hardt and Elming, 2010) propose a technically similar approach, albeit for a different purpose. Their idea is to keep post-edited translations in a dynamic corpus. This corpus grows with every sentence that is translated, and is periodically used to re-train a dynamic phrase table.

In a simulation of the approach, using reference translations instead of actual post-edited translations, they show that this dynamic phrase table helps to improve translation performance significantly in some domains. They argue for the existence of a *file-context effect*, that is, that "translation data from within a file has a striking effect on translation quality" (Hardt and Elming, 2010).

Since dynamic phrase tables are typically small, word alignment has been recognized as a major challenge in all related publications. A successfully tested solution is to train GIZA++ (Och and Ney, 2003) on the primary corpus first, then using the obtained models to align the dynamic corpus (Chen et al., 2009; Hardt and Elming, 2010). (Hardt and Elming, 2010) then apply heuristic post-processing to improve these approximate alignments.

Many approaches of combining and weighting phrase tables have been proposed. (Hardt and Elming, 2010) add the dynamic phrase table as an alternative decoding path to their Moses system, copying the parameter weights from the primary phrase table. (Chen et al., 2007) concatenate the phrase tables and augment them by adding new features as system markers. (Chen et al., 2009) propose a combination that avoids duplicate phrase pairs, giving priority to the primary phrase table.

In contrast to word alignment and phrase table combination, little attention has been paid to the issue of obtaining translation probabilities for the dynamic phrase tables. (Hardt and Elming, 2010) and (Chen et al., 2009) report that they use standard Moses procedures, i.e. Maximum Likelihood Estimation (MLE) for phrase translation probabilities, and lexical weights that are based on the word translation probabilities estimated by MLE.[1] Since MLE is unreliable for low sample sizes, we expect the performance of systems that include a dynamic phrase table to seriously degrade as the dynamic phrase table becomes smaller. We propose to mitigate the problem by grounding the MLE-based scoring of the dynamic phrase table in the frequency counts of both the dynamic and the primary phrase table. Since our implementation can be used for both system combination and online learning, we will test the effect of scoring on both approaches.

3 System description

3.1 Data and Tools

We conduct our experiments on the parallel part of the Text+Berg corpus, a collection of Alpine texts (Volk et al., 2010). The collection so far consists of the yearbooks of the Swiss Alpine Club from 1864 to 1995. Since 1957, the yearbook has been published in two parallel editions, German and French. Table 1 shows the amount of training data. Note that we use a relatively small amount of training data, but that training is in-domain with respect to the test set. As a consequence, the main weakness of the baseline system is data sparseness. In the 1000-sentence test set, 19% of the types (5% of the tokens) are out-of-vocabulary words, i.e. words that are not in the translation model. This can be mostly attributed to the morphological complexity of German, which is the source language in our experiments. Incorporating rule-based MT systems, which are able to decompose German compounds and analyse inflected forms, into the translation process promises to mitigate this problem.

Our motivation is to use external systems to fill lexical gaps in the baseline SMT system, which is trained on a relatively small amount of in-domain training data and outperforms other systems not adapted to the domain (see section 4.1). Ideally,

[1] See (Koehn et al., 2003) for the formulae.

Corpus	segments	words DE	words FR
Training	151 000	2 840 000	3 200 000
Tuning	1135	23 100	25 800
Test	991	19 200	21 600
LM	490 000	-	9 510 000

Table 1: Data used for training, tuning and testing, and for training the language model.

translations by the external systems should only be used for source words or phrases which are not well-evidenced in the primary system. For this, the glass-box approach of taking an existing SMT system and extending it with a dynamic phrase table seems better suited than a black box combination of systems, in which we cannot know how well-evidenced different translation options are.

As external SMT systems for the multi-engine translation approach, we use the rule-based Personal Translator 14[2], and Google Translate[3]. While Google Translate is a statistical system, it promises to be more robust to data sparseness than our in-domain system because the Google system is trained on significantly more training data.[4]

We build the SMT systems using Moses (Koehn et al., 2007), SRILM (Stolcke, 2002), and MGIZA++ (Gao and Vogel, 2008). In terms of configuration, we adhere to the instructions for building a baseline system by the 2010 ACL Workshop on SMT.[5] Additionally, we prune the primary phrase table using the approach by (Johnson et al., 2007).

3.2 Alignment

We compute the word alignment of the primary corpus using the default configuration in Moses, but saving all models to disk. We then force an alignment of all dynamic corpora on the basis of these models with MGIZA++. Since we do not focus on word alignment in this paper, we only compute the alignments once for each dynamic corpus, re-using the alignment when we build phrase tables from parts of the corpus. This allows us to rule out alignment differences as the reason for variation in performance.

[2] http://www.linguatec.net/products/tr/pt

[3] http://translate.google.com

[4] Even though we do not know the actual amount of training data used for Google Translate DE-FR, this is a safe assumption (see table 1).

[5] http://www.statmt.org/wmt10/baseline.html

3.3 Scoring

For each of the experiments with dynamic phrase tables, the baseline scoring sytem is vanilla Moses, i.e. a scoring of translation probabilities based on the dynamic corpus only. This is the implementation described in (Chen et al., 2007) and (Hardt and Elming, 2010). We propose to score the dynamic phrase table by also taking the primary corpus into account, since MLE is unreliable for small sample sizes.

Our modified approach to scoring is implemented as follows. The Moses training scripts are modified to not only return phrase translation probabilities and lexical weights, but also the *sufficient statistics*, i.e. word and phrase (pair) frequencies, required to recompute all parameters. Each time the dynamic corpus is updated, we train the dynamic phrase table using this modified script. Then, we rescore the translation probabilities and lexical weights in the dynamic phrase table with a client-server architecture.

The server stores all relevant frequencies of the primary corpus in memory, and upon receiving the command to rescore the dynamic phrase table, extracts the frequencies of the dynamic corpus, then computes updated translation probabilities based on the sum of the frequencies in both corpora.

We illustrate the motivation behind this modification with two examples, shown in table 2. The two sentences exemplify two different situations. In the first, the German compound *Konditionswunder* (roughly: *one who is in miraculous shape*) is unknown by the primary system. Here, the multi-engine approach is shown to work, since this lexical gap is filled with an adequate translation by one of the external systems.

The second sentence is translated well by the domain-specific system, but improperly by the external systems. Most striking is the German word *Pässe* (English: *mountain passes*), correctly translated as *cols* by the primary system, but as either *passeports* (English: *passports*) or *la passe* (English: *pass [of the ball]*) by the external ones, both possible translations of *Pass*, but unlikely ones in the mountaineering domain. *Pässe* is well-evidenced in the primary corpus (136 observations), with $p(cols|Pässe)$ estimated at $64/136$ (0.47). We find that, 2 being the frequency of *Pässe* in the dynamic corpus, estimating $p(passeports|Pässe)$ and $p(la\ passe|Pässe)$ at $1/(136+2)$ (0.007), rather than $1/2$ (0.5), better

Source	Er ist ein Konditionswunder.
	He is in miraculous shape.
Reference	C'est un miracle de condition physique.
System 1 (Moses)	C'est un Konditionswunder.
System 2 (PT 14)	C'est un miracle de condition.
System 3 (Google Translate)	Il est un miracle de remise en forme.
Multi-Engine (vanilla)	C'est un miracle de condition.
Multi-Engine (modified)	C'est un miracle de condition.
Source	Wir konnten das Aussehen **der Pässe** nur ahnen.
	We could only guess at the look **of the mountain passes**.
Reference	Nous ne pouvions que deviner l'aspect **des cols**.
System 1 (Moses)	nous ne pouvions seulement deviner l'aspect **des cols**.
System 2 (PT 14)	Nous ne pouvions que nous douter de l'air **des passeports**.
System 3 (Google Translate)	Nous ne pouvions imaginer l'aspect **de la passe**.
Multi-Engine (vanilla)	nous ne pouvions de l'air **des cols de la passe**.
Multi-Engine (modified)	nous ne pouvions l'aspect **des cols** que deviner.

Table 2: German–French translation examples.

models our expectation. The numbers are simplified, discussing only one of four scores computed for the phrase table. Also, possible errors during word alignment and/or phrase extraction are not considered. If we look at the output of the vanilla multi-engine system, we see that such a misalignment has indeed occurred, with German *nur ahnen* (English: *only guess*) being mistranslated as *de la passe*. With modified scoring, this phrase pair is given low scores[6], preventing it from being selected during decoding.

Summing the frequencies of different corpora is not a new idea. If we simply concatenated the corpora before scoring, we would achieve the same effect. However, working with two phrase tables, one static and one dynamic, has distinct advantages: since we only rescore the dynamic phrase table, training is much faster than if we had to rescore the primary model regularly. It also allows us to give different weights to the two phrase tables. We show in the evaluation that this leads to a better performance.

3.4 Combination of Phrase Tables

(Chen et al., 2009) decided against using the primary and dynamic phrase table as alternative decodings paths in Moses, since this increases the search space for MERT, especially since they extend the system with additional features, for instance to mark the origin of translation hypotheses. (Hardt and Elming, 2010), on the other hand, use alternative decoding paths and avoid the problem of tuning by using the same set of weights for both phrase tables.

For our experiments, we will use alternative decoding paths, keeping the search space under control by not adding any features, and never using more than one dynamic phrase table. In the online learning experiments, we will follow (Hardt and Elming, 2010) in using the same set of weights for both phrase tables.

4 Evaluation

Data and tools used for our experiments are described in section 3.1. For the evaluation, we use BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), applying bootstrap resampling to test for statistical significance (Riezler and Maxwell, 2005). After establishing the baseline performance of our in-domain SMT system and the two external systems (i.e. Personal Translator 14 and Google Translate), we describe three experiments.

The first is a re-implementation of the multi-engine approach described in (Chen et al., 2009), the second one of online learning by (Hardt and Elming, 2010), and the third a combination of the two. In the first two experiments, we want to address the following research questions:

- Can we reproduce the positive effect of both approaches in our evaluation scenario?

[6]This especially applies to the lexical weights; since both the source and the target phrase are rare in the primary corpus, the phrase translation probabilities are only slightly penalized.

System	BLEU	METEOR
own baseline	**17.18**	**38.28**
Personal Translator 14	13.29	35.68
Google Translate	12.94	34.36

Table 3: SMT performance DE–FR.

- How does the multi-engine approach described here compare to system combination algorithms that only use the translation hypotheses and a language model, but not a parallel corpus?
- What is the effect of dynamic phrase table size on translation performance, excluding word alignment as a factor?
- Is our proposed modification to scoring effective at improving system performance?

In the third experiment, our aim is to demonstrate that both multi-engine MT and online learning can be combined within a single framework.

4.1 Baseline Experiments

In terms of baseline performance (table 3), we find that our in-domain system obtains markedly better scores than both Personal Translator 14 and Google Translate when evaluating it on our Alpine test set.[7] However, the performance is relatively low for the language pair DE–FR – when we trained an SMT system on Europarl, we achieved 28.24 BLEU points on a Europarl test set for this language pair. This indicates that the mountaineering narratives which constitute our test set are relatively hard to translate.

4.2 Multi-Engine Translation

For the multi-engine translation experiments, we first built a dynamic phrase table encompassing both the tuning and the test set (or approximately 4000 sentence pairs).[8] We then conduct MERT with this dynamic phrase table added as an alternative decoding path to Moses.

Three alternative system combination methods are evaluated against the approach described here (called **Dynamic**): **Concat**, an SMT system trained on the concatenation of the parallel training corpus and the translation hypotheses by Google Translate and Personal Translator 14. **MANY** (Barrault, 2010) and **MEMT** (Heafield and Lavie, 2010), both open source system combination software with confusion network decoding.

For the experiments with a dynamic phrase table, we test the effect of dynamic corpus size on SMT performance. We do this by varying the number of sentences that are translated at once, each time building a dynamic phrase table that only includes the translations of the sentences needed at the time. In the extreme case, each sentence is translated independently, with a dynamic phrase table built from two sentence pairs.

4.2.1 Results

All combined systems shown in table 4 significantly outperform the baseline of 17.18 BLEU points. The score difference between MANY and MEMT is not statistically significant, but both are significantly better than the baseline and significantly worse than the approaches that make use of the in-domain parallel text. This validates our attempts to exploit the in-domain parallel corpus for system combination. We have not analyzed at which stage MANY and MEMT fail to exploit the full potential of the translation hypotheses; both alignment errors and decoding errors are conceivable.

A concatenation of the primary corpus and the translation hypotheses, with the same training procedure as in the baseline system, works surprisingly well, yielding a BLEU score of 19.11. However, this approach has little practical use, since retraining the entire SMT system is prohibitively slow. The experiments with a dynamic phrase table yield the best results. The modified scoring as proposed in section 3.3 achieves 20.06 BLEU points, as opposed to 19.33 BLEU points achieved with vanilla scoring.

One weakness of the vanilla scoring algorithm, i.e. the unreliability of MLE, is bound to become more severe when we translate the test set in smaller steps and build smaller dynamic phrase tables. The results in table 5 confirm that this holds true, although the effect is smaller than we expected. Only with a dynamic corpus size of 2, i.e. when builiding a separate dynamic phrase table for each sentence that is translated, did we observe a statistically significant drop in performance. With

[7] We are aware that BLEU scores might not give a fair assessment of rule-based MT as compared to SMT (see (Callison-Burch et al., 2006)). If the rule-based system indeed performs better than the BLEU scores suggest, this is all the more reason for tapping its knowledge in a multi-engine approach.

[8] This is twice the size of the tuning and test set: every source sentence is once paired with its translation by Personal Translator 14, once with the one by Google Translate.

Combination System	BLEU	METEOR
MANY	18.23	39.68
MEMT	18.39	39.01
Concat	19.11	39.45
Dynamic (vanilla)	19.33	40.00
Dynamic (modified)	**20.06**	**40.59**

Table 4: SMT performance DE–FR for multiple system combination approaches.

size of dynamic corpus (sentence pairs)	vanilla	modified
4000	19.33	20.06
200	19.26	19.95
20	19.17	19.96
2	18.80	19.93

Table 5: SMT performance DE–FR as a function of dynamic corpus size. BLEU scores.

our modified scoring algorithm, we successfully eliminated this dependence of SMT performance on the size of the dynamic corpus.

4.3 Online Learning

Our test set consists of 7 held-out articles of the Text+Berg corpus, spanning 991 sentences. The fact that the test sentences are not selected randomly allows us to investigate file-context effects as observed in (Hardt and Elming, 2010).

We use the same translation process as in the last experiments, with two differences: Firstly, we do not perform MERT and use the weights of the primary phrase table for the dynamic one. Secondly, the dynamic corpus is different. Instead of using external translation hypotheses for the sentences that are currently translated, we use the reference translation for all previously translated sentences of the test set. This simulates a post-editing approach where the corrected translations are re-used to improve later translations. The dynamic corpus is thus retrained after every sentence, but its size increases over time.

4.3.1 Results

The results are shown in table 6. Our vanilla reimplementation of the online learning approach is worse than the baseline. (Hardt and Elming, 2010) attributed the lack of improvement in one experiment to a weak file-context effect in one of the test sets. While our test set, which consists of mountaineering narratives, is also less repetitive than the technical domains in which (Hardt and Elming, 2010) found strong file-context effects, this is not enough to explain why the scores go down.

System	BLEU	METEOR
baseline	17.18	38.28
vanilla scoring	16.81	37.61
modified scoring	**17.57**	**38.60**

Table 6: SMT performance DE–FR with online learning system.

The reason is that translation probabilities are not estimated well, as discussed in section 3.3. To give another amusing example of the consequences, the German word *farbige* (English: *colourful*) is translated as *très colorés* by our baseline system, but as *nous déployons* (English: *we deploy*) by the experimental one. The translation, learned from a sentence about deploying parachutes, makes little sense in the context of *colourful birds*. With modified scoring, the mistranslation no longer occurs.

By using the modified scoring procedure, we significantly outperform the baseline. Considering the small amount of additional training material, and the elimination of other possible confounding factors (all systems use the same weights), we conclude that file-context effects exist in our test set. Also, the experiment validates our modifications to scoring of the dynamic phrase table, turning a loss of 0.4 BLEU points with the experimental approach into a gain of the same size.

4.4 Combining Multi-Engine MT and Online Learning

It is attractive to combine the two prior experiments, since both are based on the same architecture, with the only difference being the corpus for training the dynamic phrase table, and the parameters for the models. Sadly, we observed comparatively little gains with online learning in our test set, which limits the potential of the combined approach.

We decided against increasing the number of dynamic phrase tables, which would complicate the scoring, without offering additional benefits: the main advantage of having multiple phrase tables would be the possibility of using separate parameters for each table, but the explosion in the size of the search space makes it unlikely for MERT to

System	BLEU	METEOR
baseline	17.18	38.28
online learning	17.57	38.60
multi-engine MT	19.93	40.52
combined	20.05	40.61

Table 7: SMT performance DE–FR with system combining multi-engine MT and online learning.

find good weights.

We chose our best-performing experiment so far, the multi-engine system with modified scoring, as our new baseline. We performed retraining of the dynamic phrase table for every sentence, including the translation hypotheses by both external MT systems, and all previous translation pairs from the test set. Preliminary experiments have shown that the multi-engine system works significantly worse than our best experiment when we simply copy the MERT parameters of the primary phrase table (BLEU score: 18.04). Thus, we chose to adopt the parameters of the multi-engine experiment for this one.

4.4.1 Results

The combined system does not significantly outperform the multi-engine MT system, as table 7 shows. One problem is that the effect of online learning is already small in our test set; the other, that we cannot expect the improvements of the two approaches to be additive, since both mitigate the same weakness of the primary system.

5 Conclusion

In this paper, we reimplemented two differently motivated but technically similar approaches that use a dynamic phrase table, along with a static primary one, to provide SMT systems with information relevant to the translation task at hand. (Chen et al., 2009) built a dynamic phrase table from translation hypotheses by external MT systems, while (Hardt and Elming, 2010) used previous translation pairs from the same file to contribute to the translation of future ones. We successfully reimplemented both approaches and showed them to work in our translation setting that consists of a German–French SMT system trained on a small domain-specific translation corpus. We show that this approach can outperform system combination algorithms that only use information on the target side, i.e. the translation hypotheses and a language model. Additionally, we propose modification to the scoring procedure for dynamic phrase tables. Rather than basing MLE of translation probabilities only on the dynamically created corpus, we show that combining the frequencies of the primary and the dynamic corpus for the purpose of scoring leads to significant gains in SMT performance. We have observed an increase by 0.73 and 0.76 BLEU points over the vanilla scoring algorithm, and 2.88 BLEU points over the best individual system. We also identified the vanilla scoring procedure as the reason for a decrease in performance in one experiment, namely the reimplementation of incremental retraining by (Hardt and Elming, 2010). We conclude that it is advisable to score dynamic phrase tables with recourse to the frequencies in larger corpora. With a client-server architecture, where the frequencies are held in memory, there is little impact on translation time, albeit at the cost of memory space.

The potential performance gain of both multi-engine MT and online learning approaches varies between translation scenarios, depending on the availability and quality of training data, external MT systems, and the file-context effect. We demonstrated the feasibility of combining the two approaches, but found no statistically significant additional score increase over the multi-engine approach. We suspect that bigger gains can be achieved when translating texts of a more repetitive nature, for which (Hardt and Elming, 2010) demonstrated the beneficial effect of online learning.

So far, the only component of our experimental system that incorporates dynamic knowledge is the phrase table. For future research, we want to investigate whether dynamically retraining other components such as the language model or reordering model may lead to additional performance gains.

Acknowledgments

I would like to thank Loïc Barrault and Kenneth Heafield for providing me with the newest source code and technical support for their software. This research was funded by the Swiss National Science Foundation under grant 105215_126999.

References

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Ex-*

trinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Barrault, Loïc. 2010. MANY: Open source MT system combination at WMT'10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 277–281, Uppsala, Sweden, July. Association for Computational Linguistics.

Callison-Burch, C., M. Osborne, and P. Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *Proceedings the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy.

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.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics. Revised August 2010.

Chen, Yu, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-engine machine translation with an open-source decoder for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 193–196, Morristown, NJ, USA. Association for Computational Linguistics.

Chen, Yu, Michael Jellinghaus, Andreas Eisele, Yi Zhang, Sabine Hunsicker, Silke Theison, Christian Federmann, and Hans Uszkoreit. 2009. Combining multi-engine translations with Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 42–46, Morristown, NJ, USA. Association for Computational Linguistics.

Gao, Qin and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Hardt, Daniel and Jakob Elming. 2010. Incremental retraining for post-editing SMT. In *Conference of the Association for Machine Translation in the Americas 2010 (AMTA 2010)*, Denver, CO, USA.

Heafield, Kenneth and Alon Lavie. 2010. CMU multi-engine machine translation for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 301–306, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johnson, Howard, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic, June. Association for Computational Linguistics.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180, Prague, Czech Republic.

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

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Riezler, Stefan and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Stolcke, A. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, CO, USA.

Volk, Martin, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef. 2010. Challenges in building a multilingual alpine heritage corpus. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).