Parametric Weighting of Parallel Data for Statistical Machine Translation

Kashif Shah, Loïc Barrault, Holger Schwenk
LIUM, University of Le Mans
Le Mans, France.
FirstName.LastName@lium.univ-lemans.fr

Abstract

During the last years there is increasing interest in methods that perform some kind of weighting of heterogeneous parallel training data when building a statistical machine translation system. It was for instance observed that training data that is close to the period of the test data is more valuable than older data (Hardt and Elming, 2010; Levenberg et al., 2010). In this paper we obtain such a weighting by resampling alignments using weights that decrease with the temporal distance of bitexts to the test set. By these means, we can use all the available bitexts and still put an emphasis on the most recent one. The main idea of our approach is to use a parametric form or meta-weights for the weighting of the different parts of the bitexts. This ensures that our approach has only few parameters to optimize. We report experimental results on the Europarl corpus, translating from French to English and further verified it on the official WMT'11 task, translating from English to French. Our method achieves improvements of about 0.6 points BLEU on the test set with respect to a system trained on data without any weighting.

1 Introduction

Statistical machine translation (SMT) systems are based on two types of resources: monolingual data to build a language model (LM) and bilingual data – also called bitexts – to train the translation model (TM). The parallel data often comes from different sources, *e.g.* Europarl, UN, in-domain data in limited amounts, data crawled from the Internet or even bitexts automatically extracted from comparable corpora. It seems obvious that the appropriateness and the usefulness of this parallel data for a particular translation task may vary quite a lot. Nevertheless, the standard procedure is to concatenate all available parallel data, to perform word alignment using GIZA++ (Och and Ney, 2000) and to extract and score the phrase pairs by simple relative frequency. Doing this, the parallel data is (wrongly) considered as one homogeneous pool of knowledge. We argue that the parallel data is quite inhomogeneous in many practical applications with respect to several factors:

- the data may come from different sources that are more or less relevant to the translation task (in-domain versus out-of-domain data).
- more generally, the topic or genre of the data may be more or less relevant.
- the data may be of different quality (carefully performed human translations versus automatically crawled and aligned data).
- the recency of the data with respect to the task may have an influence. This is of interest in the news domain where named entities, etc change over time.

There have been several attempts in the literature to address some of these problems. Matsoukas et al. (2009) proposed to weight each sentence in the training bitexts by optimizing a discriminative function on a tuning set. Sentence-level features are extracted to estimate the weights that are relevant to the given task. Foster et al. (2010) proposed an extended approach by an instant weighting scheme which learns weights on individual phrase pairs instead of sentences and incorporated the instance-weighting model into a linear combination of feature functions.

The technique presented in this paper is related to these previous works as it concerns the weighting of corpora or sentences. However, it does not

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1323–1331,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

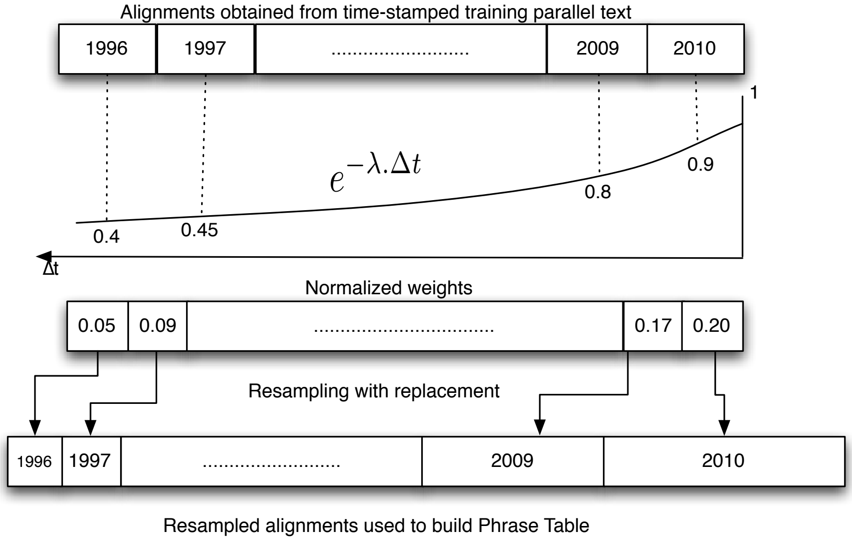

Figure 1: Overview of the weighting scheme. The alignments are weighted by an exponential decay function, parameterized by λ. Resampling with replacement is used to create a new corpus (parts with higher weight will appear more often). The phrase table is built from this corpus using the standard procedure.

require the calculation of additional sentence-level features.

In our previous work Shah et al. (2010) we proposed a technique to weight heterogeneous data by weighted resampling of the alignments. The weights were numerically optimized on development data.

Hardt and Elming (2010) has shown recency effect in terms of file-context and concluded that the data within the same file is of greater importance than the rest. Levenberg et al. (2010) proposed an incremental training procedure to deal with a continuous stream of parallel text. Word alignment was performed by the stepwise online EM algorithm and the phrase table was represented with suffix arrays. The authors showed that it is better to use parallel data close to the test data than all the available data.

The research presented in this paper is the extension of our previous work Shah et al. (2010) to weight corpora by resampling and is inspired by the work of Levenberg et al. (2010) to consider the recency of the training data. In fact, we could split the training data into several parts over time scale and use our previous resampling approach Shah et al. (2010) to automatically optimize the weights of each time period. However, this approach does not seem to scale very well when the number of individual corpora increases. Numerical optimization of more than ten corpus weights would probably need a large number of iterations, each one consisting in the creation of a complete phrase table and its evaluation on the development data.

The main idea of our work is to consider some kind of meta-weights for each part of the training data. Instead of numerically optimizing all the weights, these meta-weights only depend on few parameters that need to be optimized. Concretely, in this work we study the exponential decrease of the importance of parallel data in function of its temporal distance to the development and test data. The weighting of the parts is still done by resampling the alignments. However, our general approach is not limited to weighting the training data with respect to recency to the development and test data. Any other criterion could be used as long as it can be calculated by a parametric function, *i.e.* to measure the topic appropriateness.

2 Weighting Scheme

The main idea of our work is summarized in Figure 1. We consider that time information is available for the bitexts. If this is not the case, one can consider that the time advances sequentially with the lines in the file. First, the data is considered in parts according to the time information. In Figure 1, we group together all data within the same year, but any other granularity is possible (months, weeks, days, etc). Given the observation that more recent training data seems to be more important than older one, we apply an exponential decay function:

$$e^{-\lambda \cdot \Delta t} \qquad (1)$$

where λ is the decay factor and Δt is the discretized time distance (0 for most recent part, 1 for the next one, etc.). Therefore, our weighting scheme has only one parameter to be optimized.

Following our previous work Shah et al. (2010), we resample the alignments in order to obtain a weighting of the bitexts according to their recency. The weight of each part of the bitexts is normalized (sum to one). The normalized weights represent the percentage of final aligned corpus that is originated from each part of the source corpus: word alignments corresponding to bitexts that are close to the test period will appear more often than the older ones in the final corpus.

In addition, we considered the quality of the alignments during resampling, as described in our previous work (Shah et al., 2010). Alignments produced by GIZA++ have alignment scores associated with each sentence pair in both direction, *i.e.* source to target and target to source. Alignment scores have a very large dynamic range and are concentrated around very low values, consequently the following logarithmic mapping is applied in order to flatten the distribution:

$$\log(\alpha \cdot \frac{(\sqrt[n_{trg}]{a_{src\_trg}} + \sqrt[n_{src}]{a_{trg\_src}})}{2}) \qquad (2)$$

where a is the alignment score, n the size of a sentence and α a smoothing coefficient to optimize. We used these normalized alignment scores as confidence measurement for each sentence pair.

3 Description of the algorithm

The architecture is presented in Figure 2. The starting point is a parallel corpus. We performed word alignment in both directions using GIZA++. The corpus is then separated into several parts on the basis of a given time span. We performed experiments with different span sizes, namely year, month, week and day. The decaying function is scaled so that the range does not change when using different span sizes. A weighting coefficient obtained with the exponential decay function is then associated to each part.

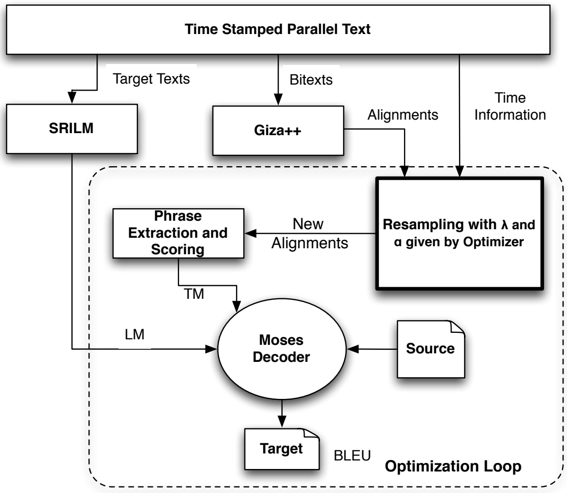

Figure 2: Architecture of SMT Weighting System.

Algorithm 1 Weighting with Exponential Decay function using resampling

```
Determine word to word alignment with GIZA++ on concatenated bitexts.
Initialize $\lambda$ and $\alpha$ with equal weights.
while not Optimized do
  Compute time-spans weights by eq. 1
  Normalize weights
  for $i = 0$ to #time-span do
    $proportion \leftarrow$ required_size[1] $*$ weights$[i]$
    j = 0
    while $j < proportion$ do
      $Al \leftarrow$ Random alignment
      $Al_{score} \leftarrow$ normalized score of $Al$
      Flatten $Al_{score}$ with $\alpha$
      $Threshold \leftarrow$ rand[0, 1]
      if $Al_{score} > Threshold$ then
        keep it
        $j = j + 1$
      end if
    end while
  end for
  Create new resampled alignment file.
  Extract phrases and build the phrase table.
  Decode
  Calculate the BLEU score on Dev
  Update $\lambda$ and $\alpha$
end while
```

Then, for each part, resampling with replacement is performed in order to select the required number of alignments and form the final corpus. The resampling is done as follows: for each alignment considered, a new random threshold is gen-

[1] required_size depends upon the number of times we resample - see section 5.

erated and compared to the alignment score. The alignment is kept only if its score is above the threshold. This ensures that all alignments have a chance to be selected, but this chance is proportional to its alignment score.

Note that some alignments may appear several times, but this is exactly what is expected as it will increase the probability of certain phrase pairs which are supposed to be more related to the test data (in terms of recency) and of better quality. The smoothing and decay factors, α and λ respectively, are optimized with a numerical optimizer called CONDOR (Berghen and Bersini, 2005). The procedure and steps involved in our weighting scheme are shown in algorithm 1.

4 Experimental evaluation

Our first experiments are based on the French-English portion of the freely available time-stamped Europarl data (Koehn, 2005) from April 1996 to December 2010. We have built several phrase-based systems using the Moses toolkit (Koehn et al., 2007), though our approach is equally applicable to any other approach based on alignments and could be used for any language pairs. In our system, fourteen feature functions are used. These feature functions include phrase and lexical translation probabilities in both directions, seven features for the lexicalized distortion model, a word and phrase penalty, and the target language model. The coefficients of these feature functions are optimized by minimum error training.

In the first experiments, the whole Europarl corpus was split into train, development and test as shown in Figure 3. The most recent 5K sentences are split into two sets of equal size, one for development and the other for testing. The remaining data was used as training bitexts to build the different systems.

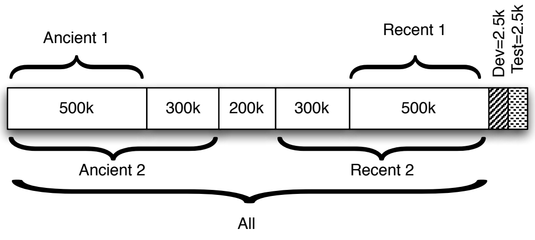

Figure 3: Data used to build the different systems (# sentences)

Since we want to focus on the impact of the weighting scheme of the bitexts, we used the same language model for all systems. It has been trained with the SRILM toolkit (Stolcke, 2002) on the target side of all the training data. In addition, the weights of the feature functions were tuned once for the system that uses all the training data and then kept constant for all the subsequent experiments, *i.e.* no tuning of the feature functions weights is done during the optimization of the weighting coefficients λ and α.

Table 1 presents the results of the systems trained on various parts of the available bitexts without using the proposed weighting scheme. The best performance is obtained when using all the data (55M words, BLEU=30.48), but almost the same BLEU score is obtained by using only the most recent part of the data (24M words, part *Recent 2*). However, if we use the same amount of data that is further away from the time period of the test data (25M words, part *Ancient 2*), we observe a significant loss in performance. These results are in agreement with the observations already described in (Levenberg et al., 2010). Using less data, but still close to the evaluation period (15M words, part *Recent 1*) results in a small loss in the BLEU score. The goal of the proposed weighting scheme is to be able to take advantage of all the data while giving more weight to recent data than to older one. By these means we are not obliged to disregard older parts of the data that may contain additional useful translations. If the weighting scheme does work correctly, we cannot perform worse than using all the data. Of course, we expect to achieve better results by finding the optimal weighting between recent and ancient data.

The amount of data per year in the Europarl data can vary substantially in function of time period since it depends on the frequency and length of the sessions of the European Parliament. As an example Figure 4 shows the histogram of the data per year.

One can ask which time granularity should be used to achieve best weights. Only one weight is given to each time span, consequently the span size will have an impact on the alignment selection process. Using smaller spans results in a more fine grained weighting scheme. We have tested different settings with different time spans to see whether the impact of weighting changes with the

Europarl	Ancient data		Recent data		
	Ancient 1	Ancient 2	Recent 1	Recent 2	All
# of sentences/words	500K/15M	800K/25M	500K/15M	800K/24M	1800K/55M
BLEU (on dev)	29.84	30.08	30.80	31.09	**31.34**
BLEU (on test)	29.30	29.43	30.32	30.44	**30.48**

Table 1: BLEU scores obtained with systems trained on data coming from different time spans.

Europarl	Weighting + alignment selection				Best+retune
Time span	Days	Weeks	Months	Years	Years
Optimized λ	0.0099	0.0109	0.0110	0.0130	0.0130
BLEU (on dev)	31.73	31.82	31.75	31.80	**31.92**
BLEU (on test)	30.94	30.97	30.92	**30.98**	**31.09**

Table 2: Results in BLEU score after weighting.

size of each span. The results are shown in Table 2.

It is observed that all four systems obtained very similar results, which indicates that the size of the spans is not very important. One surprising observation is that the optimized decay factor for all time span sizes are really close to each other. The reason to this could be the scaling of the exponential decaying function based on the time span size. In fact scaled values ensure that the oldest data point get roughly the same value independent of using years, months or days as time span. Looking at the optimized values of λ in Table 2, we can observe that the relative difference between recent and ancient data is rather small, *i.e.* the ancient data is still somehow important and cannot be neglected.

By using years as time span, we obtain an improvement of +0.50 BLEU score on the test set compared to using all data without weighting (30.48 → 30.98). It is clear that recency has a positive impact on system performance, however, weighting properly the different parts gives better performance than using the most recent or all available data.

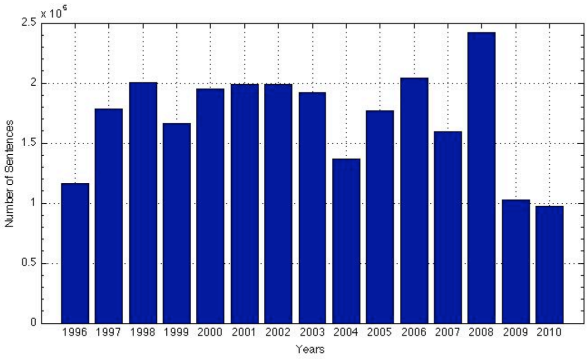

Figure 4: Amount of data available in the Europarl corpus for each year

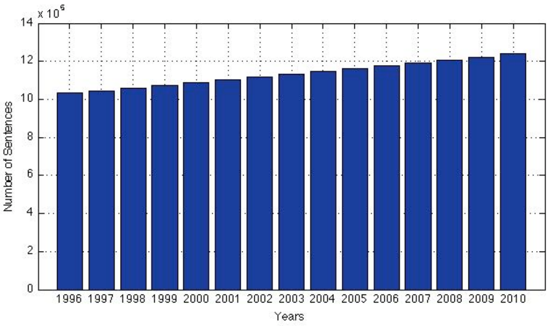

Figure 5: Distribution of data after weighting

Finally, the best system is retuned (feature functions weights) and an overall improvement of +0.61 in the BLEU score is observed on test set.

5 Discussion

The optimal decay factor of approximately 0.01 actually leads to an almost linear decrease over time. The difference in the quantity of data taken from most recent and least recent data is only 1.4% (which still represent 200k sentences). Therefore, one could think that the weighting does not favor recent data that much. This is not the case as we can see in Figure 5 where the distribution of data used to build the adapted model is presented.

Europarl	Resampling only	Weighting only	alignment selection only
BLEU (on dev)	31.36	31.69	31.45
BLEU (on test)	30.51	30.84	30.64

Table 3: Results in BLEU score with different settings.

Example 1	
A:	Mr Ribeiro e Castro, we **shall** see all this in the Conference of Presidents.
B:	Mr Ribeiro e Castro, we **will** see all this at the Conference of Presidents.
R:	Mr Ribeiro e Castro, we **will** look at all these questions in the Conference of Presidents' meeting.
Example 2	
A:	We **shall** most probably consider again lodge a complaint with the **Court of Justice of the European Communities**.
B:	We **will** most probably consider again to lodge a complaint to the **European Court of Justice**.
R:	Most probably we **will** again discuss renewed recourse to the **European Court of Justice**.
Example 3	
A:	no Member State has *not* led to **field trials** as regards the BST .
B:	no Member State has led to **tests on the ground** as regards BST .
R:	No Member State has yet carried out **field tests** with BST .

Table 4: Example translations produced by systems *All* (A) and *Best+retune* (B) versus reference (R)

When comparing to Figure 4, the overall proportion of data coming from recent years is clearly bigger when using our resampling approach. This leads to different word choices while decoding.

Note that resampling is performed several times to estimate and select the samples which better represent the target data set. The more often we resample, the closer we get to the true probability distribution. The *required-size* in algorithm 1 depends upon the number of times we resample. We resampled ten times in our experiments. It is also worth to note that, we keep the original training data along with resampled one. It ensures that no information is lost and the set of extracted phrase pairs remain the same - only the corresponding probability distributions in the phrase table are changed.

In order to get more insight in our method, we separately performed the different techniques:

- resampling the training data without weighting;
- resampling the training data using weighting only (with respect to recency);
- resampling the training data using alignment selection.

These results are summarized in Table 3.

Note that the first case does not correspond to duplicating the training data a certain amount of time (which would of course produce exactly the same phrase-table). Since we perform resampling with replacement, this procedure introduces some randomness which could be beneficial. According to our results, this is not the case: we obtained exactly the same BLEU scores on the dev and test data than with the standard training procedure. Weighting with respect to recency or alignment quality both slightly improve the BLEU, but not as much as both techniques together. The performance increase seems actually to be complementary.

Some comparative examples between the translations produced by systems *All* and *Best+retune* versus the reference translations are given in Table 4. It was noticed that a lot of occurrences of *"will"* in the reference are actually translated into *"shall"* with system *All* whereas the correct word choice is made by the system *Best+retune* as shown in Example 1. This could be explained by the fact that recently the word *"will"* is more frequently seen in the training corpus and adapting the model by weighting the most recent data pro-

WMT Task	Baseline	Receny weighting + alignment selection	Recency weighting + alignment selection + relative importance
BLEU (on dev)	26.08	26.51	**26.60**
BLEU (on test)	28.16	28.59	**28.69**

Table 5: Results in BLEU score after weighting on English to French WMT Task.

duced correct translation. Actually, it was found that the word *"will"* is 10% more frequent in recent data (*Recent 1*) than in ancient data (*Ancient 1*) while the word *"shall"* is 2% less frequent.

Another interesting example is Example 2, in which the correct name for the *European Court of Justice* is proposed by the adapted system unlike the system *All* which proposed *Court of Justice of the European Communities*. Actually, it appears that the *Court of Justice of the European Communities* is the former name of the *European Court of Justice* prior to December 2009. The use of recent data allows to correctly translate the named entities which can change over time. The correct translation proposed by System *Best+retune* could be observed in Example 3 because of alignment selection procedure.

In our experiments, we assume that the test data is in present time (the usual case in a news translation system), and consequently we decrease the weight of the bitexts towards the past. This principle could be of course adapted to other scenarios.

An alternative approach could be to directly use the time decay function as the count for each extracted phrase. However, resampling the alignments and changing the counts of extracted phrases is not exactly the same. Same phrase pairs could be extracted from different parallel sentences coming from different time spans. Furthermore, weighting the alignments with their scores has shown improvements in the BLEU score as presented in Table 3, but considering the alignment score at the phrase level is not straight forward.

6 Experiments on the WMT task

To further verify whether our results are robust beyond the narrow experimental conditions, we considered a task where the development and test data do not come from the same source than the bitexts. We took the official test sets of the 2011 WMT translation tasks as dev and test sets (Schwenk et al., 2011) *i.e* news-test09 and news-test10 respectively. We built English-French systems by using the Europarl and News-Commentary (NC) corpora, both contain news data over a long time period.

For this set-up, there are three coefficients to optimize: the decay factor for Europarl λ_1, the decay factor for the news-commentary texts λ_2 and a coefficient for the alignments α. The Europarl corpus was divided into time span according to years and *NC* corpus was assumed to be sorted over time since time-stamp information was not available for the *NC* corpus. Remaining settings are kept same as mentioned in previous experiments to build the system *Best+retune*. The results are shown in Table 5. Finally, we considered the relative importance of the Europarl and *NC* corpora. For this, a weight is attached to each corpus which represents the percentage of the final aligned corpus that comes from each source corpus. These weights are also optimized on the development data using the same technique as proposed in our previous work (Shah et al., 2010). Using all these methods, we have achieved an overall improvement of approximately +0.5 BLEU on the development and test data, as shown in Table 5.

7 Conclusion and future work

In this paper, a parametric weighting technique along with resampling is proposed to weight the training data of the translation model of an SMT system. By using a parametric weighting function we circumvented the difficult problem to numerically optimize a large number of parameters. Using this formalism, we were able to weight the parallel training data according to the recency with respect to the period of the test data. By these means, the system can still take advantage of all data, in contrast to methods which only use a part of the available bitexts. We evaluated our approach on the Europarl corpus, translating from French into English and further tested it on official English to French WMT Task. A reasonable improvement in

BLEU score on the test data was observed in comparison to using all the data or only the most recent one. We argue that weighting the training data with respect to its temporal closeness should be quite important for translating news material since word choice in this domain is rapidly changing.

An interesting continuation of this work is to consider other criteria for weighting the corpora than the temporal distance. It is clear that recency is a relevant information and this could be associated with other features, *e.g.* thematic or linguistic distance. Also, this work can be included into a stream-based framework where new data is incorporated in an existing system by exponential growth function and making use of online retraining procedure as discussed in (Levenberg et al., 2010).

Acknowledgments

This work has been partially funded by the European Commission under the project Euromatrix-Plus, the French government under the project Cosmat and by an Overseas scholarship of Higher Education Pakistan. We would like to thank the unknown reviewers for their valuable comments.

References

Frank Vanden Berghen and Hugues Bersini. 2005. CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, 181:157–175, September.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 451–459, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Hardt and Jakob Elming. 2010. Incremental re-training for post-editing smt. In *The Ninth Conference of the Association for Machine Translation in the Americas 2010.*

Philipp Koehn, 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 *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 394–402, Stroudsburg, PA, USA. Association for Computational Linguistics.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 440–447, Stroudsburg, PA, USA. Association for Computational Linguistics.

Holger Schwenk, Patrik Lambert, Loïc Barrault, Christophe Servan, Sadaf Abdul-Rauf, Haithem Afli, and Kashif Shah. 2011. Lium's smt machine translation systems for wmt 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 464–469, Edinburgh, Scotland. Association for Computational Linguistics.

Kashif Shah, Loïc Barrault, and Holger Schwenk. 2010. Translation model adaptation by resampling. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 392–399, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm—an extensible language modeling toolkit. *In Proceesings of the 7th International Conference on Spoken Language Processing (ICSLP 2002*, pages 901–904.