Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel and Alex Waibel

Interactive Systems Laboratory
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213, USA
silja@cs.cmu.edu, matteck@cs.cmu.edu, stephan.vogel@cs.cmu.edu, waibel@cs.cmu.edu

Abstract. In this paper we present experiments concerning translation model adaptation for statistical machine translation. We develop a method to adapt translation models using information retrieval. The approach selects sentences similar to the test set to form an adapted training corpus. The method allows a better use of additionally available out-of-domain training data or finds in-domain data in a mixed corpus. The adapted translation models significantly improve the translation performance compared to competitive baseline systems.

1. Introduction

The goal of this research is to improve the translation performance for a Statistical Machine Translation system. The basic approach is to adapt the translation models.

Statistical machine translation can be described in a more formal way as follows:

$$t^* = \arg\max_t P(t \mid s) = \arg\max_t P(s \mid t) \cdot P(t)$$

Here t is the target sentence, and s is the source sentence. $P(t)$ is the target language model and $P(s|t)$ is the translation model used in the decoder. Statistical machine translation searches for the best target sentence from the space defined by the target language model (LM) and the translation model (TM).

Statistical translation models are usually either phrase- or word-based and include most notably IBM1 to IBM4 and HMM (Brown et al., 1993; Vogel et al., 1996). All models use available bilingual training data in the source and target language to estimate their parameters and approximate the translation probabilities.

Typically, the more data is used to estimate the parameters of the translation model, the better it can approximate the "true" translation probabilities. This will obviously lead to a higher translation performance. However if a significant amount of out-of-domain data is added to the training data, translation quality can drop. One reason for this is that a general translation model $P(s|t)$, that was trained on in-domain and out-of-domain data, does not fit the topic or style of individual texts. Unfortunately the meaning of quite a number of words and phrases is ambiguous; this results in the fact that their translation highly depends on the topic and context they are used in.

For example the word 'leg' is usually thought of as a body part ('He broke his leg'). In sports, especially bicycling, the word 'leg' can also have the meaning of 'stage' ('US Postal wins fourth leg'). Similar to this meaning is the use in aviation with the phrase 'single leg airline'.

This fact would not be a problem if the translations for 'leg' were the same in every case. But this is rarely true. German for example uses different words for the upper three meanings of 'leg'. So a translation that might be totally acceptable for one specific topic, applied to test data in another topic will lead to an error in the translation.

1.1. Basic Idea

Our approach is similar to recent approaches to language model adaptation. We try to find sen-

tences from the training data, which are similar to the test sentences. Then we train the translation system only on this selection. This reduced training data hopefully matches the test data better in domain, topic and style thus improving translation performance.

1. for each test sentence
   - use test sentence to select n most similar sentences in the training data
2. build translation model only using the training sentences found for each test sentence
3. translate with adapted translation model

1.2. Previous Work

The main idea is based on the work that was done for language model adaptation. Mahajan et al. (1999) used similar techniques for language model adaptation in speech recognition. This was applied to Statistical Machine Translation by Eck et al. (2004) and further refined by Zhao et al. (2004). Kim and Khudanpur (2003) used a similar idea for their language model adaptation and introduced the idea to use the likelihood of their first pass speech recognition result according to the adapted language model to find the optimal number of retrieved documents to use.

There have not been a lot of publications for the adaptation of the translation model for Statistical Machine Translation yet. One method for the adaptation of the translation model was proposed by Wu and Wang (2004). Wu and Wang focus on the actual word alignment and improve it by training different alignment models from in-domain and out-of-domain data. It is necessary for this approach to have at least a small separate amount of in-domain data available.

2. Translation Model Adaptation

2.1. Selecting Sentences using Information Retrieval

For information retrieval we used the source language part of the bilingual training data as the document collection, each sentence representing one document. Using only the source language for the information retrieval has the advantage, that it is independent from the quality of the translation system, as no first pass translation is necessary.

Each sentence from the test data was used as one separate query.

For most of the experiments we used cosine distance similarity measure with TF-IDF term weights to determine the relevance of a query to a document.

TF-IDF term weighing is widely used in information retrieval. Each document D_i is represented as a vector $(w_{i1}, w_{i2}, ..., w_{ik})$ if *k* is the size of the vocabulary. The entry w_{ij} is calculated as:

$$w_{ij} = tf_{ij} * \log(idf_j) .$$

tf_{ij} is the weighted term frequency of the *j*-th word in the vocabulary in the document D_i i.e. the number of occurrences.

idf_j is the inverse document frequency of the *j*-th term, given as

$$idf_j = \frac{\#\,\text{documents}}{\#\,\text{documents containing j - th term}}$$

The similarity between two documents is then defined as the cosine of the angle between the two vectors.

2.2. Training an Adapted Translation System

We use the top *n* similar sentences for each sentence from the test data to train the translation model.

We do not train separate translation models for each sentence, but put all retrieved sentences together to form the new training set. The reason for this is that a translation model trained from only a few hundred sentences is unlikely to give robust probabilities. It can also be expected that a smaller test set will not change its domain so rapidly that a phrase or word translation, that is correct in the beginning of the document, would be wrong in later sentences. If a particular test consists of parts from different domains, a solution could be to train separate translation models for these parts of the test set.

It is also relevant to note that this training set can contain duplicate sentences as the top *n* retrieval results for different test sentences can contain the same training sentence. (It will certainly contain duplicate sentences for higher val-

ues of *n* as the adapted training set becomes larger than the amount of available sentences).

It is questionable if the duplicates help the translation performance. The duplicates force the translation probabilities towards the more often seen words which could help, but the adaptation should already take care of this.

We re-did all experiments with removed duplicates in the first experiment to see how the duplicate sentences effect the translations.

In the first experiment we always use a language model built from the entire training data. In some sense this language model is not matching the translation models, which were adapted. The general language model does not further support this adaptation. It is even possible that the general language model contradicts a correct translation for a specific topic and another – wrong – path is chosen in the decoding process.

We tried to resolve this un-matching condition by changing the language model training data as well and used the English part of the adapted training set to train a new adapted language model in a second experiment.

2.3. Language Model Perplexity for Measuring Selection Quality

One unsolved question at this point is how many sentences to select for the adapted training corpus.

As shown in the experiments (see sections 3–5) the optimal size of the adapted training corpus is different for different language pairs, training corpora or test sets. To be able to do a grid search for the optimal selection size, it is necessary to use a development test set with reference translations. The estimate for the optimal selection size then has to be transferable to the actual test set. The optimal number of sentences to select for training might also vary for each individual test sentence.

It would be very useful, not to be forced to compare translation scores for many experiments to estimate the selection size.

Following an idea introduced by Kim and Khudanpur (2003), to judge how well a selection of training data fits the test sentence, we measure the perplexity (PPL) of a language model built from this selection against the test sentence. Then we find the perplexity minimum to determine the optimal selection size for each test sentence. Still the main selection criterion is TF-IDF information retrieval, as we look only at the e.g. top 1000 sentences ranked by TF-IDF retrieval.

Diagram 1 shows the behavior of the perplexity of the language model (LM) built from the top 10, 20, 30…1000 sentences against the respective test sentence. (For the diagram we randomly chose 4 sentences from the Spanish – English experiment setting (see section 4).)

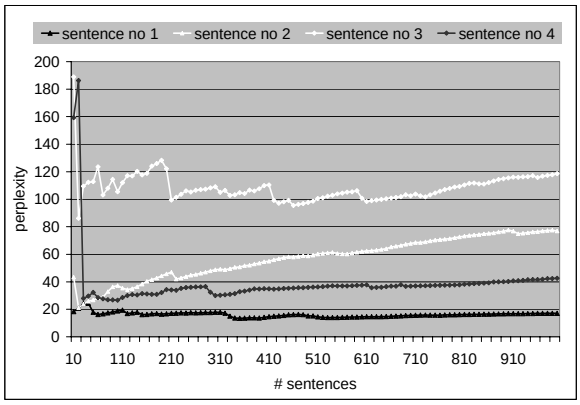

Diagram 1: LM perplexities for all selection sizes

Unfortunately the perplexity curve does not show a nice convex shape for most sentences. There are even sentences, where the perplexity minimum is at the first or second batch. The previous experiments have shown, that the optimal selection size is definitely bigger than 10 sentences per test sentence. So picking the selection with the lowest perplexity seems not to be reasonable in many cases.

Because information retrieval ranks the sentences according to their term weights, while language model perplexity gives information about matching word order, some of the 10-sentence-batches added early to the selection make the perplexity worse, while some 10-sentence-batches ranked lower in the TF-IDF retrieval improve the perplexity.

To exploit that additional information, we use the perplexity change each batch of sentences causes as an additional measure for ranking sentences on the top n sentences retrieved by TF-IDF. Because the size of the selection increases over the testing run, the changes in perplexity are not comparable, so the batches can't be completely re-ranked according to the perplexity change. The batches are only classified as 'good' or 'bad' during the pass. All the 'bad' batches

are taken out of the list and are being shuffled to the end. Among the good as well as the bad batches we keep the original TF-IDF ranking.

After re-ranking once the shape of the perplexity curve is already smoother and has a considerably lower perplexity minimum than the original order. After re-ranking a second time, the measured perplexities are even lower (Diagram 2). There are already almost no 'bad' batches before the minimum and 'good' batches after it after re-ranking twice for most sentences. So it's not worth the computation time to re-rank a third time.

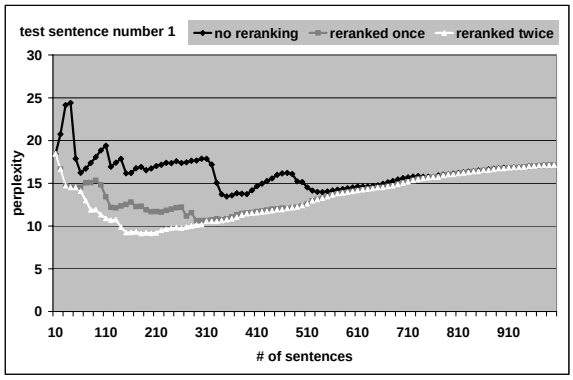

Diagram 2: Perplexity re-ranking

After re-ranking the selection size was determined for each sentence by picking the selection with the lowest perplexity.

This technique can do without any development test set, translation run or even a reference translation to adapt the translation model.

3. Experiments

3.1. Overview

We did our experiments for two different corpora and setups. The first setup translating Spanish to English in the medical domain was used to test the basic idea and check different settings.

The experiment translating Chinese to English proves that the ideas can be applied to another domain (tourism) and an overall different scenario as the out-of-domain data there is much larger than the available in-domain data.

Both experiments use a small amount of in-domain training data and an additional larger amount of out-of-domain data. In both cases just adding the out-of-domain data does not significantly improve the performance of a baseline system that was trained on the in-domain data only.

The adaptation can then be viewed in two different ways: The adaptation can improve word translations by using translations that are more appropriate for the topic. This is the case for the baseline systems that use all available in-domain and out-of-domain data. For baseline systems that have only been trained on the available small in-domain data the goal of the adaptation is to cover unknown words. Words that are covered by the available in-domain data can be translated fairly well. The hope is that the additionally selected data will cover previously unknown words.

3.2. Translation System

The applied statistical machine translation system uses IBM1 lexicon transducers and different types of phrase transducers (Zhang et al., 2003; Vogel et al., 1996; Vogel et al., 2003). The Language model is a trigram language model with Kneser-Ney-discounting built with the SRI-Toolkit (SRI, 1995-2004) using only the English part of the training data. This system was used for all experiments.

The best scores for NIST (Doddington; 2001) or BLEU (Papineni et al.; 2002) evaluation metrics are usually achieved using considerably different tuning parameters for the translation system. In the experiments for the Spanish–English translation the system was only tuned towards NIST, in the Chinese–English experiments we tuned the system towards both NIST and BLEU respectively.

4. Experiments Spanish – English

4.1. Test and Training Data

The test data for the Spanish–English experiments consisted of 329 lines of medical dialogues (6 doctor-patient dialogues). It contains 3,399 English words and 3,065 Spanish words (tokens) with one reference translation.

We had 3 different corpora of bilingual training data available. 25,077 lines of medical dialogues can be regarded as in-domain data. Additional out-of-domain data were 2,323 lines of tourism dialogues and 123,416 lines of BTEC data (also tourism domain, general tourist sentences and phrases) described in Takezawa et al (2002).

Training sets	#lines	#words (English)	#words (Spanish)
Medical dialogues (in-domain)	25,077	218,788	208,604
Tourism dialogues (out-of-domain)	2,323	26,600	24,375
BTEC data (out-of-domain)	123,416	903,525	852,364
Overall	150,816	1,148,913	1,085,343

Table 1: Training Data sizes for Experiments Spanish – English

4.2. Baseline Systems

We trained two different baseline systems. The first system only uses the medical data. In some sense this is an oracle experiment, because it might not always be known what part of the available data is the actual in-domain data. The second baseline system uses all available training data.

The scores show, that the baseline system that only uses the available in-domain data is not necessarily better than the system that uses all data. The best NIST score is actually a little higher for the second baseline system (but not statistically significant). There may be two possible reasons for the improvement using the additional data.

1. It covers 27% of the previously unknown words (36 of 132).
2. It consists of dialogues like the medical data. Those dialogues cover a different topic, but they still might be helpful for the translation as the sentence structure is fairly similar.

System	**NIST**
only in-domain data	5.1820
in-domain and out-of-domain data	5.2074

Table 2: Baseline System results

In this experiment the translation system was only tuned towards the NIST score.

4.3. Experiment 1: distinct and non-distinct retrieval

For the Spanish – English setting we built the information retrieval index using the Spanish part of all available in-domain and out-of-domain data. (We used the Lemur Toolkit (Lemur) for all Information retrieval tasks)

The top n similar sentences for each Spanish test sentence for n=30, 50, 100, 200, 300... 1000 were then retrieved from the index, using TF-IDF as the similarity measure.

For n=50 the selection for the entire test set contained 40% duplicates, 75% for n=1000.

It is also important to note that the Lemur toolkit sometimes retrieved fewer sentences than was asked for. This happens especially for short sentences, when all remaining sentences have no TF-IDF weight because not even one word matches.

This training set was used to train the new adapted translation models. The LM was trained on the entire training data.

Diagram 3 illustrates the results. The numbers in parentheses on the x-axis denote the number of distinct sentences that were used to train this particular system. The non-distinct training set contained some of those distinct sentences more than once.

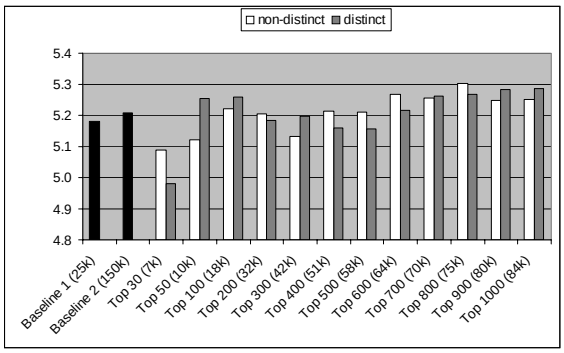

Diagram 3: Distinct and non-distinct retrieval for Spanish–English (NIST scores)

The highest NIST score for this experiment in the non-distinct case was 5.3026 at Top 800 retrieved sentences. This training set has about 250,000 sentences (with duplicates) and about 75,000 distinct sentences which is about half the size of the original training data.

In the distinct case, when the duplicate sentences were removed for the actual training the highest NIST score was 5.2878 for Top 900 (about 80,000 sentences).

4.4. Experiment 2: TM and LM Adaptation

As noted earlier, we always used the baseline (baseline 2) language model for the translations in experiment 1.

In experiment 2 we changed the language model training data as well and used the English part of the adapted training set to train the new language model. This had a bigger impact on the smaller systems, as the adapted and the general LM become more similar for larger selection sizes.

This further improved the best NIST score to 5.3264 (Top 200 with about 64,000 sentences of overall training data and just about 32,000 distinct sentences).

Diagram 4 illustrates the results in NIST score. All these experiments were done without removing the duplicate sentences.

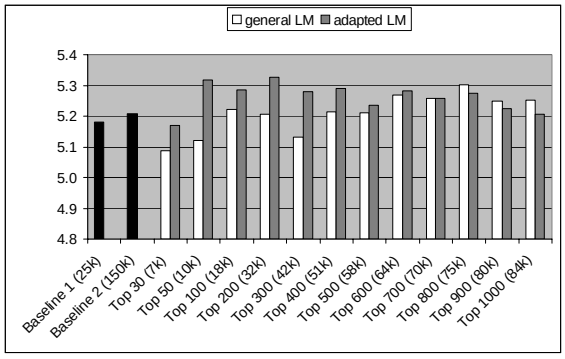

Diagram 4: TM and LM adaptation for Spanish–English (NIST scores)

4.5. Experiment 3: Perplexity based Selection Size Determination

To find the optimal selection size for the adapted training corpus we re-ranked the top 1000 sentences retrieved via TF-IDF retrieval. The perplexity was calculated after adding sentences in batches of 10.

In this experiment we always built the language model from the adapted training data, as this worked well for the previous experiments.

Diagram 5 shows the NIST scores for selection sizes picked at the perplexity minimum before re-ranking and after re-ranking once and twice in comparison to the baselines and the best scores from the previous experiments. The best NIST score of 5.3807 was reached after re-ranking twice.

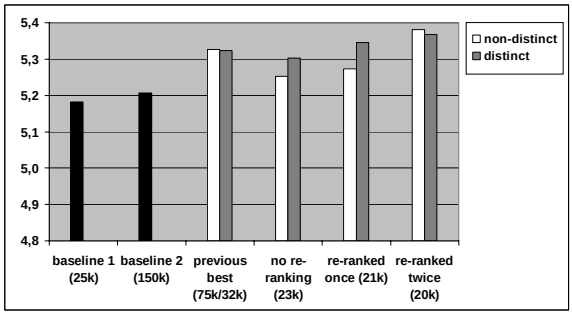

Diagram 5: PPL based selection size and re-ranking for Spanish-English (NIST scores)

4.6. Summary

The differences between the systems with or without duplicate sentences are not significant. The highest NIST score was reached using a training set that contained duplicates.

Training the language model on an matching adapted data selection clearly improves the performance.

The selection automatically found by perplexity based selection size determination was able to achieve about the same scores as the best one of a whole set of selection sizes, PPL re-ranking improved slightly over them.

System	NIST
baseline 1: in-domain data	5.1820
baseline 2: all data	5.2074
best TF-IDF with duplicates	5.3026
best TF-IDF distinct	5.2878
best with LM adaptation	5.3264
best with PPL re-ranking	5.3807

Table 3: Results for each experiment: Spanish-English

5. Experiments Chinese – English

5.1. Test and Training data

The Test Data for the Chinese–English experiments consisted of 506 lines of tourism dialogues. The test data contains 3510 Chinese words. There are 16 English references per test sentence available.

The in-domain training data consisted of exactly 20,000 lines of tourism dialogues, also from the BTEC data.

We used additional 9.1 million lines of TIDES data (mainly Chinese newswires and speeches) to build the index and retrieve the additional data.

Training sets	#lines	#words (English)	#words (Chinese)
BTEC data *(in-domain)*	20,000	188,935	175,284
TIDES data *(out-of-domain)*	9.1 million	144 million	135 million

Table 4: Training Data sizes for Experiments Chinese–English

5.2. Baseline System

The baseline system was only trained on the available in-domain data and had a NIST score of 8.1129 and a BLEU score of 0.4621.

It was known from earlier results that a system using all available training data does not improve over this baseline. The vocabulary coverage certainly improves (89 unknown words in the baseline, 4 with the complete TIDES corpus) but the out-of-domain data introduces too many wrong translations. We did not explicitly train another baseline from all data for this reason.

In this in-/out-of-domain data scenario one could argue, that adding some data to the small initial system will improve the translation performance, no matter what data is selected. So we selected different numbers of sentences randomly from the complete training corpus and compared the translation results to our adaptive selection. From different random selections only small ones could improve over the baseline (2 examples are given in table 2).

System	BLUE	NIST
only in-domain data (20k lines)	0.4621	8.1129
Randomly selected out-of-domain data 15k lines	0.4850	8,2262
Randomly selected out-of-domain data 75k lines	0,4501	7,9482

Table 2: Baseline System results: Spanish-English

This shows the trade-off between a small domain-specific model that can not cover all words and a larger system that might introduce wrong out-of-domain translations.

5.3. Experiment 4: In-domain/out-of-domain data scenario

With this small amount of in-domain training data at hand we built the index for the out-of-domain data only. The top n similar sentences for each Chinese test sentence for n=10, 20, 30, 40, 60, 70, 80, 100, 125, 150, 175, 200, 250 and 300 were then retrieved from the index, using TF-IDF as the similarity measure.

We then added the retrieved sentences from the out-of-domain data to the in-domain data for the training of the translation model.

As we felt that the available in-domain data was too poorly represented especially if we added more and more training data for a larger number of retrieved n we removed the duplicates in all cases. In additional translation runs we also weighted the in-domain data three times (instead of once) in the training to get more robust probabilities for the words already known to be in-domain (denoted by 'weight 3:1' in the diagrams). As expected this especially helped with the larger selection sizes.

The overall best scores were 8.3398 (NIST) and 0.4931 (BLEU). Both scores were accomplished with changed weight of the in-domain data, the best NIST score for the Top 60 retrieved sentences, the best BLEU score for the Top 80 retrieved sentences. Diagrams 6 and 7 illustrate the further results. (The number in parentheses on the x-axis denotes the amount of training data in lines that was added to the available in-domain data of 20,000 lines to form the overall training data.)

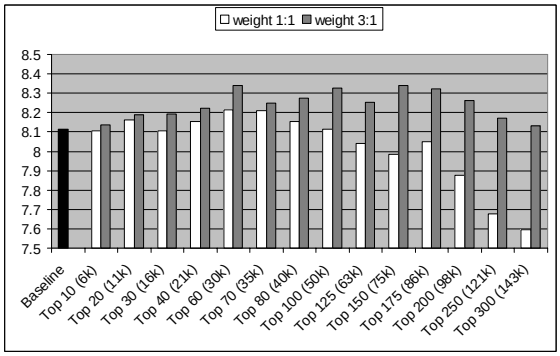

Diagram 6: Chinese-English: different selection sizes (NIST scores)

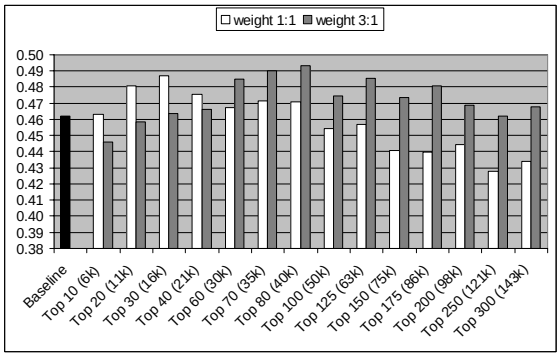

Diagram 7: Chinese-English: different selection sizes (BLEU scores)

5.4. Experiment 5: Perplexity based Selection Size Determination

In this data setting we chose a batch size of 20 and re-ranked only the top 800 retrieved sentences in the first and the top 600 in the second perplexity re-ranking run because of runtime issues due to the big data collection and vocabulary size.

Diagram 8 and 9 show the NIST and BLEU scores for selection sizes picked at the perplexity minimum before re-ranking and after re-ranking once and twice in comparison to the baselines and the best scores from the previous experiments.

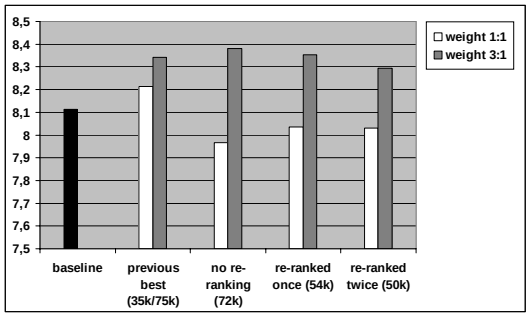

Diagram 8: Perplexity determined Selection Size and Re-ranking (NIST scores)

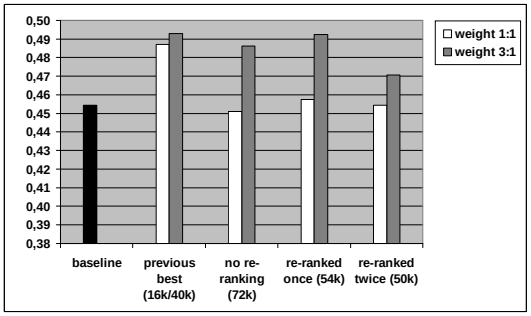

Diagram 9: Perplexity determined Selection Size and Re-ranking (BLEU scores)

In this experimental setting the re-ranking itself gave no real improvement but the automatic determination of the selection size was able to reach the same results achieved by trying various selection sizes. The reason might be, that the 3:1 weight for in- and out-of-domain data is far from optimal for these selection sizes. The weighing had a big impact on the scores, drowning out the possibly positive effect of re-ranking.

5.5. Summary & Example translations

System	BLEU	NIST
baseline: in-domain data	0.4621	8.1129
best random	0.4850	8.2262
best weight 1:1	0.4871	8.2132
best weight 3:1	0.4931	8.3398
best with PPL selection size	0.4924	8.3812

Table 5: Results for each experiment: Chinese-English

Table 6 shows some example translations comparing the reference with the baseline and best system (according to NIST score).

Reference	no-smoking, please.
Baseline	i ‘d like a seat please
Best system	i ‘d like a no smoking seat please
Reference	can i have a medical certificate?
Baseline	could you give me a medical open
Best system	could you give me a medical certificate
Reference	three glasses of melon juice, please.
Baseline	please give me three of those melon juice please
Best system	please give me three glasses of melon juice please
Reference	excuse me. could you tell me how to get to the getty museum?
Baseline	excuse me could you tell me the way to the art museum yosemite san diego please
Best system	excuse me could you tell me how to get to the museum

Table 6: Example Translations: Chinese-English

6. Further results

There are several other similarity measures that are widely used in information retrieval. We compared results using the Okapi similarity measure instead of TF-IDF and found no significant difference in translation quality. Looking at the retrieval result for the whole test set the portion of retrieved sentences from the TF-IDF retrieval that can be found in the Okapi retrieval result amounts to over 75% for the top 300 sentences per query and over 90% for the top 1000 retrieved sentences per query.

7. Future Work

Different things could be done to further investigate this approach to translation model adaptation. We already tried the TF-IDF and Okapi similarity measures but those only focus on unigrams. It could be helpful to develop a more sophisticated similarity measure that matches phrases, too. It was demonstrated in Zhao et al. (2004) that language model adaptation could benefit from such an advanced similarity measure and it is certainly possible to apply these ideas here. Other information retrieval techniques like stemmers, the usage of a stop-word list or pseudo feedback could be applied, too.

It might also be beneficial to use training algorithms that allow sentences to have fractional weights. Section 5.4 showed that tuning weights for in- and out-of domain data can give improvements. Determining the best weight in each situation would certainly be helpful and it could be interesting to further investigate this behavior.

Another possible experiment could be to train separate translation models for the in-domain and retrieved out-of-domain data and interpolate those models.

The LM adaptation in the presented experiments is always based on the source side. It is possible that target side LM adaptation approaches as presented in Eck et al. (2004) and Zhao et al. (2004) combined with the TM adaptation as presented in this paper could further improve the translation performance.

8. Conclusions

We show that it is possible to adapt translation models for statistical machine translation by selecting similar sentences from the available training data. There are improvements in translation performance on two different language pairs and overall different test conditions.

The results show that it is helpful to support this adaptation method by analogically adapting the language model as this further improves the translation quality.

Using language model perplexity to determine the selection size automatically renders a development test set with reference translations unnecessary. Re-ranking the retrieval result according to LM perplexity even improved translation quality slightly in one of the cases.

With more investigation especially into optimizing the weights between in- and out-of-domain data, it will hopefully be possible to further improve the translation performance.

References

BROWN, Peter E., DELLA PIETRA, Stephen A., DELLA PIETRA, Vincent J. and MERCER, Robert L. (1993). 'The mathematics of statistical machine translation: Parameter estimation', Computational Linguistics, 19(2), pp. 263-311

DODDINGTON, George (2001). 'Automatic Evaluation of Machine Translation Quality using n-Gram Co-occurrence Statistics'. NIST Washington, DC, USA.

ECK, Matthias, VOGEL, Stephan and WAIBEL, Alex (2004). 'Language Model Adaptation for Statistical Machine Translation based on Information Retrieval', Proceedings of LREC 2004, Lisbon, Portugal, May 2004.

KIM, Woosung and KHUDANPUR, Sanjeev (2003). 'Language Model Adaptation Using Cross-Lingual Information', Proceedings of Eurospeech 2003, Geneva, Switzerland, Sept. 2003.

'The LEMUR Toolkit for Language Modeling and Information Retrieval' http://www.cs.cmu.edu/~lemur/

MAHAJAN, Milind BEEFERMAN Doug and HUANG, X.D. (1999). 'Improved Topic-Dependent Language Modeling Using Information Retrieval Techniques', IEEE International Conference on Acoustics, Speech and Signal Processing 1999, Phoenix, AZ.

PAPINENI, Kishore, ROUKOS, Salim, WARD, Todd and ZHU, Wei-Jing (2002). 'BLEU: a Method for Automatic Evaluation of Machine Translation', Proceedings of the ACL 2002, Philadelphia, USA.

TAKEZAWA, Toshiyuki, SUMITA, Eiichiro, SUGAYA, Fumiaki, YAMAMOTO, Hirofumi, YAMAMOTO, Seiichi (2002) 'Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World', LREC 2002 (Third International Conference on Language Resources and Evaluation), Vol.1, pp.147-152

VOGEL, Stephan, NEY, Hermann and TILLMANN, Christoph (1996). 'HMM-based Word Alignment in Statistical Translation', Proceedings of COLING 1996: Proceedings of Coling 1996. Copenhagen, August 1996.

VOGEL, Stephan, ZHANG, Ying, HUANG, Fei, TRIBBLE, Alicia, VENUGOPAL, Ashish, ZHAO,

Bing, WAIBEL, Alex (2003). 'The CMU Statistical Translation System', Proceedings of MT-Summit IX 2003. New Orleans, LA. Sept. 2003.

WU, Hua and WANG, Haifeng (2004). 'Improving Domain-Specific Word Alignment for Computer Assisted Translation', Proceedings of ACL 2004, Barcelona, Spain, July 2004.

ZHANG, Ying, VOGEL, Stephan and WAIBEL, Alex (2003). 'Integrated Phrase Segmentation and Alignment Algorithm for Statistical Machine Translation', Proceedings of International Conference on Natural Language Processing and Knowledge Engineering 2003, Beijing, China, Oct. 2003.

ZHAO, Bing, ECK, Matthias and VOGEL, Stephan 2004. 'Language Model Adaptation for Statistical Machine Translation via Structured Query Models', Proceedings of Coling 2004, Geneva, Switzerland, Aug. 2004.