Domain Adaptation in Statistical Machine Translation using Factored Translation Models

Jan Niehues and Alex Waibel
Institute for Anthropomatics
Karlsruhe Institute for Technology
Karlsruhe, Germany
{jan.niehues,alex.waibel}@kit.edu

Abstract

In recent years the performance of SMT increased in domains with enough training data. But under real-world conditions, it is often not possible to collect enough parallel data. We propose an approach to adapt an SMT system using small amounts of parallel in-domain data by introducing the corpus identifier (corpus id) as an additional target factor. Then we added features to model the generation of the tags and features to judge a sequence of tags. Using this approach we could improve the translation performance in two domains by up to 1 BLEU point when translating from German to English.

1 Introduction

Statistical machine translation (SMT) is currently the most promising approach to machine translation of large vocabulary tasks. The approach was first presented in Brown et al. (1993) and has been used in many translation systems since then.

One drawback of this approach is that large amounts of training data are needed. Furthermore, the performance of the SMT system improves if this data is selected from a similar topic and from a similar genre. Since this is not possible for many real-world scenarios, one approach to overcome this problem is to use all available data to train a general system and to adapt the system using in-domain training data.

Factored translation models as presented in Koehn and Hoang (2007) are able to tightly integrate additional knowledge into a phrase-based statistical machine translation system. In most cases, the approach is used to incorporate linguistic knowledge, such as morphological, syntactic and semantic information. In contrast, we will use the approach to integrate domain knowledge into the system by introducing a corpus identifier (corpus id) tag.

Using the corpus id as a target word factor enables us to adapt the SMT system by introducting two new types of features in the log-linear model in phrase-based SMT systems. First, we will use relative frequencies to model the generation of the corpus id tags similar to the translation model features that are used to model the generation of the target words in a standard phrase-based system. Furthermore, we can use features comparable to the word count and language model features to judge the generated sequence of corpus id tags. Moreover, using the general framework of factored translation models leads to a simple integration of this approach into state-of-the-art phrase-based systems.

The remaining part of the paper is structured as follows: First, we present some related work in the next section. Afterwards, in Section 3, a motivation for and an overview over the presented model is given. In the following two sections the new types of features are introduced. In Section 6 the approaches are evaluated and in the end a conclusion is given.

2 Related Work

In recent years different methods were proposed to adapt translation systems to a domain. Some authors adapted only the language model inspired by similar approaches in speech recognition (Bulyko et al., 2007). The main advantage of language model adaptation in contrast to translation model adaptation is that only monolingual in-domain data is needed.

To be able to adapt the translation model in these conditions, other authors tried to generated a synthetic parallel text by translating the monolingual corpus with a baseline system and use this corpus to train a new system or to adapt the baseline system (Ueffing et al., 2007), (Schwenk and Senellart, 2009), (Bertoldi and Federico, 2009).

Snover et al. (2008) used cross-lingual information retrieval to find similar target language corpora. This data was used to adapt the language model as well as to learn new possible translations. Wu et al. (2008) presented an approach to adapt the system using hand-made dictionaries and monolingual source and target language text.

In cases where also in-domain parallel data is available, authors also tried to adapt the translation model. Koehn and Schroeder (2007) adapted the language model by linear and log-linear interpolation. Furthermore, they could improve the translation performance using two translation models and combining them with the alternate decoding path model as described in Birch et al. (2007). Although this approach does also use factored translation models, the way they integrate the domain and the type of features they use is different from ours.

An approach based on mixture models was presented by Foster and Kuhn (2007). They tried to use linear and log-linear, language model and translation model adaptation. Furthermore, they tried to optimize the weights for the different domains on a development set as well as to set the weights according to text distance measures. Matsoukas et al. (2009) also adapt the system by changing the weights of the phrase pairs. In their approach this is done by assigning discriminative weights for the sentences of the parallel corpus.

In contrast, Hildebrand et al. (2005) proposed a method to adapt the translation towards similar sentences which are automatically found using information retrieval techniques.

Factored translation models were introduced by Koehn and Hoang (2007) to enable the straightforward integration of additional annotations at the word-level. This is done by representing a word by a vector of factors for the different types of annotations instead of using only the word token. The additional factors can be used to better judge the generated output as well as to generate the target word from the other factors, if no direct translation of the source word is possible. Factored translation models are mainly used to incorporate additional linguistic knowledge, for example, part-of-speech information. It could be shown that the performance of SMT systems can be improved by incorporating linguistic information using this approach.

3 Factored Domain Model

In this section, we will first give a short motivation for modeling the domain of the training data. Afterwards, we will describe how to introduce the domain information into a phrase-based translation model using the framework of factored translation models.

3.1 Motivation

In the phrase-based translation approach every training sentence is equally important for generating the translations. In many cases this simplification is acceptable, but it no longer holds if the training corpus consists of an in-domain and out-of-domain set. In this case, the information learned from the in-domain set should be more important than the one from the out-of-domain set.

The simplification mentioned before does lead to many translation errors, if the size of the in-domain training data is small compared to the out-of-domain data, which is the case in most applications. In these cases, a small amount of in-domain training data is not able to improve the translations quality as much as possible. To be able to make better use of examples, the in-domain translations should be more important in the training process than the ones from out-of-domain data.

For example, for the German-English translation task, the biggest available parallel corpus are the Proceedings of the European Parliament. In this context, some words have different English translations than they would have in a news document. If all sentences are treated equally, the probability of the translations specific to the proceedings would be more probable, since they were seen more often.

3.2 Model

To be able to overcome the problems mentioned before, we tried to model the influence of the in-domain and out-of-domain data explicitly. To be able to do this, we store with every phrase pair the corpus it is extracted from. Using this information, the phrase pairs are no longer equally important, but we can, for example, prefer phrase pairs that

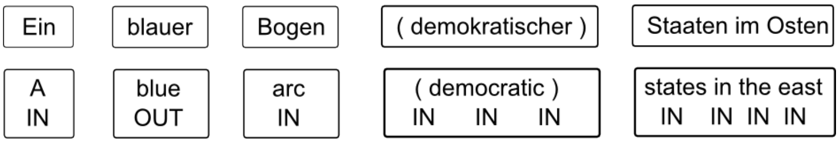

Figure 1: Example of German to English translation with corpus id factors

were extracted from the in-domain corpus.

To model this idea we used the general framework of factored translations models. In this framework we add one factor to the target words representing the corpus id.

The resulting representation of an example sentence is shown in Figure 1. In the example we use two corpus ids. One for the in-domain and one for the out-of-domain part of the corpus. Phrases extracted from the out-of-domain corpus, generate the *OUT* factor on the target side. In contrast, phrase pairs learned from the in-domain part will generate an *IN* factor on the target side.

Many phrase pairs occur in different parts of the corpus. For example, the phrase pair *Ein # A* shown in the example in Figure 1 does also occur in the out-of-domain part of the corpus. In these cases, both phrase pairs will be extracted and the decoder will select one of them depending on the models described in the following.

With this approach it is possible to see which parts of the translation are learned from the in-domain training examples and which parts are translated by using phrase pairs from the out-of-domain corpus. This information can then be used to judge the quality of the translation. That means, translations which are generated from in-domain phrase pairs will more probably be a better translations than the ones generated merely by phrase pairs extracted from the out-of-domain corpus.

To be able to model this, we add two types of features to the log-linear model used in a phrase-based translation system. The first one that we call the *Domain Factors Translation Model*, models the probability that a sequence of corpus id tags is generated. As it is done for the translation model of words, we use features based on relative frequencies to model this probability. The different features are described in detail in the next section.

A second group of features is used to judge the corpus id tag sequence similar to the target language model used in SMT systems. For example, we count the number of in-domain tags and use this as an *Domain Factors Sequence Model*. Different approaches to model this probability will be described in Section 5.

Since we use the general framework of factored translation models, the weights for these features can be optimized during the training on the development data of the log-linear model using, for example, Minimum Error Rate training. The resulting weights prefer in-domain phrase pairs in a way that leads to the best translation performance on the development data.

4 Domain Factors Translation Model

The domain factors translation model is used to describe the probability that a sequence of corpus id tags is generated. If we look at the example mentioned before, the features of the model should model the probability of generating the sequence *IN OUT IN IN IN IN IN IN IN* if the input sentence is *Ein blauer Bogen (demokratischer) Staaten im Osten*.

As mentioned before, this is similar to the phrase translation model in state-of-the-art SMT approaches. In most cases, this is described by a log-linear combination of four different probabilities. First, the probability of the target phrase given the source phrase $P(t|s)$ is approximated by the relative frequency

$$P(t|s) = \frac{cooc(s,t)}{cooc(s,*)} \tag{1}$$

where $cooc(s,t)$ is the number of cooccurences of the source phrase s and the target phrase t in the parallel training corpus and $cooc(s,*)$ the number of phrase pairs extracted from the parallel corpus with source phrase s.

The second probability $P(s|t)$, the probability of the source phrase given the target phrase approximated in the analogous:

$$P(t|s) = \frac{cooc(s,t)}{cooc(*,t)} \tag{2}$$

In addition the lexical translation probabilities of both directions are used in a default configuration of an SMT system.

In our factored model we no longer have only the coocurrence count depending on the source and target phrase $cooc(s,t)$, but, in addition, a coocurrence count depending on three parameters, $cooc(s,t,d)$, where d is the sequence of corpus id tags. Consequently, we can extend the existing probabilities by three more possible features.

First, we can use the probability of the corpus id tags given the source phrase $P(d|s,t)$ that can be approximated analogously to the existing translation probabilities by

$$P(d|s,t) = \frac{cooc(s,t,d)}{cooc(s,t,*)} \tag{3}$$

Secondly, we can also use the probability of the target phrase given the source phrase and domain tag sequence $P(t|s,d)$. Since we cannot extract a phrase pair partly from one corpus and partly from another one, the corpus id tags for all words of one phrase pair is the same. Consequently this probability equals the probability of the target phrase given the source phrase restricted to the phrases extracted from the corpus indicated by the corpus id tags. The probability can be approximated by:

$$P(t|s,d) = \frac{cooc(s,t,d)}{cooc(s,*,d)} \tag{4}$$

At last, we can define the probability with switched roles of the source and target phrase $P(s|t,d)$. This can be approximated by:

$$P(s|t,d) = \frac{cooc(s,t,d)}{cooc(*,t,d)} \tag{5}$$

5 Domain Factors Sequence Model

In the last section we described how to model the generation of the corpus id tags. After generating the corpus id tag sequence, one main advantage of this approach is that we are able to introduce models to judge the different possible domain tag sequences for a given source sentence.

In the example given in Figure 1 another possible translation of the source sentence would generate the tag sequence *IN OUT OUT IN IN IN OUT OUT OUT OUT*. By only looking at the corpus id sequence we should prefer the translation shown in the figure, since it uses more phrase pairs that occur in the in-domain corpus. To model this we propose two features.

In contrast to the translation model, in this case we cannot simply extend the language model approach used for the target words. This would mean, we train a language model on the corpus id tag sequence and then use this language model to evaluate the tag sequence. The problem is that a training sentence is always only from one document. Consequently, all words have got the same corpus id tag. But in the test case, we do not only want to generate sentences using phrase pairs extracted from the same corpus. Consequently, there is no corpus to train a language model and we have to use different types of models.

Therefore, we use a unigram model in the experiments, although the framework supports general sequence models. One possibility is to do this at phrase level leading to a model similar to the phrase count model used in nearly every phrase-based SMT decoder. Instead of counting all phrases we can just count the phrases which have got in-domain corpus id tags. In the example shown in Figure 1 this would lead to a value of 4 for this feature.

Another approach is to use a in-domain word count feature similar to the already existing word count feature. We evaluated both types of feature and present the results in Section 6.

6 Evaluation

We evaluated our approach to adapt an SMT system on the German-English translation task. For this language pair, the biggest parallel corpus available are the Proceedings of the European Parliament (EPPS).

The first system we built was designed to translate documents from the news-commentary domain. As test set we used the test set from the WMT Evaluation in 2007. As parallel training data we used the EPPS corpus as well as an in-domain news-commentary corpus with about 1M words. In contrast, the corpus from the EPPS domain has about 39M words.

In a preprocessing step we cleaned the data and performed a compound splitting on the German text based on the frequency method described in Koehn et al. (2003). We generated a word alignment for the parallel corpus using a discriminative approach as described in Niehues and Vogel (2008) trained on a set of 500 hand-aligned sentences. Afterwards, the phrase pairs were extracted using the training scripts from the Moses package (Koehn et al., 2007).

We use two language models in our SMT system. The first one, a general language model, was trained on the English Gigaword corpus. In addition, we use a second one, trained only on the English part of the parallel in-domain corpus. Using this additional language model, our baseline system was already partly adapted to the target domain.

To be able to model the quite difficult reordering between German and English we used a part-of-speech based reordering model as described in Rottmann and Vogel (2007) and Niehues and Kolss (2009). In this approach reordering rules based on part-of-speech tags are learned from the parallel corpus. For every test sentence, different possible reorderings of the source sentence are encoded in a word lattice. Then the decoder translates this lattice instead of the original input sentence.

We use a phrase-based decoder as described in Vogel (2003) using the language models, phrase scores as well as a word count and phrase count model. The optimization was done by MER training described in Venugopal et al. (2005).

We performed a second series of experiments on the translation task of lectures from German to English. The system was trained on the data from the European Parliament, the news-commentary data, German-English BTEC data and a small amount of translated lectures. The in-domain corpus contained only around 210K words.

The system was built similar to the systems in the other experiments except to some small changes due to speech translation. Instead of doing a separate compound splitting, we used the same splitting as it was used by the speech recognizer. Since the output of the speech recognition system is lower-cased, we lower-cased the source part of the phrase table. In the case where this lead to two identical phrase pairs, we kept both.

6.1 Results for News Task

We first evaluated our approach on the task of translating documents from the news-commentary domain. We performed experiments to analyze the influence of the domain sequence model and a second group of experiments to look at the domain translation model.

The results using different sequence models are shown in Table 1. The baseline system does not use any adaptation. Then we added a second in-domain target language mode. In addition, the other three systems use the domain translation model as described in Section 4 using only the domain relative frequency as introduced in Equation 3.

Table 1: Different sequence models for domain factors (BLEU)

	System	Dev	Test
1	Baseline	25.90	29.03
2	(1) + LM Adaptation	26.68	29.24
3	(2) + Domain Rel. Frequency	26.80	29.21
4	(3) + Word Count Model	27.03	29.63
5	(3) + Phrase Count Model	27.09	29.54

The first system does not apply the domain sequence model. As it can be seen in the table, in this case the approach could not improve the translation quality compared to the baseline system.

Using one of the two different types of sequence modeling as described in Section 5 could improve the performance by 0.3 or 0.4 BLEU points compared to the system where only the LM is adapted. If we compare both sequence models they perform quite similar.

Table 2: Different translation models for domain factors (BLEU)

	System	Dev	Test
1	Baseline	25.90	29.03
2	(1) + LM Adaptation	26.68	29.24
3	(2) + Word Count Model	26.13	29.17
4	(3) + Domain Frequency	27.03	29.63
5	(3) + Target Frequency	27.00	29.51
6	(3) + Source Frequency	26.95	29.84
7	(3) + All	27.07	29.69

After taking a look at the sequence model, we evaluated the different approaches to model the generation of the corpus id tags. The results using different translation model scores for the domain factors are displayed in Table 2. In this experiments we always used the word count model to judge the corpus id sequence.

The first experiment, using only a domain sequence model and no domain translation model did not lead to any improvement in the translation quality. Then we used the scores introduced in Equations 3, 4 and 5 separately. This means, that system 4 is equal to system 4 in Table 1. On the development set, this did lead to quite similar results, while the results on the test set vary a little

Figure 2: Example translations with and without domain adaptation

Input:	Ein blauer Bogen (demokatischer) Staaten im Osten, ...
Reference:	An arc of blue (Democratic) states in the East, ...
Baseline:	A blue sheet (democratic) countries in Eastern Europe, ...
Adapted:	A blue arc (democratic) states in the east, ...

bit more. But all models lead to an improvement of translation quality compared to the baseline system.

In a last experiment we use all three domain translation model scores. With the resulting system, we get the best performance on the development data, but a slightly worse performance on the test set than by using the source frequency only.

6.2 Results for Lecture Task

Table 3: Evaluation of the speech translation system (BLEU)

	System	Dev	Test
1	LM Adaptation	36.93	29.84
2	(1) + Source Frequency	37.90	31.12
3	(1) + Target Frequency	37.63	30.73
4	(1) + Domain Frequency	37.28	30.16
5	(1) + All	37.74	31.53
6	(2) + All Sep. corpus ids	38.01	31.51

A second group of experiments was performed on the lecture translation task. The results for these experiments are displayed in Table 3.

We use a baseline system that is already adapted to the target domain by an in-domain target language model. Then we add the word count domain sequence model and the different domain translations model scores separately. In the next configuration we combined the different domain translation models. The system using the Source Frequency preformed best on the development set and led to an improvement of 1.3 BLEU points on the test set compared to the baseline system.

In a last system, we extended the best preforming system to not only use two corpus id tags for the in-domain and out-of-domain part of the corpus, but we use a separate one for every part of the corpus. This led to four corpus id tags (*EPPS*, *NEWS*, *BTEC*, *LECTURE*). Then we use a different word count feature for each of this tags. If we look at the weights assigned to the different features after the optimization, we see that the system prefers phrases extracted from the lecture domain most. Then, translations from the BTEC domain are preferred over phrases from the EPPS and NEWS domain. These additional features could improve the performance even more by 0.4 BLEU points leading to a BLEU score of 31.51 BLEU on the test set.

6.3 Examples

Figure 2 shows different translations of the example sentence introduced in Figure 1. The first translation was generated by a baseline system that does not use the adaptation technique and the second translation by a system using the technique.

The translations of both systems differ in three words. Applying the adaptation model could improve the lexical selection in these cases. One example is the German word *Osten* (Engl. *East*). In the big out-of-domain corpus containing the Proceedings of the European Parliament, quite often this word is used as abbreviation for *Eastern Europe*. But in the example sentence, this is a wrong translation. In this case, we get a better translation, if we also use the information, how words are translated differently in the in-domain corpus.

Furthermore, the example shows the importance of combining the better matching in-domain knowledge with the broader knowledge of the whole corpus. Since there is no translation for *blauer* in the in-domain corpus, we use the out-of-domain phrase pair. For other phrases for which in-domain and out-of-domain phrase pairs are available, we prefer the better matching in-domain phrase pairs.

7 Conclusion

We presented a new approach to adapt a phrase-based translation system using factored translation models. Consequently, this approach is easy to integrate into state-of-the-art phrase-based translations systems. Instead of incorporating linguistic knowledge with the framework, we used it to integrate domain knowledge by introducing the corpus id as additional factor.

Using this approach we can prefer phrase pairs from a specific domain. Therefore, we introduced a model to estimate the probability of the phrase pair belonging to a certain domain and a model to judge the generated sequence of corpus id tags. The weights of the different models can be optimized using MER training leading to the best translation quality on the development data.

We could show an improvement by up to 1 BLEU point on two different tasks when translating from German to English.

In the future, we will investigate more complex domain sequence models, to judge better where to use in-domain and out-of-domain phrase pairs. Furthermore, we will try to automatically split the training corpus into segments according to different topics to obtain more fine-grained corpus ids.

Acknowledgement

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

Bertoldi, Nicola and Marcello Federico. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Fourth Workshop on Statistical Machine Translation*, Athens, Greece.

Birch, Alexandra, Miles Osborne, and Philipp Koehn. 2007. CCG Supertags in Factored Satistical Machine Translation. In *Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Bulyko, Ivan, Spyros Matsoukas, Richard Schwartz, Long Nguyen, and John Makhoul. 2007. Language Model Adaptation in Machine Translation from Speech. In *ICASSP 2007*, Honolulu, USA.

Foster, George and Roland Kuhn. 2007. Mixture-Model Adaptation for SMT. In *ACL 2007*, Prague, Czech Republic.

Hildebrand, Silja, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adapatation of the Translation Model for Statistical Machine Translation based on Information Retrieval. In *EAMT 2005*, Budapest, Hungary.

Koehn, Philipp and Hieu Hoang. 2007. Factored Translation Models. In *EMNLP-CoNLL*, Prague, Czech Republic.

Koehn, Philipp and Josh Schroeder. 2007. Experiments in Domain Adaptation for Statistical Machine Translation. In *Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *HLT/NAACL 2003*.

Koehn, P., H. Hoang, A. Birch, Ch. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, Ch. Moran, R. Zens, Ch. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Demonstration Session*, Prague, Czech Republic.

Matsoukas, Spyros, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative Corpus Weight Estimation for Machine Translation. In *Conference on Empirical Methods on Natural Language Processing (EMNLP 2009)*, Singapore.

Niehues, Jan and Muntsin Kolss. 2009. A POS-Based Model for Long-Range Reorderings in SMT. In *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece.

Niehues, Jan and Stephan Vogel. 2008. Discriminative Word Alignment via Alignment Matrix Modeling. In *Third Workshop on Statistical Machine Translation*, Columbus, USA.

Rottmann, Kay and Stephan Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *TMI*, Skövde, Sweden.

Schwenk, Holger and Jean Senellart. 2009. Translation Model Adapation for an Arabic/French News Translation System by Lightly-Supervised Training. In *MT Summit XII*, Ottawa, Canada.

Snover, Matthew, Bonnie Dorr, and Richard Schwartz. 2008. Language and Translation Model Adaptation using Comparable Corpora. In *Conference on Empirical Methods on Natural Language Processing (EMNLP 2008)*, Honolulu, USA.

Ueffing, Nicola, Gholamerza Haffari, and Anoop Sarkar. 2007. Semi-Supervised Model Adaptation for Statistical Machine Translation. *Machine Translation*, 21(2):77–94.

Venugopal, Ashish, Andreas Zollman, and Alex Waibel. 2005. Training and Evaluation Error Minimization Rules for Statistical Machine Translation. In *Workshop on Data-drive Machine Translation and Beyond (WPT-05)*, Ann Arbor, MI.

Vogel, Stephan. 2003. SMT Decoder Dissected: Word Reordering. In *Int. Conf. on Natural Language Processing and Knowledge Engineering*, Beijing, China.

Wu, Hua, Haifend Wang, and Chengqing Zong. 2008. Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Corpora. In *Coling 2008*, Manchester, UK.