Improving Statistical Machine Translation with Word Class Models

Joern Wuebker, Stephan Peitz, Felix Rietig and Hermann Ney
Human Language Technology and Pattern Recognition Group
RWTH Aachen University
Aachen, Germany
<surname>@cs.rwth-aachen.de

Abstract

Automatically clustering words from a monolingual or bilingual training corpus into classes is a widely used technique in statistical natural language processing. We present a very simple and easy to implement method for using these word classes to improve translation quality. It can be applied across different machine translation paradigms and with arbitrary types of models. We show its efficacy on a small German→English and a larger French→German translation task with both standard phrase-based and hierarchical phrase-based translation systems for a common set of models. Our results show that with word class models, the baseline can be improved by up to 1.4% BLEU and 1.0% TER on the French→German task and 0.3% BLEU and 1.1% TER on the German→English task.

1 Introduction

Data sparsity is one of the major problems for statistical learning methods in natural language processing (NLP) today. Even with the huge training data sets available in some tasks, for many phenomena that need to be modeled only few training instances can be observed. This is partly due to the large vocabularies of natural languages. One possiblity to reduce the sparsity for model estimation is to reduce the vocabulary size. By clustering the vocabulary into a fixed number of word classes, it is possible to train models that are less prone to sparsity issues. This work investigates the performance of standard models used in statistical machine translation when they are trained on automatically learned word classes rather than the actual word identities.

In the popular tooklit GIZA++ (Och and Ney, 2003), word classes are an essential ingredient to model alignment probabilities with the HMM or IBM translation models. It contains the `mkcls` tool (Och, 1999), which can automatically cluster the vocabulary into classes.

Using this tool, we propose to re-parameterize the standard models used in statistical machine translation (SMT), which are usually conditioned on word identities rather than word classes. The idea is that this should lead to a smoother distribution, which is more reliable due to less sparsity. Here, we focus on the phrase-based and lexical channel models in both directions, simple count models identifying frequency thresholds, lexicalized reordering models and an n-gram language model. Although our results show that it is not a good idea to replace the original models, we argue that adding them to the log-linear feature combination can improve translation quality. They can easily be computed for different translation paradigms and arbitrary models. Training and decoding is possible without or with only little change to the code base.

Our experiments are conducted on a medium-sized French→German task and a small German→English task and with both phrase-based and hierarchical phrase-based translation decoders. By using word class models, we can improve our respective baselines by 1.4% BLEU and 1.0% TER on the French→German task and 0.3% BLEU and 1.1% TER on the German→English task.

Training an additional language model for trans-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1377–1381,
Seattle, Washington, USA, 18-21 October 2013.

lation based on word classes has been proposed in (Wuebker et al., 2012; Mediani et al., 2012; Koehn and Hoang, 2007). In addition to the reduced sparsity, an advantage of the smaller vocabulary is that longer n-gram context can be modeled efficiently.

Mathematically, our idea is equivalent to a special case of the Factored Translation Models proposed by Koehn and Hoang (2007). We will go into more detail in Section 4. Also related to our work, Cherry (2013) proposes to parameterize a hierarchical reordering model with sparse features that are conditioned on word classes trained with `mkcls`. However, the features are trained with MIRA rather than estimated by relative frequencies.

2 Word Class Models

2.1 Standard Models

The translation model of most phrase-based and hierarchical phrase-based SMT systems is parameterized by two phrasal and two lexical channel models (Koehn et al., 2003) which are estimated as relative frequencies. Their counts are extracted heuristically from a word aligned bilingual training corpus.

In addition to the four channel models, our baseline contains binary count features that fire, if the extraction count of the corresponding phrase pair is greater or equal to a given threshold τ. We use the thresholds $\tau = \{2, 3, 4\}$.

Our phrase-based baseline contains the hierarchical reordering model (HRM) described by Galley and Manning (2008). Similar to (Cherry et al., 2012), we apply it in both translation directions with separate scaling factors for the three orientation classes, leading to a total of six feature weights.

An n-gram language model (LM) is another important feature of our translation systems. The baselines apply 4-gram LMs trained by the SRILM toolkit (Stolcke, 2002) with interpolated modified Kneser-Ney smoothing (Chen and Goodman, 1998). The smaller vocabulary size allows us to efficiently model larger context, so in addition to the 4-gram LM, we also train a 7-gram LM based on word classes. In contrast to an LM of the same size trained on word identities, the increase in computational resources needed for translation is negligible for the 7-gram word class LM (wcLM).

2.2 Training

By replacing the words on both source and target side of the training data with their respective word classes and keeping the word alignment unchanged, all of the above models can easily be trained conditioned on word classes by using the same training procedure as usual. We end up with two separate model files, usually in the form of large tables, one with word identities and one with classes. Next, we sort both tables by their word classes. By walking through both sorted tables simultaneously, we can then efficiently augment the standard model file with an additonal feature (or additional features) based on word classes. The word class LM is directly passed on to the decoder.

2.3 Decoding

The decoder searches for the best translation given a set of models $h_m(e_1^I, s_1^K, f_1^J)$ by maximizing the log-linear feature score (Och and Ney, 2004):

$$\hat{e}_1^{\hat{I}} = \arg\max_{I, e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, s_1^K, f_1^J) \right\}, \quad (1)$$

where $f_1^J = f_1 \dots f_J$ is the source sentence, $e_1^I = e_1 \dots e_I$ the target sentence and $s_1^K = s_1 \dots s_K$ the hidden alignment or derivation.

All the above mentioned models can easily be integrated into this framework as additional features h_m. The feature weights λ_m are tuned with minimum error rate training (MERT) (Och, 2003).

3 Experiments

3.1 Data

Our experiments are performed on a French→German task. In addition to some project-internal data, we train the system on the data provided for the WMT 2012 shared task[1]. Both the `dev` and the `test` set are composed of a mixture of broadcast news and broadcast conversations crawled from the web and have two references. Table 1 shows the data statistics.

To confirm our results we also run experiments on the German→English task of the IWSLT 2012 evaluation campaign[2].

[1] `http://www.statmt.org/wmt12/`
[2] `http://hltc.cs.ust.hk/iwslt/`

		French	German
train	Sentences	1.9M	
	Running Words	57M	50M
dev	Sentences	1900	
	Running Words	61K	55K
test	Sentences	2037	
	Running Words	60K	54K

Table 1: Corpus statistics for the French→German task. The running word counts for the German side of dev and test are averaged over both references.

3.2 Setup

In the French→German task, our baseline is a standard phrase-based system augmented with the hierarchical reordering model (HRM) described in Section 2.1. The language model is a 4-gram LM trained on all German monolingual sources provided for WMT 2012. For the class-based models, we run mkcls on the source and target side of the bilingual training data to cluster the vocabulary into 100 classes each. This clustering is used to train the models described above for word classes on the same training data as their counterparts based on word identity. This also holds for the wcLM, which is a 4-gram LM trained on the same data as the baseline LM. Further, the smaller vocabulary allows us to build an additional wcLM with a 7-gram context length. On this task we also run additional experiments with 200 and 500 classes.

On the German→English task, we evaluate our method for both a standard phrase-based and the hierarchical phrase-based baseline. Again, the phrase-based baseline contains the HRM model. As bilingual training data we use the *TED* talks, which we cluster into 100 classes on both source and target side. The 4-gram LM is trained on the *TED*, *Europarl* and *news-commentary* corpora. On this data set, we directly use a 7-gram wcLM.

In all setups, the feature weights are optimized with MERT. Results are reported in BLEU (Papineni et al., 2002) and TER (Snover et al., 2006), confidence level computation is based on (Koehn, 2004). Our experiments are conducted with the open source toolkit *Jane* (Wuebker et al., 2012; Vilar et al., 2010).

	dev		test	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
-TM +wcTM	21.2	64.2	24.7	59.5
-LM +wcLM	22.2	62.9	25.9	58.9
-HRM +wcHRM	24.6	61.9	27.5	58.1
phrase-based	24.6	61.8	27.8	57.6
+ wcTM	24.7	61.4	28.1	57.1
+ wcLM	24.9	61.2	28.4	57.1
+ wcHRM	25.4‡	60.9‡	28.9‡	56.9‡
+ wcLM^7	25.5‡	60.7‡	29.2‡	56.6‡
+ wcModels$_{200}$	25.5‡	60.8‡	29.3‡	56.4‡
+ wcModels$_{500}$	25.2†	60.8‡	29.0‡	56.6‡

Table 2: BLEU and TER results on the French→German task. Results marked with ‡ are statistically significant with 95% confidence, results marked with † with 90% confidence. *-X +wcX* denote the systems, where the model *X* in the baseline is replaced by its word class counterpart. The 7-gram word class LM is denoted as *wcLM*^{7}. *wcModels*$_X$ denotes all word class models trained on *X* classes.

3.3 Results

Results for the French→German task are given in Table 2. In a first set of experiments we replaced one of the standard TM, LM and HRM models by the same model based on word classes. Unsurprisingly, this degrades performance with different levels of severity. The strongest degradation can be seen when replacing the TM, while replacing the HRM only leads to a small drop in performance. However, when the word class models are added as additional features to the baseline, we observe improvements. The wcTM yields 0.3% BLEU and 0.5% TER on test. By adding the 4-gram wcLM, we get another 0.3% BLEU and the wcHRM shows further improvements of 0.5% BLEU and 0.2% TER. Extending the context length of the wcLM to 7-grams gives an additional boost, reaching a total gain over the baseline of 1.4% BLEU and 1.0% TER. Using 200 classes instead of 100 seems to perform slightly better on test, but with 500 classes, translation quality degrades again.

On the German→English task, the results shown in Table 3 are similar in TER, but less pronounced in BLEU. Here we are able to improve over the phrase-based baseline by 0.3% BLEU and 1.1% TER

	dev		test	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
phrase-based	30.2	49.6	28.6	51.6
+ wcTM	30.2	49.2	28.9	51.3
+ wcLM^7	30.5	48.3‡	29.0	50.6†
+ wcHRM	30.8	48.3‡	28.9	50.5‡
hiero	29.6	50.3	27.9	52.5
+ wcTM	29.8	50.3	28.1	52.3
+ wcLM^7	30.0	49.8	28.2	51.7

Table 3: BLEU and TER results on the German→English task. Results marked with ‡ are statistically significant with 95% confidence, results marked with † with 90% confidence.

by adding the wcTM, the 7-gram wcLM and the wcHRM. With the hierarchical decoder we gain 0.3% BLEU and 0.8% TER by adding the wcTM and the 7-gram wcLM.

4 Equivalence to Factored Translation

Koehn and Hoang (2007) propose to integrate different levels of annotation (e.g. morphologial analysis) as *factors* into the translation process. Here, the surface form of the source word is analyzed to produce the factors, which are then translated and finally the surface form of the target word is generated from the target factors. Although the translations of the factors operate on the same phrase segmentation, they are assumed to be independent. In practice this is done by *phrase expansion*, which generates a joint phrase table as the cross product from the phrase tables of the individual factors.

In contrast, in this work each word is mapped to a single class, which means that when we have selected a translation option for the surface form, the target side on the word class level is predetermined. Thus, no phrase expansion or generation steps are necessary to incorporate the word class information. The phrase table can simply be extended with additional scores, keeping the set of phrases constant.

Although the implementation is simpler, our approach is mathematically equivalent to a special case of the factored translation framework, which is shown in Figure 1. The generation step from target word e to its target class $c(e)$ assigns all probability

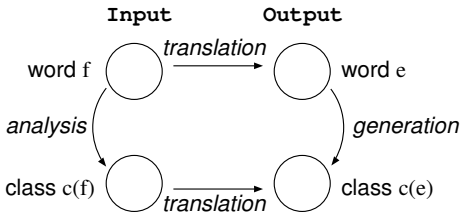

Figure 1: The factored translation model equivalent to our approach. The generation step assigns all probability mass to a single event: $p_{gen}(c(e)|e) = 1$.

mass to a single event:

$$p_{gen}(c|e) = \begin{cases} 1, \text{ if } c = c(e) \\ 0, \text{ else} \end{cases} \quad (2)$$

5 Conclusion

We have presented a simple and very easy to implement method to make use of word clusters for improving machine translation quality. It is applicable across different paradigms and for arbitrary types of models. Depending on the model type, it requires little or no change to the training and decoding software. We have shown the efficacy of this method on two translation tasks and with both the standard phrase-based and the hierarchical phrase-based translation paradigm. It was applied to relative frequency translation probabilities, the n-gram language model and a hierarchical reordering model. In our experiments, the baseline is improved by 1.4% BLEU and 1.0% TER on the French→German task and by 0.3% BLEU and 1.1% TER on the German→English task.

In future work we plan to apply our method to a wider range of languages. Intuitively, it should be most effective for morphologically rich languages, which naturally have stronger sparsity problems.

Acknowledgments

This work was partially realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. The research leading to these results has also received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

References

Stanley F. Chen and Joshuo Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, August.

Colin Cherry, Robert C. Moore, and Chris Quirk. 2012. On Hierarchical Re-ordering and Permutation Parsing for Phrase-based Decoding. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, WMT '12, pages 200–209, Montral, Canada.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 22–31, Atlanta, Georgia, USA, June.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 847–855, Honolulu, Hawaii, USA, October.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic, June.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-03)*, pages 127–133, Edmonton, Alberta.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain, July.

Mohammed Mediani, Yuqi Zhang, Thanh-Le Ha, Jan Niehues, Eunah Cho, Teresa Herrmann, and Alex Waibel. 2012. The kit translation systems for iwslt 2012. In *Proceedings of the International Workshop for Spoken Language Translation (IWSLT 2012)*, Hong Kong.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, December.

F. J. Och. 1999. An efficient method for determining bilingual word classes. In *Proc. of the Ninth Conf. of the Europ. Chapter of the Association of Computational Linguistics*, pages 71–76, Bergen, Norway, June.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, September.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden, July.

Joern Wuebker, Matthias Huck, Stephan Peitz, Malte Nuhn, Markus Freitag, Jan-Thorsten Peter, Saab Mansour, and Hermann Ney. 2012. Jane 2: Open source phrase-based and hierarchical statistical machine translation. In *International Conference on Computational Linguistics*, pages 483–491, Mumbai, India, December.