# The DCU-ICTCAS MT system at WMT 2014 on German-English Translation Task

**Liangyou Li***, **Xiaofeng Wu***, **Santiago Cortés Vaíllo***
**Jun Xie**†, **Andy Way***, **Qun Liu***†
* CNGL Centre for Global Intelligent Content, School of Computing
Dublin City University, Dublin 9, Ireland
† Key Laboratory of Intelligent Information Processing, Institute of Computing Technology
Chinese Academy of Sciences, Beijing, China
{liangyouli,xiaofengwu,scortes,away,qliu}@computing.dcu.ie
xiejun@ict.ac.cn

## Abstract

This paper describes the DCU submission to WMT 2014 on German-English translation task. Our system uses phrase-based translation model with several popular techniques, including Lexicalized Reordering Model, Operation Sequence Model and Language Model interpolation. Our final submission is the result of system combination on several systems which have different pre-processing and alignments.

## 1 Introduction

On the German-English translation task of WMT 2014, we submitted a system which is built with Moses phrase-based model (Koehn et al., 2007).

For system training, we use all provided German-English parallel data, and conducted several pre-processing steps to clean the data. In addition, in order to improve the translation quality, we adopted some popular techniques, including three Lexicalized Reordering Models (Axelrod et al., 2005; Galley and Manning, 2008), a 9-gram Operation Sequence Model (Durrani et al., 2011) and Language Model interpolation on several datasets. And then we use system combination on several systems with different settings to produce the final outputs.

Our phrase-based systems are tuned with k-best MIRA (Cherry and Foster, 2012) on development set. We set the maximum iteration to be 25.

The Language Models in our systems are trained with SRILM (Stolcke, 2002). We trained a 5-gram model with Kneser-Ney discounting (Chen and Goodman, 1996).

| Corpus | Filtered Out (%) |
|---|---|
| Bilingual | 7.17 |
| Monolingual (English) | 1.05 |

Table 1: Results of language detection: percentage of filtered out sentences

In the next sections, we will describe our system in detail. In section 2, we will explain our pre-processing steps on corpus. Then in section 3, we will describe some techniques we have tried for this task and the experiment results. In section 4, our final configuration for submitted system will be presented. And we conclude in the last section.

## 2 Pre-processing

We use all the training data for German-English translation, including Europarl, News Commentary and Common Crawl. The first thing we noticed is that some Non-German and Non-English sentences are included in our training data. So we apply Language Detection (Shuyo, 2010) for both monolingual and bilingual corpora. For monolingual data (only including English sentences in our task), we filter out sentences which are detected as other language with probability more than 0.999995. And for bilingual data, A sentence pair is filtered out if the language detector detects a different language with probability more than 0.999995 on either the source or the target. The filtering results are given in Table 1.

In our experiment, German compound words are splitted based on frequency (Koehn and



*Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 136–141,
Baltimore, Maryland USA, June 26–27, 2014. 

Knight, 2003). In addition, for both monolingual and bilingual data, we apply tokenization, normalizing punctuation and truecasing using Moses scripts. For parallel training data, we also filter out sentence pairs containing more than 80 tokens on either side and sentence pairs whose length ratio between source and target side is larger than 3.

## 3 Techniques

In our preliminary experiments, we take newstest 2013 as our test data and newstest 2008-2012 as our development data. In total, we have more than 10,000 sentences for tuning. The tuning step would be very time-consuming if we use them all. So in this section, we use Feature Decay Algorithm (FDA) (Biçici and Yuret, 2014) to select 2000 sentences as our development set. Table 2 shows that system performance does not increase with larger tuning set and the system using only 2K sentences selected by FDA is better than the baseline tuned with all the development data.

In this section, alignment model is trained by MGIZA++ (Gao and Vogel, 2008) with `grow-diag-final-and` heuristic function. And other settings are mostly default values in Moses.

### 3.1 Lexicalized Reordering Model

German and English have different word order which brings a challenge in German-English machine translation. In our system, we adopt three Lexicalized Reordering Models (LRMs) for addressing this problem. They are word-based LRM (wLRM), phrase-based LRM (pLRM) and hierarchal LRM (hLRM).

These three models have different effect on the translation. Word-based and phrase-based LRMs are focus on local reordering phenomenon, while hierarchical LRM could be applied into longer reordering problem. Figure 1 shows the differences (Galley and Manning, 2008). And Table 3 shows effectiveness of different LRMs.

In our system based on Moses, we use `wbe-msd-bidirectional-fe`, `phrase-msd-bidirectional-fe` and `hier-mslr-bidirectional-fe` to specify these three LRMs. From Table 2, we could see that LRMs significantly improves the translation.

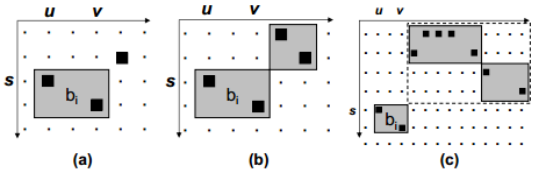


Figure 1: Occurrence of a swap according to the three orientation models: word-based, phrase-based, and hierarchical. Black squares represent word alignments, and gray squares represent blocks identified by phrase-extract. In (a), block $b_i = (e_i, f_{a_i})$ is recognized as a swap according to all three models. In (b), $b_i$ is not recognized as a swap by the word-based model. In (c), $b_i$ is recognized as a swap only by the hierarchical model. (Galley and Manning, 2008)

### 3.2 Operation Sequence Model

The Operation Sequence Model (OSM) (Durrani et al., 2011) explains the translation procedure as a linear sequence of operations which generates source and target sentences in parallel. Durrani et al. (2011) defined four translation operations: Generate(X,Y), Continue Source Concept, Generate Source Only (X) and Generate Identical, as well as three reordering operations: Insert Gap, Jump Back(W) and Jump Forward. These operations are described as follows.

- *Generate(X,Y)* make the words in Y and the first word in X added to target and source string respectively.
- *Continue Source Concept* adds the word in the queue from Generate(X,Y) to the source string.
- *Generate Source Only (X)* puts X in the source string at the current position.
- *Generate Identical* generates the same word for both sides.
- *Insert Gap* inserts a gap in the source side for future use.
- *Jump Back (W)* makes the position for translation be the Wth closest gap to the current position.
- *Jump Forward* moves the position to the index after the right-most source word.

| Systems | Tuning Set | newstest 2013 |
|---|---|---|
| Baseline | – | 24.1 |
| +FDA | – | 24.2 |
| +LRMs | 24.0 | 25.4 |
| +OSM | 24.4 | 26.2 |
| +LM Interpolation | 24.6 | 26.4 |
| +Factored Model | – | 25.9 |
| +Sparse Feature | 25.6 | 25.9 |
| +TM Combination | 24.1 | 25.4 |
| +OSM Interpolation | 24.4 | 26.0 |

Table 2: Preliminary results on tuning set and test set (newstest 2013). All scores on test set are case-sensitive BLEU[%] scores. And scores on tuning set are case-insensitive BLEU[%] directly from tuning result. Baseline uses all the data from newstest 2008-2012 for tuning.

| Systems | Tuning Set (uncased) | newstest 2013 |
|---|---|---|
| Baseline+FDA | – | 24.2 |
| +wLRM | 23.8 | 25.1 |
| +pLRM | 23.9 | 25.2 |
| +hLRM | 24.0 | 25.4 |
| +pLRM | 23.8 | 25.1 |
| +hLRM | 23.7 | 25.2 |

Table 3: System BLEU[%] scores when different LRMs are adopted.

The probability of an operation sequence $O = (o_1 o_2 \cdots o_J)$ is:

$$p(O) = \prod_{j=1}^{J} p(o_j | o_{j-n+1} \cdots o_{j-1}) \qquad (1)$$

where $n$ indicates the number of previous operations used.

In this paper we train a 9-gram OSM on training data and integrate this model directly into log-linear framework (OSM is now available to use in Moses). Our experiment shows OSM improves our system by about 0.8 BLEU (see Table 2).

### 3.3 Language Model Interpolation

In our baseline, Language Model (LM) is trained on all the monolingual data provided. In this section, we try to build a large language model by including data from English Gigaword fifth edition (only taking partial data with size of 1.6G), English side of UN corpus and English side of $10^9$ French-English corpus. Instead of training a single model on all data, we interpolate language models trained on each subset (monolingual data provided is splitted into three parts: News 2007-2013, Europarl and News Commentary) by tuning weights to minimize perplexity of language model measured on the target side of development set.

In our experiment, after interpolation, the language model doesn't get a much lower perplexity, but it slightly improves the system, as shown in Table 2.

### 3.4 Other Tries

In addition to the techniques mentioned above, we also try some other approaches. Unfortunately all of these methods described in this section are non-effective in our experiments. The results are shown in Table 2.

- *Factored Model* (Koehn and Hoang, 2007): We tried to integrate a target POS factored model into our system with a 9-gram POS language model to address the problem of word selection and word order. But experiment doesn't show improvement. The English POS is from Stanford POS Tagger (Toutanova et al., 2003).

- *Translation Model Combination*: In this experiment, we try to use the method of (Sennrich, 2012) to combine phrase tables or reordering tables from different subsets of data

to minimize perplexity measured on development set. We try to split the training data in two ways. One is according to data source, resulting in three subsets: Europarl, News Commentary and Common Crawl. Another one is to use data selection. We use FDA to select 200K sentence pairs as in-domain data and the rest as out-domain data. Unfortunately both experiments failed. In Table 2, we only report results of phrase table combination on FDA-based data sets.

- *OSM Interpolation*: Since OSM in our system could be taken as a special language model, we try to use the idea of interpolation similar with language model to make OSM adapted to some data. Training data are splitted into two subsets with FDA. We train 9-gram OSM on each subsets and interpolate them according to OSM trained on the development set.

- *Sparse Features*: For each source phrase, there is usually more than one corresponding translation option. Each different translation may be optimal in different contexts. Thus in our systems, similar to (He et al., 2008) which proposed a Maximum Entropy-based rule selection for the hierarchical phrase-based model, features which describe the context of phrases, are designed to select the right translation. But different with (He et al., 2008), we use sparse features to model the context. And instead of using syntactic POS, we adopt independent POS-like features: cluster ID of word. In our experiment *mkcls* was used to cluster words into 50 groups. And all features are generalized to cluster ID.

## 4 Submission

Based on our preliminary experiments in the section above, we use LRMs, OSM and LM interpolation in our final system for newstest 2014. But as we find that Language Models trained on UN corpus and $10^9$ French-English corpus have a very high perplexity and in order to speed up the translation by reducing the model size, in this section, we interpolate only three language models from monolingual data provided, English Gigaword fifth edition and target side of training data. In addition, we also try some different methods for final submission. And the results are shown in Table 4.

- *Development Set Selection*: Instead of using FDA which is dependent on test set, we use the method of (Nadejde et al., 2013) to select tuning set from newstest 2008-2013 for the final system. We only keep 2K sentences which have more than 30 words and higher BLEU score. The experiment result is shown in Table 4 ( The system is indicated as Baseline).

- *Pre-processing*: In our preliminary experiments, sentences are tokenized without changing hyphen. Thus we build another system where all the hyphens are tokenized aggressively.

- *SyMGIZA++*: Better alignment could lead to better translation. So we carry out some experiments on SyMGIZA++ aligner (Junczys-Dowmunt and Sza, 2012), which modifies the original IBM/GIZA++ word alignment models to allow to update the symmetrized models between chosen iterations of the original training algorithms. Experiment shows this new alignment improves translation quality.

- *Multi-alignment Selection*: We also try to use multi-alignment selection (Tu et al., 2012) to generate a "better" alignment from three alignmens: MGIZA++ with function *grow-diag-final-and*, SyMGIZA++ with function *grow-diag-final-and* and fast alignment (Dyer et al., 2013). Although this method show comparable or better result on development set, it fails on test set.

Since we build a few systems with different setting on Moses phrase-based model, a straightforward thinking is to obtain the better translation from several different translation systems. So we use system combination (Heafield and Lavie, 2010) on the 1-best outputs of three systems (indicated with $^*$ in table 4). And this results in our best system so far, as shown in Table 4. In our final submission, this result is taken as primary.

## 5 Conclusion

This paper describes our submitted system to WMT 2014 in detail. This system is based on

| Systems | Tuning Set | newstest 2014 |
|---|---|---|
| Baseline* | 34.2 | 25.6 |
| +SyMGIZA++* | 34.3 | 26.0 |
| +Multi-Alignment Selection | 34.4 | 25.6 |
| +Hyphen-Splitted | 33.9 | 25.9 |
| +SyMGIZA++* | 34.0 | 26.0 |
| +Multi-Alignment Selection | 34.0 | 25.7 |
| System Combination | – | 26.5 |

Table 4: Experiment results on newstest 2014. We report case-sensitive BLEU[%] score on test set and case-insensitive BLEU[%] on tuning set which is directly from tuning result. Baseline is the phrase-based system with LRMs, OSM and LM interpolation on smaller datasets, tuned with selected development set. Systems indicated with * are used for system combination.

Moses phrase-based model, and integrates Lexicalized Reordering Models, Operation Sequence Model and Language Model interpolation. Also system combination is used on several systems which have different pre-processing and alignment.

## Acknowledgments

This work is supported by EC Marie-Curie initial training Network EXPERT (EXPloiting Empirical appRoaches to Translation) project (`http://expert-itn.eu`). Thanks to Johannes Leveling for his help on German compound splitting. And thanks to Jia Xu and Jian Zhang for their advice and help on this paper and experiments.

## References

Amittai Axelrod, Ra Birch Mayne, Chris Callison-burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Ergun Biçici and Deniz Yuret. 2014. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions On Audio, Speech, and Language Processing (TASLP)*.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 427–436, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1045–1054, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of NAACL*.

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 848–856, Honolulu, Hawaii, October. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 321–328, Manchester, UK, August. Coling 2008 Organizing Committee.

Kenneth Heafield and Alon Lavie. 2010. Combining machine translation output with open source: The Carnegie Mellon multi-engine machine translation scheme. *The Prague Bulletin of Mathematical Linguistics*, 93:27–36, January.

Marcin Junczys-Dowmunt and Arkadiusz Sza. 2012. Symgiza++: Symmetrized word alignment models for statistical machine translation. In Pascal Bouvry, MieczysawA. Kopotek, Franck Leprvost, Magorzata Marciniak, Agnieszka Mykowiecka, and Henryk

Rybiski, editors, *Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, pages 379–390. Springer Berlin Heidelberg.

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 (EMNLP-CoNLL)*, pages 868–876.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 187–193, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Maria Nadejde, Philip Williams, and Philipp Koehn. 2013. Edinburgh's syntax-based machine translation systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 170–176, Sofia, Bulgaria, August. Association for Computational Linguistics.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France, April. Association for Computational Linguistics.

Nakatani Shuyo. 2010. Language detection library for java.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proceedings of the International Conference Spoken Language Processing*, pages 901–904, Denver, CO.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Yifan He, Josef van Genabith, Qun Liu, and Shouxun Lin. 2012. Combining multiple alignments to improve machine translation. In *COLING (Posters)*, pages 1249–1260.