The CNGL Translation System for CWMT 2015

Jian Zhang
ADAPT Centre
School of Computing
Dublin City University
zhangj@computing.dcu.ie

Qun Liu
ADAPT Centre
School of Computing
Dublin City University
qliu@computing.dcu.ie

Abstract

This report describes the CNGL (Centre for Global Intelligent Content) participation on CWMT 2015 machine translation evaluation task. We report experimental details of the system putting special emphasis on: training data selection approach, spares features and n-best list re-ranking. Our system is a hierarchical phrase-based translation system [Chiang, 2007]. Our submission improves the baseline system by +2.79 absolute BLEU score on the CWMT 2013 development data in the Chinese-to-English translation direction. We also present our experiment results on the Double Blind Evaluation task in this report.

1 Introduction

This report describes the submission of CNGL (Centre for Global Intelligent Content) to the CWMT 2015 machine translation evaluation task. We present our Chinese-to-English and Double Blind Evaluation translation systems. Our baseline system is a standard hierarchical phrase-based SMT system built with Moses toolkit [Koehn et al., 2007]. In the experiments, we combine different data selection approaches to select training instances that are close to the target domain. We then make the use of the source side of the testing data, and introduce thousands of binary type stateless sparse features into hierarchical phrase-based phrase tables. We also re-rank n-best list by interpolating perplexity scores from neural network language models and n-gram language models.

The rest of the report is organized as follows. Section 2 presents our training framework and parameters in details. Section 3 shows the training, development and testing data used in our system, and pre-/post-processing scripts. Section 4 and 5 detail our experiments and results on the Chinese-to-English and Double Blind Evaluation tasks, respectively. Finally, Section 6 reports our conclusions of this system report paper.

2 System

In our submission, we train the word alignment model using the GIZA++ Toolkit [Gao and Vogel, 2008]. We then symmetrize the word alignment models using the heuristic of `grow-diag-final-and`. Moses toolkit [Koehn et al., 2007] is used to extract translation table phrase pairs and the phrase pair probabilities are computed with $GoodTuring$ smoothing. All our submitted systems are hierarchical phrase-based translation systems [Chiang, 2007]. During phrase pair extraction, the maximum span number is set to 10 when non-terminals are generated, in order to reduce the size of trained translation table. Furthermore, phrase pairs with phrase translation probabilities less than a predefined threshold (0.0001) are also removed. The n-gram language model used during decoding is trained with KenLM toolkit [Heafield, 2011]. We also interpolate perplexity scores from neural network language models and n-gram language models to rerank n-best ($n = 100$). Our translation system is tuned (maximum tuning iteration is set to 10) using k-best batch Margin Infused Relaxed Algorithm (MIRA) [Cherry and Foster, 2012] and we evaluate our translation systems against BLEU [Papineni et al., 2002] score.

3 Data

Our word alignment model is trained using all the parallel training data listed at the Evaluation Guidelines (Appendix E, Table 1) [1], which consists of $7,777,485$ sentence pairs. However, it is

[1] `http://www.ai-ia.ac.cn/cwmt2015/file/Evaluation%20Guidelines.pdf`

ID	Systems	Tuning	Test
1	Baseline	29.35	29.39
2	Data Selection (MML)	28.72	28.86
3	Data Selection (FDA)	28.57	28.86
4	Data Selection (BLEU-selection)	21.99	22.22
5	Data Selection (BLEU-selection + MML + FDA)	29.75	29.79
7	Data Selection (BLEU-selection + MML + FDA) + Translation Sparse Features	30.10	29.92*
8	Data Selection (BLEU-selection + MML + FDA) + Translation Sparse Features + Binary Sparse Feature	31.63	31.73*
9	Data Selection (BLEU-selection + MML + FDA) + Translation Sparse Features + Binary Sparse Feature + Re-ranking	N/A	32.18*

Table 1: CWMT 2015 machine translation Chinese-to-English evaluation experiment results. System 7 is our primary submission. System 8 and 9 are the contrast system a and b in our submission, respectively. * indicates significantly improvement compare with baseline system (p = 0.01, 1000 iterations).

time-consuming to carry out experiments on such large size of data. We thus adapt data selection approaches (Section 4.2) and select $2,876,134$ parallel sentence pairs for phrase table training. CWMT2015 also provides additional monolingual English corpus, namely Reuters corpus and SogouCA (Appendix E, Table 5 at Evaluation Guidelines). We decide to not use them in our experiments due to time constraints. Our models are optimized using the development set released by the organizer. CWMT2013 development set is used for evaluation [2].

All Chinese characters in training sentences are converted into simplified Chinese characters using `java-zhconverter`[3]. ICTCLAS [Zhang et al., 2003] is then applied to segment Chinese sentences; we use `tokenizer.perl`[4] to tokenize English sentences. All training data are lower-cased before SMT systems are trained. The re-caser model for post-processing is trained with `train-recaser.perl`[5]. We also discard sentences pairs with length longer than 80 tokens. There are $6,915,337$ sentence pairs remaining after pre-processing.

[2]We noticed that the development set of CWMT 2013 and CWMT 2015 are identical when we are writing this report. Thus, our experiments can be thought as using the same data set for both tuning and testing. This find can also explain the tuning and testing results in Table 1 are comparable.

[3]https://github.com/opentalking/java-zhconverter-read-only

[4]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[5]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/train-recaser.perl

3.1 Baseline

Our baseline system is trained using all corpus listed at Evaluation Guidelines. The training parameters are illustrated at Section 2.

3.2 Data Selection

Data selection for SMT focuses on making efficient use of the general-domain training data in order to improve translation quality of SMT systems trained only on in-domain data. The core idea is to select some sentences which are similar to the in-domain training data from a large amount of general-domain training data. Data selection can also be viewed as a ranking challenge. [Biçici and Yuret, 2011] employs a feature decay algorithm which can be used in both active learning and transductive learning settings. In recent studies, a cross-entropy difference method has seen increasing interest for the problem of SMT data selection [Axelrod et al., 2011]. Another advantage of data selection is that the SMT training time can be significantly reduced. We compare three different data selection approaches, namely MML [Axelrod et al., 2011], FDA [Biçici and Yuret, 2011] and BLEU-selection.

- MML is a cross-entropy difference method which ranks training instances using the cross-entropy difference from language models trained on in-domain or general-domain sentences. The intuition is to find sentences as close to the target domain and as far from

	Language	Sentence Number	Token Number
Alignment Model Training	ZH	6,915,337	119,796,645
	EN	6,915,337	123,388,750
Phrase Table Training	ZH	2,876,134	56,757,259
	EN	2,876,134	58,786,460
Development	ZH	1,006	27,472
	EN	4,024	124,060
Test	ZH	1,006	27,472
	EN	4,024	124,885
Decoding Language Model (n-gram)	EN	2,182,227	43,896,042
Re-ranking Language Model (n-gram)	EN	6,915,337	123,388,750
Re-ranking Language Model (RNNLM)	EN	2,182,227	43,896,042

Table 2: Statistic summary of training, development and test data. We use the target side of phrase table training data (duplicated sentences are removed) as our decoding language model and RNNLM training data.

ID	Systems	Tuning	Test
1	Baseline	29.46	24.08
2	Baseline + Translation Sparse Features + Binary Sparse Feature	32.95	24.66
3	Baseline + Translation Sparse Features + Binary Sparse Feature + Re-ranking	N/A	24.92*

Table 3: CWMT 2015 machine translation Double Blind Evaluation experiment results. System 3 is our primary submission. * indicates significantly improvement compare with baseline system (p = 0.01, 1000 iterations).

the average of the general domain as possible. We select top 15% (1,383,067 sentence pairs) sentences from the MML ranked training data (Table 1, System ID 2).

- FDA data selection approach uses a feature decay algorithm which can be used in both active learning and transductive learning settings. The decay algorithm can increase the variety of training set by devaluing features that have already been seen from a training set. We select top 15% sentences (1,383,067 sentence pairs) from the FDA ranked training data (Table 1, System ID 3).

- BLEU-selection; We also use BLEU scores to select the top N similar sentences for each source sentence in testing data. We set N to be 100, which results 110,000 sentence pairs are selected from training data (Table 1, System ID 4).

- Concatenation: Table 1, System ID 5 is the system trained using the concatenation of MML, FDA and BLEU-selection selections.

Comparing between System ID 1 with System ID 2, 3 and 4 in Table 1, we observe that MML and FDA approaches are able to used 15% of the training corpus to produce a comparable performance with baseline system. There is a large 7.36 absolute BLEU difference between the BLEU-selection approach and baseline system. However, we think the result is reasonable since the training corpus sizes are significantly different. Furthermore, We concatenate the selections from our experiments and observe a 0.4 absolute BLEU improvement. We choose System ID 5 to be the output system from our data selection experiments. Table 2 summarizes the data statistic in our experiments.

3.3 Sparse Features

In SMT translation table, there are usually more than one corresponding translation options for each source phrase, which are optimal in contexts. The probability distributions, estimated by Maximum Likelihood, would make the translation to prefer the most common ones. In our system, we use additional sparse features to describe the context difference between phrases or rules. There are two types of sparse features defined in our sparse feature experiments,

- Translation Sparse Features; It is a lexical feature function, which consists of three features, word translation feature, target word insertion feature and source word deletion feature. The word translation feature captures if a specific source word is translated as a specific target word. The target word insertion feature indicates if a target translation has no alignment point during decoding. The source word deletion feature indicates if source word has no alignment point during decoding (Table 1, System ID 6).

- Domain Sparse Features; It is a binary feature function, which indicates if specific source or target words can be found in a phrase table entry (Table 1, System ID 7).

3.4 n-best list re-ranking

Instead of directly output 1-best translation from a decoder, we can also make the use of n-best list with the integration of additional features to re-rank translation options. In this experiment, we interpolate perplexity scores from two language models to re-rank 100-best list. The translation BLEU score after n-best list re-ranking is shown at Table 1, System ID 9. The first language model used for re-ranking is a n-gram language model. It is trained with all the English sentences from parallel corpus provided by the organizer, KenLM toolkit is used for language model training. The second language model is a recurrent neural network language model (RNNLM) [Mikolov et al., 2011], trained with the English sentences from the concatenated data selection corpus. Corpus statistic is shown at Table 2. The weights of two language model are learned from CWMT2013 development set.

4 Double Blind Evaluation

We also carry out experiments on the Double Blind Evaluation task. Since the training corpus is smaller (2,000,000 sentence pairs) than other evaluation tasks, we focus on spares features and n-best list re-ranking experiments in this evaluation. Because there is no testing data provided in this task, we use the "BLEU-selection" approach as we described at Section 3.2 to select 1,100 testing data for evacuation. The experiment results of our Double Blind Evaluation submission are presented at Table 3.

5 Conclusion

In this report, we present our submitted CWMT2015 SMT systems. We combine different data selection approaches and observe translation quality improvements in term of BLEU score. We use a hierarchical phrase-based translation model with thousands of additional lexical and domain type sparse features. Furthermore, we use two language models to re-rank the n-best list. Significant improvements over the baseline system are reported on both Chinese-to-English translation and the Double Blind Evaluation tasks.

6 Acknowledgements

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at at Dublin City University.

References

[Axelrod et al., 2011] Axelrod, A., He, X., and Gao, J. (2011). Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, UK.

[Biçici and Yuret, 2011] Biçici, E. and Yuret, D. (2011). Instance Selection for Machine Translation using Feature Decay Algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283, Edinburgh, UK.

[Cherry and Foster, 2012] Cherry, C. and Foster, G. (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*, pages 427–436. Association for Computational Linguistics.

[Chiang, 2007] Chiang, D. (2007). Hierarchical phrase-based translation. *computational linguistics*, 33(2):201–228.

[Gao and Vogel, 2008] Gao, Q. and Vogel, S. (2008). Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

[Heafield, 2011] Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

[Mikolov et al., 2011] Mikolov, T., Kombrink, S., Deoras, A., Burget, L., and Cernocky, J. H. (2011). RNNLM - Recurrent Neural Network Language Modeling Toolkit. IEEE Automatic Speech Recognition and Understanding Workshop.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

[Zhang et al., 2003] Zhang, H.-P., Yu, H.-K., Xiong, D.-Y., and Liu, Q. (2003). Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing - Volume 17*, SIGHAN '03, pages 184–187, Stroudsburg, PA, USA. Association for Computational Linguistics.