The DCU Multi-Engine MT System for CWMT'2011

Xiaofeng Wu, Junhui Li, Jie Jiang[+], Yifan He and Andy Way[+]
Centre for Next Generation Localisation, School of Computing
Dublin City University,
Dublin 9, Ireland
{xiaofengwu,jli,yhe}@computing.dcu.ie
[+]Applied Language Solutions, Manchester, UK[1]
{jie.jiang,andy.way}@appliedlanguage.com

Abstract: *In this paper we describe the multi-engine machine translation system we used in the Japanese-Chinese News subtask in the 7th Chinese Workshop on Machine Translation (CWMT'2011). We employed an MBR decoder and a Confusion Network decoder in our approach to combine different system outputs. We also describe the modular design of the system and report on how we pre-process the corpus.*
Keywords: *Machine Translation, System Combination*

第七届全国机器翻译研讨会（CWMT'2011）

DCU技术报告

吴晓锋 李军辉 江杰[+] 贺一帆 Andy Way[+]
Centre for Next Generation Localisation, School of Computing
Dublin City University,
Dublin 9, Irelan
{xiaofengwu,jli,yhe}@computing.dcu.ie
[+]Applied Language Solutions, Manchester, UK[1]
{jie.jiang,andy.way}@appliedlanguage.com

摘要： 本文介绍了都柏林城市大学（DCU）在第七届全国机器翻译研讨会（CWMT'2011）上的参赛系统。该系统由基本短语翻译系统、层次短语翻译系统、系统融合三部分组成。本系统在日中新闻子任务中取得了较好的成绩。
关键字： 机器翻译、系统融合

1 Introduction

In this paper, we describe the DCU submission to CWMT'2011. Our submission is a multi-engine machine translation (MT) system that follows the basic approach of [Du et al.,2009]. In the MT system, different statistical techniques as well as system combination approaches are exploited.

In this campaign we use two individual MT systems: the Moses phrase-based system [Koehn et al., 2007] and the hierarchical phrase-based (HPB) system of [Chiang,2005]. We also use a word-level combination strategy [Rosti et al., 2007] to combine the two translation hypotheses. To

---

[1] Work done while at CNGL, DCU.

combine these systems, we first use the Minimum Bayes-Risk (MBR) [Kumar and Byrne 2004] decoder to obtain the best hypothesis as the alignment reference for the Confusion Network (CN) [Mangu et al., 2000]. We then build the CN using the TER metric [Snover et al.,2006], and finally search for and generate the best translation.

The rest of this paper is organized as follows: Section 2 gives details of the components we used in the system, especially the multi-engine scheme. In Section 3 we describe the experimental set-up, including the preprocessing of the corpus, and provides our results on the development sets and test sets. We draw conclusions in Section 4.

2 MT System Description

2.1 System Outline

Unlike the MATREX system used in the WMT 2009 comptetition, our system does not use EBMT [Gough and Way, 2004] as an individual system, but only the phrase-based system and the hierarchical phrased-based system.

The aim of system combination [Rosti, 2007; Zong, 2008] is to process multiple MT system outputs and try to obtain a better translation result than any of the individual systems. It can be categorized into sentence-level combination, phrase-level combination and word-level combination. The combination scheme we used is word-level based. MBR is used for choosing backbones and CN is used for combination.

2.2 Hierarchical Phrase-based System

Our HPB translation system is a re-implementation of the hierarchical phrase translation model which is based on PSCFG [Chiang, 2005]. The PSCFG rules are recursively generated from the initial rules, as in:

$$N -> f_1 ... f_m / e_1 ... e_n$$

where N is a rule containing only non-terminals.

In addition we have:

$$M -> f_i ... f_j / e_u ... e_v$$

where $1 \leq i \leq j \leq m$ and $1 \leq u \leq v \leq n$, so now a new rule can be obtained:

$$N -> f_1^{i-1} X_k f_{j+1}^m / e_1^{u-1} X_k e_{v+1}^n$$

where X is a nonterminal.

There should be no more than two nonterminals in a rule. When extracting hierarchical rules, we set some limitations that initial rules are of no more than 7 words in length and other rules should have no more than 5 terminals and nonterminals. We also disallow rules with adjacent source-side and target-side nonterminals.

The decoder is an enhanced CYK-style chart parser that maximizes the derivation probability and spans up to 12 source words. A 5-gram language model generated by the SRI Language Modeling toolkit (SRILM) [Stolcke, 2002] is used in the cube-pruning process. The search space is pruned with a chart cell size limit of 50.

2.3 System Combination

In the system combination stage, we build an MBR-CN framework. We employ a minimum Bayes-risk decoder to select the best single system output from the merged N-best list by minimizing the loss in BLEU score [Papineni et al., 2002].

The confusion network is constructed by using the output of MBR as the backbone which determines the word order of the combination. The other hypotheses are aligned against the backbone based on the TER metric. NULL words are permitted in the alignment. Each arc in the CN represents an alternative word at that position in the sentence, and the number of votes for each word is counted when constructing the network. The features we used are as follows:

- Word posterior probability [Fiscus, 1997]
- 3 and 4-gram target language models;
- Word length penalty;
- Null word penalty.

In addition, we use MERT [Och, 2003] to tune the weights of confusion network.

3 Experimental Set-up

This section details the experimental set-up we had for the Japanese-Chinese News subtask. We also present the pre-processing and the post-processing steps we perform as well as how we build the language model.

3.1 Corpus

In this workshop we used only the data released by the organizers (Table 1). The Chinese language model is trained with the Sogou corpus.

In Table 1, *Sentences* indicates the number of sentence pairs; *Length* indicates the maximum sentence length in each corpus. In the devset, there are 4 references for Chinese as the target language.

Table 1. Corpus Statistics

Corpora	Sentences	Tokens-Ch	Tokens-Jp	Length
Training	279k	2.4m	2.9m	100
Development	500	50k	16k	100
Test	350k	---	114k	---

3.2 Language Model

The Chinese language model is trained on the Sogou corpus with SRILM. We trained a 5-gram model with Kneser-Ney discounting [Chen and Goodman, 1996].

3.3 Pre- and Post-processing

For Chinese, we first perform segmentation with Urheen2.2[2], and then perform full character to half character conversion. For Japanese we use Mecab[3] to perform segmentation. We use Giza++[4] with the grow-diag-final-and parameter to obtain the alignment.

The only post-processing we perform is to delete the spaces.

3.4 Experimental Results

The system is evaluated with respect to BLEU score. Table 2 gives the performance of our system as well as the individual systems on both development sets and the test sets.

PB is the basic phrase-based MT system; HPB is the hierarchical phrase-based MT system. We use BLEU4 as the performance criterion on the devset and HPB outperforms PB by .0182, which ia a relatively improvement of .0547. Combination in Table 2 is the output of the system described in Section 2. We observe that it performs a little better than the HPB method on the devset. From Table 2 we can see that our combination results outperform both individual systems on the devset and the test set.

Table 2. Experiment Results

System	Dev Set (BLEU4)	Test Set (BLEU5-SBP)
PB	.3145	.4306
HPB	.3327	.4689
Combination	**.3340**	**.4721**

4 Conclusions and Further Work

In this paper we briefly introduced our MT system used in CWMT'2011. We basically followed the framework DCU had successfully used in WMT2009, which combined the output results of two MT systems and generated a new N-best list after CN decoding. Then by using some global features the rescoring model generated the final translation output. The experimental results demonstrated that the combination module and rescoring module are effective in our framework.

---

[2] http://www.openpr.org.cn/index.php/NLP-Toolkit-For-Natural-Language-Processing/68-Urheen-A-Chinese/English-Lexical-Analysis-Toolkit/View-details.html

[3] http://mecab.sourceforge.net/

[4] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

Our future work is to refine our system and to investigate its effect on more tasks, and we aim to develop more powerful post-processing tools such as recasers to improve the BLEU score.

5 References

Chen, S. F. and Goodman, J. (1996). An Empirical Study of Smoothing Techniques for Language Modeling. In 34th Annual Meeting of the Association for Computational Linguistics, pp. 310–318, San Francisco, CA.

Chiang, D. (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 263–270, Ann Arbor, MI.

Du, J. Y. He, S. Penkale and A. Way. 2009. MaTrEx: the DCU MT System for WMT 2008 . In Proceedings of the Third Workshop on Statistical Machine Translation, EACL 2009, pp. 95—99, Athens, Greece.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In 1997 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 347–352, Santa Barbara, CA.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: open source toolkit for statistical machine translation. In ACL 2007: proceedings of demo and poster sessions, pp. 177—180, Prague, Czech Republic.

Kumar, S. and Byrne, W. (2004). Minimum Bayes-Risk Decoding for Statistical Machine Translation. In Proceedings of the Joint Meeting of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), pp. 169–176, Boston, MA.

Mangu, L., Brill, E. and Stolcke, A. (2000). Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. Computer Speech and Language, 14(4):373–400.

Och, F. (2003). Minimum error rate training in statistical machine translation. In 41st Annual Meeting of the Association for Computational Linguistics, pp. 160–167, Sapporo, Japan.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), pp. 311–318, Philadelphia, PA.

Rosti, A.-V. I., Xiang, B., Matsoukas, S., Schwartz, R., Ayan, N. F., and Dorr, B. J. (2007). Combining outputs from multiple machine translation systems. In Proceedings of the Joint Meeting of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007), pp. 228–235, Rochester, NY.

Snover, M., Dorr, B., Schwartz, R., Micciula, L., and Makhoul, J. (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 (AMTA 2006), pp. 223–231, Cambridge, MA.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In Proceedings of the International Conference Spoken Language Processing, pp. 901–904, Denver, CO.

Zong, C. (2008) Statistical Natural Language Processing, Tsinghua Press, Beijing, China.