MATREX: The DCU MT System for WMT 2009

Jinhua Du, Yifan He, Sergio Penkale, Andy Way

Centre for Next Generation Localisation
Dublin City University
Dublin 9, Ireland
{jdu,yhe,spenkale,away}@computing.dcu.ie

Abstract

In this paper, we describe the machine translation system in the evaluation campaign of the Fourth Workshop on Statistical Machine Translation at EACL 2009.

We describe the modular design of our multi-engine MT system with particular focus on the components used in this participation.

We participated in the translation task for the following translation directions: French–English and English–French, in which we employed our multi-engine architecture to translate. We also participated in the system combination task which was carried out by the MBR decoder and Confusion Network decoder. We report results on the provided development and test sets.

1 Introduction

In this paper, we present a multi-engine MT system developed at DCU, MATREX (Machine Translation using Examples). This system exploits EBMT, SMT and system combination techniques to build a cascaded translation framework.

We participated in both the French–English and English-French News tasks. In these two tasks, we employ three individual MT system which are 1) Baseline: phrase-based system (PB); 2) EBMT: Monolingually chunking both source and target sides of the dataset using a marker-based chunker (Gough and Way, 2004). 3) HPB: a typical hierarchical phrase-based system (Chiang, 2005). Meanwhile, we also use a word-level combination framework (Rosti et al., 2007) to combine the multiple translation hypotheses and employ a new rescoring model to generate the final result.

For the system combination task, we first use the minimum Bayes-risk (MBR) (Kumar and Byrne, 2004) decoder to select 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 and generate the translation.

The remainder of this paper is organised as follows: Section 2 details the various components of our system, in particular the multi-engine strategies used for the shared task. In Section 3, we outline the complete system setup for the shared task and provide results on the development and test sets. Section 4 is our conclusion.

2 The MATREX System

2.1 System Architecture

The MATREX system is a combination-based multi-engine architecture, which exploits aspects of both the EBMT and SMT paradigms.

This architecture includes three individual systems which are phrase-based, example-based and hierarchical phrase-based.

The combination structure is the MBR decoder and CN decoder, which is based on the word-level combination strategy.

In the final stage, we use a new rescoring module to process the N-best list generated by the combination module. See Figure 1 as a detailed illustration.

2.2 Example-Based Machine Translation

EBMT obtains resources using the Marker Hypothesis (Green, 1979), a psycholinguistic constraint which posits that all languages are marked for surface syntax by a specific closed set of lexemes or morphemes which signify context. Given a set of closed-class words we segment each sentence into chunks, creating a chunk at each new occurrence of a marker word, with the restriction that each segment must contain at least one non-marker word (Gough and Way, 2004).

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 95–99,
Athens, Greece, 30 March – 31 March 2009.

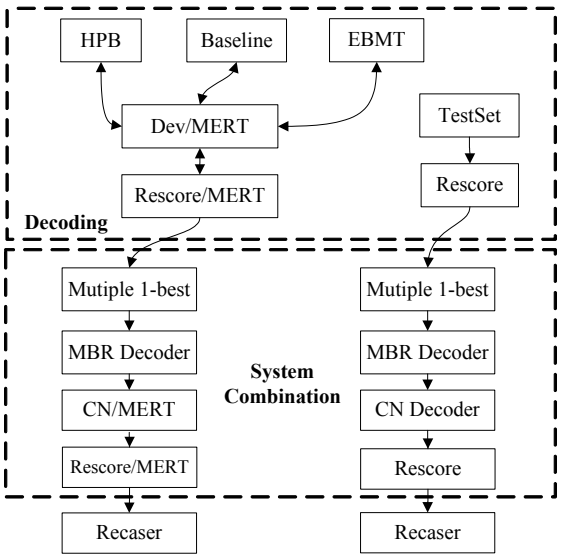

Figure 1: System Framework

We then align these segments using an edit-distance-style algorithm, in which the insertion and deletion probabilities depend on word-to-word translation probabilities and word-to-word cognates (Stroppa and Way, 2006).

We extracted phrases of at most 7 words on each side. We then merged these phrases with the phrases extracted by the baseline system adding word alignment information, and used this system seeded with this additional information.

2.3 Hierarchical Machine Translation

HPB translation system is a re-implementation of the hierarchical phrase translation model which is based on PSCFG (Chiang, 2005). We generate recursively PSCFG rules from the *initial rules* as

$$N \rightarrow f_1 \dots f_m / e_1 \dots e_n$$

where N is a rule which is initial or includes nonterminals.

$$M \rightarrow f_i \dots f_j / e_u \dots e_v$$

where $1 \leq i \leq j \leq m$ and $1 \leq u \leq v \leq n$, at which point a new rule can be obtained, named,

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

where k is an index for the nonterminal X. The number of nonterminals permitted in a rule is no more than two.

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, and we 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 4-gram language model generated by 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.

2.4 System Combination

For multiple system combination, we implement an MBR-CN framework as shown in Figure 1. Instead of using a single system output as the skeleton, we employ a minimum Bayes-risk decoder to select the best single system output from the merged N-best list by minimizing the BLEU (Papineni et al., 2002) loss.

The confusion network is built by 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 allowed 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, 4-gram target language model;
- word length penalty;
- Null word length penalty;

Also, we use MERT (Och, 2003) to tune the weights of confusion network.

2.5 Rescore

Rescore is a very important part in post-processing which can select a better hypothesis from the N-best list. We add some new global features in rescore model. The features we used are as follows:

- Direct and inverse IBM model;
- 3, 4-gram target language model;
- 3, 4, 5-gram POS language model (Ratnaparkhi, 1996; Schmid, 1994);

- Sentence length posterior probability (Zens and Ney, 2006);
- N-gram posterior probabilities within the N-Best list (Zens and Ney, 2006);
- Minimum Bayes Risk probability;
- Length ratio between source and target sentence;

The weights are optimized via MERT algorithm.

3 Experimental Setup

The following section describes the system and experimental setup for the French-English and English-French translation tasks.

3.1 Statistics of Data

Parallel Corpus
We used Europarl and Giga data for this evaluation. The statistics of parallel data are shown in Table 1.

Corpra	Sen	Token-En	Token-Fr	Len
Europarl	1.46M	39,240,672	42,252,067	80
Giga	2M	48,648,104	57,869,002	65

Table 1: Statistics of Parallel Data

In this table, *Sen* indicates the number of sentence pairs; *Len* denotes the maximum sentence length of each corpus. This year the translation task is only evaluated on *News Domain*. Experimental results showed that giga data is more correlated than Europarl and the BLEU score is significantly improved(See Table 4).

Monolingual Corpus
In this evaluation, we trained a small 4-gram language model using data in Table 1 and a large 4-gram language model using data in Table 2. We configured these two LMs for Baseline and EBMT systems while HPB only used the large one.

Language	Sen	Token	Source
English	9,966,838	240,849,221	E/N/NC
French	9,966,838	260,520,313	E/N/NC

Table 2: Statistics of Monolingual Data

In the above table, *E/N/NC* refers to Europarl/News/New_Commentary corpus.

3.2 Pre-Processing

We preprocessed both Europarl and Giga Release 1 corpus. For the Europarl corpus, we removed the reserved characters in GIZA++ and tokenized and lowercased the corpus with tools provided by WMT09. The Giga corpus was too large for our resource, so we performed sentence selection before cleaning, in the following steps.

- We split the Giga corpus into even segments, each segment consisting of 20 lines.
- We trained an SVM classifier on English side with positive examples from the monolingual news data and negative examples from noisy sentences (numbers, meaningless word combinations, and random segments) from the Giga corpus. We used "-ly" and "-ing" to approximate adverbs and present participles and did not use other POS-induced features, as in (Ferizis and Bailey, 2006). We added these features to remove noise: average length of sentences, frequency of capitalized characters, frequency of numerical characters and short word penalty (equals to 1 when average length of words < 4, and 0 otherwise). We used the classifier to remove 20% segments of lowest scores.
- We selected $1,600$ words having the highest mutual information scores with monolingual training data against the Giga corpus.
- We selected $100,000$ segments where these words occurred most frequently. However the sentence was dropped if the length ratio between English and French was larger than 1.5 or less than 0.67.

3.3 System Configuration

The two language models were done using the SRILM employing linear interpolation and modified K-N discounting (Chen and Goodman, 1996).

The configuration for the three systems is listed in Table 3.

System	P-Table	Length	LM	Features
Baseline-E	55.9M	7	2	15
Baseline-G	58.4M	7	2	15
EBMT	59.4M	7	2	15
HPB	122M	5	1	8

Table 3: Statistics of MT Systems

In this table, *E* indicates the Europarl corpus

which is used for all three systems, and *G* stands for the Giga corpus which is only used for the Baseline system. We can see from Table 3 that the size of the HPB phrase-table is more than 2 times as large as the other phrase tables. How to filter and process such a huge hierarchical table is a challenging problem.

We tuned our systems on the development set *devset2009-a* and *devset2009-b*, and performed the crossover experiment by these two devsets.

3.4 Experimental Results

The system output is evaluated with respect to BLEU score. In Table 4, we used *devset2009-b* to tune the various parameters in our three single systems and *devset2009-a* for testing. In terms of the Europarl data, we can see that the three systems we used achieved similar performance on the test set for both translation directions, with the Baseline-E system yielding slightly better results than the other two.

System	Fr-En	En-Fr
Baseline-E	22.24	22.68
Baseline-G	24.90	—[1]
EBMT	22.04	22.12
HPB	21.69	21.12
MBR	25.11	22.68
CN	25.24	22.76
Rescore	25.40	22.97

Table 4: Experimental Results on *Devset2009-a*

We then used the translations of the *devset2009-a* produced by each system to tune the parameters of our system combination module. From Table 4, we can see that using MBR and confusion network decoding leads to a slight improvement over the strongest single system, i.e. the baseline Phrase-Based SMT system. Rescoring the N-best lists yielded an increase of 0.5 (2.0 relative) absolute BLEU points over the baseline for French–English Translation and 0.29 (1.28 relative) absolute BLEU points for English–French Translation.

Table 5 is the results on 2009 *Test Data*. The scores with a slash in the last two rows are lowercased and cased respectively. From the table we can see that combination yielded 0.45 and 0.79 absolute BLEU points over the best single system for Fr-En and En-Fr direction respectively. However, 1.93 (7.2 relative) and 1.64 (6.58 relative) BLEU points are dropped between cased and lowercased results of both directions. Accordingly, training an effective recasing model is very important for our future work.

System	Fr-En	En-Fr
Baseline-E	25.64	24.47
Baseline-G	26.75	—
EBMT	25.67	24.43
HPB	25.20	24.19
Combination	27.20/25.14	25.26/22.28
Official-Auto[2]	26.86/24.93	23.78/22.14

Table 5: Summary of Results on *2009 Test Data*

4 Conclusion

This paper presents our machine translation system in WMT2009 shared task campaign. We developed a multi-engine framework which combined the output results of the three 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 result proved that the combination module and rescoring module are effective in our framework.

We also applied simple yet effective methods of genre and topical classification to remove noise and out-of-domain sentences in the Giga corpus, from which we built better translation models than from Europarl.

In future work, we will refine our system framework to investigate its effect on the tasks presented here, and we will develop more powerful post-processing tools such as recaser to reduce the BLEU loss.

Acknowledgments

This work is supported by Science Foundation Ireland (Grant No. 07/CE/I1142). Thanks also to the reviewers for their insightful comments and suggestions.

References

Chen, S. F. and Goodman, J. (1996). An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, CA.

Chiang, D. (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the*

[1] Not much time to do the experiments on English-French direction. EBMT and HPB just used the Europarl corpus.

[2] The official automatic result is scored on 2525 sentences out of the whole 3007 sentences in test set. The other 502 sentences are used as the development set for combination evaluation task.

43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 263–270, Ann Arbor, MI.

Ferizis, G. and Bailey, P. (2006). Towards practical genre classification of web documents. In *Proceedings of the 15th international conference on World Wide Web (WWW '06)*, pages 1013–1014, New York, USA.

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

Gough, N. and Way, A. (2004). Robust Large-Scale EBMT with Marker-Based Segmentation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, pages 95–104, Baltimore, MD.

Green, T. (1979). The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior*, 18:481–496.

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)*, pages 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 *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 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 *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, PA.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP)*, pages 133–142, 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)*, pages 228–235, Rochester, NY.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

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)*, pages 223–231, Cambridge, MA.

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

Stroppa, N. and Way, A. (2006). MaTrEx: the DCU machine translation system for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 31–36, Kyoto, Japan.

Zens, R. and Ney, H. (2006). N-gram Posterior Probabilities 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 2006)*, pages 72–77, New York, USA.