MATREX: The DCU MT System for WMT 2010

Sergio Penkale, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, Andy Way

CNGL, School of Computing
Dublin City University, Dublin 9, Ireland
{ *spenkale, rhaque, sdandapat, pbanerjee, asrivastava, jdu, ppecina, snaskar, mforcada, away* }*@computing.dcu.ie*

Abstract

This paper describes the DCU machine translation system in the evaluation campaign of the Joint Fifth Workshop on Statistical Machine Translation and Metrics in ACL-2010. We describe the modular design of our multi-engine machine translation (MT) system with particular focus on the components used in this participation. We participated in the English–Spanish and English–Czech translation tasks, 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.

1 Introduction

In this paper, we present the DCU multi-engine MT system MATREX (Machine Translation using Examples). This system exploits example-based MT, statistical MT (SMT), and system combination techniques.

We participated in the English–Spanish (`en–es`) and English–Czech (`en–cs`) translation tasks. For these two tasks, we employ several individual MT systems: 1) Baseline: phrase-based SMT (Koehn et al., 2007); 2) EBMT: Monolingually chunking both source and target sides of the dataset using a marker-based chunker (Gough and Way, 2004); 3) Factored translation model (Koehn and Hoang, 2007); 4) Source-side context-informed (SSCI) systems (Stroppa et al., 2007); 5) the `moses-chart` (a Moses implementation of the hierarchical phrase-based (HPB) approach of Chiang (2007)) and 6) Apertium (Forcada et al., 2009) rule-based machine translation (RBMT). Finally, we 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 translation.

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 for the best 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 evaluation results on the test set. Section 4 concludes the paper.

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. The architecture includes various individual systems: phrase-based, example-based, hierarchical phrase-based and tree-based MT.

The combination structure uses the MBR and CN decoders, and is based on a word-level combination strategy (Du et al., 2009). In the final stage, we use a new rescoring module to process the N-best list generated by the combination module. Figure 1 illustrates the architecture.

2.2 Example-Based Machine Translation

The EBMT system uses a language-specific, reduced set of closed-class *marker* morphemes or lexemes (Gough and Way, 2004) to define a way to segment sentences into *chunks*, which are then aligned using an edit-distance-style algorithm, in which edit costs depend on word-to-word transla-

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 149–154,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

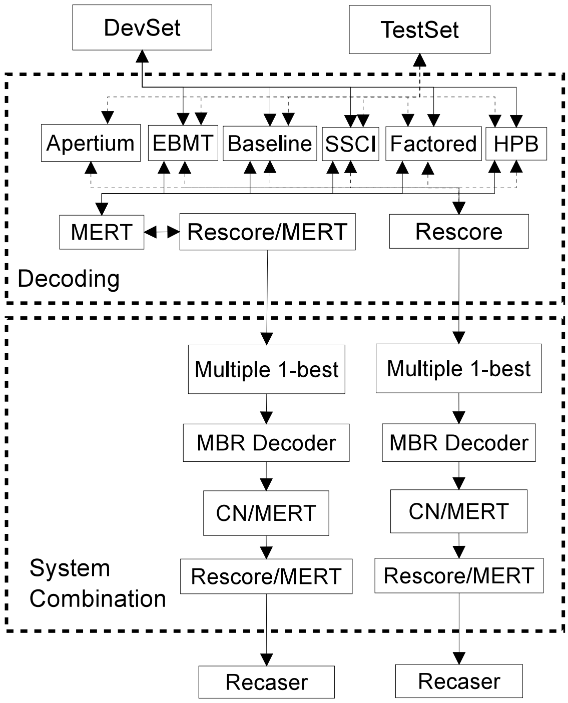

Figure 1: System Framework.

tion probabilities and the amount of word-to-word *cognates* (Stroppa and Way, 2006).

Once these phrase pairs were obtained they were merged with the phrase pairs extracted by the baseline system adding word alignment information.

2.3 Apertium RBMT

Apertium[1] is a free/open-source platform for RBMT. The current version of the en–es system in Apertium was used for the system combination task (section 2.7), and its morphological analysers and part-of-speech taggers were used to build a factored Moses model.

2.4 Factored Translation Model

We also used a factored model for the en–es translation task. Factored models (Koehn and Hoang, 2007) facilitate the translation by breaking it down into several factors which are further combined using a log-linear model (Och and Ney, 2002).

We used three factors in our factored translation model, which are used in two different decoding paths: a surface form (SF) to SF translation factor, a lemma to lemma translation factor, and a part-of-speech (PoS) to PoS translation factor.

Finally, we used two decoding paths based on the above three translation factors: an SF to SF decoding path and a path which maps lemma to lemma, PoS to PoS, and an SF generated using the TL lemma and PoS. The lemmas and PoS for en and es were obtained using Apertium (section 2.3).

2.5 Source-Side Context-informed PB-SMT

One natural way to express a context-informed feature ($\hat{h}_{\mathrm{MBL}}$) is to view it as the conditional probability of the target phrases ($\hat{e}_k$) given the source phrase ($\hat{f}_k$) and its source-side *context information* (CI):

$$\hat{h}_{\mathrm{MBL}} = \log P(\hat{e}_k|\hat{f}_k, \mathrm{CI}(\hat{f}_k)) \qquad (1)$$

We use a memory-based machine learning (MBL) classifier (TRIBL:[2] Daelemans and van den Bosch (2005)) that is able to estimate $P(\hat{e}_k|\hat{f}_k, \mathrm{CI}(\hat{f}_k))$ by similarity-based reasoning over memorized nearest-neighbour examples of source–target phrase translations. In equation (1), SSCI may include any feature (lexical, syntactic, etc.), which can provide useful information to disambiguate a given source phrase. In addition to using local words and PoS-tags as features, as in (Stroppa et al., 2007), we incorporate grammatical dependency relations (Haque et al., 2009a) and supertags (Haque et al., 2009b) as syntactic source context features in the log-linear PB-SMT model.

In addition to the above feature, we derived a simple binary feature $\hat{h}_{\mathrm{best}}$, defined in (2):

$$\hat{h}_{\mathrm{best}} = \begin{cases} 1 & \text{if } \hat{e}_k \text{ maximizes } P(\hat{e}_k|\hat{f}_k, \mathrm{CI}(\hat{f}_k)) \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

We performed experiments by integrating these two features, $\hat{h}_{\mathrm{MBL}}$ and $\hat{h}_{\mathrm{best}}$, directly into the log-linear framework of Moses.

2.6 Hierarchical PB-SMT model

For the en–cs translation task, we built a weighted synchronous context-free grammar model (Chiang, 2007) of translation that uses the bilingual phrase pairs of PB-SMT as a starting point to learn hierarchical rules. We used the open-source Tree-Based translation system moses-chart[3] to perform this experiment.

[1] http://www.apertium.org

[2] An implementation of TRIBL is freely available as part of the TiMBL software package, which can be downloaded from http://ilk.uvt.nl/timbl

[3] http://www.statmt.org/moses/?n=Moses.SyntaxTutorial

2.7 System Combination

For multiple system combination, we used an MBR-CN framework (Du et al., 2009, 2010) as shown in Figure 1. Due to the varying word order in the MT hypotheses, it is essential to define the *backbone* which determines the general word order of the CN. Instead of using a single system output as the skeleton, we employ an MBR decoder to select the best single system output E_r from the merged N-best list by minimizing the BLEU (Papineni et al., 2002) loss, as in (3):

$$r = \arg\min_{i} \sum_{j=1}^{N_s} (1 - \mathrm{BLEU}(E_j, E_i)) \qquad (3)$$

where N_s indicates the number of translations in the merged N-best list, and $\{E_i\}_{i=1}^{N_s}$ are the translations themselves. In our task, we only merge the 1-best output of each individual system.

The CN is built by aligning other hypotheses against the backbone, based on the TER metric. Null words are allowed in the alignment. Either votes or different confidence measures are assigned to each word in the network. 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;

We use MERT (Och, 2003) to tune the weights of the CN.

2.8 Rescoring

Rescoring is a very important part in post-processing which can select a better hypothesis from the N-best list. We augmented our previous rescoring model (Du et al., 2009) with more large-scale data. The features we used include:

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

3 Experimental Setup

This section describes our experimental setup for the en–cs and en–es translation tasks.

3.1 Data

Bilingual data: In the experiments we used data sets provided by the workshop organizers. For the en–cs translation table extraction we employed both parallel corpora (News-Commentary10 and CzEng 0.9), and for the en–es experiments, we used the Europarl(Koehn, 2005), News Commentary and United Nations parallel data. We used a maximum sentence length of 80 for en–es and 40 for en–cs. Detailed statistics are shown in Table 1.

Corpus	Langs.	Sent.	Source tokens	Target tokens
Europarl	en–es	1.6M	43M	45M
News-comm	en–es	97k	2.4M	2.7M
UN	en–es	5.9M	160M	190M
News-Comm	en–cs	85k	1.8M	1.6M
CzEng	en–cs	7.8M	80M	69M

Table 1: Statistics of en–cs and en–es parallel data.

Monolingual data: For language modeling purposes, in addition to the target parts of the bilingual data, we used the monolingual News corpus for cs; and the Gigaword corpus for es. For both languages, we used the SRILM toolkit (Stolcke, 2002) to train a 5-gram language model using all monolingual data provided. However, for en–es we used the IRSTLM toolkit (Federico and Cettolo, 2007) to train a 5-gram language model using the es Gigaword corpus. Both language models use modified Kneser-Ney smoothing (Chen and Goodman, 1996). Statistics for the monolingual corpora are given in Table 2.

Corpus	Language	Sentences	Tokens
E/N/NC/UN	es	9,6M	290M
Gigaword	es	40M	1,2G
News	cs	13M	210M

Table 2: Statistics of Monolingual Data. *E/N/NC/UN* refers to Europarl/News/News_Commentary/United_Nations corpora.

For all the systems except Apertium, we first lowercase and tokenize all the monolingual and bilingual data using the tools provided by the WMT10 organizers. After translation, system combination output is detokenised and true-cased.

3.2 English–Czech (en–cs) Experiments

The CzEng corpus (Bojar and Žabokrtský, 2009) is a collection of parallel texts from sources of different quality and as such it contains some noise. As the first step, we discarded those sentence pairs having more than 10% of non-Latin characters.

The CzEng corpus is quite large (8M sentence pairs). Although we were able to build a vanilla SMT system on all parallel data available (News-Commentary + CzEng), we also attempted to build additional systems using News-Commentary data (which we considered in-domain) and various in-domain subsets of CzEng hoping to achieve better results on domain-specific data.

For our first system, we selected 128,218 sentence pairs from CzEng labeled as *news*. For the other two systems, we selected subsets of 2M and 4M sentence pairs identified as most similar to the development sets (as a sample of in-domain data) based on cosine similarity of their representation in a TF-IDF weighted vector space model (cf. Byrne et al. (2003)). We also applied the pseudo-relevavance-feedback technique for query expansion (Manning et al., 2008) to select another subset with 2M sentence pairs.

We used the output of 15 systems for system combination for the en–cs translation task. Among these, 5 systems were built using Moses and varying the size of the training data (DCU-All, DCU-Ex2M, DCU-4M, DCU-2M and DCU-News); 9 context-informed PB-SMT systems (DCU-SSCI-*) using (combinations of) various context features (word, PoS, supertags and dependency relations) trained only on the News Commentary data (marked with ‡ in Table 4); and one system using the moses-chart decoder, also trained on the news commentary data.

3.3 English–Spanish (en–es) Experiments

Three baseline systems using Moses were built, where we varied the amount of training data used:

- epn: This system uses all of the Europarl and News-Commentary parallel data.
- UN-half: This system uses the data suplied to "epn", plus an additional 2.1M sentences pairs randomly selected from the United Nations corpus.
- all: This system uses all of the available parallel data.

For en–es we also obtained output from the factored model (trained only on the news commentary corpus) and the Apertium RBMT system. We also derived phrase alignments using the MaTrEx EBMT system (Stroppa and Way, 2006), and added those phrase translations in the Moses phrase table. The systems marked with ⋆ use a language model built using the Spanish Gigaword corpus, in addition to the one built using the provided monolingual data. These 6 sets of system outputs are then used for system combination.

3.4 Experimental Results

The evaluation results for en–es and en–cs experiments are shown in Table 3 and Table 4 respectively. The output of the systems marked † were submitted in the shared tasks.

System	BLEU	NIST	METEOR	TER
DCU-half †⋆	29.77%	7.68	59.86%	59.55%
DCU-all †⋆	29.63%	7.66	59.82%	59.74%
DCU-epn †⋆	29.45%	7.66	59.71%	59.64%
DCU-ebmt †⋆	29.38%	7.62	59.59%	60.11%
DCU-factor	22.58%	6.56	54.94%	67.65%
DCU-apertium	19.22%	6.37	49.68%	67.68%
DCU-system-combination †	30.42%	7.78	60.56%	58.71%

Table 3: en–es experimental results.

System	BLEU	NIST	METEOR	TER
DCU-All	10.91%	4.60	39.18%	81.76%
DCU-Ex2M	10.63%	4.56	39.12%	81.96%
DCU-4M	10.61%	4.56	39.26%	82.04%
DCU-2M	10.48%	4.58	39.35%	81.56%
DCU-Chart	9.34%	4.25	37.04%	83.87%
DCU-News	8.64%	4.16	36.27%	84.96%
DCU-SSCI-ccg‡	8.26%	4.02	34.76%	85.58%
DCU-SSCI-supertag-pair‡	8.11%	3.95	34.93%	86.63%
DCU-SSCI-ccg-ltag‡	8.09%	3.96	34.90%	86.62%
DCU-SSCI-PR‡	8.06%	4.00	34.89%	85.99%
DCU-SSCI-base‡	8.05%	3.97	34.61%	86.02%
DCU-SSCI-PRIR‡	8.03%	3.99	34.81%	85.98%
DCU-SSCI-ltag‡	8.00%	3.95	34.57%	86.41%
DCU-SSCI-PoS‡	7.91%	3.94	34.57%	86.51%
DCU-SSCI-word‡	7.57%	3.88	34.16%	87.14%
DCU-system-combination †	13.22%	4.98	40.39%	78.59%

Table 4: en–cs experimental results.

4 Conclusion

This paper presents the Dublin City University MT system in WMT2010 shared task campaign. This was DCU's first attempt to translate from en to es and cs in any shared task. We developed a multi-engine framework which combined the outputs of several individual 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 for both language pairs, and produce statistically significant improvements as measured by bootstrap resampling methods (Koehn, 2004) on BLEU over the single best system.

Acknowledgements: This work is supported by Science Foundation Ireland (Grant No. 07/CE/I1142) and by PANACEA, a 7th Framework Research Programme of the European Union, contract number 7FP-ITC-248064. M.L. Forcada's sabbatical stay at Dublin City University is supported by Science Foundation Ireland through ETS Walton Award 07/W.1/I1802 and by the Universitat d'Alacant (Spain).

References

Bojar, O. and Žabokrtský, Z. (2009). CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92:63–83.

Byrne, W., Khudanpur, S., Kim, W., Kumar, S., Pecina, P., Virga, P., Xu, P., and Yarowsky, D. (2003). The Johns Hopkins University 2003 Chinese–English machine translation system. In *Proceedings of MT Summit IX*, pages 447–450, New Orleans, LA.

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

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Daelemans, W. and van den Bosch, A. (2005). *Memory-Based Language Processing (Studies in Natural Language Processing)*. Cambridge University Press, New York, NY.

Du, J., He, Y., Penkale, S., and Way, A. (2009). MaTrEx: The DCU MT System for WMT2009. In *Proc. 3rd Workshop on Statistical Machine Translation, EACL 2009*, pages 95–99, Athens, Greece.

Du, J., Pecina, P., and Way, A. (2010). An Augmented Three-Pass System Combination Framework: DCU Combination System for WMT 2010. In *Proc. ACL 2010 Joint Workshop in Statistical Machine Translation and Metrics Matr*, Uppsala, Greece.

Federico, M. and Cettolo, M. (2007). Efficient Handling of N-gram Language Models for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 88–95, Prague, Czech Republic.

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.

Forcada, M. L., Tyers, F. M., and Ramírez-Sánchez, G. (2009). The free/open-source machine translation platform Apertium: Five years on. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation FreeRBMT'09*, pages 3–10.

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.

Haque, R., Naskar, S. K., Bosch, A. v. d., and Way, A. (2009a). Dependency relations as source context in phrase-based smt. In *Proc. 23rd Pacific Asia Conference on Language, Information and Computation*, pages 170–179, Hong Kong, China.

Haque, R., Naskar, S. K., Ma, Y., and Way, A. (2009b). Using supertags as source language context in SMT. In *EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 234–241, Barcelona, Spain.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, volume 4, pages 388–395.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P. and Hoang, H. (2007). Factored Translation Models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural*

Language Learning (EMNLP-CoNLL), pages 868–876, Prague, Czech Republic.

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. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, pages 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)*, 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.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

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.

Och, F. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*, volume 2, pages 295–302.

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., van den Bosch, A., and Way, A. (2007). Exploiting Source Similarity for SMT using Context-Informed Features. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 231–240, Skövde, Sweden.

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, NY.