

# MaTrEx: The DCU Machine Translation System for ICON 2008

Ankit Kumar Srivastava, Rejwanul Haque, Sudip Kumar Naskar, Andy Way

Centre for Next Generation Localisation

School of Computing

Dublin City University

Dublin, Ireland

{asrivastava, rhaque, snaskar, away}@computing.dcu.ie

## Abstract

In this paper, we give a description of the machine translation system developed at DCU that was used for our participation in the NLP Tools Contest of the International Conference on Natural Language Processing (ICON 2008). This was our first ever attempt at working on any Indian language. In this participation, we focus on various techniques for word and phrase alignment to improve system quality. For the English–Hindi translation task we exploit source-language reordering. We also carried out experiments combining both in-domain and out-of-domain data to improve the system performance and, as a post-processing step we transliterate out-of-vocabulary items.

## 1 Introduction

In this paper, we describe some basic experiments carried out on the English–Hindi translation task as well as test the data-driven hybrid MT system developed at DCU, MATREX (Machine Translation using examples) (Stroppa and Way, 2006), on the English–Hindi language pair.

The remainder of the paper is organized as follows. Section 2 describes the various components of the MATREX system. In Section 3, we present the system configuration for both the EILMT and TIDES data. We describe the experiments conducted in Section 4, reporting the results in Section 5, followed by discussion and avenues for future research in Section 6.

## 2 The MATREX System

The MATREX system is a modular hybrid data-driven MT system which exploits aspects of both the EBMT and SMT paradigms. It consists of a number of extendible and re-implementable modules and features, some of which are *Word Alignment Module*, *Word packing* (Ma et al., 2007), *Chunking and Chunk Alignment Module* (Tinsley et al., 2008), *Treebank-based phrase extraction* (Tinsley et al., 2007), *Supertagging* (Hasan et al., 2007), *Source-context features* (Stroppa et al., 2007), and *Decoder*.

In some cases, these modules may comprise wrappers around pre-existing software. For example, our system configuration for the translation task incorporates a wrapper around GIZA++ (Och and Ney, 2003) for word alignment and a wrapper around Moses (Koehn et al., 2007) for decoding. The system also includes language-specific extensions such as taggers, parsers, etc. used in pre-processing and post-processing modules described in the next section.

The MATREX system makes use of marker-based chunking, which is based on 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. Using a set of closed-class (or “marker”) words, such as determiners, conjunctions, prepositions, possessive and personal pronouns, aligned source-target sentences are segmented into chunks (Gough and Way, 2004) during a pre-processing step. A chunk is created at each new occurrence of a marker word, with the restriction that each chunk must contain at least one content (or non-marker) word. In order to align the chunks obtained by the chunk-

ing procedures, the system makes use of an “edit-distance style” dynamic programming alignment algorithm.

Many features available in our system which, for one reason or another, were not exploited for the purposes of this translation task.

### 3 System Description

The following section describes the system setup for the TIDES and EILMT data.

#### 3.1 Pre-processing

We carried out an analysis of the training data. We filtered out sentence pairs based on length (>100 words) and fertility (2:1 word ratio). We found the EILMT corpus to be cleaner than the TIDES corpus (highest E–H word ratio 4.7 vs. 16.5, lowest E–H word ratio .17 vs. .08). Details of these statistics are given in Table 1.

	TIDES	EILMT
Total sentences	50000	7000
$ \text{En}  > 100$ or, $ \text{Hn}  > 100$	123	1
Fertility > 2	314	32
Highest E–H word ratio	16.5	4.7
Lowest E–H word ratio	.08	.17
Highest E–H char ratio	82.5	5.6
Lowest E–H char ratio	.17	.23

Table 1: Summary of pre-processing of training data

Reordering of phrases during translation is typically managed by distortion models, which have not proved to be entirely satisfactory (Collins et al., 2006), especially for language pairs that differ a lot in terms of word order. The English–Hindi language pair exhibits longer distance SOV–SVO syntactic divergence. In this work, target language-specific heuristics are applied to reorder English syntactic trees in the pre-processing step. This reduces, and often eliminates, the *distortion load* during training. An open source probabilistic parser (Stanford Parser<sup>1</sup> version 1.6.1) was used to parse the English sentences. The parse tree is traversed for transformation of the English phrase pattern (part-of-speech tag and phrase-based schema form) to the Hindi phrase pattern. English-to-Hindi phrase transformation rules are applied to the intermediate nodes (representing

<sup>1</sup><http://nlp.stanford.edu/software/lex-parser.shtml/>

phrases). A CRF-based chunker<sup>2</sup> was used for proper reordering of VPs (usually for complex and compound sentences). We also augment the training data with an English–Hindi lexicon of 40K entries.

#### 3.2 System Configuration

As mentioned in Section 2, our word alignment module employs a wrapper around GIZA++.

We built a 5-gram language model based on the target side of the training data. This was done using the SRI Language Modelling toolkit (Stolcke, 2002) employing linear interpolation and modified Kneser–Ney discounting (Chen and Goodman, 1996).

Our phrase-table comprised of word alignment-based phrase pairs extracted using the grow-diag-final method (Koehn et al., 2003), with a maximum phrase length of 7 words. Phrase translation probabilities were estimated by relative frequency over all phrase pairs and were combined with other features, such as a reordering model, in a log-linear combination of features. We tuned our system on the development set provided using minimum error-rate training (Och, 2003) to optimise the BLEU score. Finally, we carried out decoding using a wrapper around the Moses decoder.

#### 3.3 Post-processing

The translated output contained many OOV words including named entities and unseen words. The OOV words are transliterated using a modified joint source channel-based transliteration module (Ekbal et al., 2006) in a bid to maximize matches with the reference.

### 4 Experiments

Due to time constraints, we could not run the same set of experiments on both data sets. In case of experiments on TIDES data, the 1K development set was divided into a 400-sentence development set and a 600-sentence test set for the first three experiments; only the last experiment was carried out on the released 1K test set and tuned on the the entire development set. For the EILMT data, we first ran the baseline MOSES (VANILLA) on the test data without any pre-processing or post-processing. Then we incrementally applied filtering (FIL), reordering of the source side (REO), augmentation

<sup>2</sup><http://crfchunker.sourceforge.net/>

Experiment	BLEU	NIST	METEOR	WER	PER
VANILLA	15.68	4.90	32.05	79.89	54.92
+ TRA	15.79	4.94	53.91	79.73	54.61
FIL (Baseline)	16.35	4.99	31.6	78.92	54.32
REO	17.06	4.88	32.11	76.20	55.56
+ TRA	17.14	4.92	51.61	76.02	55.27
REO + FAC	13.15	4.57	30.40	75.24	57.59
REO + LEX	17.33	4.98	35.83	76.03	54.93
+ TRA	<b>17.41</b>	<b>5.02</b>	<b>53.64</b>	<b>75.85</b>	<b>54.67</b>

Table 2: Summary of the results on EILMT test data

Experiment	BLEU	NIST	METEOR	WER	PER
VANILLA	4.87	2.69	22.86	93.83	76.65
REO	8.70	3.81	35.47	87.42	68.45
REO + LEX	8.63	3.84	35.45	87.22	68.16
REO + FAC	7.4	3.69	35.46	83.12	66.58
FIL + REO + LEX	10.47	4.23	31.50	87.61	6.34
FIL + REO + LEX + TRA	<b>10.49</b>	<b>4.24</b>	<b>37.10</b>	<b>87.57</b>	<b>66.27</b>

Table 3: Summary of the results on TIDES test data

of the training set with lexicon (LEX), and transliteration (TRA) to factor out the individual effects of each of these techniques. We also tested factored models (Koehn and Hoang, 2007)(FAC) with different factors (word, root/lemma (4 length), Part-of-Speech, morph). These were obtained using the Stanford Parser and WordNet<sup>3</sup> on the source side, and the IIIT Shallow Parser<sup>4</sup> on the target side.

## 5 Results

We initially carried out the same experiments on the Hindi data in both *wx* format and *utf* format, in which the latter format always gave us better results. Therefore we conducted all the experiments on *utf* Hindi data only. System performance is evaluated with respect to BLEU, NIST, METEOR, WER and PER. Results on the EILMT data and TIDES data are given in Tables 2 and 3 respectively. As expected, each of the pre-processing and post-processing modules proved to be useful in improving performance. MOSES without any pre-processing or post-processing produces a BLEU score of 15.68 (opposed to 17.70 reported by the organiser) on the EILMT data set. Since filtering the training data improves performance across all evaluation metrics for all the experiments, MOSES on filtered data serves as our new

baseline. Our best results (**17.41** BLEU score on EILMT data and **10.49** on TIDES data) are obtained when we apply all the pre-processing and post-processing modules collectively. On the contrary, factored models fall way short of the baseline system. The best result (13.15 BLEU score on EILMT data) produced by the factored model is obtained for the settings: alignment 1-1, translation 0-0+1-1, generation 1-0, decoding t0, t1, g0 (where 0 and 1 represents word and root factor respectively). Results on the TIDES corpus are at least 6 BLEU points lower than the results obtained on the EILMT corpus.

## 6 Discussion

We conducted a number of experiments with different settings, and observed that data in the *utf* format always gives better results. It is also evident that filtering out noise from the training data helps and that reordering is an important pre-processing step which improved the baseline system performance. In the post-processing phase, transliteration is performed which increases the accuracy slightly. All the pre-processing and post-processing modules improve system performance individually and the best result is obtained when they are applied collectively. Poor results on the TIDES data perhaps indicate the domain dependency of SMT systems. Another interesting observation is that even with 4 features the factored

<sup>3</sup><http://wordnet.princeton.edu/>

<sup>4</sup><http://ltrc.iiit.ac.in/analyzer/>

model failed to reach the baseline accuracy obtained by MOSES.

Furthermore, on manual evaluation of a random set of 60 sentences, it was found that the output translations look better than these scores suggest, i.e. given the relatively free word order (used with emphasis and complex structures) in Hindi, providing just one reference translation set for evaluation is somewhat arbitrary.

Due to time constraints we could not apply the EBMT module (which is a unique feature of the MATREX system) in the experiments, but we hope to report results using this method at the conference itself.

## Acknowledgments

This work is supported by Science Foundation Ireland (grant number: 07/CE/I1142).

## 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.
- Collins, M., Koehn, P., and Kucerova, I. (2006). Clause Restructuring for Statistical Machine Translation. In *Proceedings of ACL-2006*, pages 531–540, Sydney, Australia.
- Ekbal, A., Naskar, S.K., and Bandyopadhyay, S. (2006). A modified joint source-channel model for transliteration. In *Proceedings of the Coling/ACL 2006 Main Conference Poster Sessions*, pages 191–198, Sydney, Australia.
- Green, T. (1979). The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior*, 18:481–496.
- 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.
- Hassan, H., Sima'an, K., and Way, A. (2007). Supertagged Phrase-based Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 288–295, Prague, Czech Republic.
- Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03)*, pages 48–54, Edmonton, Canada.
- 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.
- Koehn, P. and Hoang, H. (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 '07)*, pages 868–876, Prague, Czech Republic.
- Ma, Y., Stroppa, N., and Way, A. (2007). Bootstrapping Word Alignment via Word Packing. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 304–311, Prague, Czech Republic.
- 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., Sapporo, Japan.
- Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.
- Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference Spoken Language Processing*, 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.
- 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.
- Tinsley, J., Hearne, M., and Way, A. (2007). Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07)*, pages 175–187, Bergen, Norway.
- Tinsley, J., Ma, Y., Ozdowska S., and Way, A. (2008). MaTrEx: the DCU MT System for WMT 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation, ACL 2008*, pages 171–174, Columbus, OH.