MaTrEx: the DCU MT System for WMT 2008

John Tinsley, Yanjun Ma, Sylwia Ozdowska, Andy Way
National Centre for Language Technology
Dublin City University
Dublin 9, Ireland
{jtinsley, yma, sozdowska, 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 evaluation campaign of the Third Workshop on Statistical Machine Translation at ACL 2008.

We describe the modular design of our data-driven MT system with particular focus on the components used in this participation. We also describe some of the significant modules which were unused in this task.

We participated in the *EuroParl* task for the following translation directions: Spanish–English and French–English, in which we employed our hybrid EBMT-SMT architecture to translate. We also participated in the Czech–English *News* and *News Commentary* tasks which represented a previously untested language pair for our system. We report results on the provided development and test sets.

1 Introduction

In this paper, we present the Data-Driven MT systems developed at DCU, MaTrEx (Machine Translation using Examples). This system is a hybrid system which exploits EBMT and SMT techniques to build a combined translation model.

We participated in both the French–English and Spanish–English EuroParl tasks. In these two tasks, we monolingually chunk both source and target sides of the dataset using a marker-based chunker (Gough and Way, 2004). We then align these chunks using a dynamic programming, edit-distance-style algorithm and combine them with phrase-based SMT-style chunks into a single translation model.

We also participated in the Czech–English News Commentary and News tasks. This language pair represents a new challenge for our system and provides a good test of its flexibility.

The remainder of this paper is organised as follows: Section 2 details the various components of our system, in particular the chunking and chunk alignment strategies used for the shared task. In Section 3, we outline the complete system setup for the shared task, and in Section 4 we give some results and discussion thereof.

2 The MaTrEx System

The MaTrEx system is a modular hybrid data-driven MT system, built following established Design Patterns, which exploits aspects of both the EBMT and SMT paradigms. It consists of a number of extendible and re-implementable modules, the most significant of which are:

- *Word Alignment Module*: outputs a set of word alignments given a parallel corpus,
- *Chunking Module*: outputs a set of chunks given an input corpus,
- *Chunk Alignment Module*: outputs aligned chunk pairs given source and target chunks extracted from comparable corpora,
- *Decoder*: returns optimal translation given a set of aligned sentence, chunk/phrase and word pairs.

In some cases, these modules may comprise wrappers around pre-existing software. For example, our system configuration for the shared task incorporates a wrapper around Giza++ (Och and Ney, 2003) for word alignment and a wrapper around Moses (Koehn et al., 2007) for decoding. It

Proceedings of the Third Workshop on Statistical Machine Translation, pages 171–174,
Columbus, June 2008.

should be noted, however, that the complete system is not limited to using only these specific module choices. The following subsections describe those modules unique to our system.

2.1 Marker-Based Chunking

The chunking module used for the shared task is based on the Marker Hypothesis, 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 for a particular language, such as determiners, prepositions, conjunctions and pronouns, sentences are segmented into chunks. 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. An example of this chunking strategy for English and Spanish is given in Figure 1.

2.2 Chunk Alignment

In order to align the chunks obtained by the chunking procedures described in Section 2.1, we make use of an "edit-distance-style" dynamic programming alignment algorithm.

In the following, a denotes an alignment between a target sequence e consisting of I chunks and a source sequence f consisting of J chunks. Given these sequences of chunks, we are looking for the most likely alignment $\hat{a}$:

$$\hat{a} = \underset{a}{\operatorname{argmax}}\, \mathbb{P}(a|e,f) = \underset{a}{\operatorname{argmax}}\, \mathbb{P}(a,e|f).$$

We first consider alignments such as those obtained by an edit-distance algorithm, i.e.

$$a = (t_1, s_1)(t_2, s_2) \ldots (t_n, s_n),$$

with $\forall k \in [\![1,n]\!]$, $t_k \in [\![0,I]\!]$ and $s_k \in [\![0,J]\!]$, and $\forall k < k'$:

$$t_k \leq t_{k'} \text{ or } t_{k'} = 0,$$
$$s_k \leq s_{k'} \text{ or } s_{k'} = 0,$$

where $t_k = 0$ (resp. $s_k = 0$) denotes a non-aligned target (resp. source) chunk.

We then assume the following model:

$$\mathbb{P}(a,e|f) = \Pi_k \mathbb{P}(t_k, s_k, e|f) = \Pi_k \mathbb{P}(e_{t_k}|f_{s_k}),$$

where $\mathbb{P}(e_0|f_j)$ (resp. $\mathbb{P}(e_i|f_0)$) denotes an "insertion" (resp. "deletion") probability.

Assuming that the parameters $\mathbb{P}(e_{t_k}|f_{s_k})$ are known, the most likely alignment is computed by a simple dynamic-programming algorithm.[1]

Instead of using an Expectation-Maximization algorithm to estimate these parameters, as commonly done when performing word alignment (Brown et al., 1993; Och and Ney, 2003), we directly compute these parameters by relying on the information contained within the chunks. The conditional probability $\mathbb{P}(e_{t_k}|f_{s_k})$ can be computed in several ways. In our experiments, we have considered three main sources of knowledge: (i) word-to-word translation probabilities, (ii) word-to-word cognates, and (iii) chunk labels. These sources of knowledge are combined in a log-linear framework. The weights of the log-linear model are not optimised; we experimented with different sets of parameters and did not find any significant difference as long as the weights stay in the interval $[0.5 - 1.5]$. Outside this interval, the quality of the model decreases. More details about the combination of knowledge sources can be found in (Stroppa and Way, 2006).

2.3 Unused Modules

There are numerous other features available in our system which, due to time constraints, were not exploited for the purposes of the shared task. They include:

- *Word packing* (Ma et al., 2007): a bilingually motivated packing of words that changes the basic unit of the alignment process in order to simplify word alignment.
- *Supertagging* (Hassan et al., 2007b): incorporating lexical syntactic descriptions, in the form of supertags, to the language model and target side of the translation model in order to better inform decoding.
- *Source-context features* (Stroppa et al., 2007): use memory-based classification to incorporate context-informed features on the source side of the translation model.
- *Treebank-based phrase extraction* (Tinsley et al., 2007): extract word and phrase alignments based on linguistically informed sub-sentential alignment of the parallel data.

[1] This algorithm is actually a classical edit-distance algorithm in which distances are replaced by opposite-log-conditional probabilities.

English: *[I voted] [in favour] [of the strategy presented] [by the council] [concerning relations] [with Mediterranean countries]*
Spanish: *[He votado] [a favor] [de la estrategia presentada] [por el consejo] [relativa las relaciones] [con los países mediterranéos]*

Figure 1: English and Spanish Marker-Based chunking

Filter criteria	es–en	fr–en	cz–en
Initial Total	1258778	1288074	1096941
Blank Lines	5632	4200	2
Length	6794	8361	2922
Fertility	120	82	1672
Final Total	1246234	1275432	1092345

Table 1: Summary of pre-processing on training data.

System	BLEU (-EBMT)	BLEU (+EBMT)
es–en	0.3283	0.3287
fr–en	0.2768	0.2770
cz–en	0.2235	-

Table 2: Summary of results on developments sets *devtest2006* for EuroParl tasks and *nc-test2007* for cz–en tasks.

System	BLEU (-EBMT)	BLEU (+EBMT)
es–en	0.3274	0.3285
fr–en	0.3163	0.3174
cz–en (news)	0.1458	-
cz–en (nc)	0.2217	-

Table 3: Summary of results on 2008 test data.

3 Shared Task Setup

The following section describes the system setup using the Spanish–English and French–English *EuroParl*, and Czech–English *CzEng* training data.

3.1 Pre-processing

For all tasks we initially tokenised the data (Czech data was already tokenised) and removed blank lines. We then filtered out sentence pairs based on length (>100 words) and fertility (9:1 word ratio). Finally we lowercased the data. Details of this pre-processing are given in Table 1.

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 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 a combination of marker-based chunk pairs[2], extracted as described in Sections 2.1 and 2.2, and word-alignment-based phrase pairs extracted using the "*grow-diag-final*" method of 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 functions.

We tuned our system on the development set *devtest2006* for the EuroParl tasks and on *nc-test2007* for Czech–English, using minimum error-rate training (Och, 2003) to optimise BLEU score.

Finally, we carried out decoding using a wrapper around the Moses decoder.

3.3 Post-processing

Case restoration was carried out by training the system outlined above - without the EBMT chunk extraction - to translate from the lowercased version of the applicable target language training data to the truecased version. We have previously shown this approach to be very effective for both case and punctuation restoration (Hassan et al., 2007a). The translations were then detokenised.

4 Results

The system output is evaluated with respect to BLEU score. Results on the development sets and test sets for each task are given in Tables 2 and 3 respectively, where "-EBMT" indicates that EBMT chunk modules were not used, and "+EBMT" indicates that they were used.

[2]This module was omitted from the Czech–English system as we have yet to verify whether marker-based chunking is appropriate for Czech.

4.1 Discussion

Those configurations which incorporated the EBMT chunks improved slightly over those which did not. Groves (2007) has shown previously that combining EBMT and SMT translation models can lead to considerable improvement over the baseline systems from which they are derived. The results achieved here lead us to believe that on such a large scale there may be a more effective way to incorporate the EBMT chunks.

Previous work has shown the EBMT chunks to have higher precision than their SMT counterparts, but they lack sufficient recall when used in isolation (Groves, 2007). We believe that increasing their influence in the translation model may lead to improved translation accuracy. One experiment to this effect would be to add the EBMT chunks as a separate phrase table in the log-linear model and allow the decoder to chose when to use them.

Finally, we intend to exploit the unused modules of the system in future experiments to investigate their effects on the tasks presented here.

Acknowledgments

This work is supported by Science Foundation Ireland (grant nos. 05/RF/CMS064 and OS/IN/1732). Thanks also to the reviewers for their insightful comments and suggestions.

References

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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.

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.

Groves, D. (2007). *Hybrid Data-Driven Models of Machine Translation*. PhD thesis, Dublin City University, Dublin, Ireland.

Hassan, H., Ma, Y., and Way, A. (2007a). MATREX: the DCU Machine Translation System for IWSLT 2007. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 69–75, Trento, Italy.

Hassan, H., Sima'an, K., and Way, A. (2007b). 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., 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., 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.

Ma, Y., Stroppa, N., and Way, A. (2007). Boostrapping 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., 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.

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.