MaTrEx: the DCU Machine Translation System for IWSLT 2007

Hany Hassan Yanjun Ma Andy Way
School of Computing
Dublin City University
Dublin, Ireland
{hhasan,yma,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 second participation in the evaluation campaign of the International Workshop on Spoken Language Translation (IWSLT 2007). In this participation, we focus on some new methods to improve system quality. Specifically, we try our word packing technique for different language pairs, we smooth our translation tables with out-of-domain word translations for the Arabic–English and Chinese–English tasks in order to solve the high number of out of vocabulary items, and finally we deploy a translation-based model for case and punctuation restoration.

We participated in both the classical and challenge tasks for the following translation directions: Chinese–English, Japanese–English and Arabic–English. For the last two tasks, we translated both the single-best ASR hypotheses and the correct recognition results; for Chinese–English, we just translated the correct recognition results. We report the results of the system for the provided evaluation sets, together with some additional experiments carried out following identification of some simple tokenisation errors in the official runs.

1 Introduction

In this paper, we describe some new extensions to the data-driven MT system developed at DCU, MaTrEx (Machine Translation using examples), subsequent to our participation at IWSLT 2006 (Stroppa and Way, 2006).

Firstly, we extend our word packing technique (Ma et al., 2007a) to Japanese and Arabic. Secondly, we demonstrate that smoothing the Arabic–English translation tables with out-of-domain data improves BLEU score by 9.67% relative, and 2.28% relative for Chinese–English, compared to the baseline systems. Thirdly, we treat case and punctuation as a translation task by using the translations with case differences and punctuation marks as the 'target' text, and the equivalent texts with these removed as the 'source'. The system then learns how to restore case and punctuation from this bitext for this pseudo-translation pair.

We participated in both the classical and challenge tasks for the following translation directions: Chinese–English, Japanese–English and Arabic–English. For the last two tasks, we translated both the single-best ASR hypotheses and the correct recognition results; for Chinese–English, we just translated the correct recognition results. We report the results of the system for the provided evaluation sets.

The remainder of the paper is organized as follows. In section 2, we describe the various components of the system; in particular, we give details about the various novel extensions to Ma-

TrEx as summarised above. In section 3, we report experimental results obtained for the three language pairs, while in section 4, we conclude, and provide avenues for further research.

2 The MaTrEx System

The MaTrEx system is a hybrid system which exploits both EBMT and SMT techniques to extract a dataset of aligned chunks (Armstrong et al., 2006). It is a modular data-driven MT engine, built following established Design Patterns (Gamma et al., 1995), and consists of a number of extendible and re-implementable modules (Armstrong et al., 2006; Stroppa and Way, 2006), the most significant of which are:

- *Word Alignment Module*: takes as its input an aligned corpus and outputs a set of word alignments.
- *Chunking Module*: takes in an aligned corpus and produces source and target chunks.
- *Chunk Alignment Module*: takes the source and target chunks and aligns them on a sentence-by-sentence level.
- *Decoder*: searches for a translation using the original aligned corpus and derived chunk and word alignments.

The word alignment module is one of the most important modules. (Ma et al., 2007a) has shown that word packing can improve statistical machine translation. In our experiments, we improve the word alignment module in our MaTrEx system using this word packing technique. In addition, we demonstrate the effects of smoothing the translation tables with out-of-domain data, and introduce a translation-based method for case and punctuation restoration.

2.1 Word Packing

2.1.1 Motivation

Most current statistical models (Brown et al., 1993; Vogel et al., 1996) treat the aligned sentences in the corpus as sequences of tokens that are meant to be interpreted as words; the goal of the alignment process is to find links between source and target words. Before applying such aligners, we thus need to segment the sentences into words – a task which can be quite hard for languages such as Chinese for which word boundaries are not orthographically marked. More importantly, however, this segmentation is often performed in a *monolingual* context, which makes the word alignment task more difficult since different languages may realize the same concept using varying numbers of words (Wu, 1997).[1]

Although some statistical alignment models allow for 1-to-n word alignments for those reasons, they rarely question the monolingual tokenization and the basic unit of the alignment process remains the word. In our system, we focus on 1-to-n alignments with the goal of simplifying the task of automatic word aligners by *packing* several consecutive words together when we believe they correspond to a single word in the opposite language; by identifying enough such cases, we reduce the number of 1-to-n alignments, thus making the task of word alignment both easier and more natural.

2.1.2 Bootstrapping Word Alignment

Our approach consists of using the output from an existing statistical word aligner to obtain a set of candidates for word packing. We evaluate the reliability of these candidates, using simple metrics based on co-occurrence frequencies, similar to those used in associative approaches to word alignment (Kitamura & Matsumoto, 1996; Melamed, 2000; Tiedemann, 2003). We then modify the segmentation of the sentences in the parallel corpus according to this packing of words; these modified sentences are then given back to the word aligner, which produces new alignments.

In this way, word packing can be applied several times: once we have grouped some words together, they become the new basic unit to consider, and we can re-run the same method to obtain additional groupings. However, in practice

[1] See (Ma et al., 2007b) for an investigation into different segmentations of source and target languages depending on the language pair at hand.

we have not seen much benefit from running it more than twice, i.e. few new candidates are extracted after two iterations.

As an extension to our previous work described in (Ma et al., 2007a), we allow word packing both with and without reference to the local context. If we group all the reliable n words that we believe to be a basic word unit without considering the contexts of this group of words, we call this word packing scheme *context-free word packing*; if we condition the word packing on the context, i.e. the word aligner aligns these n words in one language to one single word in the other language, we call it *context-sensitive word packing*.

2.1.3 Word Packing: Discussion

If we can pack words in an 'appropriate' way, the complexity of word alignment might hopefully be reduced; otherwise, the packed words may impact in a negative way on word alignment. Therefore, deciding which words should be packed is the most difficult part of word packing. From our experiments, we also observe that the translation results are quite sensitive to the number of packed words.

This word packing is performed based on pretokenized corpora and the word alignment using an existing word aligner, namely Giza++ (Och and Ney, 2003). An integrated model combining tokenization and word alignment is a promising novel direction for future work.

2.2 Smoothing Translation Tables

The Arabic to English task suffers from a small amount of training data, with test sets in the past containing a very high number of out of vocabulary (OOV) items. For example, the OOV ratio for the IWSLT06 test set was over 24%. It is quite a challenging task to source more training data in a similar domain, and usually data from another domain degrades translation accuracy. In the IWSLT 2004 Chinese-to-English translation task, for example, out-of-domain data consistently degraded the translation performance when added to the domain-specific data (Akiba et al., 2004). In IWSLT 2006, (Lee, 2006) combined out-of-domain data with domain-specific data by assigning a higher weight to the domain-specific training corpus than to the out-of-domain corpora. In our work presented here, we use translation models trained on out-of-domain data to smooth the domain-specific translation models. We think the cause of degradation in performance when adapting the phrase-based system with out-of-domain phrasal translations is due to two main problems:

- First, different domains indicate different phrase styles, i.e. questions versus news style;
- Second, phrases from the larger out-of-domain data usually have a higher score than in-domain phrases due to the fact that the out-of-domain data is much larger than in-domain-data. This might cause a bias toward the choice of an incorrect translation obtained via out-of-domain data of words and phrases, when these occur in both in-domain and out-of-domain training data sets(Lee, 2006).

In the current work we tried to avoid both problems by smoothing the in-domain translation tables with word translation probabilities from the out-of-domain data. In other words, we added phrases of length one from the out-of-domain data to our in-domain phrase tables. We use the out-of-domain data to obtain the lexical probabilities via Giza++ (Och and Ney, 2003) to obtain word-level alignments in both language directions. For consistency, the bidirectional alignments are used to derive word translation scores (Koehn et al., 2003). The resulting word-based translation table is combined with the in-domain translation table to construct a larger smoothed translation table. We tried to use the out-of-domain translation tables for translating OOV words only; however, we found that using the OOV translation tables helps in translating both in-vocabulary *and* OOV items. We combine the out-of-domain word-based translation table with the in-domain phrase table simply by introducing a back-off procedure. If a phrase

Smoothing	OOV Ratio
No smoothing	24.23%
Smoothing	6.42%

Table 1: Smoothing effect on OOV ratio for IWSLT06

translation exists in the in-domain phrase table we use it, otherwise we back-off to the more reliable word to word translation from the out-of-domain data.

The proposed technique improved the score on the IWSLT06 test set for Arabic to English task from 23.68 to 25.97 BLEU score, a relative improvement of 9.6% (cf. Table 2). It also improved the score on IWSLT07 test set for Chinese to English task from 30.00 to 30.53 BLEU score (cf. Table 6).

Table 1 shows how smoothing affects the OOV ratio for the IWSLT06 test set for the Arabic to English task, where it can be seen that the OOV ratio dropped from over 24% to 6.4%. This large degradation in the OOV ratio results in better translations as reflected by the automatic evaluation scores.

2.3 Case and Punctuation Restoration

Case and punctuation restoration is an important post-processing step for speech translation. For punctuation restoration, it is possible to consider punctuation marks as hidden events occurring between words, with the most likely hidden tag sequence (consistent with the given word sequence) being found using an n-gram language model trained on a punctuated text. For case restoration, the task can be viewed as a disambiguation task in which we have to choose between the (case) variants of each word of a sentence. Again, finding the most likely sequence can be done using an n-gram language model trained on a case-sensitive text.

In our experiments, we consider case and punctuation restoration as a translation process. In this model, case and punctuation restoration can be combined together. The case-sensitive text with punctuation can be considered as the target language. Then we remove the punctuation and case information in the target language and use them as the corresponding source language to construct a pseudo-'bilingual' corpus. With this 'bilingual' corpus, we can train a phrase-based statistical machine translation system to restore punctuation and case information. Naturally we can train a system to restore just punctuation information, or if required just case information.

We consider this approach to be very effective in restoring punctuation and case information especially for the ASR data (cf. Tables 2, 6 and 7).

3 Experimental Results

3.1 Data

The experiments were carried out using the provided datasets, extracted from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002). This multilingual speech corpus contains tourism-related sentences similar to those that are usually found in phrasebooks for tourists going abroad. We participated in both the classical and challenge tasks for the following translation directions: Chinese–English, for which we translated the correct recognition results, and both Japanese–English and Arabic–English, for which we translated both the single-best hypotheses and the correct recognition results.

Training was performed using the default training set, to which we added the sets devset1, devset2, and devset3 for the Chinese–English task.[2] We used devset4 for development purposes. For the Arabic to English task, training was performed using the default training set, to which we added the sets devset1, devset2, and devset3. All language models were built using the SRILM toolkit[3] (Stolcke, 2002) using only the target side of the bilingual training data.

As a pre-processing step, the English sentences were tokenized using the maximum entropy-based tokenizer of the OpenNLP[4] toolkit, and case information was removed. For training the out-of-domain word probabilities,

[2]More specifically, we chose the first English reference from the 7 references and the Chinese sentence to construct new sentence pairs.

[3]http://www.speech.sri.com/projects/srilm/

[4]http://opennlp.sourceforge.net

we used the LDC parallel news data and a large part of the UN data (2 million sentences, about 50 million words).

The Arabic data was tokenized and segmented using the ASVM toolkit[5] which is based on Support Vector Machines, a Machine Learning algorithm, and has been trained on the Arabic Treebank (Diab et al., 2004). The AVSM toolkit tokenized the Arabic data and segmented the Arabic words with the same segmentation style as in the Arabic Treebank.

3.2 Results

The system output is evaluated with respect to the BLEU automatic MT evaluation metric (Papineni et al., 2002), as computed by the IWSLT 2007 evaluation server. The official results are reported in Tables 3, 4, and 5. These results include case and punctuation information.

We used our word repacking technique (Ma et al., 2007a) for the Chinese-to-English, Japanese-to-English and Arabic-to-English translation tasks. For Chinese-to-English, we tried both context-free and context-sensitive word packing, and the best results yield a 3.4% relative increase in BLEU. In order to further investigate the effectiveness of word packing, we evaluate its effect on the lower-case translation quality, and we see a gain of 2.9% relative improvement in BLEU score. For Arabic-to-English, context-sensitive word packing showed only a minor improvement. For Japanese-to-English, we tried context-sensitive word packing, but did not see any improvement.

From Table 6, we can see while context-sensitive word packing improves BLEU, the translation quality decreases according to NIST (Doddington, 2002) and METEOR (Banerjee & Lavie, 2005). However, context-free word packing consistently improves quality according to all the automatic evaluation metrics. This shows that during word packing, maintaining consistency over the whole corpus is more important than having a mix of good and bad packings. We would like to try context-free word packing for both Arabic-to-English and Japanese-to-English translation to see whether similar improvements can be achieved as for Chinese-to-English.

[5]http://www1.cs.columbia.edu/ mdiab/

System	BLEU
Baseline	0.2253
WordPacking (WP)	0.2264
WP+Case/Punct Restoration (CP)	0.2368
WP+CP+ Smoothing for OOV	0.2453
WP+CP+Smoothing ALL	0.2597

Table 2: Arabic-to-English Results on IWSLT06

The smoothing of the phrase tables using out-of-domain data was very effective in overcoming the high ratio of OOV items for Arabic–English, with most of the gain in our view being contributed by this simple adaptation technique.

3.3 Arabic-to-English Results

For the Arabic-to-English translation task, Table 2 shows the translation performance for various configurations of the MaTrEx system on the IWSLT06 testset. As the results show, WordPacking (WP) gives a small improvement over the baseline system. Case and Punctuation restoration (CP) showed a significant improvement over baseline results. Smoothing the phrase table with word to word translation probabilities for OOV items provided further improvement, while smoothing the phrase table with word to word translation probabilities for all words , not just OOV items, resulted in our best overall score of 0.2597 BLEU, an improvement over the baseline of 0.0344 points absolute, or a 15.3% relative increase.

3.4 Comments on the Results

The word packing technique was used for all experiments. We found that both context-sensitive word packing and context-free word packing can significantly improve Chinese-to-English translation. However, if we take evaluation metrics other than BLEU into account, context-free word packing outperforms context-sensitive word packing. In addition, we find context-sensitive word packing cannot significantly improve Arabic-to-English and Japanese-to-English translation quality. This tells us that achieving consistency in the word packing pro-

Data Condition	BLEU
ASR output (1-best)	0.3942
Correct Transcripts	0.4709

Table 3: Official results – Arabic

Data Condition	BLEU
Corrected Transcripts	0.2737

Table 4: Official results – Chinese

cess is very important; even if the words are wrongly packed, if it is consistently wrong, that will lead to better performance compared to a mixture of partly wrong and partly correct word packings.

3.5 Additional Experiments

Following the submission of our official runs in the IWSLT-07 campaign, we noticed there were differences in tokenisation between the output from MaTrEx and the reference translations provided.

Accordingly, in a post-processing phase, we adapted the system output so that tokenisation was the same as that contained in the reference translations, for the Chinese–English and Japanese–English language pairs. The improved results are provided in Tables 7 and 8. For Chinese–English, we observe an increase from 0.3053 to 0.3203 for BLEU, a relative improvement of 4.91%. For Japanese–English, the increase was from 0.3959 to 0.4216 on the corrected transcripts, a relative improvement of 6.49%, while on the ASR output we improved from 0.3182 to 0.3523, a relative improvement of 10.7%.

4 Conclusion

In this paper, we described some new extensions to MaTrEx, the hybrid data-driven MT system developed at DCU. We described word packing, a technique to improve word alignment for MT and its integration into MaTrEx for new language pairs. We also introduced a new technique for adapting translation tables with out-of-domain data to help solve the OOV problem for the Arabic to English task. Finally, we handled the problems of case and punctuation restoration as a pseudo-MT problem which we believe helped restore punctuation into the ASR output with a high degree of success.

Data Condition	BLEU
ASR output (1-best)	0.3182
Corrected Transcripts	0.3959

Table 5: Official results – Japanese

Data Condition	experiments	BLEU
Corrected Transcripts	baseline	0.4216
	cs.wp	0.4208
ASR output (one-best)	baseline	0.3523
	cs.wp	0.3498

Table 8: Additional experimental results - Japanese

Acknowledgments

This work is supported by Science Foundation Ireland (grant number OS/IN/1732). We wish to thank our colleague Mary Hearne for her assistance in obtaining the output and submitting the results to IWSLT.

References

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun-ich Tsujii. 2004. Overview of the IWSLT04 Evaluation Campaign. In *IWSLT 2004 Proceedings*. Kyoto, Japan, pp.1–12.

Stephen Armstrong, Marian Flanagan, Yvette Graham, Declan Groves, Bart Mellebeek, Sara Morrissey, Nicolas Stroppa and Andy Way. 2006. MaTrEx: Machine Translation Using Examples. *TC-STAR Open-Lab Workshop on Speech Translation*. Trento, Italy.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation measures for MT and/or Summarization*, Ann Arbor, MI., pp.65–72.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic tagging of Arabic text: From raw text to

Data Condition	experiments	BLEU	NIST	METEOR	WER	PER
Corrected Transcripts	baseline	0.2897	6.23	54.18	0.5018	0.4167
	cs.wp	0.2985	5.96	53.52	0.4869	0.4138
	cf.wp	0.3000	6.23	54.40	0.4927	0.4138
	cs.w.smoothing	0.3053	6.29	54.72	0.4914	0.4127

Table 6: Additional experimental results - Chinese (true case)

Data Condition	experiments	BLEU	NIST	METEOR	WER	PER
Corrected Transcripts (lower case)	baseline	0.3051	6.23	54.17	0.5020	0.4170
	cs.wp	0.3145	5.96	53.52	0.4869	0.4138
	cf.wp	0.3145	6.23	54.40	0.4927	0.4138
	cs.w.smoothing	0.3203	6.29	54.72	0.4914	0.4127

Table 7: Additional experimental results - Chinese (lower case)

base phrase chunks. In *Proceedings of HLT-NAACL 2004*, pages 149–152, Boston, MA.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings Human Language Technology*, San Diego, CA., pp.128–132.

Erich Gamma, Richard Helm, Ralph Johnson and John Vlissides. 1995. *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley, Reading, MA.

Mihoko Kitamura and Yuji Matsumoto. 1996. Automatic Extraction of Word Sequence Correspondences in Parallel Corpora. In *Proceedings of 4th Workshop on Very Large Corpora*, Copenhagen, Denmark, pp.79–87.

Philip Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 48–54, Edmonton, Canada.

Young-Suk Lee. 2006. IBM Arabic-to-English translation for IWSLT 2006. In *Proceedings of IWSLT 2006 Workshop*, Kyoto, Japan, pp.45–52.

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007a. Boostrapping word alignment via word packing. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, Prague, Czech Republic, pp.304–311.

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007b. Alignment-Guided Chunking. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, Skövde, Sweden, pp.114–121.

I. Dan Melamed. 2000. Models of Translational Equivalence Between Words. *Computational Linguistics* **26**(2):221–249.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic (ACL 2002)*, pages 311–318, Philadelphia, PA.

Andrea Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Nicolas Stroppa and Andy Way. 2006. Matrex: DCU Machine Translation System for IWSLT 2006. In *Proceedings of IWSLT 2006 Workshop*, Kyoto, Japan, pp.31–36.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of LREC 2002*, Las Palmas, Canary Islands, Spain, pp.147–152.

Jörg Tiedemann. 2003. Combining Clues for Word Alignment. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary, pp.339–346.

Stefan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING 1996, 16th International Conference on Computational Linguistics, Proceedings of the Conference*, pages 836–841, Copenhagen, Denmark.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.