Exploiting Alignment Techniques in MATREX: the DCU Machine Translation System for IWSLT 2008

Yanjun Ma[†], *John Tinsley*[†], *Hany Hassan*[†], *Jinhua Du*[‡], *Andy Way*[†‡]

[†] National Centre for Language Technology
[‡] Centre for Next Generation Localisation
School of Computing
Dublin City University
Dublin, Ireland
{yma,jtinsley,hhasan,jdu,away}@computing.dcu.ie

Abstract

In this paper, we give a description of the machine translation (MT) system developed at DCU that was used for our third participation in the evaluation campaign of the International Workshop on Spoken Language Translation (IWSLT 2008). In this participation, we focus on various techniques for word and phrase alignment to improve system quality. Specifically, we try out our word packing and syntax-enhanced word alignment techniques for the Chinese–English task and for the English–Chinese task for the first time. For all translation tasks except Arabic–English, we exploit linguistically motivated bilingual phrase pairs extracted from parallel treebanks. We smooth our translation tables with out-of-domain word translations for the Arabic–English and Chinese–English tasks in order to solve the problem of the high number of out of vocabulary items. We also carried out experiments combining both in-domain and out-of-domain data to improve system performance and, finally, we deploy a majority voting procedure combining a language model-based method and a translation-based method for case and punctuation restoration. We participated in all the translation tasks and translated both the single-best ASR hypotheses and the correct recognition results. The translation results confirm that our new word and phrase alignment techniques are often helpful in improving translation quality, and the data combination method we proposed can significantly improve system performance.

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 [1] and IWSLT 2007 [2].

Firstly, we test our novel word and phrase alignment modules including word packing [3], syntax-enhanced word alignment [4] and parallel treebank-based phrase extraction [5]. Secondly, in addition to smoothing translation tables with out-of-domain data [2], we attempt to improve system coverage by training on a combination of both in-domain and out-of-domain data. Lastly, we deploy a majority voting procedure for punctuation restoration by combining a language model-based system and a translation-based system.

We participated in the CHALLENGE, BTEC and PIVOT tasks covering all language pairs and translation directions. This included: Chinese–English, English–Chinese, Arabic–English, Chinese–Spanish and Chinese–English–Spanish. We translated both the single-best ASR hypotheses and the correct recognition results.

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 MATREX as summarised above. In Section 3, the experimental setup is presented and experimental results obtained for various language pairs are reported in Section 4. In Section 5, 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 [6]. It is a modular data-driven MT engine, built following established designpatterns, and consists of a number of extensible and re-implementable modules [1, 6], 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.

For this participation, our system has been enriched with various novel word and phrasal alignment modules, including word packing [3], syntax-enhanced word alignment [4] and parallel treebank-based phrase extraction [5] as described in the following sections. We also developed some effective domain adaptation heuristics to facilitate the use of large-scale out-of-domain data to help improve system performance.

2.1. Improving Word Alignment via Word Packing

Most current statistical models [7, 8] treat the aligned sentences in the corpus as sequences of tokens that are meant to be 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 realise the same concept using varying numbers of words [9].

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.

Our word packing approach consists of using the output from an existing statistical word aligner (GIZA++, [10]) 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 [11, 12, 13]. 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 get additional groupings. However, in practice, we have not seen much benefit from running it more than twice (few new candidates are extracted after two iterations).

Word packing has been shown to be an effective approach to improve translation quality. However, this approach is very sensitive to the confidence measures used for packing words. Therefore, the optimisation of the confidence measures are essential to the performance of word packing approach. 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.

2.2. Syntax-enhanced Word Alignment

Syntactic dependency between words is a potentially useful source of knowledge for word alignment. Syntax-enhanced word alignment is a word alignment framework that facilitates the incorporation of syntactic dependencies encoded in bilingual dependency tree pairs. This model consists of two sub-models: an anchor word alignment model which aims to find a set of high-precision anchor links, and a syntax-enhanced word alignment model which focuses on aligning the remaining words relying on dependency information invoked by the acquired anchor links. The anchor links can be obtained using existing word aligners; the syntax-enhanced word alignment model incorporating dependency information can be estimated through discriminative training on a gold-standard word alignment corpus.

Figure 1 gives an example. Note that the link (c_2, e_4) can be easily identified, but the link involving the fourth Chinese word (a function word denoting 'time') (c_4, e_4) is hard. In such cases, we can make use of the dependency relationship ('tclause') between c_2 and c_4 to help the alignment process. Given such an observation, our model is composed of two related alignment models. The first one is an anchor alignment model which is used to find a set of anchor links; the other one is a syntax-enhanced alignment model aiming to process the words left unaligned after anchoring.

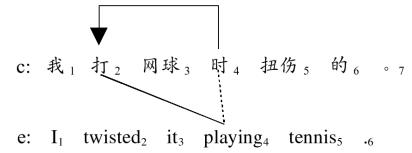


Figure 1: Syntactic dependencies can help word alignment

Formally, given a source sentence f_1^J and target sentence e_1^I, we seek to find the optimal alignment $\hat{A}$ such that:

$$\hat{A} = \underset{A}{\operatorname{argmax}} P(A|f_1^J, e_1^I) \tag{1}$$

We use a model (2) that directly models the linkage between source and target words similarly to [14]. We decompose this model into an anchor alignment model (3) and a syntax-enhanced model (4) by distinguishing anchor alignments from non-anchor alignments:

$$p(A|f_1^J, e_1^I) = \prod_{j=0}^{J} p(a_j|f_1^J, e_1^I, a_1^{j-1}) \tag{2}$$

$$= \frac{1}{Z} \cdot p_\epsilon(A_\Delta|f_1^J, e_1^I) \cdot \tag{3}$$

$$\prod_{j \in \bar{\Delta}} p(a_j|f_1^J, e_1^I, a_1^{j-1}, A_\Delta) \tag{4}$$

Here A_Δ denotes the set of anchor alignment where a set of word indice $\Delta \subset \{1, \ldots, J\}$ are involved.

The syntax-enhanced model is used to model the alignment of the words left unaligned after anchoring ($\bar{\Delta}$). We directly model the linkage between source and target words using a discriminative word alignment framework where various features can be incorporated. Given a source word f_j and the target sentence e_1^I, we search for the alignment a_j such that:

$$\hat{a_j} = \underset{a_j}{\operatorname{argmax}}\{p_{\lambda_1^M}(a_j|f_1^J, e_1^I, a_1^{j-1}, A_\Delta)\} \quad (5)$$
$$= \underset{a_j}{\operatorname{argmax}}\{\textstyle\sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I, a_1^j, A_\Delta, T_f, T_e)\}$$

In this decision rule, we assume that a set of highly reliable anchor alignments A_Δ has been obtained, and T_f (resp. T_e) is used to denote the dependency structure for source (resp. target) language. In such a framework, various machine learning techniques can be used for parameter estimation. We used Support Vector Machines (SVMs) in our experiments.

We have shown in [4] that our syntax-enhanced word alignment approach can lead to a significant reduction in alignment error rate [10]. When tested on MT tasks, this sophisticated syntax-enhanced word alignment model can also achieve competitive results.

Compared to word packing, the syntax-enhanced model has the advantage of capturing long distance dependencies between words. However, this approach relies on the quality of the dependency parsing. Furthermore, the acquisition of anchor alignments is essential for the overall performance.

2.3. Parallel Treebank-Based Phrase Extraction

Previous experiments have shown that, for a number of European language pairs, augmenting the standard phrase-based translation model with syntactically motivated phrase pairs extracted from a parallel treebank consistently improves translation accuracy [15, 16]. A parallel treebank is a linguistically annotated parallel corpus aligned at sub-sentential level. The sub-sentential alignments imply translational equivalence between the yields of the linked constituent pair. Figure 2 contrasts phrase-pair extraction from a parallel treebank with standard SMT phrase extraction based on word alignments.

In order to use this technique for the purposes of the IWSLT tasks, we needed to build a parallel treebank for each language pair from the original parallel training corpus. That is, one for Chinese–English (also serves English–Chinese), Chinese–Spanish and for the pivot task, Chinese–English and English–Spanish. The first step involved monolingually parsing each corpus. In all cases, Chinese and English data was parsed using the Berkeley parser [17] and Spanish data was parsed using Bikel's parser [18] trained on the Cast3LB treebank [19].The next step was the induction of the sub-sentential alignments between the tree pairs. This was carried out using our statistical tree alignment algorithm as described in [5]. Finally, from each parallel treebank for each task, we extracted a set of phrase pairs which we incorporated into the baseline phrase-table for each system. Table 1 shows the contribution of the parallel treebank phrase pairs to the translation model of each MT system.

We see that adding the parallel treebank phrase pairs to the translation model significantly increases the translation coverage of the system. Also the overlap between the parallel treebank phrase pairs and the baseline phrase pairs gives increased probability mass to those phrase pairs in the model. This is desirable as we may expect those phrase pairs to be more reliable having been extracted via both methods. The influence of the parallel treebank phrase pairs on coverage is especially evident in those systems with Chinese as one of the two languages. The increase is not as pronounced for the English–Spanish system (part of the pivot system) as the baseline model already achieves relatively high recall. While we will always extract a relatively consistent number of phrase pairs from a parallel treebank regardless of the language pair, phrase extraction in SMT is heavily reliant on the recall of the statistical word alignments. As word alignment involving Chinese is more difficult as we discussed in Section 2.1, the recall is lower and subsequently fewer phrase pairs are extracted.

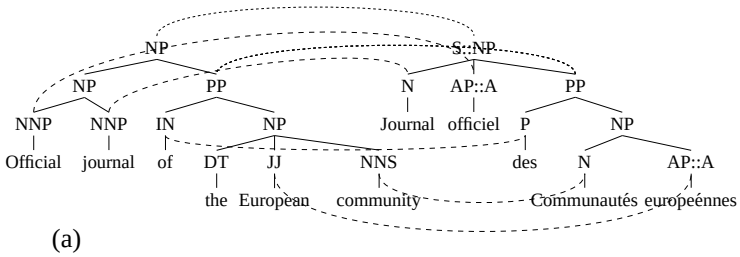
2.4. Smoothing Translation Tables

The translation tasks suffer due to the 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 20%. It is quite challenging 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. We think the cause of degradation in performance when adapting the phrase-based system with out-of-domain phrase translation is due to two main problems:

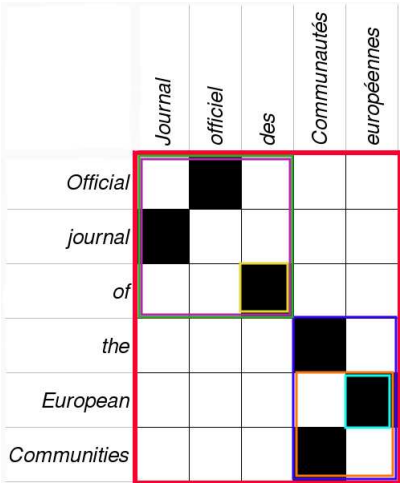
- First, different domains indicate different phrase styles, i.e. questions versus news style;
- Second, phrases from the out-of-domain data usually have a higher score than in-domain phrases due to the fact that there is much more out-of-domain data than in-domain data. This might cause a bias toward 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 [20].

To avoid these problems, [20] combined out-of-domain data with domain-specific data by assigning a higher weight

Training Sentence Pair
Official journal of the European Community ↔ Journal officiel des Communautés européennes



(a)



(b)

† Official journal	↔	Journal officiel
† Official journal of	↔	Journal officiel des
∗ Official journal of the European Communities	↔	Journal officiel des Communautés européennes
∗ of	↔	des
∗ of the European Communities	↔	des Communautés européennes
∗ the European Communities	↔	Communautés européennes
∗ European	↔	européennes
◇ Communities	↔	Communautés
◇ Official	↔	officiel
◇ journal	↔	Journal

(c)

Figure 2: Example of phrase extraction for the given sentence pair depicting: (a) the aligned parallel tree pair; (b) the word alignment matrix (the rectangled areas represent extracted phrase pairs); (c) the combined set of extracted phrase pairs where: ◇ = only extracted from (a); † = only extracted from (b); ∗ = extracted from both (a) and (b).

System	Initial Size	Treebank Size	Final Size	Coverage	Overlap
Zh–En	158,807	86,161	213,875	34.7%	19.6%
Zh–Es	101,593	68,870	151,446	49%	18.7%
Zh–En (pivot)	84,025	80,431	144,630	72.1%	23.6%
En–Es (pivot)	292,209	65,628	323,884	10.8%	11.6%

Table 1: Impact of adding parallel treebank phrase pairs to the baseline system where **Size**=the number of phrase pairs in the translation model, **Coverage**=the increase in coverage of the model given the treebank phrase pairs and **Overlap**=the percentage of phrase pairs extracted using both methods.

to the domain-specific training corpus than to the out-of-domain corpora in IWSLT 2006. In IWSLT 2007 [2], we used translation models trained on out-of-domain data to smooth the domain-specific translation models. Specifically, we smoothed the in-domain translation tables with word translation probabilities from the out-of-domain data by adding phrases of length one from the out-of-domain data to our in-domain phrase tables. 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 both for in-vocabulary and OOV items. 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).

Table 3 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 decrease in the OOV ratio results in better translations as reflected by the automatic evaluation scores.

System	BLEU
Baseline	0.2368
Smoothing for OOV	0.2453
Smoothing ALL	0.2597

Table 2: Impact of smoothing on IWSLT06 for Arabic–English

Smoothing	OOV Ratio
No smoothing	24.23%
Smoothing	6.42%

Table 3: Smoothing effect on OOV ratio for IWSLT06 Arabic–English

In our current work, we extend our approach of smoothing the translation table by combining in-domain and out-of-domain data. Differently from [20], by assigning a higher weight to the domain-specific training corpus, we only pick up a sentence from the out-of-domain data when all the words it contains occur in the bag of words from the in-domain training set. The in-domain data are all kept for training. Using this combination, we can not only overcome the OOV problems but also avoid any negative impact from the out-of-domain data.

2.5. Case and Punctuation Restoration

Case and punctuation restoration is an important component for speech translation. Punctuation restoration can be considered as a preprocessing or post-processing task, while case restoration is usually considered as a post-processing task. In order to obtain better word alignments for our MT system, we trained our system on data with punctuation. Therefore, we perform punctuation restoration as a preprocessing step preceding 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.

Punctuation restoration can also be considered as a translation process [2]. The text with punctuation can be considered as the target language. Then we remove the punctuation 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 SMT system to restore punctuation. Naturally we can also train a system to restore case information only, or if required, to restore both case information and punctuation.

We observed that the final punctuation mark is most difficult to be restored. The language model(LM)-based approach can propose two conflicting hypotheses, while the translation-based approach suffers from translation quality. In order to better restore the final punctuation mark, we combine the output of LM and translation-based approaches with a majority voting procedure. With two proposed hypotheses from the LM-based method and one from the translation-based method, we choose the hypothesis using majority voting. If no solution can be found using this approach, we choose the first hypothesis proposed by LM-based method.

3. Experimental Setup

3.1. Data

The experiments were carried out using the datasets provided, extracted from the Basic Travel Expression Corpus (BTEC) [21]. This multilingual speech corpus contains tourism-related sentences similar to those that are usually found in phrasebooks for tourists going abroad. We participated in CHALLENGE, BTEC and PIVOT tasks and covered all language pairs and translation directions in this evaluation campaign. We translated both the single-best hypotheses and the correct recognition results.

For our primary submissions, training was performed using the default training set, to which we added the data from devset1, devset2, devset3 and devset4 for the Chinese–English task,[1] and devset6 was used for development purposes. For the Arabic to English task, training was performed using the default training set, to which we added devset1, devset2, and devset3. For English–Chinese, Chinese–Spanish and Chinese–English–Spanish, training was carried out using the relevant default training sets.

For training the out-of-domain word probabilities, we used the LDC parallel news data for the Arabic–English task. The Hong Kong Parallel Text and LDC Chinese-English Name Entity Lists Version 1.0 were used for Chinese–English and English–Chinese tasks. We also added the HIT corpus as in-domain training data.

For translation-based punctuation and case restoration, we used the English side of the training corpus to train the translation system.

3.2. Tools

As a preprocessing step, the English sentences were tokenized using the maximum entropy-based tokenizer of the OpenNLP toolkit,[2] and case information was removed.

The Arabic data was tokenized and segmented using the ASVM toolkit which is based on SVMs, and has been trained on the Arabic Treebank. [22] The AVSM toolkit tokenized the Arabic data and segmented the Arabic words with the same segmentation style as in the Arabic Treebank.

[1] More specifically, we use the Chinese side of the bilingual sentence pairs together with the first English reference from the 7 references to construct new sentence pairs.

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

The Chinese data was segmented using ICTCLAS Olympic version.[3] For the numbers identified, we split them into characters.

A 5-gram language model[4] with Kneser-Ney smoothing was trained with SRILM [23] on the English side of the training data, and Moses [24] was used to decode.

4. Experimental Results

4.1. Chinese–English Translation

We participated in both CHALLENGE and BTEC tasks for translating Chinese into English. We tried out different word and phrase alignment methods including word packing, syntax-enhanced word alignment and treebank-based phrase extraction. From Table 5, we can see that syntax-enhanced word alignment leads to an 8.08% relative increase over the baseline for the CHALLENGE task and 6.34% for the BTEC task in terms of BLEU score. Compared to generative word alignment models, which require a considerably large training data to achieve good results, our discriminative syntax-enhanced word alignment model can achieve competitively good results with a relatively small amount of training data.

Our best results are achieved with a combination of in-domain and out-of-domain data, 13.96% relative increase over the baseline for the CHALLENGE task and 10.32% for BTEC task in terms of BLEU score.

Word packing is not shown to be helpful in this task. We attribute this to the small size of training data, which results in unsatisfactory word alignments using GIZA++ and consequently inappropriate word groupings are generated. The grouping of words will further worsen the data sparseness problem in the next step of word packing.

For the BTEC task adding the parallel treebank phrase pairs gives a relative improvement of 5.23% BLEU over the baseline system. For the CHALLENGE task, however, we see no significant improvement using this method.

System	CHALLENGE	BTEC
Baseline	0.3194	0.3595
Word Packing	0.2967	0.3522
Syntax-enhanced	0.3452	0.3823
Treebank	0.2881	0.3785
Smoothing for OOV	0.3259	-
Smoothing for ALL	0.3295	-
Data Combination	0.3640	0.3966

Table 5: Impact of various sub-modules in MATREX on IWSLT 2008 Chinese–English CRR (case+punc) translation in terms of BLEU score

[3] http://ictclas.org/index.html

[4] We used a 5-gram language model simply because it achieved better results than using trigram language model on development set.

4.2. English–Chinese Translation

The approaches to word and phrase alignment used in Chinese–English translation can be easily adapted for English–Chinese translation. Our syntax-enhanced word alignment model leads to a 3.99% relative improvement in BLEU over the baseline system. Again, not surprisingly, data combination is an effective approach to improve system performance giving a relative improvement of 13.48% BLEU over the baseline system. For similar reasons as described for Chinese–English translation, word packing is not shown helpful in this task.

Adding parallel treebank phrase pairs harmed translation accuracy significantly when compared to the baseline system. This was surprising as it went against all previous experiments reported using this method. One reason we can offer for this drop in accuracy has to do with the reordering model. Having manually analysed the translation output we see that, although using the treebank phrases provides greater coverage, the segments are quite often in the wrong order. Conversely, the baseline system has relatively good ordering of segments. We can attribute this to the fact that, in both systems, the reordering model only covers the baseline SMT phrase pairs. There is no model to tell us how to reorder the treebank phrase pairs and thus the system that uses them suffers from sub-standard ordering. This issue did not arise with previous translation pairs and directions we evaluated this method with. However, given the difficulty of translating into Chinese, the problem manifests itself here. In order to resolve this issue and exploit the syntax-based phrase pairs to their full potential, especially when translating between highly divergent languages, we need to somehow estimate a distortion model from the parallel treebank.

System	En–Zh
Baseline	0.4080
Word Packing	0.4004
Syntax-enhanced	0.4243
Treebank	0.3773
Data Combination	0.4630

Table 6: Impact of various sub-modules in MATREX on IWSLT 2008 English–Chinese CRR (case+punc) translation in terms of BLEU score

4.3. Chinese–Spanish Translation

In both the direct system and the PIVOT system, adding parallel treebank phrase pairs led to significant improvement over the baseline model. Interestingly, our PIVOT system significantly outperforms our direct system for Chinese–Spanish translation. We attribute this to the relative quality of the English–Spanish module of our PIVOT system. Looking back to Table 1 we see that the En–Es system's translation model is more than twice the size than that of the direct Zh–Es system. This is because the techniques used in our

Data Condition	**CHALLENGE**		**BTEC**			**PIVOT**
	Zh–En	En–Zh	Ar–En	Zh–En	Zh–Es	Zh–En–Es
CRR	0.3640	0.4630	0.4715	0.3966	0.2924	0.3292
ASR output (1-best)	0.3086	0.4022	0.3858	0.3397	0.2670	0.2948

Table 4: Official results on IWSLT 2008 (case+punc) in terms of BLEU score

MaTrEx system can produce high-recall word alignments for English–Spanish, which in turn lead to a system with broader coverage. Translating Chinese into the pivot language, English, allows us to employ this high-coverage module which outperforms direct translation.

System	**BTEC**	**PIVOT**
Baseline	0.2693	0.2832
Treebank	0.2924	0.3292

Table 7: Impact of various sub-modules in MATREX on IWSLT 2008 Chinese–Spanish CRR (case+punc) translation in terms of BLEU score

4.4. Punctuation Restoration

We tried both LM-based approach and translation-based approach to restore punctuations in Chinese. We found that combining these two using a majority voting precedure can improve the system performance further as show in Table 8. It also shows that different punctuation restoration techniques have very limited influence on the final BLEU score.

Approach	**BLEU**
LM-based	0.3171
Translation-based	0.3144
Combined	0.3194

Table 8: Impact of punctuation restoration techniques in MATREX on IWSLT 2008 Chinese–English CRR (case+punc) translation in terms of BLEU score

4.5. Discussion

We employed our new word and phrase alignment techniques for different translation tasks. However, the combination of these approaches has not yet been tested. How to combine these approaches remains a challenging yet promising research direction.

From the results, we can see that systems trained properly on larger amount of training data perform much better than those trained on smaller amount of training data. Our new alignment approaches are only applied on a small amount of training data in current work, and experiments on scaling up these word and phrase alignment techniques are required to further justify the merits of these approaches.

5. Conclusion

In this paper, we described some new extensions to MATREX, the hybrid data-driven MT system developed at DCU. We described word packing, syntax-enhanced word alignment and treebank-based phrase extraction, which are all aimed at improving word and phrasal alignment for MT, as well as their integration into MATREX. Word packing, which had been previously shown to be effective in improving translation quality, did not help in this particular data and system configuration. Treebank-based phrase extraction performed inconsistently, improving translation accuracy in 3 of the 5 tasks in which it was employed. The syntax-enhanced word alignment model consistently improved translation quality across all translation tasks. We also carried out experiments taking advantage of both in-domain and out-of-domain data to solve the pervasive OOV problems and improved translation quality. Finally, we handled the problems of case and punctuation restoration by deploying a majority voting procedure combining a language model-based system and a translation-based system, which we believe improved the quality of the output strings.

6. Acknowledgments

This work is supported by Science Foundation Ireland (grant Nos. 05/RF/CMS064, O5/IN/1732 and 07/CE/I1142) and the Irish Centre for High-End Computing.[5] We would like to thank the reviewers for their insightful comments.

7. References

[1] N. Stroppa and A. Way, "MaTrEx: the DCU machine translation system for IWSLT 2006," in *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 31–36.

[2] H. Hassan, Y. Ma, and A. Way, "MaTrEx: the DCU machine translation system for IWSLT 2007," in *Proceedings of the International Workshop on Spoken Language Translation*, Trento, Italy, 2007, pp. 21–28.

[3] Y. Ma, N. Stroppa, and A. Way, "Bootstrapping word alignment via word packing," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, 2007, pp. 304–311.

[5] http://www.ichec.ie/

[4] Y. Ma, S. Ozdowska, Y. Sun, and A. Way, "Improving word alignment using syntactic dependencies," in *Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, Columbus, OH, 2008, pp. 69–77.

[5] J. Tinsley, V. Zhechev, M. Hearne, and A. Way, "Robust language-pair independent sub-tree alignment," in *Machine Translation Summit XI*, Copenhagen, Denmark, 2007, pp. 467–474.

[6] S. Armstrong, M. Flanagan, Y. Graham, D. Groves, B. Mellebeek, S. Morrissey, N. Stroppa, and A. Way, "MaTrEx: Machine translation using examples," in *TC-STAR OpenLab on Speech Translation*, Trento, Italy, 2006.

[7] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[8] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, 1996, pp. 836–841.

[9] D. Wu, "Stochastic Inversion Transduction Grammars and bilingual parsing of parallel corpora," *Computational Linguistics*, vol. 23, no. 3, pp. 377–403, 1997.

[10] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[11] M. Kitamura and Y. Matsumoto, "Automatic extraction of word sequence correspondences in parallel corpora," in *Proceedings of the 4th Workshop on Very Large Corpora*, Copenhagen, Denmark, 1996, pp. 79–87.

[12] I. D. Melamed, "Models of translational equivalence among words," *Computational Linguistics*, vol. 26, no. 2, pp. 221–249, 2000.

[13] J. Tiedemann, "Combining clues for word alignment," in *Proceedings of the 10th Conference of European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 2003, pp. 339–346.

[14] A. Ittycheriah and S. Roukos, "A maximum entropy word aligner for Arabic-English machine translation," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, 2005, pp. 89–96.

[15] J. Tinsley, M. Hearne, and A. Way, "Exploiting parallel treebanks to improve phrase-based statistical machine translation," in *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07)*, Bergen, Norway, 2007, pp. 175–187.

[16] M. Hearne, S. Ozdowska, and J. Tinsley, "Comparing constituency and dependency representations for smt phrase-extraction," in *15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN '08)*, Avignon, France, 2008.

[17] S. Petrov and D. Klein, "Improved inference for unlexicalized parsing," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, 2007, pp. 404–411.

[18] D. Bikel, "Design of a multi-lingual, parallel-processing statistical parsing engine," in *Human Language Technology Conference (HLT)*, San Diego, CA, 2002.

[19] M. Civit and M. A. Martí, "Building cast3lb: A spanish treebank," *Research on Language and Computation*, vol. 2, no. 4, pp. 549–574, 2004.

[20] Y.-S. Lee, "IBM Arabic-to-English translation for IWSLT 2006," in *Proceedings of IWSLT 2006*, Kyoto, Japan, 2006, pp. 45–52.

[21] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," in *Proceedings of Third International Conference on Language Resources and Evaluation 2002*, Las Palmas, Canary Islands, Spain, 2002, pp. 147–152.

[22] M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic tagging of Arabic text: From raw text to base phrase chunks," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Boston, MA, 2004, pp. 149–152.

[23] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 2002, pp. 901–904.

[24] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180.