MaTrEx: the DCU MT System for NTCIR-8

Tsuyoshi Okita, Jie Jiang, Rejwanul Haque
Hala Al-Maghout, Jinhua Du, Sudip Kumar Naskar
Andy Way
Dublin City University
Center for Next Generation Localisation / School of Computing
{tokita, jjiang, rhaque, halmaghout, jdu, snaskar, away}@computing.dcu.ie

ABSTRACT
This paper gives the system description of the Dublin City University Machine Translation system MaTrEx for our participation in the translation subtask in the NTCIR-8 Patent Translation Task under the team ID of DCUMT. Four techniques are deployed in our systems: supertagged PB-SMT, context-informed PB-SMT, noise reduction, and system combination. For EN-JP, our system stood second in terms of BLEU reference score among six participants.

Categories and Subject Descriptors
I.2.7 [**Computation and Language**]: natural language processing—*machine translation*

General Terms
Languages

Keywords
Machine Translation, Supertag, Noise Reduction, System Combination

1. INTRODUCTION
This paper describes new extensions to the hybrid MT system MaTrEx (Machine Translation using Examples) developed at Dublin City University. We deployed four techniques, under the team ID of DCUMT, in this NTCIR-8 Patent Translation Task [11].

The first technique deployed in our system is to incorporate the target-side supertag information to the MT systems, which has been demonstrated to be effective in [15]. Due to the availability of supertaggers only for English, we apply this technique only in the Japanese to English direction. This technique is challenging to be applied here. Firstly, long distance dependencies in Japanese may not be captured well by incorporating the supertag information in English since this technique is often considered to be useful for better local reorderings. Secondly, we noticed that the quality of parsing outputs of patent data are sometimes error-prone. This reflects the characteristics of patent corpora that differ from the Penn-II treebanks on which English parsers are often trained: notably reference numbers, many parentheses, long sentences, technical terms, and symbols. We use the HPSG supertagger ENJU [24, 31] instead of the CCG supertagger [5] due to its robustness in these areas.

The second technique is to incorporate the source-side supertag information to the MT systems [12], which is applied only on the English to Japanese direction, due to the same reasons outlined above. Compared to European language pairs, this method is also a challenging task since Japanese has some typical characteristics which may not be captured by the source-side context information. Such characteristics include structural ambiguity in syntactic constituency rather than lexical ambiguity; the scrambling phenomenon [14] which can rearrange the order among the constituents of a sentence where the case particles can serve to identify the functions of the accompanying NPs within the sentence, and so forth.

The third technique is to reduce noisy sentences from the training corpus [27], which relies on detecting failures in word alignment. The rationale to apply this technique here is that the patent corpus may include various paraphrases, multi-word expressions and non- literal translations, which were the original motivation for using this technique. However, the challenge here is to see whether the method works under the phenomenon that error-prone word alignments are also produced by long and complex sentences, and whether it works when a few sentences are removed.

The fourth technique is to combine several translation outputs [1, 23, 8] via the MBR decoder [20]. This technique has become one of the standard techniques in various recent MT evaluation campaigns, and it is shown to often empirically improve the quality of MT.

The remainder of this paper is organized as follows. Section 2 describes four components which are extended in our systems; two syntax-based modules which incorporate supertagged information in the PB-SMT (Phrase-Based SMT) systems, the filtering module for noisy sentences, and the system combination module. In Section 3, our experimental setup is explained. In Section 4, our experimental results are presented. We conclude in Section 5.

2. THE MATREX SYSTEM

The MaTrEx system is a hybrid data-driven MT system which exploits aspects of different MT paradigms [30]. The system follows a modular design and facilitates to plug in different MT engines and novel techniques equipped with a system combination technique which enables us to combine different MT outputs [8]. In the following subsections, we describe the main techniques used in the participation of NTCIR-8 [11].

2.1 Supertagged PB-SMT: JP → EN

We employ supertaggers to enrich the English side of the parallel training corpus with a single supertag sequence per sentence. Then we extract phrase pairs together with the co-occurring English supertag sequence from this corpus via the same phrase extraction method used in the baseline model. This way we directly extend the baseline model with supertags both in the phrase translation table and in the language model. We define the probabilistic model that accompanies this syntactic enrichment of the baseline model.

Let t and s be the target and source language sentences respectively. Any (target or source) sentence x will consist of two parts: a bag of elements (word phrases etc.) and an order over that bag. In other words, $x =< \phi_x, O_x >$, where ϕ_x stands for the bag of phrases that constitute x, and O_x for the order of the phrases as given in x. Hence, we may separate order from content.

$$\begin{aligned}&\arg\max_t P(t|s)\\ =\;& \arg\max_t P(s|t)P(t)\\ =\;& \arg\max_{<\phi_t, O_t>} P(\phi_s|\phi_t)P(O_s|O_t)P_w(t)\end{aligned}$$

where $P_w(t)$ is the target-language model, $P(O_s|O_t)$ represents the conditional (order) linear distortion probability, and $P(\phi_s|\phi_t)$ stands for a probabilistic translation model from target-language bags of phrases to source-language bags of phrases using a phrase translation table.

Let ST represent a supertag sequence of the same length as a target sentence t. Hence this equation becomes

$$\begin{aligned}&\arg\max_t \sum_{ST} P(s|t, ST)P_{ST}(t, ST) \approx\\ &\arg\max_{<t,ST>} P(\phi_s|\phi_t, ST)P(O_s|O_t)^{\lambda_o}\\ &P_{ST}(t, ST)\exp^{|t|\lambda_w}.\end{aligned}$$

The approximations made in this formula are of two kinds: the standard split into components and the search for the most likely joint probability of a target hypothesis and a supertag sequence co-occurring with the source sentence (a kind of Viterbi approach to avoid the complex optimization involving the sum over supertag sequences). The distortion and word penalty models are the same as those used in the baseline PB-SMT model.

The language model $P_{ST}(t, ST)$ is a supertagger assigning probabilities to sequences of word-supertag pairs. The language model is further smoothed by log-linear interpolation with the baseline language model over word sequences.

The supertagged phrase translation probability consists of a combination of supertagged components analogous to their counterparts in the baseline model, i.e. it consists of $P(s|t, ST)$, its reverse and a word-level probability. We smooth this probability by log-linear interpolation with the factored back-off version $P(s|t)P(s|ST)$, where we import the baseline phrase table probability and exploit the probability of a source phrase given the target supertag sequence. A model in which we omit $P(s|ST)$ turns out to be slightly less optimal than this one.

As in most state-of-the-art PB-SMT systems, we use GIZA++ to obtain word-level alignments in both language directions. The bidirectional word alignment is used to obtain lexical phrase translation pairs using heuristics presented in [26] and [19]. Given the collected phrase pairs, we estimate the phrase translation probability distribution by relative frequency as follows:

$$\hat{P}_{ph}(s|t) = \frac{count(s, t)}{\sum_s count(s, t)}$$

For each extracted lexical phrase pair, we extract the corresponding supertagged phrase pairs from the supertagged target sequence in the training corpus. For each lexical phrase pair, there is at least one corresponding supertagged phrase pair. The probability of the supertagged phrase pair is estimated by relative frequency as follows:

$$P_{ST}(s|t, st) = \frac{count(s, t, st)}{\sum_s count(s, t, st)}$$

2.2 Context-Informed PB-SMT: EN → JP

One natural way to express a context-informed feature ($\hat{h}_m$) is to view the conditional probability $P(\hat{e}_k|\hat{f}_k)$ conditioned on its context information (CI) where ($\hat{e}_k$) denotes the target phrases and ($\hat{f}_k$) denotes the source phrase. That is

$$\hat{h}_m(\hat{f}_k, CI(\hat{f}_k), \hat{e}_k, s_k) = logP(\hat{e}_k|\hat{f}_k, CI(\hat{f}_k)).$$

Here, CI may include any feature (lexical, syntactic, etc.), which can provide useful information to disambiguate the given source phrase ($\hat{f}_k$). Beside supertags [12], we also replicated the experiments [29] by considering the context words and part-of-speech features. Like dependency information [13], supertags capture the long-range word-to-word dependency in a sentence and provide enough evidence to disambiguate the source phrase. The lexical and syntactic features used in our experiments are described in the following subsections.

Lexical Features. These features include the direct left and right context words of length l (resp. $f_{i_k-l}...f_{i_k-1}$ and $f_{j_k+1}...f_{j_k+l}$) of a given focus phrase $\hat{f}_k = f_{i_k}...f_{j_k}$. It forms a window of size $2l$ features. Thus lexical contextual information (CI_{lex}) can be described as follows:

$$CI_{lex}(\hat{f}_k) = \{f_{i_k-l}, ..., f_{i_k-1}, f_{j_k+1}, ..., f_{j_k+l}\}$$

Syntactic Features (Part-of-Speech tag). In addition to the lexical features, it is possible to exploit several knowledge sources characterizing the context. For example, we can consider the part-of-speech of the focus phrase and of the context words. In our model, the POS tag of a multi-word focus phrase is the concatenation of the POS tags of the words composing that phrase. Thus a window of size $2l+1$ features is formed including the POS tag of focus phrase. Thus contextual information (CI_{pos}) defining part-of-speech features is described as follows:

$$\begin{aligned} CI_{pos}(\hat{f}_k) \quad &= \{pos(f_{i_k-1}), ..., pos(f_{i_k-l}), \\ &pos(\hat{f}_k), pos(f_{j_k+1}), ..., pos(f_{j_k+l})\} \end{aligned}$$

Syntactic Features (Supertags). As well as using local words and POS-tags as features, as in [29], we incorporate supertags as a syntactic source context feature in the log-linear model of PB-SMT. In our experiments two kinds of supertags are employed: those from Lexicalized Tree-Adjoining Grammar (LTAG: [2]) and Combinatory Categorial Grammar (CCG: [6]). Both the LTAG [3] and the CCG supertag sets [16] were acquired from the WSJ section of the Penn-II Treebank using hand-built extraction rules. Here we test both the LTAG and CCG supertaggers. The contextual information (CI_{st}) for supertag is defined as

$$\begin{aligned} CI_{st}(\hat{f}_k) \quad &= \{st(f_{i_k-1}), ..., st(f_{i_k-l}), st(\hat{f}_k), \\ &st(f_{j_k+1}), ..., st(f_{j_k+l})\} \end{aligned}$$

Similar to the CI_{pos} feature, the supertag of a multi-word focus phrase is the concatenation of the supertags of the words composing that phrase. Thus the supertag-based CI forms a window of size $2l+1$ features including the supertag of the focus phrase. In our experiments, we used ± 1 and ± 2 lexical and syntactic features (i.e. $l = 1, 2$).

We refer the interested reader to [29] and [12] for more details of how memory-based features are integrated in the log-linear MT framework of Moses.

2.3 Noise Reduction

Given that the amount of training data available for these tasks is limited, developing techniques to make the best use of them is essential for the performance of the MT systems. We used a technique to improve the translation model by differentiating "good" and "bad" data, where our *good points algorithm* selects high quality parallel sentence pairs in the training data to build MT systems. This leads to better word alignments since this process can remove noisy sentence pairs (also called outliers) from training data. Given that state-of-the-art word alignment models only allows 1-to-n mappings between source and target words, those sentences which include n-to-m mappings between source and target words (for example, paraphrases, non-literal translations, and multiword expressions) are considered to be noise. The noisy sentence pairs can potentially hinder a word aligner in achieving high quality alignments; moreover, the errors in word alignment will be propagated in later stages of MT training including phrase extraction. To remove the noisy sentence pairs, we use a method as shown in Algorithm 1 [27].

Algorithm 1 Good Points Algorithm

Step 1: Train WB-SMT (Word-Based SMT) using the whole training data, and translate all the sentences in the training data to output n-best lists.
Step 2: For the n-best translations for each source sentence, obtain the (maximum) cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$ by comparing each translation against the reference target sentence. This score is used measure the quality of the current sentence pair.
Step 3: Train PB-SMT using the whole training data. Translate all training sentences to output n-best lists.
Step 4: For the n-best translations for each source sentence, obtain the (maximum) cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{PB,X}$ by comparing each translation against the reference target sentence. This score is also used measure the quality of the current sentence pair.
Step 5: Remove sentence pairs where $S_{WB,2} = 0$ and $S_{PB,2} = 0$, and sentence length is greater than 2.
Step 6: The remaining sentence pairs after removal in Step 5 are used to train the final PB-SMT systems.

2.4 Multiple System Combination

Multiple system combination [8] is deployed to combine the outputs from three different prototype Statistical Machine Translation systems, namely PB-SMT and HBP-SMT (Hierarchical Phrase-Based SMT).

For multiple system combination, we implement an Minimum Bayes-Risk-Confusion Network (MBR-CN) framework as used in [8]. Due to the varying word order in the MT hypotheses, it is essential to decide the backbone which determines the general word order of the confusion network. Instead of using a single system output as the skeleton, we employ a MBR decoder to select the best single system output from the merged n-best list by minimizing the BLEU loss, as follows:

$$\hat{e}_i = \underset{i \in \{1, \cdots, N\}}{\arg\min} \sum_{j=1}^{N} \{1 - BLEU(e_j, e_i)\}$$

where e_i and e_j are hypotheses in the n-best list, and N indicates the number of hypotheses in the merged n-best list. $BLEU(e_j, e_i)$ calculates sentence-level BLEU score of e_i with e_j as th e reference translation.

The confusion network is built using the output of MBR decoder as the backbone which determines the word order of the combination. The other hypotheses are aligned against the backbone based on the TER metric. NULL words are allowed in the alignment. Either votes or some form of confidence measures are assigned to each word in the network. Each arc in the CN represents an alternative word at that position in the sentence and the number of votes for each word is counted when constructing the network. The features we used are as follows:

- word posterior probability [10]

Systems	BLEU	#OOV
System combination	27.61*	321
HPB-SMT 1	26.86*	314
PB-SMT 1	26.51*	194
Noise reduction (PB-SMT)	24.01	443
PB-SMT 2+	23.91*	316
Preprocessing (PB-SMT)+	23.82	194
HPB-SMT 2	23.30	303
Supertag (ENJU) 1	20.68	430
Supertag (ENJU) 2	18.27	426
System combination (unofficial run)	28.43	331

Table 1: Intrinsic evaluation results (JP-EN). The HPB-SMT 1 bases on the chart-based Moses and the HPB-SMT 2 bases on joshua. The PB-SMT 1 bases on Moses with the distortion limit 12 over 600k training corpus, while the PB-SMT 2 bases on Moses with the distortion limit 6 over 3,200k training corpus. The supertag (ENJU) 1 uses the configuration of MERT process tuned with one factor, while the supertag (ENJU) 2 uses the configuration of MERT process tuned with two factors. It is noted that the official BLEU scores which have asterisk in their shoulder are evaluated after the removal of OOV words. It is noted that we trained over 3,200k training corpus for the systems marked with + and over 600k training corpus for other systems.

- trigram and 4-gram target language model
- word length penalty
- NULL word length penalty

Minimum Error-Rate Training (MERT) is used to tune the weights of the confusion network.

3. EXPERIMENTAL SETUP

Two open-source SMT systems, PB-SMT and chart-based SMT system Moses [18] and HPB-SMT system Joshua [21], are used in our experiments.

The baseline in our experiments is a standard log-linear PB-SMT system based on Moses. The GIZA++ implementation [26] of IBM Model 4 is used as the baseline for word alignment: Model 4 is incrementally trained by performing 5 iterations of Model 1, 5 iterations of HMM, 3 iterations of Model 3, and 3 iterations of Model 4. For phrase extraction the grow-diag-final heuristics described in [19] is used to derive the refined alignment from bidirectional alignments. We then perform MERT process [25] which optimizes the BLEU metric, while a 5-gram language model is derived with Kneser-Ney smoothing [17] trained with SRILM [28] on the English side of the training data. We use Moses for decoding.

For the HPB-SMT system, we use two systems: the chart-based decoder of Moses [18] and that of Joshua [21]. Most of the procedures are identical with the PB-SMT systems except the rule extraction process [4].

Additionally, we use the factored model of Moses for the first and the second technique, and we use memory-based classifiers [7] for the second technique.

Systems	BLEU
System combination	33.03
HPB-SMT 1	32.50
PB-SMT 1	30.53
PB-SMT 2+	30.08
Noise reduction	29.53
Preprocessing (PB-SMT)+	27.93
HPB-SMT 2	27.23
Context supertag (Base)	26.83
Context supertag (Superpair)	26.45
Context supertag (CCG)	26.38
Context supertag (LTAG)	26.38
Context supertag (CCG-LTAG)	26.22
Context supertag (POS)	26.21

Table 2: Intrinsic evaluation results (EN-JP). The HPB-SMT 1 bases on the chart-based Moses and the HPB-SMT 2 bases on joshua. The PB-SMT 1 bases on Moses with the distortion limit 12 over 600k training corpus, while the PB-SMT 2 bases on Moses with the distortion limit 6 over 3,200k training corpus. It is noted that we trained over 3,200k training corpus for the systems marked with + and over 600k training corpus for other systems.

In our experiments, we used data provided within the NTCIR-8 Patent Translation Task; no additional data resources we used.

4. EXPERIMENTAL RESULTS

We report the experimental results obtained on the techniques described above on three different systems, PB-SMT, HPB-SMT and WB-SMT, where WB-SMT is used internally in noise reduction technique. System combination results are finally obtained based on the single-best translation outputs. For extrinsic evaluation, we used the same four systems which we used for intrinsic evaluation.

Intrinsic Evaluation. The results for JP-EN and EN-JP directions are shown in Tables 1 and 2 respectively. The scores for the primary runs are underlined.

For the JP-EN direction, the primary run is the output of system combination of four systems (HPB-SMT 1 and 2, PB-SMT 1 and 2). This was the fourth-best scoring system among seven participants. After the submission, we conducted another run of system combination combining seven systems (HPB-SMT 1 and 2, PB-SMT 1 and 2, noise reduction, supertag (ENJU) 1, and preprocessing) that scored 28.43 BLEU points. It is to be noted that due to our internal human communication problems when we run the experiments, several runs were submitted after the removal of OOV words (Out Of Vocabulary words) from the translation outputs (indicated by asterisk in Table 1) and others were not.

For EN-JP direction, the primary run is the output of the system combination of four systems (HPB-SMT 1 and 2,

Systems	BLEU	MAP	r@100	r@200	r@500	r@1000
PB-SMT 1$^+$	24.00	0.21	0.55	0.63	0.72	0.78
HPB-SMT 2	23.71	0.18	0.53	0.59	0.68	0.73
HPB-SMT 1	23.48	0.18	0.53	0.59	0.68	0.74
PB-SMT with preprocessing$^+$	22.35	0.21	0.55	0.64	0.70	0.76

Table 3: Extrinsic evaluation results. The column shows the evaluated measure whether it is BLEU, MAP (Mean Average Precision) or Recall@N (which is abbreviated in a table as r@N). It is noted that we trained over 3,200k training corpus for the systems marked with $^+$ and over 600k training corpus for other systems.

PB-SMT 1 and 2). This run stands second among six participants. Six runs of the context-informed SMT, shown in the lower six rows in Table 2, were worse than the other seven runs. This is due to the poor quality of CCG and LTAG parser when confronted with the patent data. It is to be noted that although we used the ENJU parser for the JP-EN direction, we used the CCG / LTAG parser for the EN-JP direction. However, at the same time, we observed from our translation outputs that function words seem to be often translated differently. We might need to handle structural ambiguity as well if we want to cope well with lexical ambiguity in the case of Japanese.

Extrinsic Evaluation. Each group was requested to machine translate the search topics and the retrieval was performed by the organizers using the translated search topics [11]. Examples of such models include technology survey, in which patents related to a specific technology are searched for, and invalidity search, in which prior art related to a patent application are searched for.

Two evaluation measures for CLPR (Cross-Lingual Patent Retrieval) were additionally used where these evaluation measures discard the information about the word order: Mean Average Precision (MAP) and Recall@N which stands for Recall for the top N documents. Okapi BM25 is used as the retrieval model where documents are sorted according to the score and are retrieved up to the top 1000 documents for each topic.

Table 3 shows our extrinsic evaluation results where PB-SMT performs consistently better than HPB-SMT across all the measures, namely BLEU, MAP and r@N measures. This might indicate that the key patent terminologies are better retrieved by PB-SMT than by HBP-SMT probably because of its constituency nature.

5. CONCLUSIONS

In this paper we described four new techniques deployed in the Dublin City University MaTrEx system under the team ID of DCUMT: supertagged PB-SMT, context-informed PB-SMT, noise reduction, and system combination.

We showed that the system combination strategy is effective in both EN-JP and JP-EN directions even though we combine only four 1-best translation outputs. Improvements were 0.75 BLEU points absolute and 2.8 % relative for JP-EN (1.57 BLEU points absolute and 1.6 % relative for unofficial run which combined seven 1-best outputs) and 0.53 BLEU points absolute and 1.6 % relative for EN-JP compared to the single best outputs. Both improvements were statistically significant at the 5 % level.

Supertagged PB-SMT and context-informed PB-SMT seem to have difficulties probably due to the typical characteristics of Japanese we mentioned in section 1, since both methods are not originally intended for Japanese. On the one hand, difficulties in supertagged PB-SMT may be attributed to the better reordering model which can cope with long distance dependencies in Japanese. On the other hand, the main difficulties in context-informed PB-SMT may be ascribed to the additional components which can handle structural disambiguity in Japanese as well as lexical disambiguity.

There are several avenues for further work. Firstly, the NTCIR patent corpora include various equations, parentheses, and symbols. Since these components are often handled by preprocessing methods in a deterministic manner, this issue looks trivial at first sight for an MT system. However, to choose the correct preprocessing methods and to apply them to the corpora often determines the final quality of the translations. Since this process is often ad hoc and time-consuming, a more organized approach may be required for data similar to the NTCIR patent corpora. Secondly, Table 1 shows that there were quite a few OOV words untranslated in the translation outputs. The word lattice-based decoding approach [9, 22] may reduce this number.

6. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`http://www.cngl.ie`) at Dublin City University. We would also like to thank the Irish Centre for High-End Computing.

7. REFERENCES

[1] S. Bangalore, G. Bordel, and G. Riccardi. Computing consensus translation from multiple machine translation systems. *In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 350–354, 2001.

[2] S. Bangalore and A. Joshi. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265, 1999.

[3] J. Chen, S. Bangalore, and K. Vijay-Shanker. Automated extraction of Tree-Adjoining Grammars from treebanks. *Natural Language Engineering*, 12(3):251–299, 2006.

[4] D. Chiang. A hierarchical phrase-based model for statistical machine translation. *In Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 263–270, 2005.

[5] S. Clark. Supertagging for combinatory categorical grammar. *In Proceedings of the International Workshop on Tree Adjoining Grammars*, pages 19–24, 2002.

[6] S. Clark and J. Curran. The importance of supertagging for wide-coverage CCG parsing. *In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 282–288, 2004.

[7] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. TiMBL: Tilburg memory based learner, version 6.1, reference guide. *ILK Research Group Technical Report Series, No. 07-07*, 2007.

[8] J. Du, Y. He, S. Penkale, and A. Way. MaTrEx: the DCU MT System for WMT 2009. *In Proceedings of the Third EACL Workshop on Statistical Machine Translation*, pages 95–99, 2009.

[9] C. Dyer, S. Muresan, and P. Resnik. Generalizing word lattice translation. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, pages 1012–1020, 2008.

[10] J. Fiscus. A post-processing system to yield reduced Word Error Rates: Recogniser output voting error reduction (ROVER). *In Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–352, 1997.

[11] A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, T. Ehara, H. Echizen-ya, and S. Shimohata. Overview of the patent translation task at the NTCIR-8 workshop. *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, 2010.

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

[13] R. Haque, S. K. Naskar, A. van den Bosch, and A. Way. Dependency relations as source context in phrase-based SMT. *In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pages 170–179, 2009.

[14] S. Harada. Nihongo ni henkei wa hitsuyoo da. *Gengo 6*, pages 88–103, 1977.

[15] H. Hassan, K. Sima'an, and A. Way. Integrating supertags into phrase-based statistical machine translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 288–295, 2007.

[16] J. Hockenmaier. Data and models for statistical parsing with combinatory categorial grammar. *Ph.D Thesis, University of Edinburgh, UK*, 2003.

[17] R. Kneser and H. Ney. Improved backing-off for n-gram language modeling. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184, 1995.

[18] 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*, pages 177–180, 2007.

[19] P. Koehn, F. Och, and D. Marcu. Statistical phrase-based translation. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computationa Linguistics (HLT / NAACL 2003)*, pages 115–124, 2003.

[20] S. Kumar and W. Byrne. Minimum Bayes-Risk word alignment of bilingual texts. *In Proceedings of EMNLP 2002*, pages 140–147, 2002.

[21] Z. Li, C. Callison-Burch, C. Dyer, S. Khudanpur, L. Schwartz, W. Thornton, J. Weese, and O. Zaidan. Joshua: An open source toolkit for parsing-based machine translation. *In Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, 2009.

[22] Y. Ma, T. Okita, O. Cetinoglu, J. Du, and A. Way. Low-resource Machine Translation using MaTrEx: the DCU Machine Translation system for IWSLT 2009. *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2009)*, pages 29–36, 2009.

[23] E. Matusov, N. Ueffing, and H.Ney. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. *In Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–40, 2006.

[24] Y. Miyao and J. Tsujii. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80, 2008.

[25] F. Och. Minimum Error Rate Training in Statistical Machine Translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 2003.

[26] F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[27] T. Okita. Data cleaning for word alignment. *In Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009) Student Research Workshop*, pages 72–80, 2009.

[28] A. Stolcke. SRILM – An extensible language modeling toolkit. *In Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, 2002.

[29] N. Stroppa, A. van den Bosch, and A. Way. Exploiting source similarity for SMT using context-informed features. *In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation*, pages 231–240, 2007.

[30] N. Stroppa and A. Way. MaTrEx: DCU machine translation system for IWSLT 2006. *In Proceedings of the International Workshop on Spoken Language Translation*, pages 31–36, 2006.

[31] M. Takuya, Y. Miyao, and J. Tsujii. Efficient HPSG parsing with supertagging and CFG-filtering. *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1671–1676, 2007.