BBN's Systems for the Chinese-English Sub-task of the NTCIR-10 PatentMT Evaluation

Zhongqiang Huang
Raytheon BBN Technologies
50 Moulton St, Cambridge, MA
USA
zhuang@bbn.com

Jacob Devlin
Raytheon BBN Technologies
50 Moulton St, Cambridge, MA
USA
jdevlin@bbn.com

Spyros Matsoukas
Raytheon BBN Technologies
50 Moulton St, Cambridge, MA
USA
smatsouk@bbn.com

ABSTRACT

This paper describes the systems we developed at Raytheon BBN Technologies for the Chinese-English sub-task of the Patent Machine Translation Task (PatentMT) of the NTCIR-10 workshop. Our systems were originally built for translating newswire articles and were subsequently adapted to address some special problems of patent documents in the NTCIR-9 PatentMT evaluation. We applied some of our recent advancements in translation to the patent domain and investigated a sentence-level language model adaptation approach to take advantage of the characteristics of patent documents. These approaches contributed substantially to the improvement of translation quality and our systems achieved the best results among all submissions across all of the evaluation types and evaluation metrics.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence] Natural Language Processing - Machine Translation.

General Terms

Algorithms, Performance, Experimentation

Keywords

Machine Translation, Patent Translation

Team Name

BBN

Subtasks

Chinese to English

External Resources Used

ADSO dictionary, LDC96L15 (Chinese word segmentation lexicon)

1. INTRODUCTION

In this paper, we describe the statistical machine translation (SMT) systems developed at Raytheon BBN Technologies for the Chinese-English sub-task of the Patent Machine Translation Task (PatentMT) of the NTCIR-10 workshop [5]. We have been developing SMT systems based on the string-to-dependency translation model [16] for a variety of languages and genres and have achieved superior performance in the DARPA GALE evaluations and the recent BOLT MT evaluation. In the previous NTCIR-9 PatentMT evaluation [6], we adapted our newswire system to the patent domain to better handle some special problems in patent documents. As demonstrated in the evaluation results, these methods were effective in achieving high quality patent translation. In the current NTCIR PatentMT evaluation [5], we applied some of our recent advancements in translation to the patent domain and investigated a sentence-level language model (LM) adaptation approach to take advantage of the characteristics of patent documents. Specially, our patent MT systems benefited from the following new features and improved component models:

- Miscellaneous features (bigram lexical translation probabilities, trait features, etc.)
- Sentence-level LM adaptation
- Robust context dependent translation
- Recurrent neural network LM
- Translation-based true caser

The rest of the paper is organized as follows. We will first introduce our translation framework in Section 2 and describe the improvements we made to our systems to better handle patent documents in the NTCIR-9 PatentMT evaluation. We will then describe the new features and improved component models that were applied in the current NTCIR-10 PatentMT evaluation. Finally, we will present the evaluation results in Section 5 and conclude this paper in Section 6.

2. TRANSLATION FRAMEWORK

Our SMT systems are based on a string-to-dependency translation model [16] that employs hierarchical rules to translate strings in the source language to dependency trees in the target language. In addition to about 10 to 20 regular features [15] that involve rule and lexical translation probabilities, language model scores, etc., our systems also utilize a large number of sparse features (about 50K in total), in a way similar to the methods reported in [3, 2], with feature weights trained discriminatively to maximize expected BLEU [14]. The discriminative features come from 8 categories [3]:

1. Does the rule contain the target phrase X?

2. Does the rule translate word X to word Y?
3. Does the rule translate POS X to POS Y?
4. Was this rule seen exactly once in the training?
5. Do the two non-terminals in source switch position in the target?
6. Does the source word X align to exactly two target words?
7. How often was the lexical source-target pair (X, Y) seen in the training corpus Z?
8. Is the target non-terminal X filled by the target non-terminal Y?

Our MT systems use a target trigram language model to generate n-best hypotheses and then re-score the n-best hypotheses with a target 5-gram language model. The n-gram probabilities in the language models were smoothed using modified Kneser-Ney smoothing [1]. GIZA++ [12] is used to train word alignment models.

3. NTCIR-9 SYSTEMS

The fact that patent documents are juridical documents and thus the presence of well-structured sentences and less ambiguity of word meanings makes patent documents easier to translate for MT; however, some characteristics, such as the abundance of long and complicated sentence structures, as well as technical terminology and new terms that are originally defined by patent applicants, make patent documents challenging for MT. We developed a variety of techniques to adapt our newswire system to address some of the special challenges in patent documents. The resulting patent MT systems achieved top performance in the NTCIR-9 PatentMT evaluation. These strong systems served as the initial systems based on which we developed new systems for the NTCIR-10 PatentMT evaluation. For the completeness of this paper, we briefly describe in this section our systems built for the NTCIR-9 PatentMT evaluation. Please refer to our NTCIR-9 system description paper [10] for details. The improvements in mixed-case BLEU obtained by various methods are summarized in Table 1, where we use the sign "+" to indicate changes applied on top of the system shown in the preceding row.

System	Test
Baseline with 45M LM	34.01
+ consistent tokenization	34.56
+ more token sharing	34.97
+ patent case-LM	36.47
+ optimized word segmenter	36.95
+ top 100 features	37.71
+ 14B LM	39.14
+ LM adaptation	40.04

Table 1: Improvements (in mixed-case BLEU) obtained by various methods in NTCIR-9 PatentMT evaluation

3.1 Preparation

The training and development data released by the organizers of NTCIR-9 PatentMT evaluation [6] includes a parallel training corpus of one million (1M) Chinese-English sentence pairs with a total of roughly 45 million (45M) words, a monolingual English patent corpus of US patent documents with a total of 14 billion (14B) words, and a development data set of two thousand (2K) bilingual sentence pairs. We split this development set into two subsets, one for tuning MT systems and one for measuring performance. In order to make the two subsets similar in terms of translation difficulty, we first translated these 2K sentences with our newswire MT system, and then split the patent documents in the development set into two subsets, roughly half-and-half, of approximately equal translation error rate (TER). By this procedure, we ended up with 1039 sentences in the tuning set and 961 sentences in the test set.

We trained the translation model on the patent parallel corpus. We also trained two English LMs, one trained with only the 45M English words from the parallel corpus and the other one with the 45M words plus the 14B monolingual English corpus. We denote the former one as 45M LM and the latter one as 14B LM.

In order to save time to explore the best strategies to build patent MT systems, unless specified otherwise, we used the smaller 45M LM and only the regular features (not including the 50k sparse features) for most of the experiments. Our initial patent MT system trained under this setup has a mixed-case BLEU of 34.01 on the test set and it serves as the baseline in Table 1.

3.2 Consistent tokenization and special token sharing

We found that Chinese patent documents contain significantly more special ASCII strings[1] than Chinese newswire articles and many of these strings in the parallel corpus were not aligned properly due to inconsistency between tokenization of ASCII strings on the source and target sides[2]. In order to remove such inconsistency, we tokenized ASCII strings in Chinese sentences in the same way as in English sentences. This improved BLEU from 34.01 of the baseline system to 34.56 as shown in Table 1.

We also extended a special token sharing mechanism that were originally developed for numbers in newswire articles to 4 additional types of special tokens, namely patent identification numbers, name abbreviations, numbers with labels, and math expressions, due to their abundance in the patent data. This mechanism replaces each type of special tokens with a common token before training word alignment and language models, as well as before translating test sentences, and replacing each common token back to the original tokens after translation. The sharing of special tokens in patent documents further improves BLEU to 34.97 as shown in Table 1.

3.3 Re-training casing LM

We originally cased MT outputs based on a trigram casing

[1]Strings written in ASCII characters, e.g., English words, patent numbers, mathematical expressions and abbreviation names for materials.

[2]For example, the ASCII string "IS-1000" was tokenized as itself when occurring in the Chinese sentences but tokenized as "IS - 1000" when occurring in the English sentences.

LM that was trained on a collection of normal English text with mixed cases. Since patent text has significantly different characteristics from newswire text, we re-trained the casing LM with the mixed-case English sentences from the patent parallel corpus. This significantly improves BLEU by 1.50 to 36.47.

3.4 Optimizing Chinese word segmenter

Our Chinese word segmenter used a simple left-to-right and longest-match-first algorithm based on a Chinese lexicon. The lexicon used in the previous experiments is a subset of a big Chinese word lexicon[3] that was optimized for MT performance on newswire. The iterative optimization procedure starts with the big lexicon and gradually removes words from the lexicon that are not aligned well in the GIZA++ alignment of the parallel training data until the performance of the new MT system trained with the reduced lexicon stops improving. We re-optimized the lexicon on the patent parallel corpus, starting from the combination of the lexicon optimized for newswire and a set of words extracted from the ADSO dictionary[4]. This lexicon optimization procedure improves BLEU to 36.95 as shown in Table 1.

3.5 Using more features

We then added the 50K sparse features described in Section 2 to the system. The feature weights were trained discriminatively to maximize the expected BLEU [14]. Due to the small size of the tuning set, the addition of the 50K features resulted in a big improvement on the tuning set but only a small gain on the test set. In order to alleviate the over-fitting problem, we reduced the number of sparse features by selecting only the top 100 features based on the tuned weights of the 50K features and re-trained the system with the regular features plus these 100 sparse features. The addition of the top 100 features provides another significant improvement of 0.78 in BLEU as shown in Table 1.

3.6 LM adaptation

We also adopted an LM adaptation approach, similar to [17], to interpolate a general LM with an LM estimated from some text data that is closely related to the document being translated. The related text data was acquired through a cross-lingual information retrieval (CLIR) technique. For each test document, we used our own CLIR tool [18] to extract related patent documents in the same monolingual English patent corpus that was included to train the 14B LM, and selected the most relevant documents as the related text for LM adaptation. We use the term bias LM to refer to the LM estimated from the CLIR-retrieved text. While translating a sentence s in the test document, we compute LM probability according to Equation 1:

$$\mathrm{p}_{\mathrm{LM}}(s) = (1-\alpha)\mathrm{p}_{\mathrm{generalLM}}(s) + \alpha \mathrm{p}_{\mathrm{biasLM}}(s) \qquad (1)$$

where α is a document-dependent interpolation weight that is estimated automatically.

In order to have a fair comparison to evaluate the effectiveness of this approach, we re-trained the system to switch from the 45M LM to the 14B LM. The use of the 14B LM significantly improves BLEU from 37.71 to 39.14, as shown in the line labeled as "+ 14B LM" in Table 1. On top of the system with the 14B LM, the LM adaptation approach contributes another significant improvement of 0.9 in BLEU, achieving a final BLEU score of 40.04.

[3]It consists of words from the Chinese word lexicon released by LDC (LDC96L15) and words acquired from a few web sites.

[4]The dictionary is publicly available at www.adsotrans.com. We extracted 10k words from the ADSO dictionary that are not in the big Chinese lexicon.

4. NTCIR-10 SYSTEMS

We built the systems for the NTCIR-10 PatentMT evaluation on top of the systems that were built for the NTCIR-9 evaluation. We applied some of our recent advancements in translation to the patent domain and investigated a sentence-level LM adaptation approach to take advantage of the characteristics of patent documents. Table 2 summarizes the improvements in mixed-case BLEU obtained by applying these new methods on top of our NTCIR-9 systems. Compared to the NTCIR-9 systems, we achieved significant BLEU improvements of 2.31 and 2.09 absolute over the systems with the 45M LM and the 14B LM, respectively. We will next describe these methods.

System	Test
NTCIR-9 system with 45M LM	37.71
+ miscellaneous features	38.06
+ robust context dependent translation	38.72
+ recurrent neural network LM	39.35
+ translation-based true caser	40.02
NTCIR-9 system with 14B LM (no bias LM)	39.14
+ miscellaneous features	39.51
~~+ document-level LM adaptation~~	~~39.94~~
+ sentence-level LM adaptation[5]	40.95
+ robust context dependent translation	41.09
+ recurrent neural network LM	41.43
+ translation-based true caser	42.13

Table 2: Improvements (in mixed-case BLEU) obtained by various methods in NTCIR-10 PatentMT evaluation

4.1 Miscellaneous features

The miscellaneous features include two sets of new features and various tweaks to our system. The first set of features is an extension of context-based lexical probabilities to model the joint likelihood of every target bigram given their aligned source words and source contexts. The second set of features models MT hypotheses using traits [4], which are high-level characteristics such as the percentage of source context words that are not translated, percentage of source terminals/non-terminals that cross alignment links inside the corresponding decoding rules, and the average number of target words per rule. The traits were used as stand-alone features in this evaluation, although they can also be used to generate diverse hypotheses that could be further combined through a consensus network. Please refer to [4] for details of the trait features.

[5]The sentence-level LM adaptation was applied on top of the "+ miscellaneous features" system. The document-level LM adaptation shown crossed-out in the table is for comparison only and was not used to build our NTCIR-10 system.

In our NTCIR-9 systems, the values of each feature were modeled using a Gaussian distribution and the feature values were normalized based on the mean and variance of the distribution before they were optimized. Optimization with normalized features was found to converge faster and better fit the tuning set; however, feature normalization also caused increased risk of over-fitting. We disabled feature normalization in our NTCIR-10 systems and re-evaluated the feature selection procedure in Subsection 3.5. We observed that it is still advantageous to use the top 100 features for the system trained with the 45M LM but it is better to use all of the features for the system trained with the 14B LM.

As shown in Table 2, the addition of miscellaneous features contributes 0.35 and 0.37 in BLEU to the MT systems trained with the 45M LM and the 14B LM, respectively.

4.2 Sentence-level LM adaptation

We also applied the same LM adaptation procedure described in Subsection 3.6 to our NTCIR-10 system trained with the 14B LM and obtained a decent improvement of 0.43 in BLEU. The CLIR-based LM adaptation method uses individual Chinese patent documents as queries and retrieves the most relevant documents in the monolingual English patent corpus as related text for LM adaptation. The granularity of this LM adaptation approach is at the document level, and it is possible that the most relevant text at the document level may not be the most relevant for individual sentences in the document.

As we discussed earlier, when compared to newswire articles, sentences in patent documents are well-structured and the word meanings are less ambiguous. This is partly exemplified by the fact that patent documents tend to re-use n-grams that were used in other patent documents, as shown in Table 3 where we list the percentage of source n-grams, up to length 8, in the development set that are also observed in the corresponding parallel training data for the newswire domain in GALE (227M words in training) and the patent domain in the NTCIR-10 PatentMT evaluation. Take 4-grams for example, only 5.4% of the 4-grams (tokens) in the newswire domain are observed in training, while as many as 21% of the 4-grams in the patent domain are observed in training despite the much smaller size of the training data.

N-gram order	Newswire		Patent	
	Type	Token	Type	Token
1	0.83	0.95	0.81	0.97
2	0.55	0.68	0.78	0.83
3	0.20	0.24	0.46	0.49
4	0.047	0.054	0.19	0.21
5	0.011	0.012	0.075	0.083
6	0.0035	0.0039	0.036	0.039
7	0.0014	0.0016	0.020	0.021
8	0.0006	0.0007	0.012	0.013

Table 3: Percentage of source n-grams (measured by type or token) in the development set that are also observed in the parallel training set for GALE and NTCIR-10 PatentMT evaluation

What's more interesting is to look at the coverage of target n-grams. Since the size of the monolingual English patent corpus is a lot larger than the size of the patent parallel training corpus, a much higher percentage of the target n-grams in the development set are observed in the monolingual English patent corpus, as shown in Table 4 where we compare the parallel training corpus and the monolingual English corpus in terms of their coverage of target n-grams in the development set. In order to take advantage of the high coverage of target n-grams in the monolingual English patent corpus, we investigated a targeted LM adaptation approach at the sentence level. Instead of using individual documents as queries, as we did for the NTCIR-9 PatentMT evaluation, we used individual sentences as queries and searched for the most relevant passages, not documents, in the monolingual English patent corpus, where passages are defined as overlapping segments of a patent document with roughly 300 words in each segment [17]. The retrieved passages were used as related text for LM adaptation in Equation 1.

As shown in Table 2, sentence-level LM adaptation significantly improves the system without LM adaptation by 1.44 in BLEU, which is 1.01 higher than the improvement obtained by document-level LM adaptation.

N-gram order	Parallel corpus		English corpus	
	Type	Token	Type	Token
1	0.97	0.99	0.98	0.99
2	0.87	0.91	0.95	0.97
3	0.61	0.65	0.85	0.86
4	0.34	0.37	0.65	0.66
5	0.18	0.19	0.42	0.43
6	0.10	0.11	0.26	0.26
7	0.068	0.073	0.16	0.16
8	0.052	0.056	0.10	0.11

Table 4: Percentage of target n-grams (measured by type or token) in the development set of the NTCIR-10 PatentMT evaluation that are also observed in the patent parallel corpus and the monolingual English patent corpus.

4.3 Robust context dependent translation

Our systems employed multiple context-based lexical translation and distortion models, such as the probability of a target translation given the source word and its context and its extension to the joint probability of every target bigram as mentioned in Subsection 4.1. Central to these models is a linear interpolation method to alleviate data sparsity:

$$\mathrm{p}(e|c_1, \cdots, c_m) = \sum_i w_i \mathrm{p}(e|C_i) \quad (2)$$

where e is an event to predict, c_i's are conditionals in the context, C_i's are ordered back-off contexts, called components, based on heuristics, and w_i is the interpolation weight of backoff model $\mathrm{p}(e|C_i)$ whose probabilities are unsmoothed maximum likelihood estimates (MLE).

The problem is that, unlike language modeling, there is no clear ordering as to how this backoff should be best performed. A fixed set of interpolation weights is also suboptimal as it does not take into account of the reliability of individual contexts. In order to address these problems, we developed a framework called robust context dependent translation that considers many possible ways[6] of backing

[6]We divide the conditionals of a context into essential condi-

off the context and interpolate all of the backoff models together using optimized weights that depend on the reliability of the backoff components:

$$\mathrm{p}(e|c_1,\cdots,c_m) = \sum_i \frac{\delta_i + \gamma_i \log(\mathrm{N}(C_i))}{\sum_j \delta_j + \gamma_j \log(\mathrm{N}(C_j))} \mathrm{p}(e|C_i) \quad (3)$$

where C_i's can be any backoff components and do not need to be ordered, $\delta_i + \gamma_i \log(\mathrm{N}(C_i))$ is the weight of the back-off model with parameters δ_i and γ_i to be optimized on a held-out set, and $\mathrm{N}(C_i)$ is the count of component C_i in the training data. This model supports the intuition that the more time a component has been observed, the more reliable the corresponding back-off model is, but the increase in reliability quickly drops off once it is been seen a sufficient number of times.

Note that although we consider many possible ways of backing off the context, many of the backoff models may not contribute much useful information. In addition, it is resource intensive to store all of these backoff models. To address this problem, we used a greedy algorithm to add backoff models one at a time and prune the ones that are least effective on the held-out set. As shown in Table 2, this approach contributes an improvement of 0.66 in BLEU to the system with the 45M LM and an improvement of 0.14 in BLEU to the system with the 14B LM.

4.4 Recurrent Neural Network LM

We also investigated recurrent neural network LM [11] for language modeling in MT. In a recurrent neural network LM, the history of each word effectively includes all of the previous words in the sentence, rather than a fixed window of size n in a conventional n-gram LM. This is achieved by using the hidden layer from the $(i-1)$-th word as part of the input layer for the i-th word. We trained a recurrent neural network LM on the 45M LM training data[7] and interpolated it with a 5-gram LM model trained on the same data. The same model was used for both the MT systems with the 45M LM and the 14B LM. The scores computed from this model were used as an additional feature in n-best rescoring of MT hypotheses. As shown in Table 2, the recurrent neural network LM contributes 0.57 and 0.34,respectively, in BLEU on the systems with the 45M LM and the 14B LM.

4.5 Translation-based true caser

Up to this point we have been using a 3-gram casing LM for case restoration. The casing LM was trained on the 45M LM training data for the NTCIR-9 PatentMT evaluation. Similar to [7], we developed a new casing model that treats casing as a MT problem by translating a lower case sentence (the source) to a sentence of true cases (the target).

We extracted casing rules from the parallel data. In addition to rule probabilities and the true-case LM probability, we added several sparse feature types. These include positional features, such as *Is the target word upper cased and does it follow a period?*, and part-of-speech features, such as *Is the target word upper cased and a proper noun?*. The new true caser improves the two systems with the 45M LM and the 14B LM by 0.67 and 0.70 in BLEU, respectively, achieving a final mixed-case BLEU of 40.02 with the 45M LM and 42.13 with the 14B LM.

tionals and additional conditionals. We consider all possible ways of backing off for essential conditional but only consider additional conditions independently on top of all of the essential conditionals.

[7]We did not try to train a recurrent neural network LM on the 14B LM training data because training such a model is prohibitively computationally expensive.

5. EVALUATION RESULTS

Four types of evaluations were conducted at the NTCIR-10 Chinese-English PatentMT evaluation: Intrinsic Evaluation (IE), Patent Examination Evaluation (PEE), Chronological Evaluation (ChE), and Multilingual Evaluation (ME). Please refer to [5] for detailed descriptions of these evaluations. Our primary system, labeled[8] as BBN-1, for all of the evaluation types was trained on the provided 45M parallel training corpus and the 14B monolingual English patent corpus, following the procedures described in Sections 3 and 4 except that the system was tuned on the entire development set, not just the tuning subset. As requested from the organizers, we also trained a secondary system, labeled as BBN-2, in a similar way, except using just the 45M parallel training corpus and the 45M LM training data extracted from the target side, for intrinsic evaluation (IE). Our systems achieved the best performance among all submissions across all of the evaluation types and evaluation metrics.

System	IE	ChE	ME
BBN-1	42.68	39.44→41.09	27.62
BBN-2	39.98	36.69→38.93	N/A
BASELINE1	32.52	30.74→30.74	17.96
BASELINE2	31.34	29.34→29.34	18.05

Table 5: Automatic scoring results (BLEU) of the intrinsic (IE), chronological (ChE) and multilingual (ME) evaluations. The right arrow → in the ChE column indicates the change in BLEU from the NTCIR-9 evaluation to the NTCIR-10 evaluation.

Table 5 shows the automatic scoring results (BLEU) of our two systems[9] as well as two baseline systems[10] provided by the organizers for three evaluation types: intrinsic (IE), chronological (ChE), and multilingual (ME) evaluations. Both of our systems produced significantly better performance than the baseline systems in intrinsic evaluation and our primary and secondary systems gained an improvement of 1.65 and 2.24, respectively, in BLEU in the chronological evaluation (ChE) compared to our NTCIR-9 systems. Our primary system also achieved a significantly higher BLEU score than the baseline systems as well as other submissions (see [5] for details) in the multilingual evaluation (ME), in which a MT system translates manual Chinese translations of Japanese patent documents into English[11].

[8]The subtask name and evaluation type need to be included in the name of official submissions; they are left out in this paper for simplicity.

[9]The BBN-2 system was not officially evaluated in the chronological evaluation (ChE); we measured the performance of the BBN-2 system on the chronological evaluation (ChE) set by ourselves and included the scores in the table for comparison.

[10]BASELINE1 is Moses' hierarchical phrase-based SMT system [8] and BASELINE2 is Moses' phrase-based SMT system [9].

[11]We did not participate in the Japanese-English subtask of

Besides automatic evaluation, the organizers also carried out manual evaluation of the submissions to measure the adequacy and acceptability of the translations of 300 sentences in the intrinsic evaluation set by trained annotators. Each translation was manually examined and assigned a 5-level adequacy score, 5, 4, 3, 2, or 1, from the best to the worse, and also a 5-level acceptability score, AA, A, B, C, or F, from the best to the worst [5]. Table 6 shows the adequacy scores of our system BBN-1 and the baseline systems. As shown, our system produced significantly better translations in terms of adequacy, with more than 50% of the translations receiving the highest adequacy score and a average adequacy score of 4.15. Table 7 shows the allocation of acceptability scores of our system BBN-1 and its pairwise acceptability score [5] that is computed based on the percentage of wins and ties when comparing the acceptability score our system output with other submissions. A high pairwise score of 0.69 means that our system was on average much better than any other submission in terms of acceptability.

System	Average adequacy	Allocation of scores				
		5	4	3	2	1
BBN-1	4.15	156	66	44	34	0
BASELINE1	3.23	46	73	91	84	6
BASELINE2	2.82	38	34	75	141	12

Table 6: Manual adequacy scores of the intrinsic evaluation (IE)

System	Pairwise score	Allocation of scores				
		AA	A	B	C	F
BBN-1	0.69	81	36	50	35	98

Table 7: Manual acceptability scores of the intrinsic evaluation (IE)

The evaluation organizers also evaluated the utility of machine translation for patent examination. In patent examination evaluation (PEE) for the Chinese-English subtask, reference Japanese patents that were used to reject real patent applications were first manually translated to Chinese, which were then machine translated by the participating systems. The translation outputs were rated by two experienced patent examiners using a 6-level score, S, A, B, C, D, or F, from best to worse, based on the percentage of important facts that can be recognized from the translated text to reject the original patent application. As shown in Table 8, despite a relatively low percentage of translated documents was judged as perfect (rated S), a large portion of facts that are important for patent examination were recognized from the translation output of our BBN-1 system and the translation results were considered useful in general for patent examination[12].

the evaluation. The baseline systems and several other submissions, despite their lower performance in Chinese-English translation in the multilingual evaluation (ME), achieved higher BLEU scores than ours by directly translating the Japanese documents into English.

[12]Translated documents were rated A (or B) if at least half (or one) of the important facts were recognized and the translation results were considered useful for examination[5].

System	Allocation of scores					
	S	A	B	C	D	F
BBN-1	6	19.5	3.5	0	0	0

Table 8: Allocation of scores of the patent examination evaluation (PEE) averaged from the two patent examiners

6. CONCLUSION

We have described the work we carried out for building SMT systems for the Chinese-English sub-task of the NTCIR-10 PatentMT evaluation. Our systems were originally built for newswire and were subsequently adapted to address some special problems of patent documents in the NTCIR-9 PatentMT evaluation. We applied some of our recent advancements in translation, such as robust context dependent translation and recurrent neural network LM, to the patent domain and investigated a sentence-level LM adaptation approach to take advantage of the characteristics of patent documents. These approaches contributed substantial gain to our patent MT systems and helped to achieve promising results in patent translation.

7. REFERENCES

[1] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University, 1998.

[2] D. Chiang, K. Knight, and W. Wang. 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009.

[3] J. Devlin. Lexical features for statistical machine translation. Master's thesis, University of Maryland, Colleg Park, 2009.

[4] J. Devlin and S. Matsoukas. Trait-based hypothesis selection for machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012.

[5] I. Goto, K. P. Chow, B. Lu, E. Sumita, and B. K. Tsou. Overview of the patent machine translation task at the NTCIR-10 workshop. In *Proceeding of the 10th NTCIR Workshop Meeting on Evauation and Information Access Technologies: Information Retrieval, Question Answering, and Cross Lingual Information Access*, 2013.

[6] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceeding of the 9th NTCIR Workshop Meeting on Evauation and Information Access Technologies: Information Retrieval, Question Answering, and Cross Lingual Information Access*, 2011.

[7] 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*, 2006.

[8] H. Hoang, P. Koehn, , and A. Lopez. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, 2009.

[9] 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 *Proceeding of Annual Meeting of the Association for Computational Linguistics*, 2007.

[10] J. Ma and S. Matsoukas. BBN's systems for the chinese-english sub-task of ntcir-9 patentmt evaluation. In *Proceeding of the 9th NTCIR Workshop Meeting on Evauation and Information Access Technologies: Information Retrieval, Question Answering, and Cross Lingual Information Access*, 2011.

[11] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, 2010.

[12] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computional Linguistics*, 2003.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, 2002.

[14] A.-V. I. Rosti, B. Zhang, S. Matsoukas, and R. Schwartz. Bbn system description for wmt10 system combination task. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, 2010.

[15] L. Shen, J. Xu, and R. Weischedel. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2008.

[16] L. Shen, J. Xu, and R. Weischedel. String-to-dependency statistical machine translation. *Computational Linguistics*, 2010.

[17] M. Snover, N. Madnani, B. Dorr, and R. Schwartz. Language and translation model adaptation using comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[18] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.