BBN's Systems for the Chinese-English Sub-task of the NTCIR-9 PatentMT Evaluation

Jeff Ma
BBN Technologies
50 Moulton Street, Cambridge,
MA 01741, USA
1-617-873-3600
jma@bbn.com

Spyros Matsoukas
BBN Technologies
50 Moulton Street, Cambridge
MA 01741, USA
1-617-873-3600
smatsouk@bbn.com

ABSTRACT
This paper describes the work we conducted for building a statistical machine translation (SMT) system for the Chinese-English sub-task of the NTCIR-9 patent machine translation (MT) evaluation [17]. We first applied the various techniques on patent data that we had developed for improving SMT performance on other types of data. Our results show that most of the techniques work on patent document translation as well. Second we made changes to our SMT system training in order to address special characteristics of patent documents. The changes produced additional improvements.

Categories and Subject Descriptors
I.2.7 [Artificial Intelligence] Natural Language Processing – machine translation.

General Terms
Algorithms, Performance, Experimentation

Keywords
TeamName: [BBN]

Subtasks/Languagues: [Chinese-to-English patent MT]

External Resources Used: [Giza++, LDC96L15, BBN's CLIR tool, ADSO dictionary]

1. INTRODUCTION
In this paper we describe the efforts we took to build MT systems for the Chinese-English sub-task of the NTCIR-9 patent MT evaluation. We have been building SMT systems based on the string-to-dependency translation model [11], which employs hierarchical rules to translate strings in the source language to dependency trees in the target language. The details about this model and its implementation can be found in [11]. Recently we improved our SMT models with two techniques: use of a large number (50,000) of features, similar to the method reported in [18, 2], and discriminative training of feature weights to maximize the expected BLEU [9]. Since the expected BLEU criterion is continuous and differentiable, gradient descent may be performed, thus supporting weight tuning for a large number of features. The use of 50,000 features yielded gains similar to those reported in [2]. We use GIZA++ [6] for training word alignment models. Since all the above MT model training methods have been published before, we do not elaborate them but focus on the changes we made for building patent MT systems in this paper.

All the systems we built for different language and genres with the above method yielded superior performance in the GALE MT evaluations[1]. So, for the NTCIR-9 evaluation we first tried to build MT systems with the same method on the patent data released by the NTCIR-9.

Patent documents are juridical documents, which are typically more structured than general documents, and they have their own special characteristics. People tried to utilize these special characteristics in various applications, such as categorization of patent documents in [3, 5], and machine translation of patent documents in [4, 7, 10]. Some of the patent document characteristics make MT easier, for example, the presence of well-structured sentences and less ambiguity of word meanings. On the other hand, some characteristics become challenges for MT, for example, long and complicated sentence structures, technical terminology and new terms that are originally defined by patent applicants. Due to these challenges people have explored various strategies for improving patent MT quality, such as combining SMT with rule-based MT in [4, 14, 15], with promising results. Therefore, the second thing we did was to make changes and implement new techniques for our SMT systems to handle special characteristics of patent data better.

The paper is organized as follows: Section 2 presents preparation work we did for building the patent MT systems; Section 3 descries the training of the SMT systems and reports incremental gains from a few methods we implemented specially for patent MT; and Section 4 shows the NTCIR-9 evaluation results of our systems.

2. Preparation
We have built various Chinese-English SMT systems before this NTCIR-9 evaluation. One of them is a "newswire" MT system for translating Chinese newswire text. We used this system to help set up a development set and a test set for building our patent SMT systems.

2.1 The Chinese-English newswire MT system
This "newswire" MT system was trained on a parallel training corpus that includes 227 million (227M) words, the majority of which is newswire text. The collections in this corpus had been

[1] http://www.itl.nist.gov/iad/mig/tests/gale/index.html

released by the Linguistic Data Consortium (LDC) for the DARPA GALE project. Our MT system uses a tri-gram target language model (LM) to generate n-best hypotheses and then ranks the n-best with a 5-gram LM. We build n-gram LMs with the modified Kneser-Ney smoothing [1]. For the "newswire" MT system we trained the target (English) LMs on an English corpus that consists of more than 6 billion (6B) words of news text, out of which 227M words come from the English side of the Chinese-English parallel corpus, 2.2 billion words from the 4th edition of the LDC English Gigaword monolingual data release, 2.5 billion words from Google news and 1.6 billion words from news text that we downloaded from various websites, such as BBC, Xinhua News, and The Arab News. We denote this LM as "6B-nw-LM". This "newswire" MT system produced a BLEU [8] score of 26.22 on the GALE Phase 4 Chinese newswire evaluation test set that includes 490 sentences – with one reference translation per sentence.

2.2 Patent training and test data

For building the patent MT system, we used the data released by the NTCIR-9 evaluation organizers for the Chinese-English sub-task of the NTCIR-9 patent MT evaluation [17]. The released data includes a parallel training corpus that consists of one million (1M) Chinese-English sentence pairs and a development data set that consists of two thousand (2K) bilingual sentence pairs.

The total number of English words in the 1M parallel sentence pairs was close to 45 million (45M). The 2K sentence pairs in the development set were extracted from 103 patents. The full contents of the 103 patents, including titles, abstracts and full descriptions, were also provided in both Chinese and English languages. As the evaluation organizers stated, people could use the full contents – referred to as "context data" – for building MT systems. We explored using this context data for adapting target LMs. We split this development set into two subsets, one for tuning the MT system and one for measuring the performance. To make the two subsets similar in terms of translation difficulty, we first translated these 2K sentences with the "newswire" MT system, and then split the 103 patent documents into two subsets, roughly half-and-half, based on their translation error rate (TER) [13], resulting in two subsets of approximately equal TER scores. With this splitting, we ended up with 1039 sentences from 54 patents in the tuning set – denoted as "Tune" in this paper – and 961 sentences from 49 patents in the test set – denoted as "Test". The mixed-case BLEU scores measured on the whole 2K development set, the Tune, and the Test sets are listed in Table 1. The scores on the Tune and Test are close.

Table 1. BLEU scores measured on the 2K development set, the Tune, and the Test sets using the "newswire" MT system

Data set	2K-dev	Tune	Test
BLUE	15.38	15.87	14.77

Recall that the "newswire" system produced a BLEU score of 26.22 on the newswire evaluation test set. However, on the 2K patent sentences it performed significantly worse – a BLEU score of 15.38. We re-tuned the decoding parameters for the "newswire" MT system with the new Tune set and the re-tuning only improved the BLEU on the Test set slightly –from 14.77 to 15.64. This implies that this big performance degradation mainly resulted from mismatches between the training and test data.

Besides the parallel training corpus and the development data set, the NTCIR-9 committee also released a monolingual English patent corpus for the purpose of training English LMs. This corpus includes US patent documents published in the period 1993-2005, totaling 14 billion (14B) words. We used this corpus for training our English LMs.

Since we focused on the mixed-case performance in our work, we will report only the mixed-case BLEU scores measured on the Test set for all experiments shown below, unless specified otherwise.

3. Building patent MT systems

3.1 Training the MT system

We first re-trained the MT model with the 45M word patent parallel corpus. Before training the word alignment models, we segmented words in the Chinese sentences with a 52K lexicon by using a left-to-right and longest-match-first algorithm, which generated 41 million (41M) Chinese words in the 1M sentences.

We trained two sets of English LMs. One was trained with only the 45M English words from the 1M parallel corpus and the other one with the 45M words plus the 14B monolingual English corpus. We denote the former one as "45M-pt" LM and the latter one as "14B-pt" LM.

We looked into effects of the three different LMs, "6B-nw", "45M-pt" and "14B-pt", on the MT performance. Mixed-case BLEU scores of the re-trained MT model with the three LMs are listed in Table 2.

Table 2. Effect (BLEU scores) of the three LMs when used with the patent SMT system

MT model	227M newswire	45M patent	45M patent	45M patent
LM	6B-nw	6B-nw	45M-pt	14B-pt
Test	14.77	30.71	34.01	36.16

As can be seen, the use of the 14B monolingual patent data in the LM training helped improve the performance by about 2 BLEU points (from 34.01 to 36.16).

In the above set of experiments we used only the regular features (not including the 50K features). The main reason was to save time for exploring the best strategies to build the patent MT system. For the same reason we also used the smaller LM – "45M-pt" – in many of our following investigation experiments. Unless specified otherwise, experiments reported later in this paper used the "45M-pt" LM and the regular features. Hence, the system trained with the 45M parallel patent data (the 4th column in Table 2) serves as a baseline for our later efforts to improve the patent MT performance.

3.2 Addressing special characteristics of patent data

We first saw that, compared to the Chinese newswire text data, the Chinese patent text includes significantly more special strings that are not written in Chinese characters, such as English words, patent numbers, mathematical expressions and abbreviation names for materials. This is one characteristic of the patent data. Since all the special strings are written in ASCII characters, we call them ASCII strings. We found many of the ASCII strings occurring in the 45M word patent parallel corpus were not aligned properly during the word alignment training. The main reason was inconsistent tokenization of the ASCII strings on the source and target sides. 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. To remove such inconsistency we tokenized the ASCII strings in the Chinese sentences in the same way as we tokenized the English sentences. The system trained with this consistent tokenization of ASCII strings is denoted as "+ consistent tokenization" in Table 3, where we use the sign "+" to indicate changes applied on top of the system shown in the preceding row. Compared to the "baseline", the consistent tokenization of the ASCII strings improved the performance by about half a BLEU point.

The second thing we did was to increase sharing of translation rules and LM n-gram scores among certain types of special tokens. When training Chinese-English MT systems, we let "infrequent" numbers – all numbers except numbers in the range 1-31 – share translation rules and LM n-gram scores. The sharing mechanism is as follows:

1. Train word alignment models after replacing "infrequent" numbers on both sides of the parallel corpus with a special "number token"
2. Train LMs after applying the same number replacement on the LM training corpus
3. Before translating test sentences, conduct the same number replacement on test sentences and save replacement information that includes the original numbers and their places in the sentences
4. After translating the test sentences, replace the special number tokens occurring in the MT hypotheses with their corresponding original numbers based on the number replacement information and source-to-target word alignment information that the MT decoder outputs during the translation

This sharing mechanism improves our MT performance. Because there are more special tokens in the patent data, such as patent identification numbers and mathematical expressions, we applied the translation rule and n-gram sharing mechanism on 4 more special tokens:

1. patent identification numbers – all 7-digit whole numbers, such as 5,716,812 and 5869649
2. name abbreviations – ASCII strings occurring in the Chinese sentences that consist of only English characters and digits, such as "PMMA" and "CO2"
3. numbers with labels – numbers followed with the commonly-used unit labels, such as "1.03ml" and "20.8g"
4. math expressions – items that consists of any of the math signs, such as "x=0.25" and "a+b"

We applied this special token rule and n-gram score sharing on top of the "+ *consistent tokenization*" system, this new system is denoted as "+ *more sharing*" in Table 3. As can be seen, the rule and LM n-gram sharing on the 4 special tokens produced a 0.4 gain on the BLEU.

Table 3. Improvements (on BLEU) from addressing patent data related issues

System	Test
Baseline	34.01
+ *consistent tokenization*	34.56
+ *more sharing*	34.97
+ *patent case-LM*	36.47
+ *optimized word segmentor*	36.95

3.3 Re-training the casing LM

We case our MT outputs with a tri-gram LM that is trained with mix-cased English text. The casing algorithm searches among all the possible casing combination of the words in a sentence for the path that has the highest likelihood against the tri-gram casing LM. Our initial casing LM was trained on the mixed-case version of the newswire 6B LM training corpus. As shown before, the newswire data differs significantly from the patent data in terms of the data characteristics. Therefore, we re-trained the casing LM with the mixed-case English sentences from the 45M patent parallel corpus. This new casing LM improved the mixed-case BLEU score by 1.5 points, as shown in the row "+ *patent case-LM*" in Table 3.

We then trained another casing LM with the 14B US patent corpus (mixed-case version) added to the 45M words from the patent parallel corpus. Our results showed that the casing LM trained with the augmented data degraded the mixed-case BLEU score slightly. So we used the casing LM trained with the 45M words in all our following experiments.

3.4 Optimizing the Chinese word segmentor

In the experiments we have reported so far, we segmented the Chinese words with a 52K Chinese word lexicon by using a simple left-to-right and longest-match-first algorithm. The 52K lexicon is an optimized subset of a big Chinese word lexicon that includes 121K entries[2]. Our lexicon optimization procedure starts with a big lexicon and gradually removes words from the lexicon that are not aligned well – by measuring if the removal improves the MT performance. The procedure is as follows:

1. Segment Chinese words in the parallel corpus with an initial big word lexicon
2. Train word alignments and measure the MT performance on a test set
3. Remove from the lexicon any words that are aligned less than "Threshold" times

[2] It consists of the words from the Chinese word lexicon released by LDC (LDC96L15) and words we acquired from a few websites

4. Segment Chinese words in the parallel corpus with the reduced lexicon
5. Train a new word alignment model
6. Measure MT performance with the new word alignment model
7. If the performance gets improved, go to Step 3 with an increased value for the "Threshold". Otherwise, stop.

On the 227M Chinese-English parallel corpus we started with the 121K word lexicon and ran a few iterations of the optimization by increasing the "Threshold" value gradually – from 5 to 10 to 20 and to 30. We obtained the best MT performance when the "Threshold" was set to 20 and the lexicon size was reduced to 52K.

We ran the same lexicon optimization on the 45M patent parallel corpus, but starting with a 62K lexicon that includes the 52K lexicon and 10K new words we extracted from the ADSO word translation lexicon[3]. By increasing the threshold from 2 to 3 to 4 – much smaller values due to the significantly less amount of training data – we got the best MT performance when the lexicon was reduced to 32K (at the threshold = 3). The BLEU score on the Test set with the initial 62K lexicon was 36.07% and was increased to 36.95% with the optimized 32K lexicon. The performance of the system that used this optimized 32K lexicon to segment the Chinese words is shown in the row "+ *optimized word segmentor*" of Table 3. As can be seen, this lexicon optimization improved the MT performance by 0.5 BLEU points, compared to the system that used the 52K lexicon.

3.5 Using more features

We then added more features to the system. As mentioned before, the total number of new features we extracted was about 50,000 (50K). These features came from 8 feature categories [18]:

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?

Over-fitting is a well-known problem for tuning weights for a large number of features. We discriminatively trained feature weights to maximize the expected BLEU by using the same technique as reported in [9]. The expected BLEU was computed with the same formula as the BLEU computation in [8], but the n-gram counts and matches are expected versions that are derived from n-best hypotheses.

Table 4. System performance (BLEU scores) with different numbers of new features added

System	Tune	Test
optimized word segmentor	38.66	36.95
add 50K features	44.57	37.38
add top-100 best features	42.82	37.71

After adding the 50K features, we noticed that the gap between the BLEU scores on the Tune and Test sets got significantly large. As shown in the row "*add 50K features*" in Table 4, the gap between the Tune and Test sets was 7 BLEU points, which is much larger than that we observed when we were conducting the same tuning for the system trained on the 227M parallel corpus. So the over-fitting problem got worse in the tuning here because of the small size of the tuning set. To alleviate the over-fitting we tried to reduce the number of the new features. Based on the tuned weights for the 50K features, we selected the top-100 ones that had the highest weights and then added only these 100 features to the system. This experiment is shown in the row "*add top-100 best features*" of Table 4. As shown, the use of the top-100 best features alleviated the over-fitting and the performance on the Test set improved. Compared to the baseline – the "*optimized word segmentor*" system, the use of the 100 extra features helped produce 0.8 BLEU gain.

3.6 LM adaptation

For MT we adopted an LM adaptation approach, similar to [12], that interpolates a general LM with an LM estimated from text data closely related to the test document that is being translated. We acquire the related text data through the cross-lingual information retrieval (CLIR) technique. This LM adaptation has helped improve performance for most of our MT systems. We use the term "bias LM" to refer to the LM estimated from the CLIR-retrieved text. While translating a test document, we compute log LM scores according to,

$$\log(score_{LM}) = \log(score_{generalLM} + \alpha * score_{biasLM}) \quad \text{(Eqn.1)}$$

where "*generalLM*" denotes the general LM and "*biasLM*" the bias LM. "α" is an interpolation weight that is document-dependent and automatically estimated. For a test document – d, the adaptation procedure is as follows:

1. treat the document as a query and run a CLIR tool to extract *N* related passages from a large monolingual text corpus in the target language
2. compute the mean, $\mu(d)$, and standard deviation, $\sigma(d)$, of the CLIR scores of all the *N* passages

[3] The latest release (v5.077) of the ADSO dictionary consists of 185K entries, which is free to the public (http://www.adsotrans.com/downloads). The dictionary includes many phrasal translations. We extracted entries that have only a single word on the English side and treated the tokens on the Chinese side as Chinese words, and then we selected 10K words that are not in the 121K Chinese lexicon.

3. select passages whose CLIR scores are higher than $\mu(d) + T * \sigma(d)$
4. train a bias LM with the selected passages
5. estimate the interpolation weight, $\alpha(d)$
6. compute log LM scores according to (Eqn. 1)

The CLIR tool we used in Step 1 is the one presented in [16], where details of the CLIR score computation can be found. In Step 3, "T" is a threshold that controls the selection. In Step 4, when estimating the bias LM, we applied higher weighting on n-grams counted from more closely related passages. The weighting factor for n-gram counts from a selected passage is computed according to

$$psg\_wgt = \frac{CLIR\_scr(p) - \mu(d)}{\arg\max_{\{p' \in selected\_psgs\}} CLIR\_scr(p') - \mu(d)}$$

where "*CLIR_scr(p)*" represents the CLIR score of a selected passage.

In Step 5 we estimated the interpolation weight according to

$$\alpha(d) = \frac{\mu_s(d) - \mu(d)}{\arg\max_{\{d' \in all\_test\_docs\}} \mu_s(d') - \mu(d')}$$

where $\mu_s(d)$ represents the means of the CLIR scores of all the selected passages. The weight is normalized to be between 0 and 1.

We used our own CLIR tool to extract related passages. We found that it was a good choice to set N = 4,000 in Step 1 and set the selection threshold "T" to 1.28 in Step 3. The related passages for training the bias LM were extracted from the 14B English patent document corpus. In our case here passages are equivalent to patent documents in the English patent corpus.

To have the right baseline system, we re-ran the "*add top-100 best features*" system, shown in Table 4, but switching from the "45M-pt" LM to the "14B-pt" LM. This system is denoted as "*Add top-100 best features + 14B-pt LM*" in Table 5. Comparing these two systems, we see that the "14B-pt" LM improved the MT performance by 1.4 BLEU points (from 37.71 to 39.14). We then conducted the LM adaptation on top of the "*Add top-100 best features + 14B-pt LM*" system. The performance is shown in the "*+ LM adaptation*" row in Table 5. The LM adaptation improved the BLEU score by 0.9 points (from 39.14 to 40.04).

Table 5. . Improvements (in terms of the BLEU score) from the LM adaptation

System	Test
Add top-100 best features + 14B-pt LM	39.14
+ LM adaptation	40.04
LM adaptation (patent description)	39.97
LM adaptation (patent abstract)	40.23

As described earlier, the 2K sentences of the development set were extracted from the descriptions of 103 patents, so the patent documents in the Test set are only portions of the corresponding original patent documents. In the above LM adaptation we used the portions of the full patent descriptions as queries for the CLIR. Since the NTCIR-9 organizers also provided the full contents of the 103 patent documents, we explored uses of the abstracts and the full descriptions of the patent documents as queries for the CLIR in the LM adaptation. The scores of these two experiments are listed in the last two rows in Table 5. As shown, the uses of the abstracts and full descriptions in the LM adaptation produced similar results.

3.7 Augmenting the parallel training data

We tried to extract related documents from the 227M newswire parallel corpus to augment the patent parallel training corpus. Using documents in the 45M patent parallel training as queries, we ran our monolingual information retrieval (IR) tool to extract highly related documents from the newswire parallel corpus. We selected a 17M word subset of the 227M newswire corpus that is highly related to the patent MT corpus. Adding this 17M newswire parallel data to the training of the patent MT model hurt the performance (the BLEU score degraded from 34.97 to 34.41).

4. Evaluation results

For the NTCIR-9 Chinese-English patent MT evaluation we submitted two systems. The primary system, named as "BBN-1", was the "*+ LM adaptation*" system shown in Table 5. The secondary system, named as "BBN-2", was the "*add top-100 best features*" system shown in Table 4. The secondary system was trained only with the 1M parallel corpus, which was the "core" system the NTCIR-9 asked each participant to submit. The differences between the primary and the secondary system were the use of the 14B US patent data in the LM training and LM adaptation. As shown, the use of the 14B US patent data produced a gain of 2.3 BLEU points (37.71 vs. 40.04) on the Test set.

The NTCIR-9 Chinese-English patent MT evaluation test consisted of 2,000 sentences. Our two systems, "BBN-1" and "BBN-2", produced the best performance. Table 6 shows the automatic evaluation results (BLEU scores) of our two systems and two baseline systems that the NTCIR-9 organizers provided. The "Baseline1" system was the Moses phrase-based hierarchical SMT system and the "Baseline2" the Moses phrase-based SMT system. As can be seen, our systems produced significantly better performance. The performance difference between our two systems is 2.8 BLEU points, which is similar to that was observed on the Test set.

Table 6. Automatic evaluation results (BLEU scores) of the Chinese-English sub-task of NTCIR-9 patent MT evaluation

System	BBN-1	BBN-2	Baseline1	Baseline2
BLEU	39.44	36.64	30.72	29.32

Besides the automatic evaluation, the evaluation organizers also carried out manual evaluation of the systems – measuring the adequacy and acceptability of the translations by annotators. Translations were manually measured with 5 levels of adequacy, 1, 2, 3, 4 and 5, from the worst to the best adequacy and with also 5 levels of acceptability, AA, A, B, C and F, from the best to the worst acceptability [17]. Table 7 shows the adequacy scores of our system "BBN-1" and the baseline system "Baseline1" and Table 8 the acceptability scores. As shown, our system "BBN-1" also produced significantly better translations in terms of both the adequacy and acceptability, compared to the baseline system.

Table 7. Manual evaluation results (Adequacy) of the Chinese-English sub-task of NTCIR-9 patent MT evaluation

system	Average adequacy	Distribution of the adequacy scores				
		5	4	3	2	1
BBN-1	4.03	119	91	72	17	1
Baseline1	3.29	48	64	122	59	7

Table 8. Manual evaluation results (acceptability) of the Chinese-English sub-task of NTCIR-9 patent MT evaluation

System	AA	A	B	C	F
BBN-1	43	78	67	51	61
Baseline1	14	29	45	54	158

5. Conclusion

We have described the work we carried out for building an SMT system for the Chinese-English patent MT sub-task of the NTCIR-9 MT evaluation. First, we made changes to our SMT training procedure in order to better handle the special characteristics of patent data, and obtained incremental improvements from the various changes. Then, the re-training of the casing LM with patent text and the use of more features to the MT system improved the BLEU scores significantly. Finally, the LM adaptation improved the MT performance further – by about 1 BLEU point. Our work shows that most of the strategies for building an SMT system, such as the use of a large number of features and the LM adaptation, were easily applicable to the patent genre and produced gains. But certain techniques had to be customized in order to better handle the special characteristics of the patent genre.

6. REFERENCES

[1] Stanley F. Chen and Joshua Goodman. 1998. "An Empirical Study of Smoothing Techniques for Language Modeling", Tech. Rep. TR-10-98, Harvard University.

[2] David Chiang, Kevin Knight, and Wei Wang. 2009. "11,001 new features for statistical machine translation", in Proceedings of the Human Language Technologies, NAACL 2009.

[3] Makoto Iwayama, Fujii Atsushi, Kando Noriko, and Akihiko Takano. 2003. "Overview of patent retrieval task at NTCIR-3", in Proceedings of the third NTCIR workshop.

[4] Yaohong Jin. 2010. "A hybrid-strategy method combing semantic analysis with rule-based MT for patent machine translation", in Proceedings of 2010 International Conference on NLP-KE, 2010.

[5] Jae-Ho Kim and Key-Sun Choi. 2007. "Patent document categorization based on semantic structural in-formation", Information Processing and Management (2007), doi:10.1016/j.ipm.2007.02.002

[6] Franz J. Och and Hermann Ney. 2003. "A systematic comparison of various statistical alignment models", Computational Linguistics, 29(1):19_51.

[7] Len Offersgaard and Claus Povlsen. 2007. "Patent documentation – comparison of two MT strategies", in Proceedings of the Machine Translation Summit XI workshop on Patent Translation.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu. 2002. "BLEU: a Method for Automatic Evaluation of Machine Translation", in Proceedings of the 40th Annual conference of the association of computational linguistics.

[9] Atti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2010. "BBN System Description for WMT 10 System Combination Task", Proceed-ings of the 5th workshop on Statistical Machine Translation, 2010.

[10] Sayori Shimohata. 2005. "Finding Translation Candidates from Patent Corpus," *MT Summit X Workshop on Patent Translation, 2005*.

[11] Libin Shen, Jinxi Xu and Ralph Weischedel. 2010. "String-to-Dependency Statistical Machine Translation", in *Computational Linguistics*, pages 649-671, Vol. 36, No. 4, Dec. 2010.

[12] Matthew Snover, Bonnie J. Dorr, and Richard Schwartz. 2008. "Language and translation model adaptation using comparable corpora", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[13] Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2010. "TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate", *Machine Translation Journal, Special Issue on: Automated Metrics for MT Evaluation*.

[14] Terumasa Ehara. 2007. "Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation", in Proceeding of MT Summit XI, pp.13-18.

[15] Dan Wang. 2009. "Chinese to English automatic patent machine translation at SIPO", World Patent Information, Volume 31, Issue 2, pp. 137-139.

[16] Jinxi Xu, Ralph Weischedel and Chanh Nguyen. 2001. "Evaluating a Probabilistic Model for Cross-Lingual Information Retrieval", in SIGIR 2001: 105-110Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825.

[17] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumito and Benjamin K. Tsou. 2011. "Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop", in *Proceedings of the 9th NTCIR workshop on Evaluation of Information Access Technologies*.

[18] Jacob Devlin. 2009. Lexical features for statistical machine translation. Master's thesis, Univ. of Maryland