English to Bangla Phrase-Based Machine Translation

Md. Zahurul Islam
Dipartimento di Informatica
Università di Pisa
islam@di.unipi.it

Jörg Tiedemann
Department of Linguistics
and Philology
Uppsala University
jorg.tiedemann@lingfil.uu.se

Andreas Eisele
Language Technology Lab
DFKI GmbH, Saarbrücken
eisele@dfki.de

Abstract

Machine Translation (MT) is the task of automatically translating a text from one language to another. In this work we describe a phrase-based Statistical Machine Translation (SMT) system that translates English sentences to Bangla. A transliteration module is added to handle out-of-vocabulary (OOV) words. This is especially useful for low-density languages like Bangla for which only a limited amount of training data is available. Furthermore, a special component for handling preposition is implemented to treat systematic grammatical differences between English and Bangla. We have shown the improvement of our system through effective impacts on the BLEU, NIST and TER scores. The overall BLEU score of our system is 11.7 and for short sentences it is 23.3.

1 Introduction

SMT requires enormous amount of parallel text in the source and target language to achieve high quality translation. However, many languages are considered to be low-density languages, either because the population speaking the language is not very large, or because insufficient digitized text material is available in a language even though it is spoken by millions of people. Bangla/Bengali is one such language. Bangla, an Indo-Aryan language, is a language of Southeast Asia, which comprises present day Bangladesh and the Indian state of West Bengal. With nearly 230 million speakers, Bangla is one of the most spoken languages in the world, but only a very small number of tools and resources are available for Bangla.

Our aim in this work is to present a phrase-based SMT system for translating English sentences to Bangla. The current state-of-the-art phrase-based SMT systems available for this task is based on a log-linear translation model, which is used as our baseline system. Though we have built our system with a small amount of training data compared to MT systems of other language pairs, we have got better results than existing MT systems for this language pair when tested on the same domain. A transliteration module has been added as a component with the translation system to handle OOV words. For the transliteration module we applied the same phrase-based SMT model, but using characters instead of words. Another difference to standard SMT models is the preposition handling module we added to our system to handle prepositional diversity between English and Bangla. Instead of prepositions, as in English, Bangla uses postpositions, or attaches inflections to the head noun of the prepositional phrase (i.e., the object of the preposition) (Naskar and Bandyopadhyay, 2006b).

The rest of the paper is organized as follows. Section 2 summarizes some related past work. Section 3 shows the importance of preposition handing during translation. Section 4 describes the experiments and results. Finally, we summarize our observation, and outline future work in Section 5, followed by conclusion in Section 6.

2 Related Work

Although being among the top ten most widely spoken languages in the world, Bangla language still lacks significant research in the area of natural language processing, especifically in MT. Dasgupta et.al. (2004) propose an approach for English to Bangla MT that uses syntactic transfer of English sentences to Bangla aiming at optimal time complexity. In the beginning, they tag the English sentences and then parse those sentences in the next step. They use the the CYK algorithm, which outputs parse trees in Chomsky Normal Form (CNF) form. In the third step, they con-

vert CNF trees to normal parse trees using some conversion rules. Unlike a normal parse tree, each parent node in the CNF tree must have two children. In the next step, they use transfer rules and a bilingual dictionary to convert English parse trees to Bangla parse trees. Finally, they generate output translation with morphological generation.

Saha and Bandyopadhyay (2005) propose an English to Bangla Example Based Machine Translation (EBMT) system for translating news headlines. Here translation of source to target headlines is done in three steps. In the first step, search in made directly in the example base; if it is not found there then it searches in the generalized tagged example base. If a match is found in the second step, then it extracts the English equivalent of the Bangla words from the bilingual dictionary and applies synthesis rules to generate the surface words. If the second step fails, then the tagged input headline is analyzed to identify the constituent phrases. The target translation is then generated from the bilingual example phrase dictionary, using heuristics to reorder the Bangla phrases.

Naskar and Bandyopadhyay (2006a) present an EBMT system. This work identifies the phrases in the input through a shallow analysis, retrieves the target phrases from a set of examples and finally combines the target language phrases by employing some heuristics based on the phrase reordering rules in Bangla.

Naskar and Bandyopadhyay (2006b) show a technique of handling prepositions in an English to Bangla machine translation system. In Bangla there is no concept of preposition. In many cases English prepositions are translated to Bangla by attaching inflections to the head nouns of the prepositional phrase. The English form of preposition (preposition) (reference object) is translated to (reference object) [(inflection)] [(prepositional word)]. The reference object plays a major role in determining the correct preposition sense. For example: *at home* should be translated to the stem word বাড়ি (bari : home) and the inflection -তে (-te).

Roy (2009) proposes a semi-supervised approach for Bangla to English phrase based MT. A baseline system was built using a limited amount of parallel training data. The system randomly selects sentences from a Bangla monolingual corpus, and translates them using the baseline system. Finally, source and translated sentences were added to the existing bilingual corpora. Acquiring parallel sentences is an iterative process until a certain translation quality is achieved.

Anwar et al. (2009) propose a context-sensitive parser that is used in machine translation. Each Bangla sentence is converted to structural representation (parse tree), translation of each Bangla word is chosen from a bilingual dictionary and finally English sentences are constructed with the help of a bigram language model.

There is an open source machine translation system called Anubadok (*http://anubadok.sourceforge.net*) available for translating English to Bangla. The translation process consists of three steps. First it converts different kind of documents into XML format. Then it tokenizes, tags and lemmatizes English sentences. In the next step, it performs the translation. At the beginning of the third step, it determines the sentence type, subject, object, verb and tense, and then translates English words to Bangla using a bilingual dictionary. Finally, it joins subject, object and verbs in the SOV order.

Transliteration systems are being used nowadays in MT systems. UzZaman et. al (2006) present a phonetics based transliteration system for English to Bangla which produces intermediate code strings that facilitate matching pronunciations of input and desired output. They have used table-driven direct mapping techniques between the English alphabet and the Bangla alphabet, and a phonetic lexicon–enabled mapping.

Matthews (2007) presents a machine transliteration system of proper nouns using MOSES (Koehn et al., 2007). The system transliterates proper nouns in both directions in both English–Chinese and Arabic–English. He has achieved 43.0% accuracy of forward transliteration from Arabic to English and 37.8% from English to Chinese.

3 Importance of Preposition Handling

Prepositional systems across languages vary to a considerable degree, and this cross-linguistic diversity increases as we move from core physical senses of preposition into the metaphoric extensions of prepositional meaning (Naskar and Bandyopadhyay, 2006b). The lexical meaning of a preposition is important, because it is intended for use in a MT system, where the meaning of a sentence, a phrase or lexical entry of the source language must be preserved in the target language, even though it may take different syntactic forms

in the source and target language.

Instead of prepositions Bangla typically uses postpositions or inflectional attachment to the head noun (reference object). The noun is usually in the genitive/accusative case unless the two words are placed under the rules of সন্ধি (Sandhi)[1] or সমাস (Samas), [2] in which case the noun is not inflected. Therefore, English prepositions are translated to Bangla by attaching appropriate inflections to the head noun (reference object). For example: inflection -তে (-te) attaches to the noun বাড়ি (home) and it becomes বাড়িতে (at home), inflection -য় (-y) attaches to the noun সন্ধ্যা (the evening) and it becomes সন্ধ্যায় (in the evening). The English form of preposition (preposition) (reference object) is translated to Bangla (reference object) [(inflection)] [(postpositional-word)]. Our intuition is that handling prepositions during translation will improve the MT performance.

4 Experiment

4.1 Data and Tools

We have used a parallel corpora of South Asian languages called Enabling Minority Language Engineering (EMILLE) corpus developed by Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India. Information about EMILLE corpora is available here: *http://www.elda.org/catalogue/en/text/W0037.html*. This corpus is distributed by the European Language Resources Association. It contains 200,000 words of text in English and its accompanying translations in Hindi, Bangla, Punjabi, Gujarati and Urdu. The Bangla translation contains 189,495 words. Table 1 shows the EMILLE corpus statistics for English and Bangla. Before training the system we converted the file encoding to UTF-8. All the sentences have been extracted to text from XML mark up and aligned using an automatic sentence aligner. Finally, we tokenize the English and Bangla part of the corpus and convert the English text to lower case.

We also used KDE4 system messages as a corpus; the English and the Bangla translation of

[1]Sandhi is the euphonic change when words are conjoined. For example: ঢাকা (Dhaka) + ঈশ্বরী (Godess) = ঢাকেশ্বরী (Godess of Dhaka), বধু (bride) + উৎসব (festival) = বধূৎসব (festival to welcome a bride). Sandhi takes place on the simple joining of words in a sentence, on the formation of compound words and on the adding of affixes to noun or verbs.

[2]Samas is the rules of compounding words. For example: তুমি (you) এবং (and) আমি (I) = আমরা (we).

Table 1: EMILLE English - Bangla corpus statistics

Encoding: UTF-16
Total number of files : 72 (English) and 70 (Bangla)
Total English sentences : 12,654
Total Bangla sentences : 12,633

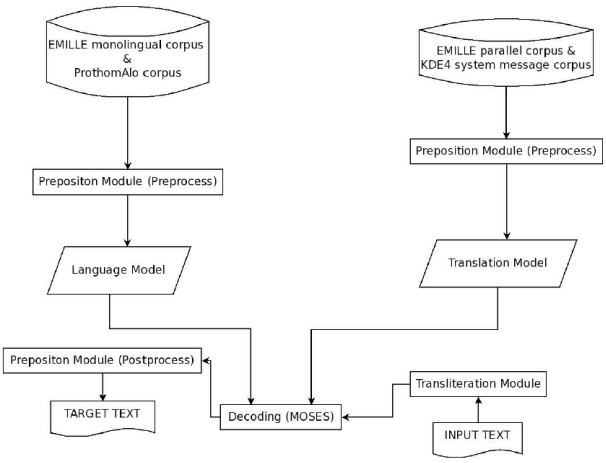

Figure 1: System architecture

BN_BD (Bangla in Bangladesh) and the BN_IN (Bangla in WestBengal/India) domains. The KDE4 corpus has been taken from the OPUS corpus (Tiedemann and Nygard, 2004) which is already aligned. This KDE4 system message corpus contains 221,409 words and 33,365 sentence pairs with UTF-8 encoding. We used the monolingual corpus from the EMILLE project and the Prothom-Alo corpus developed by BRAC University, Bangladesh. The EMILLE monolingual corpus contains 1,867,452 words and the Prothom-Alo corpus contains 19,496,884 words.

We have used freely available tools, such as MOSES, GIZA++ (Och and Ney, 2003), MERT (Och, 2003) and SRILM (Stolcke, 2002) to build the system. To evaluate a MT system a single metric is not enough. Therefore, we have used more than one metric to evaluate our translation and transliteration system. BLEU (Papineni et al. , 2001), NIST (Doddington, 2002) and TER (Snover et al. , 2006) have been used for translation evaluation and ACC (accuracy in Top-1), Top-5, Top-20 as well as Mean F–score for transliteration evaluation. ACC represents the correctness of candidate transliteration; Top-5 and Top-20 represent the percentage of correct transliteration in top 5 and top 20 candidates; Mean F–score refers to

Table 2: Sample output of baseline system

English	**Bangla**
A shoppers guide	একটি shopper এগুল
Your legal rights	আপনার আইনগত অধিকার
Office of fair trading	office of ন্যায্য ব্যবসা
DTI publications orderline	dti publications orderline
The office of fair trading also has a new director general Mr John Vickers october 2000	office - ন্যায্য ব্যবসা বাধ্যতামূলকভাবে একটি নতুন শতাংশই সাধারণ মিঃ john vickers অক্টোবর উনিশশ
This is not as difficult as it sounds and just the threat of it could be enough to resolve matters.	এই নয় হিসাবে কঠিন এইটি হিসেবে ধ্বনি এবং করতে threat সেটা যায়নি যথেষ্ট পারেন।
Mahmoud Ahmadinejad has denied the holocaust, describing it a ”myth”.	mahmoud ahmadinejad হয়েছে প্রত্যাখ্যাত যে holocaust, এটা ”myth”
Ban Ki-Moon	ban ki-moon

the closeness of candidate transliteration with the reference transliteration.

4.2 System Architecture

Section 3 describes the importance of preposition handling during translation. A module is added to the system to handle prepositions. The preposition handling module is divided into two parts: *preprocess* and *post-process*. The *preprocess* sub-module interchanges the positions of postpositional words and reference object. Moreover, it separates suffixes (inflections) from the reference object and puts that before the reference object with suffix marker. The *post-process* sub-module moves the postpositional words after the reference objects and adds suffixes to the next word (reference object). The preposition handling module is applied only on the Bangla corpus.

A transliteration module is also added, as we believe that this modules will improve the translation quality and accuracy of our MT system. It is basically responsible for identifying OOV words, and transliterates them in order to avoid the presence of English words in the target Bangla translation. Figure 1 shows the combined system architecture.

Table 3: Evaluation of basline system

Test corpus	BLEU	NIST	TER
EMILLE	1.20	1.65	0.90
KDE4	14.00	4.19	1.02
Combined	5.10	2.70	0.89

4.3 Baseline System

We constructed a parallel training corpus of 10,850 sentence pairs using the GMA sentence aligner (Dan Melamed, 1996). The English corpus contains 199,973 words and the Bangla corpus contains 189,495 words. Each side of the KDE4 corpus contains 35,366 sentence pairs. The KDE4 Bangla corpus contains 221,409 words while the English corpus contains 157,392 words. We separated 500 sentence pairs from the EMILLE corpus and 1,000 sentence pairs from the KDE4 corpus for development sets. We also separated the same number of sentences from both corpus as a test set. The *5-gram* language model was built from the EMILLE monolingual corpus, Prothom-Alo corpus and the training data, which together contain more than 21 million words. Table 2 shows sample output of the baseline system, and Table 3 shows the evaluation result of this system.

4.4 Corpus Cleaning

Table 2 shows that the output contains some English words. The reason for this is the inclusion of many English words in the Bangla side of training, development and test sets, in some cases even entirely in English. This occurs for both the EMILLE and the KDE4 corpus. Therefore, we decided to clean the corpus in order to get better translation.

First of all, we wanted to improve the automatic sentence alignment of the EMILLE corpus. For this we experimented with the Interactive Sentence Aligner (ISA) (Tiedemann, 2006) tool. ISA is an interactive tool with web interface for semi-automatic sentence alignment of parallel XML documents. It uses a standard length-based approach to align sentences and allows to manually add or remove segment boundaries to correct existing alignments in order to improve the overall quality of the parallel corpus.

However, the ISA corpus alignment tool was not enough to clean the EMILLE corpus. Most of the files vary substantially. For example, the English text about child education has a total of

Table 4: Irregularities in Bangla corpus

English Corpus	**Bangla Corpus**
We analysed relevant data and worked with the Office for National Statistics (ONS) in order to establish better estimates of the incidence of low pay	আমরা সমস্ত প্রাসঙ্গিক তথ্য বিশ্লেষণ করেছি এবং কম মজুরির প্রভাবের হিসাব আরো ভালোভাবে করতে পারার জন্য অফিস ফর ন্যাশনাল স্ট্যাটিস্টকস্ (Office for National Statistics - ONS) বা জাতীয় পরিসংখ্যান দফতরের সঙ্গে একত্রে কাজ করেছি ।

Table 5: Evaluation of new translation system

Test corpus	BLEU	NIST	TER
EMILLE	5.10	3.1	0.84
KDE4	22.50	5.18	0.65
Combined	11.10	4.24	0.78

662 sentences, whereas the Bangla translation of this file has a total of 425 sentences. In various cases translations on either side were missing or moved. Another noticeable observation was that for organization or group name in the English text, there were Bangla translation, Bangla transliteration and names in English in the Bangla corpus. Table 4 shows example of these inconsistencies, অফিস ফর ন্যাশনাল স্ট্যাটিস্টকস্ is the transliteration and জাতীয় পরিসংখ্যান দফতর is the translation of "Office for National Statistics". Finally, we cleaned the corpus manually which took more than 120 working hours. We found two among seventy files where we could not align the sentences at all which were deleted from the corpus.

The KDE4 corpus also contains many English words in Bangla side. To clean this corpus we simply extracted sentence pairs where there is no English character on the Bangla side. Finally, we have obtained in total 9,111 sentence pairs from the EMILLE corpus and 16,389 sentence pairs from the KDE4 corpus.

Table 6: Sample English OOV words and their transliteration with XML markup

<np translation="অফিস "> office </np>
<np translation = "স্পার"> shoppers </np>
<np translation = "কি-মুন"> ki-mon</np>

Table 7: Evaluation of transliteration module

Test Set	ACC	Top 5	Top 20	M. F-score
EN Names	0.187	29.68	79.68	0.797

Table 8: Evaluation of combining transliteration module with translation system

Test corpus	BLEU	NIST	TER
EMILLE	5.40	3.13	0.83
KDE4	23.20	5.16	0.63
Combined	11.40	4.25	0.77

4.5 New Translation System

From the manually cleaned and aligned corpora we extracted the same number of sentences for development and test set as the baseline system. The new language model has been cleaned as well – we deleted all the sentences which contain English characters. This time a *8-gram* language model has been used. We compared different *N-gram* language models for Bangla and observed the best BLEU score for *8-gram* even though the output subjectively looks better when we use a *5-gram* model. With all these changes we obained significantly better output compared to the previous baseline system. Table 5 shows the evaluation result of the new translation system. We have achieved more than 100% improvement over the baseline system.

4.6 Transliteration Module

Generally, SMT systems are trained using large parallel corpora. These corpora consist of several million words, still they can never be expected to have a complete coverage especially over highly productive word classes like proper nouns. When translating a new sentence, SMT systems use the knowledge acquired from the training corpora. If they come across a word not seen during training, then they will at best either drop the unknown word, or copy into the translation. Table 2 shows that there are some words (*office, Mahmoud Ahmadinejad and Ban Ki-Moon*) in the output (Bangla) text which are not translated by the baseline system, which may be OOV words or English words in the Bangla training corpus. Hence a transliteration system is an promising addition to the baseline system to handle proper nouns

or OOV words. However, there are risks of using a transliteration module for OOV words or names. When we transliterate names, the output translation contains some English words (unknown words), and when we transliterate OOV words, there are some words which should not be transliterated. We experimented with both and got better BLEU score when we transliterated OOV words.

We collected 2,200 unique names from Wikipedia and Geonames (www.geonames.org). To build this system, we tried to go a step further than the translation system and treated the words (names) as sequences of letters, which have to be translated into a new sequence of letters. We used the same tools as the translation system. For the language model (*8-gram*) we used the Prothom-Alo corpus and Bangla names from training. We just put one space between each character in the corpora. We used the same tools as the translation system (e.g. MOSES, GIZA++, MERT) and followed the same steps (e.g. training, tuning and testing) as well. Table 7 shows evaluation results of the transliteration system.

We identified OOV words by using *english.vcb* as a vocabulary list which is generated by GIZA++ during alignment. Any word not in the vocabulary list is considered as an OOV word. We transliterated all the OOV words using the transliteration module and replaced them with XML markup. Table 6 shows some sample English OOV words and their transliteration with XML markup. MOSES has an advanced feature by which we can provide external knowledge to the decoder during decoding. The *-xml-input* flag was raised with *exclusive* value so that the XML-specified translation (transliterated name) is used for the input phrase and any phrase from the phrase tables that overlaps with that span is ignored. Table 8 shows the evaluation result of combining the transliteration module with the translation system.

4.7 Preposition Handling Module

We have already mentioned the importance of handling prepositions during English to Bangla machine translation. English prepositions are translated to Bangla by attaching inflections to the head noun of the prepositional phrase, or as a postposition word after the head noun. To implement the prepositional module, we took the *intersection* of word alignment using the intersection option during the training of MOSES. Then we extracted the intersection word list from our training corpus. As there is no freely available Parts of Speech (POS) tagger for Bangla, we used the OpenNLP (http://opennlp.sourceforge.net) tool to POS tag English words and transfer the tags to the aligned Bangla words. For many English words there was more than one candidate tag. In this case we considered only the top 1 candidate. Finally, we extracted the words that are tagged as noun.

Table 9: Sample output of final combined system

English	**Bangla**
A shoppers guide	একটি স্পার এর নির্দেশিকা
Your legal rights	আপনার আইনগত অধিকার
Office of fair trading	অফিসে ন্যায্য বাণিজ্য
This is not as difficult as it sounds and just the threat of it could be enough to resolve matters.	এটি কঠিন নয় এবং এটা কেবল এই গণ্ডগোলের হুমকি এটা হতে পারে যথেষ্ট অভিযোগটির মীমাংসা করার জন্য।
Mahmoud Ahmadinejad has denied the holocaust, describing it a ”myth”.	মাহমুদ আহমেদিনাজাদ আছে যে, বর্ণনা করা হচ্ছে এটা করা হলোকাস্ট একটি " মিথ "।
Ban Ki-Moon	কি-মোন নিষেধ
In some cases it may be necessary to go to court to get the matter settled.	কোন কোন ক্ষেত্রে প্রয়োজনীয় হতে পারে কোর্টে যে সেটেল্ড পেতে পারেন।
This offers the additional benefit to consumers of a 14 day cooling-off period on most goods sold by members of the direct selling association .	আপনার ক্রেতাদের যে অতিরিক্ত বেনিফিট কোন সময়ের ১৪ দিন কলিঙ্গ-অফফ ওপর বেশীর ভাগ মালবাহী sold সদস্যদের বিক্রি করে সরাসরি অ্যাসোসিয়েশন।

We preprocessed corpora in two steps. Firstly, we come up with 19 postpositional words. We identified those postpositional words in the corpora and moved them before the reference object (head noun). Secondly, we came up with a group of 9 suffixes which can be attached to the nouns. We just stripped those suffixes from the nouns and put them in front of the noun with a suffix mark (#X#, where X is a suffix). We did these for training, development and for the monolingual corpus for the language model.

After translation, we identify suffixes (any word with suffix mark) then attach them to the next

word. Our assumption is that the word after any suffix will be a noun (reference object). We have checked the output after post processing. In very few cases, suffixes were attached to words that are not noun. We also identified each postpositional word, and move it after the next word. Table 9 shows the sample output of combining the preposition handling module with the previous combined system. Table 10 shows the evaluation result.

4.8 Comparison with Anubadok

We have compared our system with the open source MT system for English to Bangla called Anubadok. We have used the same test set as we used to evaluate our system. Our system clearly outperforms Anubadok. Table 11 shows the comparison result. The BLEU score of the Anubadok system is 1.60 while our system's BLEU score is 11.70.

5 Observation and Future Work

SMT systems require a significant amount of parallel corpora to achieve satisfactory translations. However, there are not enough parallel corpora available for English and Bangla to achieve high quality translation using only a statistical MT system. Our system works reasonably well for short sentences. After transliterating all the OOV words, there are only a few remaining English words in the translated text. Table 9 shows that *sold* is an English word in the translated text. As a consequence, we have found 53 phrase-table entries where *sold* is the part of the phrase. But there is no phrase-table entry where *sold* is a single phrase and also no entry where *sold* is following or preceding any of the adjacent words in the test sentences. We have experimented with *grow-diag-final*, *grow*, and *intersection* alignment options available in MOSES. In all the cases some English words were left in the translated text after transliterating all OOV words. Another noticeable observation in Table 9 is that the word "Ban" is wrongly translated. Here "Ban" is the part of the name "Ban Ki-Moon", but in English "Ban" is also a verb and the Bangla translation is নিষেধ which we can see in the output. This type of ambiguity will remain in our system.

Although satisfactory results (for a low–density language) were obtained with the current system and the additional modules, there is still a lot of room for improvement in several parts of the system. Obviously the largest improvements can be expected when adding more parallel data for training the system.

Table 10: Evaluation of final combined system

Test corpus	BLEU	NIST	TER
EMILLE	5.70	3.16	0.83
KDE4	23.30	5.18	0.63
Combined	11.70	4.27	0.76

Table 11: Comparison with Anubadok

System	BLEU	NIST	TER
Anubadok	1.60	1.46	1.03
Our System	11.70	4.27	0.76

A test set with more than one reference would be useful to evaluate our system. Our plan is to develop a test set for English ⇔ Bangla MT system with more than one reference translation.

Another idea for future work is to extend the preposition handling component. Adding more postpositional words and inflectional suffixes would improve the system's accuracy.

Furthermore, both English and Bangla have many compound words. Another module that could handle English compound words would be useful for an English ⇔ Bangla MT system. Finally, the integration of linguistic features is another direction for future work. We like to work on the enhancement of the phase-based SMT model with some features representing syntactic information and morphological information. We also need to add more names in the training data in order to improve the transliteration quality.

6 Conclusion

In this paper we presented an English to Bangla phrase-based statistical machine translation system. We incorporated two additional modules in the baseline translation system (transliteration and preposition handling) to improve the translation accuracy and quality. We also showed that an automatic transliteration system can be built with the phrase-based SMT model and that its performance is comparable to the state-of the-art transliteration system (Jiang et al. , 2009). Even though the transliteration module still has a rather low accuracy of 0.18, the produced transliterated words are very close to reference transliterations which gives

an improved impression of translation quality. The preposition handling module is also effective to improve translation accuracy. Even though there is not much improvement after combining the baseline system with the preposition handling module, the results suggest that prepositions should be handled separately in the English ⇔ Bangla machine translation. Overall we obtained reasonable scores for short sentences (23.30 BLEU and 0.63 TER). However, on average the scores are still much lower with BLEU: 11.70, NIST: 4.27 and TER: 0.76.

Acknowledgments

We would like to thank the Erasmus Mundus European Masters Program in Language and Communication Technologies (EM-LCT) for support. This work was carried out while the first author was an EM-LCT student at the University of Groningen. One of the authors was supported by the EuroMatrixPlus project funded by the European Commission under FP7.

References

Md. Musfique Anwar, Mohammad Zabed Anwar and Md. Al-Amin Bhuiyan. 2009. *Syntax Analysis and Machine Translation of Bangla Sentences*, International Journal of Computer Science and Network Security, 09(08),317–326.

Sajib Dasgupta, Abu Wasif and Sharmin Azam. 2004. *An Optimal Way Towards Machine Translation from English to Bengali*, Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT).

George Doddington. 2002. *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*, Proceedings of the second international conference on Human Language Technology Research.

Xue Jiang and Le Sun and Dakun Zhang. 2009. *A Sylleable-based Name Transliteration System*, Proceedings of Name Entities Workshop(NEW) 2009, ACL–IJCNLP.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL).

Dan Melamed. 1996. *A Geometric Approach to Mapping Bitext Correspondence*, Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP).

David Matthews. 2007. *Machine Translation of Proper Names*, Masters Thesis, School of Informatics, University of Edinburgh.

Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2006a. *A Phrasal EBMT System for Translating English to Bengali*, Proceedings of the Workshop on Language, Artificial Intelligence, and Computer Science for Natural Language Processing Applications (LAICS–NLP).

Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2006b. *Handling of prepositions in English to Bengali Machine Translation*, Proceedings of the EACL 2006 Workshop.

Franz Josef Och 2003. *Minimum error rate training in statistical machine translation*, Proceedings of the 4th Annual Meeting of the Association for Computational Linguistics (ACL).

Franz Josef Och and Hermann Ney 2003. *A Systematic Comparison of Various Statistical Alignment Models*, Journal of the Computational Linguistics, 29(1), 19-51.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2001. *BLEU: a Method for Automatic Evaluation of Machine Translation*, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

Maxim Roy 2009. *A Semi-supervised Approach to Bengali-English Phrase-Based Statistical Machine Translation*, Proceedings of the 22nd Canadian Conference on Artificial Intelligence.

Diganta Saha and Sivaji Bandyopadhyay. 2005. *A semantics-based English-Bengali EBMT system for translating news headlines*, Proceeding of MT Summit X second workshop on example-based machine translation.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhou. 2006. *A Study of Translation Edit Rate with Targeted Human Annotation*, Proceedings of Association for Machine Translation in the Americas.

Andreas Stolcke 2001. *SRILM–an extensible language modeling toolkit*, Proceedings of the ICSLP.

Jörg Tiedemann and Lars Nygard. 2004. *The OPUS corpus - parallel and free*, Proceedings of LREC 2004.

Jörg Tiedemann 2006. *ISA and ICA – Two Web Interfaces for Interactive Alignment of Bitexts*, Proceedings of LREC 2006.

Naushad UzZaman, Arnab Zaheen and Mumit Khan. 2006. *A Comprehensive Roman (English) to Bangla Transliteration Scheme*, Proceedings of International Conference on Computer Processing on Bangla.