

DEVELOPING LANGUAGE RESOURCES
FOR ENGLISH/বাংলা MACHINE TRANSLATION

A Thesis

Submitted to the Department of Computer Science and Engineering

of

BRAC University

By

Rabia Sultana Ummi

Student ID: 03101037

Fahmina Huda

Student ID: 04301006

In Partial Fulfillment of the

Requirements for the Degree

of

Bachelor of Science in Computer Science and Engineering

August 2008

DECLARATION

We hereby declare that this thesis is based on the results found by ourselves. Materials of work found by other researchers are mentioned by reference. This thesis, neither in whole nor in part, has been previously submitted for any degree.

Signature of
Supervisor

Signature of
Authors

ACKNOWLEDGMENTS

Special thanks to Dr. Mumit Khan for his support, teachings and supervision during the entire period of work on this paper. Special thanks to Mr. Altaf Mahmud, the developer of Bangla tagset used in this research, for supplying us with the initial resource needed for this thesis and for helping us understand the tagset better.

ABSTRACT

We developed English-Bangla parallel corpora for statistical machine translation. By hand we tagged 20,000 words of our Bangla corpus according to their particular part of speeches. In our work we also suggested a method for identifying word correspondence in parallel English-Bangla text using a translation model based on part of speech and n-gram model.

TABLE OF CONTENTS

	Page
TITLE.....	1
DECLARATION.....	2
ACKNOWLEDGEMENTS.....	3
ABSTRACT.....	4
TABLE OF CONTENTS.....	5
LIST OF TABLES.....	6
LIST OF FIGURES.....	7
1. INTRODUCTION	8
2. THE TAGSET AND SUGGETION.....	9
3. METHODOLOGY.....	12
4. SUMMARY OF RESULT.....	16
5. FUTURE WORK AND CONCLUSION.....	17
REFERENCES.....	18
APPENDICES.....	19

LIST OF TABLES

Table	Page
1. Bangla Tagset	9
2. n-Gram model for Word Alignment.....	15

LIST OF FIGURES

Figure	Page
1. Annotated English-Bangla Parallel Corpora	13
2. Cognate Lookup and Suffix Removal.....	15

1. Introduction

Machine translation investigates the use of computer software to translate text or speech from one natural language to another [3]. There are three approaches to developing machine translation. These are rule based, example based and statistical.

Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora, such as the Canadian Hansard corpus, the English-French record of the Canadian parliament. Where such corpora are available, impressive results can be achieved translating texts of a similar kind, but such corpora are still very rare [3].

In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (now usually electronically stored and processed). They are used to do statistical analysis, checking occurrences or validating linguistic rules on a specific universe.

A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus). Multilingual corpora that have been specially formatted for side-by-side comparison are called aligned parallel corpora.

In order to make the corpora more useful for doing linguistic research, they are often subjected to a process known as annotation. An example of annotating a corpus is part-of-speech tagging, or POS-tagging, in which information about each word's part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of tags.

Machine translation for English-Chinese, English-Arabic, English-French, and many other pairs of languages are relatively quite advanced. On the other hand, very little work has been done in developing machine translation software for translating English to Bangla. Whatever work has been done, it's for the Bangla of Western Bengal in India.

In response to the lack of resources for English-Bangla machine translation, our paper describes the development of a parallel English-Bangla Corpora and a recommendation of a statistical method to word-align Bangla text to corresponding English text in the corpora, using a translation model based on part of speech (POS) and n-gram model. In the process, a 20,000 words part of speech tagged Bangla corpus was developed. These resources are required to develop machine translation software for translating Bangla text or speech to English text or speech or vice versa.

To begin with, we had only a Bangla tagset which was under-development. With this meager resource, we set out to developing other resources required for implementing an English-Bangla statistical machine translation as mentioned above. In Section 2 of this paper, we describe the Bangla tagset and the major modification we have instigated to finalize documentation for the tagset. In Section 3, we describe the methodologies we have used in developing the resources mentioned above. Finally, in the next two sections, there are brief discussions on the outcome of this thesis and the future steps needed in the development of English-Bangla machine translation.

2. The tagset and Suggestions

The initial Bangla tagset, we started working on had 50 tags. After scrutinizing it, we suggested some modifications. Based on these suggestions, the developer made some other modifications to the rules. Additionally, he increased the number of tags to 55. After these modifications, a final documentation has been developed.

The final tagset has 55 tags and 17 categories. Each category includes a set of subcategories. All the 55 tags, their categories and subcategories, and examples have been included in Table 1. Level 1 of the table refers to the category of the tag while Level 2 refers to subcategory of the tag.

Table 1
55-Tag Bangla Tagset

Level 1	Level 2	Tag	Word
Noun	Common	NN	মানুষ, পানি
	Proper	NNP	মতিউর, অক্টোবর, ঢাকা, চট্টগ্রাম, শনিবার
	Compound Common Noun	NNC	ছেলে/NNC মেয়ে/NN, স্বরাষ্ট্র/NNC মন্ত্রনালয়/NN
	Compound Proper Noun	NNPC	আব্দুর/NNPC রহমান/NNPC বিশ্বাস/NNP
	Verb Root	NNV	গোসল, পান
	Temporal	NNT	গতকাল, আগামীকাল, আজ
	Question Temporal	QNT	কখন, যখন
	Locative	NNL	উপর, নিচ, আগে
	Question Locative	QNL	কোথায়, যেথায়, যেখানে

Level 1	Level 2	Tag	Word
Pronoun	Personal Pronoun	PRP	আমি, আমরা, তুমি, তোমরা, সে, তারা, আপনি, তিনি, তুই, তোরা
	Question Pronoun	QPR	কে, কারা, যে, যারা
Adjective	Simple	JJ	সুন্দর, লাল, গরম, শ্রেষ্ঠ, শ্রেষ্ঠতর, শ্রেষ্ঠতম
	Verb Root	JJV	লাল, দুর্বল
	Question Adjective	QJJ	কেমন, যেমন
Vocative	Vocative	VOC	ওপো, ওরে, ওহে
Verb	Main Finite Verb	VB	করি, কর, করে, করাই, করলাম, করলে, করেছিস, করব, করাব
	Nonfinite Nominal	VBM	করা, করানো, পরা, পরানো
	Nonfinite Conditional	VBC	করলে, করালে
	Nonfinite Perfective	VBT	করে, গিয়ে
	Nonfinite	VBF	করতে, করাতে
	Finite Existential	VBE	হয়, হবে
	Nonfinite Existential	VBEF	হতে
Adverb	Adverb	RB	দ্রুত, হয়তো, অবশ্য, না, নাই, খুব, বেশী, অনেক
	Question Adverb	QRB	কেন, কিভাবে, যেভাবে
Conjunction	Coordinating	CC	এবং, ও, কিংবা, অথবা, নতুবা
	Compound Coordinating	CCC	না/CCC হয়/CC
	Suspicion	CN	যদি, পাছে
	Eternal Joining	CET	যেমন/CET ... তেমন/CET, যেই/CET ... সেই/CET, যখন/CET ... তখন/CET
	Subordinating	CS	যে, কেননা, বলে, এইজন্য
	Compound Subordinating	CSC	তাই/CSC বলে/CS, এই/CSC কারণে/CS
Postposition	Postposition	ON	দ্বারা, কর্তৃক, হতে, থেকে, জন্য, চেয়ে, চাইতে
Interjection	Interjection	UH	বাহ্!, ওহ্!, হায়!
Particle	Particle	RP	না, তো, বটে
	Question Particle	QRP	কি
Determiner	Common	DT	ওসব, তাবৎ, কোন, যেকোন, এই, ঐ
	Singular	DTS	এটি, ওটি
	Question Determiner	QDT	কোনটা, যেটা, কোনগুলো, যেগুলো, কোনসব
Quantifier	Quantifier	QF	সব, সকল, আরও, কম, কিছু
	Quantifier Number	QFNUM	১, ২, এক, তিন, একটি, পাঁচটি
	Question Quantifier	QQF	কত, যত, কতটুকু
Foreign Word	Foreign Word	FW	যেকোন বিদেশী শব্দ
Symbol	Symbol	SYM	বৈজ্ঞানিক বা অংকশাস্ত্রীয় যেকোন চিহ্ন, অন্যান্য
List Item Marker	List Item Marker	LS	a, b, (a), 1, 2.3.1, ক, ৩.১৩

Level 1	Level 2	Tag	Word
Suffix	Postpositional	SFON	এ, য়, তে
	Accusative	SFAC	কে, রে, এরে, দিগকে, দিগেরে
	Possessive	SF\$	এর, দের
Punctuation Mark	Sentence Final Punctuation	.	।, ?, !
	Comma	,	,
	Colon, Semi-colon	:	∴, ;
	Dash, Double-Dash	-	-, --
	Left Parenthesis	({ [
	Right Parenthesis)	}]
	Opening Left Quote	LQ	' , "
	Closing Right Quote	RQ	' , "

There have been various issues where we felt modifications are needed in the tagset. These issues and some of our suggestions are discussed below.

In certain situations where a title is associated in a name such as *Dr.* in *Dr. Rabeya Khatun*, we had suggested using a new tag for the title preceding the actual name. Our suggestion was based on the tagset of C7 where the *Dr.* is tagged NNB since it is before the actual name. Similarly, we suggested using a new tag for a noun following the actual name, as in example *Dr. Rabeya Khatun M.B.B.S.* In C7, *M.B.B.S.* would be tagged as NNA since it's after the actual name. Therefore, it would be very useful to use a separate tag for nouns such as the ones mentioned above. In arguments against our suggestion, explanations terming increase in the size of the Bangla tagset as problematic have been provided. Since there was no rule assigned for these cases, the developer of the tagset modified the tagset to tag these titles as NN.

In cases of name of degree, award, etc used with a name (following it) such as *M.B.B.S* in *Dr. Rabeya Khatun M.B.B.S* and বীর প্রতীক in তারামন বিবি বীর প্রতীক, no rules have been suggested in the documentation. We suggest tagging these words as NN.

Additionally, there is no tag suggested when a title is added to a name (preceding it) such as শের-এ-বাংলা in শের-এ-বাংলা এ কে ফজলুল হক and বীর শ্রেষ্ঠ in বীর শ্রেষ্ঠ হামিদুর রহমান. So, we suggest tagging them, along with the whole name, as NNPC. This is so because these titles have become unavoidable part of the name.

According to the documentation, DTS are “Single determiners precede nouns and determinate objects those are singular in number and cannot be inflected by any

suffix.” To contradict this, we would like to give examples like ওটা আমের আঁটি. Here আমের আঁটি is a compound common noun and আমের is inflected by suffix. Thus, DTS can also be used when a singular noun is inflected. In এটা চিনি আর ওটা লবণ, চিনি and লবণ are mass nouns. Therefore, a correction has to be made so as to include mass nouns in the definition.

Since in Bangla part of speech, there is no adverb. Therefore, there are arguments favoring tagging words in Bangla, corresponding to adverbs in English as nouns or adverbs. To resolve this situation, we have been advised to tag these words as nouns (NNT or NNL). For example, উপর in তুমি উপরে যাও has been tagged as NNL instead of RB, following the above rule mentioned.

While POS-tagging the Bangla words, we have come across certain words like “মধ্যে”, “সময়”, “ভেতরে ভেতরে” and “দেখাদেখি”. There were no guidelines in the documentation for such confusing words. After a discussion with the developer of the tagset, the documentation was updated and tags for such words were included.

One other change we suggest is using separate tags for the punctuation marks in Bangla as in the case of C7. Each punctuation mark in Bangla has separate uses. For example, ‘!’ and ‘|’ should not be both be tagged as sentence final punctuation, ‘.’, because the exclamation sign has more use than just terminating a sentence.

Using the final tagset and its documentation, we started developing POS-tagged English-Bangla Parallel Corpora. In the next section, we discuss the methodologies used in reaching our goal.

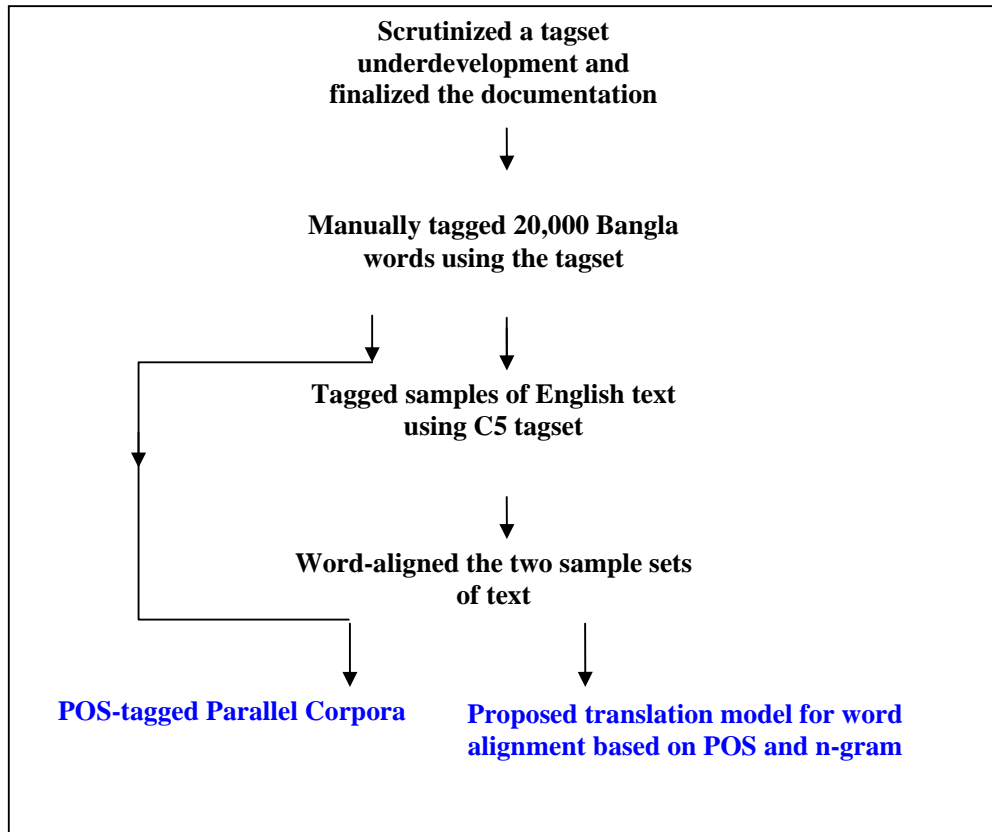
3. Methodology

An overview of the process we used to develop a POS-tagged Parallel Corpora and a proposition of translation model for word alignment based on POS and n-gram model can be seen in the form of Figure 1. As already mentioned in the previous section, we first scrutinized the initial tagset and then used the final tagset to manually tag 20,000 Bangla words. The source of these words is www.bdnews24.com, a Bangladeshi news website. Various news reports in Bangla and their corresponding reports in English were taken. Then, 20,000 words of the Bangla reports were tagged using the tagset. Combining these, tagged texts with their corresponding English texts, we developed the POS-tagged parallel corpora. To serve as examples, we have included two sets of extracts from the parallel corpora. The first set includes the English sentence “Rupali is now undergoing treatment at Dhaka Medical College

Hospital” with its corresponding Bangla sentence “রুপালি বর্তমানে ঢাকা মেডিকেল কলেজ হাসপাতালে চিকিৎসাধীন রয়েছে।” and the second one consists of “Her sister and another worker suffered injuries in the accident and were sent to hospital” as English text and “দুর্ঘটনায় তার বোন ও অপর এক শ্রমিক আহত হন এবং তাদের হাসপাতালে পাঠানো হয়েছে।” as its corresponding Bangla sentence.

Additionally, some samples of English texts were tagged using the C5 tagset for English. The tagged Bangla text was combined with the samples of tagged English text to manually word align the two set of texts. To word align the two sets of text, we used a technique, a proposition we have made for future work in the area, for word alignment using n-gram model and POS.

Figure 1: Annotated English-Bangla Parallel Corpora



In order to word align the Bangla text corresponding to the English text; we have used an algorithm called PosAlign. Chang et al suggest using this algorithm to word align English-Chinese texts. The algorithm has three steps [2]. The steps are:

1. Tag the parallel text with part-of-speeches
2. Initial alignment with the help of cognate lookup
3. Train the translation model iteratively using the unaligned part of the POS sequences

After the tagging of the two sets of texts in Bangla and English, an initial alignment using cognate look up is done. Cognate includes proper nouns, pronouns, numerical expressions and punctuation marks [2]. We also suggest using relationships (father, mother, sister, brother-in-law, etc) as cognates because size of the set of corresponding translations in Bangla is very limited.

Initially, we hand aligned some parts of the texts. Using these parts, we developed an n-gram model to devise a statistical translation model. We used this model on the unaligned part of the POS tagged set of sample texts in English and Bangla. Advantage of using POS alignment is parallel text of limited size is required for training.

To understand the process better, we present a discussion that walks through the process of word alignment of a Bangla sentence to its corresponding English sentence, using PosAlign. The sentences are pre-tagged. This paper will use the following Bangla and English tagged sentences:

“The/AT0 injured/AJ0 were/VBD admitted/VVN to/PRP Nazirhat/NP0 health/NN1 complex/NN1./.”

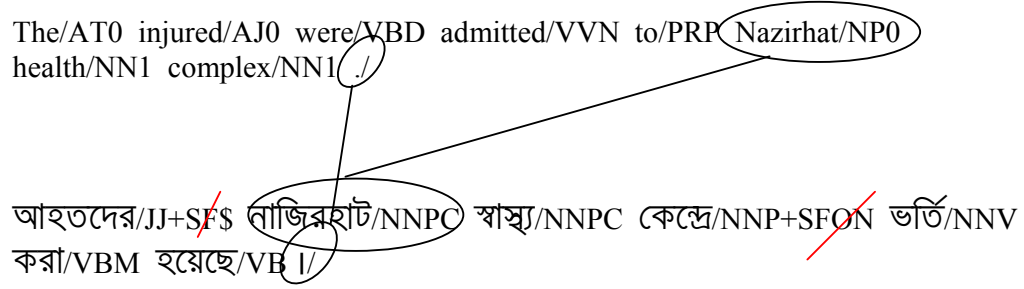
and

“আহতদের/JJ+SF\$ নাজিরহাট/NNPC স্বাস্থ্য/NNPC কেন্দ্রে/NNP+SFON ভর্তি/NNV করা/VBM হয়েছে/VB I./.”

Firstly, cognate look up is used to word align the sentences partially. As a result, we align “Nazirhat” to “নাজিরহাট” and ‘.’ to ‘I’ as in Figure 2. Next, we remove the suffix tags, also represented in the figure. Here, we notice that in the

example used only the Bangla sentence has suffixes. After these actions, the resulting tag sets of the English and Bangla sentences look like {AT0, AJ0, VBD, VVN, PRP, NN1} and {JJ, NNPC, NNP, NNV, VBM, VB} respectively.

Figure 2: Cognate Lookup and Suffix Removal



Then we use the n-gram model, to word align the rest of the sentences. A subset of the n-gram model is shown in Table 2. We shall use this subset in aligning the rest of the sentences. The table is a very simple one where there are columns titled “বাংলা/English” meaning the Bangla tag given the English tag and the conditional probability of the Bangla tag given the English tag or the conditional probability of a sequence of Bangla tags given a sequence of English tags. We look up the first tag in the Bangla set, “JJ” in the table. There are three entries for “JJ”. The first one is “JJ/ATO,AJO” with a probability of 1.00, the second one is “JJ/AJ0” with a probability of 0.67 and the third one is “JJ/NN1” with a probability of 0.33. Of these three, we choose the one with the highest probability, i.e. “JJ/ATO, AJO”. Thus “আহতদের/JJ” is aligned to “The/AT0 injured/AJ0”.

Table 2
n-Gram model for Word Alignment

বাংলা/English	Conditional Probability	বাংলা/English	Conditional Probability
NNPC/NN1	0.40	NNPC/AJ0	0.40
NNP/NN1	1.00		
NNV,VBM,VB/VBD,VVN	1.00		
VBM,VB/VBD,VVN	1.00		
JJ/ATO,AJO	1.00		
JJ/AJ0	0.67	JJ/NN1	0.33
Φ/AT0	0.33	Φ/PRP	0.50

Next we look up “NNPC” and we find two entries, “NNPC/NN1” and “NNPC/AJO” both with equal probability of 0.4. Thus, there is a tie. And still there remains a problem of aligning health with স্বাস্থ্য and complex with কেন্দ্র. To resolve this issue and also the tie, we have devised a rule. The rule suggests that if there is an NNPC aligned with a NP0 in our sentences by using the method of cognate lookup, first count the number of tags in the Bangla compound proper noun sequence (here it is 3; NNPC NNPC NNP) and then check whether NP0 is the starting tag of a sequence in the English sentence that contains the same number of tags and ends with either NP0 or NN1. If we have got a sequence like NP0.....NP0, all the tags are to be NP0s there. Else if we have a sequence like NP0.....NN1, the sequence may contain other tags like AJ0. If so, then align the two sequences with each other using the same order as they appear in the sentences as following.

AT0 AJ0 VBD VVN PRP NP0 NN1 NN1 .
 JJ NNPC NNPC NNP NNV VBM VB .

We also lookup the sequence “NNV,VBM,VB” in the table and locate the entry “NNV,VBM,VB/VBD,VVN” with a probability of 1.00. Thus, we align the rest of the words in English with the rest of the words in Bangla. This way the whole sentence in Bangla is word aligned 100% accurately to the sentence in English.

4. Summary of Result

The outcome of the thesis is the development of resources required to develop English-Bangla machine translation. A 20,000 Bangla word tagged parallel corpora has been built.

Our method of word alignment of sentences using POS and n-gram model has shown good performances, but more tests are required. Encouraged by this good performance of the alignment process, we propose this technique for word alignment of sentences.

5. Future Work and Conclusion

The 20,000 word annotated Bangla corpus will be very useful, not only in the area of machine translation, but also in various other areas of natural language processing. Moreover, the bilingual corpus can be used for statistical machine translation.

This paper encourages with suggestions and guidelines for those who wish to work in Sentence Alignment, Cognate Dictionary and similar other areas. It serves as the foundation for the development of statistical machine translation for English-Bangla text or speech processing.

After this stage of work in this area, future work is required in statistical sentence alignment. From there, an electronic cognate dictionary must be built. Once these two processes are done, advanced work in statistical machine translation using POS alignment can be done.

List of References

- [1] Daniel Jurafsky, James H. Martin, *Speech and Language Processing*, University of Colorado, Boulder, Pearson Education, Inc., 2000

- [2] Jyun-Sheng Chang, Huey-Chyun Chen, *Using Part-of-Speech Information in Word Alignment*, National Tsing Hua University, Conference of the Association for Machine Translation in the Americas (1994)

- [3] *Wikipedia, The Free Encyclopedia*,
http://en.wikipedia.org/wiki/Machine_translation

APPENENDICES

Finalized documentation of Mr. Altaf Mahmud's 55-Tag Bangla Tagset has been appended in this section.