Example Based English-Bengali Machine Translation Using WordNet

Khan Md. Anwarus Salam†, Mumit Khan* and Tetsuro Nishino†

† Department of Information and Communication Engineering, Graduate School of Electro-Communications, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, Japan.
**Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh*
Tel: +81-042-443-5246, Fax: +81-042-443-5293, Email: kmanwar@gmail.com

Abstract

In this paper we propose an architecture of English-Bengali Example Based Machine Translation (EBMT) using WordNet. The proposed EBMT system has five steps: 1) Tagging 2) Parsing 3) Prepare the chunks of the sentence using sub-sentential EBMT 4) Using an efficient adapting scheme, match the sentence rule 5) Translate from Source Language (English) to Target Language (Bengali) in the chunk and generate with morphological analysis with the help of WordNet. Using the word senses given by the WordNet we can detect the ambiguity and improve the correctness of translation.

1. Introduction

In present there are many ways of machine translation system. [2] Many researchers came up with different approaches. But still it is not possible to get the finest possible result. It has been found that EBMT has several advantages in comparison with other MT paradigms. For example, it can be upgraded easily by adding more examples to the knowledge base. It is robust because of best-match reasoning. WordNet needed for word sense disambiguation. It makes explicit the semantic relations. Using WordNet we showed a way to improve English-Bengali EBMT.

2. Background

Example-based Machine Translation makes use of past translation examples to generate the translation of a given input. An EBMT system stores in its example base of translation examples between two languages, the source language and the target language. These examples are subsequently used as guidance for future translation tasks. In order to translate a new input sentence in Source Language (SL), similar SL sentence is retrieved from the example base, along with its translation in Target Language (TL). This example is then adapted suitably to generate a translation of the given input.

It may be observed that in today's world a lot of information is being generated. However, since most of this information is in English, it remains out of reach of people at large for which English is not the language of communication. As a consequence, an increasing demand for developing machine translation systems from English to Bengali is being felt very strongly.

However, development of MT systems typically demands availability of a large volume of computational resources, which is currently not available for Bengali [5]. Moreover, generating such a large volume of computational resources (which may comprise an extensive rule base, a large volume of parallel corpora etc.) is not an easy task. EBMT scheme, on the other hand, is less demanding on computational resources making it more feasible to implement in respect of these languages.

3. Data preparation

3.1 Prepare knowledge base

Table1 shows some example English - Bengali translations. Our EBMT system will store the translation rules in the knowledge base [6]. We also prepared English to Bengali Dictionary with Morphological Analysis.

Table1: Sample knowledge base of the English-Bengali EBMT system [1]

English	English Chunk	Transfer to Bengali Chunk	Bengali
He reads a book	[NP He/PRP] [VP reads/VBZ] [NP a/DT book/NN]	[NP সে/PRP] [VP পড়ে /VBZ] [NP একটি/DTবই/NN]	সে একটি বই পড়ে
The sun Rises in the east	[NP The/DT sun/NN Rises/NNS] [PP in/IN] [NP the/DT east/JJ]	[NP /DT সূর্য /NN উদিত /NNS] [PP হয়/IN] [NP /DT পূর্বে /JJ]	সূর্য পূর্বে উদিত হয়
He is reading a book	[NP He/PRP] [VP is/VBZ reading/VBG] [NP a/DT book/NN]	[NP সে /PRP] [VP /VBZ পড়ছে /VBG] [NP এক টি /DT বই /NN]	সে এক টি বই পড়ছে
They are reading a book	[NP They/PRP] [VP are/VBP reading/VBG] [NP a/DT book/NN]	[NP তারা /PRP] [VP /VBP পড়ছে /VBG] [NP এক টি /DT বই /NN]	তারা এক টি বই পড়ছে
I have done the work	[NP I/PRP] [VP have/VBP done/VBN] [NP the/DT work/NN]	[NP আমি /PRP] [VP /VBP কাজটি /VBN] [NP টি/DT কাজ/NN]	আমি কাজটি করেছি
He has gone to Dhaka	[NP He/PRP] [VP has/VBZ gone/VBN] [PP to/TO] [NP Dhaka/NNP]	[NP সে /PRP] [VP /VBZ গিয়াছে /VBN] [PP /TO] [NP ঢাকা /NNP]	সে ঢাকা গিয়াছে
The boys were playing	[NP The/DT boys/NNS] [VP were/VBD playing/VBG] ./.	[NP /DT বালকগুলো /NNS] [VP /VBD খেলছিল /VBG] ./.	বালকগুলো খেলছিল

3.2 Using WordNet

We connect with English WordNet to get information about word senses. It helps to detect Ambiguity and translate correctly from the rule.

4. Proposed EBMT Architecture

The proposed EBMT system has five steps [1]

1. Tagging the English sentence
2. Parsing the English sentence
3. Using sub-sentential EBMT prepare the chunks of the sentence
4. Using an efficient adapting scheme match the sentence rule [3].

5. Translate from Source Language (English) to Target Language (Bengali) in the chunk and generate with morphological analysis with the help of WordNet.

Table 2: Sample word rules for plural words.

	অপ্রাণীবাচক	প্রাণীবাচক	সাধারণ	উদাহরণ
আবালি	yes	no	no	চরিতাবলি, পদাবলি
কুল	no	yes	no	
গণ	no	yes	no	মনুষগণ, দেবতাগণ
গাম	yes	yes	no	
চয়	yes	no	no	
জন	no	yes	no	বিদ্বজন, পণ্ডিতজন
দাম	yes	no	no	লতাদাম, বিদূদ্দাম
নিকর	yes	no	no	
নিচয়	yes	no	no	
মণ্ডল	yes	no	no	মেঘমণ্ডল
মণ্ডলী	no	yes	no	পণ্ডিতমণ্ডলী
মালা	yes	no	no	মেঘমালা
রাজি	yes	no	no	বৃক্ষরাজি
লোক	no	yes	no	
বর্গ	no	yes	no	নেতৃবর্গ
বৃন্দ	no	yes	no	সুধীবৃন্দ
সকল	no	no	yes	
সভা	no	yes	no	পণ্ডিতসভা যুবতীসভা
সব	no	no	yes	ভাইসব
সমূচয়	no	no	yes	
সমুই	no	no	yes	
মহল	yes	no	no	রাজনৈতিকমহল, বন্দুমহল

	রা	গণ	বৃন্দ	মন্ডলী	বর্গ
শিক্ষক	no	yes	yes	yes	no
বালিকা	yes	no	no	no	no
গরিব	yes	no	no	no	no
ধনী	yes	no	no	no	no
দেব	yes	yes	no	yes	yes
নর	yes	yes	no	no	no
জন	no	yes	no	no	no
সুধী	yes	yes	yes	yes	yes
ভক্ত	yes	yes	yes	yes	yes
সম্পাদক	yes	yes	yes	yes	yes
পণ্ডিত	yes	yes	no	no	yes

4.1 Tagging and parsing

Tagging, is the process of marking up the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e., relationship with adjacent and related words in a phrase, sentence, or paragraph. Eg. I do-> I/PRP do/VBP

Parsing, is the process of analyzing a sequence of tokens to determine grammatical structure with respect to a given formal grammar. We used the tag set of Table3 for tagging the English sentence. Eg. I am a boy-> (S (NP (PRP I)) (VP (VBP am) (NP (DT a) (NN boy))))

Table3: Tag set used in English-Bengali EBMT [10]

Level 1	**Level 2**	**Tag**
Noun	Common	NN
	Proper	NNP
	Compound Common Noun	NNC
	Compound Proper Noun	NNPC
	Verb Root	NNV
	Temporal	NNT
	Question Temporal	QNT
	Locative	NNL
	Question Locative	QNL
Pronoun	Personal Pronoun	PRP
	Question Pronoun	QPR
Adjective	Simple	JJ
	Verb Root	JJV
	Question Adjective	QJJ
Vocatives	Vocatives	VOC
Verb	Main Finite Verb	VB
	Nonfinite Nominal	VBM
	Nonfinite Conditional	VBC
	Nonfinite Perfective	VBT
	Nonfinite	VBF
	Past tense	VBD
	Gerund/present participle	VBG
	Past participle	VBN
	Non-3rd ps. sing. Present	VBP
	3rd ps. sing. Present	VBZ
	Existential	VBE
Adverb	Adverb	RB
	Question Adverb	QRB
Conjunction	Co-ordinating	CC
	Compound Co-ordinating	CCC
	Suspicion	CN
	Eternal Joining	CET
	Subordinating	CS
	Compound Subordinating	CSC
Numbers	Cardinal Numbers	CD
Interjection	Interjection	UH
Particle	Particle	RP
	Question Particle	QRP
Determiner	Common	DT
	Singular	DTS
	Question Determiner	QDT
Quantifier	Quantifier	QF
Foreign Word	Foreign Word	FW
Symbol	Symbol	SYM
List Item Marker	List Item Marker	LS
Suffixes	Adpositional	SFON
	Accusative	SFAC
	Possessive	SF$
Punctuation Marks	Sentence Final Punctuation	.
	Comma	,
	Colon, Semi-colon	:
	Dash, Double-Dash	-
	Opening Left Quote	LQ
	Closing Right Quote	RQ
	Preposition/subordinate conjunction	IN
	Adjective, superlative	JJS
	Adjective, comparative	JJR
	Modal	MD
	Proper noun, plural	NNPS
	Noun, plural	NNS
	Possessive ending	POS
	Possessive pronoun	PRP$
	Adverb, comparative	RBR
	To	TO
	wh-determiner	WDT
	wh-pronoun	WP
	Possessive wh-pronoun	WP$

4.2 Handle complex sentence using sub-sentential EBMT

Handling complex sentence in general considered to be difficult to deal with in an MT system. Since exact sentence matches only occur in special domains, we want to extend this to sub-sentence matches. For this we need to:

- Find the most similar example (involves segmenting by preparing chunks)
- Alter source side to match current input.

Similarity requires a "distance metric" in the source language (English). This can be closeness:

- of the lexical items in a hierarchy of terms/ concepts from ontology
- of the sequence of syntactic categories and function words,
- of the two syntactic structures,
- or combinations of these.

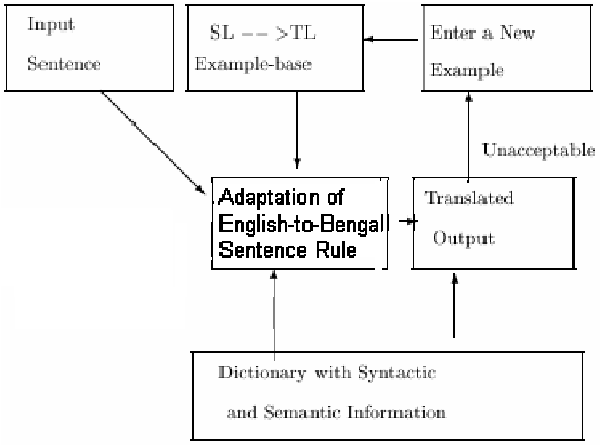

Figure 1: Role of Examples in Translation

4.3 Adapting scheme to match sentence rule

Efficient adaptation of past examples is a major aspect of an EBMT system. There are many adaptation schemes available for an EBMT system. Even an efficient similarity measurement scheme and a quite large example base cannot guarantee an exact match for a given input sentence. As a consequence, there is a need for an efficient and systematic adaptation scheme for modifying a retrieved example, and thereby generating the required translation. In section 5 we discuss details about our proposed adaptation scheme. In Table1 we gave a Sample knowledge base of the English to Bengali EBMT System. During translation our adapting scheme chooses the best rule for the source sentence.

4.3 Translate with morphological analysis with the help of WordNet.

Study of divergence for English to Bengali translation is also required. Translation divergence can be effectively handled within an EBMT framework. As in earlier step we have the sample rule and the parsed sentence. Now we can easily translate the sentence by matching the rule.

I am a boy > ami ekti chele

I am a man > ami ekjon manus

In these two examples "a" has different meaning in Bengali "jon" and "ti". Here we can see that it has same sentence rule but different translation. Depending on the quality of the word we are choosing the actual meaning. Using WordNet we are determining that word sense. This technique dramatically improves the quality of EBMT.

For all birds plural we can use -kul

- Birds are flying > pakhikul akashe urchhe
- Parrots are flying> totapakhikul akashe urchhe

But for trees we have to use -raji

- Trees give us food> brikkhoraji amader khaddo dey

From the above example we see that in Bengali based on Noun quality different pos-fix used. Using WordNet and Table2 we can easily identify the ambiguity and translate correctly.

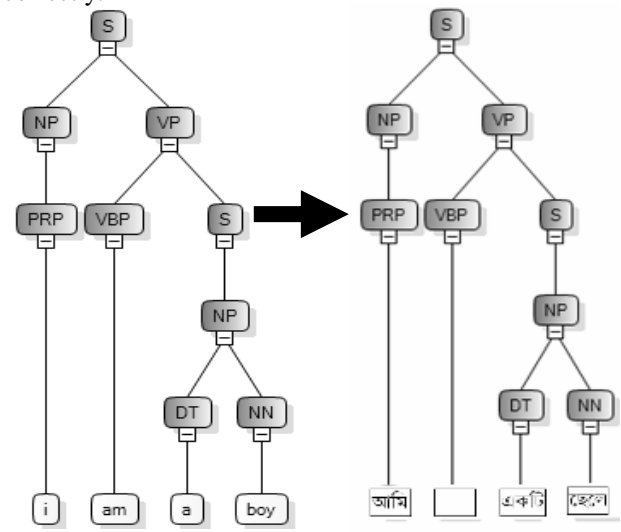

Figure 2: Translating English-Bengali

5. Adaptation in English to Bengali translation

A successful EBMT system requires a good adaptation scheme. The need for an efficient and systematic adaptation scheme arises for modifying a retrieved example, and thereby generating the required translation. Researcher came with various approaches to deal with adaptation aspect of an EBMT system. Overall the adaptation procedures employed in different EBMT systems primarily consist of four operations [4]:

- Copy, where the same chunk of the retrieved translation example is used in the generated translation;
- Add, where a new chunk is added in the retrieved translation example;
- Delete, when some chunk of the retrieved example is deleted; and
- Replace, where some chunk of the retrieved example is replaced with a new one to meet the requirements of the current input.

5.1 Adaptability/Map ability for a chunk has 4 discrete values:

Level 3: Source Language (SL): Target Language (TL) mapping is one-to-one for all words

Level 2: Syntactic Functions map, but not some POS tags

Level 1: Functions differ, but lexical correspondence still holds

Level 0: Cannot establish correspondence

5.2 Description of the adaptation operations

i. Constituent Word Replacement (WR): One may get the translation of the input sentence by replacing some words in the retrieved translation example. Suppose the input sentence is: "The bird was eating apples.", and the most similar example retrieved by the system (along with its Bengali translation) is: "The elephant was eating fruits", "haathii phol khacchilo". The desired translation may be generated by replacing "haathii" with the Bengali of "birds", i.e. "pakhi" and replacing "phal" with the Bengali of "apples", i.e. "aapel". These are examples of the operation of constituent word replacement.

ii. Constituent Word Deletion (WD): In some cases one may have to delete some words from the translation example to generate the required translation. For example, suppose the input sentence is: "Animals were dying of thirst". If the retrieved translation example is : "Birds and Animals were dying of thirst.", "pakhi ebong pashu trishnay mara jacche", then the desired translation can be obtained by deleting "pakhi ebong" (i.e the Bengali of "birds and") from the retrieved translation. Thus the adaptation here requires two constituent word deletions.

iii. Constituent Word Addition (WA): This operation is the opposite of constituent word deletion. Here addition of some additional words in the retrieved translation example is required for generating the translation. For illustration, one may consider the example given above with the roles of input and retrieved sentences being reversed.

iv. Morpho-word Replacement (MR): In this case one morpho-word is replaced by another morpho-word in the retrieved translation example. For illustration, if the input sentence is "He eats rice", and the retrieved example is: "He is reading a book.", "se akte boi porChe", then to obtain the desired translation[4] first the morpho-word "Che" is to be replaced by "e"

v. Morpho-word Deletion (MD): Here some morpho-word(s) are deleted from the retrieved translation example.

vi. Morpho-word Addition (MA): This is the opposite case of morpho-word deletion. Here some morpho-words need to be added in the retrieved example in order to generate the required translation.

vii. Suffix Replacement (SR): Here the suffix attached to some constituent word of the retrieved sentence is replaced with a different suffix to meet the current translation requirements. This may happen with respect to noun, adjective verb, or case ending.

viii. Suffix Deletion (SD): By this operation the suffix attached to some constituent word may be removed, and thereby the root word may be obtained.

ix. Suffix Addition (SA): Here a suffix is added to some constituent word in the retrieved example.

x. Copy (CP): When some word (with or without suffix) of the retrieved example is retained in to in the required translation then it is called a copy operation.

5.3 Study of adaptation procedure for morphological variation of active verbs

Verb morphology variations are divided into four groups. These are:

1. Same tense same verb form
2. Different tenses same verb form
3. Same tense different verb forms
4. Different tenses different verb forms

6. Summary

In this current research we focused to improve English to Bengali EBMT. Presently our system can deal with simple English sentence to translate into Bengali as it has limited knowledge base. By increasing the knowledge base we can improve the correctness of the translation.

7. References

[1]. "Example Based English to Bengali Machine Translation" B.Sc. Thesis of Khan Md. Anwarus Salam completed in August 2009, BRAC University

[2]. Machine Translation: An Introductory Guide , *By Doug Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys, Louisa Sadler; Colchester, August 1993*

[3]. Contributions To English To Hindi Machine Translation Using Example-Based Approach, Phd Theses in January 2005, Deepa Gupta, IIT Delhi.

[4]. D. Gupta and N. Chatterje., Study of Divergence for Example Based English-Hindi Machine Translation. STRANS-2001, IIT Kanpur, 2001 pp. 43-51.

[5]. Balanced Bengali Language Corpus: A Proposal, *By Khan Md. Anwarus Salam, S M Murtoza Habib and Dr. Mumit Khan*, Research work in BRAC University in 2008.

[6]. H.A. Guvenir and I. Cicekli., Learning Translation Templates from Examples. Elsevier Science Ltd., 1998

[7]. R. Jain, R.M.K Sinha and A. Jain., ANUBHATRI: Using Hybrid Example-Based Approach for Machine Translation.. STRANS-2001, IIT Kanpur, 2001 pp. 20-32.

[8]. Verb Transfer For English To Urdu Machine Translation, Thesis by Nayyara Karamat, FAST-Lahore, 2006

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

[10].Research Report on Bangla Tagset, Altaf Mahmud and Mumit Khan, CRBLP, BRAC University, Dhaka, Bangladesh