How to Translate Unknown Words for English to Bangla Machine Translation Using Transliteration

Khan Md. Anwarus Salam
Graduate School of Informatics and Engineering,
The University of Electro-Communications, Chofu, Tokyo, Japan.
Email: kmanwar@gmail.com

Setsuo Yamada
NTT Corporation, Tokyo, Japan
Email: yamada.setsuo@lab.ntt.co.jp

Tetsuro Nishino
Graduate School of Informatics and Engineering,
The University of Electro-Communications, Chofu, Tokyo, Japan.
Email: nishino@ice.uec.ac.jp

***Abstract*—Due to small available English-Bangla parallel corpus, Example-Based Machine Translation (EBMT) system has high probability of handling unknown words. To improve translation quality for Bangla language, we propose a novel approach for EBMT using WordNet and International-Phonetic-Alphabet(IPA)-based transliteration. Proposed system first tries to find semantically related English words from WordNet for the unknown word. From these related words, we choose the semantically closest related word whose Bangla translation exists in English-Bangla dictionary. If no Bangla translation exists, the system uses IPA-based-transliteration. If unknown word is not found in the English IPA dictionary, the system uses Akkhor transliteration mechanism. We implemented the proposed approach in EBMT, which improved the quality of good translation by 16 points.**

***Index Terms*—English – Bangla Machine Translation, Example-Based Machine Translation, Transliteration, WordNet**

I. Introduction

According to the survey of "Distribution of languages on the Internet"., 56.4% web contents are in English [15]. On the other hand, "Human Development Report 2009" of United Nations Development Program (UNDP) reported the literacy rate of Bangladesh as 53.5% [16]. That means, around half of the Bangla speaking people of Bangladesh are monolingual. To improve the information access to those Bangla speaking monolingual people, it is important to have good English to Bangla Machine Translation (MT) system. However, lack of parallel corpus makes the development of the MT system very challenging.

English has rich language resources like automated parser, tokenizer and WordNet. WordNet is a large lexical database of English, where nouns, verbs, adjectives and adverbs are grouped into clusters using <lexical filename> information [4]. On the other hand Bangla is a low-resource language.

In this situation, to utilize the available computational language resource for English, we consider to use English as source-language (SL) and Bangla as target-language (TL).

There were several attempts in building English-Bangla MT systems. The first available free MT system from Bangladesh was Akkhor Bangla Software [17]. The second available online MT system was apertium based Anubadok [18]. These systems used Rule-Based approach and did not handle unknown words considering low-resource scenario. Most recently from June 2011, Google Translation started offering MT service for Bangla language. Google is using statistical approach and they also have issues in translating unknown words.

In present, Machine Translation systems can categorize as Rule-Based MT (RBMT), Statistical MT (SMT) and Example-Based MT (EBMT). RBMT require human made rules, which are very costly in terms of time and money, but still unable to translate general-domain texts. SMT and EBMT both are data driven approach. SMT works well for close language pairs like English and French. It requires huge parallel corpus, but currently huge English-Bangla parallel corpus is not available. EBMT is better choice for Bangla language, as it is less demanding on large parallel corpus. Moreover, EBMT can translate in good quality when it has good example match. All these approach has issues on translating unknown words.

We considered EBMT approach by improving the translation quality using WordNet, IPA-based and Akkhor transliteration. For using WordNet in translation

doi:10.4304/jcp.8.5.1167-1174

rules we used chunk-string templates (CSTs). CSTs consist of a chunk in the source language (English), a string in the target language (Bangla), and the word alignment information between them. CSTs are generated from the aligned parallel corpus and WordNet, by using English chunker. For clustering CSTs, we used <lexical filename> information for each words, provided by WordNet-Online.

To improve translation quality for Bangla language, we propose a novel approach for EBMT using WordNet, IPA-based and Akkhor transliteration. Proposed system first tries to find semantically related English words from WordNet for the unknown word. From these related words, we choose the semantically closest related word whose Bangla translation exists in English-Bangla dictionary. If no Bangla translation exists, the system uses International-Phonetic-Alphabet(IPA)-based-transliteration. If unknown word is not found in the English IPA dictionary, the system uses the transliteration mechanism provided by Akkhor Bangla Software. We mention this mechanism in section VII. Based on the above methods, we built an English-to-Bangla MT system. We implemented the proposed approach in EBMT, which improved the quality of good translation by 16 points.

II. Related Works

Chunk parsing was first proposed in [1]. Although EBMT using chunks as the translation unit is not new, it has not been explored widely for low-resource Bangla language yet. [5] proposed to use syntactic chunks as translation units for improving insertion or deletion words between two distant languages. However this approach requires an example base with aligned chunks in both source and target language. In our example-base only source side contains chunks and target side contains corresponding translation string.

Gangadhariah et. al. showed that templates can be useful for EBMT with statistical decoders to obtain longer phrasal matches [10]. Their templates increased coverage and quality. However, manually clustering the words can be a time consuming task. It would be less time consuming to use standard available resources such as WordNet for clustering. That is why in our system, we used <lexical filename> information for each English words, provided by WordNet-Online for clustering the proposed CSTs.

Dasgupta et. al. proposed to use syntactic transfer in [11]. They converted CNF trees to normal parse trees and using a bilingual dictionary, generated output translation. This research did not consider translating unknown words.

Naskar et. al. reported a phrasal EBMT for translating English to Bangla in [12]. They did not provide any evaluation of their EBMT. They did not clearly explain their translation generation.

Saha et. al. reported an EBMT for translating news headlines in [3]. Their works showed that EBMT can be a good approach for Bangla language. Their approach only considered about news headlines.

English to Bangla phrase-based statistical machine translation was reported by Islam et. al. in [9]. This system achieved low BLEU score due to small parallel corpus for English-Bangla.

Salam et. al. proposed EBMT for English-Bangla language pair using WordNet to improve translation quality but did not consider about translating unknown words in [7].

III. EBMT Architecture

The Fig. 1 shows the proposed EBMT architecture. The dotted rectangles identified the main contribution area of this research. During the translation process, at first, the input sentence is parsed into chunks using OpenNLP Chunker. The output of Source Language Analysis step is the English chunks. Then the chunks are matched with the example-base using the Matching algorithm as described in section IV. This process provides the CSTs candidates from the example-base. It also mark the unknown words. In Unknown Word Translation step, using our proposed mechanism in section V, we try to find translation candidates for those unknown words. Then in Generation process WordNet helps to translate determiners and prepositions correctly to improve MT performance [7]. Finally using the generation rules we output the target-language strings. Based on the above MT system architecture, we built an English-to-Bangla MT system.

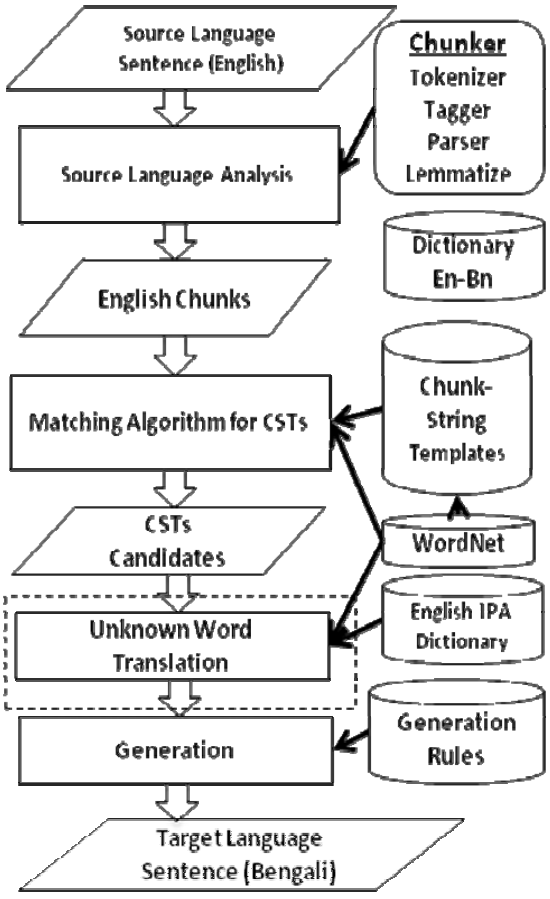


Fig. 1: Proposed EBMT Architecture

In this research we proposed EBMT based on chunk-string templates (CSTs), which is especially useful for developing a MT system for high-resource to low-resource language. CSTs consist of a chunk in the source language (English), a string in the target language (Bangla), and the word alignment information between them. From the English-Bangla aligned parallel corpus CSTs are generated au-tomatically.

Table I shows sample word-aligned parallel corpus. Here the alignment information contains English position number for each Bangla word. For example, the first Bangla word "বশ্বিব্যাপী" is aligned with the 11th word in the English sentence. That means "বশ্বিব্যাপী" is aligned with "worldwide".

TABLE I
EXAMPLE WORD-ALIGNED PARALLEL CORPUS

English	Bangla	Align
Bangla is the native language of 1 2 3 4 5 6 around 230 million people worldwide 7 8 9 10 11	বশ্বিব্যাপী বাংলা হচ্ছে প্রায় ২৩০ মিলিয়ন মানুষ এর মাতৃভাষা	11 1 2 7 8 9 10 6 4

The example-base of our EBMT is stored as CSTs. CSTs consists of <c;s;t>, where c is a chunk in the source language (English), s is a string in the target language (Bangla), and t is the word alignment information between them.

Fig. 2 shows the steps of CSTs generation. First the English chunks are auto generated from a given English sentence. Then initial CSTs are generated for each English chunks and each CSTs global alignment for complete sentences are generated using the parallel corpus. After that the system generate more CSTs, finally we generalize CSTs using WordNet to achieve wide-coverage.

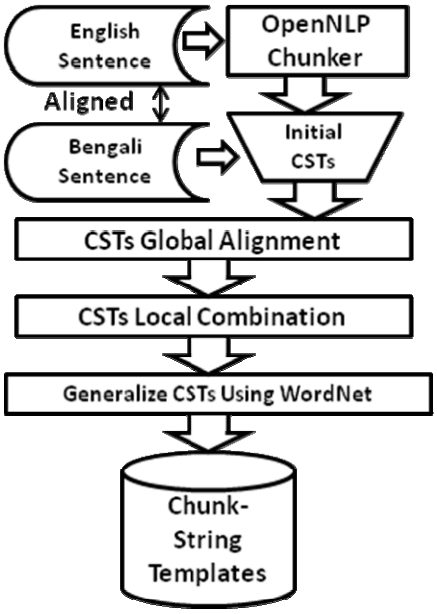


Fig. 2: Steps of CSTs generation

In the first step, using OpenNLP chunker, we prepare chunks of the English sentences from the word aligned English-Bangla parallel corpus.

In the second step, we produced CSTs from the parallel corpus. Table II shows the initial CSTs for the parallel sentence given in Table1. In Table II, c is a chunk in the source language (English), s is a string in the target language (Bangla), and t is the alignment information calculated from the original word alignment.

TABLE II
EXAMPLE OF INITIAL CSTS

CST#	English Chunk (C)	Bangla (S)	T	Align	Chunk-Start-Index
CST1	[NP Bangla/NNP]	□□□ □□	1	1	0
CST2	[VP is/VBZ]	□□□ □□	1	2	1
CST3	[NP the/DT native/JJ language/NN]	□□□ □□□ □□	2	4	2
CST4	[PP of/IN]	–□□	1	6	5
CST5	[NP around/RB 230/CD million/CD people/NNS]	প্রায় □□□ □□□ □□□ □□ □□	1 2 3 4	7 8 9 10	6
CST6	[ADVP worlwide/RB]	□□□ □□□ □□□ □□	1	11	10

In the third step we generate the global alignment information from Initial CSTs as given in Table II, based on the original word alignment as given in Table I. For example, Table III shows the chunk alignment information produced from Table I and Table II.

TABLE III
EXAMPLE OF CSTS GLOBAL ALIGNMENT

CT#	CSTs	Global Alignment
CCST1	CST1 CST2 CST3 CST4 CST5 CST6	CST6 CST1 CST2 CST5 CST4 CST3

In the fourth step CSTs Local Combination generates new possible chunk orders. The goal is to match longer phrases to achieve wide-coverage. Table IV contains the sample Combined-CSTs (CCSTs).

TABLE IV
GENERALIZED CSTS

CT#	CSTs	Local Alignment
CCST1	CST1 CST2 CST3 CST4 CST5 CST6	CST6 CST1 CST2 CST5 CST4 CST3
CCST2	CST1 CST2	CST1 CST2
CCST3	CST4 CST5	CST5 CST4
CCST4	CST3 CST4 CST5	CST5 CST4 CST3

In the fifth step CSTs are generalized by using WordNet to increase the EBMT coverage. To generalize we only consider nouns, proper nouns and cardinal number (NN, NNP, CD in OpenNLP tagset). For each

proper nouns we search in WordNet. If available we replace that NNP with <lexical filename> returned from the WordNet. For example WordNet return <noun.communication> for "Bangla". For cardinal number we simply CDs together to <noun.quantity>. We show example generalized CSTs produced using WordNet in Table V.

TABLE V
COMBINED-CSTS EXAMPLES

CST#	English Chunk (C)	Generalized Chunk
CST1	[NP Bangla /NNP]	[NP <noun.communication>/NNP]
CST5	[NP around/RB 230/CD million/CD people/NNS]	[NP around/RB <noun.quantity> people/NNS]

Finally we get the CSTs database which has three tables: initial CSTs, generalized CSTs and Combined-CSTs. From the example word-aligned parallel sentence of Table I, system generated 6 initial CSTs, 2 Generalized CSTs and 4 Combined-CSTs.

IV. MATCHING ALGORITHM FOR CSTS

From the set of all CSTs we select the most suitable one, according to the following criteria:
1. The more CSTs matched, the better;
2. Linguistically match give priority by follow-ing these ranks, higher level is better:

- Level 4: Exact match.
- Level 3: <lexical filename> of WordNet and POS tags match
- Level 2: <lexical filename> of WordNet match
- Level 1: Only POS tags match
- Level 0: No match found, all unknown words.

For the above example, it chooses CSST1 as it has more CSTs match.

V. UNKNOWN WORD TRANSLATION

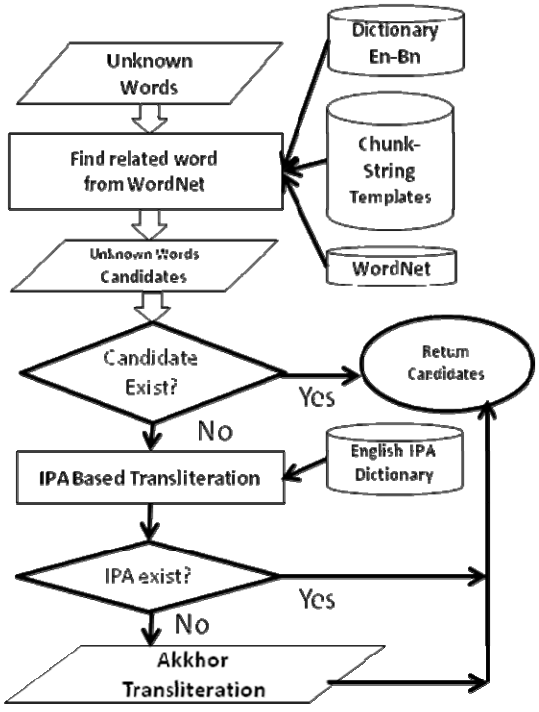


Fig. 3: Steps of Unknown Word Translation

As in our assumption, the main users of this EBMT will be monolingual people; they cannot read or understand English words written in English alphabet. However, with related word translation using WordNet and Transliteration can give them some clues to understand the sentence meaning. As Bangla language accepts foreign words, transliterating an English word into Bangla alphabet, makes that a Bangla foreign word. For example, in Bangla there exist many foreign words, so that user can identify those as foreign words.

Fig. 3 shows the unknown word translation process in a flow chart. Proposed system first tries to find semantically related English words from WordNet for the unknown word. From these related words, we choose the semantically closest related word whose Bangla translation exists in English-Bangla dictionary. If no Bangla translation exists, the system uses IPA-based-transliteration. For proper nouns, the system uses transliteration mechanism provided by Akkhor Bangla Software.

A. Find Semantically Related Word From WordNet

As the parallel corpus is small it is important to have a good method for translating unknown words. When the word has no match in the CSTs, it tries to translate using English WordNet and bilingual dictionary for English-BANGLA. WordNet provide related word for nouns, proper nouns, verbs, adjectives and adverbs. For nouns and verbs WordNet provide hypernyms, which is defined as follows:

Y is a hypernym of X if every X is a (kind of) Y.

For example "canine" is a hypernym of noun "dog", because every dog is a member of the larger category of canines. Verb example, "to perceive" is an hypernym of "to listen". However, WordNet only provides hypernym(s) of a synset, not the hypernym tree itself. As hypernyms can express the meaning, we can translate the hypernym of the unknown word. To do that, until any hypernym's Bangla translation found in the English-Bangla dictionary, we keep discovering upper level of hypernym's. Because, nouns and verbs are organized into hierarchies, defined by hypernym or IS-A-relationships in WordNet. So, we considered lower level concept is generally more suitable then the higher level words.

This process discovers the hypernym tree from WordNet in step by step. For example, from the hypernym tree of "dog" from WordNet, we only had the "animal" entry in our English-Bangla dictionary. Our system discovered the hypernym tree of "dog" from WordNet until "animal". Following is the discovered hypernym tree:

dog, domestic dog, Canis familiaris
=> canine, canid
=> carnivore => placental, placental mammal
=> mammal => vertebrate, craniate => chordate
=> animal => ...

This process search in English-Bangla dictionary, for each of the entry of this hypernym tree. As we only had the entry for "animal", we translated "dog" as the

translation of "animal", which is "পশু" [poshu] in Bangla. Similarly, for adjectives we try to find "similar to" words from WordNet. And for Adverbs we try to find "root adjectives".

This step returns Unknown Words candidates from WordNet which exist in English-Bangla dictionary.

B. Transliterate If No Candidate Found From WordNet

When unknown word is not even found in WordNet, we use IPA-Based transliteration using the English IPA Dictionary as described in section VI.

However, when unknown word is not even found in the English IPA dictionary, we use transliteration mechanism of Akkhor Bangla Software. For example, for the word "Muhammod" which is a popular Bangla name, Akkhor transliterated into "মুহাম্মদ" in Bangla.

VI. IPA-BASED TRANSLITERATION

English words pronunciations in IPA obtained from the English IPA dictionary. Output for this step is the Bangla word transliterated from the IPA of the English word. In this step, we use following English-Bangla Transliteration map to transliterate the IPA into Bangla alphabet. Table VI, VII and VIII shows our proposed English-Bangla IPA chart for vowels, diphthongs and consonants. Using rule-base we transliterate the English IPA into Bangla alphabets. The above IPA charts leaves out many IPA as we are considering about translating from English only. To translate from other language such as Japanese to Bangla we need to create Japanese specific IPA transliteration chart. Using the above English-Bangla IPA chart we produced transliteration from the English IPA dictionary. For examples: pan(pæn)*: প্যান;* ban(bæn)*: ব্যান;* might(ma**I**t)**:** মাইট .

TABLE VI
ENGLISH-BANGLA IPA CHART FOR VOWELS

Mouth narrower vertically	[i:] ই / ি sleep /sli: p/	[I] ই / ি slip /sl**I**p/	[ʊ] উ / ু book /bʊ k/	[u:]ঊ/ ূ boot /bu: t/
	[e] এ / ে ten /ten/	[ə] আ / া after /a: ftə /	[ɜ :] আ / া bird /bɜ : d/	[ɔ :] র্ bored /bɔ : d/
Mouth wider vertically	[æ]এ্যা/্যা cat /kæt/	[ʌ] আ /া cup / k^p/	[ɑ :] আ / া car / cɑ : r/	[ɒ] অ hot /hɒ t/

TABLE VII
ENGLISH-BANGLA IPA CHART FOR DIPHTHONGS

[**I**ə] ইয়া/িয়া beer /b**I**ə r/	[e**I**] এই/ েই say /se**I**/	
[ʊ ə] উয়া/ ুয়া fewer /fjʊ ə r/	[ɔ **I**] অয়/য় boy /bɔ **I**/	[ə ʊ] ও / ো no /nə ʊ /
eə ঈয়া/ীয়া bear /beə r/	[a**I**] াই / আই high /ha**I**/	[aʊ]আউ /াউ cow /kaʊ /

TABLE IX
ENGLISH-BANGLA IPA CHART FOR CONSONANTS

[p] প pan /pæn/	[b] ব ban /bæn/	[t] ট tan /tæn/	[d] ড day /de**I**/	[ʧ] চ chat /ʧ æt/	[ʤ] জ judge /ʤ^ʤ/	[k] ক key /ki: /	[g] গ get /get/
[f] ফ fan /fæn/	[v] ভ van / væn/	[θ] থ thin /θ **I**n/	[ð] দ than /ðæn/	[s] স sip /s**I**p/	[z] জ zip / z**I**p/	[ʃ] শ ship /ʃ **I**p/	[ʒ] স vision /v**I**ʒ ^n/
[m] ম might /ma**I**t/	[n] ন night /na**I**t/	[ŋ]ং/ঙ thing /θ **I**ŋ /	[h] হ height /ha**I**t/	[l] ল light /la**I**t/	[r] র right /ra**I**t/	[w] য় white /hwa**I**t/	[j]ইয়/িয়ে yes /jes/

VII. AKKHOR TRANSLITERATION

Akkhor Bangla Software first implemented Bangla phonetic input method for computers. This phonetic mapping became very popular among Bangladeshi computer users. In this research we used Akkhor phonetic mapping with Bangla Lexicon database. Table IX shows the phonetic mapping for Bengali alphabets. For example, Akkhor transliterates "onigiri" as "ওনগিরি".

TABLE IX
AKKHOR PHONETIC MAPPING FOR BENGALI ALPHABETS

বাংলা	অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
English	A	a/aa/^a	i/^i	I/ee/`I	u/^u	U/^U	ri/^ri	e/^e	oi/`oi	o/^o	ou ou

বাংলা	ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ
English	k	kh	g	gh	Ng	ch	Ch	j	jh	Y
বাংলা	ত	থ	দ	ধ	ন	ট	ঠ	ড	ঢ	ণ
English	t	th	d	dh	n	**T**	Th	D	Dh	N
বাংলা	প	ফ	ব	ভ	ম	য	র	ল	শ	ষ
English	p	f/ph	b	bh/v	m	**z**	r	l	sh	S
বাংলা	গ	ক্ষ	হ	ড়	ঢ়	য়	ং	ঃ	ঁ	
English	S	k-S	h	R	rh	y	ng	:	~	
বাংলা	১	২	৩	৪	৫	৬	৭	৮	৯	০
English	1	2	3	4	5	6	7	8	9	1
বাংলা	কা	কে	কি	কু	কো	ক্র	ক্রে	ক্রি	ক্রু	ক্রু
English	ka	ke	ki	ku	kO	kro	kre	kree	kru	krU
বাংলা	কী	চী	মী	কূ	মূ	বূ	ণূ	নূ	ক্য	ব্য
English	kI	chI	mI	kU	mU	bU	NU	nU	k-z	b-z

VIII. EXPERIMENT

We did quality evaluations for the proposed EBMT with unknown words, by comparing with baseline EBMT system. Quality evaluation measures the translation quality through human evaluation.

Baseline system architecture has the same components as described in Fig. 1, except for the components inside dotted rectangles. Matching algorithm of baseline system is that not only match with exact translation examples, but it can also match with POS tags. The Baseline EBMT use the same training data: English-BANGLA parallel corpus and dictionary, but does not use CSTs, WordNet and unknown words translation solutions.

Currently from the training data set of 2000 word aligned English-BANGLA parallel corpus, system generated 15356 initial CSTs, 543 Generalized CSTs and 12458 Combined-CSTs.

The development environment was in windows using C Sharp language. Out test-set contained 336 sentences, which are not same as training data. The test-set includes simple and complex sentences, representing various grammatical phenomena. We have around 20,000 English-BANGLA dictionary entries.

Quality evaluation measures the translation quality through human evaluation. Perfect Translation means there is no problem in the target sentence, and exact match with test-set translation. Good Translation means not exact match with test-set reference, but still understandable for human. Medium means there are several problems in the target sentence, like wrong word choice and wrong word order. Poor Translation means there are major problems in the target sentence, like non-translated words, wrong word choice and wrong word order. Table IX shows the human evaluation of current system. Currently 25.33% of the test-set translations produced by the system were good translation. Around 16 points of poor or medium translations produced by EBMT Baseline was improved using the proposed unknown word translation mechanism.

TABLE IX
HUMAN EVALUATION USING SAME TESTSET

Translation Quality	Grade	Baseline EBMT %	EBMT+ Unknown Words	Google MT (En-Bn)
Perfect Translation	A	22.67	22.67	11.00
Good Translation	B	25.33	41.33	18.67
Medium Translation	C	19.67	21.00	26.33
Poor Translation	D	32.33	15.00	44.00
Total		**100%**	**100%**	

In our test-set we included translation examples with proper noun or unknown words. From the above evaluation we can clearly see the limitation of Google MT with translating unknown words. Due to those unknown words most of the translation output of Google MT was in poor category. However, Google use statistical MT approach and obvious they have trained their system with different training data set. In the case of Bangla or other low-resource language, Google or any machine translation system needs to consider about translating unknown words. From our experiment we can say that, our proposed solution is very effective for translating unknown words.

The identified main reason for improving the translation quality is our solution for translating unknown words. For example, even though "dog" was an unknown word, using our solution, it can be translated as "animal". As a result, during quality evaluation some test-set sentence improved from "poor" or "medium" to "good" translation.

We observed some drawbacks of using WordNet as well. Sometimes our system chooses the wrong synset from the WordNet. As a result, some test-set still produced "poor" translation. On the other hand CSTs played a major role in sub-sentential match. As a result it helped to translate grammatically similar structured sentences as "perfect" or "good" translation. Drawbacks of using CSTs are high computational complexity and big memory requirement for larger parallel corpus.

In English, there are four types of sentences: Declarative, Imperative, Interrogative and Exclamatory sentences. This sentence types further fall into four basic sentence type: Simple, Compound, Complex and Compound-Complex. The Table X gives approximate status of implementation for each sentence type. It shows the performance of the translated texts by our current EBMT system for grammatical structures. Here A,B,C,D represent perfect, good, medium and poor translation same as Table IX.

TABLE X
GRAMMATICAL STRUCTURES

Sentence Type	Declarative	Imperative	Interrogative	Exclamatory
Simple	A	A	B	B
Compound	B	B	B	B
Complex	B	B	C	C
Compound - Complex	C	C	D	D

TABLE XI
SAMPLE SYSTEM PRODUCE TRANSLATIONS COMPARISON

#	English	EBMT Baseline	Proposed EBMT Unknown Words	Google
1.	Are you looking for an aardvark?	□□□□ □□ কে·ান aardvark □□□□ছনে ?(D)	□□□□ □□ কে·ান পশু □□□□ছনে?(B)	□□□□ □□□□ aardvark □□□□□□? (D)
2.	This dog is really cool.	Dog টী আসলেই □□□□□ (D)	পশুটী আসলেই □□□□□ (B)	□□ □□□□□ □□□ □□ □□□□□□ □□□□□□. (D)
3.	I am eating onigiri	আমি onigiri খাচ্ছি (D)	আমি ওনগিরি খাচ্ছি(A)	□□□ onigiri □□□ □□□□□□ (D)
4.	His name is Rupok.	□□□ □□□ Rupok. (D)	□□□ □□□ □□□□□(A)	□□□ □□□ Rupok. (D)
5.	What is abstriction?	abstriction □□? (D)	এ্যাবস্ট্রকিশান □□? (B)	abstriction □□? (D)

Table XI shows sample translation produced from English to Bangla language by EBMT Baseline, Proposed

EBMT with Unknown Words solution and Google. It also shows the translation quality in bracket (A,B,C,D: same as previous meaning) which we prepared for our quality evaluation. As "aardvark" and "dog" has no match CSTs and dictionary, EBMT baseline produced poor translation for #1 and #2. Our WordNet solution improved these two translation into good category. "onigiri" is a Japanese food name and "rupok" which is a person name are unknown words in #3 and #4. Our Akkhor transliteration solution improved these from poor to perfect translation. In the case of #5, "abstriction" is an unknown word, which the system translated using the proposed IPA-based transliteration solution. As a result the translation improved to good quality. As we can see Google with the biggest English-Bengali parallel corpus and statistical approach has poor translation quality for all these 5 examples in Table XI. All these sample examples demonstrate the effectiveness of our proposed solution for translating unknown words.

IX. Conclusion and Future Works

We proposed an EBMT system for low-resource language using CSTs in the example-base. Our EBMT system is effective for low resource language like BANGLA. Using Unknown Words solution we improved the quality of our EBMT system. We used WordNet to translate the unknown words which are not directly available in the dictionary. And then we used IPA-based transliteration mechanism for the rest unknown words.

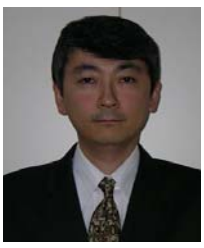
Currently we used a small parallel corpus to generate CSTs. However to increase the performance we need a balanced parallel corpus [7]. Although current system works well for small parallel corpus, the performance can decrease with larger parallel corpus. Because it will have many candidate CSTs.

In future, we want to improve the CSTs selection mechanism. We plan to use statistical language model for future improvement. It can improve the generation quality. In future we also want to evaluate the system using BLEU and other standard Machine Translation evaluation metrics.

References

[1] Abney, Steven. 1991. Parsing by chunks. In Principle-Based Parsing, pages 257–278. Kluwer Academic Publishers.
[2] Diganta Saha, Sivaji Bandyopadhyay. 2006. A Semantics-based English-Bangla EBMT System for trans-lating News Headlines. Proceedings of the MT Summit X, Second workshop on Example-Based Machine Translation Programme.
[3] Diganta Saha, Sudip Kumar Naskar, Sivaji Bandyopadhyay. 2005. A Semantics-based English-Bangla EBMT System for translating News Headlines, MT Xummit X.
[4] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.
[5] Jae Dong Kim, Ralf D. Brown, Jaime G. Carbonell. 2010. Chunk-Based EBMT. EAMT, St Raphael, France.
[6] Khan Md. Anwarus Salam, Mumit Khan and Tetsuro Nishino. 2009. Example Based English-Bangla Machine Translation Using WordNet. TriSAI, Tokyo.
[7] Khan Md. Anwarus Salam, Yamada Setsuo and Tetsuro Nishino. 2010. English-Bangla Parallel Corpus: A Proposal. TriSAI, Beijing.
[8] Khan Md. Anwarus Salam, Yamada Setsuo and Tetsuro Nishino. 2011. Translating Unknown Words Using WordNet and IPA-Based-Transliteration. ICCIT, Dhaka .
[9] Md. Zahurul Islam, Jörg Tiedemann & Andreas Eisele. 2010. English to Bangla phrase-based machine translation. Proceedings of the 14th Annual confe-rence of the European Association for Machine Translation.
[10] R. Gangadharaiah, R. D. Brown, and J. G. Carbonell. Phrasal equivalence classes for generalized corpus-based machine translation. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 6609 of Lecture Notes in-Computer Science, pages 13–28. Springer Berlin / Heidelberg, 2011.
[11] Sajib Dasgupta, Abu Wasif and Sharmin Azam. 2004. An Optimal Way Towards Machine Translation from English to Bangla, Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT).
[12] Sudip Kumar Naskar, Sivaji Bandyopadhyay. 2006a. A Phrasal EBMT System for Translating English to Bangla. Workshop on Language, Artificial Intelligence and Computer Science for Natural Language Processing applications (LAICS-NLP).
[13] Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2006b. Handling of Prepositions in English to Ben-gali Machine Translation. In the proceedings of Third ACL-SIGSEM Workshop on Prepositions, EACL 2006. Trento, Italy.
[14] Zhanyi Liu, Haifeng Wang And Hua Wu. 2006. Example-Based Machine Translation Based on Tree-string Correspondence and Statistical Generation. Machine Translation, 20(1): 25-41
[15] http://www.netz-tipp.de/languages.html
[16] http://hdr.undp.org/en/reports/global/hdr2009/
[17] http://www.akkhorbangla.com
[18] anubadok.sourceforge.net
[19] http://translate.google.com/#en|bn|
[20] http://wordnetweb.princeton.edu/perl/webwn

Khan Md. Anwarus Salam received the M.S. degree in Information and Communication Engineering from the same university in 2011. He received B.Sc. in Computer Science from BRAC University, Bangladesh in 2009. He is currently a PhD student in The University of Electro-Communications, Tokyo. His research interests include natural language processing and machine translation for Bangla language. He is a member of IEEE and AAMT.



Setsuo Yamada received the M.S. degree in Information Science from Tokyo Denki University in 1992 and the Ph.D. degree in Engineering from Tottori University in 2006. From 1992, He works for Nippon Telegraph and Telephone Corporation. From 1997 to 2000, he worked for ATR Spoken Language Translation Research

Laboratories. He currently works for Nippon Telegraph and Telephone Corporation. His research interests include natural language processing, especially machine translation. He is a member of ANLP and IPSJ.

Tetsuro Nishino graduated from department of Mathematics, Waseda University and continued his research and obtained a D.Sc. in Mathematics in 1991. He was a researcher at Tokyo Research Laboratory, IBM Japan during 1984-1987. He was an Associate Professor at School of Information Science, Japan Advanced Institute of Science and Technology, Hokuriku during 1992-1994. In 1994, he joined at Department of Communications and Systems Engineering, The University of Electro-Communications as an Associate Professor and in 2006 he became a Professor. He received The Funai Information Technology Prize (2003) and IBM Faculty Award (2008).