Statistically Motivated Example-based Machine Translation using Translation Memory

Sandipan Dandapat, Sara Morrissey, Sudip Kumar Naskar, and Harold Somers
Centre for Next Generation Localisation
Dublin City University
Glasnevin, Dublin 9, Ireland
{sdandapat, smorri, snaskar, hsomers}@computing.dcu.ie

Abstract

In this paper we present a novel way of integrating Translation Memory into an Example-based Machine Translation System (EBMT) to deal with the issue of low resources. We have used a dialogue of 380 sentences as the example-base for our system. The translation units in the Translation Memories are automatically extracted based on the aligned phrases (words) of a statistical machine translation (SMT) system. We attempt to use the approach to improve translation from English to Bangla as many statistical machine translation systems have difficulty with such small amounts of training data. We have found the approach shows improvement over a baseline SMT system.

1 Introduction

In 1988 IBM's Peter Brown presented a 'purely statistical' approach to machine translation (MT) (Brown et al., 1988) at the Second TMI conference at Carnegie Mellon University. Since then most MT research has been carried out in the SMT paradigm. In parallel a lot of research has also been rooted in the EBMT paradigm, following the framework proposed by Nagao (Nagao, 1984). Furthermore, since the development of Moses, an open-source toolkit for SMT (Koehn et al, 2007), people seem to have moved away from EBMT with the exception of a few recent works (Kim et al., 2010; Phillips and Brown, 2009; Smith and Clark, 2009; Somers et al., 2009; Farwell and Padro, 2009; Lepage et al., 2008; Bosch et al., 2007). Although both SMT and EBMT systems are data-driven approaches to MT, each of them has their own strengths and limitations. Typically, an SMT system works well when a significant amount of training data (i.e. parallel bilingual corpus) is available for the language pair. SMT approaches also been shown to have some difficulties with free word order languages (Khalilov et al., 2010). In contrast, an EBMT approach can be developed with a limited example-base; also an EBMT system works well when training and test set are quite close in nature. This is because EBMT systems reuse the segment of a test sentence that can be found in the source side of the example-base.

EBMT is often linked with another related technique, namely "Translation Memory" (TM). The commonality between EBMT and TM is the idea of the reuse of examples from the existing translations. The main difference between EBMT and TM is that EBMT is an automated technique for translation whereas TM is an interactive tool for the human translator.

In this paper, we describe our attempt to use EBMT and TM to tackle the English–Bangla language pair which had proved troublesome with low BLEU scores for various SMT approaches (Islam et al., 2010). We attempt to translate medical-receptionist dialogues, primarily appointment scheduling. Our first task was to collect an English-language corpus of patient-receptionist dialogue to develop the English–Bangla parallel corpus. A major difficulty in collecting genuine data in the medical field, or any domain where personal information is involved, is that the confidentiality and other ethical issues more or less preclude using genuine data collected *in situ*. Thus, it is important to develop an MT system based on a limited amount of data which can be collected by simulating the appointment scheduling with a medical secretary.

Keeping the above difficulties in mind, we are inspired to use an EBMT approach. In this paper we examine different experiments and report which is the most meaningful for this problem of data sparseness.

The rest of the paper is organized as follows. The next section presents related research in the

Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2010

area of TM and some research with Indian Language MT. Section 3 describes some related issues for translating medical domain data. Section 4 describes the detail of our EBMT-based approach. Section 5 presents the experimental setup, the data and the results obtained with different experiments. Section 6 presents the observations with some error analysis. We conclude in section 7.

2 Related Work

The original idea of TM was developed by Martin Kay in his well-known 'Proper Place' paper (Kay, 1980). A TM essentially stores source- and target-language translation pairs for effective reuse of the previous translations. TM is often used to store examples of EBMT systems. TM is also widely used in computer-aided translation (CAT) systems to help the professional translators. EBMT systems first find the example (or a set of examples) from the TM which most closely matches the source-language string to be translated (Somers, 2003). After retrieving a set of examples, with associated translations, EBMT systems automatically extract the translation of the proper fragment and combine them to produce a grammatical target output. On the other hand, CAT systems segment the input text to be translated and compare each segment against the translation units (TU) in the TM (Bowker, 2002). The CAT systems produce one or more target equivalent for the source segment and professional translators select and recombine them (perhaps with modification) for the desired translation. Both EBMT- and CAT- based systems (Somers, 2003) are developed based on idea similar premise but in an EBMT approach selection and recombination is done automatically to produce the translation without the help of a professional translator.

Phrase-based SMT systems (Koehn, 2009), produce a source-target aligned phrase table which could be adapted as a TM to a CAT environment (Bourdiallet, 2009). To the best of our knowledge, so far no one has used SMT phrase tables within an EBMT system as a TM. However, it is worth mentioning that some work has been carried out to integrate MT in a CAT environment to translate the whole segment using the MT system when no matching TU is found in the TM. The TransType system (Langlais et al., 2000; Langlais et al., 2002) integrates a SMT system within a text editor. Our approach attempts to integrate the TM (obtained from SMT system) within an EBMT system.

2.1 Previous Work on Indian Language MT

Although a lot of work has been carried out for English to Indian Language (IL) translation using rule-based MT, some recent works focus on data-driven MT for ILs (Islam et al., 2010; Ramanathan et al., 2009; Ramanathan et al., 2008; Ambati and Rohini, 2007; Naskar and Bandyopadhyay, 2005). However, very little work has been carried out for English to Bangla data-driven MT (Islam et al., 2010; Naskar and Bandyopadhyay, 2005). Bangla is the main spoken language in Bangladesh, the second most commonly spoken language in India and the sixth most commonly spoken language in the world. Islam et al. (2010) describes a phrase based SMT system for English to Bangla translation using different corpora. An EBMT system has been developed by identifying source phrases using syntactic processing and the target is generated using a phrasal example-base (Naskar and Bandyopadhyay, 2005). As far as we can tell, no work has been carried out for the medical domain especially translation of dialogues for appointment scheduling. Also to our knowledge, there has been no research on TM or a combination of TM with EBMT for Bangla.

3 Issues with Medical Domain

The development of an MT system for the medical-receptionist dialogue has three main potential difficulties which we will explain here.

3.1 Corpora Collection and Translation

Our first task is to collect an English-language corpus of patient-receptionist dialogue. It is difficult to collect medical data due to the involvement of personalized information. Thus there exists a lot of confidentiality and other ethical issues. This difficulty has long been recognized in medical training, where "standardized patients" (SPs) are used with medical students, that is, actors trained to simulate consistently the responses of a patient in a particular medical settings. Training SPs is of course a major undertaking in itself necessarily involving experts, so for the purpose of our work we made a compromise in that we engaged an experienced GP's receptionist to participate in a number of role-play sessions with the native English speakers. These are all recorded and later transcribed. Thus, we believe that our corpus contains samples that

are realistic, and offer a broad coverage of our target domain. Due to the involvement of the aforementioned stages, it is time-consuming and expensive to collect a large amount of medical-receptionist dialogue. Thus our corpus comprises 380 dialogue turns. In transcription, this works out at just under 3,000 words (a very small corpus by any standard), each sentence turn is on average 8 words.

The next stage in the process was to manually translate our English corpus in to Bangla. Translation between any languages, whether related or not, involves cases where closely following the source text (a "literal" translation, within the grammatical constraints of the target language) can result in a stilted, unnatural or incorrect translation. This is especially the case when translating medical-receptionist dialogue between English and Bangla which has a wide syntactic difference. This has a serious implication for our approach to MT. A good example is the dialogue in (1):

(1) a. Which doctor would you prefer?
b. I don't mind.

The response (1b) can be translated in the following ways in (2).

(2) a. আমি কিছু মনে করব না।
Ami kichhu mane karba nA.[1]
I SOME MIND DO-FUTURE NOT.
I don't mind.
b. যে কোনো একজনকে দেখালেই হবে।
ye kono ekajanake dekhAlei habe.
WHO-RELATIVE ANY "ONE PERSON"-ACCUSATIVE SHOW-CONDITIONAL PARTICIPLE-EMPHATIC BE-FUTURE.
Can see either of them.

The literal translation (2a) without the context would be misleading or meaningless. Thus, in Bangla, a literal translation (2a) is less preferable than a more explicit translation (2b). We are keeping (2b) as the translation of (1b) though none of the English words has an equivalent in the target side. This scenario might affect an SMT system trained on such a corpus. However, to ensure the closeness and fluency in the dialogue we concentrate on meaningful translation in context. This issue occurs quite frequently while translating English dialogue corpus to Bangla. Along with this we have found other difficulties such as lexical choice and translating borrowed words.

3.2 Size and Type of the Corpora

Since both SMT and EBMT are data-driven machine translation techniques, the first thing needed is the machine readable parallel corpus. To adopt any data-driven machine translation system, there is also the question of the *size* of the example database: how many examples are needed? As reported in the previous section, our corpus is particularly small in comparison with standard data-driven training corpora which can sometime enter into the millions of parallel training sentences. The SMT models are based on probability distribution which requires a significant bi-corpus otherwise it suffers from the sparse data problem. As far as we can tell, there is no SMT system developed with only 380 parallel sentences. In contrast, many EBMT systems have been developed with such a small corpus. Table 1 lists some of the EBMT systems developed using a small amount of parallel data (details can be found in Somers, 2003).

System	Language Pair	Size
TTL	English → Turkish	488
TDMT	English → Japanese	350
EDGAR	German → English	303
ReVerb	English → German	214
ReVerb	Irish → English	120
METLA-1	English → French	29
METLA	English → Urdu	7

Table 1: Size of example database in EBMT systems

The sizes of the example database in Table1 motivate us to develop an EBMT system for medical-receptionist dialogue even when a very small amount of parallel data is available.

We have found that medical-receptionist dialogue is comprised of very similar sentences. This is illustrated in example (3) and (4).

(3) a. I need a medical for my *insurance company*.
b. I need a medical for my *new job*.
(4) a. The doctor told me to *come* back *for* a *follow up appointment*.
b. The doctor told me to *call* back *in* a *week*.

[1] In this paper we use ITRANS convention to transcribe Bangla script: http://www.aczoom.com/itrans/.

The portions in italics are the only difference between (a) and (b). Thus, it might be helpful to reuse the translation of the common part while translating a new sentence. The above observation leads us to use EBMT to reuse some parts of the sentence which are common with the closest sentence in the example-base.

3.3 Building Translation Memory

Building high quality TM is an expensive and time consuming process (Marcu, 2001). As we have no access to any TM for the medical domain, we decided to build a TM automatically from our small patient-dialogue corpus. We use Moses[2] to automatically build this TM. Based on Moses word alignment (using GIZA++) and a phrase table, we construct two TMs for further use. Our first TM is based on the aligned phrase pairs from the Moses phrase table (PT). It has been observed that a source phrase has multiple target equivalents. We keep all the target equivalents in a sorted order based on the phrase translation probability. This essentially helps us in the matching procedure (see section 4.1) but during recombination we only consider the most probable target equivalent (see section 4.3). We shall call this **PT** in the rest of the paper. Another TM is created based on the word aligned file from source to target language. Like PT, we also keep the multiple target equivalents for a source word in a sorted order. This essentially creates a lexical table (LT) of source- and target-language equivalent word pairs. We shall call this **LT** in the rest of the paper.

4 Our Approach

Like all other EBMT system our particular approach is also comprised of three stages - *Matching*, *Adaptability* and *Recombination*.

4.1 Matching

The first step in an EBMT system is to find source language example(s) that closely match with the input sentence. In particular in our approach, we find *the closest sentence* (s_c) from the example-base for the input sentence (s) to be translated. We have used a word-based edit distance metric (Levenshtein, 1965; Wagner and Fischer, 1974) to find this closest match sentence from the example-base ($\{s_i\}_1^N$) based on the following equation.

$$score(s, s_i) = 1 - \frac{ED(s, s_i)}{\max(|s|, |s_i|)} \qquad (1)$$

where $|x|$ denotes the length (in words) of a sentences and ED(x,y) refers to the word based edit distance between x and y.

Based on the above fuzzy scoring criteria, we are able to choose the closest match for the input sentence to be translated. For the two input sentences in (5) the corresponding closest fuzzy-matched sentences from the example-base are given in (6) respectively.

(5) a. Ok, I have booked you in for eleven fifteen on Friday with Dr. Thomas.
b. Is it possible to book an appointment later this week?

(6) a. Ok, I have booked you in for three thirty on Thursday with Dr. Kelly.
b. Is it possible to book an appointment with the nurse?

Then we consider the associated translation (s_c^t) in (7) from the example-base for the closest match source sentence in (6) as the skeleton translation of the input sentences to be translated (5).

(7) a. আচ্ছা , আমি আপনার জন্য বৃহস্পতিবার তিনটে তিরিশে ডাঃ কেলির সাথে বুক্ করেছি ।
AchchhA , Ami ApanAra janya bRRihaspatibAra tinaTe tirishe DAH kelira sAthe buk karechhi.
WELL, I YOU-OBLIQUE FOR THURSDAY THREE THIRTY-LOCATIVE DR. KELLY-OBLIQUE WITH BOOK DO-FIRST PERSON-PRESENT PARTICIPLE.

b. নার্সের সাথে একটা অ্যাপয়ন্টমেন্ট বুক্ করা সম্ভব কি ?
nArsera sAthe ekaTA aypaYnTamenTa buk kara sambhaba ki ?
NURSE-OBLIQUE WITH ONE APPOINTMENT BOOK DO-GERUND POSSIBLE “IS IT”?

We will use some segment of the associated skeleton translation (7) to produce the new translation in the following two subsections.

[2] Moses (http://www.statmt.org/moses/) is a SMT system that automatically trains a translation model for any language pair. The *lexical translation table* (of training step 4) and *score phrases* (of training step 6) are used to build our TMs.

4.2 Adaptability

After matching and retrieving an example, with its associated translation, the next step is to extract from that translation the inappropriate fragments. In order to do that, we align the three sentences - the input (*s*), the closest source-side match (s_c) and its target equivalent (s_c^t).

First, we mark the mismatch portion between *s* and s_c while computing the edit distance in equation (1). This is shown in example (8) with angular brackets (the numbers within angular brackets index the mismatched segments). Further we align the mismatch portion of s_c with its associated translation s_c^t . We use the two translation memories PT and LT to find the match between s_c and s_c^t . Based on the source-target aligned pair from the PT and LT we mark the mismatch segment in the s_c^t as in (8c). The portions marked with angular brackets in (8c) are aligned with the mismatch portion in (8b). Here also, the numbers within the angular brackets in s_c^t is the mapping between the segments with s_c.

(8) a. Ok, I have booked you in for <0:*eleven fifteen*> on <1:*Friday*> with Dr. <2:*Thomas*>.
b. Ok, I have booked you in for <0:*three thirty*> on <1:*Thursday*> with Dr. <2:*Kelly*>.
c. আচ্ছা , আমি আপনার জন্য <1:বৃহস্পতিবার> <0:তিনটে তিরিশ>ে ডাঃ <2:কেলি>র সাথে বুক্ করেছি ।

With the help of above matching, in the recombination step, we will replace the segments within the angular brackets keeping the remaining matched fragments unchanged.

4.3 Recombination

The final step of our EBMT approach is recombination, to add or substitute the segments from the input sentence (*s*) with the skeleton translation equivalent (s_c^t). From the example (8), we need to replace the three segments in (8c) {বৃহস্পতিবার (*bRRihaspatibAra* [Thursday]), তিনটে তিরিশ (*tinaTe tirishe* [three thirty]) and কেলি (*keli* [Kelly])} with three corresponding source segments in (8a) {*eleven fifteen*, *Friday* and *Thomas*} to produce a target equivalent. Thus keeping the mapping we produce (9).

(9) আচ্ছা, আমি আপনার জন্য <1:*Friday*> <0:*eleven fifteen*>ে ডাঃ <2:*Thomas*>র সাথে বুক্ করেছি ।

Note that there might not be a one-to-one correspondence between *s* and s_c^t . If there is some extra segment in *s* which does not have any mapping in s_c^t then we add the new segment from *s* in target equivalent (s_c^t). If there is an extra segment in s_c^t then we simply delete them from the target translation. Thus we have produced the target equivalent as in (9) after adding/deleting/substituting segments form the input sentence to be translated (*s*) with the skeleton translation (s_c^t). Then, we translate the new untranslated segments in (9). We translate these new segments in two ways to understand their relative performance. First, we translate them using the Moses decoder and putting them in the desired positions we produce the target equivalent. Second, we use the PT and LT to produce the translation of the un-translated segments. The detail of the algorithm is given in Figure 1.

Algorithm 1 recombination(X, PT, LT)

In: Source-languge string to be translated $(X = x_1^n)$, Translation Memories(PT, LT)
Out: Target-language translation T_X

1: **if** the source phrase X exists in the PT **then**
2: $T_X \leftarrow PT(X)$
3: return T_X
4: **else**
5: translate longest possible matched segment x_1^m from the PT recursively until no further segments can be translated using PT
$T_X \leftarrow x_1 \ldots x_i PT(x_1^m) x_{i+m+1} \ldots x_n$
6: **if** there exist any un-translated word x_i in X **then**
7: translate using LT
$T_X \leftarrow x_1 \ldots x_{i-1} LT(x_i) x_{i+1} \ldots x_n$
8: **else**
9: return T_X
10: **end if**
11: **end if**

Figure 1: Algorithm for recombining the fragment from the aligned sentences (*s*, s_c *and* s_c^t).

Replacing the untranslated segment in (9) with the translation obtained using the PT and LT, we get the output translation of the original input sentence in (10).

(10) আচ্ছা , আমি আপনার জন্য <*শুক্রবার*> <*এগারটা আসতে*ে> ডাঃ <*Thomas*>র সাথে বুক্ করেছি ।

Note that unknown words are not translated with the approach which is mostly the case for all MT techniques. Incorrect translations are obtained due to incorrect word/phrase alignments.

5 Experiments

First, we conduct two different experiments to estimate the baseline accuracy of our approach for the English to Bangla translation task.

1. We use MaTrEx (Stroppa and Way, 2006; Dandapat et al., 2010), an open-source SMT system **OpenMaTrEx**[3] and compare the results with our approach.
2. Based on the matching step (section 4.1) of the EBMT, we obtain the closest target-side equivalent (the skeleton sentence) and consider this as the baseline output for the input to be translated. This is referred to as **EBMT** in the experiments below. We will consider this as the baseline accuracy for our approach.

We have conducted different experiments using different TM and SMT decoding. After obtaining the skeleton translation, in the *recombination step*, we use two different translation memories (PT and LT) to translate unmatched segment based on the algorithm described in Figure 1. Thus, we have two more systems:

3. The recombination step of the EBMT uses the PT as the only example-base. We refer to this as **EBMT+TM(PT)**.
4. Both PT and LT have been used as example-bases by the EBMT system. We call this **EBMT+TM(PT,LT).**

Finally, instead of using PT and LT to translate the unmatched portion of the input sentence, we use the SMT decoder to obtain the translation. Thus, we obtain the translation of X (= x_1^n) in Figure 1 using T_X = SMT(X).

5. We translate the unmatched segment of the input using SMT for the recombination. We shall call this as **EBMT+SMT.**

Thus, we have a total of 5 systems under different combination schemes of EBMT, TM and SMT. The same training and test set has been used to evaluate the systems.

[3] http://www.openmatrex.org/

5.1 Data used for the Experiments

For all the experiments we have used the same English-Bangla corpus described in section 3.1. The training data consist of 381 parallel sentences. A fixed set of 41 sentences disjoint from the training set have been used to test the systems. The test set is also a dialogue exchange between a patient and the receptionist. Note that none of the sentences from the test sentence belongs to the training set though the test examples are quite close to the example-base.

5.2 Results

We have used both automatic metrics and manual evaluation methods. The widely used BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) score is used for automatic evaluation of our systems. BLEU score captures the fluency of the translation while the translation adequacy is captured by NIST score. Table 2 summarizes the accuracy obtained with the different systems.

System	BLEU	NIST
SMT	39.32	4.84
EBMT	50.38	5.32
EBMT+TM(PT)	57.47	5.92
EBMT+TM(PT,LT)	57.56	6.00
EBMT+SMT	52.01	5.51

Table 2: System accuracies obtained by different models

Note that the baseline SMT BLEU score is 39.32 but the baseline EBMT gives a BLEU score of 50.38%. This absolute improvement of 11.06 BLEU point motivates us to use skeleton sentences and make further changes to some fragment in the skeleton sentences with respect to the original input sentence to be translated. We have achieved 57.47 BLEU score with our EBMT+TM(PT) system – an absolute improvement of 7.09 in BLEU score over the baseline EBMT score. However, the use of LT along with the PT (EBMT+TM(PT,LT)) has a little improvement in both BLEU (absolute 0.09) and NIST (absolute 0.08) score over the EBMT+TM(PT) system. On the contrary, the EBMT+SMT system has an accuracy of 52.01 BLEU score which reflects an improvement of 1.63 absolute BLEU point over the baseline EBMT system but no improvement compared to any of the TM-integrated EBMT-based systems.

In addition to the above automatic evaluations, we also performed a manual evaluation to understand the translation quality from human per-

spective. While manually evaluating the MT systems, we assign values from two five-point scales representing *fluency* and *adequacy* (LDC, 2005). The five-point scale of adequacy indicates how much meaning is conveyed in the hypothesis translation in connection to the reference translation. The five-point scale of fluency indicates the closeness of the hypothesis translation to natural speech. The two scales are explained below.

Fluency	Adequacy (meaning expressed)
5 = Flawless Bangla	5 = All
4 = Good Bangla	4 = Most
3 = Non-native Bangla	3 = Much
2 = Disfluent Bangla	2 = Little
1 = Incomprehensible	1 = None

Using the above scale 4 different native Bangla speakers were asked to score each translation produced by the different MT systems. Table 3 shows the average fluency and adequacy of our different MT systems.

System	Fluency	Adequacy
SMT	3.00	3.16
EBMT+TM(PT)	3.50	3.55
EBMT+TM(PT,LT)	3.50	3.70
EBMT+SMT	3.44	3.52

Table 3: Average Fluency and Adequacy of the MT Systems in the scale of 1-5

Like automatic evaluation, we see that the EBMT-based system uniformly improves both fluency and adequacy over SMT. The increase of 0.44 – 0.5 in fluency and 0.36 – 0.54 on SMT is noteworthy. Though there is not much difference in fluency between the three combinations but the use of LT (EBMT+TM(PT,LT)) has improvement on other two combinations (EBMT+TM(PT), EBMT+SMT).

6 Observations

We find that the baseline EBMT system has higher accuracy across all metrics compared to the baseline SMT system. This is due to the fact that large segments of the input sentences to be translated can be found with the example-base. This helps to retain the word order in the target translation which would otherwise affect the BLEU score. In contrast, the SMT-based system essentially does not use these matched segments as a whole instead the SMT decoder prefers the most probable translation. Due to the availability of a very small corpus the SMT decoder suffers with the sparse data problem.

EBMT+TM(PT) has better accuracy than the baseline SMT and EBMT systems. The use of word-level TM (LT) improves the accuracy of the translation system with respect to all other systems. The use of LT ensures that no known word is left un-translated in the target language. Altogether, the use of EBMT and TM together has shown considerable improvement over all MT evaluation metrics.

The combination of SMT with EBMT (EBMT+SMT) has some improvement over both baseline systems (SMT and EBMT). However, the EBMT+SMT system has much lower accuracy compared to the combined systems of EBMT and TM (EBMT+TM(PT) and EBMT+TM(PT,LT)). This is due to the fact that while SMT is used to translating the unmatched segments some words are left untranslated. From the example output of the system in Table 4, we can see that for the first sentence the word '*usually*' has no translation produced by the EBMT+SMT system.

Type	Sentences	
Input	1. which doctor do you usually see ?	2. i'll call you back in a few minutes .
Closest Match	1. which doctor would you like to see ?	2. i'll call you back within half and hour .
Ref	1. সাধারনত আপনি কোন ডাক্তারকে দেখান ? Usually you which doctor-accusative see?	2. আমি আপানাকে কিছুক্ষনের মধ্যে আবার কল করব । I you-accusative "some time"-oblique within again call do-future.
SMT	1. আপনি কি which ডাক্তার দেখাতে usually ? You what *which* doctor to see-causative *usually*?	2. আমি কয়েক মিনিট দেরিতে আপনি call আবার আসুন। I few minutes after-locative you *call* again come.
EBMT+ TM(PT,LT)	1. আপনি কোন ডাক্তারকে দেখাতে সধারনত ? You which doctor-accusative to see-causative usually?	2. আমি নিয়ে কয়েক মিনিট দেরিতে আবার কল করব। I with few minutes after-locative again call do-future.
EBMT+ SMT	3. আপনি কোন ডাক্তারকে দেখাতে usually ? You which doctor-accusative to see-causative *usually*?	1. আমি নিয়ে কয়েক মিনিট দেরিতে আবার কল করব। I with few minutes after-locative again call do-future.

Table 4: Translations of example sentences using three different systems. The words in *italics* in the gloss represent un-translated words.

6.1 Assessment of Error Types

The errors are propagated mostly due to the wrong source-target equivalent in the phrase table and lexical table which are being used as the translation units in our TM (PT and LT). This results in some incorrect alignment in the matching step of our EBMT system. For example, the first sentence in Table 4, the matching module gives the following alignment:

s : which doctor <do> you <usually> see ?
s_c : which doctor <would> you <like to> see ?
s_c^t : আপনি কোন ডাক্তারকে দেখাতে <1:চান> ?
Apani kona DAktArake dekhAte chAna?
YOU WHICH DOCTOR-ACCUSATIVE TO SEE-CAUSATIVE WANT?

In the above example, the word '*would*' doesn't have any alignment in s_c^t .The three target equivalents in the PT {a. হবে (*habe* [is-3Fr]) b. বলব (*balaba* [say-1Fr]) c. কি (*ki* [what])} of the word 'would' do not match with any of the word in s_c^t. Also, the system suffers when there is a mismatch either in the verb or in the subject of the sentence. This is because the inflection on the verbs depends on the morphological attributes of the subject.

The second type of error is propagated during the recombination step. In the second example in Table 4, we have successfully matched the fragments as follows:

s : i'll call you back <in a few minutes> .
s_c : i'll call you back <within half and hour> .
s_c^t : আমি <0:আধ ঘন্টা থেকে এক ঘন্টার ভিতরে> আবার কল করব।
Ami Adha ghanTA theke eka ghanTAra bhitare AbAra kala karaba.
I half an hour to one hour within again call future

However, in the recombination step, we need to generate the translation for the segment '*in a few minutes*'. We have found that the portion '*a few minutes*' has a translation equivalent 'কয়েক মিনিট দেরিতে (*kaYeka miniTa derite*)' in PT. Thus, we still need to translate the word '*in*' to generate the target equivalent for the whole segment. For the word '*in*', the PT has a separate entry with three target equivalents – a. নিয়ে (*niYe*) b. নিয়ে আসতে (*niYe Asate*) c. আসুন (*Asuna*). Picking the most probable target equivalent for '*in*' and combining with the target equivalent of '*a few minutes*', we generate 'নিয়ে কয়েক মিনিট দেরিতে (*niYe kaYeka miniTa derite*)'. This is not a fluent target equivalent because we don't need to translate the word 'in' separately as this has been already been captured in the inflection (তে – *te*[locative]) of the final word of the target equivalent of '*a few minutes*'.

7 Conclusion and Future Work

The experiments described in the paper show initial investigations for combining translation memories in an EBMT framework. The results show that all the EBMT approaches work better compared to the SMT-based system. The integration of TM with the EBMT framework has improved the translation quality and the best accuracy has been achieved while two translation memories (PT and LT) are in use along with the EBMT framework.

However, based on the observations and propagated errors discussed in the previous section, we intend to use more sophisticated *Matching* and *Adaptability* techniques considering the syntax of the examples and test sentences to overcome the false positive adaptations as in section 6.1. We also plan to use morpho-syntactic information during *recombination* to improve the translation quality.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (http://www.cngl.ie) at Dublin City University.

References

V. Ambati and U. Rohini. 2007. A hybrid approach to example based machine translation for Indian languages. In *Proceedings of 5th International Conference on Natural Language Processing* (*ICON 2007*), Hyderabad, India, pp. 4-6.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, & P. Roossin. 1988. A statistical approach to French/English translation. In *Proceedings of 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages* (*TMI 1988*), Pittsburgh, Pennsylvania, pp. 810-823.

A. van den Bosch, N. Stroppa and A. Way. 2007. A memory-based classification approach to marker-based EBMT. In *METIS-II Workshop:*

New Approaches to Machine Translation, Leuven, Belgium, pp. 63-72.

J. Bourdaillet, S. Huet, F. Gotti, G. Lapalme and P. Langlais. 2009. Enhancing the bilingual concordancer TransSearch with word-level alignment. In *Proceedings, volume 5549 of Lecture Notes in Artificial Intelligence: 22nd Canadian Conference on Artificial of Intelligence* (*Canadian AI 2009*), Springer-Verlag, pp. 27-38.

L. Bowker. 2002. Computer-aided translation technology: a practical introduction. Chapter *Translation memory systems*, University of Ottawa Press, Ottawa, pp. 92-127.

S. Dandapat, M. Forcada, D. Groves, S. Penkale, J. Tinsley and A. Way. 2010. OpenMaTrEx: A free/open-source marker-driven example-base machine translation system. In *Proceedings of Icetal 2010, the 7th International Conference on Natural Language Processing,* Reykjavik, Iceland, pp. 121-126.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram occurrence statistics. In *Proceedings of the 2nd international conference on Human Language Technology Research* (*HLT 2002*), San Diego, California, pp. 138-145.

D. Farwell and L. Padró. 2009. FreeLing: from a multilingual open-source analyzer suite to an EBMT platform. In *Proceedings of the 3rd International Workshop on Example-Based Machine Translation* (*EBMT 2009*), Dublin, Ireland, pp.37-43.

Md. Z. Islam, J. Tiedemann and A. Eisele. 2010. English to Bangla phrase-based machine translation. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation* (*EAMT 2010*), Saint-Raphaël, France, no page number.

M. Kay. 1980. The proper place of men and machine in language translation. Research Report CSL-80-11. Xerox PARC, Palo Alto, California. Reprinted in: *Machine Translation* 12(1-2), 1997, pp. 3-23; and in: Nirenburg, S. et al. (eds.) *Readings in machine translation* (Cambridge, Mass.: MIT Press, 2003), pp. 221-232.

M. Khalilov, J.A.R. Fonollosa, I. Skadina, E. Bralitis and L. Pretkalnina. 2010. English-Latvian SMT: the challenge of translating into a free word order language. In *Proceedings of the second International Workshop on Spoken Language Technologies for Under-resourced Languages* (*SLTU 2010*), Penang, Malaysia, pp. 87-94.

J.D. Kim, R.D. Brown, & J.G. Carbonell: Chunk-based EBMT. 2010. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation* (*EAMT 2010*), Saint-Raphaël, France, no page number.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of Association of Computational Linguistics* (*ACL 2007*), Prague, Czech Republic, pp. 177-180.

P. Koehn. 2009. *Statistical Machine Translation*, Cambridge University Press.

P. Langlais, G. Foster and G. Lapalme. 2000. TransType: a computer-aided translation typing system. In *Proceedings of ANLP/NAACL 2000 workshop: Embedded machine translation systems*, Seattle, Washington, pp.46-51.

P. Langlais, G. Lapalme and M. Loranger. 2002. TransType : Development-evaluation cycles to boost translator's productivity. *Machine Translation,* 15(4):77-98.

V. I. Lavenshtein. 1965. Binary Codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845-848. English translation in Soviet Physics Doklady, 10(8), 707-710.

Y. Lepage, A. Lardilleux, J. Gosme and J.L. Manguin. 2008. The GREYC machine translation system for the IWSLT 2008 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation* (*IWSLT 2008*), Hawaii, USA; pp. 39-45.

D. Marcu. 2001. Towards a unified approach to memory- and statistical-based machine translation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (*ACL 2001*), Toulouse, France, pp.386-393.

S.K. Naskar and S. Bandyopadhyay. 2005. A phrasal EBMT system for translating English to Bengali. In *Proceedings of MT Summit X*, Phuket, Thailand, pp.372-379.

M. Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In *Artificial and human intelligence*, Amsterdam: North Holland, pp. 173-180.

K. Papineni, S. Roukos, T. Ward and W. J. Zhu. 2002. "BLEU: a method for automatic evaluation of machine translation". In *Proceedings of 40th Annual meeting of the Association of Computation Linguistics* (*ACL 2007*), Philadelphia, Pennsylvania, pp. 311-318.

A.B. Phillips and R.D. Brown. 2009. Cunei machine translation platform: system description. In *Proceedings of the 3rd International Workshop*

on Example-Based Machine Translation (*EBMT 2009*), Dublin, Ireland, pp. 29-36.

A. Ramanathan, H. Choudhary, A. Ghosh and P. Bhattacharyya. 2009. Case markers and morphology: addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing* (*ACL-IJCNLP-2009*), Suntec, Singapore, pp. 800-808.

A. Ramanathan, P. Bhattacharyya, J. Hegde, R. M. Shah and M. Sasikumar. 2008. Simple syntactic and morphological processing can help English-Hindi statistical machine translation. In *Proceedings of 3rd International Joint Conference on Natural Language Processing* (*IJCNLP 2008*), January 7-12, 2008, Hyderabad, India, pp. 513-520.

J. Smith and S. Clark. 2009. EBMT for SMT: a new EBMT-SMT hybrid. In *Proceedings of the 3rd International Workshop on Example-Based Machine Translation* (*EBMT 2009*), Dublin, Ireland, pp. 3-10.

H. Somers, S. Dandapat and S.K. Naskar. 2009. A review of EBMT using proportional analogies. In *Proceedings of the 3rd International Workshop on Example-Based Machine Translation* (*EBMT 2009*), Dublin, Ireland, pp. 53-60.

H. Somers. 2003. Translation Memory Systems. In Harold Somers, editor, *Computers and translation: a translator's guide,* John Benjamins, pp. 31-47.

H. Somers. 2003. An overview of EBMT. In Michael Carl and Andy Way, editors, *Recent advances in example-based machine translation,* Kluwer Academic Publishers, pp. 3-57.

N. Stroppa and A. Way. 2006. MaTrEx: the DCU machine translation system for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation* (*IWSLT 2006*), Kyoto, Japan, pp. 31-36.

R.A. Wagner and M.J. Fischer. 1974. The string to string correction problem. *Journal of the ACM* (*JACM*), 21(1):168-173.

T. Watanabe, J. Suzuki, H. Tsukada and H. Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2007),* Prague, Czech Republic. pp. 764-773.