A Hybrid Approach to Example based Machine Translation for Indian Languages

Vamshi Ambati
Digital Library of India Project
IIIT, Hyderabad, India
vamshi@iiit.ac.in

Rohini U
Language Technologies Research Centre
IIIT, Hyderabad, India
rohini@research.iiit.ac.in

Abstract

Corpus based approaches to machine translation namely Example based machine translation and Statistical machine translation have received wide focus in the recent years. Hybrid approaches combining the two further improved the performance. Indian language machine translation has mostly focussed on rule based machine translation. We propose a hybrid approach to Example based machine translation making use of statistical machine translation methods and minimal linguistic resources. Our motive in this paper is to obtain a 'good enough' translation as opposed to a perfect translation aimed by earlier machine translation efforts. Our approach can be used for translation of english to any indian language. In this paper, we perform experiments for translation of english to hindi and report BLEU scores.

1 Introduction

Recently corpus based approaches to machine translation have received wide focus. They are namely Example Based Machine Translation (EBMT) (Nagao, 1984) and Statistical Machine Translation (SMT) (Brown et al., 1993). A combination of statistical and example-based MT approaches shows some promising perspectives for overcoming the shortcomings of each approach. Efforts have been made in this direction, using the alignments from both the methods to improve the translation (Groves and Way, 2006), to improve the alignment in the EBMT using the statistical information computed from SMT methods (Kim et al., 2005) etc. The results obtained have shown improvement in performance.

However, these approaches cannot directly be applied to Indian languages due to the small size of the parallel texts available and sparse linguistic resources. Also some of the assumptions made in some of these approaches like marker hypothesis (Gough, 2005), cannot directly be applied to translate from english to Indian languages since word order in the source and target languages is very different and sequential word orderings between source and target sentences do not exist. Consider the following example english and hindi sentences (in wx)

Example 1:

- English: All members of the Danish East India Company died of fever.
- Hindi: jvara se isake saBI saxasya mara gae.

As it can be seen, there is word re-ordering in this hindi sentence corresponding to the english sentence. Clearly, splitting the sentence at marker words (Gough, 2005) does not apply here as it assumes sequential source and target sentences.

Our approach to EBMT is in its purest sense (similar to (Brown, 2000)) which makes use of source and target parallel sentences. The advantage of this approach is that it can be language independent and makes minimal use of linguistic resources as they are sparse in our case. We aim to build rapid English to Indian language EBMT systems making the most of the available minimal resources - parallel texts and bi lingual dictionaries.

Earlier approaches to EBMT considered the longest match of the input sentence with the source sentence in the example database. Alignment is performed by computing a correspondence matrix using a manual dictionary and improved with statistical dictionary. We take a similar approach of considering the longest match and extract alignment from the matching example pair using a manual and a statistical dictionary. We construct the statistical dictionary using a statistical alignment tool GIZA++ (Och and Ney, 2003). Then, to extract an alignment for a matching sub-sentence, we first consider the best viterbi alignment for the sentence-target example pair as given by GIZA++ and enhance the alignments using the manual dictionary and statistical dictionary. Using this enhanced alignment between the source and target example pair, we extract the translation of the sub-sentential fragment identified from the longest match. We then perform a simple combination of the translation fragments by joining all the translation fragments obtained to obtain the final translation. We performed experiments using our system for translating english to hindi and the BLEU scores are reported.

The rest of the paper is organized as follows. In Section 2, we discuss the Related Work, In Section 3, we discuss the Problems in adapting an EBMT system to Indian languages, In Section 4, we discuss our System, In Section 4, we present out Experimental Results. In Section 5, we discuss our Conclusions and Future work.

2 Related Work

Machine translation of Indian Languages has been pursued mostly on the linguistic side. Hand crafted rules were mainly used for translation, (Sinha and A.Jain, 2003), (Bharati et al., 1997). Rule based approaches were combined with EBMT system to build hybrid systems (Jain et al., 2001). (Dave et al., 2002) performs interlingua based machine translation. Input in the source language is converted into UNL, the Universal Networking Language and then converted back from UNL to the target language. Recently, Gangadhariah et al (Gangadharaiah and Balakrishnan, 2006) used linguistic rules are used for ordering the output from a generalized example based machine translation (Brown, 2000).

While, in general in the machine translation literature, hybrid approaches have been proposed for EBMT primarily using statistical information most of which have shown improvement in performance over the pure EBMT system. (Vogel and Ney, 2000) automatically derived a hierarchical TM from a parallel corpus, comprising a set of transducers encoding a simple grammar. (Paul et al., 2003) used example-based re-scoring method to validate SMT translation candidates. (Imamura et al., 2004) proposed an example based decoding for statistical machine translation which outperformed the beam search based decoder (Koehn, 2004). Kim et al (Kim et al., 2005) showed improvement in alignment in EBMT using statistical dictionaries and calculating alignment scores bi-directionally. (Groves and Way, 2006) (Groves and Way, 2005) combined the sub-sentential alignments obtained from the EBMT systems with word and phrase alignments from SMT to make 'Example based Statistical Machine Translation' and 'Statistical Example based Machine Translation'.

3 Problems Adapting an EBMT System to Indian Languages

In this section, we discuss the problems faced in adapting an EBMT system to Indian language machine translation.

From a corpus of source-target sentence pairs, EBMT models of translation perform three distinct phases in order to transform a new input string into a target language translation:

- Matching Phase: Searching the source side of the parallel corpus for 'close' matches and their translations.
- Alignment Phase: Determining the sub-sentential translation links in those retrieved examples.
- Recombination Phase: Recombining relevant parts of the target translation links to derive the translation.

Firstly, it is difficult as to how to segment a sentence in order to get a good match and hence a good translation. Earlier theories assumed sequential source and target sentences (Gough, 2005) which is not true in our case as shown in Example 1. Secondly, as pointed out in the earlier, the EBMT systems require a large amount of parallel texts. Previous work on EBMT uses a parallel corpus of the order of a few hundred thousand sentences. Preparation of such large parallel texts requires a large amount of manual effort. For minority languages such as Indian languages, there is no availability of such large parallel texts. To our knowledge, the largest available parallel text consists of 54K sentences of english-hindi parallel text[1]. In the matching phase discussed above, due to small size of the parallel texts available for Indian languages, matching of the input sentence with the source language text might not result in a good match and hence no good alignment and translation.

The alignment phase is even more difficult while applying previous approaches for several reasons. Firstly, sequential sentences in the source and target sentences assumption as done in marker based EBMT approach (Gough, 2005) etc does not much hold for Indian Languages. There is a lot of word re-ordering which takes place in target language which makes it difficult for the alignment. Secondly, low coverage of dictionaries available makes it difficult. Also the word usage in the dictionary and the target sentence might be different though it means the same thing. Thirdly, due to large number of inflections in the Indian languages, there will often be matching problems. Also, two words in Indian languages combine to form a new word or inflection of an existing word (*sandhi, samasam*) which makes it even worse while matching words for performing alignment. Morphological analysis may come to rescue here but due to lack of such tools for all Indian languages, we do not use this in our current research.

In the recombination phase, the different translations extracted from different examples have to be combined together removing any boundary friction problems that crop in. This combination is difficult in Indian languages due to its rich morphology.

4 Our System

Our approach to Example based machine translation addresses some of the issues addressed in the above section. The framework is largely motivated and is similar to existing paradigms of EBMT like (Brown, 2000). The design of the proposed system is shown in the Figure 1. In this section we discuss in detail some of the important phases in the system.

4.1 Matching

Indian language parallel corpus is a scarce resource, even when considering English as the source language. The power of an EBMT system lies in the examples that it uses for the translation. Given the small size of the parallel corpus as in our case. We look for the longest possible match of the input with the source language sentence in the example base in order to preserve context.

We address the problem of how to segment a given input by first looking at the example database as to what is the longest possible fragment available. We pick the corresponding sentence. Next, we move on to the remaining fragment in the input sentence for which the match has to be found from the database. We then look in the database for the longest possible fragment available for this remaining fragment in the input sentence. We pick the corresponding sentence. Similarly we proceed until the input string terminates.

Due to lack of extensive corpus, we may not find many successful matches in the example database. Therefore, we apply the following two approaches to improve the matching phase in our system.

4.1.1 Morphological Analysis of the corpus

Morphological analysis is done on the source language of the corpus (english) and each word in the sentence is replaced with the root word corresponding to the word in the sentence. This can be considered as the first step in generalization i.e generalizing all variations of a word to the root word. When an input sentence is requested for translation, the sentence is run through a morphological analyzer to get the root words of each of the words in the sentence. This sentence is now matched against the sentence in the database which is a string containing all the root words in the sentence. We do not assume any kind of morphological analysis on the target part of the corpus.

4.1.2 Generalizing the corpus

It has been shown earlier (Brown, 2000) that generalization of the examples in the example base can reduce the corpus size by almost a sixth portion without much loss in the translation accuracy. In our approach, we generalize only nouns in the parallel corpus. We have a database of noun classes manually built using words from a dictionary. This is used to generalize all the sentences in the English side of the parallel corpus. Verbs are more complicated and we believe they should not be generalized as this might actually compromise the quality of the output.

[1] 54K sentence pairs originally collected as part of TIDES MT project and later refined at IIIT-Hyderabad, India

4.2 Alignment

Sub sentential alignment is critical in locating the correct translation for a matched fragment of the input in an EBMT System. The sub-sentential alignments are computed using a correspondence matrix. A correspondence matrix is created by first looking up each word in the source-language half of a translation example in a manually created bilingual dictionary or a statistical dictionary and marking each occurrence of any of the translations in the target-language half as a possible correspondence and then pruning it to remove any ambiguities.

We take a similar approach and construct a statistical dictionary using the word alignments given by GIZA++. We first take the best viterbi alignment given by GIZA++ for the source-target pair identified in the matching phase. Then, we further enhance the word alignments using a manual dictionary and a statistical dictionary. The statistical dictionary is constructed from the same parallel corpus used as the example database. We add a correspondence between the words in the source and target sentences whenever we find a word in the manual dictionary or in the statistical dictionary. Not all the entries in the statistical dictionaries are used. Only the entries above a certain threshold for translation probabilities are used. We set the threshold to 0.3 in our experiments. From this alignments, we construct a correspondence matrix which is used to compute the translation of the matched source fragment.

For example, consider the sentence pair (hindi sentences in wx)

- **English**: sardar gurdeet singh with his 30 companions slipped into the lanes of calcutta and was never heard of again.

- **Hindi**: apane 30 sAWiyoM ko sAWa lekara saraxAra guraxIwa siMha kalakawA kI galiyoM meM vilIna ho gae

The alignment given by GIZA++ is

- NULL (10 12 14 18) apane (5) 30 (6) sAWiyoM ( ) ko ( ) sAWa (4) lekara ( ) saraxAra (1) guraxIwa (2 7 8 9 11 15 16 17 19) siMha (3) kalakawA (13) kI ( ) galiyoM ( ) meM ( ) vilIna ( ) ho ( ) gae ( ) . (20)

For each word, its possible alignments are listed in brackets following it. The numbers in brackets correspond to the number of the token in the english sentence starting from 1. . For example, 1 refers to 'sardar', 2 refers to 'gurdeet' and so on. NULL refers to null translations. i.e the translations in the target sentence for the words (10, ie 'the' , 12 ie 'of' etc) in the source sentence are not found. Hence they are mapped to null. We enhance this alignment by adding the manual dictionary and statistical dictionary entries. The enhanced alignment is as follows (enhanced correspondences underlined)

- NULL (10 12 14 18) apane (5) 30 (6) sAWiyoM (7) ko ( ) sAWa (4) lekara ( ) saraxAra (1) guraxIwa (2 7 8 9 11 15 16 17 19) siMha (3) kalakawA (13) kI (12) galiyoM (11) meM ( ) vilIna ( ) ho ( ) gae ( ) . (20)

From the above enhanced alignment, we now extract the translation for the sub-sentential match identified in the matching phase.

4.3 Recombination

Given the limitations of the resources as in our case, the motive in this paper to obtain a 'good enough' translation if not a perfect translation. By 'good enough' translation we mean that the translation approximately conveys the meaning of the input sentence. Hence, the last step ie of combining all the alignments obtained is not particularly the focus of this paper.

In (Frederking and Nirenburg, 1994), it has been shown that the quality of MT systems is improved by using the best results obtained from a variety of systems working on the same text simultaneously. Similarly research results from other approaches aimed at improving the quality of translation using language models and linguistic rules (Gangadharaiah and Balakrishnan, 2006) can be applied here to improve the translation quality but it is not the focus of this paper. In this paper, we perform a simple combination of the translation fragments by joining all the translation fragments obtained.

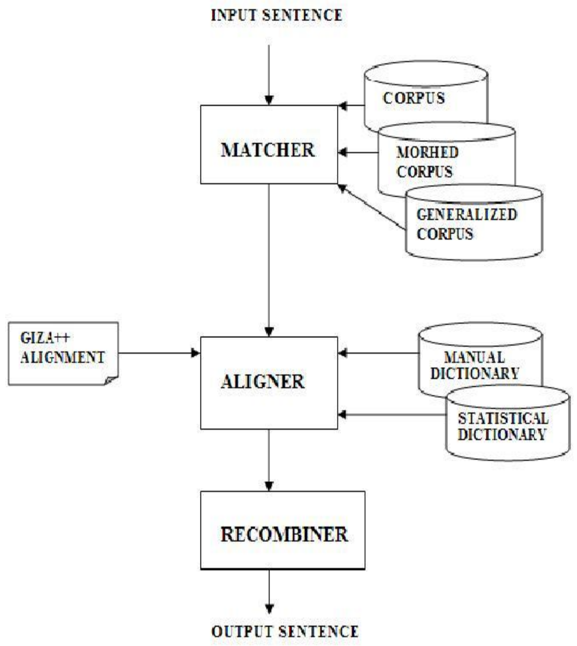

Figure 1: Proposed System

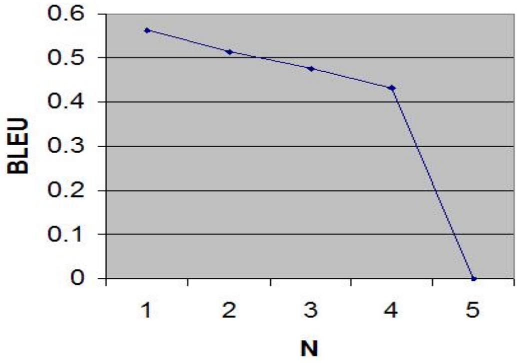

Figure 2: Variation of BLEU with N

Method	BLEU
1. word-word match (manual dict)	0.124
2. word-word match (manual+stat dict)	0.214
3. Our approach	0.432

Table 1:

5 Evaluation

In this section, we describe the experiments we performed on our system. All of the data used for the primary experiments described below consists of 54K english-hindi parallel sentences pairs, originally collected as part of TIDES MT project and later refined at IIIT-Hyderabad, India. In all the experiments reported below, the source language is english and the target language is hindi. We constructed a training set consisting of 53K sentences and the test set consisted of randomly selected 100 sentences.

For empirical evaluation, we use the metric proposed by IBM, called BLEU (Papineni et al., 2002). It tries to assess how close a machine translation is to a set of reference translations generated by humans. Our experiments use the single reference translation available in the parallel sentence pair. Table 1 shows the results of our approach and compares to two other methods. The first method is ”word to word match” using a manual dictionary. During alignment a word to word substitution is done using a manual dictionary. As it can be guessed, the BLEU score is low. Firstly because of the word to word match, the n-grams comparison that BLEU uses penalizes and gives a low score. Secondly due to missing entries in the manual dictionary. The second method in the table is the word to word match same as in the first method but a statistical dictionary build using GIZA++ is also used in addition to a manual dictionary. As it can be seen, there is a slight improvement in the score due to the high coverage of the statistical dictionary. The third method is our approach described in Section 4.2. We achieved a BLEU score of 0.432. The first and second methods form the baseline systems. For the results reported in Table 1, the value N, which represents the order of n-gram picked for calculating the BLEU scores has been set to the default value(4). We also performed an experiment varying the length from 1 to 5 and the graph is as shown in Figure 2. As it can be seen, the BLEU score dropped gradually from N equal to 1 to 2, and 3 and abruptly from 4 to 5. This is because, they were very less 5-grams found on the translation. This can be improved by improving the recombination phase using existing approaches.

As pointed out in (Gangadharaiah and Balakrishnan, 2006), BLEU scores are very harsh on Indian languages due to high infections in Indian language. This research only compares and reports the improvement in the BLEU scores over the baseline system considered. By performing an effective recombination step, our BLEU scores can go high and comparisons with other translation research can be made.

6 Conclusions

In this paper, we have presented a hybrid approach to EBMT for performing translation from english to Indian languages using statistical approaches. We performed matching by considering the longest match of the input sentence available in the example database. We performed alignment using a manual and a statistical dictionary build from GIZA++ and the best viterbi alignment given by GIZA++ for each sentence pair in the example database. For recombination step, we performed a simple combination of the translation fragments by joining all the translation fragments obtained to obtain the final translation. We performed experiments using our system for english to hindi translation and the results are reported. In future, we shall experiment our system with other Indian languages given the availability of parallel corpora.

References

Akshar Bharati, Vineet Chaitanya, Amba P Kulkarni, and Rajeev Sangal. 1997. Anusaaraka: Machine translation in stages. *A Quarterly in Artificial Intelligence*, 10(3):22–25, July.

Peter F. Brown, Stephen A. Della Pietra, Vi cent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computlational Linguistics*, 19(2).

Ralf D. Brown. 2000. Automated generalization of translation examples. In *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)*, pages 125–131, Saarbrcken, Germany.

Shachi Dave, Jignashu Parikh, Bhattacharyya, and Pushpak Interlingua. 2002. Based english hindi machine translation and language divergence. *Journal of Machine Translation*, 17, September.

Robert Frederking and Sergei Nirenburg. 1994. Three heads are better than one. In *Proceedings of the Fourth Conference on Applied Natural Language Processing, ANLP-94*, Stuttgart, Germany.

Rashmi Gangadharaiah and N. Balakrishnan. 2006. Application of linguistic rules to generalized example based machine translation for indian languages. In *First National Symposium on Modeling and Shallow Parsing of Indian Languages, (MSPIL)*, India, Mumbai, April.

N Gough. 2005. *Example-Based Machine Translation Using the Marker Hypothesis*. Ph.D. thesis, Dublin City University, Dublin, Ireland.

D Groves and A Way. 2005. Hybrid example based smt: the best of both worlds? In *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages pp. 183–190, Ann Arbor, MI.

D. Groves and A. Way. 2006. Hybridity in mt: Experiments on the europarl corpus. In *Proceedings of the 11th Conference of the European Association for Machine Translation*, Oslo, Norway.

Kenji Imamura, Hideo OKUMA, Taro Watanabe, and Eiichiro Sumita. 2004. Example-based machine translation based on syntactic transfer with statistical models. In *COLING 2004*, volume I, pages 99–105.

Renu Jain, R.M.K. Sinha, and Ajai Jain. 2001. Anubharti: Using hybrid example based approach for machine translation. In *Symposium on Translation Support Systems, SYSTRANS*, Kanpur, India, February.

Jae Dong Kim, Ralf D. Brown, Peter J. Jansen, and Jaime G. Carbonell. 2005. Symmetric probabilistic alignment for example-based translation. In *Proceedings of the Tenth Workshop of the European Assocation for Machine Translation (EAMT-05)*, pages 153–159, May.

P Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In R. Frederking and K. Taylor, editors, *Machine Translation: From Real Users to Research; AMTA 2004, LNAI 3265*, pages 115–124, Berlin/Heidelberg, Germany. Springer Verlag.

Makoto Nagao. 1984. A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and Human Intelligence*, pages 173–180.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics ACL '02*.

Michael Paul, Eiichiro Sumita, and Seiichi Yamamoto. 2003. Example-based rescoring of statistical machine translation output. In *Machine Translation Summit IX*, pages 410–417, New Orleans, Louisiana.

R.M.K. Sinha and A.Jain. 2003. Anglahindi:an english to hindi machine-aided translation system. In *MT Summit IX*, New Orleans, Louisiana, USA, September.

S. Vogel and H. Ney. 2000. Construction of a hierarchical translation memory. In *Proceedings ofthe 18th International Conference on Computational Linguistics: COLING 2000*, pages 1131–1135, Saarbrucken, Germany.