Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation

Ananthakrishnan Ramanathan,
Pushpak Bhattacharyya
Department of Computer Science
and Engineering
Indian Institute of Technology
Powai, Mumbai-400076
India
{anand,pb}@cse.iitb.ac.in

Jayprasad Hegde, Ritesh M. Shah,
Sasikumar M
CDAC Mumbai (formerly NCST)
Gulmohar Cross Road No. 9
Juhu, Mumbai-400049
India
{jjhegde,ritesh,sasi}
@cdacmumbai.in

Abstract

In this paper, we report our work on incorporating syntactic and morphological information for English to Hindi statistical machine translation. Two simple and computationally inexpensive ideas have proven to be surprisingly effective: (i) reordering the English source sentence as per Hindi syntax, and (ii) using the suffixes of Hindi words. The former is done by applying simple transformation rules on the English parse tree. The latter, by using a simple suffix separation program. With only a small amount of bilingual training data and limited tools for Hindi, we achieve reasonable performance and substantial improvements over the baseline phrase-based system. Our approach eschews the use of parsing or other sophisticated linguistic tools for the target language (Hindi) making it a useful framework for statistical machine translation from English to Indian languages in general, since such tools are not widely available for Indian languages currently.

1 Introduction

Techniques for leveraging syntactic and morphological information for statistical machine translation (SMT) are receiving a fair amount of attention nowadays. For SMT from English to Indian languages, these techniques are especially important for the following three reasons: (i) Indian languages differ widely from English in terms of word-order; (ii) Indian languages are morphologically quite rich; and (iii) large amounts of parallel corpora are not available for these languages, though smaller amounts of text in specific domains (such as health, tourism, and agriculture) are now becoming accessible. It might therefore be expected that using syntactic and morphological information for English to Indian language SMT will prove highly beneficial in terms of achieving reasonable performance out of limited parallel corpora. However, the difficulty in this is that crucial tools, such as parsers and morphological analyzers, are not widely available for Indian languages yet.

In this paper, we present our work on incorporating syntactic and morphological information for English to Hindi SMT. Our approach, which eschews the use of parsing and other tools for Hindi, is two-pronged:

1. Incorporating syntactic information by combining phrase-based models with a set of structural preprocessing rules on English

2. Incorporating morphological information by using a simple suffix separation program for Hindi, the likes of which can be created with limited effort for other Indian languages as well

Significant improvements over the baseline phrase-based SMT system are obtained using our approach. Table 1 illustrates this with an example [1].

Since only limited linguistic effort and tools are required for the target language, we believe that the framework we propose is suitable for SMT from English to other Indian languages as well.

[1] This example is discussed further in section 4

input	For a celestial trip of the scientific kind, visit the planetarium.
reference	वैग्यानिक तरीके के एक दिव्य सैर के लिए, तारामंडल आएं। vaigyaanika tariike ke eka divya saira ke lie, taaraamandala aaem *scientific kind of a celestial trip for, planetarium visit (come)*
baseline	के स्वर्गीय यात्रा के वैग़्यानिक प्रकार, का तारागृह है। ke svargiiya yaatraa ke vaigyaanika prakaara, kaa taaraagruha hai *of celestial trip of scientific kind, of planetarium is*
baseline+syn	वैग़्यानिक प्रकार के स्वर्गीय यात्रा के लिए, तारागृह है। vaigyaanika prakaara ke svargiiya yaatraa ke lie, taaraagruha hai *scientific kind of celestial trip for, planetarium is*
baseline+syn+morph	वैग़्यानिक प्रकार के स्वर्गीय यात्रा के लिए, तारागृह देखें। vaigyaanika prakaara ke svargiiya yaatraa ke lie, taaraagruha dekhem *scientific kind of celestial trip for, planetarium visit (see)*

Table 1: **Effects of Syntactic and Morphological Processing** (*reference*: human reference translation; *baseline*: phrase-based system; *syn*: with syntactic information; *morph*: with morphological information)

The rest of this paper is organized as follows: Section 2 outlines related work. Section 3 describes our approach – first, the phrase-based baseline system is sketched briefly, leading up to the techniques used for incorporating syntactic and morphological information within this system. Experimental results are discussed in section 4. Section 5 concludes the paper with some directions for future work.

2 Related Work

Statistical translation models have evolved from the word-based models originally proposed by Brown et al. (1990) to syntax-based and phrase-based techniques.

The beginnings of phrase-based translation can be seen in the alignment template model introduced by Och et al. (1999). A joint probability model for phrase translation was proposed by Marcu and Wong (2002). Koehn et al. (2003) propose certain heuristics to extract phrases that are consistent with bidirectional word-alignments generated by the IBM models (Brown et al., 1990). Phrases extracted using these heuristics are also shown to perform better than syntactically motivated phrases, the joint model, and IBM model 4 (Koehn et al., 2003).

Syntax-based models use parse-tree representations of the sentences in the training data to learn, among other things, tree transformation probabilities. These methods require a parser for the target language and, in some cases, the source language too. Yamada and Knight (2001) propose a model that transforms target language parse trees to source language strings by applying reordering, insertion, and translation operations at each node of the tree. Graehl and Knight (2004) and Melamed (2004), propose methods based on tree-to-tree mappings. Imamura et al. (2005) present a similar method that achieves significant improvements over a phrase-based baseline model for Japanese-English translation.

Recently, various preprocessing approaches have been proposed for handling syntax within SMT. These algorithms attempt to reconcile the word-order differences between the source and target language sentences by reordering the source language data prior to the SMT training and decoding cycles. Nießen and Ney (2004) propose some restructuring steps for German-English SMT. Popovic and Ney (2006) report the use of simple local transformation rules for Spanish-English and Serbian-English translation. Collins et al. (2006) propose German clause restructuring to improve German-English SMT.

The use of morphological information for SMT has been reported in (Nießen and Ney, 2004) and (Popovic and Ney, 2006). The detailed experiments by Nießen and Ney (2004) show that the use of morpho-syntactic information drastically reduces the need for bilingual training data.

Recent work by Koehn and Hoang (2007) proposes factored translation models that combine feature functions to handle syntactic, morphological, and other linguistic information in a log-linear model.

Our work uses a preprocessing approach for incorporating syntactic information within a phrase-based SMT system. For incorporating morphology, we use a simple suffix removal program for Hindi and a morphological analyzer for English. These aspects are described in detail in the next section.

3 Syntactic & Morphological Information for English-Hindi SMT

3.1 Phrase-Based SMT: the Baseline

Given a source sentence f, SMT chooses as its translation $\hat{e}$, which is the sentence with the highest probability:

$$\hat{e} = \underset{e}{arg\,max}\; p(e|f)$$

According to Bayes' decision rule, this is written as:

$$\hat{e} = \underset{e}{arg\,max}\; p(e)p(f|e)$$

The phrase-based model that we use as our baseline system (defined by Koehn et al. (2003)) computes the translation model $p(f|e)$ by using a phrase translation probability distribution. The decoding process works by segmenting the input sentence f into a sequence of I phrases $\overline{f}_1^I$. A uniform probability distribution over all possible segmentations is assumed. Each phrase $\overline{f}_i$ is translated into a target language phrase $\overline{e}_i$ with probability $\phi(\overline{f}_i|\overline{e}_i)$. Reordering is penalized according to a simple exponential distortion model.

The phrase translation table is learnt in the following manner: The parallel corpus is word-aligned bidirectionally, and using various heuristics (see (Koehn et al., 2003) for details) phrase correspondences are established. Given the set of collected phrase pairs, the phrase translation probability is calculated by relative frequency:

$$\phi(\overline{f}|\overline{e}) = \frac{count(\overline{f}, \overline{e})}{\sum_f count(f, \overline{e})}$$

Lexical weighting, which measures how well words within phrase pairs translate to each other, validates the phrase translation, and addresses the problem of data sparsity.

The language model $p(e)$ used in our baseline system is a trigram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998).

The weights for the various components of the model (phrase translation model, language model, distortion model etc.) are set by minimum error rate training (Och, 2003).

3.2 Syntactic Information

As mentioned in section 2, phrase-based models have emerged as the most successful method for SMT. These models, however, do not handle syntax in a natural way. Reordering of phrases during translation is typically managed by distortion models, which have proved not entirely satisfactory (Collins et al., 2006), especially for language pairs that differ a lot in terms of word-order. We use a preprocessing approach to get over this problem, by reordering the English sentences in the training and test corpora before the SMT system kicks in. This reduces, and often eliminates, the 'distortion load' on the phrase-based system.

The reordering rules that we use for preprocessing can be broadly described by the following transformation rule going from English to Hindi word order (Rao et al, 2000):

$$SS_mVV_mOO_mCm \rightarrow C'_mS'_mS'O'_mO'V'_mV'$$

where,
S: Subject
O: Object
V: Verb
C_m: Clause modifier
X': Corresponding constituent in Hindi, where X is S, O, or V
X_m: modifier of X

Essentially, the SVO order of English is changed to SOV order, and post-modifiers are converted to pre-modifiers. Our preprocessing module effects this by parsing the input English sentence [2] and ap-

[2] Dan Bikel's parser was used for parsing (http://www.cis.upenn.edu/d̃bikel/license.html).

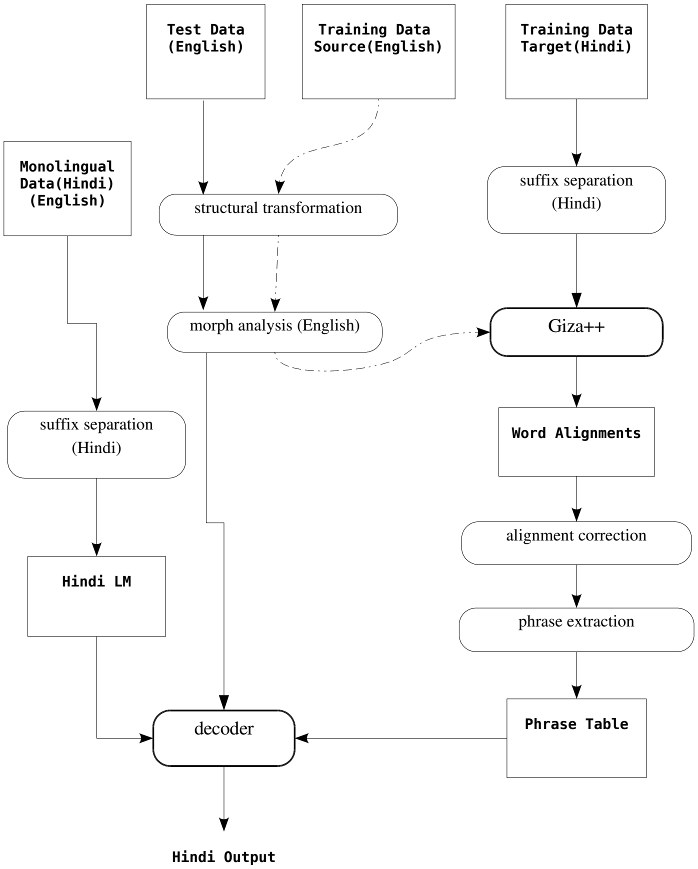
Test Data (English)
Training Data Source(English)
Training Data Target(Hindi)
Monolingual Data(Hindi) (English)
structural transformation
suffix separation (Hindi)
morph analysis (English)
Giza++
suffix separation (Hindi)
Word Alignments
alignment correction
Hindi LM
phrase extraction
decoder
Phrase Table
Hindi Output

Figure 1: **Syntactic and Morphological Processing: Schematic**

plying a handful of reordering rules on the parse tree. Table 2 illustrates this with an example.

3.3 Morphological Information

If an SMT system considers different morphological forms of a word as independent entities, a crucial source of information is neglected. It is conceivable that with the use of morphological information, especially for morphologically rich languages, the requirement for training data might be much reduced. This is indicated, for example, in recent work on German-English statistical MT with limited bilingual training data (Nießen and Ney, 2004), and also in other applications such as statistical part-of-speech tagging of Hindi (Gupta et al., 2006).

The separation of morphological suffixes conflates various forms of a word, which results in higher counts for both words and suffixes, thereby countering the problem of data sparsity. As an example, assume that the following sentence pair is part of the bilingual training corpus:

English: Players should just play.
Hindi: खिलाडियों को केवल खेलना चाहिए।
khilaadiyom ko kevala khelanaa caahie
Hindi (suffix separated): खिलाड इयों को केवल खेल ना चाहिए।
khilaada iyom ko kevala khela naa caahie

Now, consider the input sentence, “The men came across some players,” which should be translated as “आदमियों को कुछ खिलाडी मिले” (*aadmiyom ko kucha khilaadii mile*). Without using morphology, the system is constrained to the choice of खिलाडियों (*khilaadiyom*) for the word *players* (based just on the

English	$\overbrace{The\ president}^{S}\ \overbrace{of\ America}^{S_m}\ \overbrace{visited}^{V}\ \overbrace{India}^{O}\ \overbrace{in\ June}^{V_m}$
Reordered	$\underbrace{America\ of}_{S_m}\ \underbrace{the\ president}_{S}\ \underbrace{June\ in}_{V_m}\ \underbrace{India}_{O}\ \underbrace{visited}_{V}$
Hindi	अमरीका के राष्ट्रपति ने जून में भारत की यात्रा की amariikaa ke raashtrapati ne juuna mem bhaarata kii yaatraa kii

Table 2: **English and Hindi Word-Order**

आ	आएं	अता	आने	एगा
इ	आओं	अती	ऊंगा	एगी
ई	इयां	ईं	ऊंगी	आएगा
उ	इयों	अतिं	आऊंगा	आएगी
ऊ	आइयां	अते	आऊंगी	आया
ए	आइयों	आता	एंगे	आए
ओ	आँ	आती	एंगी	आई
एं	इयाँ	आतीं	आएंगे	आईं
ओं	आइयाँ	आते	आएंगी	इए
आं	अताएं	अना	ओगे	आओ
उआं	अताओं	अनी	ओगी	आइए
उएं	अनाएं	अने	आओगे	अकर
उओं	अनाओं	आना	आओगी	आकर

Table 3: **Hindi Suffix List**

evidence from the above sentence pair in the training corpus). Also, the general relationship between the oblique case (indicated by the suffix इयों (*iyom*)) and the case marker को (*ko*) is not learnt, but only the specific relationship between खिलाडियों (*khilaadiyom*) and को (*ko*). This indicates the necessity of using morphological information for languages such as Hindi.

To incorporate morphological information, we use a morphological analyzer (Minnen et al., 2001) for English, and a simple suffix separation program for Hindi. The suffix separation program is based on the Hindi stemmer presented in (Ananthakrishnan and Rao, 2003), and works by separating from each word the longest possible suffix from table 3. A detailed analysis of noun, adjective, and verb inflections that were used to create this list can be found in (McGregor, 1977) and (Rao, 1996). A few examples of each type are given below:

Noun Inflections: Nouns in Hindi are inflected based on the case (direct or oblique), the number (singular or plural), and the gender (masculine or feminine[3]). For example, लडका (*ladakaa* - boy) becomes लडके (ladake) when in oblique case, and the plural लडके (*ladake* - boys) becomes लडकों (*ladakom*). The feminine noun लडकी (*ladakii* - girl) is inflected as लडकियाँ (*ladakiyaam* - plural direct) and लडकियों (*ladakiyom* - plural oblique), but it remains uninflected in the singular direct case.

Adjective Inflections: Adjectives which end in आ (*aa*) or आं (*aam*) in their direct singular masculine form agree with the noun in gender, number, and case. For example, the singular direct अच्छा (*accha*) is inflected as अच्छे (*acche*) in all other masculine forms, and as अच्छी (*acchii*) in all feminine forms. Other adjectives are not inflected.

Verb Inflections: Hindi verbs are inflected based on gender, number, person, tense, aspect, modality, formality, and voice. (Rao, 1996) provides a complete list of verb inflection rules.

The overall process used for incorporating syntactic and morphological information, as described in this section, is shown in figure 1.

[3] Hindi does not possess a neuter gender

Technique	**Evaluation Metric**				
	BLEU	mWER	SSER	roughly understandable+	understandable+
baseline	12.10	77.49	91.20	10%	0%
baseline+syn	16.90	69.18	74.40	42%	12%
baseline+syn+morph	15.88	70.69	66.40	46%	28%

Table 4: **Evaluation Results** (*baseline*: phrase-based system; *syn*: with syntactic information; *morph*: with morphological information)

4 Experimental Results

The corpus described in the table below was used for the experiments.

	#sentences	**#words**
Training	5000	120,153
Development	483	11,675
Test	400	8557
Monolingual (Hindi)	49,937	1,123,966

The baseline system was implemented by training the phrase-based system described in section on the 5000 sentence training corpus.

For the Hindi language model, we compared various n-gram models, and found trigram models with modified Kneser-Ney smoothing to be the best performing (Chen and Goodman, 1998). One language model was learnt from the Hindi part of the 5000 sentence training corpus. The larger monolingual Hindi corpus was used to learn another language model. The SRILM toolkit [4] was used for the language modeling experiments.

The development corpus was used to set weights for the language models, the distortion model, the phrase translation model etc. using minimum error rate training. Decoding was performed using Pharaoh [5].

fnTBL (Ngai and Florian, 2001) was used to POS tag the English corpus, and Bikel's parser was used for parsing. The reordering program was written using the perl module Parse::RecDescent.

We evaluated the various techniques on the following criteria. For the objective criteria (BLEU and mWER), two reference translations per sentence were used.

- **BLEU** (Papineni et al., 2001): This measures the precision of n-grams with respect to the reference translations, with a brevity penalty. A higher BLEU score indicates better translation.

- **mWER** (multi-reference word error rate) (Nießen et al., 2000): This measures the edit distance with the most similar reference translation. Thus, a lower mWER score is desirable.

- **SSER** (subjective sentence error rate) (Nießen et al., 2000): This is calculated using human judgements. Each sentence was judged by a human evaluator on the following five-point scale, and the SSER was calculated as described in (Nießen et al., 2000).

0	Nonsense
1	Roughly understandable
2	Understandable
3	Good
4	Perfect

Again, the lower the SSER, the better the translation.

Table 4 shows the results of the evaluation. We find that using syntactic preprocessing brings substantial improvements over the baseline phrase-based system. While the impact of morphological information is not seen in the BLEU and WER scores, the subjective scores reveal the effectiveness of using morphology. The last two columns of the table show the percentage of sentences that were found by the human judges to be roughly understandable (or higher) and understandable (or higher) respectively in the evaluation scale. We find that including syntactic and morphological information brings substantial improvements in translation fluency.

[4] http://www.speech.sri.com/projects/srilm/

[5] http://www.isi.edu/licensed-sw/pharaoh/

An Example: Consider, again, the example in table 1. The word-order in the baseline translation is woeful, while the translations after syntactic preprocessing (baseline+syn and baseline+syn+morph) follow the correct Hindi order (compare with the reference translation). The effect of suffix separation can be seen from the verb form (देखें (*dekhem*) – visit or see) in the last translation (baseline+syn+morph). The reason for this is that the pair “visit → देखें” is not available to be learnt from the original and the syntactically preprocessed corpora, but the following pairs are: (i) to visit → देखना (ii) worth visiting → देखने योग्य, and (iii) can visit → देख सकते हैं. Thus, the baseline and baseline+syn models are not able to produce the correct verb form for “visit”. On the other hand, the baseline+syn+morph model, due to the suffix separation process, combines देख (*dekha*) and एं (*em*) from different mappings in the aligned corpus, e.g., “visit +ing → देख ने” and “sing → गा एं”, to get the right translation for visit (देखें) in this context.

5 Conclusion

We have presented in this paper an effective framework for English-Hindi phrase-based SMT. The results demonstrate that significant improvements are possible through the use of relatively simple techniques for incorporating syntactic and morphological information.

Since all Indian languages follow SOV order, and are relatively rich in terms of morphology, the framework presented should be applicable to English to Indian language SMT in general. Given that morphological and parsing tools are not yet widely available for Indian languages, an approach like ours which minimizes use of such tools for the target language would be quite desirable.

In future work, we propose to experiment with a more sophisticated morphological analyzer. As more parallel corpora become available, we also intend to measure the effects of using morphology on corpora requirements. Finally, a formal evaluation of these techniques for other Indian languages (especially Dravidian languages such as Tamil) would be interesting.

Acknowledgements

We are grateful to Sachin Anklekar and Saurabh Kushwaha for their assistance with the tedious task of collecting and preprocessing the corpora.

References

Ananthakrishnan Ramanathan and Durgesh Rao, A Lightweight Stemmer for Hindi, *Workshop on Computational Linguistics for South-Asian Languages*, EACL, 2003.

Stanley F. Chen and Joshua T. Goodman, An Empirical Study of Smoothing Techniques for Language Modeling, *Technical Report TR-10-98*, Computer Science Group, Harvard University, 1998.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, A Statistical Approach to Machine Translation, *Computational Linguistics*, 16(2), pages 79–85, June 1990.

Michael Collins, Philipp Koehn, and Ivona Kucerova, Clause Restructuring for Statistical Machine Translation, *Proceedings of ACL*, pages 531–540, 2006.

Jonathan Graehl and Kevin Knight, Training Tree Transducers, *Proceedings of HLT-NAACL*, 2004.

Kuhoo Gupta, Manish Shrivastava, Smriti Singh, and Pushpak Bhattacharyya, Morphological Richness Offsets Resource Poverty – an Experience in Builing a POS Tagger for Hindi, *Proceedings of ACL-COLING*, 2006.

Kenji Imamura, Hideo Okuma, Eiichiro Sumita, Practical Approach to Syntax-based Statistical Machine Translation, *Proceedings of MT-SUMMIT X*, pages

Philipp Koehn and Hieu Hoang, Factored Translation Models, *Proceedings of EMNLP*, 2007.

Philip Koehn, Franz Josef Och, and Daniel Marcu, Statistical Phrase-based Translation, *Proceedings of HLT-NAACL*, 2003.

Daniel Marcu and William Wong, A Phrase-based Joint Probability Model for Statistical Machine Translation, *Proceedings of EMNLP*, 2002.

McGregor, R. S., *Outline of Hindi Grammar*, Oxford University Press, Delhi, India, 1974.

I. Dan Melamed, Statistical Machine Translation by Parsing, *Proceedings of ACL*, 2004.

Guido Minnen, John Carroll, and Darren Pearce, Applied Morphological Processing of English, *Natural Language Engineering*, 7(3), 207–223, 2001.

G. Ngai and R. Florian, Transformation-based Learning in the Fast Lane, *Proceedings of NAACL*, 2001.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney, An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research, *International Conference on Language Resources and Evaluation*, pages 39–45, 2000.

Sonja Nießen and Hermann Ney, Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information, *Computational Linguistics*, 30(2), pages 181–204, 2004.

Franz Josef Och, Christoph Tillman, and Hermann Ney, Improved Alignment Models for Statistical Machine Translation, *Proceedings of EMNLP*, pages 20–28, 1999.

Franz Josef Och, Minimum Error Rate Training in Statistical Machine Translation, *Proceedings of ACL*, 2003.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, *IBM Research Report*, Thomas J. Watson Research Center, 2001.

Maja Popovic and Hermann Ney, Statistical Machine Translation with a Small Amount of Bilingual Training Data, *5th LREC SALTMIL Workshop on Minority Languages*, pages 25–29, 2006.

Durgesh Rao, Natural Language Generation for English to Hindi Human-Aided Machine Translation of News Stories, *Master's Thesis*, Indian Institute of Technology, Bombay, 1996.

Durgesh Rao, Kavitha Mohanraj, Jayprasad Hegde, Vivek Mehta, and Parag Mahadane, A Practical Framework for Syntactic Transfer of Compound-Complex Sentences for English-Hindi Machine Translation, *Proceedings of KBCS*, 2000.

Kenji Yamada and Kevin Knight A Syntax-based Statistical Translation Model, *Proceedings of ACL*, 2001.