A Dependency Treelet-based Phrasal SMT: Evaluation and Issues in English-Hindi Language Pair

Kalika Bali
Microsoft Research India
Bangalore 560080 INDIA
kalikab@microsoft.com

Sankaran Baskaran‡
Microsoft Research India
Bangalore 560080 INDIA
baskaran@microsoft.com

A Kumaran
Microsoft Research India
Bangalore 560080 INDIA
a.kumaran@microsoft.com

Abstract

In this paper, we present a detailed evaluation of a Dependency Treelet-based Phrasal Statistical Machine Translation (SMT) system for English-Hindi language pair. The dependency treelet-based phrasal SMT system that adds the source language syntactic information to a standard phrasal SMT has been shown to perform significantly better than surface based approaches on several well-studied European language pairs. We seek to examine if this observation holds true for languages as diverse as English and Hindi, by developing and testing such a system, for the first time in this language pair. We make baseline comparisons with a standard phrasal SMT implementation, and further study the effect of two radically different types of corpora, namely, technical text and general web text, on the performance of the dependency-treelet based phrasal system. The evaluation includes human judgment, in addition to the two standard automated metrics, namely, BLEU and METEOR. Some language-specific issues are also highlighted that provide an insight into the challenges involved in applying standard phrasal SMT techniques for translation between English and an Indic-language like Hindi.

1 Introduction

The past decade has seen a revolution in the area of Machine Translation (MT) using statistical/corpus-based approaches. The seminal work by Brown et al. (1993) caused a shift in focus for MT systems from rule-based and example-based approaches to corpus-based Statistical Machine Translation (SMT) approaches. Their system considered translation as a noisy channel problem of communication theory, and used a sentence aligned parallel corpus to model word alignments in sentence pairs. The decoder chose the most probable word-alignment path for translating an unseen sentence. While, this was a word-based framework, later phrase-based approaches (Koehn et al., 2003; Vogel et al., 2003) improved upon the word-based SMT, by modeling translations of *phrases*[1]. The state-of-the-art in Statistical SMT since then has advanced significantly from word-based approaches to phrasal SMT and subsequently to treelet based phrasal SMT systems that use source-side syntactic information to the standard phrasal SMT.

Until recently, MT research in Indian languages has largely focused on rule-based approaches (Sinha and Jain 2003; MANTRA; Kavitha et al., 2006) and the use of Interlingua (Dave et al. 2002). However, in recent times there has been a shift towards building blocks for an end-to-end SMT system. The first English-Hindi SMT system to appear in the research literature is by Udupa and Faruquie (2004) based on the IBM framework (Brown et al., 1993). This is a word-based SMT using IBM Models 1, 2 and 3 (Brown et al 1993). The major impediment to any such effort is the lack of large a parallel corpus that is essential for improving the quality of any system based on statistical tech-

‡ Currently affiliated with Simon Fraser University, Canada.
[1] Phrase here is defined as a string of adjacent words and not as a syntactic constituent.

Proceedings of ICON-2008: 6th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008

niques. More recently, Ramanathan et al. (2008) reported significant improvement over standard phrasal SMT by incorporating syntactic and morphological information in the English-Hindi language pair.

In this paper, we present a detailed evaluation of a dependency treelet-based phrasal SMT system (Quirk et al, 2005) for the English-Hindi language pair. A baseline comparison of this system with a surface-phrase system using Pharaoh (Koehn et al., 2003) evaluated with an automated quality metric, namely BLEU (Papineni et al., 2002), is also reported. Further, we report evaluation of the English-Hindi treelet system specifically on two different corpora used for training the system. The impact of these two factors on system performance is studied using different evaluation metrics like BLEU, METEOR and human judgment. Finally, we highlight some language-specific issues that provide an insight into the challenges involved in applying standard phrasal SMT techniques for translation between English and an Indic-language like Hindi.

In sections 2, we describe the surface-phrasal and dependency treelet system respectively. Section 3 presents the evaluation results and analysis. Some Hindi specific challenges are discussed in Section 4 followed by conclusion in Section 5.

2 Dependency Treelet-based Phrasal Statistical Machine Translation System

We have adapted the dependency treelet-based Phrasal SMT system described in (Quirk et al., 2005) for translation between English-Hindi language pair. Though this system has been implemented successfully for several other language pairs, primarily European and East Asian, this is the first time that such a system has been developed for an Indian language. In this section, we briefly describe the systems used and training methodologies for translation between English-Hindi language pair.

2.1 Training

The training process consists of three stages, where several statistical models such as translation model, language model and order model are learnt from suitable data sources.

Dependency parsing and alignment: The system uses a source language dependency parser to parse the source text of the parallel corpora. The parser produces directed, unlabeled, ordered dependency trees and marks the POS tag for each source word.

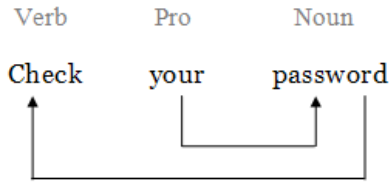

Figure 1: An example dependency tree

An example dependency tree, along with POS information for the words is shown in Figure 1, where the arrow indicates the head annotation

Simultaneously, a word-alignment component - GIZA++, is used to align the source and target language texts at surface level. It derives many-to-many alignments by running IBM models in both directions and combines the results heuristically. Before aligning the texts, a target language word-breaker is used to segment the words in target text, using language specific rules.

Projecting dependency trees: The dependency parsed sentences of the source are then projected onto the word-aligned parallel texts to produce word-aligned parallel dependency corpus. In the case of one-to-one mappings, the projections are simple and the target tree becomes isomorphic to the source. In many-to-one alignments, multiple source words that are linked in the tree get projected onto a single target word. For one-to-many alignments, a single source word corresponds to several target words that are contiguous in the tree. The system projects the source node to the rightmost word of the target phrase. Figure 2, illustrates the process of dependency tree projection.

The words that are unaligned in the target sentence are attached to the closest lower node to the left [or right] in the dependency tree. Similarly, if all the nodes to the left [or right] of a word w_j are unaligned, the word is attached to the leftmost [or right-most] word that is aligned. The resulting target dependency tree may not be in the same surface ordering as the original sentence, due to the

projection of source dependency structure onto the target.

This anomaly is rectified by a reattachment pass, which reattaches each wrong node to the lowest of its ancestors, so as to generate the correct surface ordering.

Extracting treelet pairs: Individual treelets are then extracted from this corpus, which together form a treelet translation model of source and target translation pairs. Here a treelet is defined as an arbitrary connected subgraph of the aligned parallel dependency tree. For every possible treelet in source side, all the target nodes that are aligned to the source treelet are identified and if these target nodes form a treelet by itself, then both the source and target treelets are retained along with alignment and word order details.

The frequency count for the treelets is also maintained, which is then used for maximum likelihood estimation.

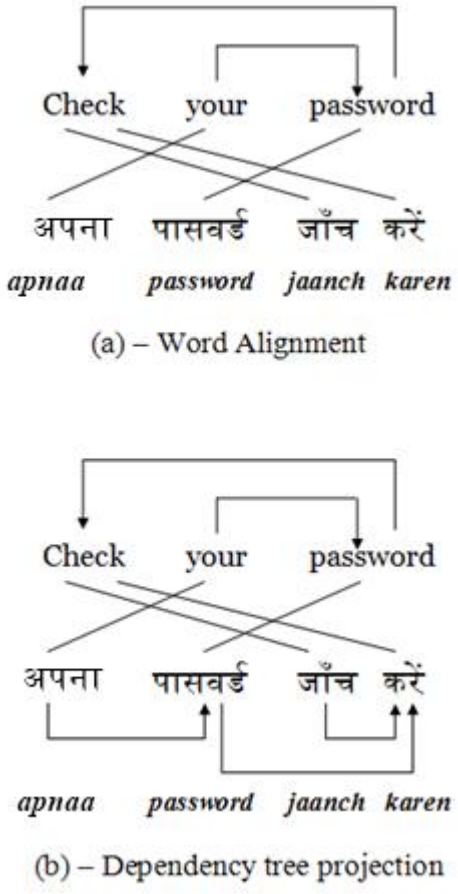

Figure 2: Projection of dependencies

A threshold is used to limit the size of the treelet, so as to limit the possible combinations and this has been fixed to maximum of four source nodes based on observation.

Statistical models: There a number of models possible to build a statistical model for the phrasal SMT system. Here we discuss a few:

Order model: Standard phrasal SMT systems often use a model to rank various possible orderings of a set of phrases using simple technique like, estimating the probability of a source phrase *i* getting translated to a target phrase in position *j*.

The present treelet system incorporates syntactic information for ordering the phrases. It assigns a probability to the order of a target treelets given the sequence of source treelets. It is simplified by an assumption that *phrases move as a whole*, which predicts the probability of each given ordering of modifiers independently. This can be represented as

$$P\,(order\,(T)|S,T) = \prod_{t \in T} P\,(\,order(c(t))\,|\,S,T)$$

where, *S* and *T* are the source and target treelet sequences respectively, and *c* is the function returning the list of nodes modifying *t*.

Further, it is assumed that the position of each child can be modeled independently in terms of a head-relative position:

$$P\,(order(T)|S,T) = \prod_{m \in 2c(t)} P\,(pos(m,t)|S,T)$$

P(pos (m,t) | S, T) is modeled using a small set of features from the dependency tree, such as,

- Lexical items corresponding to the head and the modifier
- Source lexical items aligned to the head and the modifier
- POS of the source nodes aligned to the head and the modifier
- Head relative position of source node aligned to the source modifier

The training corpus acts as a supervised training set for extracting feature vector for each node in the target tree, which are then used to train a decision tree (Quirk et al., 2005).

Channel model: The system uses two distinct channel models, a maximum likelihood estimate (MLE) model and one using IBM Model-1 word-to-word alignment probabilities. The MLE is effective in modeling idioms and other non-literal translation of phrases but suffers from data sparseness since these occur very rarely in the corpus. In contrast, word-to-word model is biased towards literal translations and is not as effective for idiomatic uses.

Target model: For improving the fluency of the translation, it uses a trigram language model with Kneser-Ney smoothing. The surface string is obtained from the ordered target dependency tree.

2.2 Decoding

Unlike string-based approaches, which use simple left-to-right decoding, the decoding of treelet-based approach is more complicated, due to the discontiguous and/or overlapping treelets and combinatorial explosion of the search space.

The decoder begins with a bottom up, exhaustive search, where all treelet translation pairs matching the input dependency trees are identified. Treelet translation pairs are selected subject to the following constraint: A treelet translation pair *x* is said to *match* the input tree *S* if and only if there is some connected subgraph $\boldsymbol{S}'$ that is identical to the source side of *x*. Then *x* is said to *cover* all the nodes in $\boldsymbol{S}'$ and is *rooted* at source node *s*, where *s* is the root of the matched sub graph $\boldsymbol{S}'$.

After identifying the treelet pairs, they are placed on a list associated with the input mode, where the match is rooted. Then moving bottom up through the input tree, a list of *candidate translations* are computed for the input sub tree rooted at each node *s*, as follows: Consider in turn each treelet translation pair *x* rooted at *s*. The treelet pair *x* may cover only a portion of the input sub tree rooted at *s*. Find all descendents $\boldsymbol{S}'$ of *s* that are not covered by *x*, but whose parent $\boldsymbol{S}''$ is covered by *x*. At each such node $\boldsymbol{S}''$ looks at all interleaving of the children of $\boldsymbol{S}''$ specified by *x*, if any, with each translation $\boldsymbol{t}'$ from the candidate translation list of each child. Each such interleaving is scored using the models previously described and added to the candidate translation list for that input node and the best scoring candidate is then taken as the resultant translation.

3 Evaluation

The surface-phrasal SMT was used as a baseline and the translation quality of the dependency treelet based system was evaluated using BLEU, METEOR and human judgment. We show that both BLEU and METEOR largely correlate with human assessment in the English-Hindi pair.

3.1 Data

We used two distinct parallel corpora to train the system separately. The technical domain corpora (Microsoft product support articles, product documentation, etc.) had ~100K parallel sentences in English and Hindi. Product names, some technical terms, commands etc. were not translated to Hindi and instead the English term/phrase was retained in the translation. For example, phrases such as *Microsoft Office*, *SMTP*, *Windows* were retained as English strings in the translations (in Hindi) and were not even transliterated.

It should be noted that in the case of technical documents the translation vocabulary and word choices are highly restricted, and the language style followed is also standardized as applicable to the formal register of the documents. Thus, the word *you* would always be translated to ***aap*** (honorific form of *you*) and never as ***tum*** (normal form of *you*).

The second corpus was representative of more real-life documents collected from WebDunia - a multilingual-portal-and it too had ~100K parallel sentences. WebDunia corpus had parallel translations in 8 different sub-domains including news, commerce, health, sports and miscellaneous. News, interview and miscellaneous formed a bulk of this with 54K, 15K and 12K parallel sentences respectively. In this corpus, the Hindi article was translated not at sentence-level but at article level. Thus, parallel sentences were carefully identified from English-Hindi article pairs manually. Hence, the WebDunia corpus used for training the systems was very rich in vocabulary and creative language style.

3.2 Baseline System

A Phrase-based system considers a sequence of words, or *phrases* as the unit for translation. First, the input is segmented into a number of phrases, which are then translated into the target language

individually and finally the translated segments are reordered based on reordering models and language models.

Most of the approaches to learn phrase translations between the source and the target languages (Och and Ney, 2003; Koehn et al., 2003; Venugopal et al., 2003) use word alignment as their basis. For our implementation of a baseline Phrasal English-Hindi SMT we use the publicly available Thot toolkit (Ortiz et al., 2005) to obtain the phrase translation model based on Koehn et al. (2003). Both Och and Ney (2003) and Koehn et al. (2003) use the space between the intersection and union of the alignments produced by GIZA++ using heuristics for expanding from the intersection.

We used Pharaoh – a beam search decoder for decoding the English sentences after learning phrase translations and target language model. A thorough description of the decoder can be obtained from (Koehn et al., 2003).

We used around 80K of sentence pairs separately for training the technical and web domain systems. 2,000 sentences were used for development testing and parameter tuning and 500 sentences for development training. We used two different testing sets consisting of 500 and 1000 sentences. However, human assessment was carried out only on the 500 sentence dataset.

The raw BLEU scores for the baseline phrasal SMT are given in Table 1.

BLEU Scores	Test Set Size	
	500	1000
Tech Doc	26.04	29.81
Web Corpus	14.78	14.72

Table 1: Baseline Phrasal SMT System (BLEU)

The web corpus of the phrase-based system gets 14.72 BLEU (for 1000 test sentences), whereas for the technical documents the score is much higher at 29.81. The earlier work on a word-based English-Hindi SMT system (Udupa and Faruquie, 2004) report BLEU scores of 13.91 on 1032 sentences from a web-corpus, but as they used a different corpus the results are not directly comparable and may taken as indicative only. However, the results of this phrase-based SMT are comparable with the results that are subsequently reported in for the dependency treelet system, as we have used common training and test data in all these experiments.

3.3 Results : Dependency Treelet System

BLEU Evaluation: Table 2 gives the BLEU scores for the same two domains on the dependency treelet-based Phrasal SMT system that we had developed.

BLEU Scores	Test Set Size	
	500	1000
Tech Doc	32.54	31.56
Web Corpus	16.16	16.46

Table 2: Dependency Treelet-based Phrasal SMT (BLEU)

We achieve BLEU scores of 32.54 and 16.16 for technical documents and web corpus for 500 test sentences. Note that the BLEU score for the web corpus is much lower than that obtained for the technical corpus. Comparing the two phrase-base systems, it can also be observed that the surface phrasal system is closely trailing the treelet-based phrasal system by about 1-2 BLEU points for the web corpus and about 2-5 BLEU points for the technical texts. It should be noted that the BLEU scores reported were the scores obtained when the treelet-based phrasal system was trained with identical training parameters (as that of phrasal system) for both technical text and web text corpora. Tuning the phrasal SMT system without such constraint resulted in even better performance; for example a BLEU score of about 36 was achieved by fine-tuning the treelet-based phrasal SMT for technical texts, which compares favorably with the BLEU score reported (40.66) for the state-of-the-art English-French SMT system trained on 1.5 million sentences of technical data (Quirk et al., 2005).

We can see from the results reported that there is a significant difference in performance between surface phrasal SMT and treelet-based phrasal SMT, and between technical and web corpus. To better understand this difference we performed METEOR evaluation as well as evaluation based on human judgment, the results of which are presented in the following sections.

METEOR Evaluation: METEOR (Banerjee and Lavie, 05) is similar to BLEU but addresses several shortcomings of the latter. Given a system translation and one or more references, it assigns a score to the system translation based on the precision and recall of overlapping unigrams. Additionally, METEOR provides for matching stemmed form of the words and for the usage of synonymous words by using a WordNet (Fellbaum, 1998). Unlike BLEU, METEOR assigns a score for every sentence and hence is expected to be useful in categorizing error patterns.

To compensate for the lack of higher order n-grams, METEOR introduces a *penalty*, accounting for the number of overlapping chunks. Finally the overall score for a sentence is then calculated from the Fmean and penalty as follows:

$$Score = Fmean * (1 - Penalty)$$

The METEOR scores presented here are obtained by using only the exact match module and without using stemming and WordNet's Synonymy modules due to the lack of availability of Hindi stemmer and Hindi WordNet respectively, at the time of conducting these experiments. Though the use of these modules can increase the scores for the system, we believe that the raw METEOR score will be sufficient for the purposes of studying the difference in BLEU of technical texts and web corpus.

As METEOR is primarily based on unigram precision and recall, it implies that function words present in the translations will be given the same significance as the content words. To avoid this, Banerjee et al. (2005) suggest the removal of function words from the translations before getting the translation score. However, in Hindi many function words also mark grammatical features like gender and number that agree with some other word(s) in the sentence (section 4.1) and hence are an important clue for syntactic wellformedness of the sentence. Thus, words such as ***haiN*** (plural copula), ***thaa*** (past tense, masculine marker), ***kaa*** (masculine singular possessive marker), etc. were retained in the evaluation texts.

To study the effect of the function words fully, we performed METEOR evaluation on three versions of the test set. The first set M1 represents the raw test data without any processing, while for set M2, we removed the function words from the test set. We then pruned the function word list to remove the words marking any agreement with other word(s) and this list was then used for creating the test set M3. The overall METEOR scores for the three test sets are given in Table 3. It may be observed that the results of all the three test sets follow the same pattern as that of BLEU, with technical system outperforming the web corpus.

Corpora	Test Set	Test Set Size	
		500	1000
Tech Doc	M1	0.699	0.713
	M2	0.739	0.748
	M3	0.708	0.722
Web Corpus	M1	0.429	0.404
	M2	0.411	0.388
	M3	0.420	0.399

Table 3: Dependency Treelet-based Phrasal SMT (METEOR Scores)

Though they may help in illustrating a trend, it is difficult to fully understand these raw scores without inspecting the Precision, Recall and Penalty for each of these cases. Table 4 gives the complete details of the METEOR experiments performed with the three versions of the test set.

METEOR		Tech Docs			Web Corpus		
		Precision	Recall	Penalty	Precision	Recall	Penalty
M1	500	.748	.723	.037	.509	.489	.127
	1000	.734	.752	.033	.466	.474	.138
M2	500	.777	.773	.045	.516	.486	.157
	1000	.790	.779	.041	.491	.464	.169
M3	500	.765	.737	.044	.512	.482	.135
	1000	.781	.748	.039	.488	.462	.145

Table 4: METEOR: Precision, Recall & Penalty

It can be seen from these results that the precision and recall has been consistently higher for the technical documents, while the web corpus system has higher penalty. Recall that, in METEOR, the penalty increases with the decrease in the number of longest matches between the reference and system translations. Between M2 and M3, M2 is penalized more; this is due to the fact that the function words marking the morphological agreements were removed from the data for M2. This further strengthens our argument that function words that mark

various morphological attributes are critical for improved accuracy. Also, on manual inspection of the results of the web corpus we found numerous sentences where the constituents did not agree with each other. This partly explains the lower precision and recall in Table 4.

Human Evaluation: Human evaluation was done for 500 sentences each from two corpora with two human judges rating the individual sentences for fluency and adequacy.

System	Evaluation	Fluency	Adequacy	Average
Tech Docs	H1	3.286	3.272	3.279
	H2	3.75	3.668	3.709
	Average	3.518	3.47	3.494
Web Corpus	H1	2.976	2.98	2.978
	H2	2.076	1.996	2.036
	Average	2.526	2.488	2.507

Table 5: Human Assessment Results

The evaluation was blind, wherein the judges were given both reference and system translations in different order, so as to avoid any bias. They were asked to score both sentences on a scale of 1-5 separately for fluency and adequacy and these were then averaged. Table 5, shows the ratings for the system translation by judges H1 and H2. The combined score in last column is the average of fluency and adequacy. While, one can see noticeable inter-annotation differences in ratings, the average fluency and adequacy for two corpora clearly point to the better translation quality of technical texts. We are not presenting the correlation between human judgment and BLEU/METEOR, because our intention is to validate the poor performance of web corpus system rather than to evaluate the automatic MT evaluation metrics.

Miscellaneous: We also analyzed other factors- average sentence length, number of unknown words and number of small sentences- that could possibly explain the poor performance of the web corpus system. All these analyses were done on the same set of sentences used for human assessment

The results (Table 6) throw-up some interesting facts about the characteristic of the two corpora. We find that about 1.63% (50 of 3060) of the words in the technical documents is unknown, while 8.55% (587 of 6861) of the words in web corpus are out-of-vocabulary (OOV) words. It may be noted that the OOV words in the test set were not translated into Hindi and were retained in the output as English strings.

System	Unknown Words (%)	Sentence Length	# Small Sentences
Tech Docs	1.63	6.12	280
Web Corpus	8.55	13.72	15

Table 6: Dataset Characteristics

We also found the web data to be more than twice in length (13.72) than the technical documents (6.12). Finally, we also observed that the technical documents had a significant percentage of smaller sentences (< 5 words) as shown in the last column. It has been observed that when the sentence length increases, the errors by the decoder, order model etc. might cascade leading to poor translation quality.

We also found the web data to be more than twice in length (13.72) than the technical documents (6.12). Finally, we also observed that the technical documents had a significant percentage of smaller sentences (< 5 words) as shown in the last column. It has been observed that when the sentence length increases, the errors by the decoder, order model etc. might cascade leading to poor translation quality.

4 Language Specific Issues

In this section we explore specific challenges encountered in Hindi, while applying the standard phrasal SMT techniques for the English-Hindi language pair; some of these issues are being addressed in our current research.

4.1 Agreement in Hindi

Hindi uses a system of postpositions and suffix morphemes to mark grammatical features such as gender, number, person etc. Agreement between several sentence constituents with respect to these features is an important characteristic of the language. Also, like French and Spanish, all Hindi nouns, including inanimate articles and abstract

nouns are assigned a masculine or a feminine gender. The agreements in Hindi are summarized below.

Example 1: Adjectives are inflected to agree with the gender of the modifying noun

लम्बी	लड़की
lambii	***laRkii***
tall	girl

In this example, the adjective ***lambii*** (लम्बी) is a feminine adjective that agrees with the feminine noun ***laRkii*** (लड़की).

Example 2: Verbs with slightly complex morphology, should agree with the subject for gender and number, while the auxiliary verbs and copula should agree with the finite main verb for tense.

काली	बिल्ली	खाना	खाती	है।
kaalii	***billii***	***khaanaa***	***khaati***	***hai.***
black	cat	food	eat	copula

In this example, the verb ***khaatii*** (खाई) agrees for gender (feminine in this case) with the Subject ***billii*** (बिल्ली), and the copula ***hai*** (है) agrees with the main verb for tense (present tense).

Example 3: Transitive verbs, when occurring in perfective tense has to agree with its object for gender.

राम	ने	रोटी	खाई।
raam	***ne***	***roTii***	***khaaii.***
Ram	erg.	bread	ate

Here, the verb ***khaaii*** (खाई) agrees with the feminine object ***roTii*** (रोटी) and not the masculine subject ***raam*** (राम).

Of these the first one is a local agreement, while the others mark long-distance agreements. In the majority of the sentences, verb morphology results in long distance agreements not only at the surface level but also in the dependency tree. The treelets, culled out of the aligned dependency tree are woefully inadequate in capturing these subtle variations expressed at morpheme level. Thus, the output translation has postpositions and affixes that do not agree with each other, seriously affecting the quality as found in both readability (human judgment) and BLEU evaluation.

Even for the cases of local agreements, such as the one between adjective and noun, the treelet system produces mismatching words, that is, feminine form of an adjective with a masculine noun.

On doing a frequency analysis of such forms in the web corpus it was found that the feminine form of adjectives is more frequent than the corresponding masculine form.

For example, the English word 'big' can have three forms in Hindi: ***baRaa*** (बड़ा), ***baRii*** (बड़ी) and ***baRe*** (बड़े) for masculine, feminine (singular and plural) and masculine plural respectively as demanded by the gender of the modifying noun. As the feminine has only one form for both the singular and plural, it is far more frequent than both the masculine singular (31% more) and plural (11% more) forms. This misleads the decoder to wrongly produce the feminine form even for masculine nouns.

Oracle Experiment: We performed an oracle experiment to quantify the effect of the agreement morphology on the translation quality. For this we developed a partial stemmer that stems only the words marking various agreements. The reference and system translations were normalized using this stemmer and then evaluated using BLEU. A significant increase in BLEU scores was observed (Table 7), for the test set of 500 sentences, BLEU increased by 37.7% and 36.6% for technical and web corpus respectively.

Surprisingly, the improvement is also equally significant for technical domain, which is not as diverse as the web corpus. This could be attributed to the correct assignment of nominal and verbal morphology. We also sought to know, whether the word with the correct affix (as found in the reference translation) had occurred in the training data.

	Test Set Size	
	500	**1000**
Tech Doc	44.81	41.18
Web Corpus	22.08	21.09

Table 7: Oracle BLEU Scores

As a part of this experiment on the test data, whenever the system failed to produce a word with the correct affix, we checked whether the correct form of the word was already seen in the training data. If the word is indeed known, then the problem lies with the decoder and a better scoring mechanism would be needed. Alternately, if the word is unknown, we would need a morphological generator to produce the word with the appropriate inflections.

In both the domains, we found that the correct word forms had always occurred in the training corpus. This underlines the need for a better scoring mechanism that would take into account the different morphological forms of the word while choosing the correct word form**.**

4.2 Relative Clauses in Hindi

The relative/conditional sentences in Hindi exhibit another form of long distance agreement where pairs of relational pronoun are used to link the two clauses.

Example 4: Consider the English sentence:

Diabetes mellitus happens when the insulin does not work.

the Hindi translation of which will look like:

जब	इनसुलिन	कार्य	नहीं	करती	तब
jab	***insulin***	***kaarya***	***nahiin***	***kartii***	***tab***
when	insulin	work	not	does	then

मधुमेह	की	बिमारी	हो	जाती	है।
madhumeh	***kii***	***bimaarii***	***ho***	***jaatii***	***hai.***
diabetes	gen.	disease		happens	

where ***jab*** (जब) (meaning, when) and ***tab*** (तब) (meaning, then) are relative pronouns.

Similarly, Hindi has several pairs of relative-correlative pronoun pairs such as: ***jo*** (जो) ... ***vo*** (वो) (meaning, who), ***jin*** (जिन) ***... ve*** (वे) (meaning, they). We can represent the above Hindi sentence by the following structure:

<RP-1> <relative clause> <RP-2> <main clause>

Example 5: The same sentence can also be written in a slightly different but lesser used form as:

<main clause> <RP-1> <relative clause>

as in the following example:

मधुमेह	की	बिमारी	हो	जाती	है
madhumeh	***kii***	***bimaarii***	***ho***	***jaatii***	***hai***
diabetes	gen.	disease		happens	

जब	इनसुलिन	कार्य	नहीं	करती।
jab	***insulin***	***kaarya***	***nahiin***	***kartii***
when	insulin	work	not	does

During real time testing, when given similar English sentences, the treelet-based phrasal system always produced the second form (5) but with the words within the two clauses in incorrect order. Hence, as the relative-correlative pronouns exhibit long distance agreement in Hindi, the treelet-based phrasal system that essentially models local syntactic constituency, fails to learn this correctly.

This result is different from the earlier experiments in French where the treelet-based phrasal system (Quirk et al., 2005) was able to handle discontiguous phrases like *ne ... pas* (not). However, it should be noted that unlike, the *ne ... pas* (not) construction in French, which can have only the conjugated verb in between, an entire clause can occur between the relative-correlative pronouns in Hindi.

In the case of relative clauses, we were unable to study the impact of discontiguous phrases on BLEU, due to the ordering issues. However, as the Oracle experiment reported in the previous section indicates, a classifier that can handle long distance agreements in Hindi will go a long way in improving the accuracy of the system.

5 Conclusion

In this paper, we presented the first ever baseline implementation of an English-Hindi dependency treelet-based SMT. In a comparison with surface-phrasal SMT using Pharaoh, the dependency treelet system clearly outperforms due to the effective use of source language syntactic information. An exhaustive evaluation of the treelet-based system in two distinct domains- technical documents and

web corpus, automated metrics, viz., BLEU, METEOR as well as a blind evaluation by humans, showed that the results are consistent across different evaluation methodologies for the two corpora. The translated output in the web corpus domain was found to be poor due to the rich and diverse nature of the corpus vis-à-vis language usage.

Interestingly, we could pinpoint some of the drop in quality due to Hindi-specific issues, such as agreements, both at the local level and at long distance. This clearly indicates that unlike the baseline implementation that has been discussed in this paper that takes into account only the source-language syntax, further research is required to use target language syntactic information as well in an integrated manner with the treelet-based phrasal SMT system.

Going forward, we would like to focus on this aspect and taking Hindi agreement as a case in point, explore means and ways of addressing target level syntactic complexities to improve the performance of such an SMT system. We believe that like the approach, like the one described in this paper, would be equally applicable across similar language pairs, resulting in a viable methodology for making statistical machine translation systems possible for all Indian languages.

Acknowledgments

We thank the invaluable help and insightful comments by Arul Menezes and Chris Quirk during the training of English-Hindi SMT system. We are also grateful to Raghavendra Udupa for many hours of discussions that benefitted the work.

References

Banerjee, S. and Lavie, A. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements*. Workshop on Intrinsic and Extrinsic Evaluation measures for MT and/or summarization in ACL-05.

Brown, P.F., Pietra, S.A.D., Pietra, V.J.D. and Mercer, R.L. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics, 19(2), pp 263-311.

Dave, S., Parikh, J. and Bhattacharyya, P. 2002. *Interlingua based English- Hindi Machine Translation and Language Divergence*. In Journal of Machine Translation. Vol17.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, ISBN-10: 0-262-06197-X.

Kavitha M, Ananthakrishnan R, Hegde, J. J., Shekhar, C., Shah, R. Bade, S. and Sasikumar M. 2006. *MaTra: A Practical Approach to Fully-Automatic Indicative English-Hindi Machine Translation*. First Natl. Symposium on Modeling and Shallow Parsing of Indian Languages, 2006.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. In ACL-02.

Koehn, P., Och, F. J. and Marcu, D. 2003. *Statistical Phrase-based Translation*. In Proceedings of the Joint HLT/NAACL Conference.

Lavie, A., Sagae, K. and Jayaraman, S. 2004. *The Significance of Recall in Automatic Metrics for MT Evaluation*. In AMTA-04.

MANTRA. Centre for Development of Advanced Computing. *http://www.cdac.in/html/aai/mantra.asp*.

Och, F. J. and Ney, H. 2003. *A Systematic Comparison of Various Statistical Alignment Models*. In Computational Linguistics, 29(1):19-52.

Ortiz-Martınez, D., Garćıa-Varea, I. and Casacuberta, F. 2005. *Thot: A Toolkit to Train Phrase based Models for Statistical Machine Translation*. In MT Summit.

Quirk, C., Menezes, A., and Cherry, C. 2005. *Dependency Treelet Translation: Syntactically Informed Phrasal SMT*. In ACL-05.

Ramanathan, A., Bhattacharyya, P., Hegde, J., Shah, R. M. and Sasikumar, M. *Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation*. The 3rd International Joint Conference on Natural Language Processing, 2008.

Sinha, R. M. K. and Jain, A. 2003. *AnglaHindi: An English to Hindi Machine-Aided Translation System*. MT Summit 2003.

Udupa R. and Faruquie T. 2004. *An English-Hindi Statistical Machine Translation System*. In Proceedings of IJCNLP-04.

Venugopal, A., Vogel, S. and Waibel, A. 2003. *Effective Phrase Translation Extraction from Alighment Models*. In ACL 2003.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B. and Waibel, A. 2003. *The CMU Statistical Machine Translation System*. In MT Summit.

Zens, R. and Ney, H. 2004. *Improvements in Phrase-based Statistical Machine Translation*. In Proceedings of the Joint HLT/NAACL Conference.