Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT

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

Abstract

We report in this paper our work on accurately generating case markers and suffixes in English-to-Hindi SMT. Hindi is a relatively free word-order language, and makes use of a comparatively richer set of case markers and morphological suffixes for correct meaning representation. From our experience of large-scale English-Hindi MT, we are convinced that fluency and fidelity in the Hindi output get an order of magnitude facelift if accurate case markers and suffixes are produced. Now, the moot question is: *what entity on the English side encodes the information contained in case markers and suffixes on the Hindi side*? Our studies of correspondences in the two languages show that case markers and suffixes in Hindi are predominantly determined by the combination of suffixes and semantic relations on the English side. We, therefore, augment the aligned corpus of the two languages, with the correspondence of English suffixes and semantic relations with Hindi suffixes and case markers. Our results on 400 test sentences, translated using an SMT system trained on around 13000 parallel sentences, show that *suffix + semantic relation → case marker/suffix* is a very useful translation factor, in the sense of making a significant difference to output quality as indicated by subjective evaluation as well as BLEU scores.

1 Introduction

Two fundamental problems in applying statistical machine translation (SMT) techniques to English-Hindi (and generally to Indian language) MT are: i) the wide syntactic divergence between the language pairs, and ii) the richer morphology and case marking of Hindi compared to English. The first problem manifests itself in poor word-order in the output translations, while the second one leads to incorrect inflections (word-endings) and case marking. Being a free word-order language, Hindi suffers badly when morphology and case markers are incorrect.

To solve the former, word-order related, problem, we use a preprocessing technique, which we have discussed in (Ananthakrishnan et al., 2008). This procedure is similar to what is suggested in (Collins et al., 2005) and (Wang, 2007), and results in the input sentence being reordered to follow Hindi structure.

The focus of this paper, however, is on the thorny problem of generating case markers and morphology. It is recognized that translating from poor to rich morphology is a challenge (Avramidis and Koehn, 2008) that calls for deeper linguistic analysis to be part of the translation process. Such analysis is facilitated by factored models (Koehn et al., 2007), which provide a framework for incorporating lemmas, suffixes, POS tags, and any other linguistic factors in a log-linear model for phrase-based SMT. In this paper, we motivate a factorization well-suited to English-Hindi translation. The factorization uses semantic relations and suffixes to generate inflections and case markers. Our experiments include two different kinds of semantic relations, namely, dependency relations provided by the Stanford parser, and the deeper semantic roles (agent, patient, etc.) provided by the universal networking language (UNL). Our experiments show that the use of semantic relations and syntactic reordering leads to substantially better quality translation. The use of even moderately accurate semantic relations has an especially salubrious effect on fluency.

2 Related Work

There have been quite a few attempts at including morphological information within statistical MT. Nießen and Ney (2004) show that the use of morpho-syntactic information drastically reduces the need for bilingual training data. Popovic and Ney (2006) report the use of morphological and syntactic restructuring information for Spanish-English and Serbian-English translation.

Koehn and Hoang (2007) propose factored translation models that combine feature functions to handle syntactic, morphological, and other linguistic information in a log-linear model. This work also describes experiments in translating from English to German, Spanish, and Czech, including the use of morphological factors.

Avramidis and Koehn (2008) report work on translating from poor to rich morphology, namely, English to Greek and Czech translation. They use factored models with case and verb conjugation related factors determined by heuristics on parse trees. The factors are used only on the source side, and not on the target side.

To handle syntactic differences, Melamed (2004) proposes 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.

Another method for handling syntactic differences is preprocessing, which is especially pertinent when the target language does not have parsing tools. 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. (2005) propose German clause restructuring to improve German-English SMT, while Wang et al. (2007) present similar work for Chinese-English SMT. Our earlier work (Ananthakrishnan et al., 2008) describes syntactic reordering and morphological suffix separation for English-Hindi SMT.

3 Motivation

The fundamental differences between English and Hindi are:

- English follows SVO order, whereas Hindi follows SOV order
- English uses post-modifiers, whereas Hindi uses pre-modifiers
- Hindi allows greater freedom in word-order, identifying constituents through case marking
- Hindi has a relatively richer system of morphology

We resolve the first two syntactic differences by reordering the English sentence to conform to Hindi word-order in a preprocessing step as described in (Ananthakrishnan et al., 2008).

The focus of this paper, however, is on the last two of these differences, and here we dwell a bit on why this focus on case markers and morphology is crucial to the quality of translation.

3.1 Case markers

While in English, the major constituents of a sentence (subject, object, etc.) can usually be identified by their position in the sentence, Hindi is a relatively free word-order language. Constituents can be moved around in the sentence without impacting the core meaning. For example, the following sentence pair conveys the same meaning (John saw Mary), albeit with different emphases.

जॉन ने मेरी को देखा
John ne Mary ko dekhaa
John-nom Mary-acc saw

मेरी को जॉन ने देखा
Mary ko John ne dekhaa
Mary-acc John-nom saw

The identity of John as the subject and Mary as the object in both sentences comes from the case markers ने (*ne* – nominative) and को (*ko* – accusative). Therefore, even though Hindi is predominantly SOV in its word-order, correct case marking is a crucial part of making translations convey the right meaning.

3.2 Morphology

The following examples illustrate the richer morphology of Hindi compared to English:

Oblique case: The plural-marker in the word "boys" in English is translated as ए (*e* – plural direct) or ओं (*on* – plural oblique):

The boys went to school.
लडके पाठशाला गये
ladake paathashaalaa gaye

The boys ate apples.
लडकों ने सेब खाये
ladokon ne seba khaaye

Future tense: Future tense in Hindi is marked on the verb. In the following example, "will go" is translated as जायेंगे (*jaaenge*), with एंगे (*enge*) as the future tense marker:

The boys will go to school.
लडके पाठशाला जायेंगे
ladake paathashaalaa jayenge

Causative constructions: The आया (*aayaa*) suffix indicates causativity:

The boys made them cry.
लडकों ने उन्हे रुलाया
ladakon ne unhe rulaayaa

3.3 Sparsity

Using a standard SMT system for English-Hindi translation will cause severe data sparsity with respect to case marking and morphology.

For example, the fact that the word *boys* in oblique case (say, when followed by ने (*ne*)) should take the form लडकों (*ladakon*) will be learnt only if the correspondence between *boys* and लडकों ने (*ladakon ne*) exists in the training corpus. The more general rule that ने (*ne*) should be preceded by the oblique case ending ओं (*on*) cannot be learnt. Similarly, the plural form of *boys* will be produced only if that form exists in the training corpus.

Essentially, all morphological forms of a word and its translations have to exist in the training corpus, and every word has to appear with every possible case marker, which will require an impossible amount of training data. Therefore, it is imperative to make it possible for the system to learn general rules for morphology and case marking. The next section describes our approach to facilitating the learning of such rules.

4 Approach

While translating from a language of moderate case marking and morphology (English) to one with relatively richer case marking and morphology (Hindi), we are faced with the problem of extracting information from the source language sentence, transferring the information onto the target side, and translating this information into the appropriate case markers and morphological affixes.

The key bits of information for us are suffixes and semantic relations, and the vehicle that transfers and translates the information is the factored model for phrase based SMT (Koehn 2007).

4.1 Factored Model

Factored models allow the translation to be broken down into various components, which are combined using a log-linear model:

$$p(\mathbf{e}|\mathbf{f}) = \frac{1}{Z} exp \sum_{i=1}^{n} \lambda_i h_i(\mathbf{e}, \mathbf{f}) \qquad (1)$$

Each h_i is a feature function for a component of the translation (such as the language model), and the λ values are weights for the feature functions.

4.2 Our Factorization

Our factorization, which is illustrated in figure 1, consists of:

1. a lemma to lemma translation factor (boy → लडक् (*ladak*))
2. a suffix + semantic relation to suffix/case marker factor (-s + subj → ए (*e*))
3. a lemma + suffix to surface form generation factor (लडक् + ए (*ladak* + *e*) → लडके (*ladake*))

The above factorization is motivated by the following:

- Case markers are decided by semantic relations and tense-aspect information in suffixes.

  For example, if a clause has an object, and has a perfective form, the subject usually requires the case marker ने (*ne*).

  John ate an apple.

  John|empty|subj eat|ed|empty an|empty|det apple|empty|obj

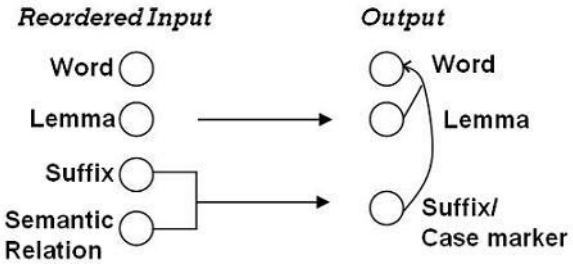

Figure 1: **Semantic and Suffix Factors**: the combination of English suffixes and semantic relations is aligned with Hindi suffixes and case markers

जॉन ने सेब खाया

john ne seba khaayaa

Thus, the combination of the suffix and semantic relation generates the right case marker (ed|empty + empty|obj → ने (*ne*)).

- Target language suffixes are largely determined by source language suffixes and case markers (which in turn are determined by the semantic relations)

  The boys ate apples.

  The|empty|det boy|s|subj eat|ed|empty apple|s|obj

  लडकों ने सेब खाये

  ladakon ne seba khaaye

  Here, the plural suffix on *boys* leads to two possibilities – लडके (*ladake* – plural direct) and लडकों (*ladakon* – plural oblique). The case marker ने (*ne*) requires the oblique case.

- Our factorization provides the system with two sources to determine the case markers and suffixes. While the translation steps discussed above are one source, the language model over the suffix/case marker factor reinforces the decisions made.

  For example, the combination लडका ने (*ladakaa ne*) is impossible, while लडकों ने (*ladakon ne*) is very likely. The separation of the lemma and suffix helps in tiding over the data sparsity problem by allowing the system to reason about the suffix-case marker combination rather than the combination of the specific word and the case marker.

5 Semantic Relations

The experiments have been conducted with two kinds of semantic relations. One of them is the relations from the Universal Networking Language (UNL), and the other is the grammatical relations produced by the Stanford parser.

The relations in both UNL and the Stanford dependency parser are strictly binary and form a directed graph. These relations express the semantic dependencies among the various words in the sentence.

Stanford: The Stanford dependency parser (Marie-Catherine and Manning, 2008) uses 55 relations to express the dependencies among the various words in a sentence. These relations form a hierarchical structure with the most general relation at the root. There are various argument relations like subject, object, objects of prepositions, and clausal complements, modifier relations like adjectival, adverbial, participial, and infinitival modifiers, and other relations like coordination, conjunct, expletive, and punctuation.

UNL: The 44 UNL relations[1] include relations such as agent, object, co-agent, and partner, temporal relations, locative relations, conjunctive and disjunctive relations, comparative relations and also hierarchical relationships like part-of and an-instance-of.

Comparison: Unlike the Stanford parser which expresses the semantic relationships through grammatical relations, UNL uses attributes and universal words, in addition to the semantic roles, to express the same. Universal words are used to disambiguate words, while attributes are used to express the speaker's point of view in the sentence.

UNL relations, compared to the relations in the Stanford parser, are more semantic than grammatical. For instance, in the Stanford parser, the agent relation is the complement of a passive verb introduced by the preposition *by*, whereas in UNL it

[1] http://www.undl.org/unlsys/unl/unl2005/

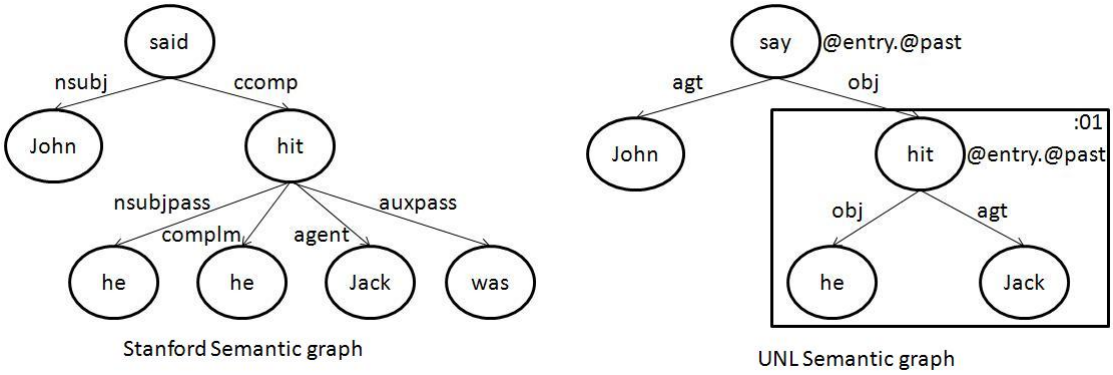

Figure 2: **UNL and Stanford semantic relation graphs for the sentence "John said that he was hit by Jack"**

	#**sentences**	#**words**
Training	12868	316508
Tuning	600	15279
Test	400	8557

Table 1: Corpus Statistics

signifies the doer of an action. Consider the following sentence:

John said that he was hit by Jack.

In this sentence, the Stanford parser produces the relation agent(hit, Jack) and nsubj(said, John) as shown in figure 2. In UNL, however, both the cases use the *agent* relation. The other distinguishing aspect of UNL is the hyper-node that represents scope. In the example sentence, the whole clause "*that he was hit by Jack*" forms the object of the verb *said*, and hence is represented in a scope. The Stanford dependency parser on the other hand represents these dependencies with the help of the clausal complement relation, which links *said* with *hit*, and uses the complementizer relation to introduce the subordinating conjunction.

The pre-dependency accuracy of the Stanford dependency parser is around 80% (Marie-Catherine et al., 2006), while the accuracy achieved by the UNL generating system is 64.89%.

6 Experiments

6.1 Setup

The corpus described in table 1 was used for the experiments.

The SRILM toolkit [2] was used to create Hindi language models using the target side of the training corpus.

Training, tuning, and decoding were performed using the *Moses* toolkit [3]. Tuning (learning the λ values discussed in section 4.1) was done using minimum error rate training (Och, 2003).

The *Stanford parser* [4] was used for parsing the English text for syntactic reordering and to generate "stanford" semantic relations.

The program for syntactic reordering used the parse trees generated by the Stanford parser, and was written in perl using the module Parse::RecDescent.

English morphological analysis was performed using *morpha* (Minnen et al., 2001), while Hindi suffix separation was done using the stemmer described in (Ananthakrishnan and Rao, 2003).

Syntactic and morphological transformations, in the models where they were employed, were applied at every phase: training, tuning, and testing.

Evaluation Criteria: Automatic evaluation was performed using BLEU and NIST on the entire test set of 400 sentences. Subjective evaluation was performed on 125 sentences from the test set.

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

- **NIST** [5]: measures the precision of n-grams. This metric is a variant of BLEU, which was

[2] http://www.speech.sri.com/projects/srilm/
[3] http://www.statmt.org/moses/
[4] http://nlp.stanford.edu/software/lex-parser.shtml
[5] www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

shown to correlate better with human judgments. Again, a higher score indicates better translation.

- **Subjective**: Human evaluators judged the fluency and adequacy, and counted the number of errors in case markers and morphology.

6.2 Results

Table 2 shows the impact of suffix and semantic factors. The models experimented with are described below:

baseline: The default settings of Moses were used for this model.

lemma + suffix: This uses the lemma and suffix factors on the source side, and the lemma and suffix/case marker on the target side. The translation steps are i) lemma to lemma and ii) suffix to suffix/case marker, and the generation step is lemma+suffix/case marker to surface form.

lemma + suffix + unl: This model uses, in addition to the factors in the lemma+suffix model, a semantic relation factor (UNL relations). The translation steps are i) lemma to lemma and ii) suffix+semantic relation to suffix/case marker, and the generation step again is lemma+suffix/case marker to surface form.

lemma + suffix + stanford: This is identical to the previous model, except that stanford dependency relations are used instead of UNL relations.

We can see a substantial improvement in scores when semantic relations are used.

Table 5 shows the impact of syntactic reordering. The surface form with distortion-based, lexicalized, and syntactic reordering were experimented with. The model with the suffix and semantic factors was used with syntactic reordering.

For subjective evaluation, sentences were judged on fluency, adequacy and the number of errors in case marking/morphology.

To judge fluency, the judges were asked to look at how well-formed the output sentence is according to Hindi grammar, without considering what the translation is supposed to convey. The five-point scale in table 3 was used for evaluation.

To judge adequacy, the judges were asked to compare each output sentence to the reference translation and judge how well the meaning conveyed by the reference was also conveyed by the output sentence. The five-point scale in table 4 was used.

Table 6 shows the average fluency and adequacy scores, and the average number of errors per sentence.

All differences are significant at the 99% level, except the difference in adequacy between the surface-syntactic model and the lemma+suffix+stanford syntactic model, which is significant at the 95% level.

7 Discussion

We can see from the results that better fluency and adequacy are achieved with the use of semantic relations. The improvement in fluency is especially noteworthy. Figure 3 shows the distribution of fluency and adequacy scores. What is worth noting is that the number of sentences at levels 4 and 5 in terms of fluency and adequacy are much higher in case of the model that uses semantic relations. That is, the use of semantic relations, in combination with syntactic reordering, produces many more sentences that are reasonably or even perfectly fluent and convey most or all of the meaning.

Table 7 shows the impact of sentence length on translation quality. We can see that with smaller sentences the improvements using syntactic reordering and semantic relations are much more pronounced. All models find long sentences difficult to handle, which contributes to bringing the mean performances closer. However, it is clear that many more useful translations are being produced due to syntactic reordering and semantic relations.

The following is an example of the kind of improvements achieved:

Input: Inland waterway is one of the most popular picnic spots in Alappuzha.

Baseline: में एक अन्तःस्थलीय जलमार्ग के सबसे प्रसिद्ध पिकनिक स्थल में जलों में दौडती है

men eka antahsthaliiya jalamaarga ke sabase prasiddha pikanika sthala men jalon men daudatii hai

gloss: in a waterway of most popular picnic spot in waters runs.

Reorder: अन्तःस्थलीय जलमार्ग आलपुया के सबसे प्रसिद्ध पिकनिक स्थल में से एक है

antahsthaliiya jalamaarga aalapuzaa ke sabase prasiddha pikanika sthala men se eka hai

Model	BLEU	NIST
Baseline (surface)	24.32	5.85
lemma + suffix	25.16	5.87
lemma + suffix + unl	27.79	6.05
lemma + suffix + stanford	28.21	5.99

Table 2: Results: The impact of suffix and semantic factors

Level	Interpretation
5	Flawless Hindi, with no grammatical errors whatsoever
4	Good Hindi, with a few minor errors in morphology
3	Non-native Hindi, with possibly a few minor grammatical errors
2	Disfluent Hindi, with most phrases correct, but ungrammatical overall
1	Incomprehensible

Table 3: Subjective Evaluation: Fluency Scale

Level	Interpretation
5	All meaning is conveyed
4	Most of the meaning is conveyed
3	Much of the meaning is conveyed
2	Little meaning is conveyed
1	None of the meaning is conveyed

Table 4: Subjective Evaluation: Adequacy Scale

Model	Reordering	BLEU	NIST
surface	distortion	24.42	5.85
surface	lexicalized	28.75	6.19
surface	syntactic	31.57	6.40
lemma + suffix + stanford	syntactic	31.49	6.34

Table 5: Results: The impact of reordering and semantic relations

Model	Reordering	Fluency	Adequacy	#errors
surface	lexicalized	2.14	2.26	2.16
surface	syntactic	2.6	2.71	1.79
lemma + suffix + stanford	syntactic	2.88	2.82	1.44

Table 6: Subjective Evaluation: The impact of reordering and semantic relations

	Baseline			Reorder			Stanford		
	F	A	E	F	A	E	F	A	E
Small (<19 words)	2.63	2.84	1.30	3.30	3.52	0.74	3.66	3.75	0.62
Medium (20-34 words)	1.92	2.00	2.23	2.32	2.43	2.05	2.62	2.46	1.74
Large (>34 words)	1.62	1.69	4.00	1.86	1.73	3.36	1.86	1.86	2.82

Table 7: Impact of sentence length (F: Fluency; A:Adequacy; E:# Errors)

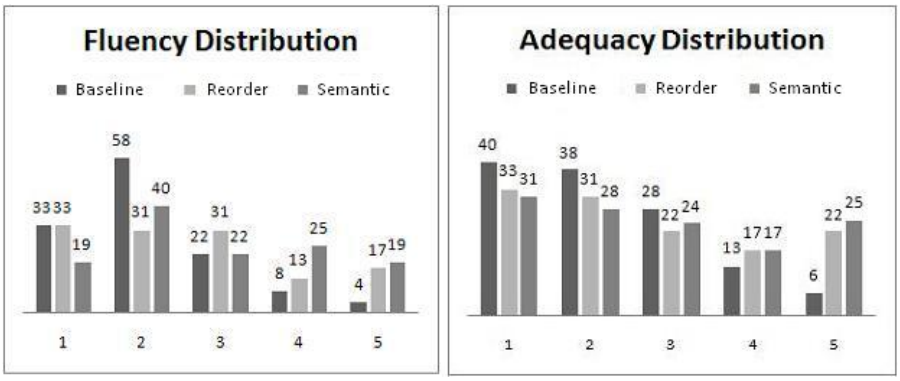

Figure 3: **Subjective evaluation: analysis**

gloss: waterway Alappuzha of most popular picnic spot of one is

Semantic: अन्तःस्थलीय जलमार्ग आलपुया के सबसे प्रसिद्ध पिकनिक स्थलों में से एक है
antahsthaliiya jalamaarga aalapuzaa ke sabase prasiddha pikanika sthalon men se eka hai
gloss: waterway Alappuzha of most popular picnic spots of one is

We can see that poor word-order makes the baseline output almost incomprehensible, while syntactic reordering solves the problem correctly. The morphology improvement using semantic relations can be seen in the correct inflection achieved in the word स्थलों (sthalon – plural oblique – *spots*), whereas the output without using semantic relations generates स्थल (sthala – singular – *spot*).

The next couple of examples illustrate how case marking improves through the use of semantic relations.

Input: Gandhi Darshan and Gandhi National Museum is across Rajghat.

Reorder: गांधी दर्शन व गांधी राष्ट्रीय संग्रहालय राजघाट में है
gaandhii darshana va gaandhii raashtriiya sangrahaalaya raajaghaata men hai

Semantic: गांधी दर्शन व गांधी राष्ट्रीय संग्रहालय राजघाट के पार है
gaandhii darshana va gaandhii raashtriiya sangrahaalaya raajaghaata ke paara hai

Here, the use of semantic relations produces the correct meaning that the locations mentioned are *across* (के पार (*ke paara*)) Rajghat, and not *in* (में (*men*)) Rajghat as suggested by the translation produced without using semantic relations.

Another common error in case marking is that two case markers are produced in successive positions in the translation, which is not possible in Hindi. The following example (a fragment) shows this error (की (*kii*) repeated) being correctly handled by using semantic relations:

Input: For varieties of migratory birds

Reorder: प्रवासी पक्षियों की की प्रकार के लिये
pravaasii pakshiyon kii kii prakaara ke liye

Semantic: प्रवासी पक्षियों की प्रकार के लिये
pravaasii pakshiyon kii prakaara ke liye

It is important to note that the gains made using syntactic reordering and semantic relations are limited by the accuracy of the parsers (see section 5). We observe that even the use of moderate quality semantic relations goes a long way in increasing the quality of translation.

8 Conclusion

We have reported in this paper the marked improvement in the output quality of Hindi translations – especially fluency – when the correspondence of English semantic relations and suffixes with Hindi case markers and inflections is used as a translation factor in English-Hindi SMT. The improvement is statistically significant. Subjective evaluation too lends ample credence to this claim. Future work consists of investigations into (i) how the internal structure of constituents can be strictly preserved and (ii) how to glue together correctly the syntactically well-formed bits and pieces of the sentences. This course of future action is suggested by the fact that smaller sentences are much more fluent in translation compared to medium length and long sentences.

References

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

Ananthakrishnan, R., Bhattacharyya, P., Hegde, J. J., Shah, R. M., and Sasikumar, M., Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation, *Proceedings of IJCNLP*, 2008.

Avramidis, E., and Koehn, P., Enriching Morphologically Poor Languages for Statistical Machine Translation, *Proceedings of ACL-08: HLT*, 2008.

Collins, M., Koehn, P., and I. Kucerova, Clause Restructuring for Statistical Machine Translation, *Proceedings of ACL*, 2005.

Imamura, K., Okuma, H., Sumita, E., Practical Approach to Syntax-based Statistical Machine Translation, *Proceedings of MT-SUMMIT X*, 2005.

Koehn, P., and Hoang, H., Factored Translation Models, *Proceedings of EMNLP*, 2007.

Marie-Catherine de Marneffe, MacCartney, B., and Manning, C., Generating Typed Dependency Parses from Phrase Structure Parses, *Proceedings of LREC*, 2006.

Marie-Catherine de Marneffe and Manning, C., Stanford Typed Dependency Manual, 2008.

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

Minnen, G., Carroll, J., and Pearce, D., Applied Morphological Processing of English, *Natural Language Engineering*, 7(3), pages 207–223, 2001.

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

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W., BLEU: a Method for Automatic Evaluation of Machine Translation, *IBM Research Report*, Thomas J. Watson Research Center, 2001.

Popovic, M., and Ney, H., Statistical Machine Translation with a Small Amount of Bilingual Training Data, *5th LREC SALTMIL Workshop on Minority Languages*, 2006.

Wang, C., Collins, M., and Koehn, P., Chinese Syntactic Reordering for Statistical Machine Translation, *Proceedings of the EMNLP-CoNLL*, 2007.