Improving Statistical Machine Translation through co-joining parts of verbal constructs in English-Hindi translation

Karunesh Kumar Arora
CDAC, Anusandhan Bhawan
C 56/1, Sector 62,
Noida, India
karunesharora@cdac.in

R Mahesh K Sinha
JSS Academy of Technical Education,
C 20/1, Sector 62,
Noida, India
sinharmk@gmail.com

Abstract

Verb plays a crucial role of specifying the action or function performed in a sentence. In translating English to morphologically richer language like Hindi, the organization and the order of verbal constructs contributes to the fluency of the language. Mere statistical methods of machine translation are not sufficient enough to consider this aspect. Identification of verb parts in a sentence is essential for its understanding and they constitute as if they are a single entity. Considering them as a single entity improves the translation of the verbal construct and thus the overall quality of the translation. The paper describes a strategy for pre-processing and for identification of verb parts in source and target language corpora. The steps taken towards reducing sparsity further helped in improving the translation results.

1 Introduction

With the availability of parallel content, increased memory and processing speed, there has been growing trend moving towards Statistical Machine Translation. Most of the phrase based machine translation systems are based on the noisy-channel based IBM models (Koehn, Och & Marcu, 2003, Zens et al., 2004). Phrases refer to a number of consecutive words that may not be a valid syntactic phrase but are learnt through the statistical alignment between two languages. English and Hindi have differing syntactical structure and pose great challenge in aligning phrases of the two languages. The former follows SVO pattern while the later adheres to the SOV pattern. Hindi being morphologically richer offers several verbal constructs governed through Tense, Aspect and Modality (TAM). The non-monotonocity between the two languages causes inferior alignment of phrases especially verbal constructs.

There have been efforts towards single tokenization of MWE parts. Ueffing and Ney, 2003 reported use of POS information for SMT to morphologically richer language. They tried to transform the source language while the approach proposed here attempts transformations on both source and target laguage sides. Recent related works use statistical measures like Mutual Information and Log Likelihood Ratio (Seretan and Wehrli, 2007) to know the degree of cohesion between constituents of a MWE. These require defining threshold value above which the extracted phrase is qualified as a MWE.

Minkov et al. (2007) utilized the rich syntactic and morphological analyzers to generate the inflections. Hindi lacks availability of robust parsers and complex morphological analyzers. The paper describes the process of identifying verbal constructs of both languages and grouping them in single units to reduce the search space. For identification of the verbal constructs, the POS information is utilized with simple combining rules to make verb phrases. This yields better alignment of verbal phrases and results in more grammatical, fluent and acceptable translations. Besides that, the data sparseness generated from chunking is

Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 95–101,
Jeju, Republic of Korea, 12 July 2012.

handled through extending the phrase table with verbal parts entries.

The paper is organized in sections, describing the phrase based SMT in brief, Hindi language and its verbal properties followed by sections describing identification of verbal constructs in English and Hindi. Further to it, corpus and pre-processing activities are detailed alongwith the experimental setup, process adopted to reduce sparcity, the translation process, observations and conclusion.

2 Overview of SMT

Candide SMT system [Brown et al., 1990], presented by the IBM researchers paved the path for statistical approach to machine translation.
In statistical machine translation, we are given a source language sentence $S = s_1^I = s_1 \ldots s_i \ldots s_I$, which is to be translated into a target language ('English') sentence $T = t_1^J = t_1 \ldots t_j \ldots t_J$. Statistical machine translation is based on a noisy channel model. It considers T to be the target of a communication channel, and its translation S to be the source of the channel. System may generate multiple translation sentences options and the problem of translation becomes identifying sentence T which fits as the best translation of the source sentence S. Hence the machine translation task becomes to recover the source from the target. So, we need to maximize P(T|S). According to the Bayes rule,

$$t^* = \arg\max_t P(t \mid s) = \arg\max_t \frac{P(s \mid t) * P(t)}{P(s)}$$

As, P(S) is constant,

$$t^* = \arg\max_t P(s \mid t) * P(t)$$

Here, P(s|t) represents Translation model and P(t) represents language model. Translation model plays the role of ensuring translation faithfulness and Language model to ensure the fluency of translated output.

3 Hindi language and its verbal properties

Indian languages are classified in four major families: Indo-Aryan (a branch of the Indo-European family), Dravidian, Austro-Asiatic (Austric), and Sino-Tibetan, with the overwhelming majority of the population speaking languages belonging to the first two families. There are 22 languages listed in eighth schedule of constitution of India. The four major families are different in their form and construction, but they share many orthographic similarities, because their scripts originate from Brahmi (Ishida, 2002).

Hindi language belongs to the Indo-Aryan language family. It is spoken in vast areas of northern India and is written in Devanagari script. In Hindi, words belonging to various grammatical categories appear in lemma and inflectional forms. Hindi Verbal constructs system is based on the TAM of the action. The Verbal costructs are formed by placement of auxiliary verbs after the main verb. The main verb that carries the lexical meaning may appear in the root or inflected form. Auxiliary verbs of the main verb denote the TAM property of the verbal construct.

Tense is a grammatization of the relations between time of some event and the refrence time. Aspect markers are semantically very sensitive and often convey subtle meanings and nuances that are not generally expressed through simple lexical words. Here we look at the two example sentences,

1. वह दिन भर बैठा रहता है

vaha din bhar baithaa rahataa hai

('He remains seated whole day').

2. वह बार-बार बैठता रहता है

vaha baar-baar baithtaa rahataa hai

('He sits frequetly')

Here, aspect marker या रह 'yaa raha' in first sentence, denotes the resultant state of the action and रह 'raha' gives perception of a longer period of time. While in a slightly modified second sentence, the aspect marker ता रह 'taa raha' gives the sense of repetition or infinity of the action and रह 'raha' gives the perception of a time spread.

The mood reflects speaker's attitude towards the action and is manifested in many ways in a language. In Hindi the moods can be of Imperative,

Subjunctive, Indefinite and definite potential, conditional and future etc. Here we look at the following three sentences.

1. तू पढ़ tu padh (‘You read’)
2. तुम पढ़ो tu padh (‘You read’)
3. आप पढ़िए tu padh (‘You read’)

All the above three sentences are imperative in nature but there is subtle difference in speaker’s attitude. The first sentence is the impolite form of expression, the second one is common form and the third sentence is the polite form of expressing the same thing.

All constituents of the verbal constructs are obligatory. Semantically TAM markers are so closely interlinked that it would be appropriate to treat them as a single entity rather than treating them sperately. Besides that, the main verb appears frequently in compound and conjunct forms in the verbal constrcuts (Singh, 2010). Compound verbs follow the pattern of verb-verb (V-V) combination while conjunct verbs are formed with either noun-verb (N-V) or adjective-verb (A-V) combinations. In V-V expressions the first verb word carries verbal stem while successive verb words play the role of auxiliary or light verbs (LV). The LVs loose their independent meaning and are used to reflect the shade of main verb. The compound and conjunct verb expressions are also referred as complex predicates (CP). The CPs are multi-word expressions (MWEs) which may be compositional or non-compositional in nature (Sinha, 2011). These should be treated as a single verbal unit to infer the intended meaning or semantics. The CP adds to the expressiveness of the expression but pose difficulty for automatic identification.

4 Identification and treatment verbal constructs

The elements of verbal constructs, if treated as individual words leave too many entries in the sentences to get aligned through statistical alignment. This makes the probability distribution unfocussed. Co-joining parts of verbal constructs reduces the sentence length and thus helps in better alignment.

4.1 English verbal constructs

The Stanford POS tagger (Kristina Toutanova et al., 2003) is used for tagging words in a sentence with their POS categories. The POS tags are based on Penn Treebank POS tagset (Mitchell et al., 1993). The verbal parts to be chunked together are identified with the help of a set of rules. Some of these rules are listed in the Table 1. As an example, the rule ‘get NP VBN’ specifies, that if Noun Phrase appears in between the word ‘get’ and VBN, this is considered as a verbal construct.

POS based Verb Chunking Rules
VBP/VBD/VBZ VBG
MD not VB
get NP VBN

Table 1: Sample rules for identiying English Verbal constructs

These rules are impletemented in the form of a Finite State Machine (FSM). The NP-phrase appearing in between the verb construct parts is identified and FSM implementation helps in achieving this. Similarly, the model auxiliaries like ‘can be’ are also co-joined with successive verbs. These simple rules help in identifying the constituents of verbal constructs. The negation markers or noun phrases that appear in between verbal constructs are moved out to reduce sparsity. Table 2 shows some English verbal constructs and how these are co-joined.

Verbal Constructs	Co-joined Verbal Constructs
is going	is_going
can not be done	not can_be_done
get the work done	get_done the work

Table 2: Sample English Verbal constructs

4.2 Identification of Hindi verbal constructs

For identifying the Hindi verbal constructs, a combination of POS tagging and presence of the TAM markers appearing as verb ending sequences are used. The POS tags are based on modified Penn Treebank POS tagset. The POS tagging identifies possible verbal parts to be chunked, while the TAM rules help in confirmation of them. Table 3 lists some of the TAM rules. Here $ indicates the presence of main verb stem.

Verbal constructs	TAM Rules
जा सकता है jaa saktaa hai	$_सकता_है $_saktaa_hai
जाने मत दो jaane mat do	मत $ने _दो mat $ne_do
खाया जा रहा होगा khaaya jaa rahaa hogaa	$या_जा_रहा_होगा $yaa_jaa_rahaa_hogaa
जा नहीं रहा है jaa nahi rahaa hai	नहीं $_रहा_है nahi $_rahaa_hai
जाता तो था jaataa to thaa	तो $ता_था to $taa_thaa

Table 3: Sample rules for identiying Hindi Verbal constructs

Table 4 shows some of the verbal constrcts and their co-joined forms after processing. The negation markers, such as, नहीं nahi ('not') and particles, such as, तो (emphatic marker) occurring in between are moved out of the verbal expressions to reduce the sparsity.

Verbal Constructs	Co-joined Verbal Constructs
जा सकता है jaa saktaa hai	जा_सकता_है jaa_saktaa_hai
जाने मत दो jaane mat do	मत जाने _दो mat jaane_do
खाया जा रहा होगा khaayaa jaa rahaa hogaa	खाया_जा_रहा_होगा khaaya_jaa_rahaa_hogaa
जा नहीं रहा है jaa nahi rahaa hai	नहीं जा_रहा_है nahi jaa_rahaa_hai
जाता तो था jaataa to thaa	तो जाता_था to jaataa_thaa

Table 4: Sample Hindi Verbal constructs

Complex Predicates are identified using the approach of Sinha (2009). Here, we make use of parallel corpus, English-Hindi dictionary of Light Verbs and TAM rules. Table below shows some sample Complex predicates in Compound and Conjuct forms and their treatment.

Compound Verbs	
Verbal Constructs	Co-joined Verbal Constructs
बैठ जा baith jaa	बैठ_जा baith_jaa
पढ़ लिया होगा padh liyaa hogaa	पढ़_लिया_होगा padh_liyaa_hogaa
कर दिया kar diyaa	कर_दिया kar_diyaa
Conjunct Verbs	
Verbal Constructs	Co-joined Verbal Constructs
परीक्षा दे parikshaa de	परीक्षा_दे parikshaa_de
बात कर रहा है baat kar rahaa hai	बात _कर_रहा_है baat_kar_rahaa_hai
बंद हो गया band ho gayaa	बंद_हो_गया band_ho_gayaa

Table 5: Sample Hindi complex predicates

5 Corpus and pre-processing

Basic Travel Expressions Corpus (BTEC) containing travel conversations is used for performing the experiments (Kikui, 2006). This contains travel expressions which are generally used when a person travels to another country and covers the utterances of potential subjects in travel situations. The expressions contained more than one sentence in single expression. These have been separated by sentence end markers (dot). Such sentences have been treated as separate sentence entities. This increased the number of independent sentences in parallel corpus. The Tables 6 and 7 list corpus statistics.

Corpus	Training	Development	Test
English:			
# sentences	19972	2343	2371
# words	153066	17806	18257
# avg words / sentence	7.7	7.6	7.7
Hindi:			
# sentences	19972	2043	2071
# words	171347	17774	17811
# avg words / sentence	8.6	8.7	8.6

Table 6: Corpus Statistics before pre-processing

Corpus	Training	Development	Test
English:			
# sentences	24056	2581	2575
# avg words / sentence	6.3	6.4	6.3
Hindi:			
# sentences	24056	2581	2575
# avg words / sentence	7.2	7.1	7.2

Table 7: Corpus Statistics after pre-processing

The average sentence length in the English corpus before pre-processing was 7.7 words per sentence and after pre-processing it came down to 6.3 words per sentence. Hindi corpus had 8.7 words per sentence and it became 7.2 words per sentence after pre-processing.

The pre-processing activity also included expanding of common abbreviated expressions e.g. I'll to 'I will' etc. This has been performed with a set of simple expansion rules. Besides that, dots appearing after titles are also replaced with hash (#), to avoid being treated them as sentence end-markers.

6 Experimental setup

For the training of the statistical models, standard word alignment GIZA++ (Och & Ney, 2003) and language modelling toolkit SRILM (Stolcke, 2002) tools were used. For translation, MOSES phrase-based SMT decoder (Koehn, 2007) has been used. For evaluation, the automatic evaluation metrics, BLEU (Papineni, 2002) was applied to the translation output.

7 Translation process

The overall process can be classified as Training and Testing processes. The training process describes the steps involved in building models. These steps include – pre-processing of training corpus, POS tagging source and target language training corpus, chunking words forming the verbal constructs, building translation and language models.

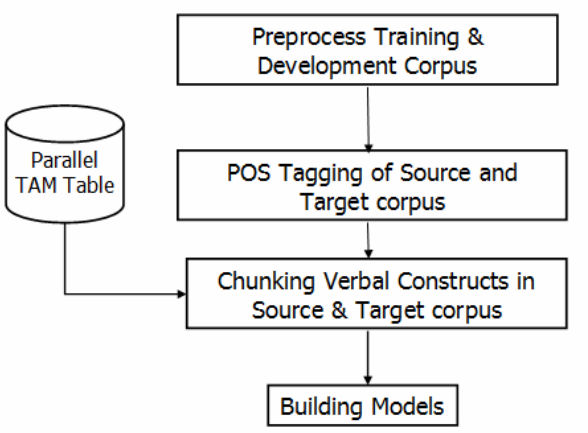

Figure 1: Training process

Testing process describes steps while translating. It involves - pre-processing of test corpus, POS tagging of test corpus, chunking the words forming the verbal constructs and searching words in the vocabulary of training models. If some words are unseen but are lexical words of verbal constructs, they are handled as described in section 8 below.

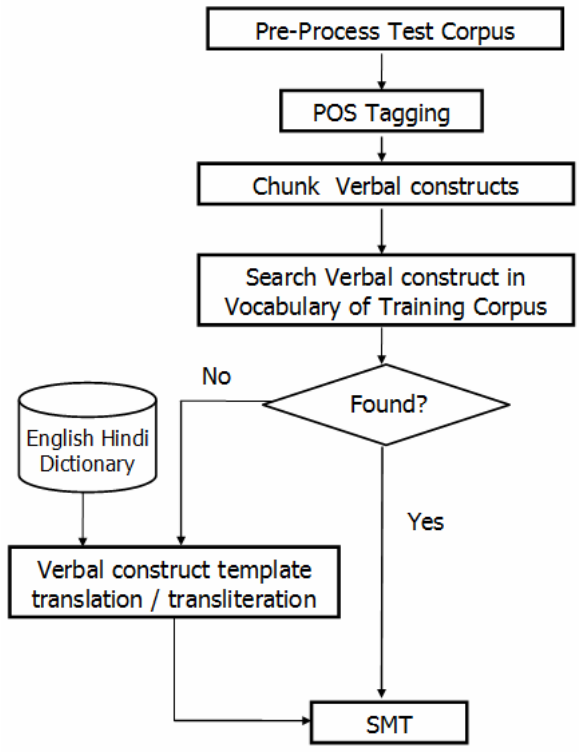

Figure 2: Testing process

8 Handling sparsity

Due to limited size of parallel corpus used for training the models, it is quite probable that some verbal constructs may appear which is unseen by the training model and is out of vocabulary (OOV). The probability of such occurrence increases due to the co-joining of words forming verbal constructs. To meet this situation, templates of different verbal constructs with their translations are used. The Table 8 shows some sample templates with their translations.

If verbal construct is OOV, it is changed to its translation template form. After that, its equivalent translation is picked up and is replaced in the sentence to be translated. As an example, if the verbal construct 'would_have_been_cleaning' is OOV. It is changed to its template form would_have_been_VBG and its respective translation VB_रहा_होगा is picked up from the translation template table. Now, with the help of English-Hindi dictionary, translation of verbal construct 'would_have_been_cleaning' in the sentence is replaced with the translated as 'साफ़_कर_रहा_होगा and is sent for final translation.

Verbal construct template	Translation template
can_VB	VB_सकता_है VB_saktaa_hai
would_have_been_VBG	VB_रहा_होगा VB_rahaa_hogaa
has_not_VBN	नहीं_VBया_है nahi_VByaa_hai

Table 8: Verbal Construct template translation

If the verb is not present in the English-Hindi dictionary too, it is translierated and 'कर' is added to it. Now, the verbal construct in the source sentence is replaced with its transliterated form before sending for translation. As an example, if word 'clean' is not found in English-Hindi dictionary, its translterated form 'क्लीन' is generated and 'कर' is added to it. The verbal construct 'would_have_been_cleaning' in the source sentence is replaced with transliterated verbal construct 'क्लीन_कर_रहा_होगा' before sending for SMT. For trnasliteration in-house statistical transliteration system is used.

9 Experiments

The experiments were carried on original, pre-processed and chunked verbal constructs based models. Table 9 below show that there is improvement in BLUE score when we pre-process the raw corpus. Better alignment is achieved due to reduced sentence length and data being in normalized form. The chunked verbal constructs corpus further improves the BLUE score. Though the BLUE score gain is marginal but on human inspection, better order and organization of Verbal constructs is observed. The table below shows the BLEU score for experiments.

Corpus	BLEU Score	Gain in BLEU score
BPP *	0.1596	
APP *	0.1672	0.0076
APP + VCC *	0.1694	0.0022

Table 9: BLEU scores for different experiments

* BPP - Before Pre-processing the corpus
* APP - After Pre-processing the corpus
* APP + VCC - After Pre-Processing corpus + Verbal Constructs Chunking

10 Conclusion and Future Work

Results show, moderate gain in BLUE score is obtained with pre-processing of the corpus. This can be attributed to better alignment due to reduced length of sentences. Marginal gain is observed with chunking of Verbal constructs, yet manual inspection show fluent translation of verbal parts.

Hindi verb forms are sensitive to gender, number and person information, which is not considered in current implementation. Work on interrogatives, prepositional phrases and other multi-word expressions, is in progress. There is scope to improve the statistical alignment using linguitic knowledge. The investigations on these are currently in progress.

Acknowledgments

We would like to thank Centre for Development of Advanced Computing (CDAC) for providing conducive environment for this work. We also would like to thank NICT, Japan for providing the English version of BTEC corpus for performing experiments. Thanks are also due to Mr Mukund Kumar Roy and Mr Pramod Kumar Gupta for setting up the software and programming efforts. Thanks are also extended to Mr VN Shukla for extending support.

References

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation, Proc. of the Human Language Technology Conference (HLT/NAACL)

Richard Zens and Hermann Ney. 2004. Improvements in Phrase-Based Statistical Machine Translation, Proc. of the Human Language Technology Conference (HLT-NAACL) , Boston, MA, pp. 257-264.

Nicola Ueffing and Hermann Ney. 2003. Using pos information for statistical machine translation into morphologically rich languages. In Proc. of the 10th Conference of the European Chapter of the ACL (EACL), Budapest, Hungary

Seretan V. and Wehrli E. 2007. Collocation translation based onalignment and parsing. Proceedings of TALN. Toulouse, France.

Einat Minkov, Krishna Toutanova and Hisami Suzuki. 2007. Generating Complex Morphology for Machine Translation, in Proc. 45th Annual Meeting of the Association for Computational Linguistics, pp 128-135.

Brown, P., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., La_erty, J. D., Mercer, R. L., and Rossin, P. 1990. A statistical approach to machine translation. Computational Linguistics, 16(2):76{85.

R. Ishida. 2002. An introduction to Indic scripts, in Proc. of the 22nd International Unicode Conference.

Singh, Suraj Bhan. 2010. A Syntactic Grammar of Hindi (first ed.), Ocean Books.

R. Mahesh K. Sinha. 2011. Stepwise Mining of Multi-Word Expressions in Hindi, ACL-HLT, Workshop on Multiword expressions, Portland, USA

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL, pp. 252-259.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank, in Computational Linguistics, Volume 19, Number 2, pp. 313--330

R. Mahesh K. Sinha. 2009. Mining Complex Predicates In Hindi Using Parallel Hindi-English Corpus, ACL-IJCNLP, Workshop on Multi Word Expression, Singapore.

G. Kikui et al. 2006. Comparative study on corpora for speech translation, IEEE Transactions on Audio, Speech and Language, vol. 14(5), pp. 1674–1682.

F. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, vol. 29(1), pp. 19–51.

A. Stolcke. 2002. SRILM -an extensible language modelling toolkit, in Proc. of ICSLP, Denver, pp. 901–904.

P. Koehn et al. 2007. Moses: Open Source Toolkit for SMT," in Proc. of the 45th ACL, Demonstration Session, Prague, Czech Republic, , pp. 177–180.

K. Papineni et al. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation, in Proc. of the 40th ACL, Philadelphia, USA, , pp. 311–318.