Statistical Machine Translation of English – Manipuri using Morpho-syntactic and Semantic Information

Thoudam Doren Singh  
Department of Computer Science and Engineering  
Jadavpur University  
Kolkata-700032, India  
thoudam.doren@gmail.com

Sivaji Bandyopadhyay  
Department of Computer Science and Engineering  
Jadavpur University  
Kolkata-700032, India  
sivaji_cse_ju@yahoo.com

Abstract

English-Manipuri language pair is one of the rarely investigated with restricted bilingual resources. The development of a factored Statistical Machine Translation (SMT) system between English as source and Manipuri, a morphologically rich language as target is reported. The role of the suffixes and dependency relations on the source side and case markers on the target side are identified as important translation factors. The morphology and dependency relations play important roles to improve the translation quality. A parallel corpus of 10350 sentences from news domain is used for training and the system is tested with 500 sentences. Using the proposed translation factors, the output of the translation quality is improved as indicated by the BLEU score and subjective evaluation.

1 Introduction

The present work reports English to Manipuri Statistical Machine Translation (SMT) system. Manipuri is a less privileged Tibeto-Burman language spoken by approximately three million people mainly in the state of Manipur in India as well as its neighboring states and in the countries of Myanmar and Bangladesh. Manipuri has little resource for NLP related research and development activities. Some of the unique features of this language are tone, the agglutinative verb morphology and predominance of aspect than tense, lack of grammatical gender, number and person. Other features are verb final word order in a sentence i.e., Subject Object Verb (SOV) order, extensive suffix with more limited prefixation. Different word classes are formed by affixation of the respective markers. In Manipuri, identification of most of the word classes and sentence types are based on the markers. All sentences, except interrogatives end with one of these mood markers, which may or may not be followed by an enclitic. Basic sentence types in Manipuri are determined through illocutionary mood markers, all of which are verbal inflectional suffixes, with the exception of the interrogatives that end with an enclitic. There are three basic forms of clausal subordination in Manipuri: subordinate clauses formed by suffixing a nominalizer to a noninflected verb; complements formed by suffixing complementizers to the nominalized clause; and adverbial clauses formed by suffixing subordinates to either nominalized clauses or complements. Two important problems in applying statistical machine translation (SMT) techniques to English-Manipuri MT are: (a) the wide syntactic divergence between the language pairs, and (b) the richer morphology and case marking of Manipuri compared to English. The first problem manifests itself in poor word-order in the output translations, while the second one leads to incorrect inflections and case marking. The output Manipuri sentences suffer badly when morphology and case markers are incorrect in this freer word order and morphologically rich language.

The parallel corpora used is in news domain which have been collected, cleaned and aligned (Singh and Bandyopadhyay, 2010) from the Sangai Express website www.thesangaiexpress.com available in both Manipuri and English. A daily basis collection was done covering the period from May 2008 to November 2008 since there is no repository.
2 Related Work

Statistical Machine Translation with scarce resources using morpho-syntactic information is discussed in (Nieβen and Ney, 2004). It introduces sentence level restructuring transformations that aim at the assimilation of word order in related sentences and exploitation of the bilingual training data by explicitly taking into account the interdependencies of related inflected forms thereby improving the translation quality. Popovic and Ney (2006) discussed SMT with a small amount of bilingual training data. Koehn and Hoang (2007) developed a framework for statistical translation models that tightly integrates additional morphological, syntactic, or semantic information. Case markers and morphology are used to address the crux of fluency in the English-Hindi SMT system (Ramanathan et al., 2009). Work on translating from rich to poor morphology using factored model is reported in (Avramidis and Koehn, 2008). In this method of enriching input, the case agreement for nouns, adjectives and articles are mainly defined by the syntactic role of each phrase. Resolution of verb conjugation is done by identifying the person of a verb and using the linguistic information tag. So far, a Manipuri to English Example Based Machine Translation system is reported in (Singh and Bandyopadhyay, 2010) on news domain. For this, POS tagging, morphological analysis, NER and chunking are applied on the parallel corpus for phrase level alignment. Chunks are aligned using a dynamic programming “edit-distance style” alignment algorithm. The translation process initially looks for an exact match in the parallel example base and returns the retrieved target output. Otherwise, the maximal match source sentence is identified. For word level mismatch, the unmatched words in the input are translated from the lexicon or transliterated. Unmatched phrases are looked into the phrase level parallel example base; the target phrase translations are identified and then recombined with the retrieved output.

2.1 Factored Model of Translation

Using factored approach, a tighter integration of linguistic information into the translation model is done for two reasons:

- Translation models that operate on more general representations, such as lemma instead of surface forms of words, can draw on richer statistics and overcome the data sparseness problem caused by limited training data

- Many aspects of translation can be best explained at a morphological, syntactic or semantic level. Having such information available to the translation model allows the direct modeling of these aspects. For instance, reordering at the sentence level is mostly driven by general syntactic principles, local agreement constraints that show up in morphology, etc.

2.2 Combination of Components in Factored Model

Factored translation model is the combination of several components including language model, reordering model, translation steps and generation steps in a log-linear model:

\[ p(e|f) = \frac{1}{Z} \sum_{i=1}^{n} \lambda_i h_i(e, f) \]  

(1)

Z is a normalization constant that is ignored in practice. To compute the probability of a translation \( e \) given an input sentence \( f \), we have to evaluate each feature function \( h_i \). The feature weight \( \lambda_i \) in the log linear model is determined by using minimum error rate training method (Och, 2003).

For a translation step component, each feature function \( h_t \) is defined over the phrase pairs \( (f_j, e_j) \) given a scoring function \( \tau \):

\[ h_t(e, f) = \sum_j \tau(f_j, e_j) \]  

(2)

For the generation step component, each feature function \( h_g \) given a scoring function \( \gamma \) is defined over the output words \( e_k \) only:

\[ h_g(e, f) = \sum_k \gamma(e_k) \]  

(3)

\[ \text{http://www.statmt.org/moses/?n=Moses.FactoredModels} \]

\[ \text{http://www.statmt.org/moses/?n=Moses.FactoredModels} \]
2.3 Morphology

In Manipuri, words are formed by three processes called affixation, derivation and compounding. The majority of the roots found in the language are bound and the affixes are the determining factor of the word class in the language. In this agglutinative language the number of verbal suffixes is larger than that of nominal suffixes. Works on Manipuri morphology are found in (Singh and Bandyopadhyay, 2006) and (Singh and Bandyopadhyay, 2008).

In this language, a verb must minimally consist of a verb root and an inflectional suffix. There are two derivational prefixes: an attributive prefix which derives adjectives from verbs and a nominalizing prefix which derives nouns from verbs. The inflectional morphology of the verb consists of eight illocutionary mood markers: the nonhypothetical –ই (i); the assertive –এ (e); the optative –ক (ke); the imperative –ও (u); the prohibitive –নু (nu); the solicitive –নি (ni); the suplicative –সি (si); and the permissive –সানু (sani). Only one inflectional morpheme may appear with a given verb root and the inflectional marker will appear after all derivational morphemes and before all enclitics.

A noun may be optionally affixed by derivational morphemes indicating gender, number and quantity. A noun may have one of the 5 semantic roles: agent (instigator of action), actor (doer of action), patient, reciprocal/goal, and theme. Actor and theme roles are not indicated morphologically, while all other semantic roles are indicated by an enclitic. For pragmatic effect, semantic role markers can be omitted or replaced by enclitics which mark contrastiveness or definiteness. Peripheral arguments may be suffixed by enclitics indicating ablative, genitive or associative case. Further, a noun may be prefixed by a pronominal prefix which indicates its possessor. The following examples explain morphological richness of this language.

(a) এই (me) রাম (Ram) লুঃজি (loves)

\[ Ey-di \ Ram-na \ Nungsi \]

Ram loves me (over all possibilities).

(b) এই (me) রাম (Ram) লুঃজি (loves)

\[ Ey-na \ Ram-da \ Nungsi \]

I (as opposed to you) love only Ram.

Words in Manipuri consist of stems or bound roots with suffixes (from one to ten suffixes), prefixes (only one per word) and/or enclitics.

(1) (only Ram) (love)

I (as opposed to you) love only Ram.

(c) পরা (Ram) লুঃজি (loves)

\[ Ey-hak-khakta \ Ram-si \ Nungsi \]

(I am the only one) (this man Ram) (loves)

I am the only one who loves this man Ram.

The identification of subject and object in both the freer word order sentences are done by the suffixes –ান (na) and –ু (du) as given by the examples (d) and (e). The case markers convey the right meaning during translation though the most acceptable order is SOV.

Thus, in order to produce a good translation output all the morphological forms of the word and its translations should be available in the training data and every word has to appear with every possible suffixes. This will require a large training data. By learning the general rules of morphology, the amount of training data could be reduced. Separating lemma and suffix allows the system to learn more about the different possible word formations.

Nouns in Manipuri are inflected by gender and number. For example, কুড়ি (nupa – man) becomes নুপাঙ্গি (nupaa-sing – men) in plural and feminine noun কুড়ী (nupii – woman) becomes নুপাঙ্গী (nupi-sing – women).

3 Stanford Dependency Parser

The dependency relations used in the experiment are generated by the Stanford dependency parser (Marie-Catherine de Marneffe and Manning, 2008). This parser uses 55 relations to express the
dependencies among the various words in a sentence. The dependencies are all binary relations: a grammatical relation holds between a governor and a dependent. 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, infinitival modifiers and other relations like coordination, conjunct, expletive and punctuation.

Figure 2. Semantic relation graph of the sentence “Sources said that Tom was shot by police” generated by Stanford Parser

Let us consider an example “Sources said that Tom was shot by police”. Stanford parser produces the dependency relations, nsubj(said, sources) and agent (shot, police) . Thus, sources\textunderscore nsubj and police\textunderscore agent are the factors used. “Tom was shot by police” forms the object of the verb “said”. The Stanford parser represents these dependencies with the help of a clausal complement relation which links “said” with “shot” and uses the complementizer relation to introduce the subordination conjunction. Figure 2 shows the semantic relation graph of the sentence “Sources said that Tom was shot by police”.

4 Lexicalized and Syntactic Reordering

4.1 Lexicalized Reordering

This method adds new features to the log-linear framework, in order to determine the order of the target phrases at decoding.

During extraction of phrases from the training corpora, the orientation of each occurrence is also extracted and the probability distribution is estimated for addition to the log-linear framework. The three different orientations are defined as:

- monotone: a word alignment point to the top left exists
- swap: an alignment point to the top right exists.
- discontinuous: no alignment points to the top left or top right exists.

Figure 1 gives the possible orientation of phrases defined over the lexicalized reordering. Finally, during decoding, automatically inferred reordering models are used to score each hypothesis according to the orientation of the used phrases.

4.2 Syntactic Reordering

This is a preprocessing step applied to the input sentences. The basic difference of Manipuri phrase order compared to English is handled by reordering the input sentence following the rule (Rao et al., 2000):

\[
SS_mV \rightarrow C'_mS'_mS'O'_mO'V'_mV'
\]

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

The program for syntactic reordering uses the parse trees generated by Stanford parser and
plying a handful of reordering rules written using perl module Parse::RecDescent. By doing this, the SVO order of English is changed to SOV order, and post modifiers are converted to pre-modifiers.

There could be two reasons why the syntactic reordering approach improves over the baseline phrase-based SMT system (Wang et al., 2007). One obvious benefit is that the word order of the transformed source sentence is much closer to that of the target sentence, which reduces the reliance on the distortion model to perform reordering during decoding. Another potential benefit is that the alignment between the two sides will be of higher quality because of fewer “distortions” between the source and the target, so that the resulting phrase table of the reordered system would be better. However, a counter argument is that the reordering is very error prone, so that the added noise in the reordered data actually hurts the alignments and hence the phrase tables.

5 Factorization approach

Manipuri case markers are decided by semantic relation and aspect information of English. Figure 3 shows the translation factors used between English and Manipuri.

(a) Tomba drives the car.

Tomba-na car-du thou-i
(Tomba) (the car) (drives)
Tomba|empty|nsubj drive|s|empty the|empty|det car|empty|dobj

A subject requires a case marker in a clause with a perfective form (such as -نا (na)
Such as, suffix + semantic relation → case marker s|empty + empty|dobj → ـنا (na)

(b) Birds are flying.

ucheksing payri
(birds are) (flying)

Thus, English-Manipuri factorization consists of

| a lemma to lemma translation factor [i.e., Bird → उचेक (ucheck)] |
| a suffix + semantic relation → suffix [i.e., s + nsubj → फ़ि (sing)] |
| a lemma + suffix → surface form generation factor [i.e., उचेक (ucheck) + फ़ि (sing) → उचेकफ़ि (ucheksing)] |

Figure 3. English to Manipuri translation factors

6 Experimental Setup

A number of experiments have been carried out using factored translation framework and incorporating linguistic information. The toolkits used in the experiment are:

- Stanford Dependency Parser6 was used to (i) generate the semantic relations and (ii) syntactic reordering of the input sentences using Parse::RecDescent module.
- Moses7 toolkit (Koehn, 2007) was used for training with GIZA++8, decoding and minimum error rate training (Och, 2003) for tuning.
- SRILM9 toolkit (Stolcke, 2002) was used to build language model of the target language using the target side of the training corpus.

6 http://nlp.stanford.edu/software/lex-parser.shtml
7 http://www.statmt.org/moses/
8 http://www.floch.com/GIZA++.html
9 http://www.speech.sri.com/projects/srilm
English morphological analyzer *morpha* (Minnen et al., 2001) was used and the stemmer from Manipuri Morphological analyzer (Singh and Bandyopadhyay, 2006) was used for the Manipuri side.

7 Evaluation

The evaluation of the machine translation systems developed in the present work is done in two approaches using automatic scoring with reference translation and subjective evaluation as discussed in (Ramanathan et al., 2009).

**Evaluation Metrics:**

- **NIST** (Doddington, 2002): A high score means a better translation by measuring the precision of n-gram.

- **BLEU** (Papineni et al, 2002): This metric gives the precision of n-gram with respect to the reference translation but with a brevity penalty.

Table 1 shows the corpus statistics used in the experiment. The corpus is annotated with the proposed factors.

<table>
<thead>
<tr>
<th></th>
<th>No of sentences</th>
<th>No of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>10350</td>
<td>296728</td>
</tr>
<tr>
<td>Development</td>
<td>600</td>
<td>16520</td>
</tr>
<tr>
<td>Test</td>
<td>500</td>
<td>15204</td>
</tr>
</tbody>
</table>

**Table 1:** Training, development and testing corpus statistics

The following models are developed for the experiment.

**Baseline:**

The model is developed using the default setting values in MOSES.

**Lemma + Suffix:**

It uses lemma and suffix factors on the source side, lemma and suffix on the target side for lemma to lemma and suffix to suffix translations with generation step of lemma plus suffix to surface form.

**Lemma + Suffix + Semantic Relation:**

Lemma, suffix and semantic relations are used on the source side. The translation steps are (a) lemma to lemma (b) suffix + semantic relation to suffix and generation step is lemma + suffix to surface form. Table 2 shows the BLEU and NIST scores of the system using these factors.

Table 3 shows the BLEU and NIST scores of handling of syntactic divergence using two different approaches.

**Sample input and output:**

(a) **Input:** Going to school is obligatory for students.

**Baseline output:**

```
 faculty _ chatpa shatra - sing - gi touda ya draba mathouni .
```

**Reorder output:**

```
 shatra school chatpa touda yadraba mathouni gloss: students to the school going is duty.
```

(b) **Input:** Krishna has a flute in his hand.

```
 Krishna _ gi khut - ta toudri ama lei .
```

---

ftp://ftp.informatics.susx.ac.uk/pub/users/johnca/morph.tar.gz
One of the main aspects required for the fluency of a sentence is agreement. Certain words have to match in gender, case, number, person etc. within a sentence. The rules of agreement are language dependent and are closely linked to the morphological structure of language. Subjective evaluations on 100 sentences have been performed for fluency and adequacy by two judges. The fluency measures how well formed the sentences are at the output and adequacy measures the closeness of the output sentence with the reference translation. The Table 4 and Table 5 show the adequacy and fluency scales used for evaluation and Table 6 shows the scores of the evaluation.

**Table 4: Adequacy scale**

<table>
<thead>
<tr>
<th>Level</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Full meaning is conveyed</td>
</tr>
<tr>
<td>3</td>
<td>Most of the meaning is conveyed</td>
</tr>
<tr>
<td>2</td>
<td>Poor meaning is conveyed</td>
</tr>
<tr>
<td>1</td>
<td>No meaning is conveyed</td>
</tr>
</tbody>
</table>

**Table 5: Fluency scale**

<table>
<thead>
<tr>
<th>Sentence length</th>
<th>Fluency</th>
<th>Adequacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>&lt;=15 words</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>&gt;15 words</td>
<td>1.49</td>
</tr>
<tr>
<td>Reordered</td>
<td>&lt;=15 words</td>
<td>2.58</td>
</tr>
<tr>
<td></td>
<td>&gt;15 words</td>
<td>1.82</td>
</tr>
<tr>
<td>Semantic</td>
<td>&lt;=15 words</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td>&gt;15 words</td>
<td>1.94</td>
</tr>
</tbody>
</table>

**Table 6: Scale of Fluency and Adequacy on sentence length basis**

The factored approach using the proposed factors show improved fluency and adequacy at the Manipuri output as shown in the Table 6. Using the Stanford generated relations shows an improvement in terms of fluency and adequacy for shorter sentences than the longer ones.

**Input:** Khamba pushed the stone with a lever.

**Reordered:** Khamba nung jamfat adu illi

**Semantic:** Khambana nung adu jamfatna illi

**Table 6: Scale of Fluency and Adequacy on sentence length basis**

By the use of semantic relation, (na) is attached to (Khamba), which makes the meaning “by Khamba” instead of just “Khamba”.

**Input:** Suddenly the woman burst into tears.

**Reordered:** Nupi thuna pirang-ga kappi

**Semantic:** Athubada nupidu kaplammi

**Table 6: Scale of Fluency and Adequacy on sentence length basis**

Here, in this example, the (nupi) is suffixed by the (du), to produce nupidi “the woman” instead of just nupi “woman”.

**9 Conclusion**

A framework for English–Manipuri Statistical Machine translation using factored model is experimented with a goal to improve the translation output and reduce the amount of training data. The
output of the translation is improved by incorporating morphological information and semantic relations by tighter integration. The systems are evaluated using automatic scoring techniques BLEU and NIST. The subjective evaluation of the systems is done to find out the fluency and adequacy. The system performs well for the suffix and semantic relations over suffixes only giving an improved output. Shorter sentences showed greater gains from using semantic relation information than larger sentences, in terms of fluency and adequacy. The improvement is statistically significant. Incorporation of more language specific information such as parts of speech (POS) is identified as future task.

References


Popovic, M., and Ney, H. 2006. Statistical Machine Translation with a small amount of bilingual training data, 5th LREC SALTMIL Workshop on Minority Languages


Singh, Thoudam D., and Bandyopadhyay, S. 2008. Morphology Driven Manipuri POS Tagger, In proceedings IJCNLP-08 Workshop on NLPLPL, pages 91-98, Hyderabad, India


