Tackling Close Cousins: Experiences In Developing Statistical Machine Translation Systems For Marathi And Hindi

Raj Dabre
CFILT
IIT Bombay
prajdabre
@gmail.com

Jyotesh Choudhari
CFILT
IIT Bombay
jyoteshrc
@gmail.com

Pushpak Bhattacharyya
CFILT
IIT Bombay
pushpakbh
@gmail.com

Abstract

In this paper we present our experiences in building Statistical Machine Translation (SMT) systems for the Indian Language pair Marathi and Hindi, which are close cousins. We briefly point out the similarities and differences between the two languages, stressing on the phenomenon of Krudantas (Verb Groups) translation, which is something Rule based systems are not able to do well. Marathi, being a language with agglutinative suffixes, poses a challenge due to lack of coverage of all word forms in the corpus; to remedy which, we explored Factored SMT, that incorporate linguistic analyses in a variety of ways. We evaluate our systems and through error analyses, show that even with small size corpora we can get substantial improvement of approximately 10-15% in translation quality, over the baseline, just by incorporating morphological analysis. We also indirectly evaluate our SMT systems by analysing and reporting the improvement in the quality of translations of a Marathi to Hindi Rule Based system (Sampark) by injecting SMT translations of Krudantas. We believe that our work will help researchers working with limited corpora on similar morphologically rich language pairs and relatable phenomena to develop quality MT systems.

1 Introduction

Marathi [1] and Hindi [2] are Indian languages ranking fourth and first [3] with respect to the number of speakers. Marathi has close to 72 million speakers whereas Hindi has close to 400 million speakers. Both Marathi and Hindi belong to the family of 'Indo-European languages' i.e. both have originated from Sanskrit and thus have some phonological, morphological and syntactic features in common.

1.1 Comparison of Marathi and Hindi

Marathi uses agglutinative, inflectional and analytic forms. It displays abundant amount of both derivational (wherein attachment of suffixes to a word form changes its grammatical category) and inflectional morphology. Hindi shares these properties of Marathi except that it is not agglutinative in nature. Both languages follow the S-O-V word order. Both languages have dative verbs. In both languages, when the agent in the sentence is in nominative case, the verb agrees with it in person, number and gender; however, when it is not in nominative case, the verb does not agree with it. These languages differ most in the participial and reported speech constructions.

1.2 Agglutination, Participials and Reported Speech constructions

The major factor in translating from Marathi (Mr) to Hindi (Hi) is handling the transfer of agglutinative morphemes. In the reverse case it is the generation of appropriate agglutinative morphological forms. Consider the translation of the Marathi word "माझ्याबरोबरच्यानेही" {majhya-barobar-chya-ne-hii} {the one with me also (nominative)} which is "मेरे साथ वाले ने भी", a whole phrase in Hindi. Marathi suffixes become Hindi post positions. Typically, in languages that have agglutinative suffixes there are millions to billions of possible surface forms and when all surface forms are not present in the parallel corpora, translation suffers from data sparsity. Also,

[1] http://en.wikipedia.org/wiki/Marathi_language
[2] http://en.wikipedia.org/wiki/Hindi
[3] http://en.wikipedia.org/wiki/Eighth_Schedule_to_the_Constitution_of_India

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 11–19,
Goa, India. December 2014.

the translation of morphemes does not merely involve independent dictionary substitution but require looking at neighboring morphemes.

An important aspect of our work involved handling of participial forms known as Krudantas and Akhyatas (Bhosale et al, 2011; Bapat et al. 2010) which are derivatives of verbs. Consider: "मी धावल्यानंतर असलेला व्यायाम करत आहे" {mi dhava-lya-nantar asa-le-la vyayam kara-ta aahe} {I am doing the exercises that come after running} in which "धावल्यानंतर" and "असलेला" (both nominal forms and are 2 consecutive Krudantas) and "करत आहे" (Verb group indicating tense, aspect and mood of action) are participial constructions. The last auxiliary verb in the verb group, "आहे" dictates the tense of the sentence; present in this case.

Consider the translation of "धावल्याने" {dhava-lya-ne} {by running (nominative)}, a Krudanta in nominal form, which in Hindi is "भागने से" {bhaagne se} or "भागने की वजह से" {bhaagne ki wajah se}. When the suffixes are to be translated, "ल्या" {lya} is translated as {ने} {ne} and "ने" {ne} is translated as से {se}. Note that "ने" {ne} is also used as a nominative case marker and will be translated as "ने" {ne} in the case of the Krudanta "धावणाऱ्याने" {dhava-narya-ne} {the runner (nominative)} which in Hindi is दौड़ने वाले ने {daudne wale ne}. "ल्या" {lya} also has an alternate translation as "हुए" {hue} {became (dead for e.g.)} in the case of "मेलेल्याला" {melelyala} {the dead person (accusative)} which in Hindi is: "मरे हुए को" {mare hue ko}. This is sufficient to indicate that translating a verb group by using rules and bilingual dictionaries is a difficult and an involved process.

A construction of reported speech contains two sentences a sentential complement and a matrix sentence. Hindi, typically, positions the sentential complement (underlined) after the matrix sentence but Marathi can place this either before or after. The sentence "He says that he comes home at 8" is written in Hindi as: "वह कहता है की <u>वह आठ बजे घर आता है</u>" {vaha kahta hai ki vaha aath baje ghar aata hai} but in Marathi as "<u>तो घरी आठ वाजता येतो</u> असे तो म्हणतो" {to ghari aath vaj-ta ye-to ase to mhana-to} or "तो म्हणतो की <u>तो आठ वाजता घरी येतो</u>" {to mhana-to ki to aath vaj-ta ghari ye-to}. Due to Krudantas a Marathi sentence can have many possible Hindi equivalents.

All these examples serve to indicate that translation between Marathi and Hindi, inspite of their closeness, is quite challenging. Due to space constraints we do not elaborate further but those interested may look at the books of M.K. Damle (1970) and Dhongde et al. (2009). We now describe the various experiments conducted and SMT systems developed.

1.3 Related Work

Nair et al. (2013) developed a basic phrase based SMT system and compared it against a Rule based system, Sampark, for Marathi to Hindi translation. Their work lacked any kind of linguistic processing leading to only simple sentences being translated well. Bapat et al. (2011) had explored the impact of handling Krudanta forms via morphological analysis during translation in Sampark by using rules. We obtained the Sampark system and its source code so that we could convert it into a Hybrid system, by SMT phrase translation injection, to get an indirect evaluation of the quality of our SMT systems. Dabre et al. (2012) had worked on improving the coverage and quality for their Marathi morphological analyzer which we exploit in the development of our SMT systems. The remainder of the paper is dedicated to the experiments conducted and evaluation.

2 SMT Systems and Experiments

We first describe the corpora used and then the SMT systems. The evaluation is in the following

Corpora	#Lines	#words
ILCI–Health–Marathi-Hindi	25000	Mr-85681 Hi -43102
ILCI–Tourism– Marathi-Hindi	25000	
DIT–Health – Marathi-Hindi	20000	
DIT–Tourism– Marathi-Hindi	20000	
Wiki+News Web- Marathi	1968907	896430
Wiki+News Web-Hindi	1538429	558206

Table 1: Corpora details

section.

2.1 Corpora details

Table 1 below describes the sources and sizes of the corpora which come from 2 major projects namely ILCI (Indian Languages Corpora Initiative) and DIT (Department of Information Technology). We also crawled the web for monolingual corpora which we used for language modelling.

The crawl of Wikipedia (Wiki) by itself provided around 500000-600000 monolingual sentences. There are many Marathi and Hindi news websites amongst which we crawled only the prominent ones for our corpora. It must be noted that the parallel corpus, which is undergoing revisions, was not of high quality and contains duplicate sentence pairs.

2.2 Training and Technical details

In order to perform training we used IBM models (Brown et al., 1993) implementation in GIZA++ for alignment and Moses (Koehn et al., 2007, 2003) for phrase table extraction and decoding. In order to obtain factors for Marathi we used the Morphological analyzer and Part of speech tagger developed under the ILMT (Indian Language Machine Translation) project. For Hindi we used a freely available tool[4] that does Morphological analysis and POS tagging simultaneously. All non-factored systems took around 15-20 minutes of training time whereas inclusion of factors increased the time to around 30-50 minutes.

We tried to tune our systems but often saw that the resulting translations were of poorer quality and thus the evaluations presented later are on our non-tuned systems. All phrase tables were binarized and provided as services using the Moses webserver daemon.

2.3 Marathi-Hindi Systems

Below are details of the development of the various systems for Marathi to Hindi translation. For each system we describe the processing steps, if any, of the corpora and give assumptions and reasons for doing so. We also indicate the pros and cons of performing these steps, most of which will be indicated by examples in the evaluation section. For factored systems we mention factors as <Factor-1 | Factor-2 | … | Factor-n>, decoding steps as <Source Factor combinations → Target Factor Combinations> and generation steps as <Target Factors → Target Surface Form>.

2.3.1 Baseline system

We trained a basic phrase based system using the full parallel corpus for training and the Hindi monolingual corpus for language modeling. This did not have substantial coverage of all word forms for Marathi and motivated us to utilize the Marathi morphological analyzer.

2.3.2 Suffix Split system

We performed morphological splitting on the Marathi corpus and replaced the surface word with its root form followed by suffixes with spaces in between. The original source sentence to target sentence length ratio was quite low (Marathi having lesser words per sentence due to agglutination) which was observed to become almost equal to 1 after morphological splitting. Keeping the words in the root form led to the loss of gender (G), number (N) and person (P) in many cases. Also in case of tense determining, auxiliary verb forms of "असणे" {asne} {to be} the root word form loses the tense information. Despite this, the morphological splitting resulted in a massive increase in translation quality (see Evaluation section). Words that were not translated in their agglutinative form get properly translated due to reduction in data sparsity.

2.3.3 Factored system – Suffixes as factors

As an initial experiment into factored models we processed the Marathi side of the corpus to have 4 factors: **<Surface Form | Root Word | Suffixes | POS Tag>**. The agglutinative suffixes for a word were grouped (separated by an underscore) and treated as a single factor. Not all suffix combinations exist in a small corpus and this grouping does not lower data sparsity much. The Hindi side is also processed to have 7 factors: **< Surface Form | Root Word | Gender | Number | Person | POS Tag | Case>**. Here Case means direct or oblique to indicate whether an inflection exists or not.

Initially we tried training a simple model in which our decoding step was: **<Marathi Factors → Hindi Factors>** which is followed by a generation step: **<Hindi Factors → Hindi Surface Word>**. But this ended up being the same as the baseline system. Since both root words and suffixes in Marathi map to words in Hindi we specified 2 additional decoding steps: **<Root Word + POS → Root Word + POS + Gender + Num-**

[4] http://sivareddy.in/downloads

ber + Person + Case> and **<Suffix → Root Word + POS + Gender + Number + Person + Case>**. The previous step we kept as a back-off. We assumed that the aligner would map Marathi root words and suffixes to Hindi root words and postpositions respectively. Since existence of suffixes are indicative of inflection in Marathi. We did not consider GNP and case information as factors. But on investigation of the phrase tables we saw that the phrase pairs recorded were of poor quality. The decoder effectively disregarded the new decoding steps and hence we refrained from pursuing this way of treating suffixes as factors.

2.3.4 Factored system – Suffixes separated from root

We realized that separating the suffixes from the roots was the best approach and thus augmented the Marathi corpus with GNP and case information. We first separated the suffixes from the root words by spaces and then added factors to the root word. The factored representation was **<Root Word | Gender | Number | Person | POS Tag | Case>**. The split suffixes were represented as **<Suffix | Gender | Number | Person | PSP | Case>**. The GNP's were copied over from the root word they were attached to and the POS was kept as PSP (postposition). The case was kept as 'd' (direct) if it was the last suffix and 'o' (oblique) otherwise. For verbs that indicated the tense (which don't have many morphological variations) of the sentence we kept them in their surface form. Finally we used our large monolingual corpora to train a generation model which would combine all factors on the Hindi side to generate the surface form.

2.3.5 Hybridized Rule based System (Sampark) – Injecting SMT into an RBMT system

The architecture of Sampark is given in Nair et al. (2013) and Bhosale et al. (2011). It is a transfer based rule based MT system and works in 3 phases: Analysis, Lexical transfer, Generation. The analysis phases generate the morphological analyses which our SMT systems can use. The original lexical transfer algorithm performed lookup in dictionaries to get root word and suffix translations. We modified the lexical transfer algorithm such that the verb groups would be translated by our best SMT system (2.3.2 in this case) by making translation requests to a Moses server daemon via a system call. This helped improve Verb Group translation quality (see Evaluation). The flow of translation is:

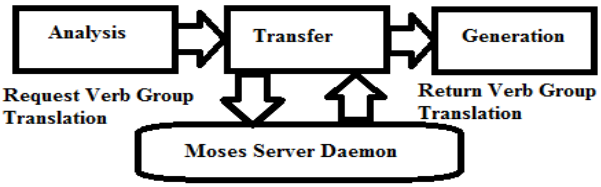

Figure 1: Modified Sampark

2.4 Hindi-Marathi Systems

The translation systems for Hindi to Marathi are given below. The terminology used is the same as before.

2.4.1 Baseline System

As before, we trained a baseline system by using the Marathi monolingual corpus for the language model. An interesting observation was that the quality of translation from Hindi to Marathi was better than the quality of translation in the reverse direction on the same corpus which furthers the belief that translation is not a bidirectional phenomenon.

2.4.2 Factored System

The factored corpus used in section 2.3.3 was reused for training this system. The decoding steps involved: **<Hindi lemma to Marathi Lemma>** and **<Hindi POS Tag + Suffix → Marathi POS Tag + suffix>**. An additional back-off decoding step was: **<Hindi Surface Word → Marathi Lemma + POS Tag + Suffix>**. Thereafter the generation step combines the Marathi factors to the surface form. However the quality of translations was not much better than the baseline system and upon investigation we observed that the decoder effectively used the back-off translation step. Once again we were faced with the same situation as in section 2.3.3. We realized that SMT should only take care of morpheme transfer as in sections 2.3.2 and 2.3.4. This led to the system below.

2.4.3 Marathi Suffix splitting + Generation

We used the processed corpus mentioned in section 2.3.2 with the modification in section 2.3.4, namely, keeping the tense determining verbs in their surface form. It is quite evident that there would be a loss of GNP but our objective at that time was to achieve morphological generation correctly. The resultant phrase based system translates a Hindi sentence into a suffix split Marathi sentence. We then wrote a simple module

which would combine the morphemes into a surface form. With this in mind we turned to our Marathi monolingual corpus, morphologically analyzed it and generated a HashMap which contained morphemes and surface forms as key value pairs. One such entry would be: "घर समोर" {ghar samor} → "घरासमोर" {gharaa-samor} {in front of the house}. Since the monolingual corpus was quite large (89.6 k unique words) we assumed that all commonly used morphological forms would be contained in it. The suffix combination method used is:

1. Consider that the morphemes of the translated sentence are in an array "**split_morph**". Let the hashmap containing morphemes to surface word mapping be called "**morpheme_to_word_map**"
2. For index = 1 to length(**split_morph**):
   a. Check the longest sequence of morphemes from current position which is present as a key in **morpheme_to_word_map**.
   a. Retrieve the surface form for this morpheme sequence.
   b. If no longest sequence can be found from current position then continue from next position.

It will be seen later that this method is quite naïve and quite drastically lowers the quality of translations.

3 Evaluation and Results

3.1 Evaluation methodology

We considered BLEU (Papineni et al.,2002) as the standard evaluation criteria which morphologically rich and free word order languages like Marathi and Hindi are not best evaluated by and hence we also performed Adequacy (meaning transfer) and Fluency (grammatical correctness) analysis as mentioned in Bhosale et al. (2011). A total of 100 sentences were translated and scores were calculated. A number of these were survival sentences. For adequacy (Ad) and fluency (Fl) the sentences were given scores from 1 to 5 (5 for best). Sentences of score 3 and above are considered acceptable. S_i is (#sentences with score i).

Adequacy / Fluency = 100 * (S_5 + 0.8 * S_4 + 0.6 * S_3) / (#Total sentences)

This gives the percentage of total sentences that the user understands and has acceptable grammar.

3.2 Results

Table 2 below gives the scores for the various systems. The best results are highlighted. In case of Marathi to Hindi translation the suffix split systems perform the best. For Hindi-Marathi the baseline and factored system do reasonably well. The suffix split + generation system was the worst.

In order to check the improvement of Krudanta (Verb Group) translation by SMT injection into the RBMT system (Sampark) we constructed a set of 52 sentences (some of which were present in the above set of 100 sentences) which contained a variety of Krudanta types. Since Krudantas can have multiple translations (see Section 1.2) BLEU is not reliable and hence we evaluated them using Adequacy and Fluency only (Table 3). It was observed that there was a substantial increment in translation quality of Sampark especially due to high quality Krudanta translations.

System	BLEU	Adequacy	Fluency
Marathi-Hindi (Baseline)	24.46	44.6	55.8
Marathi-Hindi (Suffix split)	**29.68**	**63.2**	**59.4**
Marathi-Hindi (Suffix as factor)	18.42	47.0	47.6
Marathi-Hindi (Suffix split + factored)	**30.35**	**61.8**	**66.0**
Hindi-Marathi (Baseline)	**23.93**	**77.0**	**74.4**
Hindi-Marathi (Factored)	20.06	72.6	72.2
Hindi-Marathi (Suffix split + Marathi generation)	8.4	45.0	37.8

Table 3: Results of Evaluation

System	Adequacy	Fluency
Hindi-Marathi (Sampark)	46.0	40.0
Hindi-Marathi (Hybrid Sampark – SMT injection)	**61.0**	**62.0**

Table 2: Improvement in translation quality by SMT injection

3.3 Example translations and Analysis

Consider the following examples of translations which serve to indicate that the scores above are indeed valid.

3.3.1 Marathi to Hindi (SMT systems only)

- **Input:** राज्यांतील शहरांलगत बेकायदा बांधकामांचा प्रश्न गंभीर होत चालला आहे .

  {rajyan-til shahran-lagat bekayda baandhkaaman-cha prashna gambhir hot challa ahe}

  Meaning: The situation of illegal constructions near the cities in the states is becoming severe.

- **Expected translation:** राज्यों में/के शहरों के पास अवैध/बेकायदा निर्माणों की समस्या गंभीर होती जा रही है ।

  {rajya mein/ke shaharon ke paas avaidh/bekayda nirman ki samasya gambhir hoti jaa rahi hai}

- **Baseline translation:** राज्य में शहरांलगत बेकायदा बांधकामांचा समस्या गंभीर होती जा रही हैं ।

  {rajya mein shaharan-lagat bekayda baandhkaaman-cha samasya gambhir hoti jaa rahi hai}

- **Suffix split translation:** राज्य के शहर से सटे बेकायदा निर्माण की समस्या गंभीर होती जा रही है ।

  {rajya ke shahar se sate bekayda nirman ki samasya gambhir hoti jaa rahi hai}

- **Suffix as factor translation:** शहरांलगत बेकायदा राज्यों के निर्माण का मसला गंभीर होती जा रही हैं ।

  {shaharan-lagat bekayda rajyon ke nirman ka masla gambhir hoti jaa rahi hai}

- **Suffix split + factored translation:** राज्यों के शहरों से बेकायदा निर्माण की समस्या गंभीर होती जा रहा है ।

  {rajyon ke shaharon ke bekayda nirman ki samasya gambhir hoti jaa rahi hai}

- **RBMT system Sampark (Non Hybrid version):** राज्यों के शहरों के पास गैरकानूनी बनावटों की समस्या गंभीर गुजरते गया है ।

  {rajyon ke shaharon ke paas gairkanuni banavaton ki samasya gambhir gujarte gaya hai}

- **RBMT system Sampark (Hybrid version):** राज्यों के शहरों के पास गैरकानूनी बनावटों की समस्या गंभीर होता जा रहा है ।

  {rajyon ke shaharon ke paas gairkanuni banavaton ki samasya gambhir hota jaa raha hai}

Analysis: As expected, the baseline translation is quite bad because words like "शहरांलगत" {adjoining cities} and "बांधकामांचा" {of constructions} are not translated. But the suffix split translation is almost correct except that it does not generate the plural forms "राज्यों" {states}, "शहरों" {cities} and "निर्माणों" {constructions}. This was due to the loss of GNP information due to root forms. Here "राज्यातील" {in/of the state} has 2 translations of its suffix "तील" which can be "में" {in} (generated by baseline) के {of} (generated by suffix split). Also "शहरांलगत" which was translated as "शहर से सटे" (by suffix split system) which means "clinging to the city" is also acceptable but not as natural sounding as "शहरों के पास" {near the cities}. The factored system which had the suffix as a factor, fared rather poorly; as we had anticipated. But the suffix split factored system gave a near about perfect translation except for the plural form "निर्माणों". It only missed a word "सटे" {clinging}. Also one must note that the verb group "होत चालला आहे" {is becoming} is perfectly translated in all cases. In general, however, the suffix split systems gave better translations than others.

Comparing these to the RBMT systems translations, one must note that both the Hybrid and Non-Hybrid versions of Sampark do well in handling GNP inflections since all the morphological information is retained. Moreover it can be seen that the translations of nouns and their suffixes, "शहरांलगत" as "शहरों के पास", are much more natural than those of the SMT systems.

However, the translation of the verb group "होत चालला आहे" is translated as "गुजरते गया है" which is incorrect in the non-hybrid version. This is because sense of "होत" in this sentence is that of "happening" or "becoming", but "गुजरते" is a translation of another sense of that word which is indicative of "passage of time". "होत" whose root word is "होणे" is an extremely polysemous word having more than 10 senses. RBMT systems typically make mistakes in translating the proper sense of the word unless they have very high quality WSD (word sense disambiguation) modules which in the case of Sampark are not that good when disambiguating verbs. The hybrid version gives a near about perfect translation of "होत चालला आहे" as "होता जा रहा है" (masculine inflection) instead of "होती जा रही है" (feminine inflecion). This is because the inflection of the verb group depends on the gender of the word "समस्या". But when performing SMT injection, only the phrase "होत चालला आहे" is translated and without the context "प्रश्न" the SMT system is unable to perform gender agreement thereby affecting the fluency of the translation.

3.3.2 Hindi to Marathi

- **Input:** राज्यों के शहरों के पास बेकायदा निर्माणों की समस्या गंभीर होती जा रही है ।

  {rajyon ke shaharon ke paas bekayda nirmanno ki samasya gambhir hoti jaa rahi hai}

- **Baseline translation:** राज्याच्या शहराला लागून असलेल्या बेकायदा निर्माणों समस्या गंभीर होत चालली आहे.

  {rajya-chya shahara-la laga-un aslelya bekayda nirmanon samasya gambhir hoat chala-li ahe}

- **Factored translation:** राज्यातील शहरांमध्ये जवळ बेकायदा निर्माणों समस्या गंभीर होत चालली आहे.

  {rajya-til shaharan-madhye javal bekayda nirmanon samasya gambhir hoat chala-li ahe}

- **Suffix Split Translation:** राज्य च्या शहर च्या जवळ बेकायदा निर्माणों ची समस्या गंभीर हो त आहे.

  {rajya chya shahar chya javal bekayda nirmanon chi samasya gambhir ho ta ahe}

- **Marathi Generation:** राज्यांच्या शहराच्या जवळ बेकायदा निर्माणों ची समस्या गंभीर होत आहे.

  {rajyan-chya shahara-chya javal bekayda nirmanon chi samasya gambhir hoat ahe}

Analysis: "निर्माणों" {constructions} was not translated by any of the systems. The factored system managed to handle some inflections. In the suffix split translation the Marathi morphemes were translated properly and the required surface forms were generated. The suffix split + generation system dropped the word "चालली" {going} which is acceptable. In this case the baseline translation was better than the others.

3.4 Discussion

It is quite evident that translation from Marathi to Hindi is rather easier than the reverse which involves morphological generation. Although the suffix split + factored Marathi-Hindi system did pretty well on the overall test sentences, it suffered from factor data sparsity (Tamchyna et al. 2013) and missed on generating proper inflected forms. This prevented it from outperforming the non-factored suffix split system by a large margin. More study is needed, on properly utilizing factors, to get high quality translation.

The impact of SMT injection into the RBMT system Sampark is quite interesting from the point of view of translating Krudantas. It is clear that SMT manages to capture the structure of verb group translations along with translating proper senses of the verbs. The lexical transfer algorithm of the non-hybrid Sampark would translate the words in the verb group independently, only relying on the word sense indicated by the WSD module. Since SMT works at the phrasal level the translations are better. Another interesting observation is that short distance agreements are very good for SMT and Sampark. The lack of a dependency parser leads to problems in long distance agreements. At the current moment, the injection is very naïve and further study into performing intelligent injection

needs to be done. In general, Sampark translates non-verbal class words better than the SMT systems do.

The case of Hindi-Marathi is the major challenge. The performance of the suffix split + generation method was very disappointing. But upon doing analysis we saw that the Hindi to Marathi morphemes translations were quite correct in many cases. There were some things wrong with our generation methodology. Firstly, on studying the morphemes to surface from mapping table we saw that there were many erroneous entries due to unknown lexicon words incorrectly morphologically analyzed. Secondly, sometimes suffix morphemes and root word morphemes are similar and get incorrectly joined to other words. This along with our decision to do longest morpheme sequence matching resulted in incorrect (over) generation.

"क्यों जा रहा हैं?" {kyo jaa raha hai} {why are you going} is translated as "काजा त आहे?" instead of "का जात आहे?" {kaa jaa-ta ahe}. The morpheme translation is "का जा त आहे?" which is correct. Here "काजा" is a village in India, not in the morph lexicon and incorrectly split as "का जा" and recorded in the mapping table leading to poor translation. The naïve algorithm resembles the morpheme concatenation method in Durgar et al. (2006). A proper morphological generator needs to be used.

4 Conclusion and Future work

We have presented our work and experiences in developing SMT systems for Marathi and Hindi. We have described the corpora used and the details of training the systems in necessary detail. We also have evaluated the systems and given analyses of sample translations. The Marathi to Hindi translation by SMT is more or less at a high quality owing to suffix splitting of Marathi whose morphemes map to appropriate words/post positions in Hindi. Further study into proper utilization of factors will be undertaken to improve quality. The improvement of the translation quality of the RBMT system, Sampark, by performing SMT injection of Krudanta translations is another testimony to the qualitative performance of the SMT systems. The reverse translation direction is rather difficult due to morphological generation for Marathi. Our current method is quite weak and relying on lookup is clearly not good. Currently we are working on developing a good morphological generator by reverse engineering the analyzer of Dabre et al. (2012). Their morphology grammar rules will be used. Indian Languages are all close cousins and Dravidian languages are similar to Marathi in respect to morphology. Our experiences should be applicable in the development of high quality SMT systems for these languages thereby ensuring effective sharing of knowledge written in any Indian language.

Reference

Sreelekha Nair, Raj Dabre, Pushpak Bhattacharyya. 2013. *Comparison of SMT and RBMT: The requirement of Hybridization for Marathi – Hindi MT*. ICON 2013, New Delhi, December, 2013.

Ganesh Bhosale, Subodh Kembhavi, Archana Amberkar, Supriya Mhatre, Lata Popale and Pushpak Bhattacharyya. 2011. *Processing of Participle* (*Krudanta*) *in Marathi*. ICON 2011, Chennai, December, 2011.

M.K. Damale. 1970. *Shastriya Marathii Vyaakarana*. Deshmukh and Company, Pune, India.

Dhongde and Wali. 2009. *Marathi*. John Benjamins Publishing Company, Amsterdam, Netherlands.

Peter E Brown, Stephen A. Della Pietra. Vincent J. Della Pietra, and Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimationn*. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Raj Dabre, Archana Amberkar and Pushpak Bhattacharyya. 2012. *Morphological Analyzer forAffix Stacking Languages: A Case Study of Marathi*, COLING 2012.

Mugdha Bapat, Harshada Gune and Pushpak Bhattacharyya. 2010. *A Paradigm-Based Finite State Morphological Analyzer for Marathi.* Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pages 26–34, the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010

Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. *Statistical phrase-based translation*. Association for Computational Linguistics 2003.

Philipp Koehn and Hieu Hoang. 2007. *Factored translation models*. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 868-876, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

İlknur Durgar El-Kahlout and Kemal Oflazer. 2006. Initial explorations in English to Turkish statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation* (StatMT '06). Association for Computational Linguistics, Stroudsburg, PA, USA, 7-14.

Aleš Tamchyna and Ondřej Bojar. 2013. No free lunch in factored phrase-based machine translation. In Proceedings *of the 14th international conference on Computational Linguistics and Intelligent Text Processing - Volume 2* (CICLing'13), Alexander Gelbukh (Ed.), Vol. 2. Springer-Verlag, Berlin, Heidelberg, 210-223. DOI=10.1007/978-3-642-37256-8_18