Solving Data Sparsity by Morphology Injection in Factored SMT

Sreelekha S
IIT Bombay
India

Piyush Dungarwal
IIT Bombay
India

Pushpak Bhattacharyya
IIT Bombay
India

Malathi D
SRM University
India

{*sreelekha,piyushdd,pb*}*@cse.iitb.ac.in*

{*malathi.d*}*@ktr.srmuniv.ac.in*

Abstract

SMT approaches face the problem of data sparsity while translating into a morphologically rich language. It is very unlikely for a parallel corpus to contain all morphological forms of words. We propose a solution to generate these unseen morphological forms and inject them into original training corpora. We observe that morphology injection improves the quality of translation in terms of both adequacy and fluency. We verify this with the experiments on two morphologically rich languages: Hindi and Marathi, while translating from English.

1 Introduction

Statistical translation models which translate into a morphologically rich language face two challenging tasks: 1. correct choice of inflection, and 2. data sparsity. To understand the severity of these two problems, consider an example of verb morphology in Hindi[1]. Hindi verbs are inflected based on gender, number, person, tense, aspect, and modality. Gender can be masculine or non-masculine (2). Number can be singular or plural (2). Person can be first, second or third (3). Tense can be present or non-present (2). Aspect can be simple, progressive or perfect (3). Modality can be due to shall, will, can, etc. (9). Thus, for a single root verb in Hindi, we have in total 648 (2*2*3*2*3*9) inflected forms. It is very unlikely for a Hindi corpus to have all inflected forms of each verb. Also, lesser the corpus size of morphologically richer language, more severe the problem of sparsity.

Using factored models helps in solving the problem of correct inflectional choice. But solving the sparsity problem is more challenging task. In this paper, we propose a simple and effective solution of enriching the input corpora with various morphological forms of words. We perform experiments with factored models (Koehn and Hoang, 2007) as well as unfactored models, i.e., phrase-based models (Koehn, Och and Marcu, 2003) while translating from English to Hindi and English to Marathi. Results show that morphology injection performs very well in order to solve the sparsity problem.

The rest of the paper is organized as follows: We present related work in Section 2. Then, we study the basics of factored translation models in Section 3. We also describe a general factored model for handling morphology. Then, we discuss the sparsity problem and the morphology generation, in general in Section 4, and in context of Hindi and Marathi in Section 5. Section 6 draws conclusion and points to future work.

2 Related work

Substantial volume of work has been done in the field of translation into morphologically rich languages. The source language can be enriched with grammatical features (Avramidis and Koehn, 2008) or standard translation model can be appended with *synthetic phrases* (Chahuneau et al., 2013). Also, previous work has been done in order to solve the verb morphology in English to Hindi SMT (Gandhe et al., 2011).

Although past work focuses on studying complexity (Tamchyna and Bojar, 2013) and solving morphology using factored translation models (Ramanathan et al., 2009), the problem of data sparsity is not addressed, to the best of our knowledge.

3 Factored translation models

Factored translation models can be seen as the combination of several components (language

[1]Hindi and Marathi are morphologically rich languages compared to English. They are widely spoken in Indian subcontinent.

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 95–99,
Trivandrum, India. December 2015.

model, reordering model, translation steps, generation steps). These components define one or more feature functions that are combined in a log-linear model (Koehn and Hoang, 2007):

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

Each h_i is a feature function for a component of the translation, and the λ_i values are weights for the feature functions. Z is a normalization constant.

Factored models treat each word in the corpus as vector of tokens. These tokens can provide extra linguistic information about the word. This information can be used to generate more accurate inflections compared to other unfactored models.

3.1 Factored model for handling morphology

Note that our goal is to solve the sparsity problem while translating to morphologically rich languages. Figure 1 shows a basic factored model for translation from morphologically poor language to rich language. On the source side we have: Surface word, root word, and set of factors S that affect the inflection of the word on the target side. On the target side, we have: Surface word, root word, and suffix (can be any inflection). The model has single translation (T0) and generation step (G0).

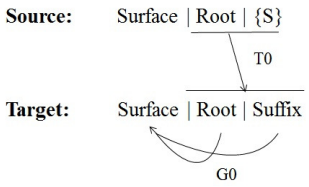

Figure 1: Factored model setup to handle inflections

4 Problem of Data Sparsity and Morphology Generation

A simple and effective solution to the sparsity problem is to generate the unseen morphological forms of words and inject them into original model. Note that, we also need to generate factors that affect the inflections of the newly generated morphological forms. For example, for the factored model described in Section 3.1, we need to generate new *Source root*|{*S*} → *Target surface word*|*Target root*|*suffix* pairs.

But then the question remains: *How do we generate these new morphological forms?* Here is the general procedure that can be adopted while translating from language X to Y:

1. We identify the factor set (S) that affects the inflections of words in language Y
2. We learn which inflection the target word will have for a particular combination of factors in S on the source side
3. We generate the surface word from the root word and inflection in language Y

In Section 5, we discuss the problem of data sparsity and morphology generation in detail, in context of Hindi and Marathi, while translating from English..

5 Morphology Generation

Hindi and Marathi are morphologically rich languages compared to English. They show morphological inflections on nouns and verbs. Before studying actual generation of various word forms, we present the factored model setup that is used for our experiments.

5.1 Factored model setup

Noun inflections in Hindi are affected by the number and case of the noun only (Singh et al., 2010). So, in this case, the set S, as in Section 3.1, consists of number and case. Number can be *singular* or *plural* and case can be *direct* or *oblique*. Example of factors and mapping steps are shown in Figure 2.

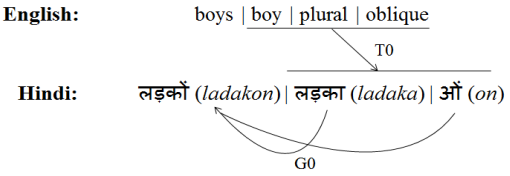

Figure 2: Factored model setup to handle nominal inflections

Similarly, verb inflections in Hindi are affected by gender, number, person, tense, aspect, and modality (Singh and Sarma, 2011). As it is difficult to extract gender from English verbs, we do not use it as a factor on English side. We just replicate English verbs for each gender inflection on

Hindi side. Hence, set S, as in Section 3.1, consists of number, person, tense, aspect, and modality only.

We build similar factored model for Marathi nouns and verbs. But, Marathi is morphologically more complex than Hindi, as multiple suffixes can be attached with Marathi root nouns and root verbs. But, still we can generate one-suffix word forms of Marathi nouns and verbs.

5.2 Building word-form dictionary

Word-form dictionary is a list consisting of all inflected forms of root words. Figure 3 shows a pipeline to generate new morphological forms for an English-Hindi/Marathi word pair. The pipeline needs the information about suffix classification based on the factors that affect those inflections. With the help of such classification, we create a list of the form: *Source root|Source S factors* $\rightarrow$ *Target root|Target suffix* by extracting source-target noun/verb pairs from the training corpus.

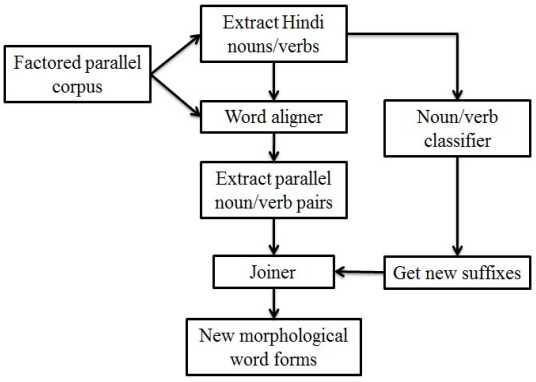

Figure 3: Pipeline to generate new morphological forms for an English-Hindi/Marathi noun/verb pair

Next step is to create a *Target surface word* from the new list of *Target root|Target suffix*. We build a joiner (reverse morphological) tool in target language, which merges root word and suffix to give target surface word. The joiner uses the ending of the root noun/verb and the class to which the suffix belongs as features. The final word-form list, thus generated, is augmented to original training data. Table 1 shows four morphological forms of *boy-लड़का (ladakaa)* noun pair. Similarly different morphological forms are created for Hindi verbs and Marathi nouns and verbs.

We also learn a factored model which combines factors on both nouns and verbs. We build word-form dictionaries separately for nouns and verbs and augment training data with both. Note that, factor normalization[2] on each word is required to maintain same number of factors.

English root|Number|Case	Hindi surface|Root|Suffix
boy|singular|direct	लड़का (ladakaa)|लड़का (ladakaa)|null
boy|singular|oblique	लड़के (ladake)|लड़का (ladakaa)|ए (e)
boy|plural|direct	लड़के (ladake)|लड़का (ladakaa)|ए (e)
boy|plural|oblique	लड़कों (ladakon)|लड़का (ladakaa)|ओं (on)

Table 1: New morphological forms of boy-लड़का (ladakaa) noun pair

We also create a word-form dictionary for phrase-based model. We follow the same procedure as described above, but we remove all factors from source and target words except the surface form.

5.3 Experiments and Evaluation

We performed experiments on ILCI (Indian Languages Corpora Initiative) En-Hi and En-Mr data set. Domain of the corpus is health and tourism. We used 44,586 sentence pairs for training and 2,974 sentence pairs for testing. Word-form dictionary was created using the Hindi and Marathi word lexicon. It consisted of 182,544 noun forms and 310,392 verb forms of Hindi and 9,869 noun forms and 101,621 verb forms of Marathi.

Moses toolkit[3] was used for training and decoding. Language model was trained on the target side corpus with *IRSTLM*[4].

For our experiments, We compared the translation outputs of: Phrase-based (unfactored) model (**Phr**), basic factored model (**Fact**) as in Section 5.1, phrase-based model trained on the corpus augmented with word-form dictionary (**Phr'**), and factored model trained on the corpus augmented with the word-form dictionary (**Fact'**).

We use *Stanford POS tagger*[5] (Toutanova et al., 2003) and *Stanford's typed dependencies* (De Marneffe et al., 2008) to extract the factors that affect the inflections (number, person, tense, etc.) from English sentence.

5.3.1 Automatic evaluation

The translation systems were evaluated by BLEU score (Papineni et al., 2002). Also, as the reduc-

[2]To use *null* when particular word can not have that factor

[3]http://www.statmt.org/moses/

[4]https://hlt.fbk.eu/technologies/irstlm-irst-language-modelling-toolkit

[5]http://nlp.stanford.edu/software/tagger.shtml

Morph. problem	Model	BLEU		# OOV		% OOV reduction		Adequacy		Fluency	
		En-Hi	En-Mr	En-Hi	En-Mr	En-Hi	En-Mr	En-Hi	En-Mr	En-Hi	En-Mr
Noun	*Fact*	22.30	8.84	2,030	1,399	**14.33**	**2.14**	3.62	2.52	3.65	2.20
	Fact'	22.41	8.85	1,739	1,369			3.73	2.55	3.66	2.23
Verb	*Fact*	23.23	9.02	1,141	1,772	**14.11**	**4.35**	3.85	2.67	3.86	2.26
	Fact'	23.26	9.02	980	1,695			3.91	2.73	3.91	2.30
Noun & Verb	*Fact*	20.93	7.55	2,193	3,137	**14.87**	**5.56**	3.89	2.69	3.92	2.28
	Fact'	21.03	7.58	1,867	2,963			4.17	2.77	4.06	2.34
Noun & Verb	*Phr*	22.87	9.27	813	1,572	**7.38**	**2.27**	4.07	2.70	3.90	2.24
	Phr'	22.89	9.28	753	1,537			4.12	2.72	3.92	2.25

Table 2: Automatic and Subjective evaluation of the translation systems

tion in number of unknowns in the translation output indicates better handling of data sparsity, we counted the number of OOV words in the translation outputs. Table 2 shows the evaluation scores and numbers.

From the evaluation scores, it is very evident that *Fact'/Phr'* outperforms *Fact/Phr* while solving any morphology problem in both Hindi and Marathi. But, improvements in *En-Mr* systems are very low. This is due to the small size of word-form dictionaries that are used for injection. % reduction in OOV shows that, morphology injection is more effective with factored models than with the phrase-based model. Also, improvements shown by BLEU are less compared to % reduction in OOV.

Why BLEU improvement is low?

One possible reason is ambiguity in lexical choice. Word-form dictionary may have word forms of multiple Hindi or Marathi root words for a single parallel English root word. Hence, many times the translation of the English word may not match the reference used for BLEU evaluation, even though it may be very similar in the meaning. Table 3 shows the number of OOVs that are actually translated after morphology injection and number of translated OOVs that match with the reference. We see that matches with the reference are very less compared to the actual number of OOVs translated. Thus, BLEU score cannot truly reflect the usefulness of morphology injection.

Morph. problem	En-Hi		En-Mr	
	#OOV translated	#Ref. Matches	#OOV translated	#Ref. Matches
Noun (*fact*)	291	105	30	5
Verb (*fact*)	436	77	77	0
Noun & Verb (*fact*)	601	137	174	20
Noun & Verb (*phr*)	124	21	35	7

Table 3: Counts of total OOVs translated after morphology injection and the matches with the reference used for BLEU evaluation

5.3.2 Subjective evaluation

As BLEU evaluation with only single reference is not a true measure of evaluating our method, we also performed human evaluation. We found out that *Fact'/Phr'* systems really have better outputs compared to *Fact/Phr* systems, in terms of both, adequacy and fluency.

For evaluation, randomly chosen 50 translation outputs from each system were manually given adequacy and fluency scores. The scores were given on the scale of 1 to 5 going from worst to best, respectively. Table 2 shows average scores for each system. We observe upto **7%** improvement in adequacy and upto **3%** improvement in fluency.

6 Conclusion

SMT approaches suffer due to data sparsity while translating into a morphologically rich language. We solve this problem by enriching the original data with the missing morphological forms of words. Morphology injection performs very well and improves the translation quality. We observe huge reduction in number of OOVs and improvement in adequacy and fluency of the translation outputs. This method is more effective when used with factored models than the phrase-based models.

Though the approach of solving data sparsity seems simple, the morphology generation may be painful for target languages which are morphologically too complex. A possible future work is to generalize the approach of morphology generation and verify the effectiveness of morphology injection on morphologically complex languages.

Acknowledgments

This work is funded by Department of Science and Technology, Govt. of India under Women Scientist Scheme- WOS-A with the project code-SR/WOS-A/ET-1075/2014.

References

Avramidis, Eleftherios, and Philipp Koehn. 2008. *Enriching Morphologically Poor Languages for Statistical Machine Translation*. ACL.

Chahuneau, Victor, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. *Translating into Morphologically Rich Languages with Synthetic Phrases*. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

De Marneffe, Marie-Catherine, and Christopher D. Manning. 2008. *Stanford typed dependencies manual*. URL http://nlp. stanford. edu/software/dependencies manual. pdf (2008).

Gandhe, Ankur, Rashmi Gangadharaiah, Karthik Visweswariah, and Ananthakrishnan Ramanathan. 2011. *Handling verb phrase morphology in highly inflected Indian languages for Machine Translation*. IJCNLP.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2007. *Statistical phrase-based translation*. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics.

Koehn, Philipp and Hieu Hoang. 2007. *Factored Translation Models*. EMNLP-CoNLL.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation*. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics.

Ramanathan, Ananthakrishnan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 2009. *Case markers and morphology: addressing the crux of the fluency problem in English-Hindi SMT*. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics.

Singh, Smriti, Vaijayanthi M. Sarma, and Stefan Müller. 2010. *Hindi Noun Inflection and Distributed Morphology*. Université Paris Diderot, Paris 7, France. Stefan Müller (Editor) CSLI Publications http://csli-publications. stanford. edu (2006): 307.

Singh, Smriti, Vaijayanthi M. Sarma. 2011. *Verbal Inflection in Hindi: A Distributed Morphology Approach*. PACLIC.

Tamchyna, Aleš, and Ondřej Bojar. 2013. *No free lunch in factored phrase-based machine translation*. Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg. 210-223.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer 2003. *Feature-rich part-of-speech tagging with a cyclic dependency network*. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics.