Morphology based Factored Statistical Machine Translation (F-SMT) system from English to Tamil

MORPHOLOGY BASED FACTORED STATISTICAL MACHINE TRANSLATION (F-SMT) SYSTEM FROM ENGLISH TO TAMIL

M. Anand Kumar, S. Rajendran , Soman K.P

Amrita Vishwa Vidyapeetham,

Coimbatore

Abstract

This paper presents a novel preprocessing methodology in factorized Statistical Machine Translation system from English to Tamil language. SMT system considers the translation problem as a machine learning problem. Statistical machine translation system for morphologically rich languages is a challenging task. Moreover it is very complex for the different word order language pair. So a simple SMT alone would not give good result for English to Tamil, which differs in morphological structure and word order. A simple SMT system performs only at the lexical level mapping. Because of the highly rich morphological structure of Tamil language, a simple lexical mapping alone will suffer a lacuna in collecting all the morphological and syntactic information from the English language. The proposed SMT system is based on factored translation models. The factored SMT uses machine learning techniques to automatically learn translation patterns from factored corpora. Using the learned model FSMT predicts the output factors for the given input factors. Using the Tamil morphological generator the factored output is synthesized.

1. Introduction

Statistical approach to machine translation learns translation patterns directly from training sentences and generalized them to handle new sentences. When translating from simple morphological language to the rich morphological language, the SMT baseline system will not generate the word forms that are not present in the training corpora. For training the SMT system, both monolingual and bilingual sentence-aligned parallel corpora of significant size are essential. The corpus size decides the accuracy of machine translation. The limited availability of parallel corpora for Tamil language and high inflectional variation increases a data sparseness problem for phrase-based SMT. To reduce the data sparseness, the words are split into lemma and their inflected forms based on their part of speech. Factored translation models [Koehn and Hoang, 2007] allow the integration of the linguistic information into a phrase-based translation model. These linguistical features are treated as separate tokens during the factored training process.

$$P(T|E) = P(T)\ P(E|T)\ /\ P(E)$$

$$T\hat{} = \underset{T}{\operatorname{argmax}}\ P(T)\ P(E|T)$$

SMT works on the above equation. Where T represents Tamil language and E represents English language. We have to find the best Tamil translation sentence (Tˆ) using P(T) and P(E|T), Where P(T) is given by the Language model and P(E|T) is given by the translation model.

2. *Factored SMT for Tamil*

Tamil language is morphologically rich language with free Word order of SOV pattern. English language is morphologically simple with the word order of SVO pattern. The baseline SMT would not perform well for the languages with different word order and disparate morphological structure. For resolving this, we go for factored SMT system (F-SMT). A factored model, which is a subtype of SMT [Koehn and Hoang, 2007], will allow multiple levels of representation of the word from the most specific level to more general levels of analysis such as lemma, part-of-speech and morphological features. A preprocessing module is externally attached to the SMT system for Factored SMT.

The preprocessing module for source language includes three stages, which are reordering, factorization and compounding. In reordering stage the source language sentence is syntactically reordered according to the Tamil language syntax using reordering rules. After reordering, the English words are factored into lemma and other morphological features. A compounding process for English language is then followed, in which the various function words are removed from the reordered sentence and attached as a morphological factor to the corresponding content word. This reduces the length of English sentence. Now the representation of the source syntax is closely related to the target language syntax. This decreases the complexity in alignment, which is also a key problem in SMT from English to Tamil language.

Parallel corpora and monolingual corpora are used to train the statistical translation models. Parallel corpora contains factored English sentences (using Stanford parser) along with its factored Tamil translated sentences (using Tamil POS Tagger [V Dhanalakshmi et.al, 2009] and Morphological analyzer [M Anand kumar et.al,2009]. Factorized monolingual corpus is used in the Language model.

The parsed source language is reordered according to the target language structure using the syntax based reordering system. A compounding process for English language is then followed, in which the various function words are removed from the reordered sentence and attached as a morphological factor to the corresponding content word. This reduces the length of English sentence. Now the representation of the source syntax is closely related to the target language syntax. This decreases the complexity in alignment, which is also a key problem in SMT from English to Tamil language.

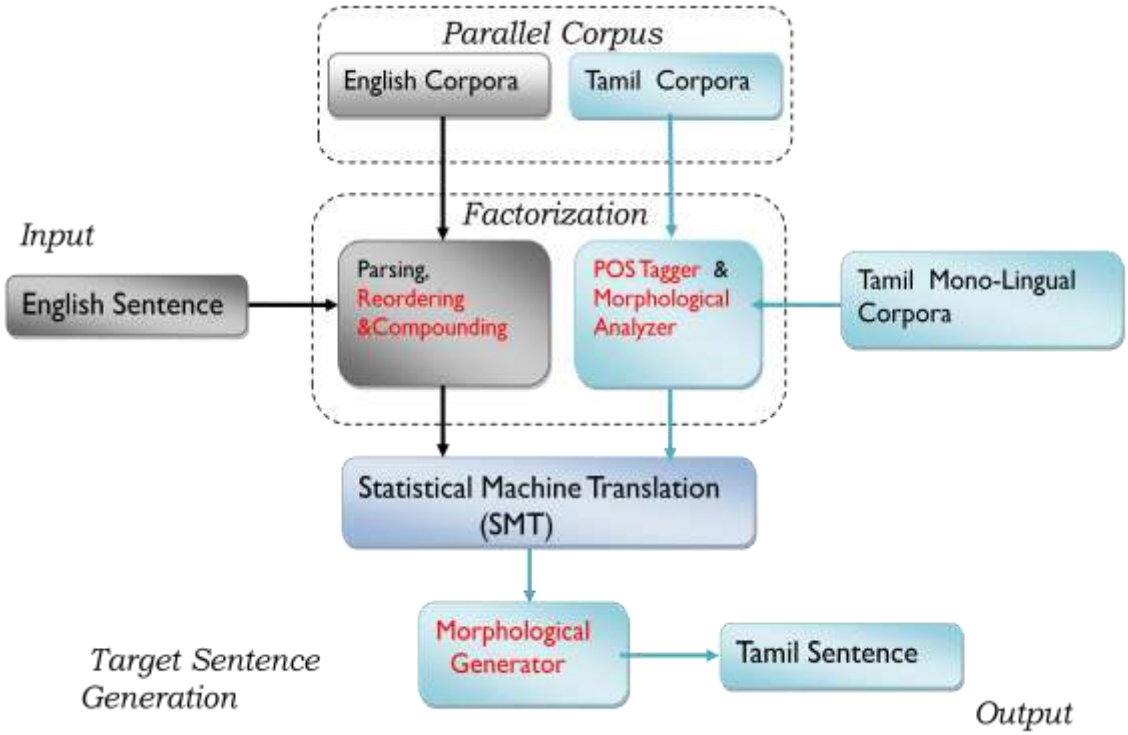

Figure.1 Architecture of the prototype factored SMT system from English to Tamil

The factored SMT system's output is post processed, where the Tamil Morphological generator is pipelined to generate the target sentence. Figure.1 shows the architecture of the prototype factored SMT system from English to Tamil.

3. Morphological models for Tamil language

Morphological models for target language Tamil are used in preprocessing as well as post processing stage. In preprocessing, Tamil POS tagger and Morphological analyzer are used to factorize the Tamil parallel corpus and monolingual corpus. Morphological generator is used in the post processing stage to generate the Tamil words from Factored SMT output.

4. Tamil POS tagger

Parts of speech (POS) tagging means labeling grammatical classes i.e. assigning parts of speech tags to each and every word of the given input sentence. POS tagging for Tamil is done using SVM based machine learning tool [V Dhanalakshmi et.al, 2009], which make the task simple and efficient. The SVM

Tool is used for training the tagged sentences and tagging the untagged sentences. In this method, one requires Part of speech tagged corpus to create a trained model.

5. Tamil Morphological Analyzer

The Tamil morphological analyzer is based on sequence labeling and training by kernel methods. It captures the non-linear relationships and various morphological features of natural language in a better and simpler way. In this machine learning approach two training models are created for morphological analyzer. These two models are represented as Model-I and Model-II. First model is trained using the sequence of input characters and their corresponding output labels. This trained model-I is used for finding the morpheme boundaries [M Anand kumar et.al,2009].

Second model is trained using sequence of morphemes and their grammatical categories. This trained Model-II is used for assigning grammatical classes to each morpheme. The SVM Tool is used for training the data. Generally SVM Tool is developed for POS tagging but here this tool is used in morphological analysis

6. Tamil Morphological Generator

The developed morphological generator receives an input in the form of lemma word class Morpholexical Information, where lemma specifies the lemma of the word-form to be generated, word class specifies the grammatical category (POS category) and Morpholexical Information specifies the type of inflection. The morphological generator system needs to handle three major things; first one is the lemma part, then the word class and finally the morpholexical information. By the way the generator is implemented makes it distinct from other morphological generator [M Anand kumar et.al,2010].

The input which is in Unicode format is first Romanized and then the paradigm number is identified by end characters. For sake of easy computation we are using romanized form. A Perl program has been written for identifying paradigm number, which is referred as column index. The morpholexical information of the required word class is given by the user as input. From the morpholexicon information list the index number of the corresponding input is identified, this is referred as row index. A verb and noun suffix tables are used in this system. Using the word class specified by the user the system uses the corresponding suffix table. In this two-dimensional suffix table rows are morpholexical information index and columns are paradigm numbers.

7. Conclusion

In this paper, we have presented a morphology based Factored SMT for English to Tamil language. The morphology based Factored SMT improves the performance of translation system for morphologically rich language and also it drastically reduces the training corpus size. So this model is suitable for languages which have less parallel corpus. Tamil morphological models are used to create a factorized parallel corpus. Source language reordering module captures structural difference between source and target language and reorder it accordingly. Compounding module converts the source language structure to fit into the target language structure. Initial results obtained from the Factored SMT are encouraging.

References

1. Philipp Koehn and Hieu Hoang (2007) , " Factored Translation Models ", Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic, June 2007.

2. V Dhanalakshmi, M Anand kumar, K P Soman, S Rajendran (2009),"POS Tagger and Chunker for Tamil language", Proceedings of Tamil Internet Conference 2009, Cologne, Germany, October 2009.

3. M Anand kumar, V Dhanalakshmi. , K P Soman, S Rajendran (2009),"A Novel Apporach For Tamil Morphological Analyzer", Proceedings of Tamil Internet Conference 2009 , Cologne, Germany, Page no: 23-35, October 2009.

4. M Anand kumar, V Dhanalakshmi, R U Rekha, K P Soman, S Rajendran (2010) ,"Morphological Generator for Tamil a new data driven approach", Proceedings of Tamil Internet Conference 2010, Coimbatore, India, 2010.

5. Jes´us Gim´enez and Llu´ıs M`arquez.(2004), "SVMTool: A general pos tagger generator based on support vector machines", Proceedings of the 4th LREC Conference, 2004.

6. Fishel,M (2009), "Deeper than words : Morph-based Alignment for Statistical Machine Translation ", Proceedings of the conference of the pacific Association for Computational Computational Linguistics (PacLing 2009) Sapporo, Japan.

✡✡✡