The Transition of Phrase based to Factored based Translation for Tamil language in SMT Systems

Dr. Ananthi Sheshasaayee[1], Angela Deepa. V.R[2]

[1]Research Supervisior, Department of Computer Science & Application, Quaid-E-Millath Government College for Women (Auonomous), Chennai

[2] Research Scholar (PG), Department of Computer Science & Application, Quaid-E-Millath Government College for Women (Auonomous), Chennai

E-mail-ananthi.research@gmail.com

Abstract— Machine translation is one of the major and the most active areas of Natural language processing. Machine translation (MT) is an automatic translation of one natural language into another using computer generated instructions. The utility and power of Statistical Machine Translation (SMT) seems destined to change our technological society in profound and fundamental ways. The current state-of-the-art approach to statistical machine translation,so-called phrase-based models is limited to linguistic information.For a highly agglutinative languages like Tamil developing a linguistic tools and machine translation system is a challenging task. Therefore, extending the phrase-based to factored based approach by tightly integrating additional annotation information at the word level which encompass not only of tokens but a vector of factor representating the levels of annotation. The additional linguistically features enabled in the toll will increase the accuracy of the SMT systems. This paper motivates the use of factored translation models for statistical machine translation systems for better reliable translation for highly morphological languages like Tamil.

Keywords— Statistical machine translation, Automata theory, Artificial intelligence, Datastructure, Morphology, Linguistics Agglutinative language,

INTRODUCTION

The performance of the SMT systems for English to Tamil language pairs is affected by two main things (i) the amount of parallel data (ii) the language difference owing to the morphological richness and word order differences due to syntactic divergence [7]. The availability of parallel data for English to Tamil language is less. The difference in word order and morphological complexity between the English and Tamil language leads in intricacy of building the translation models. In SMT systems the current state-of-the-art approach for translation model is phrase-based models. To translate morphological rich languages like Tamil there is a need to integrate linguistic information at word level in the translation model which include not only of tokens but a vector of factors representating the levels of annotation.This leads to the new approach termed as Factored based approach.

This paper motivates the importance of using factored based approach in the translation models of SMT systems for translating English and Tamil language pairs by integrating it with state-of-the-art phrase based models .The remaining part of the paper is organized as follows: Section 2 discuss about the various SMT systems build for English Tamil language pairs. Section 3 portrays the role of phrase based approach in translation models. In Section 4, 5, 6 the motivation for and an overview of the present model is given. The paper is concluded in Section 7

LITERATURE SURVEY

A Statistical Machine Translation Systems for Sinhala to Tamil language was developed by Ruvan Weerasinghe [1].In this method a small trilingual parallel corpora was formed which contains the newsevents, culture and politics of Srilanka. A semi automatic approach was employed to perform sentence boundary detection. The sentences were aligned manually and a total of 4064 sentences of Sinhala and Tamil were used in this systems.

A Statistical Machine Translation System by Ulrich Germann (2001)[18] developed a small parallel corpus Tamil-English for about 100,000 words on the Tamil side using several translators. As a part of this a simple text stemmer for Tamil was built based on the Tamil infections tables which helped to increase the performance of the systems.

In 2002 Fredric C.Gey [16] assembled a corpus of Tamil news stories from Thinaboomi website which contains nearly 3000 news stories in Tamil language. This developed corpus is been used to develop a statistical Machine translation by Information Sciences

Institute,one of the leading machine translation research organizations.
An interactive approach to develop web based English to Tamil machine translation system was proposed by Vasu Renganathan.[17]
Google developed a web based machine translation engine for English to Tamil language which is facilitated to identify the source language automatically.

TRANSLATION MODELS

(i)Word based translation

In word based translation model [19], words are the translation elements. The word based translation models rely on high fertility rates and the mapping of the single word to multiple words. Fertility is defined as the ratio of the length of sequences of translated words aiming at producing the n number of words from the source word. The Statistical machine translation is that every sentences t in a target language is a possible translation of a given sentence e in a source language. Based on the bilingual text corpus and the probability assigned to each sentence the possible translation of a given sentence is estimated. Therefore considering the words as the translation units with the applied probabilities the first Statistical machine translation models based on the words were built.

(ii)Phrase based translation

For better translation between the language pairs the translation of words is replaced with the phrase sentences. The phrase based translation models [15] aims to translate whole sequences of words based on their length. Phrases are not merely linguistic ones but the sequence of words found using statistical methods from corpora. For computation of the translation probabilities with respect to the behavior of the phrase is taken into considerations.

Steps involved in Phrase based Translation process:

a. The sentence from the source language **(E)** is grouped into phrases E_1, E_2…… the arbitrary contiguous sequences of words.
b. Each phrase E_i is translated into T_i (phrase of Tamil language)
c. The phrase in source language are reordered according to the target language (**T**) The translation model aims to assign a probability for the given Tamil sentence **T** and an English sentence **E** such that **T** generates **E**. The probability model for the phrase based translation depends on a translation and distortion probability.
The translation probability for generating source phrase T_i from target phrase E_i is $\varphi(Ti/Ei)$. The distortion probability d is responsible for the reordering of the source phrase which means that the probability of two consecutive Tamil phrases are separated in English by a distance of the English word of a particular length.

The distortion is parameterized by $d(a_i - b_{i-1})$

where a_i= start position of the source English phrase being generated by the i^{th} Tamil phrase. b_{i-1}= end position of the source English phrase generated by $i - 1^{th}$ Tamil phrase. Thus calculating the distortion probabilities handles the difference in the order of the words in the phrase based models.

The following equation(1) implies the translation model for phrase based machine translation.

$$P(T/E) = \prod \varphi(T_i/E_i)d(a_i - b_i\text{-}1)\ldots\ldots..(1)$$

Though phrase based models produce better translation than word-based models a novel approach is needed for translating longer units. The lack of linguistic information in phrase based models prone to decrease the translation quality of the systems.

MOTIVATING EXAMPLE:MORPHOLOGY

Tamil is one of the longest surviving classical languages in the world.It is morphologically rich and agglutinative.[3]. Therefore it needs deep analysis at the word level to capture the meaning of the word from its morphemes and its categories. The complex morphological structure of Tamil inflects to person, gender, and number markings and also combines with auxiliaries that indicate aspect, mood, causation, attitude etc in verb. Each root is affixed with several morphemes to generate word. Each root word can take a few thousand inflected word forms.

INCORPORATING LINGUISTIC INFORMATION IN SMT

To model translation systems for Morphological rich languages like Tamil we need to integrate linguistic information into the translation models on the levels of lemmas and grouping the different word forms from the common lemma. Factored translation models allow the integration of additional morphological and lexical information at the word level of both the source and the target languages. Factored translation models[14] is an extension to phrase-based models in which each word is substituted by a vector of factors such as word, Lemma, Part-of-speech information, morphology etc. In order to improve the translation quality the factored models can be employed with various features like morphological coherence [13,9,10,6,4,5], grammatical coherence [11], compound handling [8] or domain adaptation [2, 12].

DECOMPOSITION OF FACTORED TRANSLATION

i) Factored corpora:

Statistical machine translation system accuracy is based on the size of the parallel corpus. The scarce availability of parallel sentences between the English-Tamil language pair inhibits the efficiency of the SMT systems. Therefore in order to build the framework of factored translation the available parallel corpora is cleaned up to separate the words and punctuations. Pre-processing plays a predominant role in creating factored training corpora. For highly agglutinative languages like Tamil the pre-processing is done through the linguistic tools like POS tagger and morphological analysers(Fig1,Fig1.1).

For English the reordering and compounding steps are implemented for creation of factored corpora

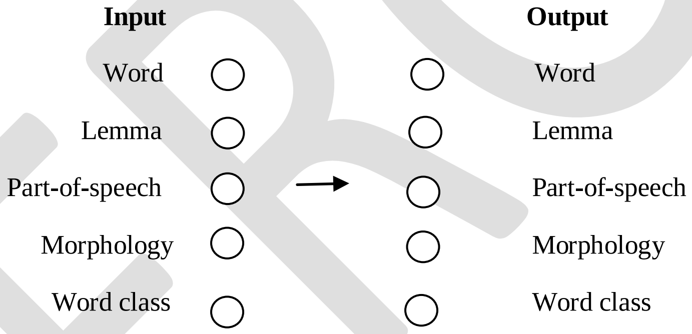


Fig1.Representing (source/target) by factors

ii) Mapping steps in Translation models:

The translation model in the factored based models is broken up into three mapping steps:

1. Translation of input lemmas into output lemmas

2. Translation of morphological and POS factors

3. Generation of surface form through the lemma and the linguistic information.

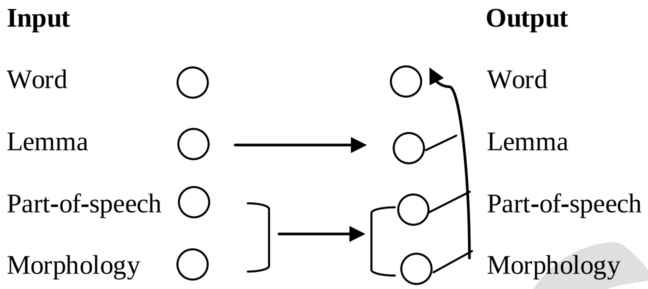
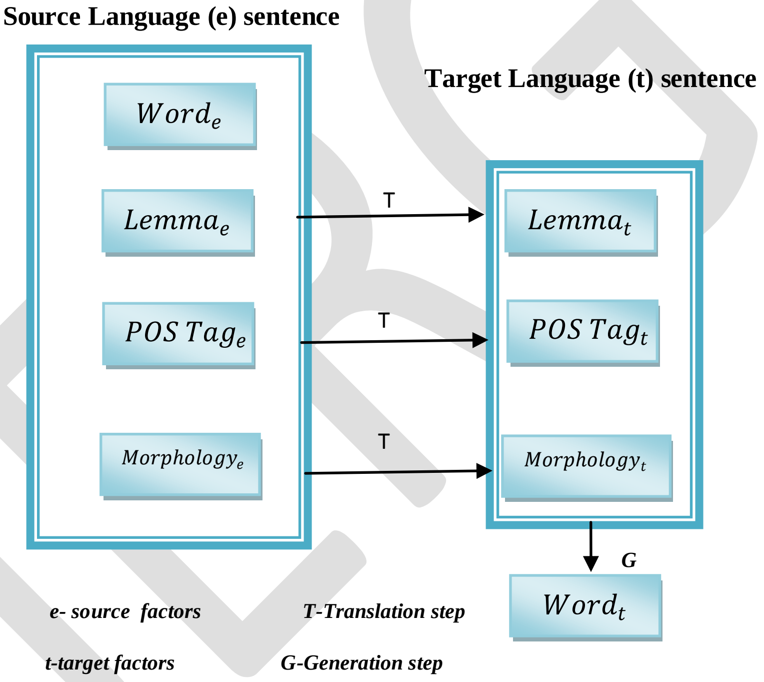


Fig.1.1.Example Factored model

Let us consider an example, the word ‘*book*’ is different from the word ‘*books*’.If the system is trained with the word *book* while translating the system identifies it but it denies to identify the word *books* (plural form)since the system is not trained with the linguistic information Though this problem does not cause much impact for the language English but shows up significant problem for morphological rich language like Tamil.



Fig(2)The annotated factors of a word in a source language(e) to that of the translated factors of source word (e) in Target Language(t)

Thus there is a need for a model in which the lemma and morphological information are separately translated which generates the output surface words on the output side instituting the obtained information. Factored translation models(Fig 2) can ultimately meet these need. Thus before training the sentences the parallel corpus should be annotated with factors which gives linguistic information such as lemma, part-of-speech, morphology etc. Translation steps compute the sentences like the phrase based models whereas the generation steps trains on the target side of the corpus.For every factor annotated additional language models are used to train the system. Models are combined in a log-linear fashion analogous to the different factors and components.

CONCLUSION

This paper describes the significance of factored translation models which is an extension of phrase based approach by integrating the additional information from linguistic tools or automated word classes. Moreover these models can be deployed in morphological rich languages like Tamil for better translation quality.

REFERENCES:

[1] Sripirakas, S., A. R. Weerasinghe, and D. L. Herath. "Statistical machine translation of systems for Sinhala-Tamil." Advances in ICT for Emerging Regions (ICTer),International Conference on IEEE,2010

[2] Niehues, J., Waibel, “A. Domain adaptation in statistical machinetranslation using factored translation models”, In :Proc.of EAMT,2010

[3] Kumar, M. Anand, et al. "A Sequence Labeling Approach to Morphological Analyzer for Tamil Language."International Journal On Computer Science and Engineering,Volume-02,Issue-06,2010

[4] Koehn, P., Haddow, B., Williams, P., Hoang, H.“More linguistic annotation for statistical machine translation”. In: Proc. of WMT And Metrics MATR,Uppsala,Sweden,ACL,Page no(115-120),2010.

[5] Yeniterzi, R., Oflazer, K..” Syntax-to-Morphology Mapping in FactoredPhrase-Based Statistical Machine Translation from English to Turkish”,In:Proc.of ACL, Uppsala, Sweden, ACL,Page no(454-464)

[6] Ramanathan, A., Choudhary, H., Ghosh, A., Bhattacharyya, P.”Case markers and morphology: addressing the crux of the fluency problem in English-Hindi SMT. In: Proc. of ACL/IJCNLP,Suntec,Singapore: Volume 2,Page no(800-808),2009

[7] Koehn, P., Birch, A., and Steinberger, R. “ 462 Machine Translation Systems for Europe”, In MT Summit XII,2009

[8] Stymne, S.,German “Compounds in Factored Statistical Machine Translation”. In Ranta,Bengt Nordstrom Aarne.”Advances in Natural Language Processing”.Volume 5221 of Lecture Notes in Computer Science. Springer Berlin/Heidelberg(2008)

[9] Avramids, E., Koehn, P.:” Enriching morphologically poor languagesfor statistical machine translation”. In: Proc. of ACL/HLT, Columbus,Ohio,ACL,Page no(763-770),2008.

[10] Badr, I., Zbib, R., Glass, J.:” Segmentation for English-to-Arabic statistical machine translation”. In: Proc. of ACL/HLT Short papers, Columbus, Ohio, ACL,Page no(153–156),2008

[11] Birch, A., Osborne, M., Koehn, P.: “CCG Supertags in Factored Statistical Machine Translation”. In: Proc. of ACL WMT, Prague,Czech Republic, ACL,Page no(9–16),2007

[12] Koehn, P., Schroeder, J.:” Experiments in domain adaptation for statistical machine translation”. In: Proc. of ACL WMT, Prague,Czech Republic, ACL ,Page no(224–227),2007

[13] Bojar, O.: “English-to-Czech Factored Machine Translation”. In: Proc.of ACL WMT, Prague, Czech Republic, ACL Page no(232–239),2007

[14] Philipp Koehn and Hieu Hoang.:” Factored translation models”. In Proc.EMNLP+CoNLL, Prague,Page no(868–876),2007

[15] Philipp Koehn, Franz Josef Och, and Daniel Marcu, “Statistical Phrase-Based Translation”, In: Proc.of HLT/NAACL,2003

[16] Fedric C. Gey,“Prospects for Machine Translation of the Tamil Language”, In:Proc. of Tamil Internet conference,

California, USA,2002

[17] Vasu Renganathan, "An interactive approach to development of English to Tamil machine translation system on the web". INFITT, (TI2002),2002

[18] Germann, Ulrich. "Building a statistical machine translation system from scratch: how much bang for the buck can we expect?."In:Proc. of the workshop on Data-driven methods in machine translation-Volume 14. Association for Computational Linguistics, 2001.

[19] Koehn, Philipp, and Kevin Knight. "Knowledge sources for word-level translation models."In: Proc.of the Conference on Empirical Methods in Natural Language Processing. 2001