



STATISTICAL MECHANICS IN ARTIFICIAL
INTELLIGENCE

PROJECT REPORT

Machine Translation System

Aman Bansal
20161181

Prajwal Singhala
20161045

Shashwat khandelwal
20161045

May 1, 2019

Contents

1	Introduction	2
2	Overview	2
3	Implementation	3
4	Experimental Results	4
5	Error Analysis	5
6	Conclusion	5

1 Introduction

Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

In word-based translation, the fundamental unit of translation is a word in some natural language. Typically, the number of words in translated sentences are different, because of compound words, morphology and idioms. The ratio of the lengths of sequences of translated words is called fertility, which tells how many foreign words each native word produces.

Simple word-based translation can't translate between languages with different fertility. Word-based translation systems can relatively simply be made to cope with high fertility, such that they could map a single word to multiple words, but not the other way about. For example, if we were translating from English to Hindi, each word in English could produce any number of Hindi words— sometimes none at all. But there's no way to group two English words producing a single Hindi word.

To counter the above problem people use phrase based models instead of word based which aim to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ. The sequences of words are called blocks or phrases, but typically are not linguistic phrases, but phrasemes found using statistical methods from corpora. It has been shown that restricting the phrases to linguistic phrases (syntactically motivated groups of words, see syntactic categories) decreases the quality of translation. Many-to-many phrase based translation can handle non-compositional phrases.

2 Overview

In the project we have implemented a Phrase based statistical machine translation system which translates sentences from English to Hindi. The Project consists of two parts where the first part included generating a phrase table consisting of a source phrase and its target-side counterparts with a log probability associated with each translation pair while the second part includes

translating English sentences into their Hindi translations using a stack decoder. The data set for training the model is IIT Bombay English-Hindi Corpus and consists of 40000 English phrases with their Hindi translations.

3 Implementation

Before we begin, we pre-process the dataset in order to remove all special characters. In order to generate phrase based model translation we first calculate the phrase translation probabilities and later employ stack based decoding to obtain the best translation. The phrase translation probabilities take components from both the training set and the language model. The PBSMT model generation consists of the following:

1. **Source Reordering**

One of the major difficulties of machine translation lies in handling the structural differences between the language pair. Translation from one language to another becomes more challenging when the language pair follows different word order. For example, English language follows subject-verb-object (SVO) whereas Hindi follows subject-object-verb (SOV) order.

2. **Word Alignment Generation**

We create the word alignment of each pair of sentences using GIZA++. Giza++ takes as input a pair of sentence (source sentence and its target sentence) and returns the alignment of the target sentence for the given source language.

3. **Phrase Transition probabilities**

The phrase translation probability calculation is a two-step process.

- **Extraction of phrases**

In the first step we generate the phrases which are consistent. We call a phrase pair (f, e) consistent with an alignment A , if all words f_1, \dots, f_n in f that have alignment points in A have these with words e_1, \dots, e_n in e and vice versa.

- **Calculation of Phrase based transition probabilities**

In order to calculate the phrase translation probabilities we for a foreign phrase f , given an English phrase e , we count the relative

occurrences of the foreign phrases given the English phrase. Thus the phrase translation probability is given by:

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \text{count}(\bar{e}, \bar{f}_i)}$$


Along with this we incorporate the probability given by the language model. In order to obtain such probabilities we use the language model given byIRSTLM.

In order to reduce very large phrases we keep an upper limit of 10 on the size of the phrase. Also log linear probabilities are used so as to avoid extremely small probability values.

4. Stack Based Decoding

After obtaining the phrase translation probabilities we apply stack based decoding to obtain the best translation. The stack based decoding algorithm takes as input the translation probability values and returns the best translation by recombining the best hypothesis.

4 Experimental Results

English	Predicted	Hindi
Who will come today?	कौन आएगा	आज कौन आएगा?
How are you?	आप कैसे हो	आप कैसे हो?
My name is khan	नाम है	मेरा नाम खान है 

5 Error Analysis

Sentence No.	Precision	Recall	Sentence BLEU score
1	0.46	0.46	0.23
2	1.0	1.0	1.0
3	0.21	0.18	0.04

	Corpus BLEU score
Baselines	10.75
PBSMT	5.32

6 Conclusion

In the project we see that phrase based statistical machine translation solves some problems which we find in word based machine translation but there are some issues which are still not addressed in PBSMT like if multiple segmentations of a phrase are possible why do we choose one over another or do we choose larger phrase pairs or multiple shorter phrase pairs and these are addressed in operation sequence model which gives significant improvements over phrase based baseline.