**CS F469**

**Information Retrieval**

**Assignment 3**

**IBM Models**

***Group Members:***

Aditya Desai - 2015A3PS0211H

Meghna Vasudeva - 2015B3A70664H

Snigdha Grandhi - 2015A8PS0495H

Aayush Barthwal - 2015A7PS0136H

***Description of each .py file:***

1. ‘assgn3\_1.py’ implements the IBM Model 1 and the EM Algorithm.

The alignments are taken as an intersection of the alignments from English to foreign language and from the foreign language to English. As a result, a word can be aligned to multiple words.

1. ‘assgn3\_2.py’ is used for comparing the alignments obtained from the results of ‘assgn3\_1.py’ with the nltk implementations of IBM Model 1 and Model 2.

The alignments from ‘assgn3\_1.py’ are used, and the alignments from nltk IBM models are directly obtained using the get\_alignment method for each sentence pair.

1. ‘assgn3\_3.py’ has the code to extract phrases using the alignments from ‘assgn3\_1.py’ and the phrased based translation module from nltk, and then ranks the phrases in decreasing order of probability.

In the end, the corresponding translated word is printed for each extracted phrase.

***Corpus*** – The corpus consists of around 30 English-to-Hindi sentence pairs. The alignments of each sentence pair has been shown in the next section with the results. The file is present in the **data** folder along with the given datasets.

Since the foreign language chosen (Hindi) follows a different order (Subject-Object-Verb) unlike English which follows Subject-Verb-Object order, the alignments from the IBM Model 1 and 2 were very different from the actual alignments. Also, words like ‘He’ can have multiple translations to ‘वह’, उसने, etc. depending on the context, this lead to errors in the output alignments.

The models work fine with the given dataset (in French) as the language follows the same order (Subject-Verb-Object), thus the output alignments are similar for both the IBM models.

***Results:***

*Running Time of IBM1 Model implementation*

* It takes around **0.025 seconds** to get the alignments from our corpus of 34 sentences.
* It takes around **0.000496 seconds** and **0.0025 seconds** to run on the ‘data1.json’ and ‘data2.json’

*Reasons for differences in the alignments*

IBM Model 1 is weak in terms of conducting reordering or adding and dropping words. In most cases, words that follow each other in one language would have a different order after translation, but IBM Model 1 treats all kinds of reordering as equally possible.

Another problem while aligning is the fertility (the notion that input words would produce a specific number of output words after translation). In most cases one input word will be translated into one single word, but some words will produce multiple words or even get dropped (produce no words at all). The fertility of word models addresses this aspect of translation. While adding additional components increases the complexity of models, the main principles of IBM Model 1 are constant. Even the IBM Model 2 suffers from the same problems as it uses IBM Model 1 for lexical translation and adds alignment on top of it.

Phrase based translation has added benefit of many to one translation over the IBM models. It accounts for complete phrases instead of going for word to word translations, leading to better translations.

*Alignments*

* English Text : ﻿He came running.

Foreign Text : वह भागते हुए आया।

IBM1 : [(0, 0), (1, 1), (2, 2)]

IBM1 (nltk) : [(0, 2), (1, 2), (2, 2)]

IBM2 (nltk) : [(0, 3), (1, 1), (2, 1)]

Actual : [(0, 0), (1, 3), (2, 1), (2, 2)]

* English Text : He can't be ill.

Foreign Text : वह बीमार नहीं हो सकता।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 0), (1, 4), (2, 4), (3, 4)]

IBM2 (nltk) : [(0, 0), (1, 2), (2, 3), (3, 2)]

Actual : [(0, 0), (1, 4), (2, 2), (3, 3), (4, 1)]

* English Text : He has a hat on.

Foreign Text : उसने टोपी पहनी हुई है।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3), (4, 1)]

IBM1 (nltk) : [(0, 0), (1, 3), (2, 4), (3, 3), (4, 3)]

IBM2 (nltk) : [(0, 0), (1, 3), (2, 4), (3, 3), (4, 3)]

Actual : [(0, 0), (1, 3), (2, 4), (3, 1), (4, 2)]

* English Text : He reads Arabic.

Foreign Text : वह अरबी पढ़ सकता है।

IBM1 : [(0, 0), (1, 1), (2, 2)]

IBM1 (nltk) : [(0, 0), (1, 3), (2, 3)]

IBM2 (nltk) : [(0, 0), (1, 3), (2, 3)]

Actual : [(0, 0), (1, 2), (1, 3), (1, 4), (2, 1)]

* English Text : He's my husband.

Foreign Text : यह मेरा पति है।

IBM1 : [(0, 0), (1, 1), (2, 2)]

IBM1 (nltk) : [(0, 2), (1, 2), (2, 2)]

IBM2 (nltk) : [(0, 2), (1, 1), (2, 1)]

Actual : [(0, 0), (1, 3), (2, 1), (3, 2)]

* English Text : How is everyone?

Foreign Text : सब लोग कैसे हैं?

IBM1 : [(0, 0), (1, 1), (2, 2)]

IBM1 (nltk) : [(0, 3), (1, 3), (2, 3)]

IBM2 (nltk) : [(0, 3), (1, 1), (2, 1)]

Actual : [(0, 2), (1, 3), (2, 0), (2, 1)]

* English Text : How is it going?

Foreign Text : कैसा चल रहा है?

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 2), (1, 2), (2, 2), (3, 2)]

IBM2 (nltk) : [(0, 0), (1, 1), (2, 1), (3, 1)]

Actual : [(0, 0), (1, 3), (2, 2), (3, 1)]

* English Text : How old are you?

Foreign Text : तुम कितने साल के हो?

IBM1 : [(0, 1), (1, 0), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 1), (1, 4), (2, 4), (3, 4)]

IBM2 (nltk) : [(0, 0), (1, 4), (2, 3), (3, 2)]

Actual : [(0, 1), (1, 2), (2, 3), (2, 4), (3, 0)]

* English Text : How rude of you!

Foreign Text : कितने बदतमीज़ हो तुम!

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 0), (1, 3), (2, 3), (3, 3)]

IBM2 (nltk) : [(0, 0), (1, 1), (2, 1), (3, 1)]

Actual : [(0, 0), (1, 1), (2, 2), (3, 3)]

* English Text : I cooked dinner.

Foreign Text : मैंने रात का खाना पकाया।

IBM1 : [(0, 0), (1, 1), (2, 2)]

IBM1 (nltk) : [(0, 0), (1, 4), (2, 4)]

IBM2 (nltk) : [(0, 0), (1, 3), (2, 3)]

Actual : [(0, 0), (1, 4), (2, 2), (2, 3), (2, 4)]

* English Text : I feel nauseous.

Foreign Text : मुझे उल्टी आ रही है।

IBM1 : [(0, 0), (1, 1), (2, 2)]

IBM1 (nltk) : [(0, 0), (1, 3), (2, 3)]

IBM2 (nltk) : [(0, 0), (1, 3), (2, 3)]

Actual : [(0, 0), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4)]

* English Text : I go every year.

Foreign Text : मैं हर साल जाता हूँ।

IBM1 : [(0, 0), (1, 1), (2, 3), (3, 2)]

IBM1 (nltk) : [(0, None), (1, 1), (2, 3), (3, 1)]

IBM2 (nltk) : [(0, 0), (1, 1), (2, 3), (3, 1)]

Actual : [(0, 0), (0, 4), (1, 3), (2, 1), (3, 2)]

* English Text : I heard a noise.

Foreign Text : मुझे कोई आवाज़ सुनाई दी।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 0), (1, 4), (2, None), (3, 4)]

IBM2 (nltk) : [(0, 0), (1, 4), (2, 3), (3, 2)]

Actual : [(0, 0), (1, 3), (2, 1), (2, 4), (3, 2)]

* English Text : I like this one.

Foreign Text : मुझे यह वाला पसंद है।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 0), (1, 3), (2, 3), (3, 3)]

IBM2 (nltk) : [(0, 0), (1, 2), (2, 3), (3, 2)]

Actual : [(0, 1), (1, 3), (2, 1), (2, 4), (3, 2)]

* English Text : I run every day.

Foreign Text : मैं रोज़ दौड़ने जाता हूँ।

IBM1 : [(0, 0), (1, 1), (2, 3), (3, 2)]

IBM1 (nltk) : [(0, None), (1, 2), (2, 3), (3, 2)]

IBM2 (nltk) : [(0, 0), (1, 2), (2, 3), (3, 2)]

Actual : [(0, 1), (0, 4), (1, 2), (2, 1), (3, 1)]

* English Text : I saw her again.

Foreign Text : मैंने उसे फिरसे देखा।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 0), (1, 3), (2, 3), (3, 3)]

IBM2 (nltk) : [(0, 0), (1, 1), (2, 1), (3, 1)]

Actual : [(0, 0), (1, 3), (2, 1), (3, 3)]

* English Text : I want a guitar.

Foreign Text : मुझे गिटार चाहिए।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 1)]

IBM1 (nltk) : [(0, 0), (1, 2), (2, None), (3, 2)]

IBM2 (nltk) : [(0, 0), (1, 2), (2, None), (3, 2)]

Actual : [(0, 0), (1, 2), (2, 0), (3, 1)]

* English Text : I was assaulted.

Foreign Text : मुझपर हमला किया गया।

IBM1 : [(0, 0), (1, 1), (2, 2)]

IBM1 (nltk) : [(0, None), (1, 3), (2, 3)]

IBM2 (nltk) : [(0, 3), (1, 1), (2, 1)]

Actual : [(0, 0), (1, 2), (1, 3), (2, 1)]

* English Text : I'm a good cook.

Foreign Text : मैं अच्छा रसोइया हूँ।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 2), (1, None), (2, 2), (3, 2)]

IBM2 (nltk) : [(0, 0), (1, 1), (2, 1), (3, 1)]

Actual : [(0, 0), (1, 3), (2, 1), (3, 2)]

* English Text : Come and join us.

Foreign Text : आओ हमारे साथ बैठो।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 3), (1, 3), (2, 3), (3, 3)]

IBM2 (nltk) : [(0, 0), (1, 1), (2, 1), (3, 1)]

Actual : [(0, 0), (1, 1), (2, 3), (3, 1)]

* English Text : Come sit with us.

Foreign Text : आओ हमारे साथ बैठो।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 3), (1, 3), (2, 3), (3, 3)]

IBM2 (nltk) : [(0, 0), (1, 1), (2, 1), (3, 1)]

Actual : [(0, 0), (1, 3), (2, 2), (3, 1)]

* English Text : What's that bird?

Foreign Text : वह कौनसी चिड़िया है?

IBM1 : [(0, 3), (1, 0), (2, 1)]

IBM1 (nltk) : [(0, 3), (1, 2), (2, 2)]

IBM2 (nltk) : [(0, 3), (1, 1), (2, 1)]

Actual : [(0, 1), (0, 3), (1, 0), (2, 2)]

* English Text : What's the story?

Foreign Text : क्या कहानी है?

IBM1 : [(0, 0), (1, 1), (2, 1)]

IBM1 (nltk) : [(0, 2), (1, 1), (2, 1)]

IBM2 (nltk) : [(0, 2), (1, 1), (2, 1)]

Actual : [(0, 0), (1, 2), (2, 1)]

* English Text : What's your name?

Foreign Text : आपका नाम क्या है?

IBM1 : [(0, 2), (1, 0), (2, 1)]

IBM1 (nltk) : [(0, 3), (1, 1), (2, 1)]

IBM2 (nltk) : [(0, 3), (1, 1), (2, 1)]

Actual : [(0, 2), (0, 3), (1, 0), (2, 1)]

* English Text : He breathed deeply.

Foreign Text : उसने गहरी साँस ली।

IBM1 : [(0, 0), (1, 1), (2, 2)]

IBM1 (nltk) : [(0, 0), (1, 3), (2, 3)]

IBM2 (nltk) : [(0, 3), (1, 1), (2, 1)]

Actual : [(0, 0), (1, 2), (1, 3), (2, 1)]

* English Text : He closed his eyes.

Foreign Text : उसने अपनी आँखें बंद करीं।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 0), (1, 4), (2, 4), (3, 4)]

IBM2 (nltk) : [(0, 0), (1, 4), (2, 3), (3, 2)]

Actual : [(0, 0), (1, 3), (1, 4), (2, 1), (3, 2)]

* English Text : He cried and cried.

Foreign Text : वह रोया और और रोया।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 0), (1, 4), (2, 4), (3, 4)]

IBM2 (nltk) : [(0, 0), (1, 4), (2, 3), (3, 1)]

Actual : [(0, 0), (1, 1), (2, 2), (2, 3), (3, 1)]

* English Text : He is a simple man.

Foreign Text : वह साधा आदमी है।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 1), (4, 2)]

IBM1 (nltk) : [(0, 0), (1, 2), (2, 3), (3, 2), (4, 2)]

IBM2 (nltk) : [(0, 0), (1, 3), (2, 3), (3, 2), (4, 2)]

Actual : [(0, 0), (1, 3), (2, 0), (3, 1), (4, 2)]

* English Text : He is sure to come.

Foreign Text : वह ज़रूर आएगा।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 1), (4, 2)]

IBM1 (nltk) : [(0, 0), (1, 2), (2, 2), (3, 2), (4, 2)]

IBM2 (nltk) : [(0, 0), (1, 2), (2, 2), (3, 2), (4, 2)]

Actual : [(0, 0), (1, 0), (2, 1), (3, 2), (4, 2)]

* English Text : We are against war.

Foreign Text : हम युद्ध का विरोध करते हैं।

IBM1 : [(0, 0), (1, 1), (2, 2), (3, 3)]

IBM1 (nltk) : [(0, 0), (1, 5), (2, 5), (3, 5)]

IBM2 (nltk) : [(0, 0), (1, 5), (2, 5), (3, 5)]

Actual : [(0, 0), (1, 4), (1, 2), (1, 5), (2, 3), (3, 1)]

* English Text : We enjoyed skating.

Foreign Text : हमे स्केटिंग करके मज़ा आया।

IBM1 : [(0, 0), (1, 1), (2, 2)]

IBM1 (nltk) : [(0, 3), (1, 3), (2, 3)]

IBM2 (nltk) : [(0, 0), (1, 3), (2, 3)]

Actual : [(0, 0), (1, 3), (1, 4), (2, 1), (2, 2)]

* English Text : We were in a hurry.

Foreign Text : हम जल्दी में थे।

IBM1 : [(0, 0), (1, 3), (2, 1), (3, 2), (4, 1)]

IBM1 (nltk) : [(0, 0), (1, 3), (2, 2), (3, None), (4, 2)]

IBM2 (nltk) : [(0, 0), (1, 3), (2, 2), (3, None), (4, 2)]

Actual : [(0, 0), (1, 3), (2, 2), (3, 3), (4, 1)]

* English Text : We were very tired.

Foreign Text : हम बहुत थक चुके थे।

IBM1 : [(0, 0), (1, 4), (2, 1), (3, 2)]

IBM1 (nltk) : [(0, 0), (1, 4), (2, 3), (3, 3)]

IBM2 (nltk) : [(0, 0), (1, 4), (2, 3), (3, 2)]

Actual : [(0, 0), (1, 3), (2, 4), (2, 1), (3, 2)]