IT412: Natural Language Processing

Assignment 7: Word Alignment for Bilingual Parallel Corpora

Instructor: Prasenjit Majumder

Lab Outcome: This assignment will introduce to machine translations and how to create alignments for parallel corpus.

1 Machine Translation

Machine Translation is the task of translating a sentence x from one language (the source language) to a sentence y in another language (the target language). One of the earliest goals for computers was the automatic translation of text from one language to another. Fig 1 shows an example of English to French Machine Translation. Few examples of translators are Google Translate, Microsoft Translate, IBM Watson etc. Automatic or machine translation is perhaps one of the most challenging tasks.

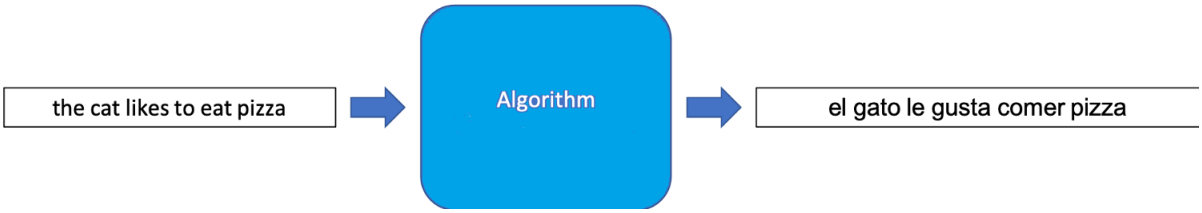

Figure 1: Example of Machine Translation

First approach to solve MT was rule-based system that rely on countless algorithms based on the grammar, syntax, and phraseology of a language. With lots of big parallel texts becoming available Statistical Machine Translation was developed for automatic Machine Translation.

1.1 Statistical Machine Translation

Statistical machine translation (SMT) learns how to translate by analyzing existing human translations (known as bilingual text corpora). SMT starts with a very large data set of good translations, that is, a corpus of texts which have already been translated into multiple languages, and then uses those texts to automatically infer a statistical model of translation. Below is equation of SMT.

$$\begin{aligned} e_{\text{best}} &= \arg\max_e P(e|f) \\ &= \arg\max_e P(f|e)P(e) \end{aligned}$$

where, $arg\max_e$is decoding algorithm
$P(f|e)$ is translation model
$P(e)$ is language model

For statistical translation word alignment function $a : i \rightarrow j$ is used. Fig 2 shows word based alignment model.

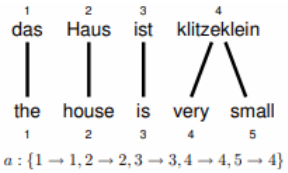

Figure 2: Example of alignment model

2 Implementation

2.1 Dataset

For this experiment you are provided two parallel corpus one in English-Hindi and another is English-French.

2.2 Exercise

1. Using Giza++ create word alignment for both the given corpora.

3 References

- A Systematic Comparison of Various Statistical Alignment Models
- http://www.ling.helsinki.fi/kit/clt310smt/SMT-2016k/f3-handout.pdf
- http://homepages.inf.ed.ac.uk/pkoehn/publications/esslli-slides-day3.pdf