Language: Czech, Hindi.

Creating a Bilingual Dictionary

Using IBM Model 1 Alignment Algorithm

Cross Lingual Alignment Similarity Score

For a given sentence in foreign language, we find (or retrieve) a candidate sentence in English language that has the highest similarity

– Translational model

- Language model

The translational model can be calculated using the IBM Model 1 Alignment model, which gives a bilingual dictionary. We need to identify possible translations for a sentence in foreign language to a sentence in English. We can modify this as a similarity measure.

Where,

The translation model is defined by,

and the length model is defined by,

Dataset

Considered 2 datasets one from Europarl and other from Parallel corpus in Indian languages

Baseline Experiments – 2 Based on Jaccard similarity

We considered the given parallel corpus and converted the sentences in English to the corresponding language using simple word substitution. For the basic experiment, the word in the foreign language is identified using a standard dictionary by choosing the word that is the most likely in the given language. We then calculated the Jaccard scores between the words in the translated sentence and the actual sentence in foreign language. This is given as “Jaccard 1”.

To bring in more domain based translations, we used the IBM M1 model dictionary trained on separate pairs of sentences and chose the most likely word to determine the translated words. The Jaccard scores are given as “Jaccard 2”

We designed the experiments as an information retrieval task; “Identify the sentence in foreign language that has the same meaning as sentences in English”. We compared every sentence (translated) with every other sentence in the foreign language and calculated the scores, in terms of Jaccard and alignment similarities. The scores are then ranked based on the most similar ones. The accuracy measure is calculated at different rank levels (1,3, 5 and 10).

**Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **Algorithm** | **Top-1** | **Top – 3** | **Top – 5** | **Top – 10** |
| Hindi | Jaccard - 1 | 0.244 | 0.322 | 0.376 | 0.430 |
| Jaccard - 2 | 0.340 | 0.460 | 0.504 | 0.538 |
| CLPD-ASA | **0.544** | **0.702** | **0.734** | **0.798** |
| Czech | Jaccard - 1 | 0.242 | 0.328 | 0.388 | 0.476 |
| Jaccard - 2 | 0.234 | 0.310 | 0.334 | 0.350 |
| CLPD-ASA | **0.402** | **0.557** | **0.624** | **0.716** |

Training:

Czech – 50000 Sentences

Hindi – 49500 Sentences

Testing:

Czech – 500 sentences

Hindi – 500 sentences