ASSIGNMENT REPORT

Hindi WikiGenerator

- Vivek Iyer (20161188)

Aim:

To automate generation of a Wikipedia article from its page on Wikidata

Parameters:

The article I chose is “India” (https://www.wikidata.org/wiki/Q668). The language of WikiGeneration used is Hindi.

Algorithm:

My algorithm can be described in detail as follows:

1. First, a ***SPARQL query*** is made to the Wikidata endpoint, in order to extract ***ALL the entities*** “India” is related to that have ***labels in Hindi*** and the corresponding relationships too.

2. Next, the ***Hindi label*** of the relationship is used if available. Else the python ***googletrans*** library is used to translate the English relationship to Hindi.

3. This gives me a list of ***subject, predicate, object*** triples where the subject is taken as “भारत”. ***Predicate*** is the ***extracted relationship*** while ***object*** is the ***entity*** “भारत” is related to.

4. **Section Classifier:** Given a relationship, this algorithm ***assigns*** it to one of pre-defined ***sections*** eg. History (इतिहास), Geography (भूगोल), Politics (राजनीति) etc - the kind of sections that are usually there on Wikipedia. If none of these are relevant, it assigns it to the intro paragraph. This is implemented as follows:

    A. For each relationship, I extract the ***Noun Phrase*** from that relationship using ***Regex Grammar Parser***.
    B. Then I check the similarity of this Noun Phrase with each of the above-mentioned sections.
    C. The ***max similarity*** is chosen and if this similarity is more than a certain ***threshold***, the relationship is assigned to the corresponding section. In other words, the sentence that will be created with this relationship and the corresponding entity should go in the chosen section.

5. **Hindi POS Tagger:** Also, for every relationship I use a **Hindi POS tagger** to get the ***POS tags*** for that relationship.

6. Based on these POS tags, I used a syntactic ***rule-based approach*** to join the subject with the predicate and object and form a grammatical sentence

7. This sentence is assigned to the section outputted by **Section Classifier.** Sentences which haven't been assigned to any section are included in the introduction.

Code:

The code for the Hindi WikiGenerator is available **here**

Analysis:

It can be observed that apart from a few grammatical errors and some misclassifications as regards the section, my algorithm is able to give a decent output that makes sense to the reader.

Future Work:

Given the current time frame, this is what I was able to implement. However, as future work, the following improvements can be made to overcome the errors in the generated article:

1. Train an **HMM model** on a **Hindi text corpus** tagged using the Hindi POS tagger and use the HMM to predict the most likely tag given the surrounding tags (those of entities and relationships), basically a CBOW (Continuous Bag of Words) model. Or even better, train a neural network to predict the most likely word given neighbouring tags. This will help in better grammar while joining sentences.

2. In **Section Classifer,** using Word2Vec similarity on the relationship alone may generate misclassifications because of lack of context in phrases. A better approach would be to input the above generated sentence to a neural network such that it can use the sequence of words to predict the corresponding section.

3. Instead of forming many short sentences, ***longer*** sentences can be formed using pronouns and conjunctions. Facts therefore need to be combined together.