Project Report
CS 215 : Group 21

Anand Dhoot 130070009 | Samarth Mishra 130260018 | Ayush Dhakar 130050033

November 16, 2014

Our final approach to the problem involved comparing values in the given knowledge base to the values appearing in the sentence, comparing each of the given values(in the sentence) for each of the countries appearing in the sentence. We relaxed exact matching by using intervals around the value to be matched (the intervals were in terms of percentages of the given value). We then searched for keywords corresponding to the given relations, in the sentence and increased confidence scores based on number of matches found.

Our step by step approach to the problem is as follows:

We, first input the data from the file kb-facts-train_SI.tsv in a map. This map was key-ed using country id and each element contained another map, key-ed using relations expressed for the country in the knowledge base and contained the corresponding data values in a vector<double>. We then made another map from country-name to country-id using the file countries_id_map.txt . Also, a map of keywords corresponding to a relation was made using a file keywords.txt (which we made from the file selected_indicators).

The code then reads the file sentences.tsv line by line and checks compares each of the values against the values available in the knowledge base for each of the countries appearing in the sentence (this is done by the function findResults()). We assign preliminary confidence scores to the sentence relation as described below:

If we are at a value X in the knowledge base (during the comparisons) and we are comparing it against a value X', then we define a relative difference between the two as

$$d = \left| \frac{X' - X}{max(X, X')} \right|$$

where d is the normalised difference. We find the confidence score(c) based on this value as follows

$$c = 100 * e^{\frac{-d^2}{2}}$$

where c is in percentage.

This value is intuitively okay in the sense that it decreases as the confidence increases and its value is 1 when the two values match exactly. The reason for choosing such a function is that it allows a certain range of values with pretty good confidence and falls sharply after a certain limit. Hence, it seemed to be better than a simple inverse relation.

We then check for the keywords which appear in the text and may indicate certain relations. For a given relation, we find the number of matches of its characteristic keywords in the given sentence and for the nth match, we increase the confidence score as

$$new \quad confidence = \frac{(confidence \quad for \quad n-1 \quad matches) * 2 + 100 * 1}{3}$$

Thus, the confidence score keeps on increasing for as many matches of the key words of the relation are found in the sentence.

Finally, we choose only those relations whose confidence score is at least as much as 50%.