Fact checker

G.Sireesha Naidu
Jahnavi Nukireddy
Spurthi Tallam

Problem Statement

- Given a document(s) and a fact, determine whether the given fact is true/false

Approach

- Summarisation of document(s) in the context of fact checking
- Finding semantic similarity of the fact with the summary
- Finding the sentiment of the fact, and each of the sentences of the summary
- Comparison of the similarities, sentiments obtained and making final decision

Extractive Summarization

- The features used for the construction of a mapping between the document sentences and a corresponding vector are :
  - Semantic similarity of the sentence with other sentences in the doc.
  - Location of the sentence in its respective paragraph (position of a sentence, first and last get higher score values, since they are far away from the mid position)
  - Length of the sentence (cut off taken as > 3 and $<= 7$)
  - No of content words in the sentence (total - no_of_stopwords)

Similarity

- To find the similarity between two sentences, we used
  - Semantic vector construction
  - Word order vector construction
- For semantic vector construction -
  - We constructed a vector each for both sentences
    - Joint set - set of words from both the sentences
    - Length of the vector = length of the joint words set
    - For a sentence, if a joint word is present, then corresponding vector[word] = 1
    - Else, vector[i] = info_content_score(word) * info_content(most_sim_word)
  - The final semantic similarity score is the dot product of the two vectors

Contd.

- Word order similarity
  - Two word order vectors are constructed
    - Joint set = union(words of sentence 1, words of sentence 2)
    - For a sentence, if the joint word is present in the sent_set, vector[i] = position(word)
    - Else try to find the most similar word of joint_word in the sent_set, and return its index
  - Then the similarity score is the normalized difference of the word order $S_r = 1 - \frac{\| \mathbf{r}_1 - \mathbf{r}_2 \|}{\| \mathbf{r}_1 + \mathbf{r}_2 \|}$.
- The total score for the similarity is the weighted average of the two score, i.e..,. λ(semantic_score) + (1-λ)word_order_score ($\lambda = 0.85$ used)

Paper reference - **Sentence Similarity Based on Semantic Nets and Corpus Statistics** by Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett

Contd.

- The final score of the sentence is given as :
  - Weighted average of the four scores
  - $\lambda 3(\lambda 2(\lambda 1(\text{similarity\_score})+(1-\lambda 1)\text{location\_score})+(1-\lambda 2)\text{length\_score})+(1-\lambda 3)\text{cuewords\_score}$
- We choose $\lambda 1 = 0.7$, $\lambda 2 = 0.6$, $\lambda 3 = 0.8$ such that
  - **Weight of similarity feature = 0.336**
  - **Weight of location feature = 0.144**
  - **Weight of length feature = 0.32**
  - **Weight given to content words = 0.2**
- We have chosen top 50% (highest final scores) of all sentences to be included in the summary

Similarity between two sentences results

I like that bachelor.	I like that unmarried man.	0.801
John is very nice.	Is John very nice?	0.660
Red alcoholic drink.	A bottle of wine.	0.477
A glass of cider.	A full cup of apple juice.	0.685
It is a dog.	That must be your dog.	0.48
It is a dog.	It is a log.	0.858
It is a dog.	It is a pig.	0.858
Dogs are animals.	They are common pets.	0.550
Canis familiaris are animals.	Dogs are common pets.	0.458
I have a pen.	Where do you live?	0.197
I have a pen.	Where is ink?	0.149

Results

- Summary = “ Blood demand has been increasing all over the world. In order to utilize the donated blood to maximum extent, generally the whole blood is divided into platelets and RBCs by a process called centrifuging the blood and stored separately. The average adult has about 10 units of blood in his body. Searching for donors and getting the blood on time is becoming a bigger problem day by day. A blood bank is an organization that collects, processes, stores, and transfuses blood. It is a perishable commodity with limited life. It is made up of various components. It is operated by medical technologists under the direction of a pathologist. If the demand - supply - storage parameters are specifically mapped, the shortfall of blood components can substantially come down. Blood donation is a simple four-step process. In most health agencies the blood bank is located in the pathology laboratory. It is in this organization that stock of blood is maintained in healthy conditions to meet the demands of common people. Currently there is a deficit of blood components in India. The blood bank is a valuable resource for the health and wealth of human beings. So, there is a need for a data science tool which will help Organisers, Blood Banks, Health departments as well as the hospital to forecast the blood components requirements in short term and accordingly plan the blood donation drives which will lead to proper utilization of blood components, demand fulfillment of patients and reducing the deficit of supply and wastage of blood components. ”

- **Fact - “Blood is an essential component” ---> Similarity score - 5.55397613097**
- **Normalization of the score with the document size, and if it is greater than a threshold value, then it is true, else false.**
- **Const. Fact - “Blood is an not essential component” ---> Similarity score - 5.13618104749**

Challenges

- The above approach gives the same result for a fact and its negation which is contradictory.
- So, to improve the performance, we considered sentiment analysis to differentiate the sentiments of the fact in the summary.
- “Blood is an essential component” - {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
- “Blood is not an essential component” - {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Thank you