Problem Statement

Design a QA system which takes a Wikipedia article and query related to that. Then predict a possible answer for that query.

Research

While research I found there is word embedding techniques which represent the word in N-dimension vector space so that semantically similar words come closer to each other.

There is one algorithm word mover distance to find distance between sentences. This algorithm distance measures the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to "travel" to reach the embedded words of another document.

There is a Smooth Inverse Frequency method which takes the average of the word embeddings in a sentence tends to give too much weight to words that are quite irrelevant, semantically speaking.

There are many pre-trained encoding methods. They have similar aim as sentence embedding, the embeddings these encoder produce can be used in a variety of applications, such as text classification, paraphrase detection, etc. This is because they have been trained on a range of supervised and unsupervised tasks, in order to capture as much universal semantic information as possible.

There were many other things that I unfold while researching on this topic.

Proposed solution

This problem statement belongs to the domain of NLP. My approach revolves around finding the similarity between the text from the link and the query.

The solution was divided into three major part. Firstly, the data extraction and preprocessing. Secondly, building the model. Thirdly using the model to predict best solution.

Here we have used the word2Vector model to develop the text specific vocabulary. Then used the model to predict answer. We used similarity score to get the best answer for a question.

Future Scope

The proposed solution is very basic solution there could be various improvements in the same. Firstly, we can improve it to answer the complex question which may have some context. Secondly, we can also improve the model and the training process. So, that it can work as generic model. Current solution takes the bag of word and not their order into matter which is important for various questions. Thus, that need to be improved.

References

[1] https://medium.com/lingvo-masino/question-and-answering-in-natural-language-processing-part-i-168f00291856

[2] https://towardsdatascience.com/building-a-question-answering-system-part-1-9388aadff507

[3] https://rare-technologies.com/sent2vec-an-unsupervised-approach-towards-learning-sentence-embeddings/

[4] http://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.XSYS1egzZPZ

[5] https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[6] https://arxiv.org/pdf/1301.3781.pdf