NLP Project

Proposal

Avinash Bagali - AE17B110
ae17b110@smail.iitm.ac.in

Advait Walsangikar - AE17B101
ae17b101@smail.iitm.ac.in

April 20, 2021

1 Limitations of the Vector Space Model

- The vector space model does not account for the meanings and relations between words based on the meanings. This will lead to retrieval of irrelevant docs in the case of polysemy and the failure to retrieve relevant docs in the case of synonymy.
- The sequence of words in Natural Language is of great importance to determine the relevance. The vector space model does not have any way of encoding the sequence of words and just uses a naive bag of words approach to create the term-space in which the docs are represented. (The words are assumed orthogonal to each other)

2 Our Hypothesis

We propose to augment the vector space model with the following two different approaches which implement different methods to include word relatedness in document representation:

- Generalized Vector Space Model : We aim to use this to incorporate external knowledge by using word similarity techniques on the WordNet corpus.
- Latent Semantic Analysis : We aim to use this to incorporate introspective knowledge of latent concepts inside the documents from the Cranfield dataset itself.

Once we have these external and introspective representations of word similarity, we will look at ways to naturally combine the two approaches. We will then use the provided cranfield dataset to compare and evaluate the hybrid approach's performance w.r.t. each of the methods individually on common evaluation metrics.

3 Realization

- We will implement the GVSM [1] technique from scratch by using similarity scores from WordNet that will be obtained through python NLTK.
- As for LSA, we will use the open source python library 'gensim' to model the latent senses.

4 Evaluation Technique

4.1 Metric

As we already have a non-binary relevance score available with the Cranfield dataset, we will continue to use the nDCG score to measure the effectiveness of the candidate IR systems

4.2 Baselines

- We use the performance of the current model (Simple Vector Space Model) implementation as a baseline to compare the new IR systems and judge the improvement and effectiveness of each approach

References

1. A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness George Tsatsaronis and Vicky Panagiotopoul