SEARCHING WITH WORD EMBEDDINGS

CS410 PROJECT PROPOSAL

BYUNG IL CHOI (EMAIL: BICHOI2@ILLINOIS.EDU, CHOI@BYUNG.ORG)

CS 410 course introduces vector space for words that can be used for text retrievals. This vector space is spatially intuitive which can be redefined in various ways. In the recent decade, some models were introduced for word embeddings such as word2vec[1] (Tomas Mikolov et al.), GloVe[2](Jeffrey Pennington et al.), etc[3]; which could enhance the VSM.

So far, it's not clear how search engines incorporate theses word embeddings. But, since words can have similarities which are not revealed in queries or documents, it will be useful to adapt word embeddings and check similarity for searching documents instead of using exact matching[4]. For example, when a user is searching for documents related to Lynx, he/she may want to include search that has Bobcat(Lynx rufus). Since Bobcat is a subspecies of Lynx, similarity can be very high, and thus should be included in the search result. But, if bobcat's document doesn't have word Lynx in it, lots documents about bobcats can be possibly excluded. And, users would not want to disregard such possibilities.

If documents with similar terms can be included or considered, general users and search engine developers can benefit from this functionality of finding similiarity[5]. How queries and documents are transformed into VSM of word embeddings and how they are used to rank the documents is the main problem to explore in this project. GloVe and word2vec are both open-sourced and spaCy and gensim is strong candidates for the implementation, their pretrained data if eligible, and corpus from the CS410 projects or corpus used for other[6] research papers[7]. This project aims to see how relevant vectors can be clustered and how the use of Rocchio Feedback[8] could show vector clusters that can be used for search engines with the emphasis on using proposed VSM.

As proposed above, particular cases of searching will be demonstrated by showing how they are clustered or searched[9], such as the case of Lynx and bobcats or searches where similar terms would impact the search result. This will be demonstrated with benchmarks, and statistical analysis if possible.

[1]https://arxiv.org/abs/1301.3781

[2]http://nlp.stanford.edu/pubs/glove.pdf

[3]Proposal guide line: What kind of tool exist?

[4]Proposal guide line: What is the function of this tool?

[5]Proposal guide line: Who will benefit from this tool?

[6]The corpus for testing are not yet decided.

[7]Proposal guide line: What existing resource can you use?

[8]Proposal guide line: What techinque would you use?

[9]Project guid line: How will you demonstrate the usefulness of my tool

PROJECT PLAN

1st week: setting tools, plan for using particular VSM(from word embedding), collect data, trained data, or even train some

2nd∼3rd week: implement necessary modules that uses word embedding except Rocchio.

4th∼5th week: implement Rocchio Feedback for project VSM, and do some experimental searching as examples

6th week: find ways to quantitatively compare results and evaluate.

7th week: polish the evaluation, make presentation.

8th week: write a report that discusses project's result and evaluation.