CSE 291: Homework 2

October 19, 2024

Instructions

In this assignment, we will analyze two phrase mining papers and use one of them to mine and organize knowledge from unstructured text with phrases as the basic unit.

Please read these two papers before starting the assignment:

- Mining Quality Phrases from Massive Text Corpora [code]
- Automated Phrase Mining from Massive Text Corpora [code]

You should submit the final writeup and code for this assignment. Please submit your solution **by Nov 14, 11:59pm PT**. Submissions should be made on **gradescope**. Please complete homework **individually**. Please include the code of your solutions in the submission with a write-up describing how to run the code.

1. **Paper Discussion (30 points):**

Please read the above two papers before answering these questions.

(a) Please outline the claims in these two papers.

(b) What is the major problem when someone is going to apply SegPhrase to a new corpus? Is there any human effort?

(c) What is the motivation of AutoPhrase? Compared with SegPhrase, which parts do you believe are novel?

(d) Why do we want to evaluate the results following the pooling strategy? Think about how much human effort is required, if we are not using pooling.

(e) What are the drawbacks of these two papers? Do you see any limitations?

(f) Can we do better in order to address these limitations? Propose a few ideas and explain how these would address the limitations.

2. **Phrase Mining experiments (30 points):**

In this question, we will use **AutoPhrase**, a data-driven phrase extraction framework to extract phrases from DBLP corpus consisting of Computer Science publications. We will then capture distributed representations for those phrases we extracted. Please download the code from here.

(a) Use AutoPhrase to extract high quality phrases on DBLP. Steps to do are mentioned below:

(i) All running commands and configurations of AutoPhrase can be found in `auto_phrase.sh`.

(ii) Download the DBLP corpus. For this, change line 24 to `DEFAULT_TRAIN=${DATA_DIR}/EN/DBLP.txt`.

(iii) Execute `bash ./auto_phrase.sh`. This will download the corpus and extract phrases.

The AutoPhrase model will automatically extract high-quality phrases, which will be stored in `AutoPhrase/models/DBLP/AutoPhrase.txt` with their respective scores.
In your submission, please include three ranked phrase lists, i.e. (`AutoPhrase.txt`, `AutoPhrase_multi-words.txt`, and `AutoPhrase_single-word.txt`).

(b) Did you find any phrases with abnormal scores (e.g. non-phrase with a high score or good phrase with a low score)? Do they show a systematic pattern? What can be the possible reason behind it and how to improve the algorithm to avoid such mistakes?

(c) Apply Word2Vec (d=100) on segmented corpus to capture semantic meaning of phrases. Steps to do are mentioned below:

(i) To modify the corpus with segmented phrases, we use the script `phrasal_segmentation.sh`.

(ii) Modify line-14 to `TEXT_TO_SEG=${TEXT_TO_SEG:- ${DATA_DIR}/EN/DBLP.txt}`

(iii) Execute `bash phrasal_segmentation.sh`. This segments the phrases and stores at `AutoPhrase/models/DBLP/segmentation.txt`. An example sentence in segmented corpus looks like:
"*< phrase >An overview< /phrase> is presented of the use of <phrase>spatial data structures< /phrase> in <phrase>spatial databases< /phrase>*".

(iv) Write your own script to parse the segmented corpus and convert each phrase into a single token.

(v) Run Word2Vec on the segmented corpus with each phrase as a single token.

(d) Run a clustering method (e.g. K-Means, GMM, etc.) on quality phrases with their semantic representation, try to name each category. List down 20 phrases from each cluster in your report. You can set the number of clusters to k=6. Do this using two different clustering methods and qualitatively compare the results between the methods.