Extractive summarisation of biomedical research articles using TextRank, WordRank, and a hybrid approach

Kristina Levina
Course code: 732A81
LiU-ID: krile102

Abstract

This project aims at building a tool to generate a summary of a biomedical research manuscript automatically based on the main text. To this end, extractive summarisation is employed. The motivation behind choosing extractive summarisation is, in essense, the neccessety to preserve key sentences from the main text. Extractive summarisation will retrieve sentences based on their importance without rephrasing them, thereby excluding misinterpretations. The meaning preservation is crucial for scientific texts. In this project, PubMed dataset is used. A subset of 100 articles and their abstracts have been manually looked through and filtered so that articles and abstracts have similar characteristincs in terms of relative number of sentences of the abstract to the main text. As a result, 24 articles and their human-written summaries (abstracts) were selected. Three algorithms were chosen for extractive summarisation: TextRank, WordRank, and their combination. Their performance was assessed using ROUGE (recall), BLEU (precision), and F1 score. The obtained results show that Y better suits for the considered task.

1 Introduction

With increasing volume of published articles in medical research, it becomes increasingly difficult for doctors, medical staff, and public health officials to stay updated. Sometimes, a

2 Theory

2.1 PageRank

2.2 ROUGE

2.3 BLEU

3 Data

Data are taken from the paper by Cohan et al. (2018). This dataset contains a large collection (100,000) of scientific articles from the biomedical domain (OpenAccess PubMed articles). Data are hosted in GitHub. Each article has the following fields:

```
{
  'article_id': str,
  'abstract_text': List[str],
  'article_text': List[str],
  'section_names': List[str],
  'sections': List[List[str]]
}
```

For this project, I have looked through 100 articles and considered only articles meeting the following assumptions: First, the provided abstract should be between 8%–12% of the main text in terms of number of sentences. This is to ensure similar conditions for generating summaries. Second, the number of sentences of the main manuscript text should be larger than 50 to meet the objective of long document summarisation. Out of 100 articles, only 24 met this conditions.

An example of the beginning of one chosen article is as follows:

```
'["anxiety affects quality of life in those living with parkinson ś disease ( pd ) more so than overall cognitive status , motor deficits , apathy , and depression [ 13 ] .", álthough anxiety and depression are often related and coexist in pd patients , recent research suggests that anxiety rather than depression is the most prominent and prevalent mood disorder in pd [ 5 , 6 ] . yet ,;
```

The beginning of the corresponding summary is as follows:

```
'["<S> research on the implications of anxiety in parkinson ś disease ( pd ) has been neglected despite its prevalence in nearly 50% of patients and its negative impact on quality of life . </S>", ć<S> previous reports have noted
```

`that neuropsychiatric symptoms impair cognitive performance in pd patients`

The data have been already tokenised. From the abstract text, <S> and </S> tags were removed. Further preprocessing included stop word removal, lemmatisation, and non-alphabetic characters' removal. This preprocessing has been done using the Spacy language model. Stop words were removed to avoid sentence ranking based on common and frequent words rather than important words. Lemmatisation was used to treat same words in an exactly same manner. Non-alphabetic characters were removed to avoid their influence on the ranking results. Punctuation and numerals should not affect the sentence importance.

4 Method

4.1 TextRank

4.2 WordRank

4.3 Hybrid

4.4 Evaluation

5 Results

The ROUGE_1, BLEU_1, and F1_1 scores respectively denote ROUGE, BLEU, and F1 scores of unigrams between the human-written and generated summaries. The results are shown in Fig. 1 and Tables 1 and 2. Both TextRank (36.0) and Hybrid80 (35.9) yield high mean F1_1 scores relative to other algorithms, but considering margin of error, the estimation by Hybrid80 is slightly more robust (35.9 $\pm$ 3.3 versus 36.0 $\pm$ 3.7). Both TextRank (27.0) and Hybrid80 (26.8) yield high mean BLEU_1 scores, but considering margin of error, again the estimation by Hybrid80 is slightly more robust (26.8 $\pm$ 3.2 versus 27.0 $\pm$ 3.7). WordRank yields the highest ROUGE_1 score (60.1 $\pm$ 4.1).

The ROUGE_2, BLEU_2, and F1_2 scores respectively denote the ROUGE, BLEU, and F1 scores of bigrams between the human-written and generated summaries. The results are shown in Fig. 2 and Tables 1 and 2. The ROUGE_3, BLEU_3, and F1_3 scores respectively denote the ROUGE, BLEU, and F1 scores of trigrams between the human-written and generated summaries. The results are shown in Fig. 3 and Tables 1 and 2. TextRank shows the highest values of ROUGE_2 and ROUGE_3 (21.2 $\pm$ 4.9 and 10 $\pm$ 3.7, respectively), BLEU_2 and BLUE_3 (10.1 $\pm$ 2.7 and 5 $\pm$ 2.0, respectively), and F1_2 and F1_3 (13.4 $\pm$ 3.2 and 6.5 $\pm$ 2.5, respectively).

The total running times of all four algorithms for summarisation of 24 data samples is shown in Table 3. TextRank (1 min) performs 16 times faster than other algorithms (16 min).

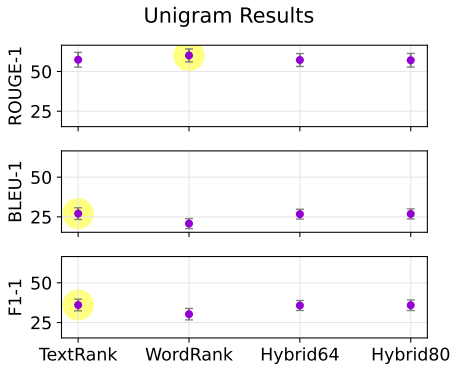

Figure 1: ROUGE, BLEU, and F1 score of unigrams between the human-written and generated summaries.

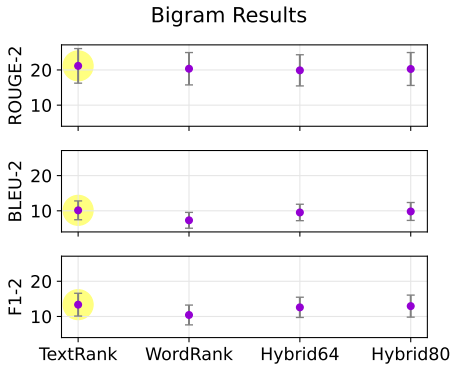

Figure 2: ROUGE, BLEU, and F1 score of bigrams between the human-written and generated summaries.

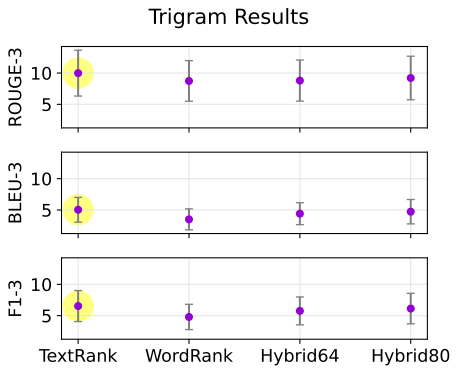

Figure 3: ROUGE, BLEU, and F1 score of trigrams between the human-written and generated summaries.

	TextRank	WordRank	Hybrid64	Hybrid80
ROUGE_1	57.4	**60.1**	57.2	57.1
ROUGE_2	**21.2**	20.3	19.9	20.3
ROUGE_3	**10.0**	8.7	8.8	9.2
BLEU_1	**27.0**	20.8	26.7	26.8
BLEU_2	**10.1**	7.3	9.5	9.8
BLEU_3	**5.0**	3.5	4.4	4.7
F1_1	**36.0**	30.3	35.7	35.9
F1_2	**13.4**	10.4	12.6	12.9
F1_3	**6.5**	4.8	5.8	6.1

Table 1: Mean values of the ROUGE, BLEU, and F1 metrics for unigrams, bigrams, and trigrams yielded by TextRank, WordRank, Hybrid64, and Hybrid80 methods.

	TextRank	WordRank	Hybrid64	Hybrid80
ROUGE_1	4.7	4.1	4.1	4.3
ROUGE_2	4.9	4.6	4.4	4.7
ROUGE_3	3.7	3.2	3.3	3.5
BLEU_1	3.7	3.2	3.1	3.2
BLEU_2	2.7	2.2	2.3	2.5
BLEU_3	2.0	1.7	1.8	1.9
F1_1	3.7	3.6	3.2	3.3
F1_2	3.2	2.8	2.9	3.1
F1_3	2.5	2.0	2.2	2.4

Table 2: Margins of error values of the ROUGE, BLEU, and F1 metrics for unigrams, bigrams, and trigrams yielded by TextRank, WordRank, Hybrid64, and Hybrid80 methods.

	TextRank	WordRank	Hybrid64	Hybrid80
Running time (min)	1	16	16	16

Table 3: Running time (min) of all four algorithms considered in this project.

6 Discussion

7 Conclusion

8 Preamble

Table 4 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by Gusfield (1997). You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (Gusfield, 1997). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. Gusfield, 1997).

References

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Output	natbib command	Old ACL-style command
(Cooley and Tukey, 1965)	`\citep`	`\cite`
Cooley and Tukey, 1965	`\citealp`	no equivalent
Cooley and Tukey (1965)	`\citet`	`\newcite`
(1965)	`\citeyearpar`	`\shortcite`
Cooley and Tukey's (1965)	`\citeposs`	no equivalent
(FFT; Cooley and Tukey, 1965)	`\citep[FFT;][]`	no equivalent

Table 4: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.