Report on various tokenization strategies for Biomedical Information Retrieval

Mallipeddi Akshay,
201301216@daiict.ac.in
DA-IICT

Problem Statement:- **Implementing various tokenization strategies for Biomedical Information Retrieval.**

Abstract:-The main challenge involved when dealing with biological names in biomedical text, is appropriate tokenization for biomedical information retrieval.In this work, I have conducted a systematic evaluation of different methods on the available TREC(2014) biomedical text collections for ad-hoc document retrieval.The main idea is that different types of text require different tokenization methods.

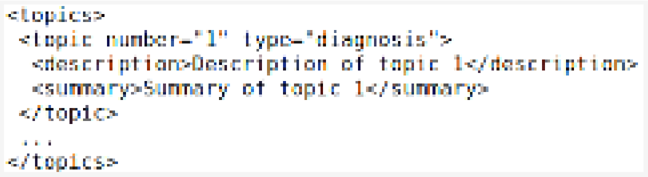
I.RESULTS

I used TREC 2014 CDS data. There are a total of 30 topics in TREC 2014 CDS dataset. The topics in TREC 2014 are medical case narratives created by expert topic developers.The information may describe patient's medical history, the patient's current symptoms, tests performed by a physician to diagnose the patient's condition ,the patient's eventual diagnosis etc. The topics are divided as shown in table-1.

TABLE-1 :-Table showing TREC 2014 topics descriptions.

Type	**Generic Clinical Question**	**Number of Topics**
Diagnosis	What is the patient's diagnosis?	10
Test	What tests should the patient receive?	10
Treatment	How should the patient be treated?	10

- Format of the topics:-



```
<topics>
 <topic number="1" type="diagnosis">
  <description>Description of topic 1</description>
  <summary>Summary of topic 1</summary>
 </topic>
 ...
</topics>
```

➔ **First trial run:-**In the first run,I considered a sample data from the TREC 2014 CDS data,which contains approximately 50,000 documents and 30 topics.The results are as follows:-

TABLE-2*:-Showing the results of first trial run

		Normalization methods		
Stemmer used		**H-NORM**	**S-NORM**	**J-NORM**
Porter	MAP	**0.0085**	**0.0086**	**0.0091**
Lovins	MAP	**0.0065**	**0.0062**	**0.0059**
S	MAP	**0.0090**	**0.0090**	**0.0093**

The baseline value used was **MAP:** 0.0075 (Default Tokenization)

The results in the first trial run were very bad and no conclusion could be drawn on the methods used because even the data set used was very limited.Based on the results in table-2 , S stemmer performed better, as it gave MAP more than the baseline value.Porter stemmer also performed better while Lovins stemmer was below the baseline.

➔ **Second final run:-** In the second run, I considered the whole TREC 2014 CDS data which contains approximately 7,50,000 documents and 30 topics.The results are as follows:-

TABLE-3*:-Showing the results of second final run

		Normalization methods		
Stemmer used		**H-NORM**	**S-NORM**	**J-NORM**
Porter	MAP	**0.0953**	**0.0954**	**0.0940**
Lovins	MAP	**0.0673**	**0.0632**	**0.0613**
S	MAP	**0.1001**	**0.1000**	**0.0976**

The baseline values used was **MAP:** 0.0954 (Default Tokenization)

The results in the final run were better than the first run and S stemmer performed better, as it gave MAP more than the baseline value.And Porter stemmer performed equally well with the baseline, while Lovins stemmer was below the baseline.There was more than **10x** improvement in my results.

II.SUMMARY AND CONCLUSION

After the first sample run, I improvised my implementation in the code.In second final run, I used a larger data set consisting of 7,50,000 documents.There was a significant improvement in my results when compared to first sample run.

Because of the irregular variants of names in biomedical text,tokenization is an important preprocessing step in biomedical information retrieval.In this report, the evaluation was conducted on available TREC biomedical information retrieval test collections.Results show that tokenizations could improve the performance.

REFERENCES

PAPER REFERENCES:-
[1] http://sifaka.cs.uiuc.edu/czhai/pub/ir-tok.pdf.
[2] http://terrierteam.dcs.gla.ac.uk/publications/ounis06terrier-osir.pdf
OTHER REFERENCES:-
[1] [Lovins 1968] Lovins, J. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics.
[2] [Porter 1980] Porter, M. F. (1997). An algorithm for suffix stripping. Program, 14(3).

***Red indicates below the baseline, green indicates greater than or equal to the baseline value.**