Passage retrieval for tweet contextualization at INEX 2012

Ayan Bandyopadhyay[1], Sukomal Pal[2], Mandar Mitra[1], Prasenjit Majumder[3], and Kripabandhu Ghosh[1]

[1] Indian Statistical Institute (Kolkata)
[2] ISM Dhanbad
[3] DAIICT Gandhinagar

Abstract. This paper describes some preliminary results obtained by treating the tweet contextualization task as a passage retrieval task. Each tweet was submitted as a query to the Indri 5.2 search engine after some preprocessing. Either paragraphs or sentences were retrieved in response to a query. Passages retrieved from the same document were concatenated. This approach does not work very well in terms of informativeness: the best of our runs was ranked 23rd out of 33 runs. Further exploration of ways to improve effectiveness is needed.

1 Introduction

The INEX tweet contextualization task at CLEF 2012 is a new task. The aim of this task is to provide some *context* for a given topic tweet [1]. For this task, the context consists of a passage of at most 500 words extracted from a cleaned dump of the English Wikipedia. It is intended to provide some background information that will help a user to better understand the tweet.

In this report, we describe our very preliminary attempts at tweet contextualization. To begin with, we have simply treated contextualization as a passage retrieval task. After some preprocessing, the textual content of a tweet is used as a query to retrieve paragraphs or sentences from the Wikipedia corpus. If multiple passages are retrieved from the same article, they are merged together.

Related work is discussed in the next section (Section 2). Our approach is described in Section 3. Section 4 presents our results and discusses some obvious limitations of our approach. Our plans for further experimentation are outlined in Section 5.

2 Related Works

The tweet contextualization task is introduced by INEX at CLEF 2012. Bellot et al. [1] describes overall report of the INEX 2011. This task is involved with tweet. Tweets are treated as topics here. `http://twitter.com` is one of the

[1] http://twitter.com

popular site of microblogging. Miles Efron [2] reveals an overview of microblog and behavior surrounding it e.g microblog retrieval, entity search, sentiment analysis. According to the passage retrieval point of view Robertson et al. [4] says why we should not use liner equation to merge passages retrieved form the same document. After the passage retrieval answer construction is the next part. Summarization and framing answer has a very important role. Salton et al. [5] says about automatic text summarization using Intra-document passage links. recent text summarization survey by Ani Nenkova et al. [3] helps to know a elaborate description of text summarization.

3 Experimental Setup

We divided each page in the corpus into separate paragraphs using the `<p>` and `</p>` tags. All text contained between these tags was indexed. Each paragraph was also split further into sentences using periods (.), question marks (?) and exclamation marks (!) as sentence delimiters. Stopwords were removed, and Porter's stemmer was used. Some statistics about the processed corpus are given below. Since any period (.) was regarded as an end-of-sentence marker, abbrevi-

Table 1. Comparison of paragraph and sentence level indexing and corpus statistics

	Paragraph Level	**Sentence Level**
Number of paragraph/sentence	8,388,955	26,039,270
Unique terms	2,878,685	2,876,680
Total terms	333,522,647	333,697,767

ations were also split up when the text was indexed at the sentence level. This is why the number of terms (total and distinct) is somewhat different when the same text is indexed at two levels of granularity.

The topic tweets (1142 in all) were provided in two formats: JSON and simple text. We used the simple text format. Stopwords, URLs, the name of the tweeting authority, and the text "RT" were removed. The remaining words were stemmed using Porter's stemmer. Using these preprocessed tweets as queries, and Indri 5.2 as the search engine, we retrieved in turn paragraphs and sentences for each query tweet. A total of three runs were submitted. Details about these runs are given below.

Run1 — Top 50 returned paragraphs were submitted. If multiple paragraphs were retrieved from a document, then those paragraphs were concatenated. The similarity scores of individual paragraphs were simply added together to obtain the score of the concatenated result. Any paragraph longer than 500 words (including those obtained by concatenation) was truncated to the first 500 words.

Run2 — Same as the Run1, except that we started with the top 100 sentences for each query.
Run3 — Same as the Run1, except that the top 100 paragraphs were used.

4 Results

Submitted summaries were evaluated according to their informativeness and readability. Table 2 compares the performance of our submitted runs (Run1, Run2, Run3) with the best run at INEX 2012.

Table 2. Comparison of submitted runs and the best run at INEX 2012

Run Name	Run ID	Rank (out of 33)	Unigram	Bigram	Skip Bigram
Run1	149	26	0.9059	0.9916	0.9916
Run2	150	23	0.9052	0.9871	0.9868
Run3	151	33	0.9223	0.9985	0.9988
Best	178	1	0.7734	0.8616	0.8623

It is clear that the overly simplistic approach that we tried did not perform well with regard to informativeness (they did obtain good readability, however). Out of these runs, the sentence-level run performs best. A number of obvious drawbacks need to be rectified.

- When multiple paragraphs / sentences from a single document are concatenated, their similarity scores are simply added together. This may lead to poor ranking [4]. The score of the combined passage needs to be calculated more carefully.
- We need to be more careful when splitting a paragraph into sentences. In particular, periods used with acronyms and abbreviations should not result in sentence breaks.
- Retrieved passages are arbitrarily truncated at 500 words, without checking for sentence boundaries.

5 Conclusion

As mentioned in Section 2, a number of query-oriented summarisation approaches have been proposed in earlier work. In future work, we intend to explore how these may be applied to the contextualization task. Also, given that the "topics" or tweets are short to start with (at most 140 characters, many of which are taken up by URLs), query expansion is likely to be beneficial. We also hope to investigate query expansion / reformulation techniques as ways to improve informativeness of the generated summaries.

References

1. Patrice Bellot, Timothy Chappell, Antoine Doucet, Shlomo Geva, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Monica Landoni, Maarten Marx, Véronique Moriceau, Josiane Mothe, G. Ramírez, Mark Sanderson, Eric SanJuan, Falk Scholer, Xavier Tannier, Martin Theobald, M. Trappett, Andrew Trotman, and Q. Wang. Report on INEX 2011. *SIGIR Forum*, 46(1):33–42, 2012.
2. Miles Efron. Information search and retrieval in microblogs. *JASIST*, 62(6):996–1008, 2011.
3. Ani Nenkova and Kathleen McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233, 2011.
4. Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Simple BM25 extension to multiple weighted fields. In *Proc. CIKM*, pages 42–49. ACM, 2004.
5. Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Inf. Process. Manage.*, 33(2):193–207, 1997.