A Pipeline Tweet Contextualization System at INEX 2013

Khaled Hossain Ansary, Anh Tuan Tran, Nam Khanh Tran

Leibniz Universität Hannover / Forschungszentrum L3S, Hannover, Germany
ansary@L3S.de, ttran@L3S.de, ntran@L3S.de

Abstract. This article describes a pipeline system and preliminary results for Tweet Contextualization at INEX 2013. The system consists of three steps: tweet analysis, passage retrieval and summarization. For each tweet, key phrases are first extracted by making use of ArkTweet toolkit and employing several heuristics. They are then submitted as queries to Indri search engine to retrieve relevant passages. Finally, a multi-document summarization system (MEAD) is used to generate the output document with a limit of 500 words. The preliminary results show that the approach does not work well where our run was ranked 22nd out of 24 runs. We discuss our observations for these results and some further possible improvements.

1 Introduction

The tweet contextualization task was first launched at INEX in 2011. The task is related with the tweets[1] which represent as short message around 140 characters. The aim of tweet contextualization is to provide automatically information as a readable summary that explains the tweet. The summary does not exceed 500 words and extracted from a cleaned dump of the English Wikipedia. The evaluation of the summaries has done by the INEX organizers to considering both informativeness and readability.

The committee of INEX has been collected about 598 tweets in English from Twitter. Tweets were selected among informative account (for example, @CNN, @TennisTweets, @PeopleMag, @science..), in order to avoid purely personal tweets that could not be contextualized. In this article, we present the experiments carried out as part of the participation in INEX 2013. We describe a pipeline system where first extracts phrases from the tweets by using ArkTweet toolkit and some heuristics; then retrieves relevant documents for these phrases from Wikipedia before summarizing those with MEAD toolkit.

2 Related Work

There has been some studies done for this task. While [1, 2] presents the improvement of the question answering techniques using information retrieval (IR), [3]

[1] https://twitter.com/

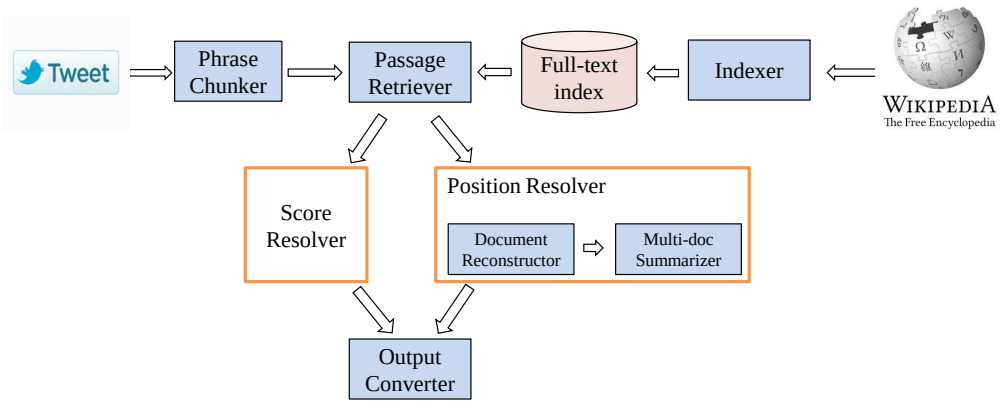

Fig. 1: Overview of the pipeline workflow

describes a hybrid tweet contextualization system using IR and automatic summarization. They used Nutch architecture and TF-IDF based sentence ranking and sentence extracting techniques for automatic summarization. An approach based on the mapping of source documents in a reduced semantic space is proposed by [4]. They estimated the words from the semantic space via a latent dirichlet allocation (LDA) algorithm. [5] developed and tested a statistical word stemmer which used by the CORTEX to preprocess input texts and generate readable summary. [6] describes a sentence retrieval technique which applied three methodologies: i) language modeling score, ii) relevance modeling score and ii) topical relevance modeling score.

Text summarization has been well-studied through several work in artificial intelligence communities, especially text mining and information retrieval. Among them, MEAD [7] is a publicly available toolkit for multi-document summarization, which generates summaries using cluster centroids produced by topic detection and tracking system.

3 A Pipeline Tweet Contextualization System

The system pipeline is described as shown in Figure 1. It consists of three components: Phrase Chunker, Passage Retriever and Summarizer.

3.1 Phrase Chunker

As shown in the system workflow, the first step is to retrieve passages from Wikipedia registered articles given a tweet of interest. As in the traditional retrieval approach, we initially used words presented in the tweet to retrieve the

relevant passages from Wikipedia. However, we observed an acceptably low performance when using original words to query the indices. This is attribute to the highly noisy nature of tweet contents, where the key phrase often mixed with non-content words such as emoticons, over-used punctuations, etc.. In addition, users employ several ad-hoc formats that are hardly found elsewhere when posting tweets. They can use hashtags (a single word starting with '#') to provide implicit context of the tweet, or use the at (@) symbol to tag other twitter accounts in the content. In many cases, words are intentionally modified, such as repeating vowels to express emotions (e.g. 'so cooooooooool this show was !! :=)'), etc. Such writing styles leads to many irrelevant results and propagates the noise to the next step.

To accommodate the passage retrieval, we tuned our phrase chunker so as to detect and extract key phrases that are more informative than the others from the tweet content. We used ArkTweet toolkit [8] to tokenize the tweet content, and to annotate each token with an adjusted Part-of-speech tags. Apart from Penn TreeBank tagset, ArkTweet introduces a number of specialized tags in Twitter domain, such as hashtag (#), at-mention (@), discourse marker (∼) to indicate the continuation of message across multiple tweets such as Retweets, URL (U), or emoticon (E). Detailed references can be found at `http://www.ark.cs.cmu.edu/TweetNLP/annot_guidelines.pdf`.

After tokenizing the tweet, we employed several heuristics to detect the key phrases as overlapping consecutive tokens. For example, we restricted that a key phrase cannot be a mix of hashtags and other words, or we skipped phrases that contain no Penn TreeBank tags. The chunker iteratively generates all n-grams, where n varies from 1 to 5. For each n-gram, it checks against each of the heuristic. We applied a dynamic programming approach to make sure two heuristics is not checked again on the subsumed grams.

3.2 Passage Retriever

We retrieved relevant Wikipedia articles for each tweet via the provided API of the track. The methodology adopted by us can be described as follows. Each extracted phrases for a given tweet was submitted as a query to Indri search engine and we obtained three different files for our purpose in the following format:

- The "docid" files contain the sentences which we retrieved from the API. The sentences which collected from the same document are merged, stored and then used as input for the summarization component.
- The docid and the phrase rank of the corresponding sentences are stored into the "docid.id" file
- The docid and the resultant scores stored into the file "docid.score". The average scores calculated for the same document phrase id. These scores use to submit as a part of our run.

3.3 Summarizer

We make use of MEAD toolkit[2] for this component. MEAD is a multi-document summarization system proposed by Radev et al. [7] implemented centroid-based approach and is then enhanced with various of features later. We adapted the system with various parameter settings including position, similarity with the first sentence, centroid, query-based features, MEAD-cosine similarity routine re-ranker with threshold value = 0.7 and enidf IDF database.

4 Results

The output summaries were evaluated according to their informativeness and readability. Table 1 and Table 2 compare the performance of our submitted run with the best one at INEX 2013 in terms of informativeness and readability, respectively.

RunID	Rank	Unigram	Bigram	Skip Bigram
266	22	0.9059	0.9824	0.9835
256	1	0.8861	0.881	0.782

Table 1: Comparison of submitted runs and the best run in terms of informativeness score at INEX 2013

RunID	Relevancy(T)	Non redundancy(R)	Soundness(A)	Syntax(S)	Mean Average
266	25.92%	25.08%	25.92%	25.92%	25.64%
275	76.64%	67.30%	74.52%	75.50%	72.44%

Table 2: Comparison of submitted runs and the best run in terms of readability score at INEX 2013

We observed that the phrases extracted from tweets contains some unexpected noises which need to be cleaner. A heuristics-based approach relies heavily on a small set of tweets to be scrutinized, and it is difficult to generalize in the arbitrary domains of tweets. This can affect the retriever components where irrelevant sentences are retrieved as results of noisy phrases. Another observation is that creating the documents by merging retrieved sentences and treating them as input for MEAD toolkit can make these documents less readable. One key point in MEAD summarization is the assumption of relatedness between

[2] We use the latest version MEAD 3.12 published at http://www.summarization.com/mead/

sentences in one documents, and build a graph of inter-references from such relatedness. This does not really fit to the re-construction of tweets as conducted in the first two steps of the pipeline. Nevertheless, this observation calls for future approaches in text summarization, where sentences are less coupled and thus should be modeled less dependently

5 Conclusion

The pipeline system has been developed as part of the participation in the Tweet Contextualization track of INEX 2013. The system was evaluated by using the evaluation metrics provided by the committees with reasonable results with its initial implementation.

Further works will be motivated towards improving the performance of the system by enhancing the quality of phrases from tweets, considering semantic similarity for retrieving relevant documents.

References

1. lvaro Rodrigo, Prez-iglesias, J., Peas, A., Garrido, G., Araujo, L.: A question answering system based on information retrieval and validation (2010)
2. Schiffman, B., Mckeown, K.R., Grishman, R.: Question answering using integrated information retrieval and information extraction. In: in Proceedings of HLT/NAACL. (2007)
3. Bhaskar, P., B.S.: A hybrid tweet contextualization system using ir and summarization. In: in Proceedings of INEX 2012. (2012)
4. Morchid, M., Linares, G.: A semantic space for tweets contextualization. In: in Proceedings of INEX 2012. (2012)
5. Torres-Moreno, J.M., Velazquez-Morales, P.: Two statistical summarizers at inex 2012. In: in Proceedings of INEX 2012. (2012)
6. Debasis Ganguly, J.L., Jones, G.J.F.: Exploring sentence retrieval for tweet contextualization. In: in Proceedings of INEX 2012. (2012)
7. Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., Zhang, Z.: MEAD — A platform for multidocument multilingual text summarization. In: Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal (2004)
8. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 42–47