IRIT at INEX 2013: Tweet Contextualization Track

Liana Ermakova, Josiane Mothe

Institut de Recherche en Informatique de Toulouse
118 Route de Narbonne, 31062 Toulouse Cedex 9, France
liana.ermakova.87@gmail.com, josiane.mothe@irit.fr

Abstract. The paper presents IRIT's approach used at INEX Tweet Contextualization Track 2013. Systems had to provide a context to a tweet. This year we further modified our approach presented at INEX 2011 and 2012 underlain by the product of scores based on hashtag processing, TF-IDF cosine similarity measure enriched by smoothing from local context and document beginning, named entity recognition and part-of-speech weighting. We assumed that relevant sentences come from relevant documents therefore we multiply sentence score by document relevance. We also used generalized POS (e.g. we merge regular adverbs, superlative and comparative into a single adverb group). We introduced sentence quality measure based on Flesch reading ease test, lexical diversity, meaningful word ratio and punctuation ratio. Our approach was ranked first, second and third over 24 runs submitted by all participants on different reference pools according to informativeness evaluation. At the same time it obtained the best readability score.

Keywords: Information retrieval, tweet contextualization, summarization, sentence extraction, readability.

1 Introduction

Twitter is an online social network and microblogging that enables to send and read text messages up to 140 characters [1]. In March 2013, the Twitter got more than 200 million active users how write more that 400 million tweet every day [2]. However, tweets are quite short and they may contain information that is not understandable to a user without some context. Therefore, providing concise coherent context seems to be helpful. INEX Tweet Contextualization Track aims to evaluate systems providing context to a tweet [3]. The context should be a readable summary up to 500 words extracted from a dump of the Wikipedia from November 2012. This year two languages were used: English and Spanish. English query set included 598 tweets in English, while Spanish subtrack was based on 354 personal tweets in Spanish.

The paper presents IRIT's approach used at INEX Tweet Contextualization Track 2013. We consider tweet contextualization task as multi-document extractive summarization. This year we further modified our approach presented at INEX 2011 [4] and 2012 [5] underlain by the product of scores based on hashtag processing, TF-IDF cosine similarity measure enriched by smoothing from local context and document

adfa, p. 1, 2011.

beginning, named entity (NE) recognition and part-of-speech (POS) weighting. We assumed that relevant sentences come from relevant documents therefore we multiply sentence score by document relevance. We also used generalized POS (e.g. we merge regular adverbs, superlative and comparative into a single adverb group). We introduced sentence quality measure based on Flesch reading ease test, lexical diversity, meaningful word ratio and punctuation ratio.

The paper is organized as follows. Firstly, we recall the principles of the 2011-2012 system we developed and describe the modifications we made. Then, we present the results and discuss them. Future development description concludes the paper.

2 Method Description

2.1 Preprocessing

Preprocessing includes several steps.

Firstly, we treat tweets themselves, i.e. special symbols like hashtags and replies.

The hashtag symbol # "is used to mark keywords or topics in a Tweet. It was created organically by Twitter users as a way to categorize messages" and facilitate a search [6]. Hashtags are inserted before relevant keywords or phrases anywhere in tweets. Popular hashtags often represents trending topics. Bearing it in mind, we put higher weight to words occurring in hashtags. Usually key phrases are marked as a single hashtag. Thus, we split hashtags by capitalized letters.

Moreover, important information may be found in @replies, e.g. when a user reply to the post of a politician or other famous person. "An @reply is any update posted by clicking the "Reply" button on a Tweet" [7]. Since people may use their names as Twitter accounts we treat them analogically to hashtags, i.e. they are split by capitalized letters.

We assume that relevant sentence come from relevant documents, so we applied a search engine to find them. We use the tweet as a query. We choose the Terrier platform [8], an open-source search engine developed by the School of Computing Science, University of Glasgow. It implements various weighting and retrieval models and allows stemming and blind relevance feedback. Terrier is suitable for different languages including English and Spanish. We choose Porter stemmer [9] for the English subtrack and Snowball stemmer [10] for the Spanish one.

The next step is to parse tweets and retrieved texts. For the English subtrack we applied Stanford CoreNLP which integrates such tools as POS tagger [11], named entity recognizer [12], parser and the co-reference resolution system. It uses the Penn Treebank tag set [13]. For the Spanish subtask we integrated Tree Tagger [14] and Apache OpenNLP [15]. Tree Tagger was used for lemmatization and POS tagging, while sentence detector, named entity recognition were performed by OpenNLP.

Then, we merged annotation obtained by parsers and Wikipedia tagging.

2.2 Searching for Relevant Sentences

We modified the extraction component developed for INEX 2011-2012. The general idea of the approach 2011 was to compute similarity between the query and sentences and to retrieve the most similar passages.

We model a sentence as a set of vectors. The first vector represents the tokens occurred within the sentence (unigram representation). Tokens are associated with lemmas. A lemma has the following features: POS, frequency and IDF. The second vector corresponds to bigrams. In both vector representation stop-words are retrieved. However, functional words, such as conjunctions, prepositions and determiners, are not taken into account in the unigram representation. NE comparison is hypothesized to be very efficient for contextualizing tweets about news. Therefore, the third vector refers to found named entities. Thereby, the same token may appear in several vectors.

For unigram and bigram vectors, we computed cosine, Jaccard and dice similarity measures, between a sentence and a target tweet. NE vectors are treated in the following way:

$$NE_{COEF} = \frac{NE_{common} + NE_{weight}}{NE_{query} + 1} \quad (1)$$

where NE_{weight} is floating point parameter given by a user (by default it is equal to 1.0), NE_{common} is the number of NE appearing in both query and sentence, NE_{query} is the number of NE appearing in the query.

Each sentence has a set of attributes, e.g. which section it belongs to, whether it is a title or header, whether it has personal verbs etc.

We introduced an algorithm for smoothing from the local context. We assumed that the importance of the context reduces as the distance increases. Thus, the nearest sentences should produce more effect on the target sentence sense than others. For sentences with the distance greater than k this coefficient was zero. The total of all weights should be equal to one. The system allows taking into account k neighboring sentences with the weights depending on their remoteness from the target sentence.

Moreover, this year we added smoothing from document beginning. Wikipedia abstracts contain the summary of the entire paper; therefore they can be also used for smoothing.

In 2013, we did not applied anaphora resolution since it did not improve much our system according to evaluation in 2012 [5]. Neither we used sentence reordering as it was not evaluated.

We assumed that relevant sentences come from relevant documents therefore we multiply sentence score by document relevance or/and by inverted document rank. We tried to use generalized POS (e.g. we merge regular adverbs, superlative and comparative into a single adverb group).

2.3 Improving Readability

We introduced sentence quality measure based on the product of the Flesch reading ease test [16], lexical diversity, meaningful word ratio and punctuation score.

Flesch Reading Ease test is a readability test designed to indicate comprehension difficulty when reading a passage (higher scores corresponds to texts that are easier to read):

$$Flesh = 206.835 - \frac{(1.015 * TokenCount)}{SentenceCount} - \frac{(84.6 * SyllableCount)}{TokenCount} \quad (2)$$

We defined lexical diversity as the number of different lemmas used within a sentence divided by the total number of tokens in this sentence.

Analogically, meaningful word ration is the number of non-stop words within a sentence divided by the total number of tokens in this sentence.

Punctuation score is estimated by the formula:

$$PunctScore = 1 - \frac{PunctuationMarkCount}{TokenCount} \quad (3)$$

In order to treat redundancy each sentence was mapped into a noun set. These sets were compared pairwise and if the normalized intersection was greater than a predefined threshold the sentences were rejected.

3 Evaluation

Summaries in English were evaluated according to their informativeness and readability [3]. Informativeness was estimated as the overlap of a summary with 3 pools of relevant passages:

1. Prior set (PRIOR) of relevant pages selected by organizers. PRIOR included 40 tweets, i.e. 380 passages or 11 523 tokens.
2. Pool selection (POOL) of most relevant passages from participant submissions for 45 selected tweets. POOL contained 1 760 passages, i.e. 58 035 tokens.
3. All relevant texts (ALL) merged together with extra passages from a random pool of 10 tweets. ALL is based on 70 tweets having 2 378 relevant passages of 77 043 tokens.

As in previous years, the lexical overlap between a summary and a pool was estimated in three terms: *Unigrams, Bigrams and Skip bigrams* representing the proportion of shared unigrams, bigrams and bigrams with gaps of two tokens respectively. Official ranking was based on decreasing score of divergence with ALL estimated by skip bigrams.

At the English subtrack we submitted 3 runs differing by sentence quality score and smoothing.

Our best run 275 was ranked first, second and third over 24 runs submitted by all participants on the PRIOR, POOL and ALL respectively (see Table 1; IRIT's runs are

set off in bold). It means that our best run is composed from the sentence of the most relevant documents. Among automatic runs our method was classified first (PRIOR and POOL) and second (ALL): the run 256 is marked as manual. It is also obvious that ranking is sensitive to not only pool selection, but also choice of divergence. According to bigrams and skip bigrams our best run is 275, while according to unigrams the best run is 273. We can also see than the runs 273 and 274 are quite close. In the run 273 each sentence is smoothed by its local context and first sentences from Wikipedia article which it is taken from. The run 274 has the same parameters except it does not have any smoothing. So, we can conclude that smoothing improves Informativeness. In our best run 275 punctuation score is not taken into account, it has slightly different formula for NE comparison and no penalization for numbers.

Readability was estimated as mean average scores per summary over soundness (no unresolved anaphora), non-redundancy and syntactical correctness among relevant passages of the ten tweets having the largest text references. According to all metrics except redundancy our approach was the best among all participants (see Table 2; IRIT's runs are set off in bold). Runs were officially ranked according to mean average scores. Readability evaluation also showed that the run 275 is the best by relevance, soundness and syntax. However, the run 274 is much better in terms of avoiding redundant information. The runs 273 and 274 are close according readability assessment as well.

Table 1. Informativeness evaluation

Rank	Run	Manual	All.skip	All.bi	All.uni	Pool.skip	Pool.bi	Pool.uni	Prior.skip	Prior.bi	Prior.uni
1	256	y	0,886	0,881	0,782	0,875	0,870	0,781	0,921	0,913	0,781
2	258	n	0,894	0,891	0,794	0,880	0,877	0,792	0,929	0,923	0,799
3	**275**	**n**	**0,897**	**0,892**	**0,806**	**0,879**	**0,875**	**0,794**	**0,917**	**0,911**	**0,790**
4	**273**	**n**	**0,897**	**0,892**	**0,800**	**0,880**	**0,875**	**0,792**	**0,924**	**0,916**	**0,786**
5	**274**	**n**	**0,897**	**0,892**	**0,801**	**0,881**	**0,875**	**0,793**	**0,923**	**0,915**	**0,787**

Table 2. Readability evaluation

Rank	Run	Mean Average	Relevancy (T)	Non redundancy (R)	Soundness (A)	Syntax (S)
1	**275**	**72.44%**	**76.64%**	**67.30%**	**74.52%**	**75.50%**
2	256	72.13%	74.24%	71.98%	70.78%	73.62%
3	**274**	**71.71%**	**74.66%**	**68.84%**	**71.78%**	**74.50%**
4	**273**	**71.35%**	**75.52%**	**67.88%**	**71.20%**	**74.96%**

4 Conclusion

This year we further developed our approach firstly introduced at INEX 2011 which is based on hashtag processing, TF-IDF cosine similarity measure enriched by smoothing from local context and document beginning, named entity recognition and part-of-speech weighting. We enriched our method by sentence quality measure based on Flesch reading ease test, lexical diversity, meaningful word ratio and punctuation ratio. We also used generalized POS (e.g. we merge regular adverbs, superlative and comparative into a single adverb group). Sentence score depends on document relevance and sentence type.

We submitted 3 runs in English differing by sentence quality score and smoothing and 1 run in Spanish.

Our approach was ranked first, second and third over 24 runs submitted by all participants on the PRIOR, POOL and ALL respectively. Among automatic runs our method was classified first (PRIOR and POOL) and second (ALL).

Readability was estimated as mean average scores per summary over resolved anaphora, non-redundancy and syntactical correctness among relevant passages of the ten tweets having the largest text references. According to all metrics except redundancy our approach was the best.

In future we plan to automatize parameter selection by machine learning methods.

5 References

1. Boyd, D., Golder, S., Lotan, G.: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. Proceedings of the 2010 43rd Hawaii International Conference on System Sciences. pp. 1–10. IEEE Computer Society (2010).
2. Celebrating #Twitter7 | Twitter Blog, https://blog.twitter.com/2013/celebrating-twitter7.
3. INEX 2013 Tweet Contextualization Track, https://inex.mmci.uni-saarland.de/tracks/qa/.
4. Ermakova, L., Mothe, J.: IRIT at INEX: Question Answering Task. Focused Retrieval of Content and Structure. pp. 219–226 (2012).
5. Ermakova, L., Mothe, J.: IRIT at INEX 2012: Tweet Contextualization, http://www.clef-initiative.eu/documents/71612/3e9ecc64-fae6-4af3-93fd-1a6a6fabb5d6, (2012).
6. Twitter Help Center | What Are Hashtags ("#" Symbols)?, https://support.twitter.com/articles/49309-what-are-hashtags-symbols.
7. Twitter Help Center | What are @Replies and Mentions?, https://support.twitter.com/groups/31-twitter-basics/topics/109-tweets-messages/articles/14023-what-are-replies-and-mentions.
8. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006). , Seattle, Washington, USA (2006).
9. Porter, M.F.: An algorithm for suffix stripping. Readings in information retrieval. Morgan Kaufmann Publishers Inc., San Francisco (1997).
10. Snowball, http://snowball.tartarus.org/.
11. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. pp. 173–180. Association for Computational Linguistics, Stroudsburg, PA, USA (2003).
12. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370. Association for Computational Linguistics, Stroudsburg, PA, USA (2005).
13. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: the Penn Treebank, (1993).
14. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing. , Manchester, UK (1994).
15. Apache OpenNLP - Welcome to Apache OpenNLP, http://opennlp.apache.org/index.html.
16. Flesch, R.: A new readability yardstick. Journal of Applied Psychology. 32, p221 – 233 (1948).