# IRIT at INEX 2014: 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 2014. Systems had to provide a context to a tweet from the perspective of the entity. This year we further modified our approach presented at INEX 2011, 2012 and 2013 underlain by the product of different measures based on smoothing from local context, named entity recognition, part-of-speech weighting and sentence quality analysis. We introduced two ways to link an entity and a tweet, namely (1) concatenation of the entity and the tweet and (2) usage of the results obtained for the entity as a restriction to filter results retrieved for the tweet. Besides, we examined the influence of topic-comment relationship on contextualization.

**Keywords:** Information retrieval, tweet contextualization, summarization, sentence extraction, readability, topic-comment relationship.

## 1 Introduction

Millions of tweets are published every day. Twitter is an online social network and microblogging that enables to send and read text messages up to 140 characters [1]. This limit provokes the fact that often tweets are not self-content and need to be explained, i.e. to be contextualized. In 2014 INEX Tweet Contextualization Track aims to evaluate systems providing context to 240 tweets in English from the perspective of the related entities [2]. These tweets were collected by the organizers of CLEF RepLab 2013. They have at least 80 characters and do not contain URLs. A tweet has the following annotation types: the category (4 distinct), an entity name from the Wikipedia (64 distinct) and a manual topic label (235 distinct).

The context has to explain the relationship between a tweet and an entity. It should be a readable summary up to 500 words extracted from a dump of the Wikipedia from November 2012.

This paper presents IRIT's approach used at INEX Tweet Contextualization Track 2014. Since the task introduced the notion of entities associated to tweets, we include this new feature and propose two ways to link an entity and a tweet:

- Making a new query that includes both the entity and the tweet;
- Using of the results obtained for the entity as a restriction to filter results retrieved for the tweet.

Moreover, we analyzed the influence of topic-comment relationship within a sentence in contextualization task.

As in previous years, we consider tweet contextualization task as multi-document extractive summarization [3, 4] underlain by

- the product of scores based on hashtag processing;
- TF-IDF cosine similarity measure;
- smoothing from local context;
- named entity (NE) recognition;
- part-of-speech (POS) weighting;
- sentence quality measure based on Flesch reading ease test, lexical diversity, meaningful word ratio and punctuation ratio.

The paper is organized as follows. The Section 2 presents our method by recalling the principles of the 2011-2013 system and describing the modifications we made. The Section 3 discusses the obtained results. The Section 4 concludes the paper and provides some perspectives.

## 2 Method Description

### 2.1 Preprocessing

Firstly, we performed query preprocessing which differs over the runs:

1. In order to link an entity and a tweet we combined the fields *entity*, *topic* and *content* into a single search query.
2. The second way is to process fields *entity* and *content* as separate queries and then use the results obtained for the entity as a restriction to filter results retrieved for the tweet. Thus, the document retrieved by using the field *content* as a query are rejected if they do not coincide with top-ranked documents retrieved by using the field *entity*.

The queries are encoded by ASCII (characters are normalized). An entity is treated as a single phrase, i.e. a document has to contain all words expressing the entity.

Document retrieval was performed by the Terrier platform [5], an open-source search engine developed by the School of Computing Science, University of Glasgow. Terrier implements various weighting and retrieval models and allows stemming and blind relevance feedback. We use Porter stemmer [6].

The next step is to parse tweets and retrieved documents by Stanford CoreNLP which integrates such tools as POS tagger [7] and named entity recognizer [8]. It uses the Penn Treebank tag set [9].

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

### 2.2 Searching for Relevant Sentences

We modified the extraction component developed for INEX 2011-2013. As in previous years, the general idea is to compute similarity between the query and sentences and to retrieve the most similar passages.

We model a sentence as a set of vectors:

- Unigram vector represents the lemmas associated with tokens occurred within the sentence. For unigram vectors we compute cosine similarity measure.
- A lemma possesses the following features: POS, frequency and IDF. Functional words, such as conjunctions, prepositions and determiners, are not taken into account. POS, frequency and IDF represents vectors of weights for the unigram vector. We used generalized POS (e.g. we merge regular adverbs, superlative and comparative into a single adverb group).
- NE vector. 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 assumed that relevant sentences come from relevant documents therefore we multiply sentence score by document relevance or/and by inverted document rank. These characteristics are used for sentence weighting.

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. Last year we added smoothing from document beginning. Wikipedia abstracts contain the summary of the entire paper; therefore they can be also used for smoothing. However, this parameter did not improve results. Therefore we didn't use it this year.

As in 2013, we did not apply anaphora resolution. Neither we used redundancy treatment nor sentence reordering since the analysis of previous results showed that their impact is small.

In 2013 we introduced sentence quality measure based on the product of the Flesch reading ease test [10], lexical diversity, meaningful word ratio and punctuation score. 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. We kept this measure.

### 2.3 Topic-comment relationship in contextualization task

Linguistics establishes the difference between the clause-level topic and the discourse-level topic. However, within the bound of this paper we are interested in clause-level topic only. The *topic* (or *theme*) is the phrase in a clause that the rest of the clause is understood to be about, and the comment (also called *rheme* or *focus*) is what is being said about the topic. In simple English clause the topic usually coincides with the subject, however it is not a case of the passive voice. In most languages the common means to mark topic-comment relation are word order and intonation. Moreover, there exist special constructions to introduce the comment. However, the tendency is to use so-called topic fronting, i.e. to place topic at the beginning of a clause.

We hypothesize that topic-comment relationship identification is useful for contextualization. Quick query analysis provides evidence that an entity is considered as a topic, while tweet content refers rather to comment, i.e. what is said about the entity. Moreover, we assume that providing the context to an entity implies that this context should be about the entity, i.e. the entity is the topic, while the retrieved context presents the comment.

We used these assumptions for candidate sentence scoring. We double the weight of sentences in which the topic contains the entity under consideration.

Topic identification is performed under assumption of topic fronting. We simplify this hypothesis by assuming that topic should be place at the sentence beginning. Sentence beginning is viewed as the first half of the sentence.

## 3 Evaluation

Summaries were evaluated according to their informativeness and readability. Informativeness was estimated as the overlap of a summary with the pool of relevant passages.

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 the gold standard estimated by skip bigrams.

The organizers used 2 gold standards:

- pool of relevant sentences per topic;
- pool of noun phrases extracted from these sentences together with the corresponding Wikipedia entry.

The gold standard thorough is a manual run on 1/5 of the 2014 topics.
We submitted 3 runs:

1. The first run (**ETC**) was performed by the system 2013. As a query three fields *entity, topic* and *content* were treated. An entity was treated as a single phrase.
2. The second run (**ETC_ENTITY**) differed from ETC by double weight for sentences where the entity represented the topic.

3. Unlike ETC, the third run (**ETC_RESTR_NOENT**) was based on document set restricted by entities (see the subsection 2.1 Preprocessing).

Table 1 and Table 2 provide evaluation results. The evaluation results presented in the Table 1 was based on the pool of relevant sentences, while the results obtained on the pool of noun phrases are given in the Table 2.

ref2013 and ref2012 are the baselines generated using 2013 and 2012 corpus. They are using the same system and index. However, they seem to be artificial. Therefore, we believe that they can be ignored in ranking.

According to the evaluation performed on the pool of sentences, our runs ETC, ETC_ENTITY and ETC_RESTR_NOENT were classified 3-rd, 4-nd and 6-th; while according to the evaluation based on noun phrases, they got slightly better ranks, namely 2, 3 and 5 respectively.

Thus, the best results among our runs were obtained by the system that merges fields *entity*, *topic* and *content* into a single query. The run #360 is better than our runs according to sentence evaluation; nevertheless, it showed worse results according to noun phrase evaluation. Our system is targeted on the nouns and especially named entities. This could provoke the differences in ranking with respect to sentences and noun phrases.

The worst results were showed by the run based on entity restriction. This could be explained by the fact that filtering out the documents that are considered irrelevant to the entity may cause the big loss of relevant documents if they are not top-ranked according to entities. ETC_RESTR_NOENT demonstrated the worst results among our runs even in the case of noun phrases. We believe that this is caused by loss in recall since the importance of noun phrases is not evaluated, but filtering out some documents could have negative effect on noun phrase recall.

The results of ETC and ETC_ENTITY are very close. However, topic-subject identification slightly decreased the performance of the system. Yet we believe that finer topic-comment identification procedure may ameliorate the results.

**Table 1.** Informativeness evaluation: pool of sentences

| Rank | Run | Unigrams | Bigrams | Skip bigrams |
|---|---|---|---|---|
| 1 | ref2013 | 0.705 | 0.794 | 0.796 |
| 2 | ref2012 | 0.7528 | 0.8499 | 0.8516 |
| 3 | 361 | 0.7632 | 0.8689 | 0.8702 |
| 4 | 360 | 0.782 | 0.8925 | 0.8934 |
| **5** | **ETC** | **0.8112** | **0.9066** | **0.9082** |
| **6** | **ETC_ENTITY** | **0.814** | **0.9098** | **0.9114** |
| 7 | 359 | 0.8022 | 0.912 | 0.9127 |
| **8** | **ETC_RESTR_NOENT** | **0.8152** | **0.9137** | **0.9154** |

| | | | | |
|---|---|---|---|---|
| 9 | 356 | 0.8415 | 0.9696 | 0.9702 |
| 10 | 357 | 0.8539 | 0.97 | 0.9712 |
| 11 | 364 | 0.8461 | 0.9697 | 0.9721 |
| 12 | 358 | 0.8731 | 0.9832 | 0.9841 |
| 13 | 363 | 0.8682 | 0.9825 | 0.9847 |
| 14 | 362 | 0.8686 | 0.9828 | 0.9847 |

**Table 2.** Informativeness evaluation: pool of noun phrases

| Rank | Run | Unigrams | Bigrams | Skip bigrams |
|---|---|---|---|---|
| 1 | ref2013 | 0.7468 | 0.8936 | 0.9237 |
| 2 | ref2012 | 0.7784 | 0.917 | 0.9393 |
| 3 | 361 | 0.7903 | 0.9273 | 0.9461 |
| **4** | **ETC** | **0.8088** | **0.9322** | **0.9486** |
| **5** | **ETC_ENTITY** | **0.809** | **0.9326** | **0.9489** |
| 6 | 360 | 0.8104 | 0.9406 | 0.9553 |
| **7** | **ETC_RESTR_NOENT** | **0.8131** | **0.936** | **0.9513** |
| 8 | 359 | 0.8227 | 0.9487 | 0.9613 |
| 9 | 356 | 0.8477 | 0.971 | 0.9751 |
| 10 | 357 | 0.8593 | 0.9709 | 0.9752 |
| 11 | 364 | 0.8628 | 0.9744 | 0.9807 |
| 12 | 358 | 0.8816 | 0.984 | 0.9864 |
| 13 | 363 | 0.884 | 0.9827 | 0.987 |
| 14 | 362 | 0.8849 | 0.9833 | 0.9876 |

Readability evaluation was performed by one assessor over a pool of 12 summaries per run. Readability was estimated as mean average scores per summary over soundness, structure (no unresolved anaphora), non-redundancy (diversity) and syntactical correctness.

The readability results are given in the Table 3. In general we can see that informativeness results are opposite to readability ones. However, our runs kept the same relative order. We received very low score for diversity and structure. This may be related to the fact that we decide not to treat this problem since in previous years their impact was small. Despite we retrieved the entire sentences from the Wikipedia, unexpectedly we received quite low score for syntactical correctness.

ETC_ENTITY demonstrated slightly higher results according to all readability measures except diversity. The differences of readability scores between ETC_RESTR_NOENT and ETC are very small since these runs are very similar. The only difference is the documents used as sources of the retrieved sentences. However, all readability scores of ETC_RESTR_NOENT are lower. This can be caused by lower quality of the documents or the influence of the informativeness on the assessor perception of readability.

**Table 3. Readability evaluation**

| Rank | Run | Readablity | Syntax | Diversity | Structure | Average |
|---|---|---|---|---|---|---|
| 1 | 358 | 94.82% | 87.31% | 72.17% | 93.10% | 86.85% |
| 2 | 356 | 95.24% | 85.19% | 70.31% | 92.40% | 85.78% |
| 3 | 357 | 94.88% | 82.53% | 71.34% | 91.58% | 85.08% |
| 4 | 364 | 88.05% | 69.94% | 63.91% | 86.92% | 77.20% |
| 5 | 360 | 92.60% | 70.35% | 58.84% | 86.33% | 77.03% |
| 6 | ref2013 | 91.74% | 69.82% | 60.52% | 85.80% | 76.97% |
| 7 | ref2012 | 91.39% | 69.58% | 60.67% | 85.56% | 76.80% |
| 8 | 359 | 93.03% | 70.64% | 53.53% | 86.34% | 75.88% |
| 9 | 363 | 83.68% | 67.92% | 61.13% | 87.55% | 75.07% |
| 10 | 362 | 83.67% | 68.00% | 60.81% | 87.59% | 75.02% |
| 11 | 361 | 93.23% | 70.41% | 50.12% | 85.97% | 74.93% |
| **12** | **ETC** | **90.88%** | **68.89%** | **56.59%** | **80.88%** | **74.31%** |
| **13** | **ETC_ENTITY** | **91.23%** | **69.47%** | **54.93%** | **81.56%** | **74.30%** |
| **14** | **ETC_RESTR_NOENT** | **90.10%** | **68.30%** | **53.83%** | **80.70%** | **73.23%** |

## 4 Conclusion

This year we further modified our approach presented at INEX 2011, 2012 and 2013 underlain by the product of different measures based on smoothing from local context, named entity recognition, part-of-speech weighting and sentence quality analysis. We introduced two ways to link an entity and a tweet, namely (1) concatena-

tion of the entity and the tweet and (2) usage of the results obtained for the entity as a restriction to filter results retrieved for the tweet. Besides, we examined the influence of topic-comment relationship on contextualization. Despite these modifications did not improve results, we believe that small changes in implementation may produce positive effect on the system performance.

## 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. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J.: Overview of the INEX 2014 Tweet Contextualization Track. CLEF 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org). 7424, (2014).
3. Ermakova, L., Mothe, J.: IRIT at INEX: Question Answering Task. Focused Retrieval of Content and Structure. pp. 219–226 (2012).
4. Ermakova, L., Mothe, J.: IRIT at INEX 2013: Tweet Contextualization Track, http://www.clef-initiative.eu/documents/71612/58a64b0a-cf0c-4751-a91f-9c8aba4312e1.
5. 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).
6. Porter, M.F.: An algorithm for suffix stripping. Readings in information retrieval. Morgan Kaufmann Publishers Inc., San Francisco (1997).
7. 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).
8. 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).
9. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: the Penn Treebank, (1993).
10. Flesch, R.: A new readability yardstick. Journal of Applied Psychology. 32, p221 – 233 (1948).