Overview of INEX Tweet Contextualization 2013 track

Patrice Bellot, Véronique Moriceau, Josiane Mothe, Eric Sanjuan, Xavier Tannier

Overview of INEX Tweet Contextualization 2013 track

Patrice Bellot[1], Véronique Moriceau[2], Josiane Mothe[3], Eric SanJuan[4], and Xavier Tannier[2]

[1] LSIS - Aix-Marseille University (France)
patrice.bellot@univ-amu.fr
[2] LIMSI-CNRS, University Paris-Sud (France)
{moriceau,xtannier}@limsi.fr
[3] IRIT, UMR 5505, Université de Toulouse, Institut Universitaire de Formation des Maitres Midi-Pyrénées (France)
josiane.mothe@irit.fr
[4] LIA, Université d'Avignon et des Pays de Vaucluse (France)
eric.sanjuan@univ-avignon.fr

Abstract. Twitter is increasingly used for on-line client and audience fishing, this motivated the tweet contextualization task at INEX. The objective is to help a user to understand a tweet by providing him with a short summary (500 words). This summary should be built automatically using local resources like the Wikipedia and generated by extracting relevant passages and aggregating them into a coherent summary. The task is evaluated considering informativeness which is computed using a variant of Kullback-Leibler divergence and passage pooling. Meanwhile effective readability in context of summaries is checked using binary questionnaires on small samples of results. Running since 2010, results show that only systems that efficiently combine passage retrieval, sentence segmentation and scoring, named entity recognition, text POS analysis, anaphora detection, diversity content measure as well as sentence reordering are effective.

Keywords: Short text contextualization, Tweet understanding, Automatic summarization, Question answering, Focus information retrieval, XML, Natural language processing, Wikipedia, Text readability, Text informativeness

1 Motivation

Text contextualization [8, 7] differs from text expansion in that it aims at helping a human to understand a text rather than a system to better perform its task. For example, in the case of query expansion in IR, the idea is to add terms to the initial query that will help the system to better select the documents to be retrieved. Text contextualization on the contrary can be viewed as a way to provide more information on the corresponding text in the objective to make it understandable and to relate this text to information that explains it.

In the context of micro-blogging, which is increasingly used for many purposes such as for on-line client and audience fishing, contextualization is specifically important since 140 characters long messages are rarely self-content. This motivated the proposal in 2011 of a new track at Clef INEX lab of Tweet Contextualization.

The use case is as follows: given a tweet, the user wants to be able to understand the tweet by reading a short textual summary; this summary should be readable on a mobile device without having to scroll too much. In addition, the user should not have to query any system and the system should use a resource freely available. More specifically, the guideline specified the summary should be 500 words long and built from sentences extracted from a dump of Wikipedia. Wikipedia has been chosen both for evaluation purpose and because this is an increasing popular ressource while being generally trustable. In this paper, details the 2013 track set up and results. The use case and the topic selection remained stable since 2011[8], so that 2011 and 2012 topics could be used as a training set. However, In 2013 we considered more diverse types of tweets for this year edition, so that participants could better measure the impact of hashtag processing on their approaches.

The remaining of the paper is organised as follows: In section 2 we describe in detail the 2013 data collection. Section 3 presents the results and Section 4 concludes this paper.

2 Data collection

This section describes the document collection that is used as the resource for contextualization, as well as the topics selected for the test set which correspond to the tweets to contextualize.

The document collection has been built based on a recent dump of the English Wikipedia from November 2012. Since we target a plain XML corpus for an easy extraction of plain text answers, like in past years, we used the same perl programs released for all participants to remove all notes and bibliographic references that are difficult to handle and keep only non empty Wikipedia pages (pages having at least one section).

Resulting automatically generated documents from Wikipedia dump, consist of a title (`title`), an abstract (`a`) and sections (`s`). Each section has a sub-title (`h`). Abstract and sections are made of paragraphs (`p`) and each paragraph can contain entities (`t`) that refer to other Wikipedia pages.

Over 2012 and 2013 editions, evaluated topics were made of 120 (60 topics each year) tweets manually collected by organizers. These tweets were selected and checked, in order to make sure that:

– They contained "informative content" (in particular, no purely personal messages); Only non-personal accounts were considered (*i.e.* @CNN, @TennisTweets, @PeopleMag, @science. . .).

– The document collection from Wikipedia contained related content, so that a contextualization was possible.

From the same set of accounts, more than 1,800 tweets were then collected automatically. These tweets were added to the evaluation set, in order to avoid that fully manual, or not robust enough systems could achieve the task. All tweets were then to be treated by participants, but only the 120 short list was used for evaluation. Participants did not know which topics were selected for evaluation.

These tweets were provided in a text-only format without metadata and in a JSON format with all associated metadata.

3 Results

This year the entire evaluation process was carried out by organizers.

Tweet contextualization [6] is evaluated on both informativeness and readability. Informativeness aims at measuring how well the summary explains the tweet or how well the summary helps a user to understand the tweet content. On the other hand, readability aims at measuring how clear and easy to understand the summary is.

Informativeness measure is based on lexical overlap between a pool of relevant passages (RPs) and participant summaries. Once the pool of RPs is constituted, the process is automatic and can be applied to unofficial runs. The release of these pools is one of the main contributions of Tweet Contextualization tracks at INEX[8, 7].

By contrast, readability is evaluated manually and cannot be reproduced on unofficial runs. In this evaluation the assessor indicates where he misses the point of the answers because of highly incoherent grammatical structures, unsolved anaphora, or redundant passages. Like in 2012, three metrics were used: **Relevancy (or Relaxed) metric**, counting passages where the T box has not been checked; **Syntax**, counting passages where the S box was not checked either, and the **Structure (or Strict) metric** counting passages where no box was checked at all.

Participant runs were ranked according to the average, normalized number of words in valid passages.

In 2013, a total number of 13 teams from 9 countries (Brasil, Canada, France, India, Ireland, Mexico, Russia, Spain, USA) submitted 24 runs to the Tweet Contextualization track in the framework of CLEF INEX lab 2013.

Infomativity results are presented in Table 1 and statistical significance of differencies between scores are indicated in Table 2. Table 1 shows readability scores.

This year, the best participating system (199) used hashtag preprocessing introduced in [1]. The best run by this participant used all available tweet features including web links which was not allowed by organisers. However his second best run without using linked web pages is ranked first among official runs. This participant also tried to weight hashtags based on 2012 results but this did not improve results. Perhaps because topics evaluated in 2012 were too specific.

Second best participant (182) in informativity and best in readability used state of the art NLP tools. This participant was first in informativity in 2011 [2]. Differences between these two best systems are not statistically significant.

Third best participant system (65) was first in 2012 [4], so the same system performs well even on a more diversify set of tweets.

Reference system by organisers (62-276) available online through an API is not more among three best systems. This systems is a robust focused information retrieval system [6] that was not smoothed for tweets. This year we also set up a baseline (62 - 278) using a state of the art IR system on sentences. Its informativity scores are high but its readability is very low.

Overall, informativity and readability scores are this year strongly correlated (Kendall test: $\tau > 90\%$, $p < 10^{-3}$) which shows that all systems have integrated this constrain. Remenber that since 2012, readability is evaluated in the context of the tweet. Passages not related to the tweet are considered as unreadable.

Rank	Participant	Run	unigram	bigram	with 2-gap
1	199	256*	0.7820	0.8810	0.8861
2	199	258	0.7939	0.8908	0.8943
3	182	275	0.8061	0.8924	0.8969
4	182	273	0.8004	0.8921	0.8973
5	182	274	0.8009	0.8922	0.8974
6	199	257*	0.7987	0.8969	0.8998
7	65	254	0.8331	0.9229	0.9242
8	62	276	0.8169	0.9270	0.9301
9	46	270	0.8481	0.9365	0.9397
10	46	267	0.8838	0.9444	0.9468
11	46	271	0.8569	0.9475	0.9500
12	62	278	0.8673	0.9540	0.9575
13	210	277	0.8995	0.9649	0.9662
14	129	261	0.8639	0.9668	0.9670
15	129	259	0.8631	0.9673	0.9679
16	129	260	0.8643	0.9677	0.9680
17	128	262	0.8738	0.9734	0.9747
18	128	255	0.8817	0.9771	0.9783
19	138	265	0.8793	0.9781	0.9789
20	138	263	0.8796	0.9785	0.9793
21	138	264	0.8790	0.9791	0.9798
22	275	266	0.9059	0.9824	0.9835
23	180	269	0.9965	0.9999	0.9999
24	180	269*	0.9981	0.9999	0.9999

Table 1. Informativeness results(official results are “with 2-gap”).

	256	258	275	273	274	257	254	276	270	267	271	278	277	261	259	260	262	255	265	263	264	266	269
256	-	1	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
258	1	-	-	-	-	-	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
275	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
273	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
274	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
257	2	-	-	-	-	-	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
254	3	3	2	2	2	3	-	-	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3
276	3	3	3	3	3	3	-	-	-	1	2	3	3	3	3	3	3	3	3	3	3	3	3
270	3	3	3	3	3	3	2	-	-	-	3	3	3	3	3	3	3	3	3	3	3	3	3
267	3	3	3	3	3	3	2	1	-	-	-	-	1	2	2	2	3	3	3	3	3	3	3
271	3	3	3	3	3	3	3	2	3	-	-	-	2	2	3	3	3	3	3	3	3	3	3
278	3	3	3	3	3	3	3	3	3	-	-	-	-	1	1	1	3	3	3	3	3	3	3
277	3	3	3	3	3	3	3	3	3	1	2	-	-	-	-	-	-	1	1	1	2	2	3
261	3	3	3	3	3	3	3	3	3	2	2	1	-	-	-	-	1	2	3	3	3	3	3
259	3	3	3	3	3	3	3	3	3	2	3	1	-	-	-	-	-	2	3	3	3	3	3
260	3	3	3	3	3	3	3	3	3	2	3	1	-	-	-	-	-	2	3	3	3	3	3
262	3	3	3	3	3	3	3	3	3	3	3	3	-	1	-	-	-	-	-	-	-	2	3
255	3	3	3	3	3	3	3	3	3	3	3	3	1	2	2	2	-	-	-	-	-	-	3
265	3	3	3	3	3	3	3	3	3	3	3	3	1	3	3	3	-	-	-	-	-	-	3
263	3	3	3	3	3	3	3	3	3	3	3	3	1	3	3	3	-	-	-	-	-	-	3
264	3	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	-	-	-	-	-	-	3
266	3	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	2	-	-	-	-	-	3
269	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-

Table 2. Statistical significance for official results in table 1 (t-test, two sided, $1 = 90\%$, $2 = 95\%$, $3 = 99\%$, $\alpha = 5\%$).

4 Conclusions

Like in 2012, almots all participants used language models.

Terminology extraction and reformulation applied to tweets was also used in 2013 like in previous editions [9]. Appropriate stemming and robust parsing of both tweets and wikipedia pages also seems to be an important issue. Most systems having a run among the top ten in informativeness used the Standford Core NLP tool or the TreeTagger.

It also seems that automatic readability evaluation and anaphora detection helps improving readability scores, but also informativeness density in summaries. It is now clear that state of the art summarization methods based on sentence scoring [5] proved to be helpful on this task even though they need to be combined with an IR engine.

Best run in 2013 also experimented a tweet hashtag scoring technique introduced in 2012 [1] while generating the summary.

Finally, this time the state of the art system proposed by organizers since 2010 combining LM indexation, terminology graph extraction and summarization based on shallow parsing was not ranked among the six best runs which shows that participant systems improved on this task over the three editions.

Rank	Run	Mean AVG	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%
5	257	69.54%	72.18%	65.48%	70.96%	72.18%
6	254	67.46%	73.30%	61.52%	68.94%	71.92%
7	258	65.97%	68.36%	64.52%	66.04%	67.34%
8	276	49.72%	52.08%	45.84%	51.24%	52.08%
9	267	46.72%	50.54%	40.90%	49.56%	49.70%
10	270	44.17%	46.84%	41.20%	45.30%	46.00%
11	271	38.76%	41.16%	35.38%	39.74%	41.16%
12	264	38.56%	41.26%	33.16%	41.26%	41.26%
13	260	38.21%	38.64%	37.36%	38.64%	38.64%
14	265	37.92%	39.46%	36.46%	37.84%	39.46%
15	259	37.70%	38.78%	35.54%	38.78%	38.78%
16	255	36.59%	38.98%	31.82%	38.98%	38.98%
17	261	35.99%	36.42%	35.14%	36.42%	36.42%
18	263	32.75%	34.48%	31.86%	31.92%	34.48%
19	262	32.35%	33.34%	30.38%	33.34%	33.34%
20	266	25.64%	25.92%	25.08%	25.92%	25.92%
21	277	20.00%	20.00%	20.00%	20.00%	20.00%
22	269	00.04%	00.04%	00.04%	00.04%	00.04%

Table 3. Readability results

References

1. Deveaud, R., Boudin, F.: Lia/lina at the inex 2012 tweet contextualization track. In: Forner et al. [3]
2. Ermakova, L., Mothe, J.: Irit at inex 2012: Tweet contextualization. In: Forner et al. [3]
3. Forner, P., Karlgren, J., Womser-Hacker, C. (eds.): CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012 (2012)
4. Ganguly, D., Leveling, J., Jones, G.J.F.: Dcu@inex-2012: Exploring sentence retrieval for tweet contextualization. In: Forner et al. [3]
5. Moreno, J.M.T., Velázquez-Morales, P.: Two statistical summarizers at inex 2012 tweet contextualization track. In: Forner et al. [3]
6. SanJuan, E., Bellot, P., Moriceau, V., Tannier, X.: Overview of the inex 2010 question answering track (qa@inex). In: Geva, S., Kamps, J., Schenkel, R., Trotman, A. (eds.) INEX. Lecture Notes in Computer Science, vol. 6932, pp. 269–281. Springer (2010)
7. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J.: Overview of the inex 2012 tweet contextualization track. In: Forner et al. [3]
8. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J.: Overview of the inex 2011 question answering track (qa@inex). In: Geva, S., Kamps, J., Schenkel, R. (eds.) Focused Retrieval of Content and Structure, Lecture Notes in Computer Science, vol. 7424, pp. 188–206. Springer (2012)
9. Vivaldi, J., da Cunha, I.: Inex tweet contextualization track at clef 2012: Query reformulation using terminological patterns and automatic summarization. In: Forner et al. [3]