Overview of INEX Tweet Contextualization 2014 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. 140 characters long messages are rarely self-content. The Tweet Contextualization aims at providing automatically information - a summary that explains the tweet. This requires combining multiple types of processing from information retrieval to multi-document summarization including entity linking. Running since 2010, the task in 2014 was a slight variant of previous ones considering more complex queries from RepLab 2013. Given a tweet and a related entity, systems had to provide some context about the subject of the tweet from the perspective of the entity, in order to help the reader to understand it.

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

1 Motivation

The task in 2014 is a slight variant of previous ones and it is complementary to CLEF RepLab. Previously, given a tweet, systems had to help the user to understand it by reading a short textual summary. This summary had to 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.

In 2014 a small variant of the task has been explored, considering more complex queries from RepLab 2013, but using the same corpus. The new use case of the task was the following: given a tweet and a related entity, the system must provide some context about the subject of the tweet from the perspective

of the entity, in order to help the reader answering questions of the form "why this tweet concerns the entity? should it be an alert?".

In the remaining we give details about the 2014 track in English language set up and results.

We refer the reader to the CLEF Working Notes for the pilot task in Spanish.

2 Data collection

The official document collection for 2014 was the same as in 2013. Between 2011 and 2013 the corpus did change every year but not the user case. In 2014, the same corpus was reused but the user case evolved. Since 2014 TC topics are a selection of tweets from RepLab 2013, it was necessary to use prior WikiPedia dumps. Some participants also used the 2012 corpus raising up the question of the impact of updating the WikiPedia over these tasks.

Let us recall that the document collection has been built based on yearly dumps of the English WikiPedia since November 2011. We released a set of tools to convert a WikiPedia dump into a plain XML corpus for an easy extraction of plain text answers. The same perl programs released for all participants have been used 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).

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

As tweets, 240 topics have been collected from RepLab 2013 corpus. These tweets have been selected in order to make sure that:

- They contained "informative content" (in particular, no purely personal messages);
- The document collections from Wikipedia had related content, so that a contextualization was possible.

In order to avoid that fully manual, or not robust enough systems could achieve the task, all tweets were to be treated by participants, but only a random sample of them was to be considered for evaluation.

These tweets were provided in XML and tabulated format with the following information:

- the category (4 distinct),
- an entity name from the wikipedia (64 distinct)
- a manual topic label (235 distinct).

The entity name was to be used as an entry point into WikiPedia or DBpedia. The context of the generated summaries was expected to be fully related to this entity. On the contrary, the usefulness of topic labels for this automatic task was and remains an open question at this moment because of their variety.

3 Evaluation

Like in 2013, the entire evaluation process was carried out by organizers.

Tweet contextualization [1] 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 Informativeness measure is based on lexical overlap between a set of relevant passages (RPs) and participant summaries based on LogSim divergence introduced in [1]. Given an integer $\lambda > 30$, and two texts T, S the LogSim divergence can be restated as:

$$LS(\Omega_S|\Omega_T) = \sum_{\omega \in \Omega_T} P(\omega|\Omega_T).\frac{\min(\Phi_T(\omega), \Phi_S(\omega))}{\max(\Phi_T(\omega), \Phi_S(\omega)} \quad (1)$$

where for any text Z, Ω_Z is the set of n-grams in Z and for any n-gram $\omega \in \Omega_Z$:

$$\Phi_Z(\omega) = \log(1 + \lambda P(\omega|\Omega_Z)) \quad (2)$$

The λ parameter used in LS formula represents the summary allowed maximal length in words (500 in our case).

Once the pool of RPs (t-rels) 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[3, 2].

In previous editions t-rels were based on a pooling of participant submitted passages. Organizers then selected among them those that were relevant. In 2013, to build a more robust reference, two manual runs by participants were added using different on line research engines to find relevant WikiPedia pages and copying the relevant passages into the reference.

This year, even though there were only five participants, the variety of submitted passages was too high compared to the number of runs. One reason was that this year topics included more facets and converting them into queries for a Research Engine was less straightforward. As a consequence, it was not possible to rely on a pooling from participant runs because it would have been too sparse and incomplete. It was finally decided to rely on a thorough manual run by organizers based on the reference system that was made available to all participants at `http://qa.termwatch.es`

A manual query in Indri language was set up for every topic over five. These queries have been refined until they provide only a set of relevant passages using the reference system on the 2013 corpus. From this RPs we extracted two t-rels, one merging all passages for each tweets, another by only considering the Noun Phrases (NPs) in the passages to reduce the risk of introducing document identifiers in the passages.

The average length of queries to build the reference is 8 tokens with a minimum of 2 and a maximum of 14. Therefore efficient queries are much shorter than tweets. The average number of relevant tokens in the t-rels based on passages is 620, and on the t-rels based on NPs is only 300.

Readability 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. Since 2012, three metrics have been used: **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.

As in previous editions, participant runs have been ranked according to the average, normalized number of words in valid passages.

4 Results

In 2014, 4 combined teams from six countries (Canada, France, Germany, India, Russia, Tunesia) submitted 12 runs to the Tweet Contextualization track in the framework of CLEF INEX lab 2014[5]. The total number of submitted passages was 54,932 with an average length of 32 tokens. The total number of tokens was 1,764,373 with an average of 7,352 per tweet.

We also generated two reference runs based one the organizer's system made available to participants using 2013 and 2012 corpus respectively.

To read the scores, the lower they are the better since these are divergences. Informativeness results based on passage t-rels are presented in Table 1, and those on NPs t-rels in Table 3. Statistical significance of differences between scores in Table 1 are indicated in Table 2. Readability results are presented in Table 4.

Both informativeness rankings in Table 1 and in Table 3 are highly correlated, however discrepancies between the two rankings show that differences between top ranked runs rely on tokens outside NPs, mainly verbs since functional words are removed in the evaluation.

Table 4 reveals that readability of reference runs is low, meanwhile they are made of longer passages than average to ensure local syntax correctness.

Since reference runs are using the same system and index as the manual run used to build the t-rels, they tend to minimize the informativeness divergence with the reference. However, average divergence remains high pointing out that selecting the right passages in the restricted context of an entity, was more difficult than previous more generic tasks. Considering readability, the fact that

[5] Two other teams from Mexico and Spain participated to the pilot task in Spanish submitting three runs not considered in this overview.

Rank	Run	unigram	bigram	with 2-gap
1	ref2013	0.7050	0.7940	0.7960
2	ref2012	0.7528	0.8499	0.8516
3	361	0.7632	0.8689	0.8702
4	360	0.7820	0.8925	0.8934
5	368	0.8112	0.9066	0.9082
6	369	0.8140	0.9098	0.9114
7	359	0.8022	0.9120	0.9127
8	370	0.8152	0.9137	0.9154
9	356	0.8415	0.9696	0.9702
10	357	0.8539	0.9700	0.9712
11	364	0.8461	0.9697	0.9721
12	358	0.8731	0.9832	0.9841
13	362	0.8686	0.9828	0.9847
14	363	0.8682	0.9825	0.9847

Table 1. Informativeness results bases on passage t-rels (official results are "with 2-gap").

	ref2013	ref2012	361	360	359	368	369	370	356	357	364	358	363	362
ref2013	-	2.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
ref2012	2.00	-	-	-	1.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
361	2.00	-	-	-	1.00	2.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00
360	2.00	-	-	-	-	-	-	-	3.00	3.00	3.00	3.00	3.00	3.00
359	3.00	1.00	1.00	-	-	-	-	-	2.00	2.00	2.00	3.00	3.00	3.00
368	3.00	3.00	2.00	-	-	-	-	2.00	3.00	3.00	3.00	3.00	3.00	3.00
369	3.00	3.00	2.00	-	-	-	-	-	3.00	3.00	3.00	3.00	3.00	3.00
370	3.00	3.00	2.00	-	-	2.00	-	-	3.00	3.00	3.00	3.00	3.00	3.00
356	3.00	3.00	3.00	3.00	2.00	3.00	3.00	3.00	-	-	-	3.00	3.00	3.00
357	3.00	3.00	3.00	3.00	2.00	3.00	3.00	3.00	-	-	-	3.00	3.00	3.00
364	3.00	3.00	3.00	3.00	2.00	3.00	3.00	3.00	-	-	-	3.00	3.00	3.00
358	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	-	-	-
363	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	-	-	-
362	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	-	-	-

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

reference runs are low ranked confirms that finding the right compromise between readability and informativeness remains the main difficulty of this task.

This year, the best participating system for informativeness used association rules. Since contextualization was restricted to some facet described by an entity, it could be that association rules helped to focus on this aspect.

The best participating system for readability used an advanced summarization systems that introduced minor changes in passages to improve readability. Changing the content of the passages was not allowed, however this tend to show that to deal with readability some rewriting is required. Moreover, since this year evaluation did not include a pool of passages from participants, systems that provided modified passages have been disadvantaged in informativeness evaluation.

Rank	Run	unigram	bigram	with 2-gap
1	ref2013	0.7468	0.8936	0.9237
2	ref2012	0.7784	0.9170	0.9393
3	361	0.7903	0.9273	0.9461
4	368	0.8088	0.9322	0.9486
5	369	0.8090	0.9326	0.9489
6	370	0.8131	0.9360	0.9513
7	360	0.8104	0.9406	0.9553
8	359	0.8227	0.9487	0.9613
9	356	0.8477	0.9710	0.9751
10	357	0.8593	0.9709	0.9752
11	364	0.8628	0.9744	0.9807
12	358	0.8816	0.9840	0.9864
13	363	0.8840	0.9827	0.9870
14	362	0.8849	0.9833	0.9876

Table 3. Informativeness results bases on NP t-rels (official results are "with 2-gap").

Rank	Run	Acceptable (T)	Syntax (S)	Structure (A)	Average
1	358	0.94822	0.722796	0.721683	0.931005
2	356	0.952381	0.650917	0.703141	0.923958
3	357	0.948846	0.578212	0.713445	0.91575
4	362	0.836699	0.366561	0.608136	0.875917
5	363	0.836776	0.363954	0.611289	0.8755
6	364	0.880508	0.337197	0.639092	0.869167
7	359	0.9303	0.258563	0.535264	0.863375
8	360	0.925959	0.258658	0.588365	0.863274
9	361	0.932281	0.247883	0.501199	0.859749
10	ref2013	0.917378	0.259702	0.605203	0.857958
11	ref2012	0.913858	0.259584	0.606742	0.855583
12	369	0.912318	0.259539	0.549334	0.815625
13	368	0.908815	0.248981	0.565912	0.80875
14	370	0.901044	0.246893	0.538338	0.806958

Table 4. Readability results

References

1. 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)
2. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J.: Overview of the inex 2012 tweet contextualization track. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF (Online Working Notes/Labs/Workshop) (2012)
3. 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)