

# Overview of the INEX 2012 Tweet Contextualization Track

Eric SanJuan<sup>1</sup>, Véronique Moriceau<sup>2</sup>, Xavier Tannier<sup>2</sup>, Patrice Bellot<sup>3</sup>, and  
Josiane Mothe<sup>4</sup>

<sup>1</sup> LIA, Université d'Avignon et des Pays de Vaucluse (France)

`eric.sanjuan@univ-avignon.fr`

<sup>2</sup> LIMSI-CNRS, University Paris-Sud (France)

`{moriceau,xtannier}@limsi.fr`

<sup>3</sup> LSIS, Université Aix-Marseille (France)

`patrice.bellot@lsis.org`

<sup>4</sup> IRT, Université de Toulouse (France)

`josiane.mothe@irit.fr`

**Abstract.** The use case of the Tweet Contextualization task is the following: given a new tweet, participating systems must provide some context about the subject of a tweet, in order to help the reader to understand it. In this task, contextualizing tweets consists in answering questions of the form “what is this tweet about?” which can be answered by several sentences or by an aggregation of texts from different documents of the Wikipedia. Thus, tweet analysis, XML/passage retrieval and automatic summarization are combined in order to get closer to real information needs.

This article describes the data sets and topics, the metrics used for the evaluation of the systems submissions, as well as the results that they obtained.

**Keywords:** Automatic Summarization, Focused Information Retrieval, XML, Twitter, Wikipedia

## 1 Introduction

The Tweet Contextualization task to be performed by the participating groups of INEX 2012 is contextualizing tweets, *i.e.* answering questions of the form “what is this tweet about?” using a recent cleaned dump of the Wikipedia. The general process involves:

- Tweet analysis,
- Passage and/or XML element retrieval,
- Construction of the context/summary.

We regard as relevant passages those that both contain relevant information but also contain as little non-relevant information as possible.

For evaluation purposes, we require that a summary uses only elements or passages previously extracted from the document collection. The correctness of summaries is established exclusively based on the support passages and documents. The summaries are evaluated according to:

- Informativeness: the way they overlap with relevant passages,
- Readability, assessed by evaluators and participants.

The paper is organized as follows. Section 2 details the collection of tweets and documents. Section 3 presents the metrics and tools used for evaluation, as well as results obtained by the participants. Finally, section 4 draws some preliminary conclusions.

## 2 Test data

Organizers provided a document collection extracted from Wikipedia, as well as 1000 topics made of tweets from several different accounts.

### 2.1 Tweets

About 1000 tweets in English were collected by the track organizers from Twitter<sup>®</sup> Search API. They were selected among informative accounts (for example, @CNN, @TennisTweets, @PeopleMag, @science...), in order to avoid purely personal tweets that could not be contextualized. Information such as the user name, tags or URLs have been provided. These tweets were available in two formats:

- a full JSON format with all tweet metadata. For example:

```
"created_at": "Wed, 15 Feb 2012 23:32:22 +0000",
"from_user": "FOXBroadcasting",
"from_user_id": 16537989,
"from_user_id_str": "16537989",
"from_user_name": "FOX Broadcasting",
"geo": null,
"id": 169927058904985600,
"id_str": "169927058904985600",
"iso_language_code": "en",
"metadata": {"result_type": "recent"},
"profile_image_url": "http://a0.twimg.com/profile_images/...",
"profile_image_url_https": "https://si0.twimg.com/profile_images/...",
"source": "&lt;a href=&quot;http://www.hootsuite.com...&quot;",
"text": "Tensions are at an all-time high as the @AmericanIdol
Hollywood Round continues, Tonight at 8/7c. #Idol",
"to_user": null,
"to_user_id": null,
"to_user_id_str": null,
"to_user_name": null
```

- a two-column text format with only tweet id and tweet text. For example:

169927058904985600	"Tensions are at an all-time high as the @AmericanIdol Hollywood Round continues, Tonight at 8/7c. #Idol"
--------------------	---

63 of these tweets were selected manually by organizers. For each of them, we checked that the document collection contained some information related to the topic of the tweet. This means that all 63 tweets had some contextualization material inside the provided collection.

From the accounts used for extraction of these 63 messages, a number of other tweets were automatically selected, bringing to 1000 the total number of tweets to be contextualized by the participants. This is done to ensure that only fully automatic and robust enough systems could accomplish the task.

However, only the 63 tweets that had been manually collected and checked have been used for informativeness evaluation; only 18 of them have been used for readability evaluation (due to the complexity of this evaluation).

## 2.2 Document collection

The document collection has been built based on a recent dump of the English Wikipedia from November 2011. Since we target a plain XML corpus for an easy extraction of plain text answers, we removed all notes and bibliographic references that are difficult to handle and kept only non empty Wikipedia pages (pages having at least one section).

Resulting documents are made of a title (**t**), 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 have entities (**e**) that refer to other Wikipedia pages. Therefore the resulting corpus has this simple DTD:

```
<!ELEMENT xml (page)+>
<!ELEMENT page (ID, title, a, s*)>
<!ELEMENT ID (#PCDATA)>
<!ELEMENT title (#PCDATA)><!ELEMENT a (p)+>
<!ELEMENT s (h, p)+>
<!ATTLIST s o CDATA #REQUIRED>
<!ELEMENT h (#PCDATA)>
<!ELEMENT p (#PCDATA | t)*>
<!ATTLIST p o CDATA #REQUIRED>
<!ELEMENT t (#PCDATA)>
<!ATTLIST t e CDATA #IMPLIED>
```

For example:

```

<?xml version="1.0" encoding="utf-8"?>
<page>
<ID>2001246</ID>
<title>Alvin Langdon Coburn</title>
<s o="1">
<h>Childhood (1882-1899)</h>
<p o="1">Coburn was born on June 11, 1882, at 134 East Springfield
Street in <t>Boston, Massachusetts</t>, to a middle-class family.
His father, who had established the successful firm of
Coburn & Whitman Shirts, died when he was seven.
[...]
</p>
<p o="2">In 1890 the family visited his maternal uncles in
Los Angeles, and they gave him a 4 x 5 Kodak camera. He immediately
fell in love with the camera, and within a few years he had developed
a remarkable talent for both visual composition and technical
proficiency in the <t>darkroom</t>. (...)</p>
(...)
</page>

```

### 2.3 Submission format

Participants could submit up to 3 runs. One run out of the 3 had to be completely automatic: participants had to use only the Wikipedia dump and possibly their own resources (even if the texts of tweets sometimes contain URLs, the Web must not be used as a resource). That is, a participant could not submit more than 3 runs in total.

A submitted summary has the following format:

```

<tid> Q0 <file> <rank> <rsv> <run_id> <text of passage 1>
<tid> Q0 <file> <rank> <rsv> <run_id> <text of passage 2>
<tid> Q0 <file> <rank> <rsv> <run_id> <text of passage 3>
...

```

where:

- The first column `tid` is the topic number.
- The second column is currently unused and should always be `Q0`. It is just a formatting requirement used by the evaluation programs to distinguish between official submitted runs and q-rels.
- The third column `file` is the file name (without `.xml`) from which a result is retrieved, which is identical to the `<id>` of the Wikipedia document. It is only used to retrieve the raw text content of the passage, not to compute document retrieval capabilities. In particular, if two results only differ by their document id (because the text is repeated in both), then they will be considered as identical and thus redundant.

- The fourth column rank indicates the order in which passages should be read for readability evaluation, this differs from the expected informativeness of the passage which is indicated by the score `rsv` in the fifth column. Therefore, these two columns are not necessarily correlated. Passages with highest scores in the fifth column can be scattered at any rank in the result list for each topic.
- The sixth column `run_id` is called the “run tag” and should be a unique identifier for the participant group and for the method used.
- The remaining column gives the result passage in raw text without XML tags and without formatting characters. The only requirement is that the resulting word sequence appears at least once in the file indicated in the third field.

Here is an example of such an output:

```
167999582578552 Q0 3005204 1 0.9999 I10UniXRun1 The Alfred Noble
Prize is an award presented by the combined engineering societies
of the United States, given each year to a person not over
thirty-five for a paper published in one of the journals of the
participating societies.
167999582578552 Q0 3005204 2 0.9998 I10UniXRun1 The prize was
established in 1929 in honor of Alfred Noble, Past President of
the American Society of Civil Engineers.
167999582578552 Q0 3005204 3 0.9997 I10UniXRun1 It has no connection
to the Nobel Prize, although the two are often confused due to
their similar spellings.
```

### 3 Evaluation

In this task, readability of answers [9] is as important as the informative content. Summaries must be easy to read as well as relevant. Following INEX 2011 Question-Answering task [1], these two properties have been evaluated separately by two distinct measures: informativeness and readability.

This section describes the metrics and tools used to perform the evaluation and gives results obtained by participating systems.

#### 3.1 Baseline System

A baseline XML-element retrieval/summarization system has been made available for participants. This baseline is the same as 2011 QA@INEX task, and has been described in [1]. It relies on the search engine Indri<sup>5</sup> and a fast summarizer algorithm [2]. The system was available to participants through a web interface<sup>6</sup> or a perl API. Its default output has been added to the pool of submitted runs.

<sup>5</sup> <http://www.lemurproject.org/>

<sup>6</sup> <http://qa.termwatch.es>

### 3.2 Submitted Runs

33 valid runs by 13 teams from 10 countries (Canada, Chile, France, Germany, India, Ireland, Mexico, Russia, Spain, USA) were submitted.

This year only three teams used the provided perl API and Indri index of the collection.

The total number of submitted passages is 671,191 (31 596 328 tokens). The median number of distinct passages per tweet is 79.5 and the average is 146.5. Only passages starting and ending by the same 25 characters have been considered as duplicated, therefore short sub-passages could appear twice in longer ones.

### 3.3 Informativeness Evaluation

Informativeness evaluation has been performed by organizers on a pool of 63 tweets. For each tweet, we took the 60 best passages based on the rsv score in the fifth column of the runs from all participants. After removing duplicates per tweet, 16,754 passages were evaluated by organizers. The median number of passages per tweet is 273 and the average is 265.9. Passages have been merged and displayed to the assessor in alphabetical order. Therefore, each passage informativeness has been evaluated independently from others, even in the same summary. The structure and readability of the summary was not assessed in this specific part, and assessors only had to provide a binary judgement on whether the passage was worth appearing in a summary on the topic, or not. 2,801 passages among 16,754 have been judged as relevant, with a median of 50 passages per tweet and an average of 55.1. The average length of a passage is 30.03 tokens.

**Metrics** Systems had to make a selection of the most relevant information, the maximal length of the abstract being fixed. Therefore focused IR systems could just return their top ranked passages meanwhile automatic summarization systems need to be combined with a document IR engine. In this task, readability of answers [3] is as important as the informative content. Both need to be evaluated. Therefore answers cannot be any passage of the corpus, but at least well formed sentences. As a consequence, informative content of passages cannot be evaluated using standard IR measures since QA and automatic summarization systems do not try to find all relevant passages, but to select those that could provide a comprehensive answer. Several metrics have been defined and experimented with at DUC [4] and TAC workshops [5]. Among them, Kullback-Leibler ( $KL$ ) and Jenson-Shanon ( $JS$ ) divergences have been used [6, 7] to evaluate the informativeness of short summaries based on a bunch of highly relevant documents.

In previous 2010 and 2011 INEX Question Answering tracks, evaluations have been carry out using FRESA package which includes a special lemmatizer. In 2011 we provided the participants with a standalone evaluation toolkit based on

Porter stemmer and implementing a new normalized ad-hoc dissimilarity defined as following:

$$Dis(T, S) = \sum_{t \in T} \frac{f_T(t)}{f_T} \times \left( 1 - \frac{\min(\log(P), \log(Q))}{\max(\log(P), \log(Q))} \right) \quad (1)$$

$$P = \frac{f_T(t)}{f_T} + 1 \quad (2)$$

$$Q = \frac{f_S(t)}{f_S} + 1 \quad (3)$$

where  $T$  is the set of terms in the reference and for every  $t \in T$ ,  $f_T(t)$  is its frequency in the reference and  $f_S(t)$  its frequency in the summary.

The idea was to have a dissimilarity which complement has similar properties to usual IR Interpolate Precision measures. Actually,  $1 - Dis(T, S)$  increases with the Interpolated Precision at 500 tokens where Precision is defined as the number of word n-grams in the reference. The introduction of the log is necessary to deal with highly frequent words.

As previously announced, we used this software to evaluate informativeness and like in INEX QA tracks, we considered as  $T$  three different sets based on Porter stemming:

- Unigrams made of single lemmas (after removing stop-words).
- Bigrams made of pairs of consecutive lemmas (in the same sentence).
- Bigrams with 2-gaps also made of pairs of consecutive lemmas but allowing the insertion between them of a maximum of two lemmas.

Bigrams with 2-gaps appeared to be the most robust metric. Sentences are not considered as simple bags of words and the measure is less sensitive to sentence segmentation than simple bi-grams. This is why bigrams with 2-gaps is our official ranking metric for informativeness.

Bigrams with 2-gaps appeared to be the most robust metric in previous INEX QA tracks, however in this edition where topics are real tweets, measures based on bigrams with or without 2-gaps are strongly correlated. Meanwhile the measure based on simple uni-grams is also stable but gives a different ranking. This will be discussed during the CLEF workshop.

**Results** Results are presented in Table 1. The 3 top ranked runs improved the baseline. Runs with (\*) have been submitted as “manual”.

Dissimilarity values are very closed, however differences are often statistically significant as shown in table 2.

### 3.4 Readability evaluation

**Human assessment** Each participant had to evaluate readability for a pool of summaries of a maximum of 500 words each on an online web interface. Each summary consisted in a set of passages and for each passage, assessors had to tick four kinds of check boxes. The guideline was the following:

Rank	Run	unigram	bigram	with 2-gap
1	178	0.7734	0.8616	0.8623
2	152	0.7827	0.8713	0.8748
3	170*	0.7901	0.8825	0.8848
4	Baseline	0.7864	0.8868	0.8887
5	169	0.7959	0.8881	0.8904
6	168	0.7972	0.8917	0.8930
7	193	0.7909	0.8920	0.8938
8	185	0.8265	0.9129	0.9135
9	171	0.8380	0.9168	0.9187
10	186	0.8347	0.9210	0.9208
11	187	0.8360	0.9235	0.9237
12	154	0.8233	0.9254	0.9251
13	162	0.8236	0.9257	0.9254
14	155	0.8253	0.9280	0.9274
15	153	0.8266	0.9291	0.9290
16	196b	0.8484	0.9294	0.9324
17	196c	0.8513	0.9305	0.9332
18	196a	0.8502	0.9316	0.9345
19	164*	0.8249	0.9365	0.9368
20	197	0.8565	0.9415	0.9441
21	163	0.8664	0.9628	0.9629
22	165	0.8818	0.9630	0.9634
23	150	0.9052	0.9871	0.9868
24	188	0.9541	0.9882	0.9888
25	176	0.8684	0.9879	0.9903
26	149	0.9059	0.9916	0.9916
27	156	0.9366	0.9913	0.9916
28	157	0.9715	0.9931	0.9937
29	191	0.9590	0.9947	0.9947
30	192	0.9590	0.9947	0.9947
31	161	0.9757	0.9949	0.9950
32	177	0.9541	0.9981	0.9984
33	151	0.9223	0.9985	0.9988

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

- *Syntax* (S): tick the box if the passage contains a syntactic problem (bad segmentation for example),
- *Anaphora* (A): tick the box if the passage contains an unsolved anaphora,
- *Redundancy* (R): tick the box if the passage contains a redundant information, i.e. an information that has already been given in a previous passage,
- *Trash* (T): tick the box if the passage does not make any sense in its context (*i.e.* after reading the previous passages). These passages must then be considered as trashed, and readability of following passages must be assessed as if these passages were not present.
- If the summary is so bad that you stop reading the text before the end, tick all trash boxes until the last passage.

For each summary, the text without tags of the tweet was displayed, thus this year readability was evaluated in the context of the tweet, and passages not related to the tweet could be considered as trash even if there were readable.

**Metrics and results** To evaluate summary readability, we consider the number of words (up to 500) in valid passages. We used three metrics based on this:

- **Relevancy or Relaxed metric:** a passage is considered as valid if the T box has not been ticked,
- **Syntax:** a passage is considered as valid if the T or S boxes have not been ticked,
- **Structure or Strict metric:** a passage is considered as valid if no box has been ticked.

In all cases, participant runs are ranked according to the average, normalized number of words in valid passages.

A total of 594 summaries from 18 tweets have been assessed. The resulting 18 tweets are included in those used for informativeness assessment. Results are presented in Table 3. The last column gives the number of evaluated summaries for corresponding run. Only runs that were evaluated on more than 6 summaries, are ranked following the relaxed metric. Missing evaluations were due to formatting problems, too long passages (more than 500 tokens) or missing summaries in the submitted runs.

## 4 Conclusion

In 2011 we experimented using the wikipedia to contextualize twitted New York times paper titles. There was a large overlapping between the two vocabularies. This year we selected a larger pool of public factual tweets with a much more diversified vocabulary. The robust baseline we provided was difficult to outperform on the average. This needs further analysis and will be discussed during the workshop. One reason could be that the baseline approach removes all non-nominals from tweet texts, keeping only nouns and adjectives and this can help

in wikipedia search. However, for specific tweets, to retrieve relevant information from the wikipedia, it was necessary to expand the tweet vocabulary or to use tags inside the tweet.

## References

1. 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: 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011. Volume 7424 of Lecture Notes in Computer Science., Saarbrcken, Germany, Springer Verlag, Berlin, Heidelberg (2012) 188–206
2. Chen, C., Ibekwe-Sanjuan, F., Hou, J.: The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *JASIST* **61**(7) (2010) 1386–1409
3. Pitler, E., Louis, A., Nenkova, A.: Automatic evaluation of linguistic quality in multi-document summarization. In: *ACL*. (2010) 544–554
4. Nenkova, A., Passonneau, R.: Evaluating content selection in summarization: The pyramid method. In: *Proceedings of HLT-NAACL*. Volume 2004. (2004)
5. Dang, H.: Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In: *Proc. of the First Text Analysis Conference*. (2008)
6. Louis, A., Nenkova, A.: Performance confidence estimation for automatic summarization. In: *EACL, The Association for Computer Linguistics* (2009) 541–548
7. Saggion, H., Torres-Moreno, J.M., da Cunha, I., SanJuan, E., Velázquez-Morales, P.: Multilingual summarization evaluation without human models. In Huang, C.R., Jurafsky, D., eds.: *COLING (Posters)*, Chinese Information Processing Society of China (2010) 1059–1067

	178	152	170	baseline	169	168	193	185	171	186	187	154	162	155	153	196b	196c	196a	164	197	165	163	150	188	176	156	149	157	191	192	161	177	151			
178	-	2	-	1	2	2	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
152	2	-	-	-	-	-	-	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
170	-	-	-	-	1	-	-	2	2	2	3	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
baseline	1	-	-	-	-	-	-	2	3	2	2	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
169	2	-	1	-	-	-	-	2	1	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
168	2	-	-	-	-	-	-	1	-	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
193	1	-	-	-	-	-	-	1	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
185	3	3	2	2	2	1	1	-	-	1	1	-	-	-	1	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
171	3	2	2	3	1	-	2	-	-	-	-	-	-	-	-	-	-	-	-	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
186	3	3	2	2	2	2	2	1	-	-	-	-	-	-	-	-	-	-	-	1	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
187	3	3	3	2	2	2	2	1	-	-	-	-	-	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
154	3	3	2	3	2	2	2	-	-	-	-	-	-	-	2	-	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
162	3	3	2	2	2	2	2	-	-	-	-	-	-	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
155	3	3	3	3	3	2	3	-	-	-	-	-	-	-	-	-	-	-	-	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
153	3	3	3	3	3	2	3	1	-	-	-	2	-	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
196b	3	3	3	3	3	3	3	2	-	-	-	-	-	-	-	-	-	-	-	3	-	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3
196c	3	3	3	3	3	3	3	2	-	-	-	-	-	-	-	-	-	-	-	-	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
196a	3	3	3	3	3	3	3	2	-	1	-	-	-	-	3	-	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
164	3	3	3	3	3	3	3	2	-	1	-	2	2	1	2	-	-	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
197	3	3	3	3	3	3	3	3	1	2	2	3	3	3	3	1	-	-	-	1	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
165	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	1	-	-	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3
163	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	-	-	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
150	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	3	-	-	-	1	1	1	1	2	3	3	3	3	3	3	3
188	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
176	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
156	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
149	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
157	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
191	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
192	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
161	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
177	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
151	3	3	3	3	3	3	3	3	3	3	3	3	3	3	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 informativeness evaluation (t-test, 1 : 90%, 2 = 95%, 3 = 99%,  $\alpha = 5\%$ ).

Rank	Run	Relevancy	Syntax	Structure	Nb
1	185	0.7728	0.7452	0.6446	17
2	171	0.6310	0.6060	0.6076	10
3	168	0.6927	0.6723	0.5721	15
4	Baseline	0.6975	0.6342	0.5703	13
5	186	0.7008	0.6676	0.5636	18
6	170*	0.6760	0.6529	0.5611	16
7	165	0.5936	0.6049	0.5442	10
8	152	0.5966	0.5793	0.5433	16
9	155	0.6968	0.6161	0.5315	16
10	178	0.6336	0.6087	0.5289	17
11	169	0.5369	0.5208	0.5181	16
12	193	0.6208	0.6115	0.5145	13
13	163	0.5597	0.5550	0.4983	12
14	187	0.6093	0.5252	0.4847	18
15	154	0.5352	0.5305	0.4748	13
16	196b	0.4964	0.4705	0.4204	16
17	153	0.4984	0.4576	0.3784	14
18	164*	0.4759	0.4317	0.3772	15
19	162	0.4582	0.4335	0.3726	17
20	197	0.5487	0.4264	0.3477	15
21	196c	0.4490	0.4203	0.3441	16
22	196a	0.4911	0.3813	0.3134	15
23	176	0.2832	0.2623	0.2388	13
24	156	0.2933	0.2716	0.2278	9
25	188	0.1542	0.1542	0.1502	11
26	157	0.1017	0.1045	0.1045	13
27	161	0.0867	0.0723	0.0584	14
-	151	0.8728	0.8728	0.8720	5
-	150	0.8493	0.8493	0.7270	3
-	192	0.6020	0.6020	0.6020	2
-	191	0.6173	0.5540	0.5353	3
-	177	0.5227	0.4680	0.4680	3
-	149	0.1880	0.0900	0.0900	4

**Table 3.** Readability results with the relaxed and strict metric.