Overview of the INEX 2011 Question Answering Track (QA@INEX)

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

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

Abstract. The INEX QA track aimed to evaluate complex question-answering tasks where answers are short texts generated from the Wikipedia by extraction of relevant short passages and aggregation into a coherent summary. In such a task, Question-answering, XML/passage retrieval and automatic summarization are combined in order to get closer to real information needs. Based on the groundwork carried out in 2009-2010 edition to determine the sub-tasks and a novel evaluation methodology, the 2011 edition experimented contextualizing tweets using a recent cleaned dump of the Wikipedia. Participants had to contextualize 132 tweets from the New York Times (NYT). Informativeness of answers has been evaluated, as well as their readability. 13 teams from 6 countries actively participated to this track. This tweet contextualization task will continue in 2012 as part of the CLEF INEX lab with same methodology and baseline but on a much wider range of tweet types.

Keywords: Question Answering, Automatic Summarization, Focus Information Retrieval, XML, Natural Language Processing, Wikipedia, Text Readability, Text informativeness.

1 Introduction

Since 2008, Question Answering (QA) track at INEX [7] moved into an attempt to bring together Focused Information Retrieval (FIR) intensively experimented in other INEX tracks (previous ad-hoc tracks [4] and this year snippet track) on the one hand, and topic oriented summarization tasks as defined in NIST Text Analysis Conferences (TAC) [3] on the other hand. Like in recent FIR INEX tasks, the corpus is a clean XML extraction of the content of a dump from Wikipedia. However QA track at INEX differs from current FIR and TAC summarization tasks on the evaluation metrics they use to measure both informativeness and readability. Following [5,10], informativeness measure is based

S. Geva, J. Kamps, and R. Schenkel (Eds.): INEX 2011, LNCS 7424, pp. 188–206, 2012.

on lexical overlap between a pool of relevant passages (RPs) and participant summaries. Once the pool of relevant passages 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 INEX QA track. By contrast, readability evaluation is completely manual and cannot be reproduced on unofficial runs. It is based on questionnaires pointing out possible syntax problems, broken anaphora, massive redundancy or other major readability problems.

Therefore QA tasks at INEX moved from the usual IR *query / document* paradigm towards *information need / text answer*. More specifically, the task to be performed by the participating groups of INEX 2011 was contextualizing tweets, *i.e.* answering questions of the form "what is this tweet about?". The general process involved:

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

We target systems efficient on small terminals like smart phones, based on local resources that do not require a network access, gathering non factual contextual information that is scattered around local resources. Off-line applications on portable devices are useful to reduce the network load and safer.

Answers could contain up to 500 words. It has been required that the answer uses only elements previously extracted from the document collection. Answers needed to be a concatenation of textual passages from the Wikipedia dump.

To constitute the pool of RPs, the informativeness of all returned passages for a subset of 50 tweets has been assessed by organizers. The pool of RPs included all passages considered as relevant by at least one assessor (each passage being submitted to two assessors). We regarded as informative passages that both contain relevant information but also contained as little non-relevant information as possible (the result is specific to the question). This year, long passages including several sentences have often been considered as uninformative because they included too much non relevant information. Furthermore, informativeness of a passage was established exclusively based on its textual content, and not on the documents from which it was extracted. Despite the use of a pool of RPs, the informativeness value of answers did not only rely on the number of its RPs, but also on lexical overlap with other RPs. We found out that evaluating informativeness based on lexical overlap with a pool of RPs is robust if the variety of participant systems is large enough and includes strong baselines.

The paper is organized as follows. Section 2 presents the description of the task. Section 3 details the collection of tweets and documents. Section 4 describes the baseline system provided by the track organizers. Section 5 presents the techniques and tools used for manual evaluation, explains the final choice of metrics and presents results. Finally, Section 6 presents 2012 CLEF "tweet contextualization" task before drawing some conclusions in Section 7.

2 Task Description

The underlying scenario is to provide the user with synthetic contextual information when receiving a message like a tweet. The task is not to find an exact answer in a database of facts, but to bring out the background of the message exclusively based on its textual content. Therefore the answer needs to be built by aggregation of textual passages grasped from the resource (Wikipedia in our case). For some topics, there can be too many relevant passages that cannot be all inserted in the answer, requiring some summarization process that preserves overall informativeness. For others, only few information can be available and the answer should be shorter than expected pointing out the lack of available information.

In this edition, we have considered a recent dump of the Wikipedia. Since we target non factual answers but short contextualizing texts, we removed all the info boxes and the external references, leaving only the textual content with all its document structure (title, abstract, sections, subtitles and paragraphs) and its internal references (links towards other pages).

We wanted to consider only highly informative tweets. In this attempt to define a contextualizing task, we chose to follow the New York Times (NYT) Twitter account. As soon as the NYT publishes an article on its website, it tweets the title of this article, with its URL. We thus considered these tweets. Therefore the task had become *"given a NYT title, find and summarize all available background information in the Wikipedia"*. We also added the first sentence of the related NYT article as a hint, but only few runs used this hint and none of the participants reported using NYT paper content: all tried to tackle the contextualization task in an off-line approach using only the available corpus.

The aggregated answers had a maximum of 500 words each and have been evaluated according to:

- Their informativeness (how much they overlap with relevant passages, Section 5.2),
- Their readability (assessed by evaluators, Section 5.3).

The informativeness of a summary cannot be evaluated without its readability since informative content measures tend to favor syntactic dense summaries. It is often possible to increase an informativeness score by weakening its discursive structure and thus its readability [9].

We provided the participants with a state of the art system derived from [12,2]. Participants had to improve its informative performance without weakening too much the readability of its results.

It was initially announced that readability would be evaluated by participants according to the "last point of interest", *i.e.* the first point after which the text becomes unreadable because of:

- syntactic incoherence,
- unsolved anaphora,

– redundancy,
– other problems.

After discussion between organizers and participants at the INEX 2011 workshop, it was finally decided to disclaim considering only the last point of interest because it relied too much on assessors' subjectivity but to mark all readability issues for every sentence in a summary. It was also decided to evaluate the readability independently from the topic to be contextualized and to read all passages, even if redundant. This increased the workload left to participants in readability evaluation but resulted in a much more refined analysis.

3 Track Data

From 2009 to 2010, QA track at INEX worked on the ad-hoc Wikipedia document collection. In 2009 we considered questions related to ad-hoc topics, and in 2010, real-user, non factual questions from the OverBlog platform[1]. Best performing systems on this task were state of the art automatic summarizers that pick up few Wikipedia pages related to the question and provided a summary as answer.

In 2011, the QA track started experimenting tweets instead of real questions. There the overlap between topics and Wikipedia content becomes much weaker than previously. It was thus decided to move to a more recent and simplified dump of Wikipedia. The new corpus was made available in October 2011 leaving two months to participants for their experiments. This corpus generation process has been completely automatized and can be apply to any XML Wikipedia dump.

3.1 Questions

The question data set was composed of 132 tweets by the NYT released on the July 20th 2011 and having a URL towards the NYT website. Each topic includes the tweet which is often the title of an article just released and the first sentence of the related article. An example is provided below:

```
<topic id="2011005">
  <title>Heat Wave Moves Into Eastern U.S</title>
  <txt>The wave of intense heat that has enveloped much of the
    central part of the country for the past couple of weeks is
    moving east and temperatures are expected to top the 100-degree
    mark with hot, sticky weather Thursday in cities from
    Washington, D.C., to Charlotte, N.C.</txt>
</topic>
```

[1] http://www.over-blog.com/

All these topics were twitted three months after the Wikipedia dump used to build the corpus, therefore we had to manually check if there was any related information in the document collection[2]

3.2 Document Collection

The document collection has been built based on a dump of the English Wikipedia from April 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 the 3,217,015 non empty Wikipedia pages (pages having at least one section).

Resulting documents are made 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 have entities (`t`) that refer to other Wikipedia pages.

Therefore the resulting corpus follows this 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>
```

Figure 1 shows an example of such a cleaned article. We have released the scripts used to generate this corpus. They process any recent XML dump of the Wikipedia in two steps:

- a light `awk` command to remove in a single pass all external references, info boxes and notes using a fast substring extraction function based on index function (GNU implementation of strchr C ISO function).
- a `perl` program that generates the XML using regular expressions to detect and encapsulate document structure and internal links. It also works in a single pass.

[2] The resulting 132 topics come from an initial set of 205 tweets after removing duplicates due to single subjects producing several papers (like different testimonies and opinion papers about the same subject) and only few tweets for which there was no overlap with the Wikipedia. Hence, the 132 selected topics represent more than 64% of the tweets released by the NYT in one day.

```
<?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 &amp;amp; Whitman Shirts, died when he was seven. After that he
was raised solely by his mother, Fannie,  who remained the primary
influence in his early life, even though she remarried when he was
a teenager. In his autobiography, Coburn wrote, &amp;quot;My mother was
a remarkable woman of very strong character who tried to dominate
my life. It was a battle royal all the days of our life
together.&amp;quot;</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>
```

Fig. 1. An example of a cleaned Wikipedia XML article

Once generated, it is necessary to check if the resulting large XML file (between 8 and 12 Gb for recent Wikipedia dumps) is valid. We use the Perl TWIG library by Michel Rodriguez[3] for that. This is a robust library that can process large XML files page by page and fix eventual illformed ones.[4] Current indexers like Indri do not parse such a large XML file and require to split it into pages organized in some folder structure avoiding too large folders. We also made available a Perl program that dispatches Wikipedia pages in 1000 folders. This process can take hours because of numerous file operations.

A complementary list of non-Wikipedia entities has also been made available. The named entities (person, organisation, location, date) of the document collection have been tagged using XIP [1]. For example, for the previous documents, the extracted named entities are:

[3] `http://search.cpan.org/ mirod/`

[4] We had to manually correct few errors on the April 2011 Wikipedia dump due to encoding errors in the original dump file itself, but we did not have error anymore in the last Wikipedia dump from November 2011. For the 2011 INEX edition, we used the corrected April 2011 dump.

```
Alvin Langdon Coburn
1882-1899
Coburn
June 11, 1882
134 East Springfield Street
Boston, Massachusetts
Coburn Whitman
Fannie
Coburn
1890
Los Angeles
Kodak
```

This can be used for participants willing to use named entities in texts but not having their own tagger.

3.3 Submission Requirements

Participants could submit up to three runs. Despite the fact that manual runs were allowed if there was at least one automatic, all submitted official runs have been registered as fully automatic.

Results were lists of passages extracted from the corpus. Two non consecutive passages had to be presented separately. Results in a single run could not include more than 500 words per topic. Any string of alphanumeric characters outside XML tags, without space or punctuation, was considered as a single word.

The format for results was a variant of the familiar TREC format with additional fields:[5]

```
<qid> Q0 <file> <rank> <rsv> <run_id> <column_7> <column_8>
```

where:

- The first column `qid` is the topic number.
- The second column is currently unused and should always be `Q0`. It is just a formating 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.

[5] The XML format to submit results originally proposed in 2010 was dismissed since it was never used by participants because of its useless extra complexity. However if the task evolves in the following years towards more complex results, TREC-like formats will not be sufficient and some XML formating will be required.

- 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 who 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 two columns indicate the selected passage in the document mentioned in the third field. Participants could refer to these passages as File Offset Lengths (FOL) like in usual INEX FIR tasks or directly give the raw textual content of the passage. However, computing character offsets can be tricky dependent on the text encoding and Wikipedia often mixes different encodings. Therefore all participants to this edition chose the alternative raw text format. In this format, each result passage is given as raw text without XML tags and without formatting characters. The only requirement is that the resulting word sequence has to appear at least once in the file indicated in the third field.

Here is an example of such an output:

```
2011001 Q0 3005204 1 0.9999 I10UniXRun1 The Alfred Noble Prize is ...
2011001 Q0 3005204 2 0.9998 I10UniXRun1 The prize was established in ...
2011001 Q0 3005204 3 0.9997 I10UniXRun1 It has no connection to the ...
```

4 Baselines

A baseline XML-element retrieval/summarization system has been made available for participants. The 2011 INEX QA baseline relies on:

- An index powered by Indri[6] that covers all words (no stop list, Krowetz stemming) and all XML tags.
- A PartOfSpeech tagger powered by TreeTagger[7].
- A fast summarizer algorithm powered by TermWatch[8] introduced in [2].
- A summary content evaluation based on FRESA[10].

The Indri index allows to experiment different types of queries to seek for all passages in the Wikipedia involving terms in the topic. Queries can be usual bag of words, sets of weighted multi-word phrases or more complex structured queries using Indri Language[6]. All extracted passages are segmented into sentences and PoS tagged using the TreeTagger. Sentences are then scored using TermWatch based on their *nominals* (i.e. its nouns and adjectives). Let Φ be the set of

[6] http://www.lemurproject.org/
[7] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[8] http://data.termwatch.es

sentences. If for each sentence $\phi \in \Phi$, we denote by φ_ϕ the set of its nominals, then the sentence score Θ_ϕ computed in [2] is:

$$\Theta_\phi = \sum_{\varphi_\phi \cap \varphi_\tau \neq \emptyset}^{\tau \in \Phi} \sum_{\varphi_\tau \cap \varphi_\sigma \neq \emptyset}^{\sigma \in \Phi} |\varphi_\phi \cap \varphi_\tau| \times |\varphi_\tau \cap \varphi_\sigma| \quad (1)$$

The idea is to weight the sentences according to the number of sentences in their neighborhood (sentences sharing at least one nominal). This gives a fast approximation of TextRank or LexRank scores[2]. Sentences are then ranked by decreasing score, only the top ranked are used for a summary of less than 500 words. The selected sentences are then re-ordered following the Indri score of the passage from which they have been extracted and the order of the sentences in these passages. This baseline summary can be computed on the fly, generating the summary taking less time than processing the query by Indri.

This system has been made available online to participants through a web interface[9]. A Perl API running on Linux to query the server was also released. By default, this API takes as input a tabulated file with three fields: topic names, selected output format and query. The output format can be the baseline summary or the first 50 retrieved documents in raw text, PoS tagged or XML source. An example of such a file allowing to retrieve 50 documents per topic based on their title was also released.

The web interface also allows to evaluate the resulting summary or user's one against the retrieved documents using Kullback-Leibler (KL) measure. This content summary evaluation also gives a lower bound using a random set of 500 words extracted from the texts and an upper bound using an empty summary. Random summaries naturally reach the closest word distributions but they are clearly unreadable.

Two baselines were then computed using the approach described in [2] and added to the pool of official submissions:

- Baseline_sum using only topic titles as bag of word queries and top ranked 50 full documents retrieved by Indri to build the summary.
- Baseline_mwt using the same process but returning only the Noun Phrases in the selected sentences to simulate a baseline run for Automatic Terminology Extractors.

5 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. These two properties have been evaluated separately by two distinct measures: *informativeness* and *readability*.

[9] `http://qa.termwatch.es`

5.1 Submitted Runs

23 valid runs by 11 teams from 6 countries (Brasil, Canada, France, India, Mexico, Spain) were submitted. All runs are in raw text format and almost all participants used their own summarization system. Only three participants did not use the online Indri IR engine. Some participants used the Perl API to query the Indri Index with expanded queries based on semantical resources. Only one participant used XML tags.

The total number of submitted passages is 37,303. The median number of distinct passages per topic is 284.5 and the median length in words is 26.9. This relative small amount of distinct passages could be due to the fact that most of the participants used the provided Indri index with its Perl API.

5.2 Informativeness Evaluation

Informativeness evaluation has been performed by organizers on a pool of 50 topics. For each of these topics, all passages submitted have been evaluated. 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. For each topic, all passages from all participants 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 judgment on whether the passage was worth appearing in a summary on the topic, or not. This approach handicaps runs based on short passages extracted from the Wikipedia, since very short passages can be difficult to assess on their own and tend not to be included in the pool of relevant passages.

To check that the resulting pool of relevant answers is sound, a second automatic evaluation for informativeness of summaries has been carried out with respect to a reference made of the NYT article corresponding to the topic. Official evaluation could not be based on these references since most of these articles were still available on the NYT website or could have been used by participants who are NYT readers. Nevertheless, a strong correlation between the ranking based on the assessed pool of relevant passages and the one based on NYT articles would be an indication of assessment soundness.

Metrics. Systems had to make a selection of the most relevant information, the maximal length of the abstract being fixed. Focused IR systems could just return their top ranked passages meanwhile automatic summarization systems need to be combined with a document IR engine. 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 answers 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 [8] and

TAC workshops [3]. Among them, Kullback-Leibler (KL) and Jenssen-Shanon (JS) divergences have been used [5,10] to evaluate the informativeness of short summaries based on a bunch of highly relevant documents.

In this edition we intended to use the KL one with Dirichlet smoothing, like in the 2010 edition[11], to evaluate the informative content of answers by comparing their n-gram distributions with those from all assessed relevant passages. However, in 2010, references were made of complete Wikipedia pages, therefore the textual content was much longer than summaries and smoothing did not introduce too much noise.

This is not the case with the 2011 assessments. For some topics, the amount of relevant passages is very low, less than the maximal summary length. Therefore using any probabilistic metric requiring some smoothing produced very unstable rankings. We thus simply considered absolute log-diff between frequencies. Let T be the set of terms in the reference. For every $t \in T$, we denote by $f_T(t)$ its frequency in the reference and by $f_S(t)$ its frequency in the summary. Adapting the FRESA package available to participants, we computed the divergence between reference and summaries as:

$$Div(T,S) = \sum_{t \in T} |\log(\frac{f_T(t)}{f_T} + 1) - \log(\frac{f_S(t)}{500} + 1)| \quad (2)$$

As T we considered three different sets based on the FRESA sentence segmentation, stop word list and lemmatizer:

- 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.

As in 2010, bigrams with 2-gaps appeared to be the most robust metric. Sentences are not considered as simple bag of words and it is less sensitive to sentence segmentation than simple bi-grams. This is why bigrams with 2-gaps is our official ranking metric for informativeness.

Results. All passages within a consistent pool of 50 topics were thoroughly evaluated by organizers. This represents 14,654 passages, among which 2,801 have been judged as relevant.

This assessment was intended to be quite generous towards passages. All passages concerning a protagonist of the topic are considered relevant, even if the main subject of the topic is not addressed. The reason is that missing words in the reference can lead to artificial increase of the *divergence*, which is a known and not desirable side effect of this measure. Results are presented in Table 1 and statistical significance of gaps between runs are indicated in Table 2.

All systems above the baseline combine a full document retrieval engine with a summarization algorithm. The three top ranked runs, all by IRIT, did not use the API provided to participants meanwhile all other runs improving the

baseline used it only to query the Indri Index, some applying special query expansion techniques. None of the participants used this year the baseline summarization system which ranks 7th among all runs when returning full sentences (`Baselinesum`) and 19th when returning only noun phrases (`Baselinemwt`).

Table 1. Informativeness results from manual evaluation using equation 2 (official results are "with 2-gap")

Rank	Run	unigram	bigram	**with 2-gap**	Average
1	ID12_IRIT_default	0.0486	0.0787	**0.1055**	0.0787
2	ID12_IRIT_07_2_07_1_dice	0.0488	0.0789	**0.1057**	0.0789
3	ID12_IRIT_05_2_07_1_jac	0.0491	0.0792	**0.1062**	0.0793
4	ID129_Run1	0.0503	0.0807	**0.1078**	0.0807
5	ID129_Run2	0.0518	0.0830	**0.1106**	0.0830
6	ID128_Run2	0.0524	0.0834	**0.1110**	0.0834
7	ID138_Run1	0.0524	0.0837	**0.1115**	0.0837
8	ID18_Run1	0.0526	0.0838	**0.1117**	0.0839
9	ID126_Run1	0.0535	0.0848	**0.1125**	0.0848
10	Baselinesum	0.0537	0.0859	**0.1143**	0.0859
11	ID126_Run2	0.0546	0.0863	**0.1144**	0.0863
12	ID128_Run3	0.0549	0.0869	**0.1151**	0.0868
13	ID129_Run3	0.0549	0.0869	**0.1152**	0.0869
14	ID46_JU_CSE_run1	0.0561	0.0877	**0.1156**	0.0876
15	ID46_JU_CSE_run2	0.0561	0.0877	**0.1156**	0.0876
16	ID62_Run3	0.0565	0.0887	**0.1172**	0.0887
17	ID123_I10UniXRun2	0.0561	0.0885	**0.1172**	0.0885
18	ID128_Run1	0.0566	0.0889	**0.1174**	0.0889
19	Baselinemwt	0.0558	0.0886	**0.1179**	0.0887
20	ID62_Run1	0.0566	0.0892	**0.1180**	0.0892
21	ID123_I10UniXRun1	0.0567	0.0895	**0.1183**	0.0894
22	ID62_Run2	0.0572	0.0900	**0.1188**	0.0899
23	ID124_UNAMiiR12	0.0607	0.0934	**0.1221**	0.0933
24	ID123_I10UniXRun3	0.0611	0.0946	**0.1239**	0.0945
25	ID124_UNAMiiR3	0.0628	0.0957	**0.1248**	0.0957

Dissimilarity values are very closed, however differences are often statistically significant as shown in table 2. In particular, top four runs are significantly better than all others. It seems that these runs carried out specific NLP post-processing. It also appears that almost all runs above Baselinesum are significantly better than those under the same baseline, meanwhile differences among runs ranked between the two baselines are rarely significant.

To check that this reference was not biased, the same 50 topics have been also automatically evaluated against the corresponding NYT article, *i.e.* taking as reference the article published under the tweeted title. None of the participants reported having used this content even though part of it was publicly available on the web.

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

	ID12_IRIT_default	ID12_IRIT_07_2_07_1_dice	ID12_IRIT_05_2_07_1_jac	ID129_Run1	ID129_Run2	ID128_Run2	ID138_Run1	ID18_Run1	ID126_Run1	Baselinesum	ID126_Run2	ID128_Run3	ID129_Run3	ID46_JU_CSE_run1	ID46_JU_CSE_run2	ID62_Run3	ID123_I10UniXRun2	ID128_Run1	Baselinemwt	ID62_Run1	ID123_I10UniXRun1	ID62_Run2	ID124_UNAMiiR12	ID123_I10UniXRun3	ID124_UNAMiiR3
ID12_IRIT_default	-	-	1	-	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID12_IRIT_07_2_07_1_dice	-	-	1	-	1	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID12_IRIT_05_2_07_1_jac	1	1	-	-	1	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID129_Run1	-	-	-	-	2	1	3	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID129_Run2	2	1	1	2	-	-	-	-	-	3	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
ID128_Run2	2	2	2	1	-	-	-	-	-	1	2	3	2	2	2	3	3	3	3	3	3	3	3	3	3
ID138_Run1	2	2	2	3	-	-	-	-	-	1	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
ID18_Run1	3	2	2	2	-	-	-	-	-	-	-	-	1	1	1	3	3	3	3	3	3	3	3	3	3
ID126_Run1	3	3	3	2	-	-	-	-	-	-	-	-	-	-	-	2	2	3	3	3	3	3	3	3	3
Baselinesum	3	3	3	3	3	1	1	-	-	-	-	-	-	-	-	-	2	1	3	2	2	3	3	3	3
ID126_Run2	3	3	3	3	2	2	2	-	-	-	-	-	-	-	-	-	1	2	2	2	2	3	3	3	3
ID128_Run3	3	3	3	3	2	3	2	-	-	-	-	-	-	-	-	-	-	1	1	1	1	2	3	3	3
ID129_Run3	3	3	3	3	2	2	2	1	-	-	-	-	-	-	-	-	-	-	1	1	1	2	3	3	3
ID46_JU_CSE_run1	3	3	3	3	2	2	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	3	3
ID46_JU_CSE_run2	3	3	3	3	2	2	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	3	3
ID62_Run3	3	3	3	3	3	3	3	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	2	3	3
ID123_I10UniXRun2	3	3	3	3	3	3	3	3	2	2	1	-	-	-	-	-	-	-	-	-	-	1	3	3	3
ID128_Run1	3	3	3	3	3	3	3	3	3	1	2	1	-	-	-	-	-	-	-	-	-	-	2	3	3
Baselinemwt	3	3	3	3	3	3	3	3	3	3	2	1	1	-	-	-	-	-	-	-	-	-	3	3	3
ID62_Run1	3	3	3	3	3	3	3	3	3	2	2	1	1	-	-	-	-	-	-	-	-	-	2	3	3
ID123_I10UniXRun1	3	3	3	3	3	3	3	3	3	2	2	1	1	-	-	-	-	-	-	-	-	-	2	3	3
ID62_Run2	3	3	3	3	3	3	3	3	3	3	3	2	2	-	-	-	1	-	-	-	-	-	1	3	3
ID124_UNAMiiR12	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	3	2	3	2	2	1	-	-	3
ID123_I10UniXRun3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	-	-
ID124_UNAMiiR3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	-	-

Results are presented in Table 3. It appears that correlation between the two rankings is quite high (Kendall's $\tau = 0.67$, Pearson's product-moment correlation = 88%, p-value $< 9.283e^{-9}$) suggesting that our approach of selecting reference text from a pool of participant runs plus the baselines is sufficient.

All previous evaluations have been carry out using FRESA package which includes a special lemmatizer. We provided the participants with a standalone evaluation toolkit based on Potter Stemmer. Based on participant feedback after

the release of the official results, we introduced in this package a normalized ad-hoc dissimilarity defined as following using the same notations as in equation 2:

$$Dis(T,S) = \sum_{t \in T} \frac{f_T(t)}{f_T} \times \left(1 - \frac{\min(\log(P), \log(Q))}{\max(\log(P), \log(Q))}\right) \quad (3)$$

$$P = \frac{f_T(t)}{f_T} + 1 \quad (4)$$

$$Q = \frac{f_S(t)}{f_S} + 1 \quad (5)$$

The idea is 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.

Table 4 shows results using this evaluation toolkit implementing basic stemming and normalized dissimilarity 3. Again, the correlation with official results in

Table 3. Informativeness results automatic evaluation against NYT article using equation 2

Rank	Run	unigram	bigram	**with 2-gap**	Average
1	ID12_IRIT_05_2_07_1_jac	0.0447	0.0766	**0.1049**	0.0766
2	ID12_IRIT_07_2_07_1_dice	0.0447	0.0767	**0.1049**	0.0766
3	ID12_IRIT_default	0.0447	0.0767	**0.1049**	0.0767
4	ID129_Run1	0.0456	0.0777	**0.1060**	0.0777
5	ID18_Run1	0.0462	0.0779	**0.1061**	0.0779
6	Baselinesum	0.0460	0.0781	**0.1065**	0.0781
7	ID126_Run1	0.0460	0.0781	**0.1065**	0.0781
8	ID128_Run2	0.0461	0.0782	**0.1066**	0.0782
9	ID138_Run1	0.0461	0.0782	**0.1066**	0.0782
10	ID129_Run2	0.0468	0.0788	**0.1071**	0.0787
11	ID129_Run3	0.0468	0.0789	**0.1072**	0.0788
12	ID126_Run2	0.0469	0.0789	**0.1073**	0.0789
13	ID128_Run3	0.0469	0.0789	**0.1073**	0.0789
14	ID123_I10UniXRun1	0.0471	0.0791	**0.1075**	0.0791
15	Baselinemwt	0.0475	0.0794	**0.1077**	0.0794
16	ID62_Run1	0.0473	0.0793	**0.1077**	0.0793
17	ID128_Run1	0.0475	0.0795	**0.1079**	0.0795
18	ID62_Run3	0.0476	0.0796	**0.1080**	0.0796
19	ID62_Run2	0.0477	0.0797	**0.1080**	0.0797
20	ID123_I10UniXRun2	0.0477	0.0797	**0.1080**	0.0797
21	ID123_I10UniXRun3	0.0483	0.0804	**0.1087**	0.0803
22	ID46_JU_CSE_run1	0.0487	0.0807	**0.1089**	0.0806
23	ID46_JU_CSE_run2	0.0487	0.0807	**0.1090**	0.0807
24	ID124_UNAMiiR12	0.0493	0.0812	**0.1094**	0.0812
25	ID124_UNAMiiR3	0.0505	0.0823	**0.1104**	0.0823

Table 4. Informativeness results from manual evaluation using Potter stemmer and normalized dissimilarity 3

Rank	Run	unigram	bigram	**with 2-gap**
1	ID12_IRIT_default	0.8271	0.9012	**0.9028**
2	ID126_Run1	0.7982	0.9031	**0.9037**
3	ID12_IRIT_07_2_07_1_dice	0.8299	0.9032	**0.9053**
4	ID129_Run1	0.8167	0.9058	**0.9062**
5	ID12_IRIT_05_2_07_1_jac	0.8317	0.9046	**0.9066**
6	ID128_Run2	0.8034	0.9091	**0.9094**
7	ID138_Run1	0.8089	0.9150	**0.9147**
8	ID129_Run2	0.8497	0.9252	**0.9253**
9	ID126_Run2	0.8288	0.9306	**0.9313**
10	ID128_Run3	0.8207	0.9342	**0.9350**
11	Baselinesum	0.8363	0.9350	**0.9362**
12	ID18_Run1	0.8642	0.9368	**0.9386**
13	ID129_Run3	0.8563	0.9436	**0.9441**
14	ID46_JU_CSE1	0.8807	0.9453	**0.9448**
15	ID46_JU_CSE2	0.8807	0.9452	**0.9448**
16	ID128_Run1	0.8379	0.9492	**0.9498**
17	ID62_Run3	0.8763	0.9588	**0.9620**
18	ID123_I10UniXRun2	0.8730	0.9613	**0.9640**
19	ID62_Run1	0.8767	0.9667	**0.9693**
20	ID62_Run2	0.8855	0.9700	**0.9723**
21	ID123_I10UniXRun1	0.8840	0.9699	**0.9724**
22	ID124_UNAMiiR12	0.9286	0.9729	**0.9740**
23	Baselinemwt	0.9064	0.9777	**0.9875**
24	ID124_UNAMiiR3	0.9601	0.9896	**0.9907**
25	ID123_I10UniXRun3	0.9201	0.9913	**0.9925**

Table 1 is quite high (Kendall's τ =89%, Pearson's product-moment correlation = 96%, p-value $< 4e^{-11}$).

This normalized metric does not allow to distinguish between top ranked runs above the baseline as shown by statistical significance tests reported in table 5 but it does among runs between the two baselines.

5.3 Readability Evaluation

Human Assessment. Each participant had to evaluate readability for a pool of around 50 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:

- *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,

Table 5. Statistical significance for manual evaluation using Potter stemmer and normalized dissimilarity in table 4 (t-test, 1 : 90%, 2 = 95%, 3 = 99%, $\alpha = 5\%$)

	ID12_IRIT_default	ID126_Run1	ID12_IRIT_07_2_07_1_dice	ID129_Run1	ID12_IRIT_05_2_07_1_jac	ID128_Run2	ID138_Run1	ID129_Run2	ID126_Run2	ID128_Run3	Baseline_sum	ID18_Run1	ID129_Run3	ID46_JU_CSE_run2	ID46_JU_CSE_run1	ID128_Run1	ID62_Run3	ID123_I10UniXRun2	ID62_Run1	ID62_Run2	ID123_I10UniXRun1	ID124_UNAMiiR12	Baseline_mwt	ID124_UNAMiiR3	ID123_I10UniXRun3
ID12_IRIT_default	-	-	-	-	1	-	-	-	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3
ID126_Run1	-	-	-	-	-	-	-	1	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3
ID12_IRIT_07_2_07_1_dice	-	-	-	-	-	-	-	-	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
ID129_Run1	-	-	-	-	-	-	-	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID12_IRIT_05_2_07_1_jac	1	-	-	-	-	-	-	-	1	1	1	2	2	2	2	3	3	3	3	3	3	3	3	3	3
ID128_Run2	-	-	-	-	-	-	-	1	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3
ID138_Run1	-	-	-	-	-	-	-	-	2	1	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID129_Run2	-	1	-	2	-	1	-	-	-	-	-	-	1	2	2	2	3	3	3	3	3	3	3	3	3
ID126_Run2	2	2	1	2	1	2	2	-	-	-	-	-	-	-	-	2	2	3	3	3	3	3	3	3	3
ID128_Run3	2	2	1	2	1	2	1	-	-	-	-	-	-	-	-	1	2	3	3	3	3	3	3	3	3
Baseline_sum	2	2	2	3	1	2	2	-	-	-	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3
ID18_Run1	2	2	2	3	2	2	3	-	-	-	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3
ID129_Run3	2	3	2	3	2	3	3	1	-	-	-	-	-	-	-	-	1	2	3	3	3	3	3	3	3
ID46_JU_CSE_run2	3	3	2	3	2	3	3	2	-	-	-	-	-	-	-	-	1	2	3	3	3	2	3	3	3
ID46_JU_CSE_run1	3	3	2	3	2	3	3	2	-	-	-	-	-	-	-	-	1	2	3	3	3	2	3	3	3
ID128_Run1	3	3	3	3	3	3	3	2	2	1	-	-	-	-	-	-	-	2	3	3	3	2	3	3	3
ID62_Run3	3	3	3	3	3	3	3	3	2	2	2	2	1	1	1	-	-	-	-	-	-	-	3	3	3
ID123_I10UniXRun2	3	3	3	3	3	3	3	3	3	3	3	3	2	2	2	2	-	-	1	2	2	-	3	3	3
ID62_Run1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	1	-	-	-	-	3	3	3
ID62_Run2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	2	-	-	-	-	2	3	3
ID123_I10UniXRun1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	2	-	-	-	-	2	3	3
ID124_UNAMiiR12	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	2	-	-	-	-	-	-	1	3	2
Baseline_mwt	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	1	-	-	-
ID124_UNAMiiR3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	-	-
ID123_I10UniXRun3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	-	-	-

– *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 at 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.

The assessors did not know the topic corresponding to the summary, and were not supposed to judge the relevance of the text. Only readability was evaluated.

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

- **Relaxed metric:** a passage is considered as valid if the T box has not been ticked,
- **Strict metric:** a passage is considered as valid if no box has been ticked.

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

A total of 1,310 summaries, 28,513 passages from 53 topics have been assessed. All participants succeeded in evaluating more than 80% of the assigned summaries. The resulting 53 topics include all of those used for informativeness assessment. Results are presented in Table 6.

None of the submitted participant runs outperformed Baselinesum (Baseline with complete summaries). This can be explained by the fact that formula 1 favors sentences with numerous Multi Word Noun Phrases. These particular sentences tend to be long, with few pronouns, thus few broken anaphora. The drawback of this Baseline is that building an extract of 500 words made of long sentences will be always less informative than a dense coherent summary made of non redundant short sentences. Therefore participants runs had to improve informativeness without hurting readability too much.

The other baseline restricted to Multi Word Noun Phrases was considered as unreadable by most assessors except by one who is a specialist in terminology and considered as acceptable any NP that corresponds to a real Multi Word Term.

6 2012 "Tweet Contextualization" Campaign

In 2012, this campaign will be integrated into CLEF (Conference and Labs of the Evaluation Forum) under the title "tweet contextualization". The aim of this task will be close to 2011 campaign, still using the most recent cleaned dump of the Wikipedia (November 2011).

About 100 tweets will be collected manually by the organizers from Twitter. They will be 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 will be provided. These tweets will be used for manual evaluation, but will be scatterred into 1000 other tweets, automatically collected from Twitter Search API. This will ensure that systems provide fully automatic runs.

The tweets will be made available in a JSON format, as shown in Figure 2.

In 2012, there will be no more automatic evaluation of informativeness since we do not have any reference, as it was the case in 2011 using NYT articles. However we showed that results between manual and automatic evaluations were pretty close.

The informativeness evaluation will be performed by organizers. The readability evaluation will still be performed by organizers and participants. Only the

Table 6. Readability results with the relaxed and strict metric

Relaxed metric			**Strict metric**		
Rank	Run id	Score	Rank	Run id	Score
1	Baseline_sum	447.3019	1	Baseline_sum	409.9434
2	ID46_JU_CSE_run1	432.2000	2	ID129_Run1	359.0769
3	ID128_Run2	417.8113	3	ID129_Run2	351.8113
4	ID12_IRIT_default	417.3462	4	ID126_Run1	350.6981
5	ID46_JU_CSE_run2	416.5294	5	ID46_JU_CSE_run1	347.9200
6	ID129_Run1	413.6604	6	ID12_IRIT_05_2_07_1_jac	344.1154
7	ID129_Run2	410.7547	7	ID12_IRIT_default	339.9231
8	ID12_IRIT_05_2_07_1_jac	409.4038	8	ID12_IRIT_07_2_07_1_dice	338.7547
9	ID12_IRIT_07_2_07_1_dice	406.3962	9	ID128_Run2	330.2830
10	ID126_Run1	404.4340	10	ID46_JU_CSE_run2	330.1400
11	ID138_Run1	399.3529	11	ID129_Run3	325.0943
12	ID128_Run1	394.9231	12	ID138_Run1	306.2549
13	ID129_Run3	393.3585	13	ID128_Run3	297.4167
14	ID126_Run2	377.8679	14	ID126_Run2	296.3922
15	ID128_Run3	374.6078	15	ID62_Run2	288.6154
16	ID62_Run2	349.7115	16	ID128_Run1	284.4286
17	ID62_Run1	328.2245	17	ID62_Run3	277.9792
18	ID62_Run3	327.2917	18	ID62_Run1	266.1633
19	ID18_Run1	314.8980	19	ID18_Run1	260.1837
20	ID123_I10UniXRun2	304.1042	20	ID123_I10UniXRun1	246.9787
21	ID123_I10UniXRun1	295.6250	21	ID123_I10UniXRun2	246.5745
22	ID123_I10UniXRun3	272.5000	22	ID123_I10UniXRun3	232.6744
23	ID124_UNAMiiR12	255.2449	23	ID124_UNAMiiR12	219.1875
24	ID124_UNAMiiR3	139.7021	24	Baseline_mwt	148.2222
25	Baseline_mwt	137.8000	25	ID124_UNAMiiR3	128.3261


```
{"created_at":"Fri, 03 Feb 2012 09:10:20 +0000",
"from_user":"XXX",
"from_use_id":XXX,
"from_use_id_str":"XXX",
"from_use_name":"XXX",
"geo":null,
"id":XXX,
"id_str":"XXX",
"iso_language_code":"en",
"metadata":"result_type":"recent",
"profile_image_url":"http://XXX",
"profile_image_url_https":"https://XXX",
"source":"<a href=''http://XXX",
"text":"blahblahblah",
"to_user":null,
"to_use_id":null,
"to_use_id_str":null,
"to_use_name":null}
```


Fig. 2. Example of a tweet, in JSON format

textual content of the tweet should be contextualized. For example, contextualization of a tweet from Barack Obama account, concerning war in Syria, should not come into details about Obama's life, or US elections, but only on the tweet text.

7 Conclusion

This track that brings together the NLP and the IR communities is getting more attention. The experimented measures used for evaluation based on textual content more than passage offsets seem to reach some consensus between the two communities. Taking into account readability of summary also encourages NLP and linguistic teams to participate. Next edition will start much earlier, the corpus generation from a Wikipedia dump being now completely automatic. We plan to propose a larger variety of questions from twitter. We also would like to encourage XML systems by providing more structured questions with explicit name entities and envisage to open the track to terminology extractor systems.

References

1. At-Mokhtar, S., Chanod, J.P., Roux, C.: Robustness beyond shallowness: Incremental deep parsing. Natural Language Engineering 8, 121–144 (2002)
2. Chen, C., Ibekwe-Sanjuan, F., Hou, J.: The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. JASIST 61(7), 1386–1409 (2010)
3. Dang, H.: Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In: Proc. of the First Text Analysis Conference (2008)
4. Geva, S., Kamps, J., Trotman, A. (eds.): INEX 2009. LNCS, vol. 6203. Springer, Heidelberg (2010)
5. Louis, A., Nenkova, A.: Performance confidence estimation for automatic summarization. In: EACL, pp. 541–548. The Association for Computer Linguistics (2009)
6. Metzler, D., Croft, W.B.: Combining the language model and inference network approaches to retrieval. Inf. Process. Manage. 40(5), 735–750 (2004)
7. Moriceau, V., SanJuan, E., Tannier, X., Bellot, P.: Overview of the 2009 qa track: Towards a common task for qa, focused ir and automatic summarization systems. In: Geva et al. [4], pp. 355–365
8. Nenkova, A., Passonneau, R.: Evaluating content selection in summarization: The pyramid method. In: Proceedings of HLT-NAACL, vol. 2004 (2004)
9. Pitler, E., Louis, A., Nenkova, A.: Automatic evaluation of linguistic quality in multi-document summarization. In: ACL, pp. 544–554 (2010)
10. 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), pp. 1059–1067. Chinese Information Processing Society of China (2010)
11. 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 2010. LNCS, vol. 6932, pp. 269–281. Springer, Heidelberg (2011)
12. SanJuan, E., Ibekwe-Sanjuan, F.: Combining language models with nlp and interactive query expansion. In: Geva, et al. [4], pp. 122–132