Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 00 (2016) 1–23
Inf. Proc. Man.

INEX Tweet Contextualization Task: Evaluation, Results and Lesson Learned

Patrice Bellot[a], Véronique Moriceau[b], Josiane Mothe[d], Eric SanJuan[c], Xavier Tannier[b]

[a]*Aix-Marseille Université - CNRS - Univ. Toulon - ENSAM (LSIS - UMR 7296), France, patrice.bellot@lsis.org*
[b]*Univ. Paris-Sud, Université Paris-Saclay, LIMSI, CNRS, France, FirstName.Name@limsi.fr*
[c]*LIA, Université d'Avignon, France, eric.sanjuan@univ-avignon.fr*
[d]*Institut de Recherche en Informatique de Toulouse, UMR5505 CNRS, Université de Toulouse, France, josiane.mothe@irit.fr*

Abstract

Microblogging platforms such as Twitter are increasingly used for on-line client and market analysis. This motivated the proposal of a new track at CLEF INEX lab of *Tweet Contextualization*. The objective of this task was to help a user to understand a tweet by providing him with a short explanatory summary (500 words). This summary should be built automatically using resources like Wikipedia and generated by extracting relevant passages and aggregating them into a coherent summary.

Running for four years, results show that the best systems combine NLP techniques with more traditional methods. More precisely the best performing systems combine passage retrieval, sentence segmentation and scoring, named entity recognition, text part-of-speech (POS) analysis, anaphora detection, diversity content measure as well as sentence reordering.

This paper provides a full summary report on the four-year long task. While yearly overviews focused on system results, in this paper we provide a detailed report on the approaches proposed by the participants and which can be considered as the state of the art for this task. As an important result from the 4 years competition, we also describe the open access resources that have been built and collected. The evaluation measures for automatic summarization designed in DUC or MUC were not appropriate to evaluate tweet contextualization, we explain why and depict in detailed the LogSim measure used to evaluate informativeness of produced contexts or summaries. Finally, we also mention the lessons we learned and that it is worth considering when designing a task.

Keywords: Short text contextualization, tweet contextualization, tweet understanding, automatic summarization, contextual information retrieval, question answering, focus information retrieval, natural language processing, Wikipedia, text readability, text informativeness, textual references, Kullback-Leibler divergence

1. Introduction

1.1. Contextualization

Generally speaking context is defined as the circumstances that surround a particular idea, event, or situation; the settings that make it understandable.

Context is a key component for understanding language-in-use, since as pointed out by Gee, "*When we speak or write we never say all that we mean. Spelling everything out in words explicitly would take far too long. [...] Speakers and writers rely on listeners and readers to use the context in which things are said and written to fill in meanings that are left unsaid, but assumed to be inferable from context.*" [36]. This is definitively true when it comes to tweets.

Tweet authors rely on their followers to find out or know about the context to complete the 140-character messages and understand them.

In discourse analysis, context refers to the fact that texts are historically produced and interpreted, they are located in time and space. Both linguistic and extralinguistic clues help in understanding speeches and writings such as culture, society and ideology [85]. Indeed, social representation which relies upon collective frames of perception plays an important role and Van Dijk suggests context models which encapsulate knowledge, either personal, from a group, or cultural knowledge, attitudes, and ideologies [81].

According to Wodak 's triangulatory approach, the concept of context takes into account four levels [85]:

1. the text-internal co-text (Internal references, logically sequenced events ...),
2. the intertextual and interdiscursive relationship between utterances, texts, genres and discourses,
3. the extralinguistic (social) level,
4. the socio-political and historical contexts.

In the same way, in Information Retrieval (IR), "context" often refers to surrounding environment in which a search occurs. The search environment consists of the user (his knowledge, search skills, level of education...), the task the search is done for, the domain, the system used... [44, 20, 42]. Context-sensitive IR systems correspond to systems that use other elements than just the current query and the document collection [74] and this brings specific challenges when it comes to evaluation [75]. Advanced user location or users' interest [10], query personalization [17], or previous queries from the same user can be used to retrieve documents she or he has already seen and wants to find again [28] or on the contrary to diversify the retrieved items [13]. Otherwise, the documents previously selected by users for similar queries (also called collaborative filtering) [83, 45] and latent interests for users or groups [84] among others can also be used in contextual IR.

Text contextualization differs from text expansion in that it aims at helping a human to understand a text rather than a system to better perform an automated task. For example, in the case of query expansion in IR, the idea is to add terms to the initial query that help the system to better retrieve relevant documents. Even if term selection can be based on contextual information to be more efficient, the main objective is not to present the expanded query to the user [61, 15]. Text contextualization on the contrary can be viewed as a way to provide more information on the corresponding text in the objective of making it understandable and to relate this text to information that explains it.

Context is being invoked to interpret the textual context of a passage or co-text which is defined as words or sentences surrounding a given unit that help to understand the meaning of the unit. In the following we define contextualization as **providing the information that is useful to fully understand that text**.

1.2. Contribution

Contextualizing micro-blogging is particularly useful since 140-character long messages are rarely self-explanatory. This motivated the proposal of a new track of Tweet Contextualization at INEX[1] in 2011 [71] which became a CLEF Lab[2] in 2012 [72] and that we fully depict in this paper. In addition, we provide a deep analysis of the results and lessons learned.

The use case is as follows: given a tweet, the user wishes to get a better understanding of it by reading a short textual summary. In addition, the user should not have to query any system and the system should use a resource freely available. More specifically, the guidelines 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 resource while being generally reliable.

In this paper, we detail the task and the evaluation collection as well as the approaches and the results obtained by the participants between 2011 and 2014 with a focus on sentence retrieval. While yearly overviews focused on system results, the long-term survey presented in this paper allows to go beyond them and to reinforce their results. We provide a detailed report on the approaches proposed by the participants and which can be considered as the state of the art for this task. The main results are as follows:

[1] INitiative for the Evaluation of XML Retrieval, `http://inex.mmci.uni-saarland.de/`.
[2] Conference and Labs of the Evaluation Forum, `http://www.clef-initiative.eu/`.

1. Effective sentence retrieval, as defined in [4], is a challenging key sub-task for tweet contextualization when achieved over a large resource.
2. Sentences are too short to be considered as atomic documents and contextual on-line text analysis is more efficient than extensive off-line sentence pre-processing and indexing.
3. Best systems have to combine passage retrieval, sentence segmentation and scoring, named entity recognition, text part-of-speech (POS) analysis, anaphora detection, diversity content measure as well as sentence reordering.
4. Global informativeness of summaries can be evaluated through sentence pooling and n-gram measures, if readability is assumed.
5. The robustness of the results against query variability, measure choice and reference incompleteness suggest that this evaluation methodology could be extended to other resources than microblogs and the Wikipedia.

The remaining of this paper is organized as follows: in Section 2 we present the Tweet Contextualization task as well as related tasks. In Section 3 we describe the evaluation metrics, both from text summarization and the new LogSim measure used to evaluate tweet contextualization. Section 4 presents the variety of the approaches the participants developed. Section 5 discusses the results and draw some conclusions on the choices made by the participants. Section 6 concludes the paper.

2. Contextualization of Tweet using Wikipedia Resource

This section presents in details the CLEF Tweet Contextualization which ran for 4 years in the framework of INEX and CLEF evaluation forum. We also and first present the related tasks that ran in various evaluation programs.

2.1. Related Tasks from other Evaluation Programs

A typical tweet contextualizer needs at least to be able to identify the key elements from a tweet, to find related external texts (here Wikipedia pages) and to summarize them. The user is not required to submit a query to any system: rather, the contextualizer must be able to find out useful information, starting from the tweet only. Tweet Contextualization is defined by providing the users with useful information to make the tweet more understandable, that is to say to clarify it. This task corresponds to a combination of information extraction, information retrieval, entity linking, search in structured collections and topic-oriented summarization. For this reason, this task is linked to some other tasks organized in evaluation campaigns such as TExt Retrieval Conference (TREC), Document Understanding Conferences (DUC), Text Analysis Conferences (TAC) and Cross-Language Evaluation Forum now Conference and Labs of the Evaluation Forum (CLEF), or Initiative for the Evaluation of XML Retrieval (INEX).

Topic-oriented summarization tasks had been evaluated in NIST TAC (2008-2011)[3] [22] and during the previous DUC (2001-2007)[4]. One of the goals of TAC 2011 Summarization Tasks was to write a 100-word summary covering all the important aspects (who, why, how...) of newswire articles for a given topic and to write a summary update of subsequent newswire articles, preventing redundancy. Summaries were evaluated for readability and fluency (grammaticality, referential clarity, focus, coherence...), as well as content and overall responsiveness. However the size of the document collections considered in these tracks were small. It was then possible, even though time consuming, to build extensive manual references to evaluate informativeness. Handling very large document collections was out of the scope of these tasks meanwhile it was one of the main aims of INEX tracks. In 2012, TAC focused on Knowledge Base Population by means of three tasks. Among them, the aim of the Entity-Linking Task was to determine the node of a collection derived from Wikipedia that is related to a given name, itself illustrated by a given document. Tweet Contextualization also used a XML corpus derived from the Wikipedia including an annotation of entities and document structure.

Other tasks involve several types of treatments: the purpose of TREC 2012 Contextual Suggestion Track[5] [24] for example was to provide users with suggestions (for example places to visit or where to have a drink) according to

[3] http://www.nist.gov/tac/
[4] http://duc.nist.gov
[5] http://sites.google.com/site/treccontext/

user's profiles (preferences) and a context defined as geotemporal location. Tweets used as queries in Tweet Contextualization often included a location or a temporal information. However, the output needed to be a readable summary and not a list of entities. In 2012, TREC Knowledge Base Acceleration Track[6] started. It aims at filtering a document stream to extract the relevant documents to a set of entities. It had much in common with document filtering, except that the input was entities (described by articles from Wikipedia) rather than topics. Moreover, the streaming organization of the collection (time-stamped news, blogs and web pages) involved the possibility of evolution of the target entities (modification of their attributes or relations). In the past, the purpose of the TREC Entity Track was to perform entity-oriented searches in the Web [6]. For the last run in 2011, the track was subdivided into two tasks: related entity finding (return a list of entities of a specified type and related to a source entity) and entity list completion (return a list of entities from a set of example entities) [5]. The corpus used for the Entity Track was the ClueWeb09 English [19] (more than one billion web pages). Compared to TREC, the Tweet Contextualization corpus is much more reduced (6M documents) and easy to index using state of the art tools like Indri or Terrier. However the number of queries to process in Tweet Contextualization has been much higher since 2012 (between 400 and 1200 tweets). So only automatic methods could participate but extensive Natural Language Processing (NLP) and intelligent indexing could be experimented on the entire document collection. However, corpus pre-processing needed also to be fully automatic since the corpus was updated every year and best results always relied on the most recent corpus because the tweets were selected few months after the dump of the Wikipedia used to build the document collections.

It is worth mentioning the High Accuracy Retrieval from Documents (HARD) TREC track that was last run in 2005 [2]. It took into account some contextual information and the expertise level of the searcher and of the documents. Its purpose was to achieve high accuracy retrieval from topics describing the searcher and the context of the query. This context was the purpose of the search (looking for background, details or a precise answer), the familiarity of the searcher with the topic, the granularity of the retrieved items (documents, passages, sentences...).

Lastly, in INEX, Focused Information Retrieval track (FIR) considers information extraction and search in structured collections. FIR has been intensively experimented in the previous ad-hoc track [38] and the more recent Snippet Track [80].

Tweet Contextualization followed these ideas of providing a background explaining a short message. However, it differed from a usual FIR task in the sense that participants had to produce a coherent summary. The overall informativeness of the summary could be much higher than the sum of each single item informativeness.

2.2. CLEF Tweet Contextualization Task

The CLEF Tweet Contextualization task we introduced in 2011 focuses on tweets. The task aims at providing the tweet reader with information that makes it understandable. The resulting information is provided to the reader under the form of a 500 words long summary, built from textual resource extracts.

For instance, the tweet:

```
Bobby Brown -- Fighting #WhitneyHouston's Family to See Bobbi Kristina
```

could be contextualized by the summary presented in Figure 1. Such a summary is expected to provide some background about all elements from the tweet, that could require an explanation.

In the following, we describe 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. We also present the reference system we used in the evaluation. All the resources are made available for research purpose at `http://tc.talne.eu`.

2.3. Dataset

2.3.1. Document Collection

In 2011, the document collection has been built based on a dump of the English Wikipedia from April 2011. For the 2012 edition a dump from November 2011 was used whereas for 2013 and 2014 a dump of the English Wikipedia from November 2012 was used. The date of the dump was anterior to all selected tweets.

[6] `http://trec-kba.org/`

Whitney Elizabeth Houston (August 9, 1963 – February 11, 2012) was an American recording artist, actress, producer, and model. Houston was one of the world's best-selling music artists, having sold over 170 million albums, singles and videos worldwide. Robert Barisford "Bobby" Brown (born February 5, 1969) is an American R&B singer-songwriter, occasional rapper, and dancer. After a three-year courtship, the two were married on July 18, 1992. On March 4, 1993, Houston gave birth to their daughter Bobbi Kristina Houston Brown, her only child, and his fourth. With the missed performances and weight loss, rumors about Houston using drugs with her husband circulated. Following fourteen years of marriage, Brown and Houston filed for legal separation in September 2006. Their divorce was finalized on April 24, 2007, with Houston receiving custody of their then-14-year-old daughter. On February 11, 2012, Houston was found unresponsive in suite 434 at the Beverly Hilton Hotel, submerged in the bathtub.

Figure 1. Example of contextualization summary for the tweet *"Bobby Brown – Fighting #WhitneyHouston's Family to See Bobbi Kristina"*. All sentences come from different Wikipedia pages.

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 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. Figure 2 shows an 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 &amp;amp; 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>
```

Figure 2. Extract of a Wikipedia page.

2.3.2. Tweets as Topics

In 2011, topics were made of 53 tweets from New York Times (NYT). The text of each tweet was actually the title of a newly published NYT article, along with the URL of this article.

In 2012 and 2013, the task was made more realistic and evaluated topics were made of 63 (2012) and 70 (2013) tweets each year, manually collected by the 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.

For example, the following tweet was considered as contextualizable:

```
Very cool!  An interactive animation of van Gogh's "The Starry Night." http://t.co/ErJCPObh
(thanks @juliaxgulia)
```

while this one was not (not informative, not contextualizable from the Wikipedia):

```
Mom:  "All you do is sit on that computer all day!" Me:  "Lies!  I sit on the chair."
```

From 2012, from the same set of accounts, about 1,000 tweets each year were then collected automatically. These tweets were added to the evaluation set, in order to avoid that fully manual, or not enough systems could achieve the task. All tweets were to be treated by participants, but a short list only was used for evaluation. Indeed, human evaluation is time consuming and can only hardly be done at large scale with trustable results. 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. Figure 3 gives an example of such topic.

```
"created_at":"Wed, 15 Feb 2012 23:32:22 +0000",
"from_user":"FOXBroadcasting",
"from_user_id":16537989,
"from_user_name":"FOX Broadcasting",
"id":169927058904985600,
"text":"Tensions are at an all-time high as the @AmericanIdol
Hollywood Round continues, Tonight at 8/7c.  #Idol",
"to_user":null,
(...)
```

Figure 3. Extract of a tweet given as topic.

In 2013 we considered more diverse types of tweets than in 2012, so that participants could better measure the impact of hashtag processing on their approaches for example.

Between 2011 and 2013 the corpus did change every year but not the use case. In 2014, the official document collection was the same as in 2013 but the user case evolved. Indeed the 2014 tweet contextualization topics were a selection of 240 tweets from RepLab 2013 [3], it was thus necessary to use prior Wikipedia dumps. Some participants also used the 2012 corpus raising up the question of the impact of updating Wikipedia over these tasks.

Again, these tweets were 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.

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 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.

2.4. Reference System

A state of the art XML-element retrieval/summarization system called qa.termwatch (qaTW)[69] combining a Question-Answering engine based on multiword term extraction [70] powered by Indri with a summarization system based on terminology graphs [18] has been made available for participants through an online interface and a perl API both available at `http://tc.talne.eu`.

The system is available online through a web interface[7] that allows to query:

- an index powered by Indri[8] that covers all words (no stop list, no stemming) and all XML tags,
- a PartOfSpeech tagging powered by TreeTagger [73][9],
- the state of the art summarization algorithm used in [18].

Three kinds of queries can be used: standard bag of words, sets of multiword terms or more complex structured queries using Indri language [55]. The system returns the first 50 documents retrieved by Indri in raw text format, PoS tagged text or XML document source.

3. Evaluation Metrics

In this section we present both related evaluation metrics used in text summarization and the new metrics we have defined. We also explain why the metrics from the literature were not fully relevant for Tweet Contextualization.

3.1. Evaluation Metrics for Summaries

As we explained in the previous section, Tweet Contextualization is clearly related to text summarization. We present here the measures commonly used for evaluating textual summaries. We also point out the advantages and drawbacks of the different metrics.

The evaluation of summaries includes at least two basic dimensions: the information content of the summary or its informativeness and the linguistic quality which is related to readability. Informativeness and linguistic quality are two dimensions which are equally important even if somehow independent: as informative as it is, in the sense of the *amount of information* it provides, the informativeness of a totally illegible text (incomprehensible by a human) should not be considered as relevant, and, conversely, readability should not be considered in isolation but in the limited context of a corpus of documents to summarize. Evaluation of these two facets of the quality of a summary cannot be fully automated.

However, it became possible to automatically evaluate summary informativeness by comparing it with either one or more reference summaries or directly with the original texts. When reference summaries are used, they are generally produced manually and constitute a *gold standard* like the relevance judgments in *ad-hoc* IR collections (TREC Qrels[10] for example). This type of assessment has been widely used especially in DUC surveys, then in TAC [23].

The evaluation of the informativeness of produced summaries can be done by estimating the percentage of information from the reference summaries which is actually in the summary constructed automatically as it was done in DUC 2003 using the SEE tool[11] or in more sophisticated approaches like in DUC from 2004. Indeed, from 2004, informativeness evaluation at DUC was based on the sentences and n-grams that compose both the reference summaries and the automatically built summary to evaluate. *ROUGE* (Recall-Oriented Understudy for Gisting Evaluation) measure [48] uses this principle and was then also used in TAC program. *Pyramid* is another method to assess summaries [59] in which, instead of comparing distributions of n-grams, key concepts are first manually extracted from the reference summaries, then the occurrence of these key concepts in the automatically generated summary indicates the quality of the produced summary.

[7] `http://qa.termwatch.es/`
[8] `http://www.lemurproject.org/`
[9] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
[10] `http://trec.nist.gov/data/qrels_eng/`
[11] `http://www.isi.edu/licensed-sw/SEE/`

These evaluation measures, however, are based on the existence of reference summaries produced by humans. This feature limits the scalability of this approach.

In the case of a very large number of documents to sum up, it is useful to find a measure that can be used automatically to compare the contents of the produced summary with the one of the full set of documents to sum up. In this framework, the measures used in TAC compare the distributions of words or word sequences between the produced summary and the documents to sum up using measures such as *Kullback-Leibler* (KL) and *Jensen-Shannon* (JS) divergence measures.

An interesting point on ROUGE, KL, and JS measures is that it has been shown that they correlate with a fully manual evaluation -under the condition the summaries are readable. Without this condition, it is possible to improve artificially the results obtained when using these measures and produce a summary that does not make sense to a human (sequence of terms, for example).

In the following, we present the Pyramid, ROUGE, KL, and JS measures. We will explain why they are not convenient to the tweet contextualization task in section 3.2 ; we will present a new measure we have proposed and which fits better the task in the same section.

3.1.1. Identifying Informational Nuggets

In the context of complex Query Answering tasks, [59] introduced the pyramidal evaluation; this measure has then been extended to evaluate summary informativeness. The pyramidal method is based on SCUs (Summary Content Units) or Nuggets that are originally defined manually by annotators. SCUs are information units gathering various linguistic expressions that correspond to the same content. These nuggets are weighted according to the number of annotators who identified them. Initially, the construction of the SCUs was manual and dependent on the inter-annotator agreement [49]. Recent work in IR has shown that SCUs can be generalized to the automatic creation of references [63], based on more general information units or nuggets extracted automatically [41] [29]. For example, in the case of Wikipedia, inner links to other pages can be considered as natural nuggets as the target pages often contain reformulations of the target entity. However, adopting such an approach as the measure of informativeness tends to reduce the task of creating a summary to a simple entity search and extraction. This explains why we rather focus on measures based on n-gram distributions that we detail below; these measures do not depend on the definition of nuggets or SCUs.

3.1.2. ROUGE Measures

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) matches a set of measures that compare a summary automatically produced with one (or more) reference summary generally produced manually [48]. The general principle used in ROUGE is to count the number of items, such as n-grams, that are common between the reference summaries and the summary to be evaluated.

A first set of ROUGE variants is the $ROUGE_n$ family that compares the n-grams as follows [48]:

$$ROUGE(S) = \frac{\sum_{R \in \mathcal{R}} |gram_n(R) \cap gram_n(S)|}{\sum_{R \in \mathcal{R}} |gram_n(R)|}$$

where:

- $\mathcal{R}$ is the set of reference summaries,
- $gram_n(T)$ is the set of n-grams in a text T.

Usually $n \leq 2$ but skip-grams (i.e. ordered pairs of words in a sliding window of n consecutive words) can also be considered with $2 \leq n \leq 4$.

$ROUGE_n$ is recall oriented and $ROUGE_1$ is similar to the cosine measure, but for $n > 1$, because the word order is taken into account, ROUGE is more strict than cosine [1]. Summaries evaluated with $ROUGE_n$ are supposed to all have the same length k, often fixed between 100 and 150 words. This tends to favour summaries that maximize the number of different words (vocabulary diversity) but include the largest number of common words shared by several reference summaries (general vocabulary) against summaries that use low frequency but more specific vocabulary. Selecting the top k frequent words in document sources allows to optimize $Rouge_n$ score. To avoid this drawback it

is necessary to simultaneously use extended stop word lists that remove too frequent words while manually checking that summaries are readable.

Under the triple assumption of completeness of the set of reference summaries, readability of summaries to be evaluated and the existence of an appropriate stop word list, the correlation between manual and automatic evaluations using ROUGE on summaries varies from 0.80 to 0.94 [51]. When manual complete references cannot be produced because the number of source documents is too large as in IR, the use of ROUGE is less appropriated.

Indeed, DUC and after that TAC used ROUGE but evaluation was on no more than fifty documents. Assessors had to read all the documents and produce a summary as comprehensive as possible and of a predefined size. For the assessment to be trustworthy with ROUGE, it is necessary to produce at least five reference summaries prepared by different experts [51]. Moreover, it is not possible to use ROUGE to assess the informativeness of a document for summaries of various size. These drawbacks motivated the following new alternative evaluation approaches.

3.1.3. Kullback-Leibler and Jensen-Shannon Divergences

In the case of a very large number of documents to sum up, a measure should be found that compares the content of automatic summary with the set of documents to be summarized. The hypothesis is that the closer a summary is to the documents, the more informative it is.

A quite intuitive solution is to compare distributions of words or word sequences between the summary and documents [57]. Kullback-Leibler (KL) and Jensen-Shannon (JS) divergence measures calculate the similarity between two distributions of probabilities $P(i)$ and $Q(i)$.

More precisely, KL measure is defined by:

$$D_{KL}(P||Q) = \sum_i \ln \frac{P(i)}{Q(i)} P(i)$$

JS measure is a symetric and smoothed variant of KL and is defined by:

$$D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$$

where

$$M = \frac{1}{2}(P + Q)$$

JS correlates the best with ROUGE [51], but JS is sensitive to the size of the summaries. It is therefore difficult to apply it when the size of the produced summaries may vary depending on the amount of relevant information found in the document collection. On the other hand, KL is not defined for zero probability, thus it is necessary to introduce smoothing techniques initially introduced in language models.

ROUGE relies on a small set of reference summaries and thus can be very sensible to the absence / presence of words in the reference summaries. On the contrary, in KL there is no reference and the whole document is taken into account, any word counting somehow. Even though ROUGE measures may be too sensitive to the presence / absence of word sequences that occur both in references and produced summary, KL has the opposite disadvantage not to allow to distinguish between a very low frequency and a total absence. Moreover, KL and JS only allow to assess generic summaries and are difficult to adapt to the case when summaries are guided by a query or a topic, as is the case for tweet contextualization. To apply these measures in the context of tweet contextualization, it would be necessary to create summary references using relevant passages. This reference would be made up of extracts from documents and would not be a summary as in the case of ROUGE. The difficulty of adapting KL and JS to the tweet contextualization case is that there is a large variability in terms of the size of the references themselves. When just few relevant passages in documents exist, estimating probabilities using smoothing becomes very difficult [69]. We thus proposed new measures that were used to evaluate tweet contextualization. They are presented in the next section.

3.2. Evaluation of Tweet Contextualization

Tweet contextualization 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.

Following [51, 68], informativeness measure is based on lexical overlap between a pool of relevant passages and participants' automatically generated summaries. Once the pool of relevant passages is constituted, the process is automatic and can be applied to unofficial runs. The definition of this new and more adapted informativeness measure and the release of these pools are two of the main contributions of Tweet Contextualization track at INEX [69].

By contrast, readability is evaluated manually 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.

3.2.1. LogSim to Evaluate Informativeness

According to the task definition, systems have to make a selection of the most relevant information, the maximal length of the abstract being fixed. Therefore outputs cannot be any passage of the corpus, but at least well formed sentences. As a consequence, summary informativeness cannot be evaluated using standard IR measures since tweet contextualization systems do not try to find all relevant passages but to select those that could provide a comprehensive answer.

To build a textual reference, we both considered:

- passages pooled from participants,
- text from manual runs by the organizers.

Passages from participant runs have been filtered out by the organizers, removing manually non informative ones. For a passage to be included in the textual reference it was sufficient to have been considered by one of the assessors. Identical sentences were merged together. Therefore each passage informativeness has been evaluated independently from others, even if they were pooled from the same summary.

Most of passages submitted by participants were well formed sentences. This method favored long self content sentences. However short sentences could be partially included in longer ones. Informativeness of runs is then evaluated based on their overlap with the resulting textual reference. To measure this overlap we remove functional words, apply Porter stemmer and consider stemmed word n-grams.

In Section 3.1, we detailed the various measures that have been defined and experimented with at DUC (Document Understanding Conferences) [60] and TAC (Text Analysis Conferences) workshops [22]. Among them, Kullback-Leibler (KL) and Jensen-Shannon (JS) divergences have been used [51, 68] to evaluate the informativeness of short summaries based on a bunch of highly relevant documents.

However, JS is very sensitive to the length of summaries meanwhile this is an important parameter in Tweet Contextualization. If there is little relevant information in the corpus, participants are supposed to submit abstracts with less than 500 words, the maximal length. KL is non symmetric and therefore is supposed to be less sensitive to abstract length variability except in the case where the reference made of a bunch of relevant passages is shorter than the abstract to be evaluated. In this case ranking is unstable and relies on the smoothing method used to compute probabilities. As a result, differences between scores are rarely significant.

We first intended to use the KL measure with Dirichlet smoothing to evaluate the informative content of answers by comparing their n-gram distributions with those from all assessed relevant passages. However, summaries in the Tweet Contextualization task are short and thus smoothing resulted in too much noise.

Moreover, for some topics or tweets, the amount of relevant passages can be very low, less than the maximal summary length. Therefore using any probabilistic metric requiring some smoothing would produce very unstable rankings.

After experimenting these metrics in the Question-Answering track of INEX (QA@INEX 2009 and 2010) [57], [69], a variant of absolute log-diff between frequencies was proposed in the Tweet Contextualization task introduced in 2011 at INEX workshop: it approximates KL when the reference is large and remains stable when the reference is shorter than the maximum summary size. This ad-hoc metric emerged as the most consensual among participants and has become in 2011 the official metrics for informativeness evaluation for the Tweet Contextualization track at CLEF INEX lab 2012 as well as in subsequent years.

Let R be the set of terms in the reference summaries and A the set of terms in the abstract to be evaluated. For every $t \in R$, we denote by $f_R(t)$ its frequency in the reference, by $f_A(t)$ its frequency in the summary and for $X \in \{R, A\}$ we

denote by $P(t|X)$ the conditional probability $\frac{f_X(t)}{f_X}$. The metric then stands as the following similarity measure between abstract and reference:

$$LogSim(R, A) = \sum_{t \in R} P(t|R) \times \frac{\min(\log(1 + P(t|R)), \log(1 + P(t|A)))}{\max(\log(1 + P(t|R)), \log(1 + P(t|A)))} \quad (1)$$

Let us point out that this similarity is highly correlated to the number of tokens that appear both in the reference and in the abstract $|R \cap A|$ since for any token t not in A, $P(t|A) = 0$ so the denominator in the LogSim formula is null. Since summaries have a maximal length of 500 tokens, LogSim can be considered as a graded Interpolate Precision measure. Indeed, $LogSim(R, A)$ increases with the Interpolated Precision at 500 tokens where precision is defined as the number of n-grams word in the reference. It appears that all reported participant rankings would have been similar using Precision instead of LogSim. However the introduction of the log increases robustness to deal with highly frequent words and the use of a graded degree of informativeness per token $P(t|R)$ improves robustness against the introduction of too specific vocabulary in the reference.

Three different sets have been considered as R based on Porter stemmer:

- 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 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.

3.2.2. Readability

We believe that the evaluation of readability requires a human evaluation. For Tweet Contextualization, this evaluation was made by the organizers and participants. Assessors indicate when readability was weak because of highly incoherent grammatical structures, unsolved anaphora, or redundant passages. These criteria are inspired from the coherence and cohesion criteria defined by [40], [65] and [33].

Each participant had to evaluate readability for a pool of abstracts on an online web interface. Each summary consisted of a set of passages and for each passage, assessors had to check four kinds of check boxes. The guideline was the following:

- *Syntax* (Syn): check the box if the passage contains a syntactic problem (bad segmentation for example).
- *Anaphora* (Ana): check the box if the passage contains an unsolved anaphora.
- *Redundancy* (Redun): check the box if the passage contains redundant information, i.e. information that has already been given in a previous passage.
- *Trash* (Trash): check the box if the passage does not make any sense in its context (*i.e.* after reading the previous passages). Then, these passages must not be considered 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, check all trash boxes until the end of the summary.

To evaluate summary readability, the number of words (up to 500) in valid passages was considered. Three metrics were used:

- **Relevancy (or Relaxed) metric**, counting passages where the Trash box has not been checked
- **Syntax**, counting passages where the Syn box was not checked either.
- The **Structure (or Strict) metric** counting passages where no box was checked at all.

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

Table 1. Some features about runs.

Year	#runs	#teams	median # dist. passages per topic	median length in words
2011	23	11	284.5	26.9
2012	33	13	171.5	28
2013	24	13	295.5	31

4. Participant Hits and Failures

4.1. Data

Main features about data collected between 2011 and 2013 are provided in Table 1.

A total number of 16 independent teams from 12 countries submitted at least one run between 2011 and 2013: Brazil (2), Canada (1), Chile (1), France (2), Germany (1), India (2), Ireland (1), Mexico (2), Russia (1), Spain (2), USA (1), Vietnam (1). The diversity of participant teams ensured the diversity of methods, systems and resources used to produce the runs. However, if none of the participant teams did use the reference summarizer system available through a web API, half of them did use the reference search engine provided by organizers. This points out that half of the participant teams were pure NLP research groups. A possible drawback of having half of the participants using the same reference IR engine could have been a biased reference because based on participant passage pooling. However, we shall show that this was avoided.

Two third of the teams participated twice and thus had the opportunity to tune their systems. In 2011, topics were news titles twitted by the New York Times and therefore easy to analyze but in 2012 topics were diversified including real tweets by individuals and the median number of distinct passages submitted per topic marked a significant drop.

Almost all systems that participated to Tweet Contextualization track adopted a three-steps approach:

1. Wikipedia documents are indexed.
2. Tweets are translated into queries to be submitted to an IR engine and a subset of documents is selected. The first retrieved documents are then split into sentences.
3. Some of the sentences are selected; they are ordered to produce the summary.

However, a variety of techniques have been used in these three common steps. We give hereafter some of the hits and failures in these three steps and point the most original approaches. In the result section (section 5), we provide detailed results and summarize the most effective approaches encapsulating the various steps.

4.2. Indexing

All participants but one used Indri[12], Terrier[13] or Lucene[14] to index the Wikipedia but in different ways.

The Université de Toulouse suggests in 2011 an approach based on an index which includes not only lemmas, but also named entities (NE) [32] and other NLP-based features which results in a theoretical representation based on several vectors for each sentence. Combined with a specific sentence ranking and ordering, this model got the best effectiveness in 2011 and inspired other works.

Only the speech group from Avignon (LIA) did not use an index but experimented in 2012 tweet reformulation based on Latent Dirichlet Allocation (LDA) [56]. For that, they build a complete LDA model of the English Wikipedia that achieved similar performances than the reference system as reported in the section results, Table 3. Meanwhile, systems that relied on external indexing tools not closely related to the NLP modules were outperformed after 2011.

[12]Lemur project search engine based on language modeling, a cooperative effort between the University of Massachusetts Amherst and Carnegie Mellon University. `http://ciir.cs.umass.edu/research/indri/`

[13]Terrier IR Platform developed at the School of Computing Science, University of Glasgow http://terrier.org/

[14]The Apache Lucene search engine `https://lucene.apache.org/core/`

4.3. Querying

The robustness and the speed of the online reference system mostly relies on simple multiword term extraction based on POS tagging, however without tweet preprocessing it misses important entities hidden in hashtags or acronyms. Hashtags are authors' annotation on key terms of their tweets. Many participants went further and tried to tackle these issues.

For example, LINA group used its own improved Wikipedia index combined with advanced hashtag preprocessing and URL analysis [26, 27].

The IULA group of the Universitat Pompeu Fabra experimented in 2011 a graph-based approach, the REG system [79] based on Indri. In their best try, queries are composed with the title of the tweet expanded with title-related terms and named entities. Terms and named entities are extracted manually from tweets, as well as relatedness among Wikipedia pages. Even though, author describes an algorithm for computing automatically relatedness (simple cooccurrence-based measure), which was not used in the runs. The authors sent two other runs that got lower performances: in the first one, tweets are simply considered as queries and submitted to the system to retrieve the documents that will be used to build the summary; in the last one, only terms and named entities related to those from the title as well as redirection from Wikipedia were used.

In Université de Toulouse (IRIT) runs, the first retrieved documents as well as tweets were parsed using the Stanford CoreNLP [52][15]. Tweets were transformed into queries through POS tagging and recognized named entities. This allows taking into account different weights for different tokens within a query, *e.g.* named entities were considered to be more important than common nouns; nouns were more significant than verbs; punctuation marks were considered as not valuable, ...

The Leibniz Universität Hannover participated in 2013 and used the ArkTweet [39] toolkit for tweet tokenization, POS tagging and phrase chunking. For a given tweet, each phrase was then submitted as a query to Indri in order to retrieve relevant documents.

4.4. Summarizing

Most participants used state of the art automatic summarizers like Cortex system [68] in 2011. Cortex was initially developed at the École Polytechnique de Montréal (EP) and then at Avignon (LIA). It was first used for complex question answering task in [78]. The authors then released the toolkit under GPL license and it has been used by other participants to build their own summaries, often in combination with the Indri index provided by the organizers. Cortex combines several measures (frequency, entropy, Hamming measures, title similarity) in order to score sentences and finally choose the best sentences and combine them into a summary. After 2011, EP and LIA runs based on Cortex experimented a new stemmer based on morphological segmentation and affixality calculation. This stemmer needs an unsupervised training on a corpus of raw text, which size varied from 100,000 to 500,000 words. The system has also been updated with special focus on the readability of summaries which is considered as the weaker point of statistical summaries [77]. Among the users of Cortex system, there are :

- In 2011, the UAM team improved sentence selection based on combinatoric analysis [47]. More specifically, the authors applied the greedy optimization algorithm on the traversal graph of sentences without improving at this time general Cortex results.
- In 2012, the UNAM team [54, 53] analyzed the impact of stemming and found that ultra-stemming did not downgrade Cortex results.
- In 2013, the Universitat Federal do Ceará [50] tried to improve Cortex and related systems by filtering and simplifying tweets but this lead to remove topics selected by the organizers to be included in the reference pool and lose non textual tweet features.

Other existing systems were experimented.

[15] The Stanford NLP-based parser `nlp.stanford.edu:8080/corenlp/`

- In 2011, the TALN group from the Universitat Pompeu Fabra used Indri to retrieve documents and then, the SUMMA toolkit was used to select the relevant sentences based on features such as similarity between the sentence and the tweet/the document, term frequency, etc [67]. Results were closed to the reference system as reported in next Table 2.
- In 2011, Jadavpur University from Kolkata [12] considered retrieving paragraphs instead of sentences. The idea was to adapt the system introduced in [62] and used as part of the participation in the Paragraph Selection (PS) Task and Answer Selection (AS) Task of QA@CLEF 2010 – ResPubliQA [64]. It appeared that the use of paragraphs improves readability but does not allow informativeness optimization [72]. The following year, the Indian Statistical Institute of Kolkata [7] reached similar conclusions about passage retrieval informativeness performance. Direct sentence retrieval was also experimented in 2013 by IRIT as an extra run. It was only efficient on retrieving sentences overlapping tweet terms [8].
- The Leibniz Universität Hannover participated in 2013. After using the ArkTweet [38] toolkit for tweet tokenization, POS tagging and phrase chunking, for a given tweet, each phrase was submitted as a query to Indri. The selected sentences were given as an input to the MEAD toolkit, a multi-document summarization system which implements a centroid-based approach [66].
- In 2013, Jadavpur University participated with an offline graph-based multi-document summarization [11]. They first used the Lucene search engine to retrieve 10 most relevant Wikipedia documents for each entity, independently from the tweet topics. Then they generated a multi-document summary for each entity. This process does not rely on the tweets and that is why they call it an offline process. The online process consisted in extracting entities from the query tweet, and considering offline-generated summaries for these entities as the selected sentences. Finally, these sentences were re-ranked by a weighting heuristic based on three values: 1/ the frequence of terms from the sentences and from the query, 2/ the number of common entities between the sentences and the query, and 3/ a specific weight for the title field.

However, it appeared that the best performing systems for Tweet Contextualization re-implemented special summarizers.

- At IRIT, the similarity between the query and sentences were computed using *tf-idf* measure considering multi-vector sentence representation. Sentence similarity was smoothed by local context: the nearest sentences produce more effect on the target sentence sense than others. The authors tried various weighting as well as various similarity measures. Three runs were submitted. In their best run, which was also the best run over the participants that year, named entities were considered with a coefficient 1.0; abstract had weight equal to 1.0, sections had score 0.8; headers, labels, and other components were not taken into account; stop-words were removed ; cosine similarity was applied; POS were ranked; each term frequency was multiplied by its inverse document frequency (*idf*). Changing the similarity measure to Dice or Jaccard and reducing the weight of the section part result in decreasing effectiveness. These results can be explained by different language models and by the features of the pool of the relevant passages. Moreover, in 2011, the sentences with the highest score were simply added to the summary until the total number of words exceeds 500. For 2012, the method was modified by adding bigram similarity, anaphora resolution, hashtag processing and sentence reordering [31]. Sentence ordering task was modeled as a sequential ordering problem, where vertices corresponded to sentences and sequential constraints were represented by sentence time stamps. More specifically, for each query and each sentence the linear combination of the unigram and bigram cosine was computed by assigning the weight 0.3 and 0.7 to unigram and bigram similarity measure respectively. This process did not result in improving the results; maybe because the sentence ordering is not strongly considered in informativeness and readability measures.
- DCU Dublin [35], University of Minnesota Duluth[21] and University of Nantes [25] also re-implemented similar approaches to IRIT.

4.5. Checking Readability

The impact of some classical measures of readability in the selection of sentences answering to topics has been studied by several participants [76, 32] since 2011. They considered the readability as an indicator of cohesion of the

summaries extracted from documents. Related work concerned the estimation of linguistic quality, or the fluency, of a text such as employed in some automatic summarization systems [58]. However, they were more interested in the detection of text easier to understand than text well written. There are many challenges: determining a good measure of readability for the task, achieving a good balance between a selection of phrases according to their informational relevance (quantity of new information) and their readability. For example, the LSIS from the University of Marseille [76] tried to improve summary readability by combining informativeness measures with Flesch and Dale-Chall readability measures [16] but found out that these document oriented measures are not efficient at sentence level. Ermakova also considers Flesch reading ease test in addition to lexical diversity, meaningful word ratio and punctuation ratio [30].

5. Results

5.1. Reusable Informativeness Assessments

One of the main outcomes of Tweet Contextualization tracks has been the release of reusable assessments to evaluate text informativeness. This has been performed by the organizers on a pool of 53 topics in 2011, 63 topics in 2012, and 70 topics in 2013. For each topic (tweet) in the evaluation pool, best 60 passages based on the provided Retrieval Status Value (RSV) score [43] of all submitted runs were extracted and merged together. After removing duplicates, all 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 judgment on whether the passage was worth appearing in a summary contextualizing the tweet, or not. For example, in 2012 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 relevant passage was 30.03 tokens. Dissimilarity values are very close, however differences are often statistically significant (details are not provided here but can be found in [72] for 2012 and in [8] for 2013).

The soundness of this procedure has been verified on 2011 data. For the 2011 track, topics were tweets coming from New York Times, and a full article was associated to each tweet. To check that the resulting pool of relevant answers was appropriate, a second automatic evaluation for informativeness of summaries was then carried out, taking as the reference the text content of the NYT article. None of the participants reported any use of the complementary information available on the NYT website. Both rankings, one based on submitted run pooling and the second one based on NYT articles, appeared to be highly correlated (Pearson's product-moment correlation = 88%, p-value$< 10^{-6}$). Details are provided in [71].

5.2. Informativeness Results

Table 2 presents the results obtained by the best runs for the best 5 systems in 2011 using informativeness as presented in Section 3.2. or top 5 best automatic participant systems. Runs are ranked by decreasing informativeness dissimilarity. The scores of the reference system described in 2.4 is denoted by *qaTW*. The three first system results are not significantly different one from the other; but they are different from the following ones. In the same way UPF and Avignon systems are not statistically different from the reference system qaTW. The best system IRIT used a Vector Model (VM) IR system different from the Language Model (LM) based system proposed by organizers as reference system. Moreover, it uses special weighting for POS, named entities, document structural elements and a smoothing from local context when calculating sentence/tweet cosine similarity. The second system (Montréal) combined the reference LM engine with Cortex [79], an advanced statistical summarizer system, meanwhile the third system (UAM) combined the same LM engine with a combinatoric optimization approach to select a cluster of sentences to be included in the summary minimizing a Jensen-Shannon divergence with the retrieved documents. These three systems significantly outperformed the provided reference. The approach by Avignon based on XML passage retrieval (Wikipedia article abstracts, sections and paragraphs) instead of sentence retrieval produced results lower than the baseline but these differences are not statistically significant.

In 2012, the task became more complex with a high diversity of tweets. Table 3 presents the 2012 results in terms of informativeness. Only one system (Duluth) outperformed the reference qaTW provided by organizers. The Duluth system also used a LM based IR and textrank like summarizer based on sentence scoring as the reference one but it also included simple rules to handle tweet specific contents, meanwhile the reference system was kept unchanged

Table 2. Informativeness results 2011 (official results are "with 2-gaps"): VM = Vector Model, B = provided Indri Index , CO = Combinatorial optimization, LM = Language Model). Horizontal line indicates strong statistical significance (t-test p-value < 0.001)

Rank	Run	Team	System	unigram	bigram	with 2-gap
1	143	IRIT	VM + Parser	0.8271	**0.9012**	**0.9028**
3	129	Montréal	B +*Cortex*	**0.8167**	0.9058	0.9062
5	131	UAM	B+CO	0.8034	0.9091	0.9094
9	-	-	**qaTW**	0.8363	0.935	0.9362
10	133	UPF	B+*SumUM*	0.8818	0.9630	0.9634
13	139	Avignon	LM + XML	0.8767	0.9667	0.9693

Table 3. Informativeness results 2012 (official results are "with 2-gaps") for top 10 best automatic participant systems (LM= Language Model, SIM= Text Similarity Metrics, Pre = Tweet preprocessing, LDA=Latent Dirichlet Allocation, VM=Vector Model, B= provided Indri Index, MWT = MultiWord Terms, SMA = Social Media Analysis). Horizontal lines indicate strong statistical significance (t-test p-value < 0.001)

.

Rank	Run	Team	System	unigram	bigram	with 2-gap
1	178	Duluth	LM + Sim	**0.7734**	**0.8616**	**0.8623**
4	-	-	**qaTW**	0.7864	0.8868	0.8887
4	169	Nantes	Pre + LM	0.7959	0.8881	0.8904
6	193	Avignon	LDA	0.7909	0.8920	0.8938
7	185	Dublin	LM + Parser	0.8265	0.9129	0.9135
11	154	UNAM	LM + Sim	0.8233	0.9254	0.9251
12	162	Montréal	B + *Cortex*	0.8236	0.9257	0.9254
15	196b	IRIT	VM + Parser	0.8484	0.9294	0.9324
21	165	IULA	MWT + LM + *REG*	0.8818	0.9630	0.9634
22	150	Kolkata	VM + Parser	0.9052	0.9871	0.9868
27	157	Konstanz	SMA	0.9715	0.9931	0.9937

Table 4. Informativeness results 2013 (official results are "with 2-gaps") for top 10 best automatic participant systems (Pre = Tweet preprocessing, LM = Language Model, Sim = Text Similarity Metrics, VM = Vector Model, TA = Syntax and Discourse Text Analysis, B = provided Indri Index, MWT = Multiword Terms). Horizontal line indicates strong statistical significance (t-test p-value < 0.001)

Rank	Run	Team	System	unigram	bigram	with 2-gap
2	258	Nantes	Pre + LM	**0.7939**	**0.8908**	**0.8943**
3	275	IRIT	VM + Parser	0.8061	0.8924	0.8969
7	254	Duluth	LM + Sim	0.8331	0.9229	0.9242
8	-	-	**qaTW**	0.8169	0.9270	0.9301
9	270	Jadavpur	VM + TA	0.8481	0.9365	0.9397
13	277	Kolkata	VM + Parser	0.8995	0.9649	0.9662
14	261	Montréal	B + *Cortex*	0.8639	0.9668	0.9670
17	262	IULA	MWT + LM + *REG*	0.8738	0.9734	0.9747
19	265	Ceará	Pre + B	0.8793	0.9781	0.9789
22	266	Leibniz	*LS3*	0.9059	0.9824	0.9835
23	269	Chaoyang	VM	0.9965	0.9999	0.9999

over the 2011-2013 tracks. Two other system reached scores closed to the reference. The Nantes system went further on tweet pre-processing combined with a specific LM IR engine. The Avignon system tried a complete different approach based on extensive LDA model of Wikipedia to extract sentences. All other systems were significantly outperformed including all those that tried to combine the reference IR engine provided by organizers with advanced summarizer methods suggesting that optimizing the integration of IR and NLP components was a key point. It seems at least necessary that all components use the exact same text representation (tokens, punctuation, numbers etc.) for query processing (tweet), passage retrieval (Wikipedia articles) and summarization. The best system (IRIT) in 2011 was outperformed in 2012 due to the handling of tweet specificity. A better tuning provided again significantly outstanding results in 2013.

Table 4 presents the informativeness results in 2013 track. Tweets proposed as topics in 2013 were as numerous than in 2012 but the organizers manually removed spams. Most of the selected tweets required however specific preprocessing like hashtag analysis meanwhile in 2011 they were press titles. Average performance of systems relying on the Indri index provided by organizers is also significantly lower in 2013. Indeed, the reference system that does not implement special tweet preprocessing and which is based on the provided Indri index, is significantly outperformed by best participant systems. Best systems closely integrates tweet preprocessing based on rules (pre), LM document search and a simple summarizer based sentence scoring. Second best system combined a lighter tweet pre-processing with a VM document search but used syntax analysis (Parser) to improve summary readability as reported in the following section. Systems relying on the provided LM model were outperformed by the reference system using the same index.

Overall, all participants but two used language models, however informativeness of runs that only used passage retrieval is under 5% [7]. Combining document retrieval with sentence segmentation performs better than passage retrieval and reaches 12% of informativeness with an acceptable level of readability (above 70%). Terminology extraction and reformulation applied to tweets were also used in 2011 and 2012 [82]. The resulting run being among the best 10 runs for informativeness and readability. But appropriate stemming [54] and robust parsing of both tweets and Wikipedia pages were also an important issue in 2012 [35]. All best systems in informativeness used the Stanford Core NLP tool or the TreeTagger.

5.3. Readability Evaluation by Participants

Runs are ranked by increasing readability (see Section 3.2). In 2011, a total of 1,310 summaries, 28,513 passages from 53 tweets have been assessed by participants. All participants succeeded in evaluating more than 80% of the assigned summaries. The resulting 53 tweets include all of those used for informativeness assessment. The assessors did not know the tweet corresponding to the summary, and were not supposed to judge the relevance of the text. Only readability was evaluated and assessors had to read and point out syntax errors for every sentence or passage. This was time consuming and favored pure passage retrieval runs based on paragraph marks <p>. Yet, informativeness

Table 5. Readability results 2011 with the relaxed and strict metric.

Relaxed metric			**Strict metric**		
Run	team	Score	Run	team	Score
143	IRIT	83.4%	129	Montréal	71.8%
129	Montréal	82.8%	131	UAM	70.2%
131	UAM	80.8%	143	IRIT	68.8%
139	Avignon	70%	139	Avignon	57.8%
133	UPF	63%	133	UPF	52%

Table 6. Readability results in 2012 with the relaxed and strict metric.

Relaxed metric			**Strict metric**		
Run	team	Score	Run	team	Score
185	Dublin	77.28%	185	Dublin	64.46%
178	Duluth	63.36%	165	IULA	54.42%
193	Avignon	62.08%	178	Duluth	52.89%
165	IULA	59.36%	169	Nantes	51.81%
169	Nantes	53.69%	193	Avignon	51.45%
154	UNAM	53.52%	154	UNAM	47.48%
196b	IRIT	49.64%	196b	IRIT	42.04%
162	Montréal	45.82%	162	Montréal	37.26%
157	Konstanz	10.17%	157	Konstanz	10.45%

of a single paragraph was generally low, relevant passages being scattered among several Wikipedia articles. The readability results for best informativeness runs in Table 2 are presented in Table 5.

In 2012, a total of 594 summaries from 18 tweets have been assessed. The resulting 18 tweets are included in those used for informativeness assessment. For each summary, the tweet was displayed, and readability was thus evaluated in the context of the tweet. Passages not related to the tweet could be considered as trash even if they were readable. The average time required to each assessor was less than 20 hours. Table 6 presents the results for readability for Tweet Contextualization 2012 most informative systems introduced in Table 3.

Finally, in 2013 participants obtained the results presented in Table 7 for Tweet Contextualization 2012 most informative systems introduced in Table 4.

When comparing the results in the various years of the task, it appears that the two metrics relaxed and strict are correlated (Kendall test: $\tau > 90\%$, $p < 10^{-3}$) . In fact, few systems set up complex text analysis like deep syntax analysis or anaphora solving. Among these few, there is the IRIT [32] system that reached best readability scores in

Table 7. Readability results in 2013 with the relaxed and strict metric.

Relaxed metric			**Strict metric**		
Run	team	Score	Run	team	Score
275	IRIT	76.64%	275	IRIT	67.3%
254	Duluth	73.3%	258	Nantes	61.52%
258	Nantes	68.36%	254	Duluth	64.52%
270	Jadavpur	46.84%	270	Jadavpur	41.2%
265	Ceará	39.46%	265	Ceará	36.46%
261	Montréal	36.42%	261	Montréal	35.14%
262	IULA	33.34%	262	IULA	30.38%
266	Leibniz	25.92%	266	Leibniz	25.08%
277	Kolkata	20%	277	Kolkata	20%
269	Chaoyang	4%	269	Chaoyang	4%

2013 as shown in Table 7 and also in 2014 as reported in [9]. This system also obtained the best informativeness scores in 2011 and 2013. The low readability scores in 2012 are correlated to lower informativeness [31]. The IRIT system tried to use automatic readability evaluation and anaphora detection to improve readability scores. One interesting finding is that these methods also help to improve informativeness density in summaries. However, they could not be fully evaluated in the Tweet Contextualization track lab because it was required not to rewrite Wikipedia passages.

5.4. Combined Evaluation

Meanwhile both evaluations were carried out independently, it appears that informativeness and readability scores are strongly correlated over runs (Kendall test: $\tau > 66\%$ in 2011 and in 2013 with $p < 10^{-3}$and $\tau > 31\%$ with $p < 10^{-2}$). From 2012, readability has been evaluated in the context of the tweet. Passages not related to the tweet were considered as unreadable. Moreover, a passage to be included in the reference had to be selected by at least one assessor, and this decision had to be only based on the passage contents. For these reasons, systems that submitted summaries made of long readable passages improved their chances to have a passage picked up, to be included in the final test reference to evaluate informativeness meanwhile systems that tried to combine short passages without checking readability were disadvantaged. The overall best system by IRIT used automatic readability evaluation measures to avoid this drawback.

Fig. 4 shows results obtained by participants in the 2011, 2012 and 2013 tracks. Reference system runs are circled.

6. Conclusion

About 340 millions of tweets are written every day. However, 140 characters long messages are rarely self-content. The Tweet Contextualization track 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 between 2010 and 2014, results show that only systems that 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. Evaluation considers both informativeness and readability. Informativeness is measured as a variant of absolute log-diff between term frequencies in the text reference and the proposed summary. Maximal informativeness scores obtained by participants from 19 different groups are between 10% and 14%. There is also room for improving readability. The goal of the task and the evaluation metrics have been kept unchanged between 2011 and 2013 allowing multiple comparative analysis. After 2011, the tweets we selected as topics reflected more the diversity of tweets. The various collections we used are available at *http://tc.talne.eu* and *http://qa.termwatch.es/data* as well as the runs from the participants and the papers they published at INEX. The evaluation tool is also available at the same address.

After having analyzed the four editions of the tweet contextualization task, we can summarize the lessons we have learned as follows:

- Considering the collections we built: while the simulated tweets we used the first year were very convenient for evaluation purposes (we built an automatic reference that makes sense) and thus for the organizers, the content and format of the tweets were too different from real tweets (no URL, no hashtag) which has confused somehow the participants. The evaluation framework should not be too different from real use cases so that participants can develop realistic models as well. This aspect has been solved from the second year of the evaluation campaign.

- We went deeper in the tweet complexity each year. On one hand, it was appropriate because the participants who participated several years could focus on different aspects of their model rather than trying to solve all the aspects of their system at the same time. On the other hand, those who participated only one year may have had the feeling that the task was incomplete. It is thus probably more appropriate to define a complete task and keep it for several years while adding sub-tasks as trial for new challenges.

- It is not always appropriate to define a new evaluation measure for a new task. However, after having deeply analyzed and tried existing measures, it appeared that defining a measure with new properties was necessary.

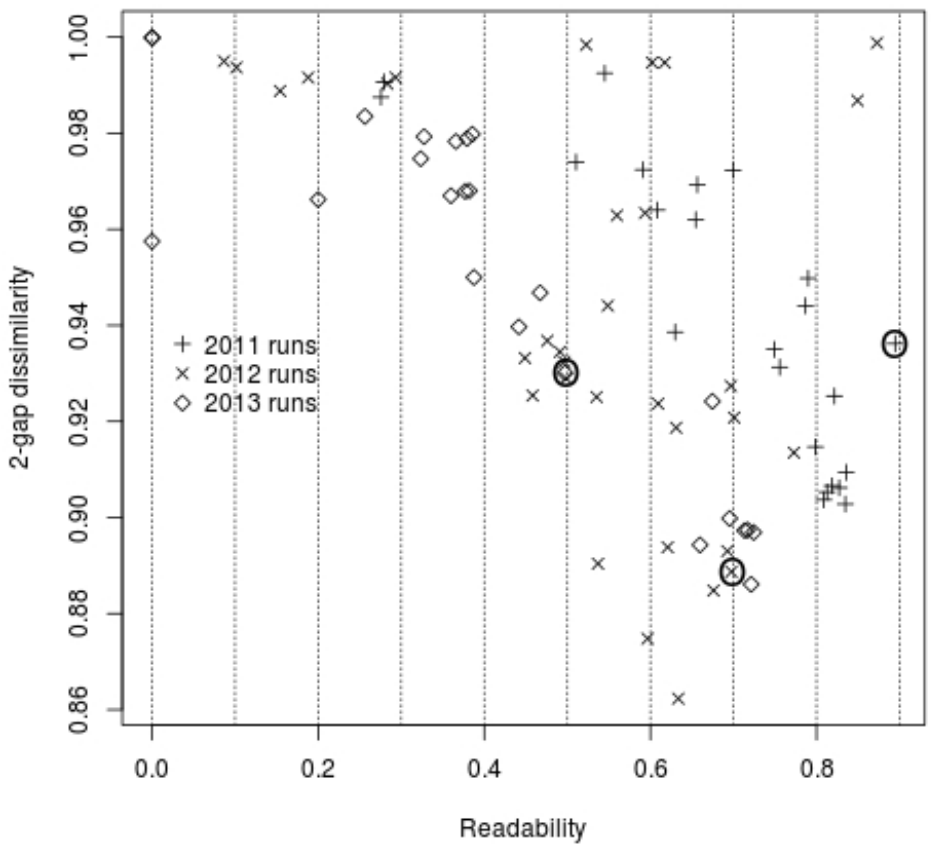

Figure 4. Informativeness and relaxed Readability scores for all 2011 - 2013 official runs (Best runs are on the bottom right: low 2-gap dissimilarity and high readability).

We wanted an effectiveness measure that is as much as possible independent from the size of the reference and that would work for new runs without penalizing too much runs that have not been used to build the reference. LogSim fits well with these constraints and it remained unchanged during the four years. However, as far as summary readability is concerned, the measures we proposed may not be enough. Psychologists and linguists' point of view could be worth using for this type of evaluation. This point remains a future work.

- Tweets are so linguistically complex and so contextual that we believe it could be interesting to consider a given, specific domain only. We do think that it would be worth considering a target domain and study whether the context can be extracted from other tweets or if external resources are necessary. This is a track to be investigated in the future and that could have many applications in various domains such as marketing for example.

Acknowledgments. We would like to acknowledge the French ANR agency for their support through the CAAS-Contextual Analysis and Adaptive Search project (ANR- 10-CORD-001 01).

References

[1] Alguliev, R. M., Aliguliyev, R. M., Hajirahimova, M. S., Mehdiyev, C. A., 2011. MCMR: Maximum coverage and minimum redundant text summarization model. Expert Syst. Appl. 38 (12), 14514–14522.

[2] Allan, J., 2005. HARD Track Overview in TREC 2005 High Accuracy Retrieval from Documents. In: The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings, NIST Special Publication: SP 500-266.

[3] Amigò, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., de Rijke, M., Spina, D., 2013. Overview of replab 2013: Evaluating online reputation monitoring systems. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Vol. 8138 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 333–352.
URL http://dx.doi.org/10.1007/978-3-642-40802-1_31

[4] Balasubramanian, N., Allan, J., Croft, W. B., 2007. A comparison of sentence retrieval techniques. In: [46], pp. 813–814.
URL http://doi.acm.org/10.1145/1277741.1277922

[5] Balog, K., Serdyukov, P., Vries, A. P. d., 2011. Overview of the trec 2011 entity track. Tech. rep., The Twentieth Text REtrieval Conference (TREC 2011).

[6] Balog, K., Serdyukov, P., Vries, A. P. d., Thomas, P., Westervelf, T., 2009. Overview of the trec 2010 entity track. Tech. rep., The Eighteenth Text REtrieval Conference (TREC 2009).

[7] Bandyopadhyay, A., Pal, S., Mitra, M., Majumder, P., Ghosh, K., 2012. Passage retrieval for tweet contextualization at inex 2012. In: [34].

[8] Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., Tannier, X., 2013. Overview of INEX tweet contextualization 2013 track. In: Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.
URL http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-BellotEt2013.pdf

[9] Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., Tannier, X., 2014. Overview of INEX tweet contextualization 2014 track. In: [14], pp. 494–500.
URL http://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-BellotEt2014.pdf

[10] Bennett, P. N., Radlinski, F., White, R. W., Yilmaz, E., 2011. Inferring and using location metadata to personalize web search. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. SIGIR '11. ACM, New York, NY, USA, pp. 135–144.

[11] Bhaskar, P., Banerjee, S., Bandyopadhyay, S., 2013. Tweet Contextualization (Answering Tweet Question) – the Role of Multi -document Summarization. In: CLEF (Working Notes).

[12] Bhaskar, P., Banerjee, S., Neogi, S., Bandyopadhyay, S., 2011. A hybrid QA system with focused IR and automatic summarization for INEX 2011. In: Focused Retrieval of Content and Structure, 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011, Saarbrücken, Germany, December 12-14, 2011, Revised Selected Papers. pp. 207–218.
URL http://dx.doi.org/10.1007/978-3-642-35734-3_18

[13] Candillier, L., Chevalier, M., Dudognon, D., Mothe, J., 2011. Diversity in recommender systems: Bridging the gap between users and systems. In: Proceedings of the International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services. CENTRIC11. pp. 48–58.

[14] Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (Eds.), 2014. Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. Vol. 1180 of CEUR Workshop Proceedings. CEUR-WS.org.
URL http://ceur-ws.org/Vol-1180

[15] Carpineto, C., Romano, G., 2012. A survey of automatic query expansion in information retrieval. ACM Comput. Surv 44 (1), 1.
URL http://doi.acm.org/10.1145/2071389.2071390

[16] Chall, J. S., Dale, E., 1995. Readability revisited: The new Dale-Chall readability formula. Brookline Books.

[17] Chatzopoulou, G., Eirinaki, M., Polyzotis, N., 2009. Query recommendations for interactive database exploration. In: Proceedings of the 21st International Conference on Scientific and Statistical Database Management. SSDBM 2009. Springer-Verlag, Berlin, Heidelberg, pp. 3–18.

[18] Chen, C., Ibekwe-Sanjuan, F., Hou, J., 2010. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. JASIST 61 (7), 1386–1409.

[19] Clarke, C. L., Craswell, N., Soboroff, I., 2009. Overview of the trec 2009 web track. In: TREC.

[20] Crestani, F., Ruthven, I., 2007. Introduction to special issue on contextual information retrieval systems. Information Retrieval 10 (2), 111–113.

[21] Crouch, C. J., Crouch, D. B., Chittilla, S., Nagalla, S., Kulkarni, S., Nawale, S., 2012. The 2012 inex snippet and tweet contextualization tasks. In: [34].

[22] Dang, H., 2008. Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In: Proc. of the First Text Analysis Conference.

[23] Dang, H., 2008. Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In: Proceedings of the First Text Analysis Conference.

[24] Dean-Hall, A., Clarke, C. L., Kamps, J., Thomas, P., Voorhees, E., 2012. Overview of the trec 2012 contextual suggestion track. Tech. rep., The Twenty-First Text REtrieval Conference (TREC 2012).

[25] Deveaud, R., Boudin, F., 2012. Lia/lina at the inex 2012 tweet contextualization track. In: [34].

[26] Deveaud, R., Boudin, F., 2013. Contextualisation automatique de tweets à partir de wikipédia. In: Berrut, C. (Ed.), CORIA 2013 - Conférence en Recherche d'Infomations et Applications - 10th French Information Retrieval Conference, Neuchâtel, Suisse, April 3-5, 2013. UNINE, pp. 125–140.
URL http://asso-aria.org/coria/2013/coria/coria2013_23.pdf

[27] Deveaud, R., Boudin, F., 2013. Effective tweet contextualization with hashtags performance prediction and multi-document summarization. In: Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.
URL http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-DeveaudEt2013.pdf

[28] Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., Robbins, D. C., 2003. Stuff i've seen: a system for personal information retrieval and re-use. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. SIGIR '03. ACM, New York, NY, USA, pp. 72–79.
[29] Ekstrand-Abueg, M., Pavlu, V., Aslam, J. A., 2013. Live nuggets extractor: a semi-automated system for text extraction and test collection creation. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 1087–1088.
[30] Ermakova, L., 2015. A Method for Short Message Contextualization: Experiments at CLEF/INEX. In: Experimental IR meets Multilinguality, Multimodality, and Interaction - Sixth International Conference of the CLEF Association, Proceedings. LNCS Vol.9283, Springer.
[31] Ermakova, L., Mothe, J., 2012. Irit at inex 2012: Tweet contextualization. In: CLEF (Online Working Notes/Labs/Workshop).
[32] Ermakova, L., Mothe, J., 2012. Irit at inex: question answering task. In: Focused Retrieval of Content and Structure. Springer, pp. 219–226.
[33] Feng, L. ans Jansche, M., Huenerfauth, M., Elhadad, N., 2010. A Comparison of Features for Automatic Readability Assessment. In: Proceedings of COLING 2010, Poster Volume.
[34] Forner, P., Karlgren, J., Womser-Hacker, C. (Eds.), 2012. CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012.
[35] Ganguly, D., Leveling, J., Jones, G. J. F., 2012. Dcu@inex-2012: Exploring sentence retrieval for tweet contextualization. In: [34].
[36] Gee, J. P., 2014. An introduction to discourse analysis: Theory and method. Routledge.
[37] Geva, S., Kamps, J., Schenkel, R., Trotman, A. (Eds.), 2011. Comparative Evaluation of Focused Retrieval - 9th International Workshop of the Inititative for the Evaluation of XML Retrieval, INEX 2010, Vugh, The Netherlands, December 13-15, 2010, Revised Selected Papers. Vol. 6932 of Lecture Notes in Computer Science. Springer.
[38] Geva, S., Kamps, J., Trotman, A. (Eds.), 2010. Focused Retrieval and Evaluation, 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009, Brisbane, Australia, December 7-9, 2009, Revised and Selected Papers. Vol. 6203 of Lecture Notes in Computer Science. Springer.
[39] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N., 2011. Part-Of-Speech tagging for Twitter: annotation, features and experiments. In: Proceedings of HLT, volume 2.
[40] Halliday, M., Hasan, R., 1976. Cohesion in English. London: Longman.
[41] Hennig, L., D., L. E. W., S., A., 2010. Learning Summary Content Units with Topic Modeling. In: Proceeding of COLING '10 (Posters). pp. 391–399.
[42] Hernandez, N., Mothe, J., Chrisment, C., Egret, D., 2007. Modeling context through domain ontologies. Information Retrieval 10 (2), 143–172.
[43] Imafouo, A., Tannier, X., 2005. Retrieval status values in information retrieval evaluation. In: Consens, M. P., Navarro, G. (Eds.), String Processing and Information Retrieval, 12th International Conference, SPIRE 2005, Buenos Aires, Argentina, November 2-4, 2005, Proceedings. Vol. 3772 of Lecture Notes in Computer Science. Springer, pp. 224–227.
URL http://dx.doi.org/10.1007/11575832_25
[44] Ingwersen, P., Järvelin, K., 2006. The turn: Integration of information seeking and retrieval in context. Vol. 18. Springer.
[45] Koren, Y., Apr. 2010. Collaborative filtering with temporal dynamics. Commun. ACM 53 (4), 89–97.
[46] Kraaij, W., de Vries, A. P., Clarke, C. L. A., Fuhr, N., Kando, N. (Eds.), 2007. SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007. ACM.
[47] Laureano-Cruces, A. L., Ramírez-Rodríguez, J., 2011. A graph-based summarization system at qa@inex track 2011. In: Focused Retrieval of Content and Structure, 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011, Saarbrücken, Germany, December 12-14, 2011, Revised Selected Papers. pp. 227–231.
URL http://dx.doi.org/10.1007/978-3-642-35734-3_20
[48] Lin, C. Y., July 25-26 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS).
[49] Lin, J. J., Zhang, P., 2007. Deconstructing Nuggets: the Stability and Reliability of Complex Question Answering Evaluation. In: [46], pp. 327–334.
[50] Linhares, A. C., 2013. An automatic greedy summarization system at INEX 2013 tweet contextualization track. In: Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.
URL http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-CarneiroLinhares2013.pdf
[51] Louis, A., Nenkova, A., 2009. Performance confidence estimation for automatic summarization. In: EACL. The Association for Computer Linguistics, pp. 541–548.
[52] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky, D., 2014. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60.
URL http://www.aclweb.org/anthology/P/P14/P14-5010
[53] Méndez-Cruz, C., Torres-Moreno, J., Urrea, A. M., Sierra, G., 2012. Extrinsic evaluation on automatic summarization tasks: Testing affixality measurements for statistical word stemming. In: Batyrshin, I. Z., González-Mendoza, M. (Eds.), Advances in Computational Intelligence - 11th Mexican International Conference on Artificial Intelligence, MICAI 2012, San Luis Potosí, Mexico, October 27 - November 4, 2012. Revised Selected Papers, Part II. Vol. 7630 of Lecture Notes in Computer Science. Springer, pp. 46–57.
URL http://dx.doi.org/10.1007/978-3-642-37798-3_5
[54] Méndez-Cruz, C.-F., Soriano-Morales, E.-P., Urrea, A. M., 2012. Testing a statistical word stemmer based on affixality measurements in inex 2012 tweet contextualization track. In: [34].
[55] Metzler, D., Croft, W. B., 2004. Combining the language model and inference network approaches to retrieval. Inf. Process. Manage. 40 (5), 735–750.
[56] Morchid, M., Linarès, G., 2012. Inex 2012 benchmark a semantic space for tweets contextualization. In: [34].
[57] Moriceau, V., SanJuan, E., Tannier, X., Bellot, P., 2009. Overview of the 2009 QA Track: Towards a Common Task for QA, Focused IR and Automatic Summarization Systems. In: Proceedings of INEX.

[58] Nenkova, A., Chae, J., Louis, A., Pitler, E., 2010. Structural features for predicting the linguistic quality of text - applications to machine translation, automatic summarization and human-authored text. In: Krahmer, E., Theune, M. (Eds.), Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation. Vol. 5790 of Lecture Notes in Computer Science. Springer, pp. 222–241.
URL http://dx.doi.org/10.1007/978-3-642-15573-4_12
[59] Nenkova, A., Passonneau, R., 2004. Evaluating content selection in summarization: The pyramid method. In: Proceedings of HLT-NAACL. Vol. 2004.
[60] Nenkova, A., Passonneau, R., 2004. Evaluating content selection in summarization: The pyramid method. In: Proceedings of HLT-NAACL. Vol. 2004.
[61] Ogilvie, P., Voorhees, E., Callan, J., Dec. 2009. On the number of terms used in automatic query expansion. Inf. Retr. 12 (6), 666–679.
[62] Pakray, P., Bhaskar, P., Pal, S., Das, D., Bandyopadhyay, S., Gelbukh, A. F., 2010. JU_CSE_TE: System Description QA@CLEF 2010 - ResPubliQA. In: Cross-Language Evaluation Forum.
[63] Pavlu, V., Rajput, S., Golbus, P. B., Aslam, J. A., 2012. IR system evaluation using nugget-based test collections. In: Proceedings of the fifth ACM international conference on Web Search and Data Msagining (WSDM). pp. 393–402.
[64] Peñas, A., Forner, P., Rodrigo, Á., Sutcliffe, R. F. E., Forascu, C., Mota, C., 2010. Overview of respubliqa 2010: Question answering evaluation over european legislation. In: Braschler, M., Harman, D., Pianta, E. (Eds.), CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy. Vol. 1176 of CEUR Workshop Proceedings. CEUR-WS.org.
URL http://ceur-ws.org/Vol-1176/CLEF2010wn-MLQA10-PenasEt2010.pdf
[65] Pitler, E., Nenkova, A., 2008. Revisiting readability: A unified framework for prediction text quality. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.
[66] Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., Zhang, Z., 2004. MEAD: a platform for multidocument multilingual text summarization. In: Proceedings of LREC.
[67] Saggion, H., 2011. SUMMA Content Extraction for INEX2011. In: Proceedings of INEX2011.
[68] Saggion, H., Torres-Moreno, J.-M., da Cunha, I., SanJuan, E., Velázquez-Morales, P., 2010. Multilingual Summarization Evaluation without Human Models. In: COLING 2010, 23rd International Conference on Computational (poster session). pp. 1059–1067.
[69] SanJuan, E., Bellot, P., Moriceau, V., Tannier, X., 2010. Overview of the inex 2010 question answering track (qa@inex). In: [37], pp. 269–281.
[70] SanJuan, E., Ibekwe-Sanjuan, F., 2010. Multi word term queries for focused information retrieval. In: Gelbukh, A. F. (Ed.), Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Proceedings. Vol. 6008 of Lecture Notes in Computer Science. Springer, pp. 590–601.
URL http://dx.doi.org/10.1007/978-3-642-12116-6_50
[71] SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J., 2012. In: Geva, S., Kamps, J., Schenkel, R. (Eds.), Focused Retrieval of Content and Structure. Vol. 7424 of Lecture Notes in Computer Science. pp. 188–206.
[72] SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J., 2012. Overview of the inex 2012 tweet contextualization track. In: [34], p. 148.
[73] Schmid, H., Sep. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: International Conference on New Methods in Language Processing.
URL http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[74] Shen, X., Tan, B., Zhai, C., 2005. Context-sensitive information retrieval using implicit feedback. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '05. ACM, New York, NY, USA, pp. 43–50.
[75] Tamine-Lechani, L., Boughanem, M., Daoud, M., 2010. Evaluation of contextual information retrieval effectiveness: overview of issues and research. Knowledge and Information Systems 24 (1), 1–34.
[76] Tavernier, J., Bellot, P., 2011. Flesch and dale-chall readability measures for INEX 2011 question-answering track. In: Focused Retrieval of Content and Structure, 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011, Saarbrücken, Germany, December 12-14, 2011, Revised Selected Papers. pp. 235–246.
URL http://dx.doi.org/10.1007/978-3-642-35734-3_22
[77] Torres-Moreno, J., 2014. Three statistical summarizers at CLEF-INEX 2013 tweet contextualization track. In: [14], pp. 565–573.
URL http://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-TorresMoreno2014.pdf
[78] Torres-Moreno, J., Gagnon, M., 2010. The cortex automatic summarization system at the qa@inex track 2010. In: [37], pp. 290–294.
URL http://dx.doi.org/10.1007/978-3-642-23577-1_26
[79] Torres-Moreno, J., Ramirez, J., 2010. REG : un algorithme glouton appliqué au résumé automatique de texte. In: Proceedings of JADT.
[80] Trappett, M., Geva, S., Trotman, A., Scholer, F., Sanderson, M., 2012. Overview of the inex 2012 snippet retrieval track. In: [34].
[81] Van Dijk, T. A., 2009. Society and discourse: How social contexts influence text and talk. Cambridge University Press.
[82] Vivaldi, J., da Cunha, I., 2012. Inex tweet contextualization track at clef 2012: Query reformulation using terminological patterns and automatic summarization. In: [34].
[83] Wang, J., Robertson, S., Vries, A. P., Reinders, M. J., Dec. 2008. Probabilistic relevance ranking for collaborative filtering. Inf. Retr. 11 (6), 477–497.
[84] Wang, J., Zhao, Z., Zhou, J., Wang, H., Cui, B., Qi, G., Jun. 2012. Recommending flickr groups with social topic model. Inf. Retr. 15 (3-4), 278–295.
[85] Wodak, R., Meyer, M., 2001. Methods of critical discourse analysis. Sage Publications.