Tweet Contextualization Based on Wikipedia and Dbpedia

Meriem Amina Zingla[*] [**] — Chiraz Latiri [***] — Yahya Slimani[*] — Catherine Berrut[****] — Philippe Mulhem[****]

[*] *University of Carthage, INSAT, LISI research Laboratory, Tunis, Tunisia*
[**] *University of Tunis El Manar, Faculty of Sciences of Tunis*
[***] *University of Tunis El Manar, Faculty of Sciences of Tunis, LIPAH research*
[****] *Grenoble Alpes University, LIG laboratory, MRIM group, Grenoble, France*

ABSTRACT. *Bound to 140 characters, tweets are short and not written maintaining formal grammar and proper spelling. These spelling variations increase the likelihood of vocabulary mismatch and make them difficult to understand without context. This paper falls under the tweet contextualization task that aims at providing, automatically, a summary that explains a given tweet, allowing a reader to understand it. We propose different tweet expansion approaches based on Wikipeda and Dbpedia as external knowledge sources. These proposed approaches are divided into two steps. The first step consists in generating the candidate terms for a given tweet, while the second one consists in ranking and selecting these candidate terms using a similarity measure. The effectiveness of our methods is proved through an experimental study conducted on the INEX 2014 collection.*

RÉSUMÉ. *La taille des tweets est limitée à un nombre maximum de caractères. Cette contrainte liée à la taille du message entraîne l'utilisation d'un vocabulaire particulier rendant le tweet difficile à comprendre. La tâche de contextualisation des tweets vise à fournir, automatiquement, un résumé qui explique un tweet donné, ce qui permet au lecteur de bien le comprendre. Nous proposons pour cela différentes méthodes basées sur deux énormes sources de connaissances à savoir, Wikipédia et Dbpedia. L'efficacité de notre méthode est prouvée par une étude expérimentale menée sur la collection d'INEX 2014.*

KEYWORDS: *Tweet contextualization, Association rules, INEX, Explicit Semantic Analysis, Query Expansion.*

MOTS-CLÉS : *Contextualisation des tweets, Expansion de requêtes, INEX Analyse Sémantique Explicite, Règles d'association*

1. Introduction

Microblogging has emerged as one of the primary social media platforms for users to submit, in real-time, short messages, to report an idea or an opinion. Twitter is an online social networking service that enables users to tweet about any topic within the 140-character limit called *tweets*. However, this limit causes users to employ different strategies such as abbreviations and slangs in order to compress more information in minimum of characters. Tweets are, therefore, often misspelled or truncated and especially hard to understand.

To study this problem, INEX (Initiative for the Evaluation of XML retrieval) launched the tweet contextualization track for the first time in 2011. Thus, the tweet contextualization track proposed to answer questions of the form "What is this tweet about?" using a cleaned dump of Wikipedia, in order to allow the reader a better understanding of the tweet. The principle of this task is first to find documents that are related to the tweet (using a information retrieval system IRS), and then to generate an accurate summary of such documents (using an automatic summarization system ASS). INEX organizers provide the task participants with a baseline system that combines IRS and ASS. This baseline system takes as input a text query (tweet). This task allows participants to focus on the best tweet formulation for the IRS, since it has a direct impact on the summary quality.

Our goal here is to study tweet expansions as presented in Figure 1, and to evaluate the results using the INEX baseline system, on the INEX 2014 test collection. Several works have already proposed the use of the tweet expansion methods for the target task such as in (Morchid *et al.*, 2013), where the authors proposed to use Latent Dirichlet Analysis to expand the original tweets, and in (Zingla *et al.*, 2014), where the authors proposed to use association rules inter-terms, but these approaches do not include a term ranking step. The absence of this ranking step resulted in noisy queries containing unrelated terms to the original tweet. In this paper, we propose to expand the original tweets, using two external knowledge sources, namely, Wikipedia and Dbpedia. We opted to use Wikipedia because it is currently the largest knowledge repository on the Web. The use of this huge source is fruitful by allowing a massive knowledge representation of a tweet. We also use Dbpedia because it provides vast amounts of structured knowledge extracted from Wikipedia info boxes, hence, allowing to augment tweet representation with massive amounts of related information. Taking into account the weaknesses of the existing works cited before, we propose to enhance the tweet expansion process so that it will be composed of two steps : Tweet candidate terms generation, and candidate terms ranking and selection. While the first step extracts related information, i.e. terms, from Wikipedia and Dbpedia, the second step computes the semantic relatedness score between the original tweet and the candidate terms, using a new measure (ESAC) that relies on Explicit Semantic Analysis (ESA) and association rules.

The remaining of this paper is organized as follows: In section 2, our work is put in the context of related works, while section 3 gives a detailed description of our tweet expansion methods. Section 4 presents our experimentations and results. Finally, section 5 is dedicated to the conclusion of this work and gives future works.

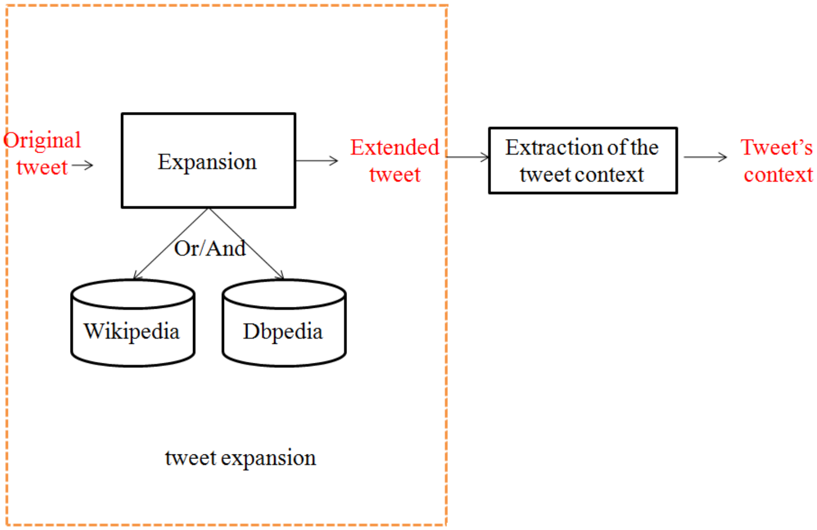

Figure 1. *Explanatory schema of the proposed work*

2. Related Work

In this section we review some related works, referring to query expansion and tweet contextualization.

2.1. *Query expansion*

Several works in the literature are proposed for the query expansion task, such as in (Song *et al.*, 2007) where the authors proposed a novel semantic query expansion technique that combines association rules with ontologies and Natural Language Processing techniques. This technique uses the explicit semantics as well as other linguistic properties of unstructured text corpus, it incorporates contextual properties of important terms discovered by association rules, and ontology entries are added to the query by disambiguating word senses. In (Latiri *et al.*, 2003), authors addressed query expansion by considering the term-document relation as fuzzy binary relations. Their approach to extract fuzzy association rules is based on the closure of an extended fuzzy Galois connection, using different semantics of term membership degrees. In (Shekarpour *et al.*, 2013), authors proposed an approach based on performing an initial retrieval of resources according to the original keyword query, the proposed process is divided into three main steps. In the first step, all words closely related to the original keyword are extracted based on two types of features linguistic and semantic. In the second step, various introduced linguistic and semantic features are weighted using learning approaches. In the third step, they assign a relevance score to the set of the related words. Using this score they prune the related word set to achieve a balance between precision and recall. In these two previous works, relations weighting aspect are important. Authors in (Tan *et al.*, 2013) proposed a semantic approach that expands short queries by semantically related terms extracted from Wikipedia, they incorporate the expansion terms into the original query and adapt language models to evaluate the expanded queries.

The proposal of the paper is therefore, in a certain way, a continuation of these works, but in the case of tweet contextualization.

2.2. *Query expansion for microblog retrieval*

Query expansion techniques are also used for microblog retrieval, authors in (Bandyopadhyay *et al.*, 2012), for example, used external corpora as a source for query expansion terms. Specifically, they used the Google Search API (GSA) to retrieve pages from the Web, and expanded the queries employing their titles. In (Lau *et al.*, 2011), authors proposed a twitter retrieval framework that focuses on topical features, combined with query expansion using Pseudo-Relevance Feedback (PRF) to improve microblogs retrieval results.

2.3. *Tweet contextualization*

Despite the fact that the idea to contextualize tweets is quite recent, there are several works in this context. Recently, authors of (Ermakova and Mothe, 2012) proposed a method based on the local Wikipedia dump, they used the Term Frequency-Inverse Document Frequency TF-IDF cosine similarity measure enriched by smoothing from local context, named entity recognition and Part-Of-Speech weighting presented at INEX 2011. They modified this method by adding bigram similarity, anaphora resolution, hashtag processing and sentence reordering. The sentence ordering task was modeled as a sequential ordering problem, where vertices corresponded to sentences and sentence time stamps represented sequential constraints. They proposed a greedy algorithm to solve the sequential ordering problem based on chronological constraints. While in (Deveaud and Boudin, 2013a), authors used a method that allows to automatically contextualize tweets by using information coming from Wikipedia. They treated the problem of tweets contextualization as an automatic summarization task, where the text to resume is composed of Wikipedia articles that discuss the various pieces of information appearing in a tweet. One of the limitations of this approach is that the number of Wikipedia articles used to extract the candidate sentences is set manually. They explore the influence of various tweet-related articles retrieval methods as well as several features for sentence extraction. Whereas, in (Deveaud and Boudin, 2013b),

authors added a hashtag performance prediction component to the Wikipedia retrieval step. They used all available tweet features including web links which were not allowed by INEX's organizers.
In (Linhares, 2013), authors used an automatic summarizer named REG based on a greedy optimization algorithm to weigh the sentences. The summary is obtained by concatenating the relevant sentences weighed in the optimization step. In (Morchid *et al.*, 2013), authors used Latent Dirichlet Analysis (LDA) to obtain a representation of the tweet in a thematic space. This representation allows the finding of a set of latent topics covered by the tweet, this approach gives good results for the tweet contextualization task. Authors in (Zingla *et al.*, 2014) use the association rules between terms to extend the tweet, they project the terms of tweet on the rules' premises and add the conclusions to the original tweets.

Closely similar to our task, the authors in (Meij *et al.*, 2012), aim at adding semantics to microblog posts. They proposed a method that uses machine learning, and is based on a high-recall concept ranking and a high-precision concept selection step.

Finally, in (Torres-Moreno, 2014) authors developed three statistical summarizer systems the first one called Cortex summarizer, that uses several sentence selection metrics and an optimal decision module to score sentences from a document source, the second one called Artex summarizer, that uses a simple inner product among the topic-vector and the pseudo-word vector and the third one called REG summarizer which is a performant graph-based summarizer.

3. Tweet Expansion

The tweet expansion aims at augmenting the thematic space of a given tweet by a massive amount of related terms. This is done to improve the IRS performance by giving more chance for a relevant document, which does not contain the original tweet terms, to be retrieved. In our work, we expand the original tweets in two steps (*cf.* Figure 2), namely:

– Step 1: Candidate terms generation.

– Step 2: Candidate terms ranking and selection.

The first step generates the candidate terms, for a given tweet, from Wikipedia and Dbpedia, the second one ranks these candidate terms according to their similarity to the tweet, and selects the best ones to be added. Three alternatives are proposed for the candidate terms generation step. Two are based on Wikipedia and one is based on Dbpedia.

3.1. *Step 1: Candidate Terms Generation*

Wikipedia is a huge knowledge source growing every day due to the contribution of people all around the world, it represents a very large, high quality, and valuable

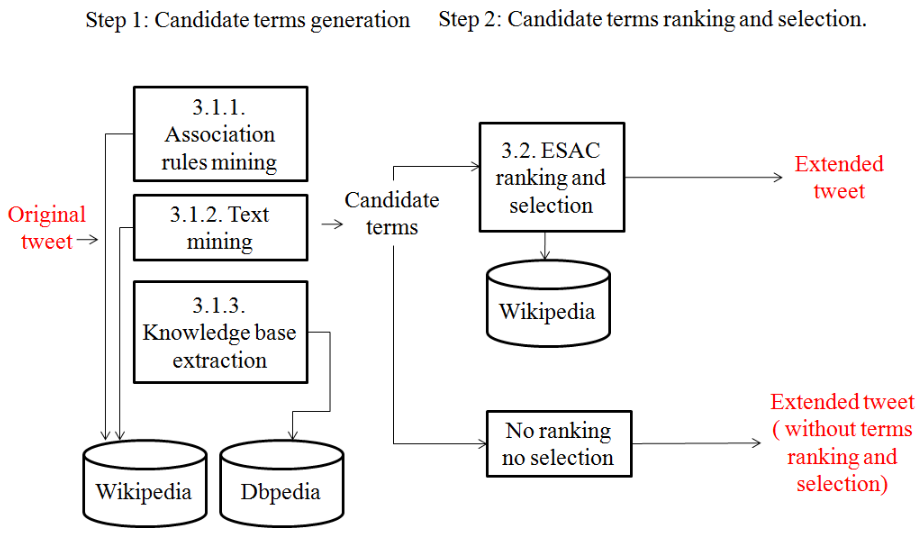
Step 1: Candidate terms generation
Step 2: Candidate terms ranking and selection.
3.1.1. Association rules mining
3.1.2. Text mining
3.1.3. Knowledge base extraction
Original tweet
Candidate terms
3.2. ESAC ranking and selection
Extended tweet
Wikipedia
Wikipedia
Dbpedia
No ranking no selection
Extended tweet (without terms ranking and selection)

Figure 2. *The Proposed Methods for Tweet Expansion*

source in natural language, this is why we opted it to extend the tweets. We proposed to use two different methods to extract information related to a given tweet from this source. The first one is a statistical method based on association rules mining, the second one based on text mining. Authors, in (Tan *et al.*, 2013), have proved, in another frame, that the use of WIkipedia is interesting for the query expansion.

Dbpedia is a project aiming to represent Wikipedia content in RDF triples. It plays a central role in the Semantic Web, due to the large and growing number of resources linked to it.

We claim that the use of these knowledge sources will augment the tweet representation by a massive amount of related information. As DBpedia is structured and filtered, we expect less noisy expansions than with Wikipedia. However, using DBpedia may generate smaller expansions than Wikipedia. Using both sources is expected to select precise terms (from DBpedia) as well as broad terms (from Wikipedia). We propose to generate candidate terms for a given tweet using different methods based on these knowledge sources, Wikipedia and Dbpedia, namely: Association rules mining, text mining, and knowledge base extraction.

3.1.1. *Association Rules Mining*

An association rule binds two sets of terms, which respectively constitute its premise (T_1) and conclusion (T_2) parts. Thus, a rule approximates the probability of having T_2 in a document, given that those of the T_1 are already there. Compared to simple co-occurrences measures, association rules are then oriented, and we believe that this point is a must in our case, as we try to find new query terms from the initial

tweet.
The rule R is said to be based on the termset T equal to $T_1 \cup T_2$. The support of T is equal to the number of documents in C containing all the term of T. The support is formally defined as follows:

$$Supp(T) = |\{d|d \in C \wedge \forall t \in T : (d,t) \in I\}| \qquad [1]$$

Where:

– C is the whole set of documents which form the collection

– d is a single document of the collection $(d \in C)$

– τ the whole set of distinct terms of the collection C

– T a set of terms of the collection $(T \subseteq \tau)$

– t a single term of the collection $(t \in \tau)$

– $I \subseteq C \times T$ is a binary (incidence) relation. Each couple $(d,t) \in I$ indicates that the document $d \in C$ has the term $t \in \tau$.

The *confidence* of a rule R: $T_1 \Rightarrow T_2$ is computed as:

$$\mathit{Conf}(R) = \frac{\mathit{Supp}(T)}{\mathit{Supp}(T_1)}. \qquad [2]$$

An illustrative example of association rules is highlighted in Table 1.

Table 1. *Association Rules examples taken from Wikipedia articles*

Premise	Conclusion	Support	Confidence
manufacture	car	356	0.8921
manufacture	motor	301	0.9102
campus	university	279	0.7431

An association rule R is said to be *valid* if its confidence value, *i.e.*, *Conf*(R), is greater than or equal to a user-defined threshold denoted *minconf*. This confidence threshold is used to exclude non valid rules.
The process of generating candidate terms for a tweet is performed in the following steps:

– Selecting a sub-set of articles, from Wikipedia, similar to the tweet, using an algorithm based on the TF-IDF measure (Xia and Chai, 2011).

– Annotating the selected articles using TreeTagger. The choice of TreeTagger was based on the ability of this tool to recognize the nature (morpho-syntactic category) of a word in its context. TreeTagger uses the recursive construction of decision trees with a probability calculation to estimate the Part-Of-Speech of a word.

– Extracting specific terms, we extract terms that are nouns.

– Generating the association rules using an efficient algorithm called CHARM. We adapted the algorithm CHARM (Zaki and Hsiao, 2002), because it allows to generate non-redundant association rules (Yahia and Nguifo, 2004). As parameters, CHARM takes $minsupp$ as the relative minimal support and $minconf$ as the minimum confidence of the rules. While considering the $Zipf$ distribution of the selected sub-set of Wikipedia articles, the minimal threshold of the support value is experimentally set in order to spread trivial terms which occur in the most of the documents, and are then related to too many terms.

– Obtaining the candidate terms for a given tweet, the candidate terms are the terms that appear in the conclusions of the irredundant association rules whose premise is included in the original tweet.

3.1.2. *Text Mining*

The second method to generate candidate terms for a given tweet from Wikipedia consists in exploring the Wikipedia articles related to the tweet, especially, the articles' first sentences that we call: definitions.

To achieve this, we use some heuristics, namely:

– Given a tweet, first, we search, in Wikipedia, all articles that correspond to the tweet's words. (Since the tweets are shorts, we consider all the words of tweet) This is done using WikipediaMiner [1], which is a toolkit developed for tapping the rich semantics encoded within Wikipedia. This toolkit helps to integrate Wikipedia's knowledge into applications , by:

- Providing simplified, object-oriented access to Wikipedia's structure and content.
- Measuring how terms and concepts in Wikipedia are connected to each other.
- Detecting and disambiguating Wikipedia topics when they are mentioned in documents.

– We extract, from these articles, their corresponding definitions.

– We annotate these definitions using TreeTagger, then, we extract specific terms (nouns) from these annotated definitions. They are the candidate terms for the original tweet.

3.1.3. *Knowledge Base Extraction*

The Dbpedia Ontology currently contains about 4,233,000 instances. The table 2 below lists the number of instances for several classes withClass Instances

Dbpedia concepts are described by short and long abstracts in 13 different languages. These abstracts have been extracted from the English, German, French, Spanish, Italian, Portuguese, Polish, Swedish, Dutch, Japanese, Chinese, Russian, Finnish and Norwegian versions of Wikipedia

[1] http://wikipedia-miner.cms.waikato.ac.nz/

Table 2. *Instances per class.*

Resource(overall)	4,233,000
Place	735,000
Person	1,450,000
Work	411,000
Species	251,000
Organisation	241,000

We use information coming from Dbpedia to extend the original tweet. This method consists in extracting for each term in the original tweet, a set of related information from the Dbpedia ontology. These related information present the candidate terms for the original tweet. This is done using a SPARQL query: by $rdf : type$. SPARQL query is a semantic query language for databases, able to retrieve and manipulate data stored in Resource Description Framework (RDF) format.

3.2. *Step 2 : Candidate Terms Ranking and Selection*

This step consists in ranking the candidate terms according to their semantic relatedness to the given tweet, and selecting the best ones to be added. To achieve this, we propose a new semantic relatedness measure (*E*SAC) that combine the Wikipedia-based Explicit Semantic Analysis (*ESA*) measure and the association rules' confidence value.
Explicit Semantic Analysis (ESA) is a promising approach that calculates semantic relatedness proposed by Gabrilovich and Markovitch (Gabrilovich and Markovitch, 2007). It is a vectorial representation of text that uses Wikipedia as a knowledge base. Specifically, in ESA, a word is represented as a column vector in the TF–IDF matrix of the corpus and a text is represented as the centroid of the vectors representing its words. Semantic relatedness of two given texts can be obtained by calculating the correlation between two high dimensional vectors generated by *ESA*.

– The ESAC measure between a tweet tw and a term w is defined as:

$$ESACtw, w = \begin{cases} \alpha\, ESAtw, w \;+\; (1-\alpha)\, Conf_{R(tw,w)} \;\; if\ \exists\, R(tw, w); \\ ESAtw, w,\ otherwise. \end{cases} \qquad [3]$$

Where

- $ESAtw, w$ is the score of relatedness between the tweet tw and the candidate term w calculated as follow:

$ESAtw, w = \frac{tw.w}{\|tw\|^2 \|w\|^2}$.

- $Conf_{R(tw,w)} = MAX \ \{Conf_{R_j(wt,w)}\}$ with $MAX\{A\} = m \ , \ m \in A$ is a maximal element of A if for all $s \in A, m \leq s$ implies $m = s$.

- $Conf_{R(wt,w)}$ is the confidence of the rule R that express the association between the candidate term w and a word in tweet wt.

- α is a weighting parameter $\in [0, 1]$.

We used the *ESA* implementation described in (Gabrilovich and Markovitch, 2007), we realized our runs with the best parameter value obtained by experiments.
Once we calculate the semantic relatedness between the tweet and the candidate terms, we selecting the most related ones that have semantic relatedness score greater than a determined threshold and adding them to the original tweet.

4. Experimentations

Once, as described above, we have generated the extended tweets, we need to extract the tweet context. This context takes the form of an easy-to-read summary, composed of passages from a provided Wikipedia corpus.

To achieve this, as described in the Introduction, we use the system provided by the INEX 2014 Tweet Contextualization organizers[2] composed of : an Information Retrieval System (IRS) to find the most relevant Wikipedia articles, and an Automatic Summarizer System (ASS) to extract, from the relevant Wikipedia articles, the passages most representative of the tweet. This system was available to participants through a web interface or a Perl API. The system receives as input a query and returns a context. This latter consists of Part-Of-Speech (POS) sentences annotated with TreeTagger . This annotation process allows to assign a score for each sentence using TermWatch. The set of sentences, not exceeding 500 words (this limit is established by the organizers), defines the context of the tweet.

4.1. *Data Collection*

The tested INEX 2014 collections contains:

1) A corpus of 3 902 346 articles rebuilt in 2013 from a dump of the English Wikipedia of November 2012. All notes and bibliographic references that are difficult to handle are removed and only non-empty Wikipedia pages (pages having at least one section) are kept. Resulting documents are made of a title (title), an abstract (a) and sections (s). Each section has a sub-title (h). The abstract and sections are made of paragraphs (p) and each paragraph can have entities (t) that refer to Wikipedia pages. Each document is provided in XML format.

[2] http://qa.termwatch.es/data

2) A collection of English tweets, composed of 240 tweets selected from the CLEF RepLab 2013. To focus on content analysis alone, urls are removed from the tweets.

4.2. *Runs*

We conducted different runs (*cf.* Table 3), namely:

(a) run-Wikipedia-textmining: Tweet expansion based on text mining without terms ranking and selecting step.

(b) run-ESAC-Wikipedia-textmining: Tweet expansion based on text mining with terms ranking and selecting step.

(c) run-Wikipedia-RA: Tweet expansion based on association rules mining without terms ranking and selecting step.

(d) run-ESAC-Wikipedia-RA: Tweet expansion based on association rules mining with terms ranking and selecting step.

(e) run-Dbpedia: Tweet expansion based on knowledge base (Dbpedia) without terms ranking and selecting step.

(f) run-ESAC-Dbpedia: Tweet expansion based on knowledge base (Dbpedia) with terms ranking and selecting step.

(g) run-ESAC-Wikipedia-Dbpedia: Combining run-ESAC-Wikipedia-textmining, run-ESAC-Wikipedia-RA and run-ESAC-Dbpedia.

Table 3. *The conducted runs.*

	ESAC Ranking and selection		No ranking, no selection
Association rules mining	g	d	c
Text mining		b	a
Knowledge base extraction		f	e

The parameters fixed for the experiments are:
$\alpha = 0.5$, $minsupp = 15$,$minconf = 0.7$.

4.3. *Evaluation Metric*

We have evaluated our runs according to the **Informativeness** metric, this latter is proposed by the INEX organizers, it aims at measuring how well the summary helps a user understand the tweets content. There are actually 47 tweets, from the 240 tweets, used for the INEX 2014 evaluation. For each tweet, each passage will be evaluated independently from the others, even in the same summary. The results are based on a thorough manual run on 1/5 of the 2014 topics using the baseline system. From this run two types of references were extracted, namely:

– a list of relevant sentences per topic.

– extraction of Noun Phrases from these sentences together with the corresponding Wikipedia entry.

The dissimilarity between a reference text and the proposed summary is given by:

$$Dis(T,S) = \sum_{t \in T} (P-1) \times \left(1 - \frac{min(log(P), log(Q))}{max(log(P), log(Q))}\right) \qquad [4]$$

where :

– T, a set of query terms present in reference summary.

– S, a set of query terms present in a submitted summary.

– $f_T(t)$, the frequency of term t in reference summary.

– $f_S(t)$, the frequency of term t in a submitted summary.

– $P = \frac{f_T(t)}{f_T} + 1$.

– $Q = \frac{f_S(t)}{f_S} + 1$.

There are different distributions for the reference summaries, namely:

– 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 (Also referred to as skip distribution).

4.4. ***Results***

proportion of the improvement / descent in the Bigrams with 2-gaps compared to run 361

4.4.1. *Within INEX 2014*

Table 4 presents the official runs submitted by INEX 2014 participants from six countries (Canada, France, Germany, India, Russia, Tunisia).

We submitted the run:*run-Wikipedia-RA (361) (c)* to take part of the INEX 2014 competition, and we achieved the best informativeness results (the results are sorted by performance on Bigrams with 2-gaps and the lowest scores represent the best runs).

4.4.2. *After INEX 2014*

After winning the INEX 2014 competition for the tweet contextualization task, we continued to improve our results.

Table 5 depicts our obtained results. As seen in the table, *run-ESAC-Wikipedia-Dbpedia (g)* has achieved the best informativeness results and has outperformed the

Table 4. *INEX Tweet Contextualization 2014 official informativeness results based on sentences.*

Run	**Unigrams**	**Bigrams**	**Bigrams with 2-gaps**	INEX system
361 (c)	0.7632	0.8689	0.8702	YES
360	0.782	0.8925	0.8934	YES
368	0.8112	0.9066	0.9082	NO
369	0.814	0.9098	0.9114	NO
359	0.8022	0.912	0.9127	YES
370	0.8152	0.9137	0.9154	NO
356	0.8415	0.9696	0.9702	NO
357	0.8539	0.97	0.9712	NO
364	0.8461	0.9697	0.9721	-
358	0.8731	0.9832	0.9841	NO
363	0.8682	0.9825	0.9847	-
362	0.8686	0.9828	0.984	-

Table 5. *The obtained informativeness results based on sentences.*

Run	**Unigrams**	**Bigrams**	**Bigrams with 2-gaps**
g	0.7494	0.8520	0.8535
b	0.7613	0.8629	0.863
f	0.7610	0.8629	0.8638
a	0.7665	0.8661	0.8668
d	0.7612	0.8671	0.8695
c	0.7632	0.8689	0.8702
e	0.7940	0.8822	0.8831

other runs, this is due to the combination of the information coming from the two knowledge sources, Wikipedia and Dbpedia.
run-ESAC-Wikipedia-textmining (b), run-ESAC-Wikipedia-RA (d), run-ESAC-Dbpedia (f) outperform *run-Wikipedia-textmining (a), run-Wikipedia-RA (c), run-Dbpedia (e)* this is due to the term ranking step that reduced the noise in the extended tweets by eliminating the non related terms, and fine-grained the semantic representation of the tweets, indeed, the use of association rules that led to the enforcement of the relatedness score between the candidate terms and the tweet, ensured that the extended tweets contain adequate correlating terms with the initial ones and helped avoid inclusion of non-similar terms in them as much as possible, so the extended tweets were, to some extent, clean. We performed a bilateral paired Student t-test to evaluate the statistical significance of the differences of the averages between our best official run from INEX 2014 run-Wikipedia-RA (361) (c) from Table 4 and the best run from our proposals, run-ESAC-Wikipedia-Dbpedia (g) from table 5.

The differences, respectively for Unigrams, Bigrams or Bigrams with 2 gaps, are not significant according to the significance threshold of 0.05. However, we noticed that, for three topics (with tweet ids 257798105473380352, 262290292173045762 and 276815901897146368), the run- ESAC-Wikipedia-Dbpedia (g) largely underperforms the run-Wikipedia-RA (361) (c).

Table 6 presents the tweet with id 257798105473380352. Table 7 shows the expansions used for the runs run-Wikipedia-RA (361) (c) and run-ESAC-Wikipedia-Dbpedia (g) for the topic 257798105473380352. From this table, we find that the number of terms added for run-Wikipedia-RA (361) (c) is much smaller than for run-ESAC-Wikipedia-Dbpedia (g). In this large number of terms (exceeds 30) for the run-ESAC-Wikipedia-Dbpedia (g), noisy terms appear like "revenue", "english" and "country". The same observation is made on the two others topics mentioned above. We believe that this problem may be corrected in the future by limiting the number of added terms for the query expansion.

We propose to compare the runs run-Wikipedia-RA (361) (c) and run-ESAC-Wikipedia-Dbpedia (g), when our proposal run-ESAC-Wikipedia-Dbpedia (g) does not fail. We obtain the results presented in table 8. The respective differences between the run-Wikipedia-RA (361) (c) and our proposal run-ESAC-Wikipedia-Dbpedia (g) on these 44 topics are all statistically significant (noted † in table 8) according to bilateral paired Student t-tests with significance threshold of 0.05. This proves that, if we are able to detect in advance when our proposal fails, we can greatly improve the state of the art results.

Table 6. *The tweet with id 257798105473380352.*

automotive Fiat S.p.A CNH-Fiat merger CNH rejects merger proposal from Fiat Industrial: US farm and industrial vehicle group CNH has rejected a merger proposal from its pa... 2012-10-15-13:00

5. Conclusion and Perspectives

In this paper, we presented our works in the tweet contextualization field. We proposed different methods, based on Wikipedia and Dbpedia, to expand the tweets. Our proposed methods are divided into two steps, the first step generates the candidate terms and the second one ranks them and expends the original tweets by the most related ones. We conducted our experimentations on the INEX 2014 collection. The results we obtained through the different performed runs showed a significant improvement in the informativeness of the contexts, and have outperformed the winning run, which we submitted to the INEX 2014 tweet contextualization task. In our future work, we intend to extend our approach to take into account the tweets specificities (#,@,...). Furthermore, we mean to generalize our approach will

Table 7. *The tweet (id= 257798105473380352) expansions by run-Wikipedia-RA (361) (c) and run-ESAC-Wikipedia-Dbpedia (g)*

Tweet id	Expansion for run c	Expansion for run g
257798105473380352	car vehicle government acquisition model market automaker auto production america state	business car vehicle manufacturing engine group turin acquisition model production market auto fiat industry design manufacture world sector revenue factory torino italy automotive automobile manufacturer sector revenue usa company organisation country english automaker government world

Table 8. *Informativeness results based on sentences, on the 44 selected runs.*

Run	**Unigrams**	**Bigrams**	**Bigrams with 2-gaps**
g	0.7364 †	0.8439 †	0.8454 †
c	0.7800	0.8912	0.8923

contextualize normal (regular) queries by applying it on other data collections such as Cultural Heritage in CLEF collection (ChiC). This latter contains short queries that have no sufficient information to express their semantic.

Acknowledgments. This work is partially supported by the French-Tunisian project PHC-Utique RIMS-FD 14G 1404.

6. References

Bandyopadhyay A., Ghosh K., Majumder P., Mitra M., "Query expansion for microblog retrieval", *IJWS*, vol. 1, n^{o} 4, p. 368-380, 2012.

Deveaud R., Boudin F., "Contextualisation automatique de Tweets à partir de Wikipédia", *CORIA 2013 - Conférence en Recherche d'Infomations et Applications - 10th French Information Retrieval Conference, Neuchâtel, Suisse, April 3-5, 2013.*, p. 125-140, 2013a.

Deveaud R., Boudin F., "Effective Tweet Contextualization with Hashtags Performance Prediction and Multi-Document Summarization", *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*, 2013b.

Ermakova L., Mothe J., “IRIT at INEX 2012: Tweet Contextualization”, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, 2012.

Gabrilovich E., Markovitch S., “Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis”, *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, p. 1606-1611, 2007.

Latiri C. C., Yahia S. B., jean-pierre chevallet, Jaoua A., “Query expansion using fuzzy association rules between terms”, *JIM’2003, France.*, 2003.

Lau C. H., Li Y., Tjondronegoro D., “Microblog Retrieval Using Topical Features and Query Expansion”, *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, November 15-18, 2011*, 2011.

Linhares A. C., “An Automatic Greedy Summarization System at INEX 2013 Tweet Contextualization Track”, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*, 2013.

Meij E., Weerkamp W., de Rijke M., “Adding semantics to microblog posts”, *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*, p. 563-572, 2012.

Milne D. N., Witten I. H., “An open-source toolkit for mining Wikipedia”, *Artif. Intell.*, vol. 194, p. 222-239, 2013.

Morchid M., Dufour R., Linéars G., “LIA@inex2012 : Combinaison de thèmes latents pour la contextualisation de tweets”, *13e Conférence Francophone sur l’Extraction et la Gestion des Connaissances*, Toulouse,France, 2013.

Shekarpour S., Höffner K., Lehmann J., Auer S., “Keyword Query Expansion on Linked Data Using Linguistic and Semantic Features”, *2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16-18, 2013*, p. 191-197, 2013.

Song M., Song I., Hu X., Allen R. B., “Integration of association rules and ontologies for semantic query expansion”, *Data Knowl. Eng.*, vol. 63, n^o 1, p. 63-75, 2007.

Tan K. L., Almasri M., Chevallet J., Mulhem P., Berrut C., “Multimedia Information Modeling and Retrieval (MRIM) /Laboratoire d’Informatique de Grenoble (LIG) at CHiC2013”, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*, 2013.

Torres-Moreno J., “Three Statistical Summarizers at CLEF-INEX 2013 Tweet Contextualization Track”, *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, p. 565-573, 2014.

Xia T., Chai Y., “An improvement to TF-IDF: Term Distribution based Term Weight Algorithm”, *JSW*, vol. 6, n^o 3, p. 413-420, 2011.

Yahia S. B., Nguifo E. M., “Approches d’extraction de règles d’association basées sur la correspondance de Galois”, *Ingénierie des Systèmes d’Information*, vol. 9, n^o 3-4, p. 23-55, 2004.

Zaki M., Hsiao C.-J., “An efficient algorithm for closed itemset mining”, *Second SIAM International Conference on Data Mining*, 2002.

Zingla M. A., Ettaleb M., Latiri C. C., Slimani Y., “INEX2014: Tweet Contextualization Using Association Rules between Terms”, *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, p. 574-584, 2014.