Statistical and Semantic Approaches for Tweet Contextualization

Meriem Amina Zingla, Latiri Chiraz, Yahya Slimani, Catherine Berrut

Statistical and Semantic Approaches for Tweet Contextualization

Meriem Amina Zingla[a], Latiri Chiraz[b], Yahya Slimani[c], Catherine Berrut[d]

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

Abstract

Microblogging sites, like Twitter, have emerged as a popular platform for expressing opinions. Bound to 140 characters, Twitter's publications (tweets) are very short and not always written maintaining formal grammar and proper spelling. The spelling variations increase the likelihood of vocabulary mismatch and make the tweets difficult to understand without some kind of context. The task of tweet contextualization was organized around these issues. It aims to provide an automatic readable context explaining a given tweet, in order to help the reader understand this latter. In this paper, we describe statistical and semantic approaches for the tweet contextualization task. While the statistical one is based on association rules mining, the semantic one uses Wikipedia as an external knowledge source. The effectiveness of our approaches is proved through an experimental study conducted on the INEX 2013 collection.

1. Introduction and motivations

Microblogging has emerged as one of the primary social media platforms for users to submit, in real-time, short messages to report an idea, an actual interest, or an opinion [1].
Twitter is one of the most popular microblogging services providers, it has become a source for discovery, with a focus on sharing relevant information and engaging in conversations. Indeed, Twitter allows to post short messages, not exceeding 140 characters, called tweets. The aim is to exchange a maximum of information in as little characters as possible [2]. However, this limit causes users to employ different strategies such as abbreviations and slangs in order to compress more information. Tweets are, therefore, often misspelled or truncated and especially hard to understand. Hence, to make them understandable by readers, it is necessary to find out their contexts.

* Corresponding author. Tel.: +21652695369.
E-mail address: zinglameriem@gmail.com

To address these issues in an efficient and effective manner, INEX launched the tweet contextualization track in 2011. Thus, the INEX 2013 tweet contextualization track proposed to answer questions of the form "What is this tweet about?" using a recent cleaned dump of the Wikipedia in order to allow the reader a better understanding of the tweet. The general process involves three steps, namely :

- Tweet analysis.
- Passage and/or XML elements retrieval, using an information retrieval system (IRS) based on the Indri[1] search engine.
- Construction of the answer, using an automatic summarization system (ASS) based on an efficient summarization algorithm created by TermWatch[2].

A baseline system composed of an IRS and an ASS has been made available online[3]. The system was available to participants through a web interface or a perl API.
The system receives as input a query in the Indri language 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. This set of sentences, not exceeding 500 words, forms the context of the tweet.
Despite the fact that the idea to contextualize tweets is quite recent, there are several works in this field, such as in[2] where authors used latent Dirichlet analysis to extend the initial query. Authors in[3] used an automatic greedy summarization system to select the most relevant sentences, while authors in[4] developed three statistical summarizers to build the context of the tweet. In[5], authors proposed a contribution composed of three main components: preprocessing, Wikipedia articles retrieval and multi-document summarization.
More closer to our work, in[6], authors presented a novel approach for mining knowledge supporting query expansion based on association rules. The key feature of this approach is a better trade-off between the size of the mining result and the conveyed knowledge. An experimental study has examined the application of association rules to some real collections, whereby automatic query expansion has been performed. The results showed a significant improvement in the performances of the information retrieval system, both in terms of recall and precision. Authors of[7] used Wikipedia as a semantic source to extend the original query by adding the titles of best k related articles to the given query.
Inspired by these two last cited works, we propose two approaches based on association rules mining and Wikipedia as an external knowledge source in order to extend the original tweets. We claim that the use of a semantic knowledge source for the tweet contextualization will be fruitful for the task of tweet contextualization, because it allows to interrogate a big source of knowledge, in our case Wikipedia.
We opted to use Wikipedia because it is currently the largest knowledge repository on the Web. Wikipedia is available in dozens of languages, while its English version is the largest of all with 4,848,394 articles increased every day with over 800 new articles.
It is worth noting that a synergy with some advanced text mining methods, especially association rules[8], is particularly appropriate for tweets expansion and contextualization. Applying association rules for tweet contextualization is far from being a trivial task, mostly because of the huge number of potentially interesting rules that can be drawn from a Wikipedia document collection.
The remainder of the paper is organized as follows: Section 2 cites some related works. Section 3 explains the problem and introduces the main definitions related to tweet contextualization task. In section 4 a detailed description of our approaches for tweet contextualization is presented. Experimental settings and results are given in section 5. Finally, section 6 concludes this article and introduces future works.

[1] http://www.lemurproject.org/indri.php
[2] http ://data.termwatch.es
[3] http://qa.termwatch.es/data

2. Related Works

Several works in the literature are proposed in response to the tweet contextualization task. Recently, authors in [9] proposed a new 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.
In [2], 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 gave good results for the tweet contextualization task. While in [10], 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. The main drawback 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 [5], authors added a hashtag performance prediction component to the Wikipedia retrieval step. They used all available tweet features including web links which was not allowed by INEX's organisers.
In [3], 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 the last edition of INEX, authors in [4] tried to improve upon the tweet contextualization system by developing three statistical summarizer systems. The first one is called Cortex summarizer which uses several sentence selection metrics and an optimal decision module to score sentences from a document source. The second one is called Artex summarizer and uses a simple inner product among the topic-vector and the pseudo-word vector. And, the third one is called Reg summarizer which is a performant graph-based summarizer. These three summarizers achieved satisfactory results.
On another side, query expansion using association rules was already used in the classical Information Retrieval, such as in [11], authors proposed a noval semantic query expansion technique that combines association rules with ontologies and information retrieval, they used the association rules discovery to find good candidate terms to improve the retrieval performance. In [12], 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. Moreover, authors of [6] proposed an approach for mining knowledge supporting query expansion that is based on association rules. The key feature of this approach is a better trade-off between the size of the mining result and the conveyed knowledge using a generic basis of association rules.

3. Problem Statement and Basic Definitions

After introducing some notations, we state the formal definitions of the concepts used in the remainder of the paper. In this respect, we shall use in text mining field, the theoretical framework of Formal Concept Analysis (FCA) presented in [13]. First, we formalize the tweet contextualization task.

So, the tweet contextualization task is concerned with contextualizing a set of n tweets $\tau = \{tw_1, \ldots, tw_n\}$ using a collection of m Wikipedia articles $\sigma = \{d_1, \ldots, d_m\}$ by providing a context c_i for each tweet $tw_i \in \tau$. For a given tweet tw_i, we retrieve a sub-set σ_p of relevant articles from σ, then we select the most relevant passages from the articles in σ_p. These passages define the context c_i.

3.1. Tweet Representation

We represent a tweet as bag of words. We do not access the tweet directly in our proposed approaches, but apply a preprocessing step first, wich removes all stop-words and twitter's specific stop-words such as (RT, @username). Formally, we have:

$$tw_i = \left\{wt_1 \dots wt_j\right\} \tag{1}$$

where
wt_j : *is a word in* tw_i
$i, j \in \mathbb{N}$

3.2. Tweet Context Representation

For a given tweet tw_i, a context is a concatenation of passages from the σ_p sub-set. These passages are composed of set of words. Formally, we have:

$$c_i = \{wc_1, \dots, wc_k\} \tag{2}$$

where
wc_k : *is a word in* c_i associated to the tweet tw_i
$k \in \mathbb{N}$ *and* $k \leq 500$

3.3. Association Rules Mining

An association rule, *i.e.*, between terms or between syntagms, is an implication of the form R: $T_1 \Rightarrow T_2$, where T_1 and T_2 are subsets of $\mathcal{I}$, where $\mathcal{I} := t_1, ..., t_l$ is a finite set of l distinct terms in the Wikipedia document collection and $T_1 \cap T_2 = \emptyset$. The termsets T_1 and T_2 are, respectively, called the *premise* and the *conclusion* of R. The rule R is said to be based on the termset T equal to $T_1 \cup T_2$. The *support* of a rule R: $T_1 \Rightarrow T_2$ is then defined as:

$$Supp(R) = Supp(T), \tag{3}$$

while its *confidence* is computed as:

$$Conf(R) = \frac{Supp(T)}{Supp(T_1)}. \tag{4}$$

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.

4. The Proposed Approaches for Tweets Contextualization

The tweet contextualization task is used to extract a context for a given tweet. The main goal is to enhance the quality of this context, *i.e.*, ensuring that the context summaries contain adequate correlating information with the tweets and avoiding the inclusion of non-similar information. In the previous editions of INEX, almost all participants used language models [14]. To address tweet contextualization in an efficient manner, we propose two approaches namely: A Statistical Approach based on Association Rules inter-Terms and inter-Syntagms (ARTS), and a Semantic Approach based on Wikipedia as a Semantic Source (WSS).
Our proposed approches for tweet contextualization offer an interesting solution to obtain relevant context . This mainly relies on an accurate choice of the added terms to an initial tweet. Interestingly enough, tweet contextalization takes advantage of large text volumes provided by wikipedia articles by extracting statistical and semantic information.

4.1. Statistical Approach based on Association Rules inter-Terms and inter-Syntagms (ARTS)

The main idea of this approach is to extract a set of non redundant rules, representing inter-terms and inter-syntagms correlations in a contextual manner. We use these rules that convey the most interesting correlations amongst terms and syntagms, to extend the initial tweets. Then, we extract the context of each tweet using the baseline system. To achieve this, we use some heuristics, namely:

- Selecting a sub-set of Wikipedia articles, from the documents collection, according to the tweet's subject, using an algorithm based on the TF-IDF measure [15].
- Annotating the selected Wikipedia 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. (*cf.* Table 1).

Table 1. TreeTagger results

Word	**POS Tag**	**Description**	**Lemma**
Film	NN	noun, singular or mass	film
movie	NN	noun, singular or mass	movie
Skyfall	NP	proper noun, singular	Skyfal
is	VBZ	verb be, pres, 3rd p. sing	be
the	DT	determiner	the
name	NN	noun, singular or mass	name

- Extracting of terms and syntagms from the annotated Wikipedia articles, where terms can be a proper noun (singular), or a proper noun (plural), or a noun. Thus, we consider syntagms as sequences of:
  - Noun-noun
  - Noun-noun (plural)
  - Adjective-noun
  - Adjective (superlative)-noun
  - Proper noun (singular)-proper noun (singular)
  - Adjective-noun-noun
  - Noun-adjective-noun
  - Adjective-adjective-noun (plural)
  - Noun-preposition-noun
  - Noun-preposition-noun (plural)
  - Noun (plural)-preposition-noun (plural)
  - Noun (plural)-preposition-noun
- Generating the association rules using an efficient algorithm for mining all the closed frequent termsets. We adapted the algorithm CHARM [16], because it allows to generate non-redundant association rules [17]. As parameters, CHARM takes *minsupp* as the relative minimal support [6] and *minconf* as the minimal threshold to derive valid association rules and gives as output, a set of association rules with their appropriate support and confidence.
- Obtaining the thematic space of each tweet by projecting the tweets on the set of the association rules. This is done by projecting the terms of the tweet, (for example, the term*"tunisia"*, (*cf.* Figure1) on the premises of the association rules, (*tunisia ==> state*), and enriching the tweet using their conclusions, (add the term *state* to the initial tweet).
- Creating the query from the terms of the tweet and the thematic space. This query is then transformed to its Indri format.
- Sending the query to the baseline system to extract from a provided Wikipedia corpus a set of sentences representing the tweet context not exceeding 500 words (this limit is established by the INEX organizers). (*cf.* Figure 1).

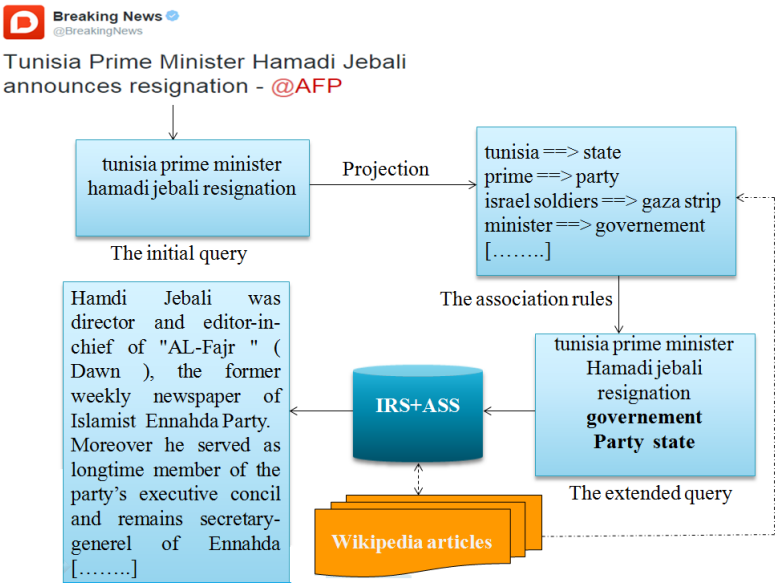

Fig. 1. An illustrative example of how to contextualize a tweet using association rules.

The advantage of the insight gained through association rules is in the contextual nature of the discovered inter-term and inter-syntagm correlations. Indeed, more than a simple assessment of pair-wise term occurrences, 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 the terms of the conclusion in a document, given that those of the premise are already there. The use of such dependencies in a tweet expansion process should significantly increase the quality of the derived context.

4.2. Semantic Approach based on Wikipedia as a Semantic Source (WSS)

We propose another method that allows to select related terms for a given tweet using Wikipedia as an external knowledge source. This latter has become a huge phenomenon among Internet users. Wikipedia is about as accurate in covering scientific topics as the Encyclopedia Britannica[4]. It covers concepts of various fields such as Arts, Geography, History, Science, Sports and Games becoming a database storing all kinds of human knowledge [18].

Our approach consists in exploring the Wikipedia articles related to the query, and using the terms appearing in these articles' first sentences, we call them: *definitions* (*cf.* Figure 2), to extend the original query. The following steps detail our proposed approach.

- Given a tweet tw_i, first, we search for all articles that correspond to each word $wt_j \in tw_i$ in Wikipedia, for example, for BMW we find the following articles: BMW, BMW motorrad, BMW in formula one, BMW in motorsport, etc.
- We select the most referenced articles, these latter follow a predictable layout, which allows to provide a short definitions of wt_j by extracting the corresponding article's first sentence and paragraph. For the previous example, we retrieve the following definition :
  "(BMW), () is a German automobile, motorcycle and engine manufacturing company founded in 1916." from the BMW article that has the highest relatedness with the query.
- We annotate these definitions using TreeTagger (*cf.* Table 1).

[4] http://www.britannica.com/

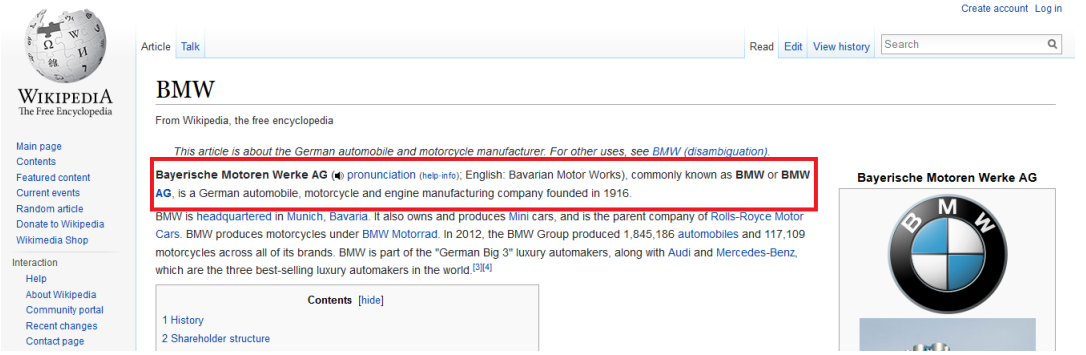

Fig. 2. Article's definition example

- We extract the nouns from these annotated definitions and add them to the original tweet.
- We transform the query (tweet) to its Indri format.
- We send the query to the baseline system, like in the last step of the previous approach, to extract the context of the tweet.

5. Experimentations, Results and Discussion

In this section, we detail the experimental study of applying our proposed approaches, ARTS and WSS, on the issue of tweets contextualization. We validated our approaches over INEX 2013 [14] collection which contain:

1. A collection of articles, that has been rebuilt based on a recent dump of the English Wikipedia from November 2012. It is composed of 3 902 346 articles, where 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). Abstract and sections are made of paragraphs (p) and each paragraph can have entities (q) that refer to Wikipedia pages. Each document is provided in XML format and respects the Document Type Definition (DTD) described in Table 2.

Table 2. DTD of Wikipedia pages.

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

2. 598 English tweets collected by the organizers from Twitter. 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, @Tennis Tweets, @PeopleMag, @science…).

- The document collection from Wikipedia contained related content, so that a contextualization was possible.

We evaluated our proposed contexts according to the **Informativeness Evaluation metric**[14]. This latter aims at measuring how well the context helps a user understand the tweet content. Therefore, for each tweet, each passage will be evaluated independently from the others, even in the same context. This measure is based on lexical overlap between a pool of relevant passages (RPs) and proposed contexts. Once the pool of RPs is constituted, the process is automatic and can be applied to unofficial runs. The release of these pools is one of the main contributions of tweet contextualization tracks at INEX. The dissimilarity between a reference text and the proposed context is given by:

$$Dis(T,S) = \sum_{t \in T} (P-1) \times \left(1 - \frac{min(log(P), log(Q))}{max(log(P), log(Q))}\right) \quad (5)$$

where :

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

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

T, a set of query terms present in reference context and for each $t \in T$.

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

S, a set of query terms present in a submitted context and for each $t \in S$.

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

T can take three distinct forms, 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).

We conducted two (2) runs, namely:

1. *run-ARTS:* In this run, we used the association rules inter-terms and inter-syntagms to extend the original tweet. We have applied CHARM with the following parameters: *minsupp*= 15 and *minconf*=0.7. While considering the $Zipf$ distribution of the collection, 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.
2. *run-WSS:* In this run, we used the terms appearing in the definition of the Wikipedia articles to extend the tweet. This experiment is done using WikipediaMiner[5]; which is a toolkit developed for tapping the rich semantics encoded within Wikipedia. It 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 have compared our runs with those submitted by INEX 2013 participants:

1. In *Pipeline-system run 266,* participants[19] used a pipeline system that consists of three components: Phrase Chunker, Passage Retriever and Summarizer.

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

2. In *REG run 265,* participants[3] used an automatic greedy summarizer named REG (REsumeur Glouton) which uses graph methods to spot the most important sentences in the document.
3. In *Best run 256,* participants[5] used hashtag preprocessing. They also used all available tweet features including web links.

We notice that these runs used the same data collection aforementioned, the reason for which we chose them to compare our results with.

Table 3 describes our obtained results where the lowest scores represent the best runs. This is justified by the fact that the results are diverging. Though, they did not achieve the best score, our approaches were quite close to the best performing system (run 256).

Table 3. Informativeness evaluation based on all overlapping INEX 2013 tweet contextualization track.

Run	Unigrams	Bigrams	Bigrams with 2-gaps
Best run 256	0.7820	0.8810	0.8861
run-ARTS	**0.8279**	**0.9356**	**0.9362**
run-WSS	**0.8259**	**0.9362**	**0.9404**
REG run 265	0.8793	0.9781	0.9789
Pipeline-system run 266	0.9059	0.9824	0.9835

5.1. Discussion

As seen in Table 3, our approach performed better than REG run 265 and Pipeline-system run 266, but did not exceed the best run (best run 256). Furthermore, our ARTS approach performed better than WSS approach, since it decreased the dissimilarity between the Bigrams with 2-gaps included in the proposed contexts and those included in the reference contexts (0.9404 *vs* 0.9362). This can be explained by the fact that the added terms are chosen based on the confidence of the association rule where only terms associated with the highest rules confidence are picked *(minconf =0.7)*. In other words, association rules allowed us to find the terms having a strong correlation with the tweet's terms. We also note that the use of syntagms in the ARTS has introduced a linguistic aspect to the approach, while in WSS, we added the terms appearing in the definitions of the most related articles to the tweet, ensuring the relatedness of the definition to the tweet. The main drawback is the quality of these terms, since there is no measure that ranks and filtrates them. Moreover, we noticed that our results suffer from a noise problem, *i.e.*, some contexts contain no information that relate to its associated tweet. This is justified by the fact that the terms in our queries are not ranked beforehand and also by the huge number of a derived association rules from the Wikipedia collection. The obtained results could be more competitive with some improvements on the queries sent to the baseline system by adding a disambiguation phase based on Explicit Semantic Analysis (ESA)[20] which measures the strength of the relatedness of a term to a query and allows us to prune those which are non-related.

6. Conclusion

In this paper, we proposed to use statistical and semantic approaches for the tweet contextualization task. while the statistical one is based on the association rules mining, the semantic one uses Wikipedia as an external knowledge source. The experimental study was conducted on INEX 2013 collections. The obtained results through the different performed runs highlighted a satisfactory improvement in the informativeness of the derived contexts. In our future work, we propose to add a disambiguation phase that aims at determining which sense of a word is activated by its use in a particular context. This phase will allow us to fine-grain the tweets by eliminating non-related terms. We also propose to use other structured and semantically enriched data sources, such as DBPedia, UMBEL, Freebase, WordNet etc, as external resources.

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

References

1. Ben-jabeur, L.. *Leveraging social relevance: Using social networks to enhance literature access and microblog search*. Ph.D. thesis; Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier); 2013.
2. Morchid, M., Dufour, R., Linéars, G.. Lia@inex2012 : Combinaison de thèmes latents pour la contextualisation de tweets. In: *13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances*. Toulouse,France; 2013, .
3. Linhares, A.C.. An automatic greedy summarization system at INEX 2013 tweet contextualization track. In: *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. 2013, URL: `http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-CarneiroLinhares2013.pdf`.
4. Torres-Moreno, J.. Three statistical summarizers at CLEF-INEX 2013 tweet contextualization track. In: *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*. 2014, p. 565–573.
5. Deveaud, R., Boudin, F.. Effective tweet contextualization with hashtags performance prediction and multi-document summarization. In: *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. 2013, .
6. Latiri, C.C., Haddad, H., Hamrouni, T.. Towards an effective automatic query expansion process using an association rule mining approach. *J Intell Inf Syst* 2012;**39**(1):209–247. URL: `http://dx.doi.org/10.1007/s10844-011-0189-9`. doi:10.1007/s10844-011-0189-9.
7. 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. In: *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. 2013, URL: `http://ceur-ws.org/Vol-1179/CLEF2013wn-CHiC-TanEt2013.pdf`.
8. Agrawal, R., Imielinski, T., Swami, A.N.. Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993*. 1993, p. 207–216. URL: `http://doi.acm.org/10.1145/170035.170072`. doi:10.1145/170035.170072.
9. Ermakova, L., Mothe, J.. IRIT at INEX 2012: Tweet contextualization. In: *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. 2012, URL: `http://ceur-ws.org/Vol-1178/CLEF2012wn-INEX-ErmakovaEt2012.pdf`.
10. Deveaud, R., Boudin, F.. Contextualisation automatique de tweets à partir de wikipédia. In: *CORIA 2013 - Conférence en Recherche d'Infomations et Applications - 10th French Information Retrieval Conference, Neuchâtel, Suisse, April 3-5, 2013*. 2013, p. 125–140.
11. Song, M., Song, I., Hu, X., Allen, R.B.. Semantic query expansion combining association rules with ontologies and information retrieval techniques. In: *Data Warehousing and Knowledge Discovery, 7th International Conference, DaWaK 2005, Copenhagen, Denmark, August 22-26, 2005, Proceedings*. 2005, p. 326–335.
12. Latiri, C.C., Yahia, S.B., jean-pierre chevallet, , Jaoua, A.. Query expansion using fuzzy association rules between terms. In: *JIM'2003, France, 2003*. ????, .
13. Ganter, B., Wille, R.. *Formal concept analysis - mathematical foundations*. Springer; 1999. ISBN 978-3-540-62771-5.
14. Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., Tannier, X.. Overview of INEX tweet contextualization 2013 track. In: *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. 2013, URL: `http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-BellotEt2013.pdf`.
15. Xia, T., Chai, Y.. An improvement to TF-IDF: term distribution based term weight algorithm. *JSW* 2011;**6**(3):413–420. URL: `http://dx.doi.org/10.4304/jsw.6.3.413-420`. doi:10.4304/jsw.6.3.413-420.
16. Zaki, M., Hsiao, C.J.. An efficient algorithm for closed itemset mining. In: *Second SIAM International Conference on Data Mining*. 2002, .
17. Yahia, S.B., Nguifo, E.M.. Approches d'extraction de règles d'association basées sur la correspondance de galois. *Ingénierie des Systèmes d'Information* 2004;**9**(3-4):23–55. URL: `http://dx.doi.org/10.3166/isi.9.3-4.23-55`. doi:10.3166/isi.9.3-4.23-55.
18. Nakayama, K., Hara, T., Nishio, S.. Wikipedia link structure and text mining for semantic relation extraction. In: *Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008), Tenerife, Spain, June 2nd, 2008*. 2008, p. 59–73. URL: `http://ceur-ws.org/Vol-334/paper-05.pdf`.
19. Ansary, K.H., Tran, A.T., Tran, N.K.. A pipeline tweet contextualization system at INEX 2013. In: *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. 2013, .
20. Gabrilovich, E., Markovitch, S.. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*. 2007, p. 1606–1611. URL: `http://dli.iiit.ac.in/ijcai/IJCAI-2007/PDF/IJCAI07-259.pdf`.