Using Association Rules between Terms and Nominal Syntagms for Tweet Contextualization

Meriem Amina Zingla [1]

University of Carthage, INSAT, LISI research Laboratory, Tunis, Tunisia
zinglameriem@gmail.com

ABSTRACT. *The goal of the tweet contextualization track is providing, automatically, a summary that explains a given tweet. This paper introduces a new approach for tweet contextualization based on association rules. The proposed approach allows the extension of the tweet's vocabulary by a set of thematically related words using mining association rules between terms, and between syntagms. To prove the effectiveness of our approach, an experimental study is conducted on the INEX 2013 collection.*

RÉSUMÉ. *Le but de la tâche de contextualisation des tweets organisée par INEX est de fournir, automatiquement, un résumé qui explique un tweet donné. Cet article présente une nouvelle approche de contextualisation des tweets basée sur les règles d'association entre syntagmes, et entre termes. Cette approche permet d'enrichir le vocabulaire de tweets par un ensemble de mots thématiquement proches. L'approche proposée est validée par une étude expérimentale sur la collection INEX 2013.*

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

MOTS-CLÉS : *Contextualisation des tweets, règles d'association, INEX, Expansion de requêtes.*

[1] Supervisors:

- Chiraz Latiri, University of Tunis El Manar, Faculty of Sciences of Tunis, LIPAH research
- Yahya Slimani, University of Carthage, INSAT, LISI research Laboratory, Tunis, Tunisia

1. Introduction and Motivations

At the beginning, Twitter[1] was designed to allow users to share the answer to the "what are you doing?" question. Today, however, it had evolved into a source of news and novel discoveries as the subscribers answer the "what's going on?" question. Many people now think of Twitter as more of a news source than a social network, using it for networking and discussion based on their own interests by sending short text messages, called ”tweets”. The size of these tweets is limited by a maximum number of characters (not exceeding 140) (Ben-jabeur, 2013).
The limit on the length of a tweet causes a wide use of a particular non standard vocabulary (Choudhury *et al.*, 2007). They are often misspelled or truncated making them hard to understand. To make them understandable by readers, it is necessary to find out their contexts.

The task of tweet contextualization (Bellot *et al.*, 2013) was organized around these issues by INEX[2]. The objectif of this task is to provide some context (summary) about the subject of a given tweet (query) in order to help the reader understand it. This context will take the form an easy-to-read summary, not exceeding 500 words, composed of passages from a provided Wikipedia corpus. *i.e.*, answering questions of the form ”what is this tweet about?” using a recent cleaned dump of Wikipedia.
These questions can be answered by several sentences or by an aggregation of texts. This task can be divided into two subtasks, the first is to find the most relevant Wikipedia articles using an Information Retrieval System (IRS), and the second is to extract, from the relevant Wikipedia articles, the passages most representative of the tweet using an Automatic Summarizer System (ASS). (*cf.* Figure 1).

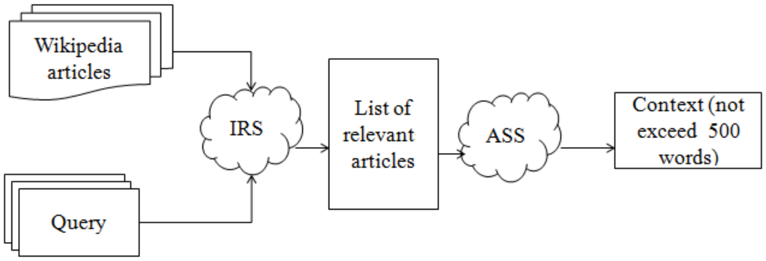


Figure 1. *Information Retrieval System + Automatic Summarizer System*

Several works in the literature are proposed in response to this task, such as (Morchid *et al.*, 2013) where they used latent Dirichlet analysis to extend the initial query,

[1] https://twitter.com/
[2] https://inex.mmci.uni-saarland.de/tracks/qa/

authors of (Linhares, 2013) used an automatic greedy summarization system to select the most relevant sentences, while authors of (Torres-Moreno, 2014) developed three statistical summarizers to build the context of the tweet. In (Deveaud and Boudin, 2013b), authors proposed a contribution that is composed of three main components: preprocessing, Wikipedia articles retrieval and multi-document summarization.

In this paper, we propose to use statistical approaches based on association rules mining (Latiri *et al.*, 2012) between terms and between syntagms[3] to extend the tweets. The main thrust in the proposal is that the proposed approaches gather a minimal set of rules allowing an effective selection of rules to be used in the expansion process.
The remainder of the paper is organized as follows: Section 2 explains the INEX 2013 tweet contextualization task and describes the data collection. Section 3 cites some related works. In Section 4, a detailed description of our approaches for tweet contextualization based on association rules is presented. Experimental settings and results obtained with our approaches are presented in section 5. Finally, in the last section, some concluding remarks and future works are presented.

2. INEX 2013 Tweet Contextualization Track

In the context of micro-blogging, contextualization is specifically important since 140 characters long messages are rarely self-content. This motivated a tweet contextualization track, in 2011, at the Clef INEX lab.
Like in INEX Question Answering track in 2011 and 2012, the present task is about contextualizing tweets, *i.e.* answering questions of the form "What is this tweet about?" using a recent cleaned dump of the Wikipedia. As organizers state[4], the general process involves three steps:

– Tweet analysis.

– Passage and/or XML elements retrieval, using an IRS based on the Indri[5] search engine , (*cf.* Figure 1).

– Construction of the answer, using an efficient summarization algorithm created by TermWatch[6] (ASS), (*cf.* Figure 1).

A baseline system composed of an IRS and an ASS has been made available online[7]. The system was available to participants through a web interface or a perl API.

[3] Syntagm is a sequence of words that constitues what we call a functional (and semantic) unit in a sentence that can be composed of nominal or verbal syntagms.

[4] See the official INEX 2013 Tweet Contextualization Track Website: https://inex. mmci.uni-saarland.de/tracks/qa/.

[5] http://www.lemurproject.org/indri.php

[6] http ://data.termwatch.es

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

2.1. *Data Collection*

In this section, we will briefly describe the tested INEX 2013 (Bellot *et al.*, 2013) collection which contains:

1) A corpus of 3 902 346 articles has been rebuilt in 2013 from a dump of the English Wikipedia from November 2012. 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

2) 598 tweets in English have been 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.

3. Related Works

Despite the recentness of the idea, tweet contextualization has gathered a lot of interest leading to the emergence of several works in this context. Recently, the authors of (Morchid *et al.*, 2013) 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. While in (Deveaud and Boudin, 2013a), authors used a method that allows to automatically contextualize tweets by using information coming from Wikipedia. they treat 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. 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) they added a hashtag performance prediction component to the Wikipedia retrieval step.
The approach proposed in (Lau *et al.*, 2011) is based on a twitter retrieval framework that focuses on using topical features, combined with query expansion using pseudo-relevance feedback (PRF) to improve microblogs retrieval results.
In (Ermakova and Mothe, 2012), a new method based on the local Wikipedia dump is proposed, authors used 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 (Torres-Moreno, 2014) authors developed three statistical summarizer systems the

first one called Cortex summarizer, it uses several sentence selection metrics and an optimal decision module to score sentences from a document source, the second one called Artex summarizer, it 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.
In (Linhares, 2013), authors used an automatic greedy summarizer named REG, the REG summarizer uses a greedy optimization algorithm to weigh the sentences. The summary is obtained by concatenating the relevant sentences weighed in the optimization step.

4. The Proposed Approach for Tweets Contextualization

The tweet contextualization track is used to extract a context for a given query. 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, we proposed two statistical approaches, based on association rules mining (Agrawal *et al.*, 1993; Latiri *et al.*, 2012): Statistical Approach based on Association Rules between Terms (SAART) andStatistical Approach based on Association Rules between Syntagms (SAARS).

4.1. *Statistical Approach based on Association Rules between Terms (SAART)*

Our proposed approach SAART is handled on the following steps :

1) Selection of a sub-set of 50000 articles, from the Wikipedia documents collection, according to the tweet's subject, using an algorithm based on the TF-IDF measure(Xia and Chai, 2011).

2) Annotating the selected Wikipedia articles using TreeTagger [8]. The choice of TreeTagger was based on the ability of this tool to recognize the nature (morphosyntactic 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.

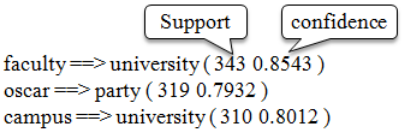
3) Extraction of nouns (terms) from the annotated Wikipedia articles, and removing the most frequent ones.

4) Generating the association rules using an efficient algorithm: Closed Association Rule Mining (CHARM) [9](Zaki and Hsiao, 2002) for mining all the closed frequent termsets, As parameters, CHARM takes *minsupp* = 15 as the relative minimal support (Latiri *et al.*, 2012) and *minconf* = 0.7 as the minimum confidence of the rules

[8] The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart; http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/.

[9] it is an open source project downloaded at http://www.cs.rpi.edu/ zaki/www-new/pmwiki.php/Software/Software

(Latiri *et al.*, 2012), While considering the $Zipf$ distribution of the collection, the maximum threshold of the support values 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. On the other hand, the minimal threshold allows eliminating marginal terms which occur in few documents, and are then not statistically important when occurring in a rule. CHARM gives as output, the association rules with their appropriate support and confidence. Figure 2 describes the output of CHARM.



```
faculty ==> university ( 343 0.8543 )
oscar ==> party ( 319 0.7932 )
campus ==> university ( 310 0.8012 )
```

Figure 2. *An example of association rules between terms extracted by CHARM*

5) Obtaining the thematic space of each tweet by projecting the tweets on the set of the association rules. (*cf.* Figure3).

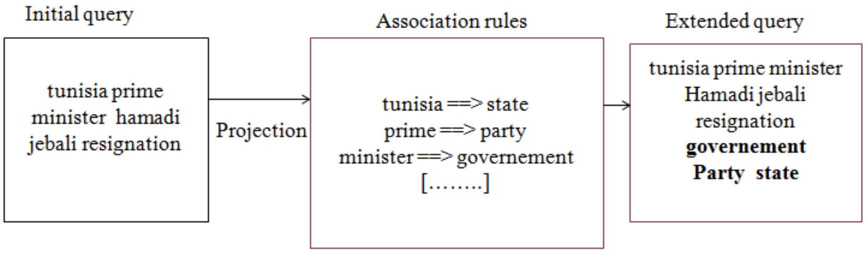


Figure 3. *Projection of the tweet on the set of the association rules to obtain its thematic space*

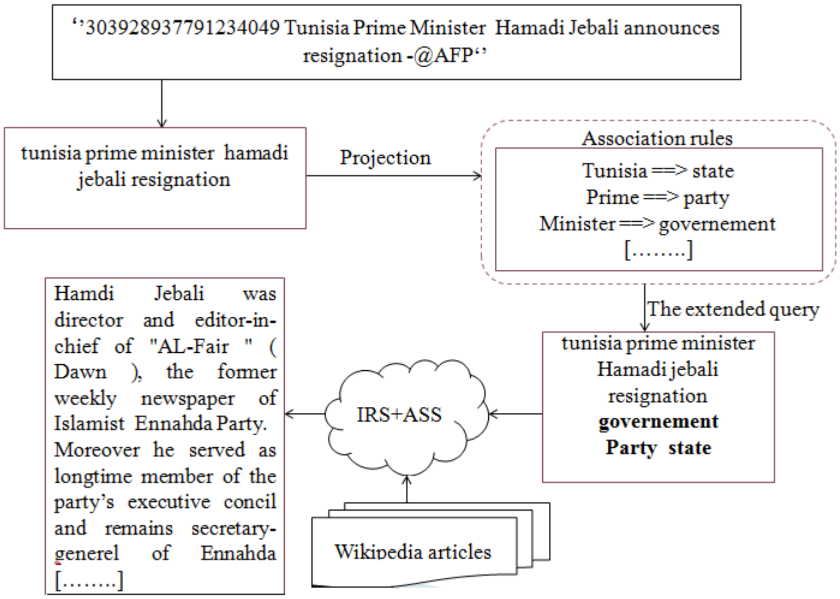
6) Creating the query from the terms of the tweet and the thematic space, this query is then transformed to its Indri format.

7) Sending the query to the baseline system, (*cf.* Figure 1), 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 organizers).

An illustrative example is depicted in Figure 4

4.2. *Statistical Approach based on Association Rules between Syntagms (SAARS)*

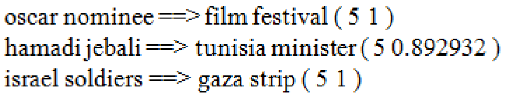
In this approach, we followed the same principle a as the previous one (4.1), the only difference is the extraction of nominal syntagms from the annotated articles.


‘’303928937791234049 Tunisia Prime Minister Hamadi Jebali announces resignation -@AFP‘’
tunisia prime minister hamadi jebali resignation
Projection
Association rules
Tunisia ==> state
Prime ==> party
Minister ==> governement
[……..]
The extended query
tunisia prime minister Hamadi jebali resignation **governement** **Party state**
IRS+ASS
Wikipedia articles

Hamdi Jebali was director and editor-in-chief of "AL-Fair " (Dawn), the former weekly newspaper of Islamist Ennahda Party. Moreover he served as longtime member of the party’s executive concil and remains secretary-generel of Ennahda [……..]

Figure 4. *An illustrative example of how to contextualize a tweet using association rules.*

Association rules between syntagms were, then, generated by applying the CHARM algorithm with the following parameters : *minsupp* = 15, and *minconf* = 0.7.
An example of association rules between syntagms is given in Figure 5.



oscar nominee ==> film festival (5 1)
hamadi jebali ==> tunisia minister (5 0.892932)
israel soldiers ==> gaza strip (5 1)

Figure 5. *An example of association rules between syntagms extracted by CHARM*

5. Experiments, Results and Discussion

In this section we detail the experimental results of applying our proposed approach on the issue of tweets contextualization. We conducted three runs, namely:

1) run-**SAART:** In this run, we used the association rules between terms to extend the original query, *i.e.* , the original tweet.

2) run-**SAARS:** In this run, we used the association rules between nominal syntagms.

3) run-**SAARTS:** In this run, we combined the association rules between terms, with those between syntagms to extend the original tweet.

We have compared our runs with the following different runs submitted by INEX 2013 participants:

1) In *Best run 256,* participants (Deveaud and Boudin, 2013b) used hashtag preprocessing. They also used all available tweet features including web links.

2) In *REG run 265,* participants (Linhares, 2013) used an automatic greedy summarizer named REG (REsumeur Glouton) which uses graph methods to spot the most important sentences in the document.

We evaluated our proposed summaries according to the **Informativeness Evaluation metric**. This latter (Bellot *et al.*, 2013) aims at measuring how well the summary helps a user understand the tweet content. It is based on lexical overlap between a pool of relevant passages (RPs) and participant summaries. 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

Run	Unigrams	Bigrams	Bigrams with 2-gaps
Best run 256	0.7820	0.8810	0.8861
run-SAART	**0.8408**	**0.9406**	**0.9441**
run-SAARS	**0.8364**	**0.9362**	**0.9395**
run-SAARTS	**0.8279**	**0.9356**	**0.9362**
REG run 265	0.8793	0.9781	0.9789

Table 1. *Informativeness evaluation based on all overlapping INEX 2013 tweet contextualization track.*

Regarding this evaluation metric, we observed that our results suffer from too much noise and need to be cleaned. We also noticed that the Run-SAARS performed better than Run-SAART, (*cf.* Table3). This can be explained by the high number of terms that exist in the selected articles compared to the number of syntagms, leading to a query drift. The use of syntagms instead of terms improved our results by decreasing the dissimilarity between the Bigrams with 2-gaps included in the submitted summary and those included in the reference summary (0.9441 *vs* 0.9395), while using both the association rules between terms and between syntagms decreased this dissimilarity even further(0.9441 *vs* 0.9362). In addition to the statistical aspect, the use of the association rules between syntagms has introduced a linguistic one (aspect) in the SAARS which led to the betterment of our results.

Furthermore, the SAART was applied on INEX 2014 tweet contextualization track and it performed the best results in the informativeness evaluation as it depicted in Table 4 (Zingla *et al.*, 2014). This improvement in the SAART results is dependent on the number of the selected relevant articles per tweet. We note that the INEX 2014 tweet collection is divided into four distinctive categories, while the INEX 2013 tweet collection is divided into a larger number of categories which lead to a smaller number of relevant articles associated with each tweet, contrary to INEX 2014.

To study the impact of the selected articles used to extract the association rules on the proposed approach, three (3) runs were submitted, namely:

– In *run-SAART-2014-50000*, the top 50000 relevant articles were selected to extract the association rules.

– In *run-SAART-2014-240*, 240 relevant articles were selected to extract the association rules.

– In *run-SAART-2014-12000*, the top 12000 relevant articles were selected to extract the association rules.

We noticed that the best results (run-SAART-2014-12000) come from the compromise between the relevance and the number of the selected articles.
The extracted association rules are influenced not only by the relevance of the selected articles, but also by their number.

Run	Unigrams	Bigrams	Bigrams with 2-gaps
run-SAART-2014-12000	**0.7632**	**0.8689**	**0.8702**
run-SAART-2014-240	**0.782**	**0.8925**	**0.8934**
run-SAART-2014-50000	**0.8022**	**0.912**	**0.9127**

Table 2. *Informativeness evaluation for INEX 2014 tweet contextualization track*

6. Conclusion

In this paper, we propose to use statistical approaches based on association rules mining between terms and between syntagms to extend the tweets. The experimental study was conducted on INEX 2013 collections. The obtained results confirmed that the synergy between association rules and tweet contextualization is fruitful. Indeed, experimental results through the different performed runs showed an satisfactory improvement in the informativeness of the contexts. In our future work we hope to add a disambiguation phase to reduce the noise in our results by eliminating the non related terms.

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

7. References

Agrawal R., Imielinski T., Swami A. N., “Mining Association Rules between Sets of Items in Large Databases”, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993.*, p. 207-216, 1993.

Ansary K. H., Tran A. T., Tran N. K., “A Pipeline Tweet Contextualization System at INEX 2013”, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*, 2013.

Bellot P., Moriceau V., Mothe J., SanJuan E., Tannier X., “Overview of INEX Tweet Contextualization 2013 Track”, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*, 2013.

Ben-jabeur L., Leveraging social relevance: Using social networks to enhance literature access and microblog search, PhD thesis, Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier), 2013.

Choudhury M., Saraf R., Jain V., Mukherjee A., Sarkar S., Basu A., “Investigation and modeling of the structure of texting language”, *IJDAR*, vol. 10, n^o 3-4, p. 157-174, 2007.

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.

Latiri C. C., Haddad H., Hamrouni T., “Towards an effective automatic query expansion process using an association rule mining approach”, *J. Intell. Inf. Syst.*, vol. 39, n^o 1, p. 209-247, 2012.

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.

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.

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.

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.