Thematic Representation of Short Text Messages with Latent Topics: Application in the Twitter context

Mohamed Morchid, Richard Dufour and Georges Linarès
Laboratoire Informatique d'Avignon (LIA)
University of Avignon
France
{mohamed.morchid,richard.dufour,georges.linares}@univ-avignon.fr

***Abstract*—The amount of information exchanged over the Internet is continuously growing, taking the form of short text messages on microblogging platforms such as Twitter. Due to the limited size of these types of messages, their understanding may require to know the context of their occurrence. In this paper, we propose a higher-level representation of short text messages based on a thematic model obtained by a Latent Dirichlet Allocation (LDA). We propose to evaluate the effectiveness of this short text message representation by using it in the experimental setup of the INEX 2012 tweet contextualization task. This topic-based representation allows to extend the message vocabulary by searching a set of thematically-related words. Results demonstrated the interest of this topic-space based approach for the tweet contextualization task.**

***Keywords*-Short text message, Thematic representation, Latent Dirichlet Allocation, Keyword extraction, Twitter**

I. Introduction

The exponential growth of available data on the Web enables users potentially access to a large amount of information. Micro-blogging platforms evolve in the same way, offering to users an easy way to disseminate ideas, opinions or common facts under the form of short text messages. Depending on the sharing platform used, the size of these messages can be limited to a maximum number of words or characters[1]. This constraint causes the use of a particular vocabulary that is often unusual, noisy, full of new words, including misspelled or even truncated words [1]. Indeed, the goal of these messages is to share the maximum amount of information with a small number of characters. It may thus be difficult to understand the meaning of a short text message (STM) without knowing the general context of its realization.

It is therefore necessary to identify the keywords that represent, as well as possible, the STM content, since these keywords will give information about its meaning. They could directly be chosen from the STM lexicon but, unfortunately, this word set offers only a poor representation of message semantics. This is due to the compactness constraint and to the fact that short messages may be written in a non-standard language. This phenomenon is, for example, a frequent case on the micro-blogging *Twitter* [2] platform.

To overcome this limit, we propose a higher-level representation of a STM by identifying its main topics in addition to its lexicon. This topic-space based approach may be viewed as a short message expansion process that aims at improving the message characterization. We propose a thematic representation using a Latent Dirichlet Allocation (LDA) approach: the STM is mapped into a topic space estimated on a large text corpus, allowing to identify its latent topics. As a result, a short message is represented by its initial written words and by a set of topics, which should help to better understand it.

To assess the interest of this approach, we evaluated it in the context of the INEX 2012 evaluation campaign [3]. The aim of this campaign is to search the context associated with a *tweet* (*i.e.* a STM) in order to help the reader in understanding it. We proposed to extract a list of keywords that will be used in the Information Retrieval (IR) and Automatic Summarization (AS) systems provided by the INEX 2012 organizers to provide a context for a considered *tweet*. The list of keywords is composed by the initial *tweet* word set and by an additional keyword set extracted from the proposed topic-based approach.

In the next section, a related work about keyword extraction and topic modeling is detailed. Section III presents the proposed thematic representation of short text messages. The INEX 2012 *tweet* contextualization task in which the topic-space approach is involved is described in section IV. Experiments and results are reported in section V. Finally, section VII concludes this work and proposes some perspectives.

II. Related Work

The classical bag-of-words approach [4] is usually used for text document representation in the context of keyword extraction. This method estimates *Term Frequency-Inverse Document Frequency* (TF-IDF) of the document terms. Although this unsupervised approach is effective for a large collection of documents, it seems inapplicable in the particular case of short messages as most of the words occur only

[1] For example, the *Twitter* service does not allow the sending of messages whose size exceeds 140 characters.

once (*hapax legomena* [5]). Other studies proposed to use binary classifiers to determine if a word can be considered as a keyword [6], [7]. These supervised methods are hardly applicable to the STM context since the vocabulary used is unpredictable and in a constant evolution (see section I).

Other approaches proposed to consider the document as a mixture of latent topics. These methods build a higher-level representation of the document in a topic space. All of these methods are commonly used in Information Retrieval (IR) field. They consider documents as a bag-of-words without taking account of word order; nevertheless, they demonstrated their performance on various tasks. Several approaches were proposed such as Latent Semantic Analysis (LSA) [8], [9], Probabilistic LSA (PLSA) [10] or Latent Dirichlet Allocation (LDA) [11]. LDA is a generative model which considers a document, seen as a bag-of-words [12], as a mixture probability of latent topics. In opposition to a multinomial mixture model, LDA considers that a theme is associated to each occurrence of a word composing the document, rather than associate a topic with the complete document. Thereby, a document can change topics from a word to another. However, it is noted that the word occurrences are connected by a latent variable which controls the global respect of the distribution of the topics in the document. These latent topics are characterized by a distribution of word probabilities which are associated with them.

PLSA and LDA models have been shown to generally outperform LSA on IR tasks [13]. Moreover, LDA provides a direct estimate of the relevance of a topic knowing a word set. In [14], authors propose to describe a STM (*tweet*) by a single topic. This study showed its effectiveness to identify a unique subject to describe the main idea of a *tweet*. Nonetheless, a single topic seems too limited to describe a short message. Indeed, a STM can spread more than one topic: we think that a set of topics should be considered to fully represent a short message.

III. Short text message topic-based representation

The major limit of STMs lies in their limited and non-standard vocabulary: users can not easily understand the meaning of a short message with its written lexical content only. We then propose to enrich a STM with a higher-level representation by identifying its main topics. We choose to compose this semantic representation by a topic space using a LDA approach (see figure 1). Each topic associated to a STM contains a list of weighted words depending on its relevance to the considered message. The following sections are devoted to the detailed description of our two-steps STM topic-based representation approach.

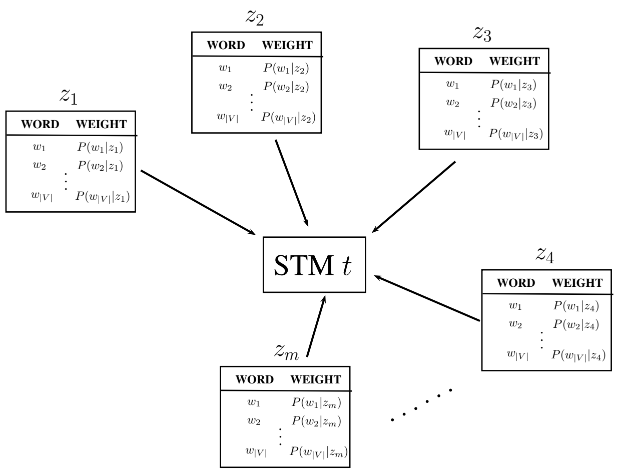

Figure 1. Short text message topic-based representation.

A. Topic space representation

A LDA model is estimated off-line using a large corpus of documents D. This step produces a topic space of size n with a vocabulary $\mathbf{V}$ and a vector V^w representing the distribution of the classes for each word w of $\mathbf{V}$. Each feature V_i^w is the probability of the word w knowing the class z_i stemming from the LDA ($P(w|z_i)$).

B. Estimation of nearest topics

To represent a STM t with its nearest topics from the LDA topic space, the Gibbs sampling algorithm [15] is applied. This algorithm is based on the Markov Chain Monte Carlo (MCMC) method. Thus, the Gibbs sampling allows to obtain samples of the distribution parameters θ knowing a word w of a test document and a given topic z_i. A feature vector V^t is then obtained. The i^{th} feature V_i^t (where $i = 1, 2, \ldots, n$) is the probability of the topic z_i knowing the short message t:

$$V_i^t = P(z_i|t) \tag{1}$$

A set of the m ($m \leq n$) topics with the higher prior probability V_i^t is chosen to compose the topic-based representation of the short message t (Figure 1).

IV. Tweet contextualization task

Our short message topic-based representation method requires an experimental application to evaluate its interest and its effectiveness. The proposed STM representation was involved in the INEX 2012 evaluation campaign [3]. The aim of this campaign is to search the context associated to a short message (*tweet*) in order to help the reader in understanding it.

This task can be divided into three sub-tasks:

1) Defining a query from a considered *tweet*. This query is composed by a set of weighted words depending to its relevance to the *tweet*.
2) Applying the Information Retrieval (IR) system [16], [17] provided by the organizers from the previous query to choose the most relevant documents of a corpus for the considered *tweet*.
3) Extracting the most representative sentences of the *tweet* from the relevant *Wikipedia* documents using the Automatic Summarization (AS) system [18] also provided by the organizers to obtain the context of the *tweet*.

Figure 2 presents the different modules used in the INEX 2012 *tweet* contextualization task. As the IR and AS systems are provided by the organizers, a particular attention should be carried to the composition of the query. We propose to use our STM topic-based representation approach for choosing the weighted words composing the query. The estimated topic space is the higher semantic representation of a *tweet*. Since this space already contains a set of weighted words (see section III), we propose to choose the terms thematically close to the *tweet* by computing a relevance score to each word of the vocabulary **V** representing the short message. The score s of the word w is the prior probability that w is generated by the short message t:

$$\begin{aligned} s(w) &= P(w|t) \\ &= \sum_{i=1}^{m} P(w|z_i)P(z_i|t) \\ &= \sum_{i=1}^{m} V_i^w \times V_i^t \\ &= \langle V^w, V^t \rangle \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the inner product between the vector representation of the word w in the topic space (V^w) and the vector representation of the *tweet* t in the topic space (V^t). As a result, a score $s(w)$ is associated to each word of a STM belonging to the vocabulary **V**.

Finally, we choose to compose the query q by unifying the initial lexical content contained in a *tweet* t (with a weight set to 1) with the word set obtained with our STM topic-based representation weighted with their score $s(w)$. This query q is then sent to the IR and AS systems provided by the evaluation campaign organizers to obtain a set of sentences that represents the query in the best possible way. This set of sentences will compose the context of the *tweet* t as shown in figure 2. An example of a *tweet* contextualization using our method is presented in the figure 3.

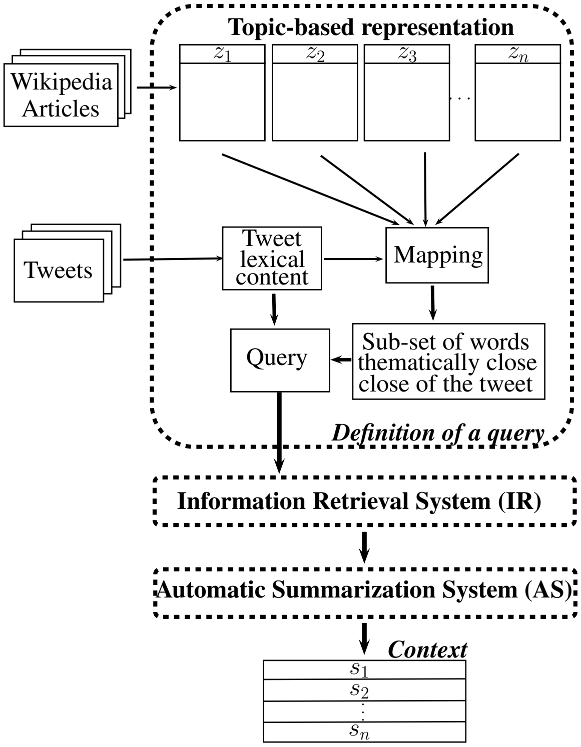
Topic-based representation
z_1
z_2
z_3
z_n
. . .
Wikipedia Articles
Tweets
Tweet lexical content
Mapping
Query
Sub-set of words thematically close close of the tweet
Definition of a query
Information Retrieval System (IR)
Automatic Summarization System (AS)
Context
s_1
s_2
s_n

Figure 2. General description of the INEX 2012 *tweet* contextualization task.

Through the examples described in the table I, we note that the words contained in a *tweet* do not necessarily appear in the vocabulary obtained with the proposed topic-based approach. These examples illustrate our initial motivation: to find a set of thematic words describing a *tweet* and which are not used in the *tweet* itself (*i.e.* its lexical content). Thus, the proposed approach allows to enrich the vocabulary associated to a *tweet*. For example, we can notice in the *tweet* (2) of the table I that some generic terms describing the event (*army, war, muslim* or *islamic*) are absent from the initial *tweet* content.

Table I
EXAMPLES OF TWEETS WITH THE 10 MOST REPRESENTATIVE WORDS. IN **bold** SOME INTERESTING WORDS WHICH DO NOT APPEAR IN THE TWEET.

Tweet	**10 highest score** $s(w)$
celtics blog welcome to the garden celtics (1)	**nba** season game team points **basketball** games time year played
syrian troops attack residential areas in hama and homs (2)	**battle army** street forces troop troops **war muslim** men **islamic** city
bras for after breast implant surgery 3 tips (3)	blood **heart** surgery **pain** body pressure patient patients muscle **tissue**
did you know that 2012 is the international year of sustainable energy for all you canfind out more at our (4)	development international **world environmental global** public human national **policy government**
wow childhood abuse disrupts brain formation study (5)	children **disorder mental** child **therapy syndrome treatment** disorders people symptoms

V. EXPERIMENTS

A. *Experimental protocol*

The proposed system was involved in the INEX 2012 evaluation campaign [3], [19]. A corpus including a set

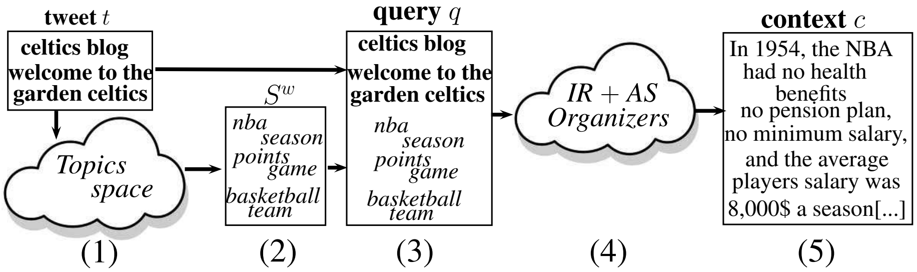
tweet t
celtics blog welcome to the garden celtics
Topics space
S^w
nba season points game basketball team
query q
celtics blog welcome to the garden celtics
nba season points game basketball team
$IR + AS$ Organizers
context c
In 1954, the NBA had no health benefits no pension plan, no minimum salary, and the average players salary was 8,000$ a season[...]
(1)
(2)
(3)
(4)
(5)

Figure 3. Example of a tweet contextualization.

of *tweets* and a set of *Wikipedia* articles to build their context was provided. The INEX 2012 corpus contains 1,142 *tweets* extracted from the *Twitter* platform, which represents 16,263 word occurrences for a vocabulary of 5,287 unique words. Each *tweet* contains an identifier (Id) and its textual content, and do not exceed 140 characters.

A corpus of recent English *Wikipedia* articles (November 2011) is also provided and is composed of about 3.7 million of articles. All notes and bibliographical references were removed from this corpus. Each document is supplied in XML format and follows the *Wikipedia* Document Type Definition (DTD). Finally, this corpus contains around 26 million sentences for a total of about 333 million word occurrences. The *Wikipedia* vocabulary contains 2.8 million of unique words (at least one occurrence in the corpus).

This *Wikipedia* corpus was used to estimate the LDA model. As a result, a space of 400 topics was estimated from which a set of 30 weighted words is selected for each *tweet* from its 5 closest topics ($m = 5$). These words are considered as the best thematically close words of a *tweet* (see section IV).

As specified in the INEX 2012 benchmark, the context associated to each *tweet* should contain almost 500 words. It is obtained using an Information Retrieval (IR) system coupled to an Automatic Summarization (AS) system supplied by the INEX organizers [3]. This one includes:

- An *Indri*[2] index which recovers all words (without the use of a stop-list or *stemming*) and all the XML tags.
- A part-of-speech (POS) system based on *TreeTagger*[3].
- A performant automatic summarization algorithm created by *TermWatch* [18]. This system is accessible by querying a CGI interface with a perl script[4].
- An evaluation tool of summarization systems based on FRESA [20].

This system uses an *Indri* query [21] as input and provides a summary as output. A summary consists of sentences annotated in POS with *TreeTagger*. This annotation process allows to attribute a score to each sentence by using *TermWatch*. This set of sentences constitutes the context of the *tweet*.

B. Results

Results obtained for the INEX 2012 *tweet* contextualization task with our system, the best system and the organizer's system, are detailed in this section. Three different evaluation methods are proposed to evaluate the system performance.

1) Evaluation of the Informativeness: The objective of this metric is to evaluate the selection of relevant sentences [3]. In this particular case, a set of 63 *tweets* composes the evaluation corpus. The 60 best sentences for each *tweet* are selected for the evaluation. This choice is made according to the score attributed by the automatic *tweet* contextualization system, by choosing the highest ones. The lower dissimilarity, the more the proposed summary is similar to the reference text. Terms can take three different forms:

- *Uni-gram*: a unique lemma (base form of the term).
- *Bi-gram*: two successive lemmas in the same sentence.
- *Bi-gram 2-gaps*: identical to the *bi-gram* form, but can be separated by two other lemmas.

Results of our system (run 193) as well as those obtained by the *baseline* system (run 194, supplied by the organizers) and the system which obtained the best score (run 178) are given in the table II.

Table II
OFFICIAL RESULTS FOR THE TWEET CONTEXTUALIZATION TASK AT INEX 2012 WITH THE INFORMATIVENESS METRIC.

Run Id	Description of run	Rank	information metric		
		(in 33)	*Uni-gram*	*Bi-gram*	*Skip-gram*
193	Topic space	7	0.7909	0.8920	0.8938
178	Best run	1	0.7734	0.8616	0.8623
194	Organizer's system	4	0.7864	0.8868	0.8887

2) Evaluation of the Readability: This metric requires the collaboration of the task participants to evaluate all the contexts automatically attributed to the 63 *tweets*. Let us remind that each context can not exceed 500 word occurrences as specified by the organizers [3]. For each sentence to evaluate, participants have to judge if the sentence contains:

[2] http://www.lemurproject.org/indri/

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

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

- *Syntax* (S): a syntax error in the sentence.
- *Anaphora* (A): repetitions of a previous element.
- *Redundancy* (R): a redundant information.
- *Trash* (T): no link with the previous sentence.

Table III presents the results obtained with the readability metric by our system (run 193) as well as those obtained by the *baseline* system (run 194, supplied by the organizers) and the best participant system (run 185).

Table III
OFFICIAL RESULTS FOR THE TWEET CONTEXTUALIZATION AT INEX 2012 WITH THE READABILITY METRIC.

run Id	Description of run	Rank	**read ease metric**		
		(*in 33*)	*Relevance*	*Syntax*	*Structure*
193	Topic space	12	0.6208	0.6115	0.5145
185	Best run	1	0.7728	0.7452	0.6446
194	Organizer's system	4	0.6975	0.6342	0.5703

3) Non-official evaluation of the title context accuracy: Every context is composed by a *Wikipedia* article title. This metric allows to measure the similarity between the reference and the context titles. The results are strongly correlated to those obtained on the informative evaluation of the context (table II). Three classic methods were chosen for the evaluation: the accuracy (measure of the noise), the recall (measure of the silence) and the F-measure (arithmetic mean between the accuracy and the recall). The results obtained with this metric by our system (run 193) as well as those obtained by the organizer's system (run 194) and the best participant system (run 152), are given in the table IV.

Table IV
NON-OFFICIAL RESULTS FOR THE TWEET CONTEXTUALIZATION AT INEX 2012 WITH THE TITLE ACCURACY METRIC.

run Id	Description of run	Rank	**title accuracy metric**		
		(*in 33*)	*Accuracy*	*Recall*	*F-measure*
193	Topic space	10	0.156219	0.442979	0.198238
152	Best run	1	0.321815	0.455337	0.323508
194	Organizer's system	8	0.153116	0.462193	0.210242

VI. RELATION TO OTHER WORKS

Topic space representation of a *tweet* has also been employed by [22]. The authors proposed to use *hashtags* to build a tweet representation without stop-words or useless mentions. Then, they use this short representation as a query to retrieve a set of Wikipedia sentences to build the context of the *tweet*.

Different approaches, such as [23], exploit the tweet content to retrieve a set of passages from Wikipedia to compose the context of a *tweet*. In [24], authors estimate the context of a *tweet* through two processes: an Information Retrieval system developed using the *Nutch* architecture[5] and based on the *Lucene* architecture[6], and an Automatic Summarization system. The proposed module is split into two sub-processes: a Text Filterization and a Sentence Extraction. The authors in [25] propose to exploit the tweet content as unigram, bigram, mention and hashtags in a TF-IDF vector representation, to extract relevant sentences from Wikipedia corpus. The similarity measure between *tweet* content and Wikipedia sentences is the cosine. The next process is crucial in the contextualization task. This process is the sentence ordering that permits to reorder the sentences to increase the readability of the context. The readability is an important aspect in [26]. They obtained good results during this evaluation and more precisely with the readability metric. The authors submitted a run for each of the three sentence selection strategies at the INEX 2012 evaluation campaign: Language Modelling (LM), Relevance Model Similarity (RLM) and Topical Relevance Model Similarity (TRLM).

Table V
ILLUSTRATION OF 6 TOPICS FROM THE SEMANTIC SPACE OF 400 TOPICS. EACH TOPIC IS SHOWN WITH ITS 10 WORDS HAVING THE HIGHEST PROBABILITY.

TOPIC 0

WORD	**PROB.**
LIFE	0.0116
WORLD	0.0106
HUMAN	0.0103
NATURE	0.0067
PHILOSOPHY	0.0066
MIND	0.0057
MAN	0.0048
MORAL	0.0047
THOUGHT	0.0045
GOOD	0.0041

TOPIC 6

WORD	**PROB.**
CODE	0.0192
LANGUAGE	0.0128
DATA	0.0124
PROGRAMMING	0.0108
OBJECT	0.0089
TYPE	0.0085
LANGUAGES	0.0076
PROGRAM	0.0067
FUNCTION	0.0057
CLASS	0.0053

TOPIC 66

WORD	**PROB.**
JAPANESE	0.0280
ATTACK	0.0083
HARBOR	0.0079
PEARL	0.0077
ISLAND	0.0067
AIRCRAFT	0.0066
ENEMY	0.0065
CARRIER	0.0063
ISLANDS	0.0061
SHIPS	0.0060

TOPIC 71

WORD	**PROB.**
TEMPLE	0.0689
MAYA	0.0150
TEMPLES	0.0112
GOD	0.0089
SHRINE	0.0084
GODS	0.0069
ANCIENT	0.0068
SITE	0.0064
SACRED	0.0060
DEITY	0.0059

TOPIC 129

WORD	**PROB.**
NBA	0.0259
SEASON	0.0237
GAME	0.0226
TEAM	0.0187
POINTS	0.0144
BASKETBALL	0.0135
GAMES	0.0118
YEAR	0.0089
PLAYED	0.0081
PLAYER	0.0074

TOPIC 132

WORD	**PROB.**
CHILDREN	0.0155
DISORDER	0.0141
MENTAL	0.0118
CHILD	0.0116
THERAPY	0.0095
SYNDROMS	0.0089
TREATMENT	0.0084
DISORDERS	0.0081
SYMPTOMS	0.0073
PATIENTS	0.0059

VII. DISCUSSIONS AND CONCLUSIONS

We note that the proposed short text message (STM) topic-based representation applied in the *tweet* contextualization task obtains good results when evaluating in terms of informativeness (table II). Furthermore, 1,174 of the 2,146 words stemming from the topic space used for the constitution of the *Indri* query, do not appear in the *tweet* vocabulary (54%). This finding demonstrates the contribution of a topic-based representation in addition to the STM lexical content. These word sub-sets are often absent from the initial content, even if they characterize the idea conveyed by the STM, as detailed in table I. Having chosen to retain the closest topics of a *tweet*, with a weight which depends on the importance of the topic knowing the STM and the word knowing the topic, results in promoting terms strongly correlated thematically. For example, a very close topic of a STM (high probability $P(z_i|t)$) will allow its vocabulary to benefit of this very strong weighting. The request q which will result from it,

[5] http://nutch.apache.org/

[6] http://lucene.apache.org/

will mainly contain terms close to its topics. The obtained results, in terms of readability (table III), take into account redundancy and other anaphora. They can be then influenced by this vocabulary thematically close to the *tweet*. The system reaches rather comparable results for the relevance of the extracted *Wikipedia* titles (table IV) compared to the ones obtained in the evaluation of the informativeness. Our system ranked between the 7^{th} and the 12^{th} position at the INEX 2012 benchmark (33 participants). The performance of our system shows that this approach allows a good higher-level representation of STMs. This task is made even more difficult as messages from *Twitter* do not use a standard vocabulary.

The main advantage of the proposed thematic approach is its direct application to a various kind of tasks (keyword extraction, document classification, ...). Moreover, none of the system parameters require an adaptation. The obtained results allow to glimpse new possibilities and perspectives. Efforts may be concentrated on several points: choosing the weight between topics and words for the scoring of a word of the topic vocabulary, or still modifying the characteristics of the index by removing grammatical words. It would also be interesting to study our system behavior by substituting the words by their lemma forms to more effectively distribute scores between thematic vocabulary (table V), or by modifying the characteristics of the topic space (number of topics composing the space, use of a corpus different from *Wikipedia* documents ...).

Acknowledgements

This work was funded by the SuMACC project of the French National Research Agency (ANR) under the program ANR-10-CORD-007.

References

[1] M. Choudhury, R. Saraf, V. Jain, S. Sarkar, and A. Basu, "Investigation and modeling of the structure of texting language," in *IJCAI-Workshop on Analytics for Noisy Unstructured Text Data*, 2007, pp. 63–70.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *ACM International Conf. on World Wide Web*, 2010, pp. 591–600.

[3] E. SanJuan, V. Moriceau, X. Tannier, P. Bellot, and J. Mothe, "Overview of the inex 2012 tweet contextualization track," in *INEX 2012 conference book*, 2012, p. 148.

[4] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval* 1," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[5] A. R. R. H. Baayen, "Aviating among the hapax legomena: Morphological grammaticalisation in current british newspaper english," *Explorations in corpus linguistics*, no. 23, p. 181, 1998.

[6] E. Frank, G. Paynter, I. Witten, C. Gutwin, and C. Nevill-Manning, "Domain-specific keyphrase extraction," in *International joint conference on artificial intelligence*, vol. 16. Citeseer, 1999, pp. 668–673.

[7] P. D. Turney, "Learning algorithms for keyphrase extraction," *Information Retrieval*, vol. 2, no. 4, pp. 303–336, 2000.

[8] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[9] J. Bellegarda, "A latent semantic analysis framework for large-span language modeling," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[10] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence, UAI ' 99*. Citeseer, 1999, p. 21.

[11] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] G. Salton, "Automatic text processing: the transformation," *Analysis and Retrieval of Information by Computer*, 1989.

[13] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.

[14] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*. Springer, 2011, pp. 338–349.

[15] T. Griffiths and M. Steyvers, "A probabilistic approach to semantic representation," in *Proceedings of the 24th annual conference of the cognitive science society*. Citeseer, 2002, pp. 381–386.

[16] B. Schiffman, K. McKeown, R. Grishman, and J. Allan, "Question answering using integrated information retrieval and information extraction," in *Proceedings of NAACL HLT*, 2007, pp. 532–539.

[17] P. Pakray, P. Bhaskar, S. Banerjee, B. Pal, S. Bandyopadhyay, and A. Gelbukh, "A hybrid question answering system based on information retrieval and answer validation," in *CLEF 2011 Workshop on QA4MRE*, 2011.

[18] C. Chen, F. Ibekwe-SanJuan, and J. Hou, "The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 7, pp. 1386–1409, 2010.

[19] M. Morchid and G. Linares, "INEX 2012 Benchmark A semantic space for tweets contextualization," in *INEX 2012*, 2012, p. 203.

[20] H. Saggion, J. Torres-Moreno, I. Cunha, and E. SanJuan, "Multilingual summarization evaluation without human models," in *International Conference on Computational Linguistics*, 2010, pp. 1059–1067.

[21] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, "Indri: A language model-based search engine for complex queries," in *International Conference on Intelligence Analysis*, 2005.

[22] R. Deveaud, F. Boudin *et al.*, "LIA/LINA at the INEX 2012 Tweet Contextualization track," *INEX'2012*, 2012.

[23] A. Bandyopadhyay, S. Pal, M. Mitra, P. Majumder, and K. Ghosh, "Passage retrieval for tweet contextualization at INEX 2012," in *INEX'2012*, 2012, p. 160.

[24] P. Bhaskar, S. Banerjee, and S. Bandyopadhyay, "A Hybrid Tweet Contextualization System using IR and Summarization," in *INEX'2012*, 2012, p. 164.

[25] L. Ermakova and J. Mothe, "IRIT at INEX 2012: Tweet Contextualization," in *INEX'2012*, 2012, p. 181.

[26] D. Ganguly, J. Leveling, and G. J. Jones, "DCU@ INEX-2012: Exploring Sentence Retrieval for Tweet Contextualization," in *INEX'2012*, 2012, p. 188.