INEX 2012 Benchmark
A semantic space for Tweet contextualization

Mohamed Morchid and Georges Linares*

Univérsite d'Avignon, Laboratoire d'Informatique d'Avignon,
339 chemin des Meinajaries, Agroparc BP 1228, 84911 Avignon cedex 9, France
{mohamed.morchid, georges.linares}@univ-avignon.fr
http://www.lia.univ-avignon.fr

Abstract. In this paper, we present a method of tweet contextualization by using a semantic space to extend the tweet vocabulary. This method is evaluated on the tweet contextualization benchmark. Contextualization is build with the sentences from English Wikipedia. The context is obtained by querying a baseline system of summary. The query is made with words from a semantic space that is estimated via a latent dirichlet allocation (LDA) algorithm. Our experiment demonstrate the effectiveness of the proposal.

Keywords: LDA, tweet, contextualization, INEX, benchmark, 2012

1 Introduction

Microblogging, provided by several services as Twitter[1] or Jaiku[2], is a new phenomenon. This form of communication enables users to broadcast their daily activities or opinions. This new communication vector, describe Internet users status in short posts disseminated in the Web. Twitter is the most popular microblogging tool. This study deals with the tweet contextualization with Wikipedia sentences. This task met two main problems: The vocabulary style and size.
Note that it is difficult to contextualize a tweet, since on at following features: a tweet has few words and the vocabulary used is quit different that the vocabulary used in Wikipedia articles.
These difficulties increase with the Web size, the dispersion and the fragmentation of the Web information. We evaluate the proposed method in the INEX2012 benchmark [2].
Different aspects of Twitter have been studied recently, as a case study [4] or as compact swap highly reactive space which can extract some descriptors of opinions or public cares [5].
We propose an approach based on the mapping of source documents in a reduced semantic space in which some words could be found by a LDA analysis

* This work was funded by the ANR project SuMACC (ANR-10-CORD-007) in CONTINT 2010 program.

[1] http://www.twitter.com

[2] http://www.jaiku.com

[1]. Other approaches like LSI/LSA [6,7] or [8] are based on statistical models that demonstrated their efficiency on various speech processing tasks. [9] uses the LSA (Latent Semantic Analysis) technique to extract the most relevant phrases from a spoken document. In [10], the authors apply LSA to an encyclopedic database for keyword extraction. We hope this method will permit to extend tweet vocabulary with others relevant words.
The remainder of the paper is organized as follows: the proposed approach is formulated in Section 2; the experimental protocol is described in Section 3; and concluding remarks are given in Section 4.

2 Tweet contextualization system

The tweet contextualization system can be decomposed as two steps. The first one is to build the query of a tweet, then, send this query to the summary system to receive the tweet context.
Concretely, the proposed method proceeds with 5 successive steps:

1. estimate off-line an LDA model on a large corpus of document D; this step produces a topic space T_{spc} of size $n^{T_{spc}}$ with a vocabulary $v^{T_{spc}}$
2. use Gibbs sampling to infer a topic distribution for a tweet t with T_{spc} to obtain a features vector V^z of the LDA classes distribution (each of these classes being implicitly associated to a topic)
3. map V^z and $v^{T_{spc}}$ to obtain a score $s(w)$ of popularity for each word w. Then, a subset S^w is composed with the words that have obtained the best score.
4. create a query q with the words of t and S^w
5. send q to the summary baseline system to receive the context c of t.

Fig. 1. Architecture of the tweet contextualization system

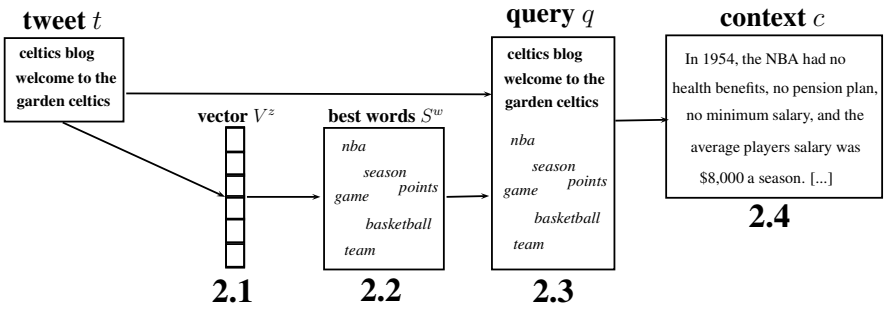


Figure 1 presents the tweet contextualization system. It can be decomposed as follows:

2.1 build a features vector V^z of a tweet by mapping t and T_{spc}
2.2 calculate the score of each word of $v^{T_{spc}}$ and extract a subset S^w of the words with best score
2.3 compose a query q with the words of t and S^w
2.4 send q to the baseline summary system and receive the context c the tweet t.

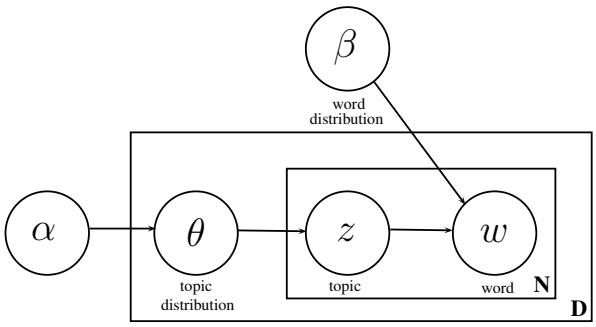
The next sections describe in-depth the main 4 parts of this process.

2.1 Features vector V^z

The Twitter language is quite unusual and sometimes constrained by the limit of the 140 characters. Using the conventional keywords, tweet query q can be affected by these features. We propose to pass through the semantic space T_{spc} from a LDA to increase the robustness of the method. Then, a features vectors V^z is calculated.The next sections describe this process.

Semantic space T_{spc}: LDA model considers a document (viewed as a *bag of words* [11]) as a probabilistic mixture of latent topics. These latent topics are characterized by a probability distribution of words associated with this topic. At the end of LDA analysis, we obtain n_{spc} classes with a set of its characteristic words and their emission probabilities.

Fig. 2. The LDA model



LDA formalism is described in Figure 2. To generate a word w in a document, a hidden topic z is sampled from a multinomial distribution defined by a vector θ of that document. Knowing z, the distribution over words is multinomial with parameters β. The parameter θ is drawn for all document from a common Dirichlet prior parameterized α. θ permit to tie the parameters between different documents. See [1] for more details.

In our experiments LDA is applied on a corpus D composed from English Wikipedia (7.8GB) of $3,691,092$ articles. This set of documents represents about 1 billion words. A semantic space of 400 topics is obtained. This number of topics is set empirically. For each LDA class, we select the 20 words with the maximum weight.

After the estimate of the background topic model T^{spc}, we have to project the tweet in this semantic space and build a features vector V^z.

Topic distribution V^z of t: We use Gibbs sampling to infer a topic distribution for the tweet t [12]. Then, a features vector V^z is obtained where the ith feature V_i^z $(i = 1, 2, \ldots, n^{T_{spc}})$ is the probability of the topic z_i knowing t:

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

2.2 Best words from vocabulary $v^{T_{spc}}$

This method allows a simple extraction of a subset S^w of the most representative words of the topic space vocabulary $v^{T_{spc}}$ knowing V^z. The system extracts $|S^w|$ (In our experiments, $|S^w| = 30$) words that obtain the highest score s. This score is the prior probability that a word can be generated by the tweet t:

$$s(w) = P(w|t) \tag{2}$$

$$= \sum_{i=1}^{n^{T_{spc}}} P(w|z_i)P(z_i|t) \tag{3}$$

$$= \sum_{i=1}^{n^{T_{spc}}} P(w|z_i)V_i^z \tag{4}$$

where $P(w|z_i)$ is the probability that the word w $(w \in v^{T_{spc}})$ was generated by the topic z_i. The score s is normalized by the highest that a word have obtained:

$$0 \leq s(w) \leq 1 \ . \tag{5}$$

Table 1 shows that the words of the tweet don't appears necessairly in S^w. That is what motivated this approach: find some others word to extend the tweet vocabulary. For example, the tweet (2) do not contain some relevant words like *army, war,muslim* or *islamic*.

2.3 Query q

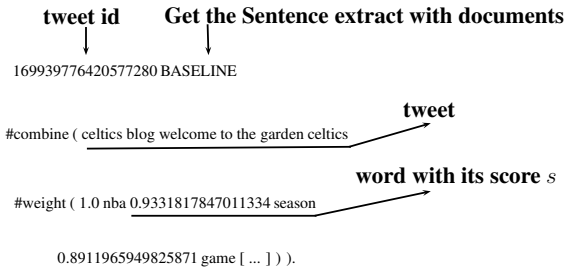
The subset S^w is used to compose the query q with the words of the tweet t. This query q is also send to the baseline XML-element retrieval system powered by Indri [13] to receive a context c of t.

Table 1. Examples of tweets with the 10 words with the best score. On **bold** some interesting words that do not appear in the tweet vocabulary.

tweets	10 best words of S^w ($\|S^w\| = 30$)
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 can find 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

The initial query is composed with the words of the tweet only. But tweets are limited by their size of 140 words and by their vocabulary. For these reasons, we extend this Indri query with the words of S^w weighted by their score s as shows in Figure 1. Figure 3 shows the different element of a query q of a tweet t. q is

Fig. 3. Example of a query q



composed by an *id*, *format* and a indri query. This query is the association of the tweet words and the S^w words weighted by their score s.

2.4 Context *c*

The query q is sent to the baseline XML-element retrieval system. The system return a context c. This context is build with the English Wikipedia sentences [2]. The index of the retrieval system covers all words (no stop list, no stemming)

and all XML tags. We query this baseline system in batch mode using the perl APIs [3].

Example of a tweet context c:

tweet t: celtics blog welcome to the garden celtics.
context c: In later life, Cousy was Commissioner of the American Soccer League from 1974 to 1979, and he has been a color analyst on Celtics telecasts since the 1980s. Today, he is a marketing consultant for the Celtics, and occasionally makes broadcast appearances with Mike Gorman and ex-Celtic teammate Tom Heinsohn. In 1954, the NBA had no health benefits, no pension plan, no minimum salary, and the average players salary was $8,000 a season. [...] 147 Boston Celtics season was the 1st season of the Boston Celtics in the Basketball Association of America (BAA/ NBA).

3 Experiments and results

1,142 tweets [2] are used for this task. Each tweets have a *i*d and at most 140 words. The first step is to create a semantic space T^{spc} with LDA. LDA need a large corpus of documents. English Wikipedia articles form this corpus. Then, the topic space T^{spc} is composed with 400 topics of 20 words.

Table 2. Results of the run.

Unigramme	Bigramme	Skip	Relevance	Syntax	Structure
0.7909	0.8920	0.8938	0.6208	0.6115	0.5145

Table 2 presents the results of the INEX 2012 benchmark. The score of unigramme, bigramme and skip are evaluated by INEX 2012 organizers. These measures do not take into account readability. The readability is the measures of relevance, syntax and structure. These evaluations are estimated on the same pool of tweets.

3.1 Conclusions

In this paper we present a method to extend tweet vocabulary. This method have been experimented in the INEX 2012 benchmark. To measure the effectiveness of our proposed method, we have to compare this results to the results of a run using just the tweet words.

Acknowledgements. We want to thanks Eric SanJuan for the baseline XML-element retrieval system.

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

References

1. D.M. Blei and A.Y. Ng and M.I. Jordan Latent dirichlet allocation The Journal of Machine Learning Research,3,993–1022,JMLR. org (2003)
2. Eric SanJuan, Véronique Moriceau, Xavier Tannier, Patrice Bellot and Josiane Mothe, Overview of the INEX 2012 Tweet Contextualization Track, Working Notes for the CLEF 2012 Workshop, Roma, Italy (to appear)
3. Clarke, F., Ekeland, I.: Nonlinear oscillations and boundary-value problems for Hamiltonian systems. Arch. Rat. Mech. Anal. 78, 315–333 (1982)
4. Z. Yang and J. Guo and K. Cai and J. Tang and J. Li and L. Zhang and Z. Su Understanding retweeting behaviors in social networks Proceedings of the 19th ACM international conference on Information and knowledge management,1633–1636, ACM(2010)
5. F. Larceneux Buzz et recommandations sur Internet: quels effets sur le box-office? Recherche et applications en marketing,45–64,JSTOR (2007)
6. S. Dumais Latent semantic indexing (LSI) and TREC-2 NIST SPECIAL PUBLICATION SP,105–105,NATIONAL INSTIUTE OF STANDARDS & TECHNOLOGY (1994)
7. J.R. Bellegarda A latent semantic analysis framework for large-span language modeling bookFifth European Conference on Speech Communication and Technology (1997)
8. T. Hofmann Probabilistic latent semantic indexing bookProceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval,50–57,ACM (1999)
9. J.R. Bellegarda Exploiting latent semantic information in statistical language modeling Proceedings of the IEEE,88,,8,1279–1296,,IEEE (2000)
10. Y. Suzuki and F. Fukumoto and Y. Sekiguchi Keyword extraction using term-domain interdependence for dictation of radio news bookProceedings of the 17th international conference on Computational linguistics-Volume 2,1272–1276,organization = "Association for Computational Linguistics (1998)
11. G. Salton Automatic text processing: the transformation Analysis and Retrieval of Information by Computer,S. l.]: Addison-Wesley Publishing Co (1989)
12. G. Casella and E.I. George Explaining the Gibbs sampler American Statistician,167–174,JSTOR (1992)
13. Strohman, T. and Metzler, D. and Turtle, H. and Croft, W.B. Indri: A language model-based search engine for complex queries Proceedings of the International Conference on Intelligent Analysis (2005)