Tweet Contextualization using Continuous Space Vectors: Automatic Summarization of Cultural Documents

Elvys Linhares Pontes*[1], Juan-Manuel Torres-Moreno[1,2], Stéphane Huet[1], and Andréa Carneiro Linhares[3]

[1] LIA, Université d'Avignon et des Pays de Vaucluse, Avignon, France
elvys.linhares-pontes@alumni.univ-avignon.fr,
{juan-manuel.torres,stephane.huet}@univ-avignon.fr
[2] École Polytechnique de Montréal, Montréal, Canada
[3] Universidade Federal do Ceará, Sobral-CE, Brasil
andreaclinhares@gmail.com

Abstract. In this paper we describe our participation in the INEX 2016 Tweet Contextualization track. The tweet contextualization process aims at generating a short summary from Wikipedia documents related to the tweet. In our approach, we analyzed tweets and created a query to retrieve the most relevant Wikipedia article. We combine Information Retrieval and Automatic Text Summarization methods to generate the tweet context.

Keywords: Text Contextualization, Automatic Text Summarization, Word Embedding, Wikipedia

1 Introduction

Twitter[4] is a social network used to diffuse news quickly using a so-called "tweet". Many newspapers and magazines use Twitter to diffuse relevant events. A tweet is composed of hashtags, usernames, words and punctuation marks. These symbols make it possible to identify Twitter's user accounts, keywords and emotions. However, a tweet is limited to 140 characters and it is complicated to describe completely an event in a single tweet. A way to overcome this problem is to get more information from another source to better explain the tweet.

Several papers concerning the tweet summarization have been developed. For example, the work of Liu et al. introduces a graph-based multi-tweet summarization system [7]. This graph integrates the functionalities of social networks, solving partially the lack of information contained in tweets. Chakrabarti and

* This work was partially financed by the French ANR project GAFES of the Université d'Avignon et des Pays de Vaucluse (France).

[4] https://twitter.com/

Punera use a Hidden Markov Model in order to enable the temporal events of sets of tweets to be modeled [3].

Inspired by the problem of tweet contextualization, the Cultural Microblog Contextualization based on Wikipedia track aims to generate short summaries which provide the background information of tweets to help users to understand them. The main idea of this task can also be forward in the French project Project "*Galerie des festivals" (Gafes)* (Gallery of Festivals), which is a collaboration between sociologists and computer scientists[5] and is carried by the Université d'Avignon (Centre Norbert Elias and Laboratoire Informatique d'Avignon). Indeed, in this paper, we contextualize a set of tweets by constructing a summary by extraction [13] guided by the "festival" mentioned in the tweet.

The summary must contain some context information about the event in order to help answering questions such as "what is this tweet about?". The context should take the form of a readable summary, not exceeding 500 words, composed of passages from the provided Wikipedia corpus. This INEX task has been described in the paper [11]. The INEX's organizers selected a set of tweets to be contextualized by the participants using the English version of Wikipedia. These tweets are collected from a set of public micro-blogs posted on Twitter and are related to the keyword "festival".

This paper is organized as follows. In Section 2 we describe our approach to contextualize the tweet. Then, we present the process of document retrieval on Wikipedia and the summarization systems in Sections 3 and 4, respectively. Finally, the conclusions are described in Section 6.

2 System Architecture

Most studies to contextualize a tweet using the Wikipedia's corpus separates this task in two parts: Information Retrieval (IR) to get the Wikipedia's documents and Automatic Text Summarization (ATS) to generate a short summary about these documents [1]. Our system is also based on these two tasks to analyze and to create the summaries (Figure 1). The first part is responsible to get the Wikipedia's document that best describes the festival mentioned in the tweet (Section 3). Initially, our system normalizes and removes the punctuation marks from each tweet to create an Indri query. Then, the Lemur system retrieves the 50 Wikipedia's documents related to the query. Finally, the system scores these documents based on the tweets and selects the document with the highest score as best description of the tweet.

The second part analyzes the selected document and creates its summary using different ATS systems (Section 4). We use the framework word2vec[6] to create the Continuous Space Vector (CSV) representation using the corpus Gigaword. Then, we create a context vocabulary of the selected document using

[5] A description for the GaFes Project is available on the website: `https://mc2.talne.eu/gafes`

[6] Site: `https://code.google.com/archive/p/word2vec/`.

the CSV representation. Finally, we use Artex and Sasi systems to summarize the selected document using the original vocabulary and the context vocabulary.

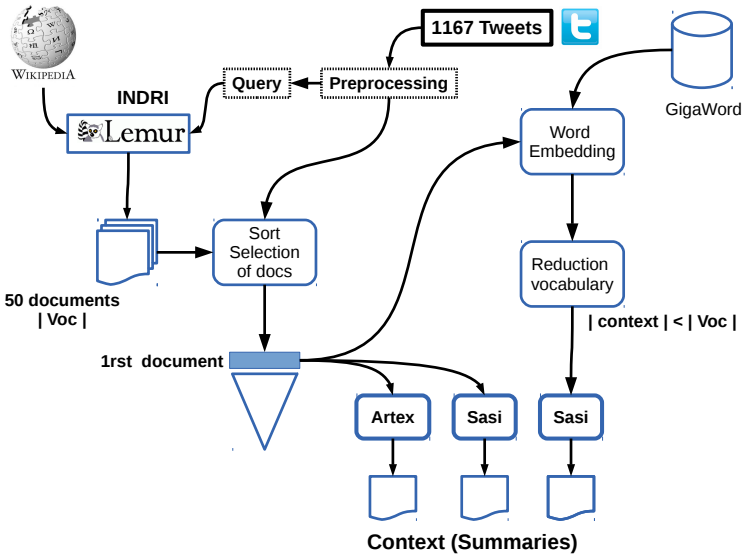


Fig. 1. Our system architecture to contextualize the tweet using the Wikipedia.

3 Wikipedia's Document Retrieval

Using the list of tweets given by the INEX's organization, we attributed different scores for the hashtags, usernames and text in the tweet. We consider the hashtags as the tweet's keywords, because they normally are names or places of cultural events. The usernames represent links to other Twitter's accounts (sometimes the festival's account) and text have few relevant words about the cultural event. Although the punctuation marks are relevant to get the semantic of the tweet, they are irrelevant to identify the festival's name. So, we remove all the punctuation marks and the stopwords.

For each tweet, we created an Indry query composed of the hashtags, the usernames and the words. Then, we used the Lemur system to find the 50 Wikipedia's documents related to this query.

As the 50 documents can have different subjects, we analyze these documents to find the document most related to the tweet. For each Wikipedia's document, we analyze the title and the text in relation to the tweet's elements (hashtag, username and word). Normally, the title of the Wikipedia's document has few words and contains the main information, while the text of the Wikipedia's

document is large and the relevance of words are small. So, we consider Equation 3 describing the score of the Wikipedia's document D based on the tweet T.

$$\text{score}_{title} = \alpha_1 \times \text{occ}(ht, title) + \alpha_2 \times \text{occ}(un, title) + \alpha_3 \times \text{occ}(nw, title) \quad (1)$$

$$\text{score}_{text} = \beta_1 \times \text{occ}(ht, text) + \beta_2 \times \text{occ}(un, text) + \beta_3 \times \text{occ}(nw, text) \quad (2)$$

$$\text{score}_{doc} = \text{score}_{title} + \text{score}_{text} \quad (3)$$

where ht are the hashtags of the tweet T, un are the usernames of the tweet T, nw are the normal words of the tweet T and $\text{occ}(ht, title)$ is the sum of occurrences of the hashtags in the title.

We analyzed a subset of tweets and we set up empirically the parameters: $\alpha_1 = 120, \alpha_2 = 80, \alpha_3 = 80, \beta_1 = 2, \beta_2 = 2, \beta_3 = 1$. For each tweet, we chose the Wikipedia's document with the biggest score to be analyzed by the ATS systems.

4 Automatic Text Summarization

The Automatic Text Summarization (ATS) systems analyze the sentences and create a short summary with the main information of the text. In order to better analyze the selected document (section above), we create two types of vocabulary (Subsection 4.1) and we use Artex (Sect. 4.2) and Sasi (Sect. 4.3) systems in order to summarize this document.

4.1 Word Representation

The word representation is very important to analyze a text. The standard word representation is an one-hot vector using a Discrete Space Vector (DSV), where each word is represented by a vector composed of zeros and only one. In this representation, all the words are independent from one another, e.g. "car", "house", "bigger" and "biggest" have different representations. In this case, we can not analyze well the sentences because we consider similar words or words with the same context as independent words.

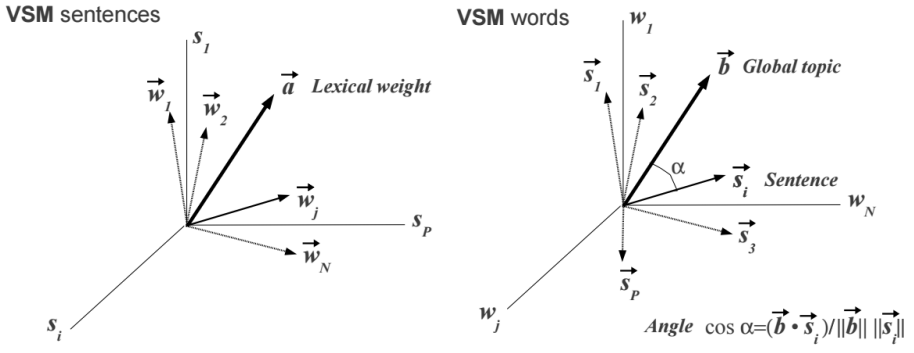
We developed a better representation to create a context vocabulary based on context of words [5]. We represent the words by the context using CSVs [9]. In this representation, two words with same context have similar representations. They devised the greedy algorithm 1 to find the similar words of word w in the texts among a pre-compiled list lcs of CSVs generated on a large corpus. If two words have a similar context, they are clustered in a same set and replaced by the most frequent word of this set. As the clusters can represent synonyms and/or words with the same idea, we can better calculate the similarity between the sentences and the metrics as Term Frequency-Inverse Document Frequency (TF-IDF).

Algorithm 1 Context vocabulary of *text*

Input: n (neighborhood size), *lcs* (list of words inside continuous space), *text*
for each word w_t in *text* **do**
  if w_t is in *lcs* **then**
    $nset \leftarrow \{w_t\}$
    $nlist \leftarrow [w_t]$
    while *nlist* is not empty **do**
      $w_l \leftarrow nlist.pop(0)$
      $nw \leftarrow$ the n nearest words of w_l in *lcs*
      $nlist.add((nw \cap$ vocabulary of $text) \setminus nset)$
      $nset \leftarrow nset \cup (nw \cap$ vocabulary of $text)$
    end while
    Replace in *text* each word of *nset* by the most frequent of *nset*
  end if
end for
Return *text*

4.2 Artex summarizer system

The Artex system [12] is an ATS system, which models a text with the sentences $s_1, s_2, \ldots, s_P$ and vocabulary size N in a Vector Space Model (VSM)[7]. Then, it calculates an average document vector that represents the average of all sentences vectors. Additionally, the system calculates the "lexical weight" for each sentence, i.e. the number of words in the sentence (Figure 2).



(a) Lexical weight (b) Global topic

Fig. 2. Artex system.

The score of sentence s_i is calculated using the proximity with the "global topic" and the "lexical weight" (Equation 4).

[7] We used the DSVs but could be expanded with CSVs.

$$\text{score}(s_i) = (\boldsymbol{s_i} \times \boldsymbol{b}) \times \boldsymbol{a} \tag{4}$$

where $\boldsymbol{s_i}$ is the vector of the sentence i, $\boldsymbol{a}$ is the average pseudo-word vector (i.e. the average number of occurrences of N words used in the sentence i) and $\boldsymbol{b}$ is the average pseudo-sentence vector (i.e. the average number of occurrences of each word j used trough the P sentences).

Finally, the summary is generated concatenating the sentences with the highest scores.

4.3 Sasi summarizer system

The Sasi system [6] is an ATS system that models the text as a graph of sentences $G = (V, E)$, where V is associated with the sentences of document (set of vertices) and E represents the similarity between two sentences (set of edges). Two sentences, which are represented by the vectors A and B, are similar if the cosine similarity between them (Equation 5) is higher than the average value of the similarity between all the sentences of the document.

$$sim(A, B) = \frac{A \times B}{||A|| \times ||B||} \tag{5}$$

From the graph G, the system calculates the independent subset[8] in order to find the most relevant non-redundant sentences. Therefore, this system creates an independent subset prioritizing the most relevant sentences based on the TF-IDF metric. Finally, the summary is composed of the most relevant sentences of the independent subset.

5 Experimental Settings and Evaluation

The set of tweets collected by INEX's organizers mentions different festivals in the world. So, it is not possible to have neither the reference summaries nor the source document about each festival. In order to evaluate the quality of the summaries, the ROUGE system [4] needs a set of reference summaries to estimate the quality of a candidate summary. In order to avoid the references, we have chosen an approach without human references [8, 10, 2] that evaluates the relevance of a candidate summary in relation to the source. In our experiments we consider the first retrieved text by the Lemur system as a "source text" for each tweet.

We use the FRESA system [10, 14] to compute the relevance of the summary based on the intersection of the n-grams between the candidate summary and the "source text". For each tweet, we generated a summary (less than 500 words) using the following systems: Artex summarizer, Sasi with the original vocabulary (Sasi_OV) and Sasi with the context vocabulary (Sasi_CV). Table 1 shows the

[8] An independent subset of a graph G is a subset of the vertices such that there is no edges between these vertices.

FRESA results about the quality of the summaries using 1-grams (FRESA-1), 2-grams (FRESA-2), skip 2-grams (FRESA-4) and their average values (FRESA-M). Results were computed using the Kullback-Leibler modified divergence [13].

Table 1. FRESA Evaluation Results.

System	FRESA-1	FRESA-2	FRESA-4	FRESA-M
Artex	0.14733	0.07701	0.07708	0.10047
Sasi_OV	0.15056	0.07679	0.07667	0.10134
Sasi_CV	0.14959	0.07665	0.07660	0.10095

From Table 1, we can not distinguish the best system, because all scores are too close. Therefore, the FRESA evaluation without references is not yet sufficient to identify the quality of the best ATS system. In fact, the first document retrieved by the IR system of INDRI may not contain the most relevant information about the festival that was mentioned in the tweet. So, establishing what is the "correct" source to evaluate a system without human references, is not a simple task. A manual evaluation is required in order to analyze correctly the source and the summaries.

6 Conclusion and Perspectives

In this paper, we presented our contributions to the INEX 2016 Tweet Contextualization Track. We considered different scores for each tweet's element to retrieve the most related Wikipedia's document with respect to a tweet. Then, we used two types of vocabularies to analyze the selected documents and to create their summary using different ATS systems. Finally, we created the summaries using two ATS systems.

In future work, summaries can be generated or can resort to a strategy to fusion multidocument sentences and preserve the grammaticality of each summary.

The evaluation using standard methods (ROUGE, FRESA,...) is probably not the most appropriate approach to measure the quality of this task of contextualization. It is possible to make a more interactive evaluation issue allowing visualization methods. We believe this evaluation, using human interaction, should correspond to a better evaluation of the results. We also want to investigate the improvement of this type of evaluation.

References

1. Bhaskar, P., Banerjee, S., Bandyopadhyay, S.: A hybrid tweet contextualization system using ir and summarization. In: INEX (2012)
2. Cabrera-Diego, L.A., Torres-Moreno, J.M., Durette, B.: Evaluating Multiple Summaries Without Human Models: A First Experiment with a Trivergent Model, pp. 91–101. Springer International Publishing, Proceedings in NLDB (2016), http://dx.doi.org/10.1007/978-3-319-41754-7_8
3. Chakrabarti, D., Punera, K.: Event Summarization using Tweets. In: 5th International Conference on Weblogs and Social Media (ICWSM). Association for the Advancement of Artificial Intelligence (2011)
4. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Moens, M.F., Szpakowicz, S. (eds.) Workshop Text Summarization Branches Out (ACL'04). pp. 74–81. ACL (2004)
5. Linhares Pontes, E., Huet, S., Torres-Moreno, J.M., Linhares, A.C.: Automatic Text Summarization with a Reduced Vocabulary Using Continuous Space Vectors, pp. 440–446. Springer International Publishing, Proceedings in NLDB (2016), http://dx.doi.org/10.1007/978-3-319-41754-7_46
6. Linhares Pontes, E., Linhares, A.C., Torres-Moreno, J.M.: Sasi: sumarizador automático de documentos baseado no problema do subconjunto independente de vértices. In: XLVI Simpósio Brasileiro de Pesquisa Operacional (2014)
7. Liu, X., Li, Y., Wei, F., Zhou, M.: Graph-Based Multi-Tweet Summarization using Social Signals. In: International Conference on Computational Linguistics (COLING'12). pp. 1699–1714 (2012)
8. Louis, A., Nenkova, A.: Automatically Assessing Machine Summary Content Without a Gold Standard. Computational Linguistics 39(2), 267–300 (2013)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. pp. 3111–3119 (2013)
10. Saggion, H., Torres-Moreno, J.M., da Cunha, I., SanJuan, E.: Multilingual summarization evaluation without human models. In: 23rd International Conference on Computational Linguistics (COLING'10). pp. 1059–1067. ACL, Beijing, China (2010)
11. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J.: Overview of the INEX 2012 Tweet Contextualization Track. In: CLEF (Online Working Notes/ Labs/ Workshop) (2012)
12. Torres-Moreno, J.M.: Artex is another text summarizer. Computing Research Repository (CoRR) (2012)
13. Torres-Moreno, J.M.: Automatic Text Summarization. John Wiley & Sons (2014)
14. Torres-Moreno, J.M., Saggion, H., da Cunha, I., SanJuan, E.: Summary Evaluation With and Without References. Polibits: Research journal on Computer science and computer engineering with applications 42, 13–19 (2010)