Microblog Contextualization using Continuous Space Vectors: Multi-Sentence Compression of Cultural Documents

Elvys Linhares Pontes*[1], Stéphane Huet[1], Juan-Manuel Torres-Moreno[1,2], 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 work for the MC2 CLEF 2017 lab. We participated in the content analysis task that involves filtering, language recognition and summarization. We combine Information Retrieval with Multi-Sentence Compression methods to contextualize microblogs using Wikipedia's pages.

Keywords: Microblog Contextualization, Multi-Sentence Compression, Word Embedding, Wikipedia

1 Introduction

Many newspapers use microblogs (Twitter, Facebook, Instagram, etc.) to disseminate news quickly. These microblogs have a limited length (e.g. a tweet is limited to 140 characters) and contain few information about an event. Therefore, it is complicated to describe an event completely in a single microblog. A way to overcome this problem is to get more information from another source to better explain the microblog.

Several studies on tweet contextualization have been done on this topic. To just name a few, Liu et al. introduced a graph-based multi-tweet summarization system [8]. This graph integrates the functionalities of social networks, solving partially the lack of information contained in tweets. Chakrabarti and Punera used a Hidden Markov Model in order to model the temporal events of sets of tweets [4]. Linhares Pontes et al. used Word Embedding to reduce the vocabulary size and to improve the results of Automatic Text Summarization (ATS) systems [7].

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

MC2 CLEF 2017 [1] lab analyzes the context and the social impact of of a microblog at large. This lab is composed of three main tasks: Content Analysis, Microblog Search and Time Line Illustration. We participated in the Content Analysis task that involves classification, filtering, language recognition, localization, entity extraction, linking open data, and summarization of Wikipedia's pages and microblogs. Specifically, we worked on the following subtasks: filtering, language recognition and automatic summarization.

The filtering subtask analyzes whether a tweet describes an existing festival or not (values are between 0 and 1, 1 for the positive case and 0 otherwise). The language recognition subtask consists in identifying the language of a microblog. Finally, the summarization task is to generate a summary (maximum of 120 words) in four languages (English, French, Portuguese and Spanish) of Wikipedia's pages describing a microblog.

This paper is organized as follows. In Section 2 we describe the architecture of our system to solve the tasks of MC2 CLEF 2017 lab. Then, we present the process of document retrieval on Wikipedia and the summarization system in Sections 3 and 4, respectively. Finally, we conclude in Section 5.

2 System Architecture

The CLEF's organizers selected a set of microblogs (tweets) with the keyword "festival" to be contextualized by the participants using four versions of Wikipedia (English, French, Portuguese, and Spanish).

For the language identification task, we pre-processed microblogs to remove all punctuation and emoticons. Then, we use the library LANGDETECT[9] to detect the languages of microblogs.

For filtering and summarization tasks, we divided our system in two parts (Fig. 1). The first part aims at retrieving the Wikipedia's pages that best describes the festival mentioned in a microblog (Section 3). We scored the Wikipedia's pages according to their relevance with respect to a microblog, which corresponds to the filtering subtask.

The second part of our system analyzes the 3 best scored pages and creates clusters of similar sentences with relevant information. Then, we use an Automatic Text Compression (ATC) system (Section 4) to compress the clusters and to generate summaries in four languages describing the festival mentioned in a microblog (summarization subtask). The algorithm 1 describes how our method analyzes the microblog, selects the 3 best Wikipedia's pages and generates the summaries.

3 Wikipedia's Document Retrieval

The set of CLEF's microblogs is composed of tweets in different languages related to festivals in all the world. Wikipedia provides a more thorough description of a given festival according to the selected language (e.g. The festival of Avignon

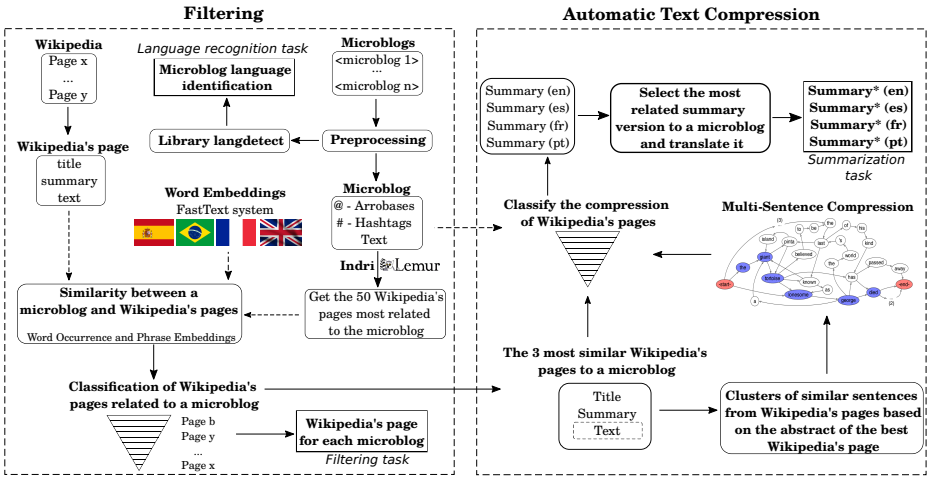

Fig. 1. Our system architecture to contextualize the microblogs.

Algorithm 1 Automatic Summarization

for each *tweet* **do**
  for *lang* **in** English, French, Portuguese and Spanish **do**
    Analyze the 3 *lang* Wikipedia's pages with the highest scores (Equation 4) using LEMUR system (*lang* version of Wikipedia)
    for each sentence of the abstract of the first Wikipedia's page (highest score) **do**
      Create the clusters of similar sentences analyzing the 3 highest scored pages
    end for
    for each cluster **do**
      Create the Word Graph (Section 4.1)
      Generate compressed sentences (Section 4.1)
    end for
    Generate the summary (*lang* language) with the compressed sentences that are the most similar to the *tweet*
  end for
  Select the best version of the summaries (most similar to the *tweet*)
  Translate the best summary version with Yandex translator to other languages
end for

is better described in the French Wikipedia). We independently analyze the four versions of Wikipedia (en, es, fr, and pt) for each microblog, repeating the whole process to first retrieve the best Wikipedia's pages and then to summarize the pages for the four versions of Wikipedia.

Our system retrieves the most related Wikipedia's pages to a microblog using a method similar to our previous work [7]. We assume that the hashtags and usernames represent the keywords of a tweet, and are independent of the language[4]. From hashtags, usernames, and the plain text (i.e. the tweet without hashtags, usernames and punctuation), we create INDRY queries to retrieve 50 Wikipedia's documents per each tweet. For each of these documents, we analyze the title and the summary 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 core information, while the summary of the document, which is made of the first paragraphs of the article before the start of the first section, is larger and provide more information[5]. Therefore, we consider Equation 4 to compute the relevance score of the Wikipedia's document D with respect to the microblog T.

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

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

$$\text{sim}(x, y) = \gamma_1 \times \text{cosine}(x, y) + \gamma_2 \times \text{occur}(x, y) \quad (3)$$

$$\text{score}_{doc} = \text{score}_{title} + \text{score}_{summary} \quad (4)$$

where ht are the hashtags of the tweet T, un the usernames of T, nw the normal words of T, and sum the summary of D. $\text{occur}(x, y)$ represents the number of occurrences of x in y, while $\text{cosine}(x, y)$ is the cosine similarity between x and y using Continuous Space Vectors[6] [3].

We set up empirically the parameters as follows: $\alpha_1 = \alpha_2 = 0.1, \alpha_3 = 0.01, \beta_1 = \beta_2 = 0.05, \beta_3 = 0.005, \gamma_1 = 1$ and $\gamma_2 = 0.5$. These coefficients give more weights to hashtags than usernames and the tweet text and compensate the shorter length of Wikipedia's article titles with respect to their summary. For each tweet, we finally keep in each language the 3 Wikipedia's documents with the highest scores to be analyzed by the ATC system.

[4] The LANGDETECT library is only used for the language recognition subtask.

[5] We did not consider the whole text of Wikipedia's page because it is sometimes huge and we preferred to rely on the work of the contributors to build the summary of the article.

[6] We used the pre-trained word embeddings (en, es, fr, and pt) of FastText system [3] that is available in `https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`.

4 Automatic Text Compression

The summary provided at the start of Wikipedia's pages is assumed to be good enough to be coherent and to provide the basic information. However, relying only of this part of the article may lead to miss relevant information about the festival that could be obtained from other sections or even pages in Wikipedia. For this reason, we preferred to use the summary of the top article as a basic abstract and to improve its quality with relevant information using Multi-Sentences Compression (MSC) (i.e. generate sentences that are shorter and more informative than the original sentences of the summary). Therefore, we consider the sentences of the summary of the best scored page as key sentences. Then for each of these sentences, we create a cluster made of the sentences of the complete 3 retrieved Wikipedia's pages which are similar; to do this, the cosine similarity is used as metrics and we empirically set up a threshold of 0.4 to consider two sentences as similar.

Then, for each cluster MSC generates a shorter and hopefully more informative compression (Section 4.1). Next, we generate the summary concatenating the most similar compression to the microblog[7].

Some language versions of Wikipedia do not have a page or they have a small description describing an specific festival. Therefore, we analyzed the summaries of each microblog obtained in the four studied languages and only retain the abstract with contains the best description of the microblog, which is estimated through the similarity between each summary and the microblog. So, we used the Yandex library[8] to translate the kept summary to others languages (en, es, fr, and pt).

4.1 Word Graph and Optimization

Our MSC system adopts the approach proposed by Filippova [5] to model a document D as a Word Graph (WG), where the vertices represent the words and arcs represent the cohesion of the words (more details in [6]). The weights of the arcs represent the level of cohesion between the words of two vertices based on the frequency and the position of these words in the sentences (Equation 5).

$$w(e_{i,j}) = \frac{\text{cohesion}(e_{i,j})}{freq(i) \times freq(j)}, \tag{5}$$

$$\text{cohesion}(e_{i,j}) = \frac{freq(i) + freq(j)}{\sum_{f \in D} dist(f,i,j)^{-1}}, \tag{6}$$

$$dist(f,i,j) = \begin{cases} pos(f,i) - pos(f,j), \text{ if } pos(f,i) < pos(f,j) \\ 0, \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{7}$$

In a previous study, we proposed to extend this approach with the analysis of the keywords and the 3-grams of the document to generate a more informative

[7] The summary is composed of the first 120 words.

[8] `https://tech.yandex.com/translate/`

MSC. Since each cluster to compress is composed of similar sentences, we consider that there is only one topic; the Latent Dirichlet Allocation (LDA) method is used to identify the keywords of this topic[2].

From the weight of 2-grams (Equation 5), the relevance of a 3-gram is based on the relevance of the two 2-grams, as described in Equation 8:

$$\text{3-gram}(i,j,k) = \frac{qt_3(i,j,k)}{\max_{a,b,c \in GP} qt_3(a,b,c)} \times \frac{w(e_{i,j}) + w(e_{j,k})}{2}, \tag{8}$$

In order to generate a better compression, the objective function expressed in Equation 9 is minimized in order to improve the informativeness and the grammaticality.

$$\text{Minimize } \Big(\alpha \sum_{(i,j)\in A} b_{i,j} \cdot x_{i,j} - \beta \sum_{k \in K} c_k \cdot w_k - \gamma \sum_{t \in T} d_k \cdot z_t\Big) \tag{9}$$

where x_{ij} indicates the existence of the arc (i,j) in the solution, $w(i,j)$ is the cohesion of the words i and j (Equation 5), z_t indicates the existence of the 3-gram t in the solution, d_t is the relevance of the 3-gram t (Equation 8), c_k indicates the existence of a word with color (keyword) k in the solution and β is the geometric average of the arc weights in the graph (more details in [6]). Finally, they calculate the 50 best solutions according to the objective (9) and we select the sentence with the lowest final score (Equation 10) as the best compression.

$$score_{norm}(f) = \frac{e^{score_{opt}(f)}}{||f||}, \tag{10}$$

where $score_{opt}(f)$ is the value of the path to generate the compression f from Equation 9. As Linhares Pontes et al. [6], we set up the parameters to $\alpha = 1.0$, $\beta = 0.9$ and $\gamma = 0.1$.

5 Conclusion

In this paper, we presented our contributions to the MC2 CLEF 2017 lab in the Content Analysis task. We considered different scores for each microblog element (hashtags, arrobases, and text) to retrieve in four languages (en, es, fr, and pt) the Wikipedia's pages most related to a microblog. Then, we generated summaries using MSC from clusters initially made of the abstract of the top retrieved article and extended with similar sentences from the 3 top retrieved articles per language. Finally, we analyzed summaries of each microblog obtained in the four languages to select the one most similar to the microblog; the kept summary is translated to other languages (en, es, fr and pt).

References

1. Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, vol. 10456 (2017)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal Machine Learning Research 3, 993–1022 (Mar 2003), http://dl.acm.org/citation.cfm?id=944919.944937
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. Chakrabarti, D., Punera, K.: Event Summarization using Tweets. In: 5th AAAI International Conference on Weblogs and Social Media (ICWSM). Association for the Advancement of Artificial Intelligence (2011)
5. Filippova, K.: Multi-sentence compression: Finding shortest paths in word graphs. In: COLING. pp. 322–330 (2010)
6. Linhares Pontes, E., da Silva, T.G., Linhares, A.C., Torres-Moreno, J.M., Huet, S.: Métodos de Otimização Combinatória Aplicados ao Problema de Compressão MultiFrases (2016)
7. Linhares Pontes, E., Torres-Moreno, J.M., Huet, S., Linhares, A.C.: Tweet contextualization using continuous space vectors: Automatic summarization of cultural documents. In: CLEF Workshop on Cultural Microblog Contextualization (2016)
8. Liu, X., Li, Y., Wei, F., Zhou, M.: Graph-Based Multi-Tweet Summarization using Social Signals. In: COLING. pp. 1699–1714 (2012)
9. Shuyo, N.: Language detection library for java (2010), http://code.google.com/p/language-detection/