Re-Ranking Microblogs Using Word2Vec in Microblog Search Task, University of Avignon

Mathias Quillot, Alexandre Delorme

Re-Ranking Microblogs Using Word2Vec in Microblog Search Task, University of Avignon

Mathias Quillot and Alexandre Delorme

LIA, University of Avignon, France
{firstname.lastname}@univ-avignon.fr

Abstract. Working note aimed at proposing improved system from Indri Search Engine for the *Microblog Search Task* of *MC2 CLEF 2017 lab*. This improvement is tried thanks to Word2Vec model used in re-classing results in comparing them with query.

Keywords: Word Embedding, Word2Vec, Indri, Index, Search, Microblog, Cultural

1 Introduction

The goal of this working note is to introduce a Microblog Search System for the MC2 task *Microblog Search*. The task consists in searching for the 64 most relevant microblogs in a collection covering 18 months of news about festivals in all languages. The given queries can be *Microblog Search Task* in Arabic, English, French and Spanish.

The performance of the system will be evaluated by the organizers of the conference. In the meantime, we worked with sociologists on the performance evaluation of the system. Our goal was to estimate if a microblog is relevant to a given micro-critic in French, based on the register (a linguistic function) of the microblog.

This working note is organized as follows. Section 2 2.1 presents the system we modeled using *Word2Vec* and with which we generated the run for the microblog search task. We then introduce the evaluation protocol established with the sociologists. Finally, we conclude and describe the next step emerging from the experiment.

2 Proposed Approach

The goal of the search task is, given a topic, to find the 64 most relevant microblogs. To achieve this, we assume that: 1) relevant microblogs are the semantically closest to the topic ; 2) *Word2Vec* has the property of additivity that allows us to represent a given microblog by adding all its word vector representations ; and 3) cosine similarity give us an appropriate score of similarity between two *Word2Vec* vector representation. Thanks to these hypotheses, we are able

to compare microblogs by their *Word2Vec* representation, and build a list of microblogs which are ranked by their similarity with respect to the microblog reference.

2.1 Model

The figure 1 shows how our system works. 1) Firstly, the topic (our microblog reference) is given to the Indri Search Engine, which looks for a subset of corresponding microblogs. Thanks to *Word2Vec*, we then represent each microblog and the input topic as a vector, by adding the vectors of each word found from their phrase (the content for a microblog and itself for the topic). 2) The system computes the cosine similarity between each microblog vector representation and the input topic. Finally, it re-ranks the microblogs by order of descending cosine similarity with the topic.

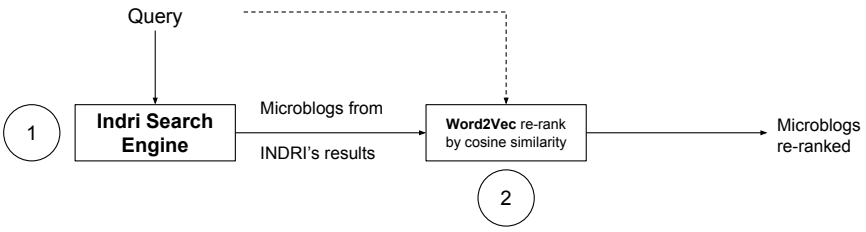

Fig. 1. Microblog Search System with Indri Search Engine and *Word2Vec*

Indri [1] is a search engine made from the *Lemur* project between the University of Massachusetts and Carnegie Mellon University, a collaboration for which the goal is to build language modeling information retrieval tools.

Word2Vec Neural Network *Word2Vec* models [3] are based on the hypothesis that semantically similar words tend to have similar contextual distributions. Concretely, this context is a window whose size is expressed in words, and which is centered on the word of interest. In this work, we use the Continuous Bag of Words (CBOW) learning method that seeks to predict the reference word given its context. The neural network model takes as input the context w_{i-2}, w_{i-1}, w_{i+1} and w_{i+2}, while it outputs the reference word w_i. We only use the hidden layer of the neural networks, which means each word is represented by a vector. The length of this vector is specified by the user as a parameter d, and the method therefore outputs an $N \times d$ matrix. More information about *Word2Vec* models may be found in [3].

[1] https://www.lemurproject.org/indri/

2.2 Experimental Setup

Now we have a methodology, we will explain how we obtain our system and how we configured our tools and softwares. This will be done by describing our data and then explain how we built our *Word2Vec* model.

Data We use the corpus provided for the *MC2 CLEF 2017 lab*, which contains 70 million tweets. It covers a period from May 2015 to November 2016, and contains tweets in 134 different languages.

Word2Vec model We firstly tried to use Gensim, which provides a lot of methods implemented as *Latent Semantic Indexing* or *Word2Vec*, but we encounterdc technical difficulties arising from the size of the corpus. Indeed,Th *Word2vec* creates a matrix of $m \times m$ where m is the size of the vocabulary. In the case of our corpus, m is very significant, so the software uses a lot of memory. Gensim does not support this kind of problem where Tmikolov's Word2Vec tool seems to solve. The former tries to take more and more *RAM* even if it is full where the second one just uses your *RAM* as possible without trying to get more and to ask swap memory space. Since we had this problem, Gensim seams to have fixed this problem.

We finally decided to build the *Word2Vec* model from the microblog data thanks to Tmikolov's Word2Vec software[2]. This one worked very perfectly. We use the following configuration:

- cbow (use of CBOW) : 1 ;
- size (desired vector dimensionality) : 200 ;
- windows (the size of the context windows) : 5 ;
- negative (negative sampling) : 25 ;
- hs (hierarchical softmax) : 0 ;
- sample : 10^-4 ;
- iter : 15.

3 Evaluation Proposal

The MC2 organizers will later evaluate the results generated from our system. In the meantime, we began a reflexion with sociologists, aiming at defining an evaluation protocol of our system. In order to highlight the necessary conditions to consider microblog as well-classified, the sociologist has analyzed the results of the French baseline system given by the MC2 organizers. We then have imagined a protocol supervised by human, to evaluate the performances of a microblog search engine considering a micro-critic given as reference (input of the system).

[2] `https://github.com/tmikolov/word2vec`

3.1 Baseline French Analysis

We used the baseline french system to generate a list of microblogs for each topic and then we analyze these results to define which microblogs are well classed and which are not.

What we highlight is that some microblogs don't share the same register. Some have sarcasm, ironical or second level of interpretation. So, even the subject seems to be the same, the register could derive from one microblog to the other. For instance, for a given micro-critic on Vodcaster which concern the best film of the *Festival de Cannes*, one of the microblogs automatically returned by the system was about a movie which was never selected at the festival but cover a subject about the festival.

This led us to refer rank the different micro-critics and microblogs by register. To attempt this goal, we referred to Jakobson and his six language functions (expressive, phatic, conative, metalinguistic, poetic, referential). [2]. Let us give more example of the illustration of Jakobson's language functions on the baseline system.

Some microblogs have subtleties that only sophisticated reader are able to understand. For example "Regarder par la fentre et se dire: tiens il pleut !!? " L'impression d'tre dans un film slectionner pour le festival de Cannes, to read this microblog, reader needs to have a priori knowledge of what look like a selected movie at can and to be in capacity to represent himself "can" and understand properly the meaning of the sentence. In this example, this microblog could correspond to the poetic and expressive functions because the rain is associated by a figure of speech at the Cannes Festival.

Let's take an example of microblog with mainly phatic register. C'est vrai c'est le dbut du Festival de Cannes aujourd'hui. Cest vrai serve to have attention, and the information in the message it is like a personal reminder that an information for all. It seems to remind a presence more than give an information to an audience. Maybe the kind of reaction expected is me to I have forgot the day of begining or This year, I think the festival like an open question very large.

The type of message that have more chance to be linked with another micrcritic or microblog its -ma claque inattendue du festival c'est jessica93 Because Its expressive and also referential. It is an advice that could help. Also this type of microblog, more referential and informational could be usefull Bebel de retour au Festival Lumire sous une standing ovation. L'motion Because the advise is related to an information the standing ovation. Thereby, we can have a advise and a tangible information that can help to get our own opinion.

There are three types of micro-critics register that cannot be related easily with another one. The poetic one, the conative one and the metalinguitic one. Because all this functions require a knowledge about the context of production of the tweet. The conative implies to know the receiver. The mtalinguistic implies to know well the subject (festival or movie) and can compare all editions of a festival for example. And the poetic function implies to know well the subject to understand the link between the word game or figure of speech and what it mean in relation to the referential subject.

Sometimes, even the same term than discover dont recover the same meaning. Because we can expected that a microblog which speak about a discover on a festival of cinema will make reference to a movie or something related, but the festivals gather also personality of TV, of music for example and people speak about them in relation to the festival. Some microblogs recalls a dicover to speak about this celebrities without relation to the cinema.

3.2 Protocol Proposal

We have seen that a microblog could be characterized by its linguistic function. So, we propose to build on this observation an evaluation protocol. This latter one is supervised by humans and consists in annotating microblogs depending on whether they have the same register that the referenced input topic.

As the figure 2 illustrates, the first step consists in searching the n corresponding microblogs to the given topic. In the case of the MC2's task, $n = 64$. In the second step, an expert annotates each microblog to define if its register corresponds with that of the topic. Finaly, we compare the expert's annotation with the ranked list given by the system.

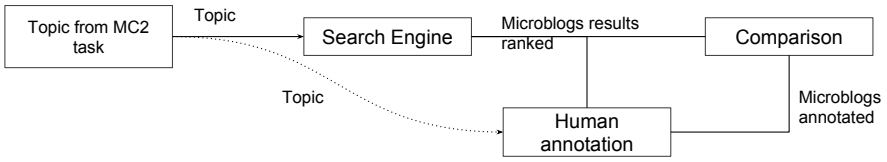

Fig. 2. Evaluation system for Microblog Search Engine

The comparison between the expert's annotation and the ranked list produced by the system is performed by calculating the precision of the m first microblogs of the list. The precision is expressed as: $g \div m$, where g is the number of microblogs from the m first chosen microblogs of the list which are annotated as having the same register than the topic. In this approach, we propose to vary m between 1 and n and then analyze this variability to choose the best m value or to interpret the performance of the system.

4 Conclusion

In this preliminary work, our objective was to make a proposal of search engine and of evaluating microblogs searching. For the former, we proposed a model based on *Word2Vec* to re-rank microblogs. For the latter, sociologists manually assessed the results and defined which microblogs are well classified using the baseline system and what discriminate thereof. Our system will be evaluated by

the MC2 task organizers. We will then evaluate it thanks to our evaluation proposal and analyze our results by evaluating their performances and comparing the results between the two evaluation systems. Moreover, this proposal may correspond with our *Word2Vec* training method which considers stopwords. These words are usually removed in search task, but their information is important on our register based approach. [1] Finally, we will propose to other candidate to test our evaluation protocol et check if our assumptions are confirmed.

References

1. Castro, D., Adame, Y., Pelaez, M., Munoz, R.: Authorship verification, combining linguistic features and different similarity functions. In: Conference and Labs Evaluation Forum – PAN labs (2015)
2. Jakobson, R.: Closing statements: Linguistics and poetics. In: Sebeok, T.A. (ed.) Style In Language, pp. 350–377. MIT press (1960)
3. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv:1301.3781 (2013)