UNED-READERS: Filtering Relevant Tweets using Probabilistic Signature Models

Henry Anaya-Sánchez, Anselmo Peñas, and Bernardo Cabaleiro

IR & NLP Group, UNED,
Juan del Rosal 16, 28040 Madrid, Spain
{henry.anaya,anselmo,bcabaleiro}@lsi.uned.es
http://nlp.uned.es/

Abstract. This paper describes the (usupervised) knowledge-based approach to filter relevant tweets for a given entity that is followed by the UNED-READERS system at RepLab 2013. The approach relies on a new way of contextualizing entity names from relative large and broad collections of texts using probabilistic signature models (i.e., discrete probability distributions of words lexically related to the knowledge or topic underlying set of entities in background text collections). The contextualization is intended to recover relevant information about the entity (specifically, lexically related words) from background knowledge. Results obtained in the filtering task are presented.

1 Introduction

In the last few years, Twitter has become an increasingly popular platform for users to communicate with each other by sharing their latest updates in the form of short messages called tweets. [1] Because tweets are compact and fast, Twitter has become widely used to spread and share many kinds of information, such as breaking news, personal updates, spontaneous ideas, or opinions about brands, services, celebrities, etc. [6].

This phenomenon has attracted the attention of information analysts, who see in tweet datasets a valuable input source to carry out a plethora of studies. The large and ever increasing volume of available tweets obliges to make use of automatic methods and tools to interpret and analyze them. Nevertheless, tweets are particularly challenging for being interpreted and analized automatically, since they are shorts piece of texts (i.e., 140 characters) usually posted using a nonstandard language with similarities to SMS style [3, 4].

Currently, a major problem concerning the analysis of tweets is that of deciding whether a given tweet containing the name of an entity is actually refering or not to the entity. This is because entity names are often ambiguous. For instance, a tweet containing the word "Ford" might refer to the motor company, but also might be about the Ford Models (the modeling agency), Tom Ford (the film director), etc. [5].

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

In this paper, we address the above problem by proposing a methodology to contextualize entity names with probabilistic models from relative large and broad text collections, so that a model of these entities can be applied to decide whether a tweet contains or not a reference to a given entity. By "contextualize" we refer to the process of recovering relevant information about the entity (specifically, lexically related words) from background knowledge.

2 Contextualizing Entities with Probabilistic Signature Models

Given a collection of text documents $\mathcal{C} = \{d_1, \ldots, d_N\}$ with vocabulary $\mathcal{V} = \{w_1, \ldots, w_M\}$, we propose to contextualize a set of entity names $q = \{w_{i_1}, \ldots, w_{i_n}\}$ in the same domain (e.g., automotive, education, music/artists) according to $\mathcal{C}$ ($\forall j \in \{1, \ldots, n\}$, we assume that $w_{i_j} \in \mathcal{V}$) by means of a probabilistic signature model; that is, a discrete probability distribution of words lexically related to the knowledge or topic underlying the set of entities in the collection.

To define such a probabilistic signature model for the text element q (hereafter referred to as the signature model for q), we rely on a posterior probability distribution of words $\{\mathrm{p}(w|q)\}_{w\in\mathcal{V}}$ that is defined from the following equation:

$$\mathrm{p}(w|q) = \mathrm{p}(q|w)\,\mathrm{p}(w) \tag{1}$$

where $\mathrm{p}(w)$ represents a prior for word w in $\mathcal{C}$, and $\mathrm{p}(q|w)$ is a conditional probability estimated from a statistical (stochastic) mapping between words $\tau = \{\mathrm{p}(w_i|w_j)\}_{1\leq i\leq M, 1\leq j\leq M}$ ($\forall j \in \{1, \ldots, M\}$, $\sum_{i=1}^{M} \mathrm{p}(w_i|w_j) = 1$) that is based on word co-occurrences from local contexts in $\mathcal{C}$, as in [2].

The aim is to base the signature model on the likelihood with which words in the vocabulary $\mathcal{V}$ are mapped to (or entail) words in q.

Nevertheless, as the posterior $\{\mathrm{p}(w|q)\}_{w\in V}$ can actually assign high probability values to either meaningless words (such as prepositions, conjunctions, common verbs, etc.) or words with a relative high prior or frequency in $\mathcal{C}$ (for example, ambiguous words with different interpretations in the collection), we propose to learn the signature model that contextualizes q from a refined version of the posterior distribution of words $\{\mathrm{p}(w|q)\}_{w\in V}$, namely, $\{\mathrm{p}_q(w)\}_{w\in V}$, that is obtained by minimizing the following cross entropy value:

$$H_q = -\sum_{w\in V} \mathrm{p}(w|q) \log\big((1-\lambda)\,\mathrm{p}_q(w) + \lambda\,\mathrm{p}(w)\big) \tag{2}$$

where λ is a mixture weight that accounts for the proportion of "background noise" in $\{\mathrm{p}(w|q)\}_{w\in V}$ and $\{\mathrm{p}(w)\}_{w\in V}$ is a background probability distribution of words (e.g., the distribution of words in $\mathcal{C}$).

Let z represents the topic underlying q in the context of $\mathcal{C}$. Then, we define the signature model of q as the discrete distribution:

$$s_q(w) \propto \mathrm{p}_q(w) \tag{3}$$

where w ranges over the set of words $\mathcal{V}_q$ ($\mathcal{V}_q \subseteq \mathcal{V}$) such that a word w' is in $\mathcal{V}_q$ if both:

(i) w' is among the set of words with the higher probablity values under p_q, and
(ii) the probability value of including w' in z (i.e., $\mathrm{p}(z|w') = (1-\lambda)\,\mathrm{p}_q(w')/((1-\lambda)\,\mathrm{p}_q(w')+\lambda\,\mathrm{p}(w'))$) is greater than a threshold (for example, the probability of including w' in complement of z that can be defined as $\mathrm{p}(\bar{z}|w') = 1 - \mathrm{p}(z|w')$).

To estimate the refined that minimize $H_{s,q}$, we rely on an Expectation Maximization procedure that starting from initial values for $\{\mathrm{p}_q(w)\}_{w\in\mathcal{V}}$, namely $\{\mathrm{p}_q^{(0)}(w)\}_w$, it iteratively approximates the values of $\{\mathrm{p}_s(w)\}_w$ for each $w \in \mathcal{V}$ until convergence by means of the following updates in the kth iteration:

$$\mathrm{p}_q^{(k)}(w) = \frac{\mathrm{p}(w|q)Z_w}{\sum_{w'} \mathrm{p}(w'|q)Z_{w'}} \tag{4}$$

where Z_w is:

$$Z_w = \frac{(1-\lambda)\,\mathrm{p}_q^{(k-1)}(w)}{(1-\lambda)\,\mathrm{p}_q^{(k-1)}(w) + \lambda\,\mathrm{p}(w)} \tag{5}$$

3 RepLab 2013 Results

From the above formulation of signature models, we devise an (unsupervised) knowledge-based system for filtering tweets as it was asked by the Replab 2013 filtering task [1].

The RepLab dataset comprises about 105,000 tweets in the test set, each one containing the canonical name of at least one entity in a selected set of 61 entities from four domains: automotive, banking, universities and music/artists. The source languages of the tweets in this dataset include English and Spanish.

The idea underlying our system is to consider –as positive evidence– the overlapping between the vocabularies of the tweets and the vocabulary of the topic signature models obtained for the four different sets of entity names that comprise each domain of entities in Replab 2013 (i.e., automotive, banking, universities and music/artists).

The signature models were obtained by contextualizing each set of entity names according to webtext (a collection of about 1,700,000 text documents in English).

We have submitted the following four runs:

(1) The system considers only **tweets in English**, and regards that the input tweet is related to the given entity if it contains **at least 2 words** in the signature obtained for the corresponding set of domain entities.
(2) The system considers only **tweets in English**, and regards that the input tweet is related to the given entity if it contains **at least one word** in the signature obtained for the corresponding set of domain entities **after removing from the tweet the name of the entity**.

(3) The system considers **tweets in any language** and regards that the input tweet is related to the given entity if it contains **at least 2 words** in the signature obtained for the corresponding set of domain entities.
(4) The system considers only **tweets in any language** and regards that the input tweet is related to the given entity if it contains **at least one word** in the signature obtained for the corresponding set of domain entities **after removing from the tweet the name of the entity**.

Since signature models wer obtained from a collection of texts in English, the aim of runs 3 and 4 is to measure the quality of signatures build from an English corpus to filter tweets in Spanish.

We present the obtained results in terms of the following measures: Precision, Recall, F1(Precision,Recall), Reliability, Sensitivity and F1(Reliability,Sensitivity). The last three of these measures were considered as the official mesaures of RepLab 2013 [1].

Tables 1 and 2 sumarize the results obtained by our four runs. As it can be appreciated, our run relative good values of Precision, Recall and F1 if we take into account that our signature models are base only in English texts. The fact that recall values are lower than those of precision can be explained since the proposed contextualization is base on a completely different collection of documents and in this way the learned models focus on modeling domain aspects rather than typical aspects observed in tweets such as irony and sarcasm-based aspects related to entities.

It is worth mentioning that our runs are completely unsupervised, and do not use the training data provided to this task at all.

Results obtained by runs 1 and 2 can not be compared to those ones obtained by runs 3 and 4 because they are applied to different input data.

Table 1. Overall values of Precision, Recall and F1(Precision,Recall) obtained by the different runs of our system.

Run id.	Precision	Recall	F1(Precision,Recall)
1	0.8946	0.5627	0.6909
2	0.9112	0.5195	0.6618
3	0.8915	0.7039	0.7866
4	0.9038	0.6438	0.7520
1(vs. test in Eng.)	0.8946	0.7249	0.8009
2(vs. test in Eng.)	0.9112	0.6693	0.7717

Interestingly, when regarding the official RepLab 2013 measure of Reliability, which is a precision-based measure, it can be seen that the results obtained by runs 3 and 4 are smaller than those obtained by runs 1 and 2 even when the values of Precision of runs 1 to 4 are similar. Besides, these values of Reliability are smaller than those of Sensitivity, which is a recall-based measure, while Recall values are always smaller than Precision values.

Table 2. Averaged values of Reliability, Sentitivity and F1(Reliability,Sensitivity) obtained by the different runs of our system.

Run id.	Reliability	Sensitivity	F1(Reliability,Sensitivity)
1	0.3734	0.3162	0.2677
2	0.3824	0.3305	0.2838
3	0.1894	0.4191	0.2071
4	0.1955	0.4339	0.2211

4 Conclusions

In this paper, we have described the UNED-READERS system that participated in the filtering task at RepLab 2013. This knowledge-based system relies on a new way of contextualizing entity names from relative large and broad collections of texts using probabilistic signature models. The learned contexts can be shown useful to model domain vocabulary if we consider the obtained values of Precision and Recall. The submitted system is completely unsupervised. It does not use the training set provided by the competition.

Acknowledgments. This work has been partially supported by the Spanish Ministry of Science and Innovation, through the project Holopedia (TIN2010-21128-C02), and by the Spanish Government (MINECO) in the framework of CHIST-ERA program (READERS project).

References

1. Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Fourth International Conference of the CLEF initiative, CLEF 2013, Valencia, Spain. Proceedings*, Springer LNCS, 2013.
2. Lisette García-Moya, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori. A language model approach for retrieving product features and opinions from customer reviews. *IEEE Intelligent Systems*, 99(PrePrints):1, 2013.
3. Max Kaufmann and Jugal Kalita. Syntactic normalization of twitter messages. In *International Conference on Natural Language Processing*, Kharagpur, India, 2010.
4. Gustavo Laboreiro, Luís Sarmento, Jorge Teixeira, and Eugénio Oliveira. Tokenizing micro-blogging messages using a text classification approach. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 81–88. ACM, 2010.
5. Damiano Spina, Julio Gonzalo, and Enrique Amigó. Discovering filter keywords for company name disambiguation in twitter. *Expert Systems with Applications*, 40:4986–5003, 2013.
6. Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.