CAp 2017 challenge: Twitter Named Entity Recognition

Abstract

The paper describes the CAp 2017 challenge. The challenge concerns the problem of Named Entity Recognition (NER) for tweets written in French. We first present the data preparation steps we followed for constructing the dataset released in the framework of the challenge. We begin by demonstrating why NER for tweets is a challenging problem especially when the number of entities increases. We detail the annotation process and the necessary decisions we made. We provide statistics on the inter-annotator agreement, and we conclude the data description part with examples and statistics for the data. We, then, describe the participation in the challenge, where 8 teams participated, with a focus on the methods employed by the challenge participants and the scores achieved in terms of F$_1$ measure. Importantly, the constructed dataset comprising $\sim$6,000 tweets annotated for 13 types of entities, which to the best of our knowledge is the first such dataset in French, is publicly available at \url{http://cap2017.imag.fr/competition.html} .

Introduction

The proliferation of the online social media has lately resulted in the democratization of online content sharing. Among other media, Twitter is very popular for research and application purposes due to its scale, representativeness and ease of public access to its content. However, tweets, that are short messages of up to 140 characters, pose several challenges to traditional Natural Language Processing (NLP) systems due to the creative use of characters and punctuation symbols, abbreviations ans slung language.

Named Entity Recognition (NER) is a fundamental step for most of the information extraction pipelines.

Importantly, the terse and difficult text style of tweets presents serious challenges to NER systems, which are usually trained using more formal text sources such as newswire articles or Wikipedia entries that follow particular morpho-syntactic rules. As a result, off-the-self tools trained on such data perform poorly BIBREF0 . The problem becomes more intense as the number of entities to be identified increases, moving from the traditional setting of very few entities (persons, organization, time, location) to problems with more. Furthermore, most of the resources (e.g., software tools) and benchmarks for NER are for text written in English. As the multilingual content online increases, and English may not be anymore the lingua franca of the Web. Therefore, having resources and benchmarks in other languages is crucial for enabling information access worldwide.

In this paper, we propose a new benchmark for the problem of NER for tweets written in French. The tweets were collected using the publicly available Twitter API and annotated with 13 types of entities. The annotators were native speakers of French and had previous experience in the task of NER. Overall, the generated datasets consists of INLINEFORM0 tweets, split in training and test parts.

The paper is organized in two parts. In the first, we discuss the data preparation steps (collection, annotation) and we describe the proposed dataset. The dataset was first released in the framework of the CAp 2017 challenge, where 8 systems participated. Following, the second part of the paper presents an overview of baseline systems and the approaches employed by the systems that participated. We conclude with a discussion of the performance of Twitter NER systems and remarks for future work.

Challenge Description

In this section we describe the steps taken during the organisation of the challenge. We begin by introducing the general guidelines for participation and then proceed to the description of the dataset.

Guidelines for Participation

The CAp 2017 challenge concerns the problem of NER for tweets written in French. A significant milestone while organizing the challenge was the creation of a suitable benchmark. While one may be able to find Twitter datasets for NER in English, to the best of our knowledge, this is the first resource for Twitter NER in French. Following this observation, our expectations for developing the novel benchmark are twofold: first, we hope that it will further stimulate the research efforts for French NER with a focus on in user-generated text social media. Second, as its size is comparable with datasets previously released for English NER we expect it to become a reference dataset for the community.

The task of NER decouples as follows: given a text span like a tweet, one needs to identify contiguous words within the span that correspond to entities. Given, for instance, a tweet “Les Parisiens supportent PSG ;-)” one needs to identify that the abbreviation “PSG” refers to an entity, namely the football team “Paris Saint-Germain”. Therefore, there two main challenges in the problem. First one needs to identify the boundaries of an entity (in the example PSG is a single word entity), and then to predict the type of the entity. In the CAp 2017 challenge one needs to identify among 13 types of entities: person, musicartist, organisation, geoloc, product, transportLine, media, sportsteam, event, tvshow, movie, facility, other in a given tweets. Importantly, we do not allow the entities to be hierarchical, that is contiguous words belong to an entity as a whole and a single entity type is associated per word. It is also to be noted that some of the tweets may not contain entities and therefore systems should not be biased towards predicting one or more entities for each tweet.

Lastly, in order to enable participants from various research domains to participate, we allowed the use of any external data or resources. On one hand, this choice would enable the participation of teams who would develop systems using the provided data or teams with previously developed systems capable of setting the state-of-the-art performance. On the other hand, our goal was to motivate approaches that

would apply transfer learning or domain adaptation techniques on already existing systems to adapt them for the task of NER for French tweets.

The Released Dataset

For the purposes of the CAp 2017 challenge we constructed a dataset for NER of French tweets. Overall, the dataset comprises 6,685 annotated tweets with the 13 types of entities presented in the previous section. The data were released in two parts: first, a training part was released for development purposes (dubbed “Training” hereafter). Then, to evaluate the performance of the developed systems a “Test” dataset was released that consists of 3,685 tweets. For compatibility with previous research, the data were released tokenized using the CoNLL format and the BIO encoding.

To collect the tweets that were used to construct the dataset we relied on the Twitter streaming API. The API makes available a part of Twitter flow and one may use particular keywords to filter the results. In order to collect tweets written in French and obtain a sample that would be unbiased towards particular types of entities we used common French words like articles, pronouns, and prepositions: “le”,“la”,“de”,“il”,“elle”, etc.. In total, we collected 10,000 unique tweets from September 1st until September the 15th of 2016.

Complementary to the collection of tweets using the Twitter API, we used 886 tweets provided by the “Société Nationale des Chemins de fer Français” (SNCF), that is the French National Railway Corporation. The latter subset is biased towards information in the interest of the corporation such as train lines or names of train stations. To account for the different distribution of entities in the tweets collected by SNCF we incorporated them in the data as follows:

For the training set, which comprises 3,000 tweets, we used 2,557 tweets collected using the API and

443 tweets of those provided by SNCF.

For the test set, which comprises 3,685 consists we used 3,242 tweets from those collected using the API and the remaining 443 tweets from those provided by SNCF.

Annotation

In the framework of the challenge, we were required to first identify the entities occurring in the dataset and, then, annotate them with of the 13 possible types. Table TABREF12 provides a description for each type of entity that we made available both to the annotators and to the participants of the challenge.

Mentions (strings starting with @) and hashtags (strings starting with #) have a particular function in tweets. The former is used to refer to persons while the latter to indicate keywords. Therefore, in the annotation process we treated them using the following protocol: A hashtag or a mention should be annotated as an entity if:

For a hashtag or a mention to be annotated both conditions are to be met. Figure FIGREF16 elaborates on that:

We measure the inter-annotator agreement between the annotators based on the Cohen's Kappa (cf. Table TABREF15) calculated on the first 200 tweets of the training set. According to BIBREF1 our score for Cohen's Kappa (0,70) indicates a strong agreement.

In the example given in Figure FIGREF20 :

[name=M1, matrix of nodes, row sep=10pt, column sep=3pt,ampersand replacement=&] schema)

[text=black] Il; & schema-spezifisch) [text=black] rejoint; & nutzerinfo) [text=frenchrose] Pierre; & host) [text=frenchrose] Fabre;

query) [text=black] comme; & fragment) [text=black] directeur; & text=black] des; & text=black] marques;

ducray) [text=magenta] Ducray; & text=black] et; & a) [text=magenta] A;& text=magenta] -; & derma) [text=magenta] Derma;

; [overbrace style] (nutzerinfo.north west) – (host.north east) node [overbrace text style,rectangle,draw,color=white,rounded corners,inner sep=4pt, fill=frenchrose] Group; [underbrace style] (ducray.south west) – (ducray.south east) node [underbrace text style,rectangle,draw=black,color=white,rounded corners,inner sep=4pt, fill=magenta] Brand; [underbrace style] (a.south west) – (derma.south east) node [underbrace text style,rectangle,draw,color=white,rounded corners,inner sep=4pt, fill=magenta] Brand;

A given entity must be annotated with one label. The annotator must therefore choose the most relevant category according to the semantics of the message. We can therefore find in the dataset an entity annotated with different labels. For instance, Facebook can be categorized as a media (“notre page Facebook") as well as an organization (“Facebook acquires acquiert Nascent Objects").

Event-named entities must include the type of the event. For example, colloque (colloquium) must be annotated in “le colloque du Réveil français est rejoint par".

Abbreviations must be annotated. For example, LMP is the abbreviation of “Le Meilleur Patissier" which is a tvshow.

As shown in Figure 1, the training and the test set have a similar distribution in terms of named entity types. The training set contains 2,902 entities among 1,656 unique entities (i.e. 57,1%). The test set contains 3,660 entities among 2,264 unique entities (i.e. 61,8%). Only 15,7% of named entities are in both datasets (i.e. 307 named entities). Finally we notice that less than 2% of seen entities are ambiguous on the testset.

Description of the Systems

Overall, the results of 8 systems were submitted for evaluation. Among them, 7 submitted a paper discussing their implementation details. The participants proposed a variety of approaches principally using Deep Neural Networks (DNN) and Conditional Random Fields (CRF). In the rest of the section we provide a short overview for the approaches used by each system and discuss the achieved scores.

Submission 1 BIBREF2 The system relies on a recurrent neural network (RNN). More precisely, a bi-directional GRU network is used and a CRF layer is adde on top of the network to improve label prediction given information from the context of a word, that is the previous and next tags.

Submission 2 BIBREF3 The system follows a state-of-the-art approach by using a CRF for to tag sentences with NER tags. The authors develop a set of features divided into six families (orthographic, morphosyntactic, lexical, syntactic, polysemic traits, and language-modeling traits).

Submission 3 BIBREF4 , ranked first, employ CRF as a learning model. In the feature engineering process they use morphosyntactic features, distributional ones as well as word clusters based on these learned representations.

Submission 4 BIBREF5 The system also relies on a CRF classifier operating on features extracted for

each word of the tweet such as POS tags etc. In addition, they employ an existing patterntn mining NER system (mXS) which is not trained for tweets. The addition of the system's results in improving the recall at the expense of precision.

Submission 5 BIBREF6 The authors propose a bidirectional LSTM neural network architecture embedding words, capitalization features and character embeddings learned with convolutional neural networks. This basic model is extended through a transfer learning approach in order to leverage English tweets and thus overcome data sparsity issues.

Submission 6 BIBREF7 The approach proposed here used adaptations for tailoring a generic NER system in the context of tweets. Specifically, the system is based on CRF and relies on features provided by context, POS tags, and lexicon. Training has been done using CAP data but also ESTER2 and DECODA available data. Among possible combinations, the best one used CAP data only and largely relied on a priori data.

Submission 7 Lastly, BIBREF8 uses a rule based system which performs several linguistic analysis like morphological and syntactic as well as the extraction of relations. The dictionaries used by the system was augmented with new entities from the Web. Finally, linguistics rules were applied in order to tag the detected entities.

Results

Table TABREF22 presents the ranking of the systems with respect to their F1-score as well as the precision and recall scores.

The approach proposed by BIBREF4 topped the ranking showing how a standard CRF approach can

benefit from high quality features. On the other hand, the second best approach does not require heavy feature engineering as it relies on DNNs BIBREF2 .

We also observe that the majority of the systems obtained good scores in terms of F1-score while having important differences in precision and recall. For example, the Lattice team achieved the highest precision score.

Conclusion

In this paper we presented the challenge on French Twitter Named Entity Recognition. A large corpus of around 6,000 tweets were manyally annotated for the purposes of training and evaluation. To the best of our knowledge this is the first corpus in French for NER in short and noisy texts. A total of 8 teams participated in the competition, employing a variety of state-of-the-art approaches. The evaluation of the systems helped us to reveal the strong points and the weaknesses of these approaches and to suggest potential future directions.