Named Entity Recognition in Twitter using Images and Text

Abstract

Named Entity Recognition (NER) is an important subtask of information extraction that seeks to locate and recognise named entities. Despite recent achievements, we still face limitations with correctly detecting and classifying entities, prominently in short and noisy text, such as Twitter. An important negative aspect in most of NER approaches is the high dependency on hand-crafted features and domain-specific knowledge, necessary to achieve state-of-the-art results. Thus, devising models to deal with such linguistically complex contexts is still challenging. In this paper, we propose a novel multi-level architecture that does not rely on any specific linguistic resource or encoded rule. Unlike traditional approaches, we use features extracted from images and text to classify named entities. Experimental tests against state-of-the-art NER for Twitter on the Ritter dataset present competitive results (0.59 F-measure), indicating that this approach may lead towards better NER models.

Introduction

Named Entity Recognition (NER) is an important step in most of the natural language processing (NLP) pipelines. It is designed to robustly handle proper names, which is essential for many applications. Although a seemingly simple task, it faces a number of challenges in noisy datasets and it is still considered an emerging research area BIBREF0 , BIBREF1 . Despite recent efforts, we still face limitations at identifying entities and (consequently) correctly classifying them. Current state-of-the-art NER systems typically have about 85-90% accuracy on news text - such as articles (CoNLL03 shared task data set) - but they still perform poorly (about 30-50% accuracy) on short texts, which do not have implicit linguistic formalism (e.g. punctuation, spelling, spacing, formatting, unorthodox capitalisation, emoticons, abbreviations and hashtags) BIBREF2 , BIBREF3 , BIBREF4 , BIBREF1 . Furthermore, the lack of external knowledge resources is an important gap in the process regardless of writing style BIBREF5 . To face these problems, research has been focusing on microblog-specific information extraction techniques BIBREF2 , BIBREF6 .

In this paper, we propose a joint clustering architecture that aims at minimizing the current gap between world knowledge and knowledge available in open domain knowledge bases (e.g., Freebase) for NER systems, by extracting features from unstructured data sources. To this aim, we use images and text from the web as input data. Thus, instead of relying on encoded information and manually annotated resources (the major limitation in NER architectures) we focus on a multi-level approach for discovering named entities, combining text and image features with a final classifier based on a decision tree model. We follow an intuitive and simple idea: some types of images are more related to people (e.g. faces) whereas some others are more related to organisations (e.g. logos), for instance. This principle is applied similarly to the text retrieved from websites: keywords for search engines representing names and surnames of people will often return similarly related texts, for instance. Thus, we derive some indicators (detailed in sec:finalclassifier which are then used as input features in a final classifier.

To the best of our knowledge, this is the first report of a NER architecture which aims to provide a priori information based on clusters of images and text features.

Related Work

Over the past few years, the problem of recognizing named entities in natural language texts has been addressed by several approaches and frameworks BIBREF7 , BIBREF8 . Existing approaches basically adopt look-up strategies and use standard local features, such as part-of-speech tags, previous and next words, substrings, shapes and regex expressions, for instance. The main drawback is the performance of those models with noisy data, such as Tweets. A major reason is that they rely heavily on hand-crafted features and domain-specific knowledge. In terms of architecture, NER algorithms may also be designed based on generative (e.g., Naive Bayes) or discriminative (e.g., MaxEnt) models. Furthermore, sequence models (HMMs, CMM, MEMM and CRF) are a natural choice to design such systems. A more recent study proposed by Lample et al., 2016 BIBREF9 used neural architectures to solve this problem. Similarly in terms of architecture, Al-Rfou et al., 2015 BIBREF10 had also proposed a model (without dependency) that learns distributed word representations (word embeddings) which encode semantic and syntactic features of words in each language. Chiu and Nichols, 2015 BIBREF11 proposed a neural network architecture that automatically detects word and character-level features using a hybrid bidirectional LSTM and CNN. Thus, these models work without resorting to any language-specific knowledge or resources such as gazetteers. They, however, focused on newswire to improve current state-of-the-art systems and not on the microblogs context, in which they are naturally harder to outperform due to the aforementioned issues. According to Derczynski et al., 2015 BIBREF1 some approaches have been proposed for Twitter, but they are mostly still in development and often not freely available.

Conceptual Architecture

The main insight underlying this work is that we can produce a NER model which performs similarly to state-of-the-art approaches but without relying on any specific resource or encoded rule. To this aim, we propose a multi-level architecture which intends to produce biased indicators to a certain class (LOC, PER or ORG). These outcomes are then used as input features for our final classifier. We perform clustering on images and texts associated to a given term INLINEFORM0 existing in complete or partial sentences INLINEFORM1 (e.g., “new york” or “einstein”), leveraging the global context obtained from the Web providing valuable insights apart from standard local features and hand-coded information. fig:architecture gives an overview of the proposed architecture.

In the first step (A), we simply apply POS Tagging and Shallow Parsing to filter out tokens except for those tagged as INLINEFORM0 or INLINEFORM1 and their INLINEFORM2 (local context). Afterwards, we use the search engine (B) to query and cache (C) the top INLINEFORM3 texts and images associated to each term INLINEFORM4 , where INLINEFORM5 is the set resulting of the pre-processing step (A) for each (partial or complete) sentence INLINEFORM6 . This resulting data (composed of excerpts of texts and images from web pages) is used to predict a possible class for a given term. These outcomes are then used in the first two levels (D.1 and D.2) of our approach: the Computer Vision and Text Analytics components, respectively, which we introduce as follows:

Computer Vision (CV): Detecting Objects: Function Description (D.1): given a set of images INLINEFORM0 , the basic idea behind this component is to detect a specific object (denoted by a class INLINEFORM1) in each image. Thus, we query the web for a given term INLINEFORM2 and then extract the features from each image and try to detect a specific object (e.g., logos for ORG) for the top INLINEFORM3 images retrieved as source candidates. The mapping between objects and NER classes is detailed in tab:tbempirical.

Training (D.1): we used SIFT (Scale Invariant Feature Transform) features BIBREF12 for extracting image descriptors and BoF (Bag of Features) BIBREF13 , BIBREF14 for clustering the histograms of extracted features. The clustering is possible by constructing a large vocabulary of many visual words and representing each image as a histogram of the frequency words that are in the image. We use k-means BIBREF15 to cluster the set of descriptors to INLINEFORM0 clusters. The resulting clusters are compact and separated by similar characteristics. An empirical analysis shows that some image groups are often related to certain named entities (NE) classes when using search engines, as described in tab:tbempirical. For training purposes, we used the Scene 13 dataset BIBREF16 to train our classifiers for location (LOC), “faces” from Caltech 101 Object Categories BIBREF17 to train our person (PER) and logos from METU dataset BIBREF18 for organisation ORG object detection. These datasets produces the training data for our set of supervised classifiers (1 for ORG, 1 for PER and 10 for LOC). We trained our classifiers using Support Vector Machines BIBREF19 once they generalize reasonably enough for the task.

Text Analytics (TA): Text Classification - Function Description (D.2): analogously to (D.1), we perform clustering to group texts together that are “distributively” similar. Thus, for each retrieved web page (title and excerpt of its content), we perform the classification based on the main NER classes. We extracted features using a classical sparse vectorizer (Term frequency-Inverse document frequency - TF-IDF. In experiments, we did not find a significant performance gain using HashingVectorizer) - Training (D.2): with this objective in mind, we trained classifiers that rely on a bag-of-words technique. We collected data using DBpedia instances to create our training dataset (INLINEFORM0) and annotated each instance with the respective MUC classes, i.e. PER, ORG and LOC. Listing shows an example of a query to obtain documents of organizations (ORG class). Thereafter, we used this annotated dataset to train our model.

where INLINEFORM0 and INLINEFORM1 represent the INLINEFORM2 and INLINEFORM3 position of INLINEFORM4 and INLINEFORM5 , respectively. INLINEFORM6 represents the n-gram of POS tag. INLINEFORM7 and INLINEFORM8 (INLINEFORM9) represent the total objects found by a classifier INLINEFORM10 for a given class INLINEFORM11 (INLINEFORM12) (where N is the total of retrieved images INLINEFORM15). INLINEFORM16 and INLINEFORM17 represent the distance between the two higher predictions (INLINEFORM18), i.e. INLINEFORM19 . Finally, INLINEFORM20 represents the sum of all predictions made by all INLINEFORM21 classifiers INLINEFORM22 (INLINEFORM23). - Training (E): the outcomes of D.1 and D.2 (INLINEFORM26) are used as input features to our final classifier. We implemented a simple Decision Tree (non-parametric supervised learning method) algorithm for learning simple decision rules inferred from the data features (since it does not require any assumptions of linearity in the data and also works well with outliers, which are expected to be found more often in a noisy environment, such as the Web of Documents).

Experiments

In order to check the overall performance of the proposed technique, we ran our algorithm without any further rule or apriori knowledge using a gold standard for NER in microblogs (Ritter dataset BIBREF2), achieving INLINEFORM0 F1. tab:performance details the performance measures per class. tab:relatedwork presents current state-of-the-art results for the same dataset. The best model achieves INLINEFORM1 F1-measure, but uses encoded rules. Models which are not rule-based, achieve INLINEFORM2 and INLINEFORM3 . We argue that in combination with existing techniques (such as linguistic patterns), we can potentially achieve even better results.

As an example, the sentence “paris hilton was once the toast of the town” can show the potential of the proposed approach. The token “paris” with a LOC bias (0.6) and “hilton” (global brand of hotels and resorts) with indicators leading to LOC (0.7) or ORG (0.1, less likely though). Furthermore, “town” being correctly biased to LOC (0.7). The algorithm also suggests that the compound “paris hilton” is more likely to be a PER instead (0.7) and updates (correctly) the previous predictions. As a downside in this example, the algorithm misclassified “toast” as LOC. However, in this same example, Stanford NER annotates (mistakenly) only “paris” as LOC. It is worth noting also the ability of the algorithm to take advantage of search engine capabilities. When searching for “miCRs0ft”, the returned values strongly indicate a bias for ORG, as expected (INLINEFORM0 = 0.2, INLINEFORM1 = 0.8, INLINEFORM2 = 0.0, INLINEFORM3 = 6, INLINEFORM4 = -56, INLINEFORM5 = 0.0, INLINEFORM6 = 0.5, INLINEFORM7 = 0.0, INLINEFORM8 = 5). More local organisations are also recognized correctly, such as “kaufland” (German supermarket), which returns the following metadata: INLINEFORM9 = 0.2, INLINEFORM10 = 0.4, INLINEFORM11 = 0.0, INLINEFORM12 = 2, INLINEFORM13 = -50, INLINEFORM14 = 0.1, INLINEFORM15 = 0.4, INLINEFORM16 = 0.0, INLINEFORM17 = 3.

Discussion

A disadvantage when using web search engines is that they are not open and free. This can be circumvented by indexing and searching on other large sources of information, such as Common Crawl and Flickr. However, maintaining a large source of images would be an issue, e.g. the Flickr dataset may not be comprehensive enough (i.e. tokens may not return results). This will be a subject of future work. Besides, an important step in the pre-processing is the classification of part-of-speech tags. In the Ritter dataset our current error propagation is 0.09 (107 tokens which should be classified as NOUN) using NLTK 3.0. Despite good performance (91% accuracy), we plan to benchmark this component. In terms of processing time, the bottleneck of the current implementation is the time required to extract features from images, as expected. Currently we achieve a performance of 3~5 seconds per sentence and plan to also optimize this component. The major advantages of this approach are: 1) the fact that there are no hand-crafted rules encoded; 2) the ability to handle misspelled words (because the search engine alleviates that and returns relevant or related information) and incomplete sentences; 3) the generic design of its components, allowing multilingual processing with little effort (the only dependency is the POS tagger) and straightforward extension to support more NER classes (requiring a corpus of images and text associated to each desired NER class, which can be obtained from a Knowledge Base, such as DBpedia, and an image dataset, such as METU dataset). While initial results in a gold standard dataset showed the potential of the approach, we also plan to integrate these outcomes into a Sequence Labeling (SL) system, including neural architectures such as LSTM, which are more suitable for such tasks as NER or POS. We argue that this can potentially reduce the existing (significant) gap in NER performance on microblogs.

Conclusions

In this paper we presented a novel architecture for NER that expands the feature set space based on feature clustering of images and texts, focused on microblogs. Due to their terse nature, such noisy data often lack enough context, which poses a challenge to the correct identification of named entities. To address this issue we have presented and evaluated a novel approach using the Ritter dataset, showing consistent results over state-of-the-art models without using any external resource or encoded rule, achieving an average of 0.59 F1. The results slightly outperformed state-of-the-art models which do not rely on encoded rules (0.49 and 0.54 F1), suggesting the viability of using the produced metadata to also boost existing NER approaches. A further important contribution is the ability to handle single tokens and misspelled words successfully, which is of utmost importance in order to better understand short texts. Finally, the architecture of the approach and its indicators introduce potential to transparently support multilingual data, which is the subject of ongoing investigation.

Acknowledgments

This research was supported in part by an EU H2020 grant provided for the HOBBIT project (GA no. 688227) and CAPES Foundation (BEX 10179135).