Named Entity Recognition in Twitter using Images and Text

Diego Esteves[1], Rafael Peres[2], Jens Lehmann[1,3], and Giulio Napolitano[1,3]

[1] University of Bonn, Germany
`{surname}@cs.uni-bonn.de,`
[2] Federal University of Rio de Janeiro, Brazil
`rafaelperes@ufrj.br`
[3] Fraunhofer IAIS, Bonn, Germany
`giulio.napolitano@iais.fraunhofer.de`

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.

Keywords: NER, Short Texts, Noisy Data, Machine Learning, Computer Vision

1 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 [2, 5]. 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) [5, 10, 12, 19]. Furthermore, the lack of external knowledge resources is an important gap in the process regardless of writing style [18]. To face these problems, research has been focusing on microblog-specific information extraction techniques [19, 22].

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 Section 3 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.

2 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 [15, 20]. 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 [11] used neural architectures to solve this problem. Similarly in terms of architecture, Al-Rfou et al., 2015 [1] 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 [4] 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 [5] some approaches have been proposed for Twitter, but they are mostly still in development and often not freely available.

3 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* t existing in complete or partial sentences S (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. Figure 1 gives an overview of the proposed architecture.
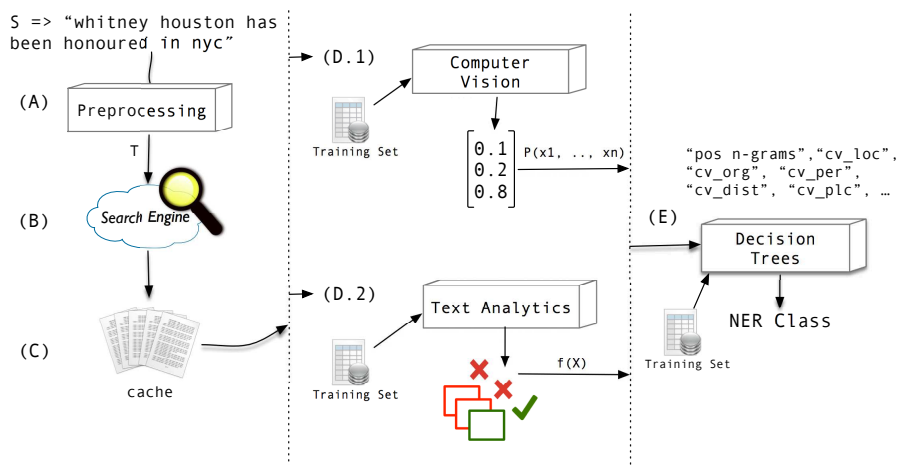

Fig. 1: Overview of the approach: combining computer vision and machine learning in a generic NER architecture

In the first step `(A)`, we simply apply *POS Tagging* and *Shallow Parsing* to filter out tokens except for those[4] tagged as *propn* or *nouns* and their *compounds* (*local context*). Afterwards, we use the search engine `(B)` to query and cache `(C)` the top N texts and images associated to each term $t \in T$, where T is the set resulting of the pre-processing step `(A)` for each (partial or complete) sentence S. 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 $\mathcal{I}$, the basic idea behind this component is to detect a specific object (denoted by a class c) in each image. Thus, we query the web for a given term t and then extract the features from each image and try to detect a specific object (e.g., logos for *ORG*) for the top N images[5] retrieved as source candidates. The mapping between objects and NER classes is detailed in Table 1. `Training (D.1):` we used

---

[4] State-of-the-art POS tagging systems still do not have exceptional performance in short texts.

[5] We set $N = 10$ in our experiments and used Microsoft Bing as the search engine.

NER	Images Candidates (number of trained models)
LOC	Building, Suburb, Street, City, Country, Mountain, Highway, Forest, Coast and Map (10)
ORG	Company Logo (1)
PER	Human Face (1)

Table 1: NER classes and respective objects to be detected in a given image. For LOC we trained more models due to the diversity of the object candidates.

SIFT (Scale Invariant Feature Transform) features [13] for extracting image descriptors and BoF (Bag of Features) [17, 21] 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* [14] to cluster the set of descriptors to k 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 Table 1. For training purposes, we used the *Scene 13* dataset [8] to train our classifiers for *location* (LOC), "faces" from *Caltech 101 Object Categories* [7] to train our *person* (PER) and logos from *METU dataset* [16] 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 [9] once they generalize reasonably enough for the task[6].

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 ($N = 15000$) and annotated each instance with the respective MUC classes, i.e. *PER*, *ORG* and *LOC*. Listing 1.1 shows an example of a query to obtain documents of organizations (ORG class). Thereafter, we used this annotated dataset to train our model.

```
SELECT ?location, ?abstract FROM <http://dbpedia.org>
WHERE {?location rdf:type dbo:Location .
       ?location dbo:abstract ?abstract .
FILTER (lang(?abstract) = 'en')} LIMIT 15000
```

Listing 1.1: SPARQL: an example of querying DBPedia to obtain LOC data for training

Final Classifier (E) - `Function Description (E):` we use the outcomes of (`D.1` and `D.2`) as part of the input to our final model. The final set of indicators is

---

[6] *scikit-learn*: svm.NuSVC(nu=0.5, kernel='rbf', gamma=0.1, probability=True).

defined as follows: let W_s be a set of *tokens* existing in a given sentence $s \in S$. We extract the POS tag (using Stanford POS Tagger) for each token w and filter out any token classified other than *PROP-NOUN* and existing *compounds* as entity candidates ($t \in S'$). The result is a simple structure:

$$M_i = \{j, t, ng_{pos}, C_{loc}, C_{per}, C_{org}, C_{dist}, C_{plc}, T_{loc}, T_{per}, T_{org}, T_{dist}\} \qquad (1)$$

where i and j represent the ith and j^{th} position of $s \in S$ and $w \in W_s$, respectively. ng_{pos} represents the n-gram[7] of POS tag. C_k and T_k ($k \in \{loc, per, org\}$) represent the total objects found by a classifier Φ for a given class k ($\sum_{n=1}^{N} \Phi(k, img_n)$)[8] (where N is the total of retrieved images $\mathcal{I}$). C_{dist} and T_{dist} represent the distance between the two higher predictions ($\mathcal{P} = \{C_k \forall K\}$), i.e. $\max(\mathcal{P}) - \max(\mathcal{P}')|\mathcal{P}' = \mathcal{P} - \{\max(\mathcal{P})\}$. Finally, C_{plc} represents the sum of all predictions made by all *LOC* classifiers $\mathcal{CL}$ ($\sum_{l=1}^{L} \sum_{n=1}^{N} \mathcal{CL}_l(loc, img_n)$)[9]. - `Training (E)`: the outcomes of `D.1` and `D.2` (M) are used as input features to our final classifier. We implemented a simple *Decision Tree*[10] (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).

4 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 [19]), achieving 0.59 F1. Table 2 details the performance measures per class. Table 3 presents current state-of-the-art results for the same dataset. The best model achieves 0.8 F1-measure, but uses encoded rules. Models which are not rule-based, achieve 0.49 and 0.56. We argue that in combination with existing techniques (such as linguistic patterns), we can potentially achieve even better results. As

NER Class	Precision	Recall	F-measure
Person (PER)	0.86	0.53	0.66
Location (LOC)	0.70	0.40	0.51
Organisation (ORG)	0.90	0.46	0.61
None	0.99	1.0	0.99
Average (PLO)	0.82	0.46	**0.59**

Table 2: Performance measure for our approach in Ritter dataset: 4-fold cross validation

---

[7] *bigram*, in our experiments.

[8] $pos = +1$, $neg = -1$.

[9] $pos = +1$, $neg = 0$.

[10] *scikit-learn*: criterion='entropy', splitter='best'.

NER System	Description	Precision	Recall	F-measure
Ritter et al., 2011 [19]	LabeledLDA-Freebase	0.73	0.49	0.59
Bontcheva et al., 2013 [3]	Gazetteer/JAPE	0.77	0.83	0.80
Bontcheva et al., 2013 [3]	Stanford-twitter	0.54	0.45	0.49
Etter et al., 2013 [6]	SVM-HMM	0.65	**0.49**	0.54
our approach	Cluster (images and texts) + DT	**0.82**	0.46	**0.59**

Table 3: Performance measures (PER, ORG and LOC classes) of state-of-the-art NER for short texts (Ritter dataset). Approaches which do not rely on hand-crafted rules and *Gazetteers* are highlighted in gray. Etter et al., 2013 trained using 10 classes.

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 (C_{loc} = 0.2, C_{org} = 0.8, C_{per} = 0.0, C_{dist} = 6, C_{plc} = -56, T_{loc} = 0.0, T_{org} = 0.5, T_{per} = 0.0, T_{dist} = 5). More local organisations are also recognized correctly, such as "*kaufland*" (German supermarket), which returns the following metadata: C_{loc} = 0.2, C_{org} = 0.4, C_{per} = 0.0, C_{dist} = 2, C_{plc} = -50, T_{loc} = 0.1, T_{org} = 0.4, T_{per} = 0.0, T_{dist} = 3.

5 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[11]. 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)

---

[11] http://commoncrawl.org/ and https://www.flickr.com/

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.

6   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

References

1. AL-RFOU, R., KULKARNI, V., PEROZZI, B., AND SKIENA, S. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada* (2015), SIAM.
2. BASAVE, A. E. C., VARGA, A., ROWE, M., STANKOVIC, M., AND DADZIE, A.-S. Making sense of microposts (#msm2013) concept extraction challenge. A. E. Cano, M. Rowe, M. Stankovic, and A.-S. Dadzie, Eds., vol. 1019 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 1–15.
3. BONTCHEVA, K., DERCZYNSKI, L., FUNK, A., GREENWOOD, M. A., MAYNARD, D., AND ASWANI, N. Twitie: An open-source information extraction pipeline for microblog text. In *RANLP* (2013), pp. 83–90.
4. CHIU, J. P., AND NICHOLS, E. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308* (2015).

5. DERCZYNSKI, L., MAYNARD, D., RIZZO, G., VAN ERP, M., GORRELL, G., TRONCY, R., PETRAK, J., AND BONTCHEVA, K. Analysis of named entity recognition and linking for tweets. *Information Processing & Management 51*, 2 (2015), 32–49.

6. ETTER, D., FERRARO, F., COTTERELL, R., BUZEK, O., AND VAN DURME, B. Nerit: Named entity recognition for informal text. *The Johns Hopkins University, the Human Language Technology Center of Excellence, HLTCOE 810Wyman Park Drive Baltimore, Maryland 21211, Tech. Rep.* (2013).

7. FEI-FEI, L., FERGUS, R., AND PERONA, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding 106*, 1 (2007), 59–70.

8. FEI-FEI, L., AND PERONA, P. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (2005), vol. 2, IEEE, pp. 524–531.

9. FLETCHER, T. Support vector machines explained. *Online]. http://sutikno. blog. undip. ac. id/files/2011/11/SVM-Explained. pdf.[Accessed 06 06 2013]* (2009).

10. GATTANI, A., LAMBA, D. S., GARERA, N., TIWARI, M., CHAI, X., DAS, S., SUBRAMANIAM, S., RAJARAMAN, A., HARINARAYAN, V., AND DOAN, A. Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. *Proc. VLDB Endow. 6*, 11 (Aug. 2013), 1126–1137.

11. LAMPLE, G., BALLESTEROS, M., SUBRAMANIAN, S., KAWAKAMI, K., AND DYER, C. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).

12. LIU, X., ZHOU, M., WEI, F., FU, Z., AND ZHOU, X. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (2012), Association for Computational Linguistics, pp. 526–535.

13. LOWE, D. G. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on* (1999), vol. 2, pp. 1150–1157 vol.2.

14. MACQUEEN, J., ET AL. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (1967), vol. 1, Oakland, CA, USA., pp. 281–297.

15. NADEAU, D., AND SEKINE, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes 30*, 1 (2007), 3–26.

16. O. TURSUN, S. K. A challenging big dataset for benchmarking trademark retrieval. In *IAPR Conference on Machine Vision and Applications* (2015).

17. PHILBIN, J., CHUM, O., ISARD, M., SIVIC, J., AND ZISSERMAN, A. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (2007), IEEE, pp. 1–8.

18. RATINOV, L., AND ROTH, D. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (2009), Association for Computational Linguistics, pp. 147–155.

19. RITTER, A., CLARK, S., ETZIONI, O., ET AL. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), Association for Computational Linguistics, pp. 1524–1534.

20. ROBERTS, A., GAIZAUSKAS, R. J., HEPPLE, M., AND GUO, Y. Combining terminology resources and statistical methods for entity recognition: an evaluation. In *LREC* (2008).

21. SIVIC, J., AND ZISSERMAN, A. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (2003), IEEE, pp. 1470–1477.

22. VAN ERP, M., RIZZO, G., AND TRONCY, R. Learning with the web: Spotting named entities on the intersection of nerd and machine learning. In *# MSM* (2013), pp. 27–30.