Weakly supervised named entity classification

Edouard Grave

Weakly supervised named entity classification

Edouard Grave
EECS Department
UC Berkeley
grave@berkeley.edu

Abstract

In this paper, we describe a new method for the problem of named entity classification for specialized or technical domains, using distant supervision. Our approach relies on a simple observation: in some specialized domains, named entities are almost unambiguous. Thus, given a seed list of names of entities, it is cheap and easy to obtain positive examples from unlabeled texts using a simple string match. Those positive examples can then be used to train a named entity classifier, by using the PU learning paradigm, which is learning from positive and unlabeled examples. We introduce a new convex formulation to solve this problem, and apply our technique in order to extract named entities from financial reports corresponding to healthcare companies.

1 Introduction

We are interested in extracting named entities from specialized texts, such as financial reports. Most state-of-the-art approaches to named entity recognition are based on supervised machine learning. In the case of specialized or technical domains, these methods suffer from several limitations. First, they rely on labeled data, which are often scarce. Thus, for many specialized domains, practitioners face a lack of labeled data. Second, these methods suffer from domain shift. Even for extracting named entities belonging to general categories, such as *person*, *organization* or *location*, a NER system trained on news articles will make mistakes that could be easily avoided by using knowledge from the domain. For example, when extracting named entities from financial reports corresponding to healthcare companies, an open domain NER system systematically labels *Henry Schein* as a person and *Aspen* as a location, while both are healthcare companies.

On the other hand, obtaining a seed list of named entities that one wants to extract from a given specialized domain is usually cheap and easy. Moreover, we argue that for many domains, mentions of those entities are almost not ambiguous. Thus, using a simple string match between the seed list of named entities and unlabeled text from the specialized domain, it is easy to obtain positive examples of named entity mentions. Then, we propose to train a named entity classifier from those positive and unlabeled examples, a problem referred to as *PU learning* (Liu et al., 2002, 2003). Our approach is also related to *distant supervision*, which has been successfully applied to relation extraction (Craven and Kumlien, 1999; Mintz et al., 2009).

Contributions. In this paper, we make the following contributions:

- We propose to frame the problem of learning to extract named entities using distant supervision as a PU learning problem (learning from positive and unlabeled examples only);
- We introduce a new method to solve this learning problem, based on a convex formulation;
- We apply our proposed method on a dataset of financial reports (10-Q) corresponding to healthcare companies, in order to automatically extract names of companies and drugs.

1.1 Related work

Learning to extract named entities from a small list of seeds and unlabeled texts has already been considered in the past. Riloff and Jones (1999) introduced a method called *mutual bootstrapping*: given a small list of entities, a set of context in which those entities appear can be extracted from unlabeled text. In turn, this set of contextual rules can be used to extract new named entities. This procedure is then applied iteratively, in order to grow the sets of entities and rules. Collins and Singer (1999) proposed a closely related method, based on *co-training*, in which the two views of the data correspond to *spelling* rules and *contextual* rules. Similar kinds of techniques have been proposed to extract relations (Brin, 1999; Agichtein and Gravano, 2000). One of the limitations of bootstrapping algorithms is semantic drift, which happens when labeling errors accumulate during the iterative learning process (Curran et al., 2007). Finally, bootstrapping has been successfully applied to Web scale information extraction (Etzioni et al., 2005; Carlson et al., 2010).

Another line of research strongly related to our approach is *distant supervision* for information extraction. Distant supervision refers to a set of methods that use a knowledge base, instead of labeled examples, to learn to extract information. Different knowledge bases have been proposed to learn to extract relations, such as the Yeast Protein Database (Craven and Kumlien, 1999), BibTex entries (Bellare and McCallum, 2007), Wikipedia infoboxes (Wu and Weld, 2007) or Freebase (Mintz et al., 2009). In the context of named entity recognition, Talukdar and Pereira (2010) proposed a graph-based semisupervised learning method, Ritter et al. (2011) introduced a technique based on LabeledLDA and Ritter et al. (2013) described a method based on an undirected graphical model.

The method described in this paper is strongly influenced by the framework of discriminative clustering (Xu et al., 2004). The goal of discriminative clustering is to assign labels to points, such that learning a discriminative classifier using those labels would lead to low empirical risk. Different loss functions have been considered, such as the hinge loss (Xu et al., 2004), the squared loss (Bach and Harchaoui, 2007) or the logistic loss (Joulin et al., 2010). Discriminative clustering has recently been applied to weakly supervised face recognition (Bojanowski et al., 2013; Ramanathan et al., 2014) and weakly supervised relation extraction (Grave, 2014).

2 Description of our approach

In this section we describe our approach to the problem of named entity classification using distant supervision. First, we extract named entity mentions, for example by extracting sequences of contiguous tokens with the part-of-speech NNP or NNPS[1]. Second, we try to match each of these mentions to a named entity from our seed list, using an exact string match. Thus, we obtain a set of positive and unlabeled examples, from which we train a multiclass classifier. It must be noted that contrary to classical semi-supervised problems, in our case, the labeled set contains *only positive examples* and no negative examples. This problem is an instance of PU learning (Liu et al., 2003). In the following, we describe a new approach to learn a multi-class classifier from positive and unlabeled examples, inspired by discriminative clustering.

2.1 Notations

We start by setting up some notations. We suppose that we have N named entity mentions represented by the feature vectors $(\mathbf{x}_n)_{1 \leq n \leq N}$. Let $\mathcal{P}$ be the set of indices of positive examples and $\mathcal{U}$ be the set of indices of unlabeled examples. For each positive example $n \in \mathcal{P}$, we denote by the integer $c_n \in \{1, ..., K\}$ its corresponding label. The integer $K + 1$ will represent the class OTHER.

Our goal is to jointly infer a binary matrix $\mathbf{Y} \in \{0, 1\}^{N \times (K+1)}$, such that

$$Y_{nk} = \begin{cases} 1 & \text{if named entity mention } n \text{ is of type } k, \\ 0 & \text{otherwise.} \end{cases}$$

and a corresponding multiclass classifier f such that $f(\mathbf{x}_n) = \mathbf{y}_n$, where $\mathbf{y}_n$ is the nth line of the matrix $\mathbf{Y}$. Of course, thanks to the distant supervision, we already know the part of the matrix $\mathbf{Y}$

[1] Alternatively, we could also classify each token individually, a possibility we would like to explore in future work.

corresponding to the positive examples. We will thus add constraints to impose the matrix $\mathbf{Y}$ to agree with the labels obtained using distant supervision.

2.2 Distant supervision by constraining $\mathbf{Y}$

First, each named entity mention belongs to exactly one class, which means that each line of the matrix $\mathbf{Y}$ contains exactly one 1. This is equivalent to the linear constraints:

$$\forall n \in \{1, ..., N\}, \quad \sum_{k=1}^{K+1} Y_{nk} = 1.$$

Second, for all the positive examples, we impose the matrix $\mathbf{Y}$ to agree with the labels obtained through distant supervision:

$$\forall n \in \mathcal{P}, \quad Y_{nc_n} = 1.$$

These constraints can be relaxed, by enforcing only a certain percentage or number of positive examples to be classified according to the labels obtained using distant supervision. This new kind of constraint is related to the *at-least-one* constraint described by Riedel et al. (2010). In our convex modeling framework, it can be formulated as follow:

$$\forall k \in \{1, ..., K\}, \quad \sum_{n \in \mathcal{P}_k} Y_{nk} \geq Z_k,$$

where $\mathcal{P}_k$ represents the set of positive examples corresponding to class k. In particular, using this kind of constraints, our approach can leverage ambiguous named entities. We reserve to future work the application of our approach to ambiguous named entities.

Finally, we want the percentage of examples to be classified as OTHER to be at least p, in order to avoid semantic drift. This is equivalent to imposing the linear constraint:

$$\sum_{n \in \mathcal{U}} Y_{n(K+1)} \geq pN.$$

The set of matrices $\mathbf{Y}$ verifying those constraints is denoted by $\mathcal{Y}$. It should also be noted that all those constraints are linear, thus defining a *convex set*.

2.3 Problem formulation and convex relaxation

Inspired by the discriminative clustering framework, introduced by Xu et al. (2004), we jointly learn the classifier f and the matrix $\mathbf{Y}$ by minimizing the cost function:

$$\begin{aligned} \min_{\mathbf{Y}, f} \quad & \sum_{n=1}^{N} \ell\left(\mathbf{y}_n, f(\mathbf{x}_n)\right) + \Omega(f) \\ \text{s.t.} \quad & \mathbf{Y} \in \{0,1\}^{N \times (K+1)}, \ \mathbf{Y} \in \mathcal{Y}, \end{aligned}$$

where ℓ is a multiclass loss function, Ω is a regularizer and $\mathcal{Y}$ is the set defined by the constraints introduced in section 2.2. Following Bach and Harchaoui (2007), we use linear classifiers $\mathbf{W} \in \mathbb{R}^{d \times (K+1)}$, the squared loss and the ℓ_2 norm regularizer, and thus obtain the following cost function:

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{W}} \quad & \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \frac{\lambda}{2}\|\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \mathbf{Y} \in \{0,1\}^{N \times (K+1)}, \ \mathbf{Y} \in \mathcal{Y}. \end{aligned}$$

We can then relax the constraint $\mathbf{Y} \in \{0,1\}^{N \times (K+1)}$ into $\mathbf{Y} \in [0,1]^{N \times (K+1)}$, in order to obtain the problem

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{W}} \quad & \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \frac{\lambda}{2}\|\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \mathbf{Y} \in [0,1]^{N \times (K+1)}, \ \mathbf{Y} \in \mathcal{Y}, \end{aligned}$$

which is jointly convex in $\mathbf{Y}$ and $\mathbf{W}$. Following Grave (2014), we solve the dual problem using the accelerated projected gradient descent algorithm (Nesterov, 2007; Beck and Teboulle, 2009).

	COMPANIES			DRUGS			ALL		
	P	R	F1	P	R	F1	P	R	F1
Stanford NER	N/A	52.6	N/A	N/A	N/A	N/A	N/A	N/A	N/A
String match	98.9	44.2	61.1	100	32.3	48.8	99.2	39.6	56.6
SVM (asym)	87.0	92.8	89.8	86.5	79.2	82.7	86.8	87.6	87.2
This work	82.9	95.8	88.9	87.4	94.0	90.6	84.6	95.1	89.5

Table 1: Results for different methods.

3 Experiments

In this section, we describe the application of our method to the extraction of the named entities from financial reports corresponding to healthcare companies. In particular, we are interested in extracting mentions of companies and drugs from those reports.

3.1 Data

Our dataset consists of quarterly financial reports (form 10-Q), filed by publicly traded healthcare companies, corresponding to the years 2013 and 2014. Our dataset contains $2,588$ reports, corresponding to 578 companies. We use the names of those 578 companies as the seed list for companies and the list of the 200 most searched drugs on the website `http://www.rxlist.com` as the seed list for drugs. We tagged and parsed this dataset using the Stanford part-of-speech tagger (Toutanova et al., 2003) and the MaltParser (Nivre et al., 2007). We manually labeled a set of $1,659$ named entity mentions corresponding to one 10-Q report, for evaluation. 405 of those named entity mentions are labeled as COMPANY and 251 are labeled as DRUG. Thus, 60% of named entity mentions correspond to the class OTHER. The long term goal of this project is to extract relations between companies, such as collaborations to develop new drugs.

3.2 Features

Each named entity mention is represented by the following features: the (lowercased) tokens that form the mention, a window of k words to the left and a window of k words to the right of the mention and the k ancestors, with their syntactic roles, of the mention in the dependency tree, for $k \in \{1, 2, 3\}$. Finally, we obtain a vectorial representation of each named entity mention by using the distributional model of semantics introduced by Grave et al. (2014). We trained a model with 100 classes on the full dataset of financial reports corresponding to healthcare companies.

3.3 Baselines

Our first baseline is a string match between named entity mentions and our seed lists. The second baseline, which can only be evaluated on the extraction of company names, is the Stanford named entity recognizer (Finkel et al., 2005), which was trained on the CoNLL and MUC datasets. Finally, we compare our method with a linear SVM with asymmetric costs for positive and unlabeled examples. This approach is very competitive, obtaining state-of-the-art results for text classification using PU learning (Liu et al., 2003).

3.4 Results

We trained our classifiers on $10,000$ examples, 593 of which being positive and the rest being unlabeled. We report precision, recall and F1 measures for the different methods in Table 1. First, we observe that the *string match* baseline achieves a high precision of 99.2, as expected, while having a low recall. Second, we observe that the Stanford named entity recognizer has a low recall on companies, retrieving only half of their mentions. An example of error is labeling *Merck* as PERSON. Third, we observe that both PU learning approaches outperform the two baselines by a large margin: SVM and our approach achieve F1 scores of 87.2 and 89.5 respectively. In particular,

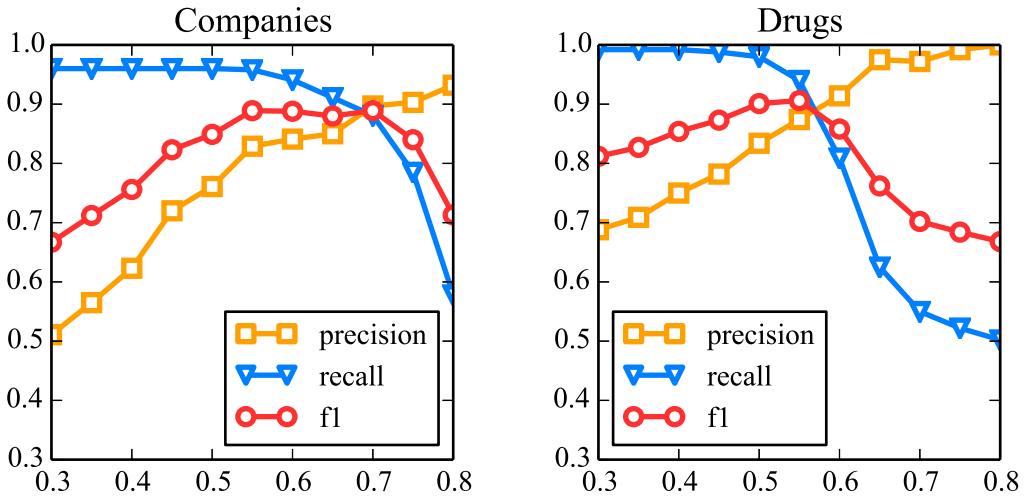

Figure 1: Influence of the parameter p on the precision and recall of our method.

our method outperforms SVM for extracting mentions of drugs, obtaining a F1 score of 90.6 while SVM have a F1 score of 82.7.

In Figure 1, we plot the precision, recall and F1 measures achieved by our method for different values of the parameter p, to illustrate its influence. We recall that the higher p, the more unlabeled examples are classified as negative. Unsurprisingly, we observe that when p increases, the precision increases while the recall drops. The parameter p thus allows to control the trade-off between precision and recall, and therefore to limit the semantic drift.

4 Discussion

In this paper, we introduced a novel method for weakly supervised named entity classification, based on discriminative clustering. We demonstrated that our approach seems competitive, on the task of extracting mentions of companies and drugs from financial reports corresponding to healthcare companies. As future work, we plan to evaluate our approach more extensively, on a larger number of classes and in other domains, such as biomedical articles. We would also like to explore the possibility of classifying all the tokens (instead of classifying only named entity mentions), in order to train a full named entity recognizer (and not only a named entity classifier). Finally, we also want to generalize this approach to ambiguous named entities, by relaxing the constraints that enforce the infered labels to be equal to the labels obtained using the seed list.

Acknowledgments

The author is supported by a grant from Inria (Associated team STATWEB) and would like to thank the anonymous reviewers for their helpful feedback.

References

Agichtein, E. and Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*.

Bach, F. and Harchaoui, Z. (2007). Diffrac: a discriminative and flexible framework for clustering. In *NIPS*.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*.

Bellare, K. and McCallum, A. (2007). Learning extractors from unlabeled text using relevant databases. In *IIWeb*.

Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. (2013). Finding actors and actions in movies. In *ICCV*.

Brin, S. (1999). Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases*.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *AAAI*.

Collins, M. and Singer, Y. (1999). Unsupervised models for named entity classification. In *EMNLP*.

Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *ISMB*.

Curran, J. R., Murphy, T., and Scholz, B. (2007). Minimising semantic drift with mutual exclusion bootstrapping. In *PACLING*.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*.

Finkel, J., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*.

Grave, E. (2014). A convex relaxation for weakly supervised relation extraction. In *EMNLP*.

Grave, E., Obozinski, G., and Bach, F. (2014). A markovian approach to distributional semantics with application to semantic compositionality. In *COLING*.

Joulin, A., Bach, F., and Ponce, J. (2010). Discriminative clustering for image co-segmentation. In *CVPR*.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In *ICDM*.

Liu, B., Lee, W. S., Yu, P., and Li, X. (2002). Partially supervised classification of text documents. In *ICML*.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP*.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*.

Ramanathan, V., Joulin, A., Liang, P., and Fei-Fei, L. (2014). Linking people in videos with their names using coreference resolution. In *ECCV*.

Riedel, S., Yao, L., and McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *ECML / PKDD*.

Riloff, E. and Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI*.

Ritter, A., Clark, S., and Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In *EMNLP*.

Ritter, A., Zettlemoyer, L., Mausam, and Etzioni, O. (2013). Modeling missing data in distant supervision for information extraction. *TACL*.

Talukdar, P. P. and Pereira, F. (2010). Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *ACL*.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*.

Wu, F. and Weld, D. S. (2007). Autonomously semantifying wikipedia. In *CIKM*.

Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. (2004). Maximum margin clustering. In *NIPS*.