Conditional Random Fields: Discriminative Training over Statistical features for Named Entity Recognition

Truc-Vien T. Nguyen, Alessandro Moschitti, and Giuseppe Riccardi

Department of Information Engineering and Computer Science,
University of Trento,
38050 Povo (TN), Italy
{nguyenthi,moschitti,riccardi}@disi.unitn.it

Abstract. We describe the experiments of the two learning algorithms for Named Entity Recognition. One implements Conditional Random Fields (CRFs), another makes use of Support Vector Machines (SVMs). Both are trained with a large number of features. While SVMs employ purely input features, CRFs also exploit statistical aspects in terms of unigram and bigram of both features and output tags. The main characteristic of our approach is the use of different learning models for the task.

Key words: Named Entity Recognition, Conditional Random Fields, Support Vector Machines.

1 Introduction

Named-entities (NEs) are a basic and important part for defining the semantics of a document. NEs are objects that can be referred by names ([2]), such as people, organizations, and locations. Our method aims at investigating the use of learning approaches in named entity recognition (NER). We developed a system which identifies NEs in text and experimented with two current off-the- shelf learning models: Conditional Random Fields and Support Vector Machines.

2 Recognition Method and Resource Collection

We selected Conditional Random Fields ([5]) as the learning algorithm. Conditional random fields (CRFs) are a probabilistic framework for labeling and segmenting sequence data. It presents several advantages over other purely generative models such as Hidden Markov models (HMM) by relaxing the independence assumptions required by HMM. Besides this, Maximum Entropy Markov models (MEMM) and other discriminative Markov models are prone to the label bias problem, which is solved effectively by CRFs.

2.1 CRFs for labelling

The task of assigning label sequences to a set of observation sequences arises in many fields, including bioinformatics, computational linguistics and speech recognition ([4, 6, 7]). For example, consider the natural language processing task of labeling the words in a sentence with their corresponding named-entity (NE) tags. In this task, each word is labeled with a tag indicating its appropriate named-entity, resulting in annotated text, such as:

[O I] [O conti] [O del] [O semestre] [O sono] [O stati] [O presentati] [O luned] [O sera] [O dal] [O direttore] [O generale] [O dell'] [O Istituto] [O ,] [B-PER Vittorio] [I-PER D'] [I-PER Angelantonio] [O ,] [O al] [B-ORG consiglio] [I-ORG di] [I-ORG amministrazione] [O ,] [O presieduto] [O da] [B-PER Italo] [I-PER Garbari] [O .]

Much like a Markov random field, a CRF is a form of undirected graphical model that defines a single log-linear distribution over label sequences given a particular observation sequence. Let us assume an input sequence X representing the sequence of observations and a sequence Y representing the hidden (or unknown) state variable that needs to be inferred given the observations. In a CRF, the distribution of each discrete random variable Y in the graph is conditioned on an input sequence X instead of a joint distribution as in HMMs. The conditional nature of such models means that no effort is wasted on modeling the observation sequence, and one is free from having to make unwarranted independence assumptions about these sequences. CRFs has proved to outperform both MEMMs and HMMs on a number of real-world sequence labeling tasks.

2.2 Software and features used

For experimentation, we used 3-fold cross-validation with CRF++ ([3]) to build our recognizer, which is a model trained discriminatively with the unigram and bigram features.

Features are extracted from a window at k words centered in the target word w (i.e. the one we want to classify with the B, O, I tags). More in detail such features are:

- The **word itself**, its **prefixes**, **suffixes**, and **part-of-speech**
- **Orthographic/Word features**. These are binary and mutually exclusive features that test whether a word contains *all upper-case, initial letter upper-case, all lower-case, roman-number, dots, hyphens, acronym, lonely initial, punctuation mark, single-char, and functional-word.*
- **Gazetteer features**. Class (geographical, first name, surname, organization prefix, location prefix) of words in the window.
- **Left Predictions**. The tags being predicted of the word on the left in the current classification.

The gazetteer lists are built with names imported from different sources. The company data is included with all the publicly traded companies listed in Google directory, Hoovers Online, the European business directory. Generic proper nouns are extracted from Wikipedia and various Italian sites. Moreover, the gazetteer lists are extracted partially from La Repubblica corpus ([1]), a large corpus of Italian newspaper text by using rule-based approach with patterns tuned specifically for each NE class. Our current gazetteer contains 11200 geographical names, 3500 locations, 7000 organizations, and 150000 person names.

3 System results and Discussion

In addition to the base CRF classifier we trained a second one in which we employed Support Vector Machines (SVMs). Although the second model performed worse than the base model in cross-validation, we reported the results for completeness.

Table 1. Results on the Development Set with External Resource

	CRFs		
Category	**Pr**	**Re**	**F_1**
All	83.23	78.45	80.76
GPE	83.63	85.58	84.55
LOC	76.79	46.71	57.79
ORG	72.54	61.40	66.50
PER	90.39	90.11	90.25
	SVMs		
Category	**Pr**	**Re**	**F_1**
All	82.56	78.83	80.64
GPE	82.49	86.20	84.28
LOC	78.68	50.31	61.01
ORG	70.93	62.42	66.38
PER	90.89	89.55	90.22

Table 1 summarizes the systems results on the development set for the categories to be annotated. Table 2 reports the same measures when external resources are not employed. We found that, with the same set of features, the accuracy of the NE classifiers upon two models are rather competitive. Table 3 shows the final results on the test set.

We found that the NE classes GPE and PER reach quite good F1 values, while the recognition of ORG and LOC seems problematic. This is in line with previous results in which ORG seems to be the most difficult to learn. Lack of resource (the gazetteer for LOC is the least) may stand for this low accuracy of LOC class.

Table 2. Results on the Development Set without External Resource

CRFs

Category	Pr	Re	F_1
All	77.05	70.65	73.70
GPE	80.77	78.12	79.30
LOC	73.69	32.77	46.14
ORG	63.42	54.47	58.61
PER	84.15	81.95	83.03

SVMs

Category	Pr	Re	F_1
All	76.61	71.01	73.69
GPE	78.53	79.31	78.84
LOC	67.30	35.55	46.54
ORG	67.10	56.37	61.28
PER	82.26	80.40	81.32

Table 3. CRFs Results on the Test set

Category	Pr	Re	F_1
All	82.26	77.43	79.77
GPE	83.93	81.80	82.85
LOC	71.21	30.13	42.34
ORG	70.05	65.87	67.89
PER	88.26	84.69	86.44

Acknowledgments. This work has been partially funded by the LiveMemories project (www.livememories.org) and Expert System (www.expertsystem.net) research grant.

References

1. Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., Mazzoleni, M.: Introducing the la Repubblica corpus: A large, annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In: Proceedings of LREC 2004, pp. 1771-1774. ELDA, Lisbon (2004)
2. Chinchor N., Robinson P.: MUC-7 Named Entity Task Definition. In: Proceedings of the MUC (1998)
3. CRF++: Yet Another CRF toolkit, http://crfpp.sourceforge.net/
4. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press (1998)
5. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282-289. Morgan Kaufmann, San Francisco, CA (2001)

6. McCallum, A., Freitag, D., Pereira, F.: Maximum entropy Markov models for information extraction and segmentation. In: Proceedings of International Conference on Machine Learning (2000)
7. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall Signal Processing Series. Prentice-Hall, Inc. (1993)