Named Entity Recognition with Semi-supervised learning

Introduction

Named entity recognition (NER) or tagging is an important subtask of information extraction task in natural language processing which find names such as organizations, persons, locations, etc. Automatically tagging named entities (NE) with high precision and recall requires a large amount of hand-annotated data which is very expansive to obtain. Thus semi-supervised learning is a common approach for this tagging process as generally we will have a small labeled data and a large unlabeled data. With semi-supervised learning we are trying to make both labeled data and unlabeled data contribute to our trained model and thus get a well-performed model with relatively low cost. There are different kinds of general approaches to SSL such as Clustering, Co-training, and Self-training.

Objectives

In this project, our basic goal is to experiment several different semi supervised learning approach on a same data set with labeled and unlabeled data and explore their performances. The labeled data is a subset of conll2003 and the unlabeled data is from Penn TreeBank.

The implemented algorithms:

1. Yarowsky and Yarowsky-cautious algorithm (Yarowsky, 1995 [1]; Collins and Singer, 1999 [2])
2. DL-CoTrain algorithm (Collins and Singer, 1999 [2])
3. Semi-supervised Conditonal Random Field (CRF) (Liao and Sriharsha, 2009[3])

Method-Yarowsky and Yarowsky-cautious

- ***(Yarowsky 95) Algorithm*:** Supervised Version

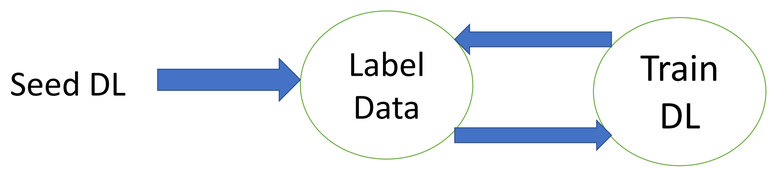
Input: n labeled training data of form (x_i, y_i) $(y_i \in Y=\{1 \ldots k\})$

Output: a function h: $X \times Y$ -> [0,1] $h(x,y) = \frac{Count(x,y)+\alpha}{Count(x)+k\alpha}$

The Decision Rule Scores: $\theta_{fj} \propto \frac{|\text{Å}_{fj}|+\epsilon}{|\text{Å}_{f}|+L\epsilon}$

- ***Yarowsky-cautious Algorithm***: Unsupervised Version

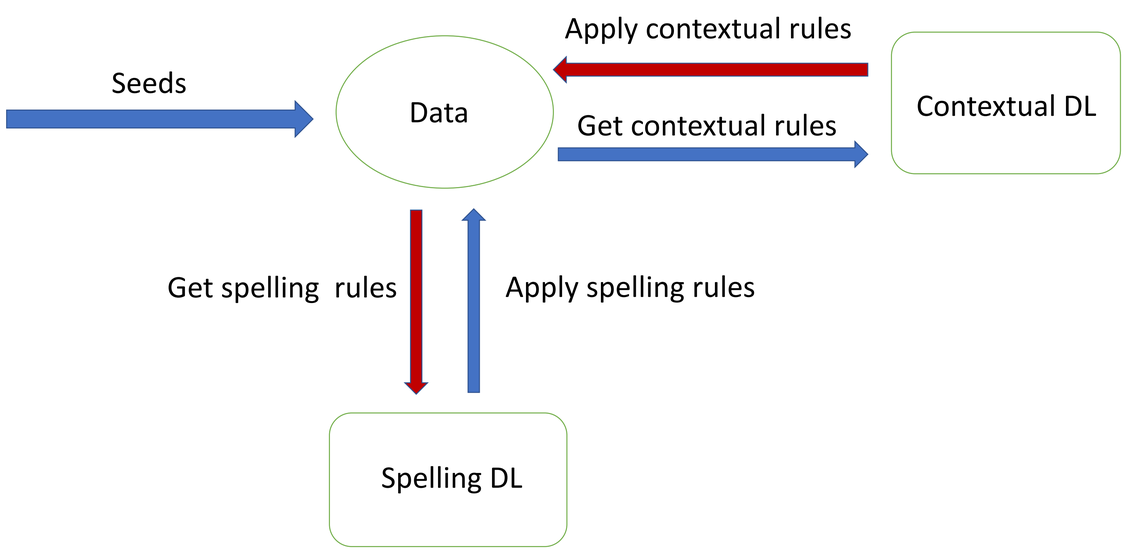
Compared to (Yarowsky 95), it is an unsupervised version. It starts with a small number of training data as seeds. ***Yarowsky-cautiou*** has a limit on the number of rules added at each stage. At the final iteration Yarowsky-cautious uses the current labelling to train a DL without a threshold or cautiousness



Method-DL_CoTrain

- ***Key Features*:**
  - Semi-supervised
  - Powerful in reducing the complexity
  - Decent performance with only a few labeled data
- ***Basic Idea***:

Each data example can be viewed via two perspectives. In NER, the two perspectives are *spelling features* and *contextual features*. Given some seeds (the labeled data), we derive new contextual rules which can be represented as (*contextual features* -> *label*) from spelling rules (*spelling features* -> *label*) and vice versa. This is what "Co-train" stands for. Here is a figure further illustrate this:



Method-Semi-supervised CRF

We use CRF to perform classification. CRFs are undirected graphical models trained to maximize the conditional probability $P(Y|X)$ of a sequence of labels $Y = y_1 \ldots y_N$ given the corresponding input sequence $X = x_1 \ldots x_N$.

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{n=1}^{N} \sum_{k} \lambda_k f_k(y_{n-1}, y_n, X, n)\right)$$

,where $Z(X)$ is normalization term and f_k is feature function, λ_k is learned by maximizing likelihood $\sum_i logP(Y_i|X_i) - \sum_k \frac{\lambda_k^2}{2\sigma_k^2}$ where σ_k^2 is the regularization parameter for f_k.

Features we used in our CRF classifier are: Token itself; Token capitalized or not; Token is tile or not; Token is digit or not; Token's position of sentence; Token's neighbors' feature, etc.

We then designed some rules to classify each token based on its features, which will take precedence over the CRF. We have a small amount of labeled data L and a large unlabeled corpus U from the test domain. At each iteration, the classifier trained on the previous training data is used to tag the unlabeled data.

Our final semi-supervised algorithm is:

Input:

L - a small set of labeled training data U - unlabeled data

Loop for k iterations:

Train a classifier Ck based on L;

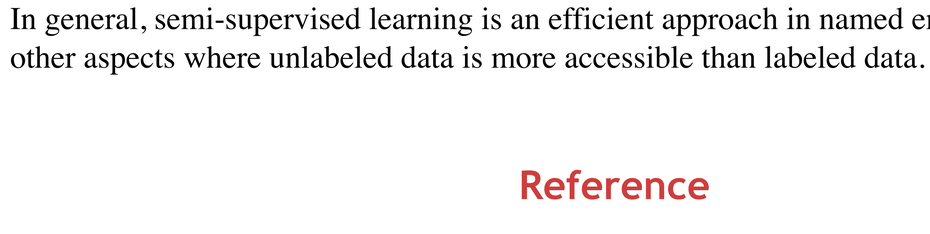
Extract new data D based on Ck;

Add D to L;

Results

To experiment all these methods' performances, we used the data set that contains around 200k labeled words and 800k unlabeled words. To evaluate our result we used accuracy of predicting all tags (I-PER, I-LOC, I-ORG, O) as our criteria:

Accuracy for each of the method：



Conclusion

As showed in our results above, results with only labeled data are lower than that with labeled and unlabeled data. The reason is that semi-supervised learning enable us to utilize large unlabeled data as if it is labeled in our models' training and thus get a larger training set and result to better performance.

Semi-supervised conditional random field achieved better result than yarowsky algorithm and dl_cotrain algorithm as it is relatively more complex than the other two algorithm. However, this improvement is not significant as the yarowsky and dl_cotrain also performed quite well. Since Yarowsky algorithm is a straightforward method, we can do some improvement on it to get better performance.

In general, semi-supervised learning is an efficient approach in named entity recognition and other aspects where unlabeled data is more accessible than labeled data.

Reference

1. Yarowsky, David. "Unsupervised word sense disambiguation rivaling supervised methods." 33rd annual meeting of the association for computational linguistics. 1995.
2. Collins, Michael, and Yoram Singer. "Unsupervised models for named entity classification." 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 1999.
3. Liao, Wenhui, and Sriharsha Veeramachaneni. "A simple semi-supervised algorithm for named entity recognition." Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. 2009.