The Yarowsky algorithm applied to Named Entity Classification

Max Whitney Anoop Sarkar

SFU

November 20, 2008

The Paper

- M. Collins and Y. Singer. Unsupervised Models for Named Entity Classification. Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. EMNLP-VLC 1999.
  http://www.aclweb.org/anthology-new/W/W99/W99-0613.pdf

The Data

- 971,746 sentences of the NYT were parsed using the Collins parser
- The task is to identify three types of named entities:
  - 1 : Location (LOC)
  - 2 : Person (PER)
  - 3 : Organization (ORG)
- -1 : Either not a NE, or a label that indicates a *don't know*

The Data

- Noun phrases (NPs) were extracted that met the following conditions:

1. The NP contained only words tagged as NNP or NNPS (proper nouns)
   (... (NP International/NNP Business/NNP Machines/NNPS) ...)
2. The NP appeared in the following two contexts:
   - Modified by an appositive whose "head" is a singular noun (tagged NN):
     ..., says Maury Cooper, a vice president at S. & P.
     produces: *Maury Cooper* $\Rightarrow$ NE and context $\Rightarrow$ *president*
   - In a prepositional phrase (PP) modifying an NP whose "head" is a singular noun *..., fraud related to work on a federally funded sewage plant in Georgia* produces: *Georgia* $\Rightarrow$ NE and context $\Rightarrow$ *plant-in*

Features

```
-1 X0_Maury-Cooper X01_president X2_Maury X2_Cooper X3_LEFT
-1 X0_S.&P. X11_president-at X7_.&. X3_RIGHT

-1 X0_I.B.M. X11_agreement-with X6_ALLCAP2 X7_... X3_RIGHT

-1 X0_NATO X11_independence-from X5_ALLCAP X3_RIGHT
```

- X0-w : the NP *w* in question
- X01-w : head noun *w* of the appositive phrase
- X11-w : the head noun *w* being modified plus the preposition
- X2-w : unigrams *w* from the NP
- X3-x : appositive (X3-LEFT) or PP (X3-RIGHT)
- X5 / X6 : all caps / all caps with intervening symbols like periods
- X7-x : non-alphabetic symbols *x* in the NP

The Task

- Classify NPs into one of the three named entity classes (LOC, PER, ORG)
- 89,305 unlabeled training data examples.
  $\Rightarrow$ How to get 88,962 examples? 343 examples extra.
- 1000 test data examples annotated by hand with the right answer.
  - Test data includes NPs that are not LOC, PER, ORG.
  - There are some easily identifiable NPs that are not named entities: month names
  - Removing all month names leaves 962 examples
- Two types of evaluation: *clean* (on 962 examples) and *noisy* (removing 85 cases that are noise, leaving 877 that are either LOC, PER, ORG)

The Yarowsky Algorithm

- The Yarowsky algorithm starts with a set of seed rules:

```
Score NP                  NE-Type
--------------------------------
1.0   X0_New-York         1
1.0   X0_California       1
1.0   X0_U.S.             1
1.0   X2_Mr.              2
1.0   X0_Microsoft        3
1.0   X0_I.B.M.           3
1.0   X2_Incorporated     3
```

- Apply the seed rules on the unlabeled training data.
- Learn a decision list from the "labeled" data.
- A decision list looks just like a set of seed rules.
- Bootstrap.

Decision list classifier

- $\mathcal{X}$ is the set of features
- $\mathcal{Y}$ is the set of output labels. In this task: LOC, PER, ORG.
- Input: $\mathbf{x} = \{x_1, \ldots, x_m\}$ where $x_i \in \mathcal{X}$. Output: $y \in \mathcal{Y}$
- Define f:

$$f(\mathbf{x}) = \underset{x \in \mathbf{x}, y \in \mathcal{Y}}{\text{argmax}}\, h(x, y)$$

- Where $h : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ and $h(x, y)$ can be seen as the conditional probability $p(y \mid x)$

$$\begin{aligned} h(x, y) &= \frac{\mathit{Count}(x, y) + \alpha}{\mathit{Count}(x) + |\mathcal{Y}| \cdot \alpha} \\ &= \frac{\mathit{Count}(x, y) + \alpha}{\sum_y \mathit{Count}(x, y) + |\mathcal{Y}| \cdot \alpha} \end{aligned}$$

- Let $(x, y) = f(\mathbf{x})$. The classifier reports label y if $h(x, y) > 0.95$ else reports -1.

Decision list classifier

```
1: LOC, 2: PER, 3: ORG

0.999928528035 X2_Mr. 2
0.999876183991 X01_analyst 2
0.999812611262 X2_York 1
0.999795354548 X2_John 2
0.999764511951 X0_U.S. 1
0.999733439957 X11_analyst-at 3
0.999686176055 X2_Stock 1
0.999636561875 X2_Michael 2
0.999615606381 X11_trading-on 1
0.99960262269 X2_Securities 3
0.99957292334 X0_New-York-Stock-Exchange 1
0.999546793564 X0_Japan 1
0.999530846821 X01_officer 2
0.999521874253 X0_California 1
0.999471319059 X0_New-York 1
```

Decision list classifier

```
1: LOC, 2: PER, 3: ORG

0.333333333333 X11_campaign-outside 1
0.333333333333 X11_stock-except 1
0.333333333333 X2_Zilber 1
0.333333333333 X0_Seiko 1
0.333333333333 X01_collateral 1
0.333333333333 X0_Polish-Revolution 1
0.333333333333 X0_non-aligned-Afghans 1
0.333333333333 X11_tent-in 1
0.333333333333 X0_Adam-Corporation\/Group 1
0.333333333333 X11_triumph-in 1
0.333333333333 X0_Associated 1
0.333333333333 X0_Fofo-Sunia 1
0.333333333333 X0_Associates 1
0.333333333333 X0_Lanham-Act 1
```

Yarowsky variants: Being Cautious

- Introduced by (Collins and Singer, 1999)
- A trick from the Co-training paper (Blum and Mitchell, 1998) is to be cautious.
- Don't add all rules that are above 0.95 confidence score.
- Add only n rules (say 5) and increase this amount by n in each iteration.
- This gives us two variants: Basic-Yarowsky and Yarowsky-cautious.

Yarowsky variants: Ambiguous seed rules

- Certain seed rules can be ambiguous.
- Mr. $\rightarrow$ PER vs. Incorporated $\rightarrow$ ORG
- We use two variants:
  - Mr. before Incorporated: Yarowsky-cautious(1)
  - Incorporated before Mr.: Yarowsky-cautious(2)

Seed Selection

- Selecting seed rules: what was the Collins and Singer strategy?
- Different ways to select seed rules:

Frequency	Sorting by frequency of feature occurrence.
Contexts	Sorting by number of other features a feature was seen with.
Weighted	Sorted by a weighted count of the other features a feature was seen with. Feature weights are their overall frequency of occurrence as a fraction of the total number of features seen.

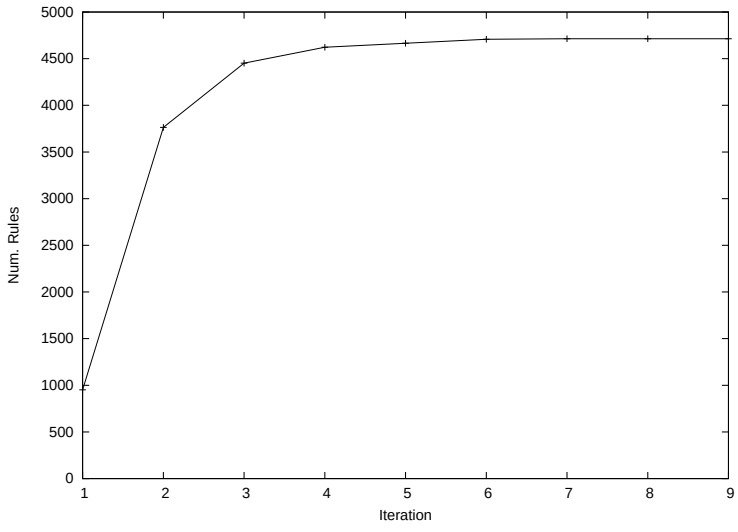
- In each case, the evaluation was done of the training data, and seeds were extracted from the sorted list of features by manual evaluation.
- Location features appear infrequently in the top features in these orderings, and it is possible that some good location seeds were overlooked in the manual evaluation stage.

Algorithm	Clean	Noisy	Last Iter		
			Iters	Rules	Coverage
Baseline (all organization)	45.8%	41.8%			
Basic Yarowsky	82.2%	74.9%	9	4622	70%
Yarowsky-cautious(1)	90.4%	82.4%	202	3030	73%
Yarowsky-cautious(2)	91.1%	83.1%	448	6720	77%

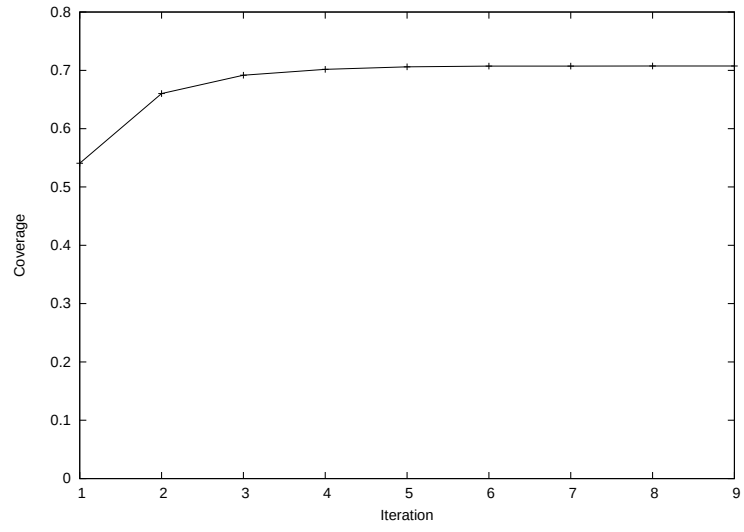
Table: The accuracy for the Yarowsky algorithms found here so far, (1) is where the "Mr." seed rule is applied before the "Incorporated" one, and (2) is where the "Incorporated" seed rule is applied before the "Mr." one.

Algorithm	Accuracy (Clean)	Accuracy (Noise)
Baseline (all organization)	45.8%	41.8%
Basic Yarowsky	81.3%	74.1%
Yarowsky-cautious	91.2%	83.2%

Table: The accuracy for the Yarowsky algorithms reported by (Collins and Singer, 1999)

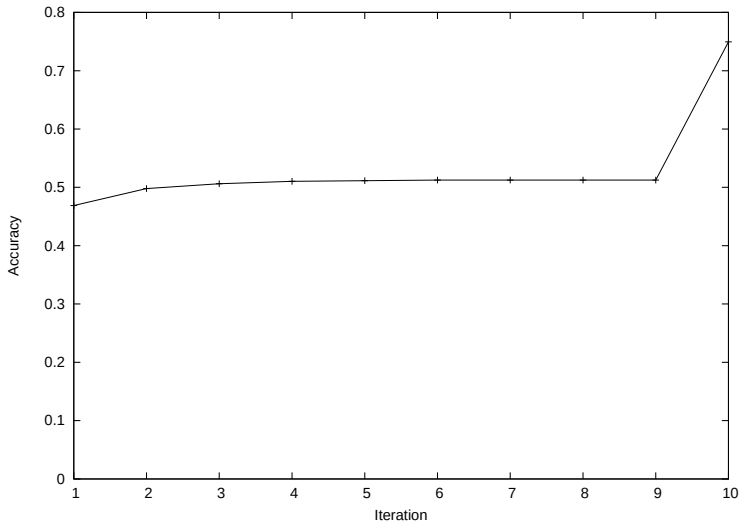


(a) Number of rules

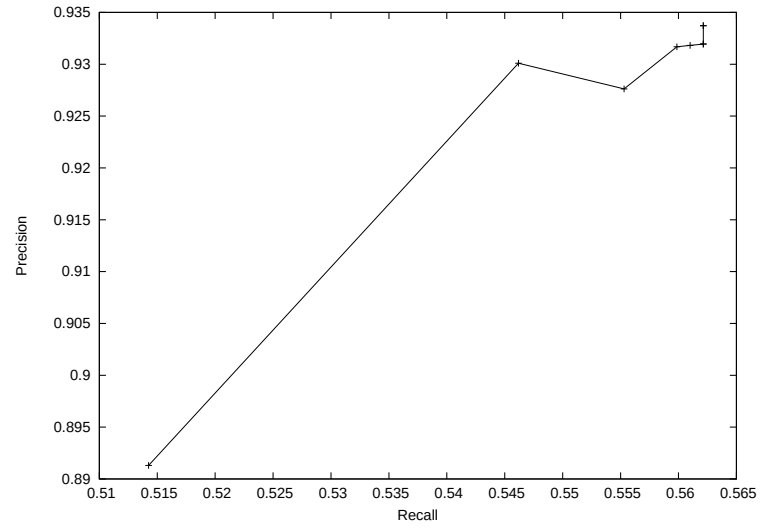


(b) Coverage on training data

Figure: Results for the basic Yarowsky algorithm (2)

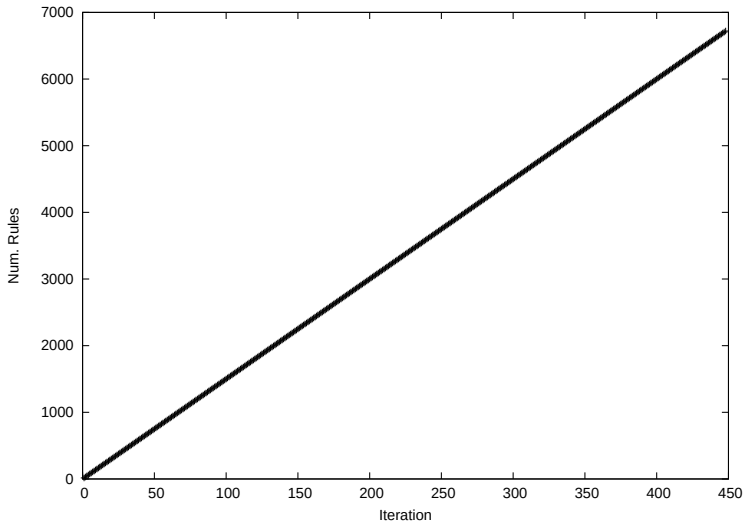


(a) Accuracy on test data

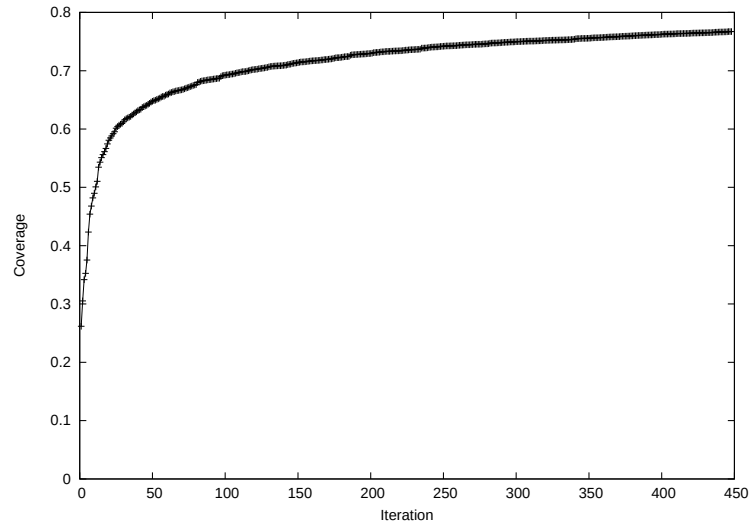


(b) Precision and recall on test data

Figure: Results for the basic Yarowsky algorithm (2)

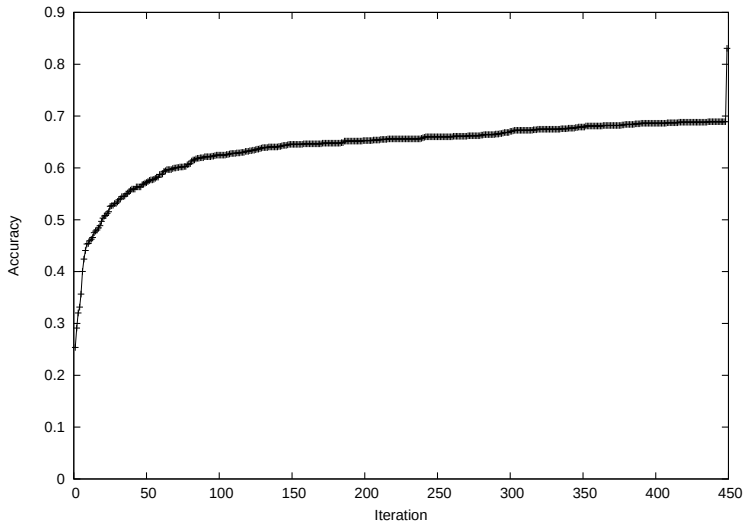


(a) Number of rules

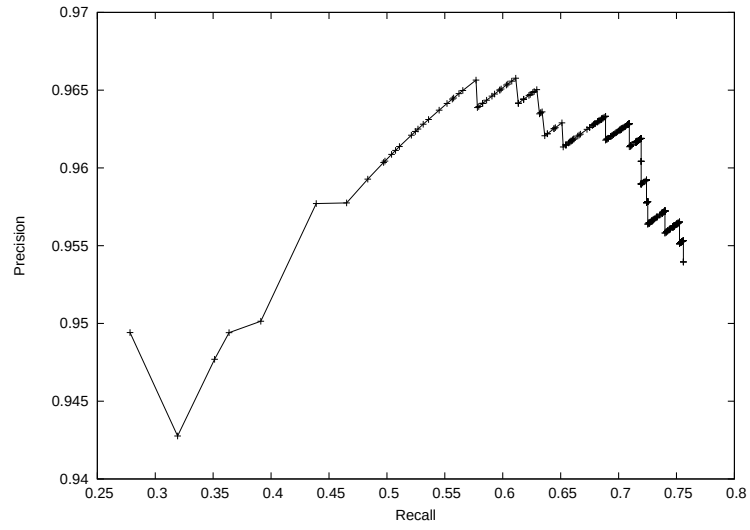


(b) Coverage on training data

Figure: Results for the Yarowsky-cautious (2)



(a) Accuracy on test data



(b) Precision and recall on test data

Figure: Results for the Yarowsky-cautious (2)

Seed selection: Basic Yarowsky

Num. rules	Frequency		Contexts		Weighted	
3	0.76	0.69	0.76	0.69	0.71	0.64
9	0.82	0.74	0.79	0.72	0.77	0.70
15	0.86	0.78	0.84	0.77	0.79	0.72

Seed selection: Yarowsky-Cautious

Num. rules	Frequency		Contexts		Weighted	
3	0.84	0.77	0.84	0.77	0.88	0.80
9	0.91	0.83	0.90	0.82	0.82	0.74
15	0.91	0.83	0.91	0.83	0.85	0.77