Semi-Supervised Active Learning for Sequence Labeling

MA.Mehralian
Outline

• Active learning
  – Introduction
  – Algorithm
  – Query strategy
• Conditional Models VS HMM
• Active Learning for Sequence Labeling
  – Fully-Supervised
  – Semi-Supervised
• Experiments for Named Entity Recognition (NER)
Active learning-intro

• Active learning is a supervised learning method in which the learner is in control of the data from which it learns.

• the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data.
Active learning - Algorithm

• **Given:**
  – $B$: number of examples to be selected
  – $L$: set of labeled examples
  – $b$: set of unlabeled examples
  – $U_M$: utility function

• **Algorithm:**
  loop until stopping criterion is met
  1. learn model $M$ from $L$
  2. for all $p_i \in P: U_{pi} \leftarrow U_M(p_i)$
  3. select $B$ examples $p_i \in P$ with highest utility $U_{pi}$
  4. query human annotator for labels of all $B$ examples
  5. move newly labeled examples from $P$ to $L$

return $L$
Active learning-Algorithm

Selected by query strategy

Labeled Pool \(\xrightarrow{L} \) (ML Model) \(\xrightarrow{M} \) Unlabeled Pool

[annotator]
Active learning-Algorithm

\[ L \rightarrow (ML \ Model) \rightarrow M \]

Labeled Pool \[ \rightarrow \] \[ annotator \] \[ \leftarrow \] Unlabeled Pool

Select by query strategy
Active learning-Algorithm

Labeled Pool $\rightarrow$ (ML Model) $\rightarrow$ Unlabeled Pool

Labeled Pool $\rightarrow$ [annotator] $\rightarrow$ Select by query strategy
Active learning-Algorithm

L

(ML Model)

M

Labeled Pool

Unlabeled Pool

[annotator]

Select by query strategy
1. **Uncertainty Sampling**
   - In this framework, an active learner queries the instances about which it is least certain how to label.
   - When using a probabilistic model for binary classification, uncertainty sampling simply queries the instance whose posterior probability of being positive is nearest 0.5
Active learning-Query strategy

1. Uncertainty Sampling- Uncertainty measure

- Least confident

\[ x_{LC}^* = \arg \max_x 1 - p_\theta(\hat{y}|x) \]

- Margin sampling

\[ x_M^* = \arg \min_x p_\theta(\hat{y}_1|x) - p_\theta(\hat{y}_2|x) \]

- Entropy

\[ x_H^* = \arg \max_x - \sum_i p_\theta(y_i|x) \log p_\theta(y_i|x) \]
2. Query-By-Committee (QBC)

- The QBC approach involves maintaining a committee $C = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$ of models which are all trained on the current labeled set, but represent competing hypotheses. Each committee member is then allowed to vote on the labeling of query candidates. The most informative query is considered to be the instance about which they most disagree.
2. Query-By-Committee (QBC)
2. Query-By-Committee (QBC)

   – In order to implement a QBC selection algorithm, one must:
     1. be able to construct a committee of models that represent different regions of the version space, and
     2. have some measure of disagreement among committee members.
Active learning-Query strategy

2. Query-By-Committee (QBC) - Committee of models
   - For generative model classes, this can be done more generally by randomly sampling an arbitrary number of models from some posterior distribution $p(\theta|L)$.
   - For other model classes, such as discriminative or non-probabilistic models, have proposed query-by-boosting and query-by-bagging.
Active learning-Query strategy

2. Query-By-Committee (QBC)-disagreement measure
   – For measuring the level of disagreement, two main approaches have been proposed.

1. vote entropy
   \[ x^*_E = \arg\max_x - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C} \]

2. Kullback-Leibler
   \[ x^*_H = \arg\max_x \frac{1}{C} \sum_{c=1}^C D(p_\theta \parallel P_C) \]
   \[ D(p_\theta \parallel P_C) = \arg\max_x - \sum_i p_{\theta(c)}(y_i|x) \log \frac{p_{\theta(c)}(y_i|x)}{p_C(y_i|x)} \]
Conditional Models **VS** HMM

- HMMs are a form of generative model, that defines a joint probability distribution \( p(X,Y) \) where \( X \) and \( Y \) are random variables respectively ranging over observation sequences and their corresponding label sequences.

- In order to define a joint distribution of this nature, generative models must enumerate all possible observation sequences – a task which, for most domains, is intractable.
Conditional Models VS HMM

- A conditional model specifies the probabilities of possible label sequences given an observation sequence.
- Furthermore, the conditional probability of the label sequence can depend on arbitrary, nonindependent features of the observation sequence without forcing the model to account for the distribution of those dependencies.
Conditional Models VS HMM

✓ no effort is wasted on modeling the observations.

✓ free from having to make unwarranted independence.

✓ arbitrary attributes of the observation data may be captured by the model, without the modeler having to worry about how these attributes are related.
ME Markov Model (MEMM) VS HMM

HMM: joint probability to paired observation and label sequences

MEMM: transition and observation functions are replaced by a single function

the probability of the transition from state $s'$ to state $s$ on input $o$. 

$$P(s|s',o)$$
Active Learning for Sequence Labeling- Fully-Supervised (FuSAL)

- a sentence is usually considered as proper sequence unit.
- Sequences for which the current model is least confident on the most likely label sequence are preferably selected.
- Use Query By Sampling \(-\) least confident strategy.
Active Learning for Sequence Labeling-Semi-Supervised (SeSAL)

• Within many sequences of natural language data, there are probably large subsequences on which the current model already does quite well and thus could automatically generate annotations with high quality.

• Only those tokens remain to be manually labeled on which the current model is highly uncertain regarding their class labels, while all other tokens are automatically tagged.
Active Learning for Sequence Labeling-Semi-Supervised (SeSAL)- parameters

• confidence threshold $t$, which directly influences the portion of tokens to be manually labeled.

• delay factor $d$, can be specified which channels the amount of manually labeled tokens obtained with FuSAL before SeSAL is to start.

*Let's go Experiments!*
Experiments for NER

• All experiments start from a seed set of 20 randomly selected examples and, in each iteration, 50 new examples are selected using AL.

| corpus     | entity classes | sentences | tokens     |
|------------|----------------|-----------|------------|
| MUC7       | 7              | 3,020     | 78,305     |
| PENNBIOIE  | 3              | 10,570    | 267,320    |

• The results reported below are averages of 20 independent runs. For each run, we randomly split each corpus into a *pool* of unlabeled examples to select from (90% of the corpus), and a complementary *evaluation set* (10% of the corpus)
Experiments for NER

- most of the tokens belong to the OUTSIDE class so that SeSAL can be expected to be very beneficial.

- SeSAL is run with a delay rate of \( d = 0 \) and a very high confidence threshold of \( t = 0.99 \) so that only those tokens are automatically labeled on which the current model is almost certain.
Experiments for NER-1

• the distribution of the model’s confidence values over all tokens of the sentences on MUC7
Experiments for NER-2

Fully Supervised vs. Semi-Supervised AL.

PennBioIE

MUC7

F-score

SeSAL
FuSAL
RAND

F-score

SeSAL
FuSAL
RAND

manually labeled tokens

manually labeled tokens
Experiments for NER-3

Detailed Analysis of SeSAL.

| manual | automatic | Σ   | AR(%) | errors | ACC  |
|--------|-----------|-----|-------|--------|------|
| 1,000  | 253       | 1,253 | 79.82 | 6      | 99.51|
| 5,000  | 6,207     | 11,207 | 44.61 | 82     | 99.27|
| 10,000 | 25,506    | 34,406 | 28.16 | 174    | 99.51|
| 12,800 | 57,371    | 70,171 | 18.24 | 259    | 99.63|
Experiments for NER-4

Impact of the Confidence Threshold.

**learning curves**

- F-score vs. manually labeled tokens
- Curves for different confidence thresholds: t=0.99, t=0.95, t=0.90, t=0.70

**error curves**

- Errors vs. all labeled tokens
- Different confidence thresholds: t=0.99, t=0.95, t=0.90, t=0.70
Experiments for NER-5

Impact of the Delay Rate.
My Experiments ;)

- Confident

![Probability Distribution](image)

![Probability Percent](image)
My Experiments ;)

- Margin

![Diagram showing margin distribution and margin percent with corresponding values and colors.]

A: 86% Margin Distribution

B: 70% Margin Percent

Colors:
- Dark Blue: 0.1
- Blue: 0.2
- Light Blue: 0.3
- Cyan: 0.4
- Green: 0.5
- Yellow: 0.6
- Orange: 0.7
- Red: 0.8
- Pink: 0.9
- Dark Pink: 1.0
My Experiments ;)

- Entropy

**A**

- Entropy Distribution
  - 10%
  - 5%
  - 4%
  - 4%
  - 5%

- Entropy Percent
  - 10%
  - 5%
  - 4%
  - 4%
  - 5%

**B**

- Entropy Percent
  - 1.0
  - 0.9
  - 0.8
  - 0.7
  - 0.6
  - 0.5
  - 0.4
  - 0.3
  - 0.2
  - 0.1

- Entropy Distribution
  - 0.1
  - 0.2
  - 0.3
  - 0.4
  - 0.5
  - 0.6
  - 0.7
  - 0.8
  - 0.9
  - 1.0
My Experiments ;)

![Bar Graph]

- Axis X: 1 to 11
- Axis Y: 0.1 to 0.9
- Two categories: A and B

The graph shows a distribution of values with a large peak at category 1 in category A.
My Experiments ;)

| Parameter value | Accuracy(%) | Used(%) |
|-----------------|------------|--------|
| **Least confident** |            |        |
| 0.6             | 96.72      | 42.72  |
| 0.4             | 96.52      | 16.27  |
| **Margin Sampling** |        |        |
| 0.4             | 96.73      | 25.45  |
| 0.2             | 96.51      | 13.38  |
| **Entropy**     |            |        |
| 0.8             | 96.73      | 37.63  |
| 0.9             | 96.64      | 26.48  |
References

• [1] K. Tomanek and U. Hahn, "Semi-supervised active learning for sequence labeling," in , 2009, pp. 1039--1047.

• [2] H. M. Wallach, "Conditional random fields: An introduction," Rapport technique MS-CIS-04-21, Department of Computer and Information Science, University of Pennsylvania, vol. 50, 2004.

• [3] B. Settles, "Active Learning Literature Survey," techreport, 2009,..

• [4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in , 2001, pp. 282--289.

• [5] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in , 2000, pp. 591—598.
MISSION ACCOMPLISHED