Multi-Criteria-based Active Learning for Named Entity Recognition

Dan Shen
Outline

- Introduction
- SVM-based NER system
- Multiple Criteria for Active Learning
  - Informativeness
  - Representativeness
  - Diversity
- Active Learning Strategies
- Experiments and Results
- Conclusion
Motivation

- **Named Entity Recognition (NER)**
  - most of current work: supervised learning
  - a large annotated corpus
    - MUC-6 / MUC-7 corpus (newswire domain)
    - GENIA corpus (biomedical domain)

- **Limitation of supervised NER**
  - corpus annotating: tedious and time-consuming
  - adaptability: in limited level

- **Target of our work**
  - explore active learning in NER
  - minimize the human annotation effort
  - without degrading performance
Active Learning Framework

- **Given**
  - an small labeled data set \( L \)
  - a large unlabeled data set \( U \)

- **Repeat**
  - Train a model \( M \) on \( L \)
  - Use \( M \) to test \( U \)
  - select the most useful example from \( U \)
  - require human expert to label it
  - add the labeled example to \( L \)

- **Until** \( M \) achieves a certain performance level
Active Learning Criteria

- **Active learning with informativeness**
  - most of current work
  - committee-based and certainty-based

- **Active learning with representativeness**
  - [McCallum and Nigam 1998] and [Tang et al. 2002]

- **Active learning with diversity**
  - [Brinker 2003]

- **NO works explored multiple criteria in active learning**
Active Learning in NLP

- Explored in a number of NLP tasks
  - POS Tagging
  - Scenario Event Extraction
  - Text Classification
  - Statistical Parsing
  - ...

- NO works explored active learning for NER
Outline

- Introduction
- SVM-based NER system
- Multiple Criteria for Active Learning
  - Informativeness
  - Representativeness
  - Diversity
- Active Learning Strategies
- Experiments and Results
- Conclusion
SVM-based NER system

- Recognize one class of NEs at a time
  - Best performance in BioCreAtIve Competition 2003

- Features
  - Binary feature vector
  - Different from supervised model
    - Cannot be produced statistically from training data set
    - No gazetteer or dictionaries

- Effort of human experts
  - Provide the basic knowledge for certain NE class
    - E.g. semantic triggers
  - Label the selected examples iteratively
Active Learning for NER

- Example unit in NER
  - Word-based
    - Select most useful word
    - Not reasonable: manually label a single word without any contexts
  - Sentence-based
    - Select most useful sentence
    - Don’t need to read the whole sentence to annotate one NE
  - Named entity-based
    - Select a word sequence (a named entity and its context)

- Active Learning for NER
  - Only word-based score is available from SVM
  - Measurements: extend from words to NEs
Outline

- Introduction
- SVM-based NER system
- **Multiple Criteria for Active Learning**
  - Informativeness
  - Representativeness
  - Diversity
- Active Learning Strategies
- Experiments and Results
- Conclusion
1. Informativeness Criterion

Most informative example: most uncertain in existing model

Most previous works are only based on this criterion.
In SVM, only support vectors are useful

Informativeness degree of a word

- How it will make effect on support vectors by adding it to training data set
- Distance of its feature vector to the separating hyperplane

\[ Dist(w) = \sum_{i=1}^{M} \alpha_i y_i k(s_i, w) + b \]

- the closer the word is to the hyperplane, the more informative the word is for the existing model.
Informativeness Measurement for NE

- NE -- a sequence of words
  - \( NE = w_1w_2...w_N \), \( w_i \) is the \( i^{th} \) word of \( NE \)

- Three scoring functions
  - Info_Avg: 
    \[
    Info(NE) = 1 - \frac{\sum_{w_i \in NE} Dist^*(w_i)}{N}
    \]
  - Info_Min: 
    \[
    Info(NE) = 1 - \min_{w_i \in NE} \{ Dist^*(w_i) \}
    \]
  - Info_InclRate: 
    \[
    Info(NE) = \frac{NUM(Dist^*(w_i) < \alpha)}{N}
    \]
Most representative example: represent most examples

Only few works [McCallum and Nigam 1998; Tang et al. 2002] consider this criterion.
Similarity Measurement between Words

- **Cosine-similarity Measurement**
  - The smaller the angle is, the more similar the vectors are

- **Cosine-similarity Measurement in SVM**
  - kernel function $k(w_i, w_j)$: replace the inner $w_i \cdot w_j$ product

\[
\text{Sim}(w_i, w_j) = \frac{|k(w_i, w_j)|}{\sqrt{k(w_i, w_i)k(w_j, w_j)}}
\]
Similarity Measurement between NEs

- Dynamic Time Warping (DTW) algorithm
  - Alignment of two word sequences

- Given
  - point-by-point distance

- To find an optimal path
  - Minimize accumulated distance along the path

\[ (w_{1n}, w_{2n}) \]
\[ (w_{14}, w_{23}) \]
An Example -- similarity between “Oct 1 binding protein” and “NF kappa B binding protein”

<table>
<thead>
<tr>
<th>Protein</th>
<th>binding</th>
<th>Oct 1</th>
<th>NF</th>
<th>kappa</th>
<th>B</th>
<th>binding</th>
<th>protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>0.5</td>
<td>0.5</td>
<td>0.71</td>
<td>0.25</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binding</td>
<td>0.5</td>
<td>0.5</td>
<td>0.71</td>
<td>0</td>
<td>0.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accumulated distances</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>2.5</td>
<td>2.5</td>
<td>2.71</td>
<td>1.92</td>
<td>1.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binding</td>
<td>2</td>
<td>2</td>
<td>2.21</td>
<td>1.67</td>
<td>1.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct</td>
<td>1.5</td>
<td>1.5</td>
<td>1.67</td>
<td>2.67</td>
<td>2.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NF</td>
<td>0.5</td>
<td>1</td>
<td>1.71</td>
<td>1.96</td>
<td>2.21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Distance between the two NEs: 1.67
Representativeness Measurement for NE

- **Representativeness of NE \(i\) in NESet**
  - NESet = \{\(NE_1, \ldots, NE_i, \ldots, NE_N\}\}
  - Quantified by its density
  - The average similarity between \(NE_i\) and the other \(NE_j\) \((j \neq i)\) in NESet

\[
Rep(NE_i) = \frac{\sum_{j \neq i} Sim(NE_i, NE_j)}{N - 1}
\]

- **Most representative NE**
  - Largest density among all NEs in NESet
  - Centroid of NESet
3. Diversity Criterion

Maximize the training utility of a batch: the members in the batch have high variance to each other.

Only one work [Brinker 2003] considered this criterion.
Global Consideration

- Consider the examples in a whole sample space
- K-Means Clustering
  - Cluster all named entities in NESet
  - Suppose:
    - the examples in one cluster are quite similar to each other
  - Select the examples from different clusters at a time
- Time consuming
  - Compute the centroids of clusters
  - Repartition examples
- For efficiency, filter out NEs before clustering
Local Consideration

- Consider the examples in a batch
- **For an example candidate:**
  - Compare it with all previously selected examples in the batch one by one
  - Add it into the batch
    - If the similarity between all of them is below a threshold
- **Threshold:**
  - The average of the pairwise similarities in NESet
- **Example candidate selection:**
  - Certain measurement
- **More efficient!**
Outline

- Introduction
- SVM-based NER system
- Multiple Criteria for Active Learning
  - Informativeness
  - Representativeness
  - Diversity
- Active Learning Strategies
- Experiments and Results
- Conclusion
Strategy 1

Unlabeled Data Set

Select M most informative examples (Informativeness Criterion)

Intermediate Set

Select centroid of each cluster (Representativeness Criterion)

Clustering (K clusters) (Diversity Criterion)

Batch
Strategy 2

Select example with max score: $\text{Info} + (1 - \text{Info}) \times \text{Rep}$
(Informativeness & Representativeness Criteria)

Compare the candidate with each example in Batch
IF any of the similarity values > threshold
THEN reject
ELSE add to Batch
(Diversity Criterion)
Outline

- Introduction
- SVM-based NER system
- Multiple Criteria for Active Learning
  - Informativeness
  - Representativeness
  - Diversity
- Active Learning Strategies
- Experiments and Results
- Conclusion
Data Set

- **Newswire Domain**
  - MUC-6 Corpus
  - 438 Wall Street Journal articles
  - To recognize Person, Location and Organization

- **Biomedical Domain**
  - GENIA Corpus V1.1
  - 670 MEDLINE abstracts
  - To recognize Protein
Experimental Setting 1

- **Corpus Split**
  - Initial training data set
  - Test data set
  - Unlabeled data set
  - Size of each data set

- **Batch size K**
  - = 50 in biomedical domain
  - = 10 in newswire domain

- **Example unit**
  - a named entity
  - its context (previous 3 words and next 3 words)
## Corpus Split

<table>
<thead>
<tr>
<th>Domain</th>
<th>Class</th>
<th>Corpus</th>
<th>Initial Training Set</th>
<th>Test Set</th>
<th>Unlabeled Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bio</td>
<td>PRT</td>
<td>GENIA 1.1</td>
<td>10 Sent. (277 words)</td>
<td>900 Sent.</td>
<td>8004 Sent.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(26K words)</td>
<td>(223K words)</td>
</tr>
<tr>
<td>News</td>
<td>PER</td>
<td>MUC-6</td>
<td>5 Sent. (130 words)</td>
<td>602 Sent.</td>
<td>7809 Sent.</td>
</tr>
<tr>
<td></td>
<td>LOC</td>
<td></td>
<td></td>
<td>(14K words)</td>
<td>(157K words)</td>
</tr>
<tr>
<td></td>
<td>ORG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experimental Setting 2

- **Supervised learning**
  - trained on the entire annotated corpus.
  - Newswire: 408 WSJ articles
  - Biomedical: 590 MEDLINE abstracts

- **Random Selection**
  - a batch of examples is randomly selected in each round

- **F-Measurement**
Experimental Results 1

Effectiveness of Single-Criterion-based Active Learning

- Supervised (223K)
- Random (83K)
- Info-based (52K, 62%, 23%)
Experimental Results 2

- Overall Results of Multi-Criteria-based Active Learning

<table>
<thead>
<tr>
<th>Domain</th>
<th>Class</th>
<th>Supervised</th>
<th>Random</th>
<th>Strategy1</th>
<th>Strategy2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bio</td>
<td>PRT</td>
<td>223K (F=63.3)</td>
<td>83K</td>
<td>40K</td>
<td>31K</td>
</tr>
<tr>
<td>News</td>
<td>PER</td>
<td>157K (F=90.4)</td>
<td>11.5K</td>
<td>4.2K</td>
<td>3.5K</td>
</tr>
<tr>
<td>LOC</td>
<td></td>
<td>157K (F=73.5)</td>
<td>13.6K</td>
<td>3.5K</td>
<td>2.1K</td>
</tr>
<tr>
<td>ORG</td>
<td></td>
<td>157K (F=86.0)</td>
<td>20.2K</td>
<td>9.5K</td>
<td>7.8K</td>
</tr>
</tbody>
</table>
Experimental Results 3

- Effectiveness of Multi-Criteria-based Active Learning
Outline

- Introduction
- SVM-based NER system
- Multiple Criteria for Active Learning
  - Informativeness
  - Representativeness
  - Diversity
- Active Learning Strategies
- Experiments and Results
- Conclusion
**Contribution 1**

- **Multi-Criteria-based active learning**
  - The *first work* -- incorporate the informativeness, representativeness and diversity criteria all together
  - Effective strategies: combine the criteria
    - Strategy 1: Info. + clustering (Rep. & Div.)
    - Strategy 2: Linear interpolation (Info. & Rep.) + pair-wise comparison in a batch (Div.)
  - Outperform single-criterion-based method
    - 60% of training data are required
Contribution 2

- Active learning for NER
  - The first work -- incorporate active learning in NER
  - Various measurements: quantify the criteria
    - Informativeness, Representativeness and Diversity
  - Compare with supervised learning and random selection:

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biomedical</td>
<td>37%</td>
<td>14%</td>
</tr>
<tr>
<td>Newswire</td>
<td>28%</td>
<td>5%</td>
</tr>
</tbody>
</table>
Contribution 3

- General measurements and strategies
  - Measurements: for word sequence
  - Active learning strategy: task independent
  - Can be easily adapted to other NLP tasks
    - Text chunking
    - POS tagging
    - Statistically parsing
    - ...
  - Can be applied to other machine learning approaches
    - Boosting algorithm
    - ...

Future Work

- How to automatically decide the optimal value of these parameters?
  - Batch size K
  - Linear interpolation parameter?

- When to stop the active learning process?
  - the change of support vectors
The End

Thank You!