Semi-Supervised Learning methods for Named Entity Recognition

Arvind R Neelakantan

Advanced Machine Learning

April 30, 2013

Named Entity Recognition (NER)

- Input: This site is conserved among isolates of HIV-2 and the closely related simian immunodeficiency virus .
- Output: This site is conserved among isolates of [Virus: HIV-2] and the closely related [Virus: simian immunodeficiency virus] .
- Problem: State of the art methods based on Conditional Random Fields require thousands of human annotated sentences.
- Idea: Applying unsupervised or semi-supervised approaches on large amounts of unlabeled text available in the web or scientific publications.

Overall Method

- We will be using a collection of more than 110,000 biomedical publications available for free.
- Step1 : Use simple rules to generate a list of candidate phrases.

  For example, use the pattern "the ... virus " to get a noisy list of virus names.

  Noun phrases following "diseases like ... ", "diseases such as ..." to get a list of disease mentions, "the ... algorithm" to get a list of algorithms.
- We obtain a high recall-low precision list of names.

  Output of "the ... virus" is influenza, human immunodeficiency, Epstein-Barr, same, new, whole and so on.

Overall Method (contd.)

- Step2 : gather (spelling,context) pairs from the corpus for every instance of a candidate phrase.

  string=new prev_word_3=the prev_word_2=development prev_word_1=of right_word_1=therapeutic

  right_word_2=strategies right_word_3=to

  string=whole prev_word_3=Next, prev_word_2=we prev_word_1=cultured right_word_1=lung

  right_word_2=homogenates right_word_3=from

  string=Epstein-Barr prev_word_3=of prev_word_2=replication prev_word_1=of right_word_1=virus

  right_word_2=DNA right_word_3=by

  string=human immunodeficiency prev_word_3=bronchoalveolar prev_word_2=lavage prev_word_1=from

  right_word_1=virus-infected right_word_2=patients right_word_3=during

- Step 3: Apply semi-supervised or unsupervised learning algorithms on the above collection.

Approach 1: Co-Training

- Input:(spelling, context) pairs for every instance of a candidate phrase in the corpus, m , ϵ , 10 spelling seed rules, i=1.
- Algorithm:
  - 1. Train a classifier using the spelling features with the current labeled data and label the entire collection.
  - 2. Add i*m labeled examples for which the classifier predicted with confidence $\geq \epsilon$.
  - 3. Train a classifier using the context features with the current labeled data and label the entire collection.
  - 4. Add i*m labeled examples for which the classifier predicted with confidence $\geq \epsilon$. Set i=i+1, goto step 1.

Co-Training (contd.)

- Advantages:
  - Has nice sample complexity results from learning theory.
  - The performance of co-training in identifying names of people, location etc. using few seed rules was comparable to sequence taggers which use thousands of human annotated sentences.
- Disadvantages:
  - Strong independence assumption (i.e.) the spelling and context of an instance are independent given the label. This independence assumption is violated in natural language.
  - The number of labeled examples to be added at every iteration and the confidence threshold need to be heuristically set. The relation between these parameters and the performance of the algorithm are not theoretically understood.

Approach 2: Using Canonical Correlation Analysis (CCA)

- Input: (spelling, context) pairs for every instance of a candidate phrase in the corpus, k
- Method:
  - 1. Define Feature Mappings $\Phi_1(spelling) - > \{0,1\}^{d_1}$, $\Phi_2(context) - > \{0,1\}^{d_2}$.
  - 2. Obtain matrix representations L_{n*d_1}, R_{n*d_2} for spelling and context views using the feature definitions.
  - 3. Get Projections Matrices Θ_1, Θ_2 by solving $max_{\Theta_1,\Theta_2}\ Correlation(L\Theta_1, R\Theta_2)$ where Θ_1, Θ_2 are projection matrices of dimension $d1 * k$ and $d2 * k$ respectively.
  - 4. Output the k-dimensional real valued representation of candidate phrases, $L\Theta_1$.

Using CCA (contd.)

- Use the CCA projections to train a SVM with 10 labeled examples.
- Advantages:
  - CCA is more stable than other dimensionality reduction techniques like PCA since CCA is scale invariant.
  - No independence assumptions between spelling and context views.
  - The only parameter to fit is SVM's regularizer which is much easier to handle compared to the parameters in co-training.

Thank You!