Homework 4 - Structured Learning

Manasa Bollavaram

March 19, 2017

1 Feature Templates

Feature Templates used for implementing the structured perceptron classifier are as follows:
w_0
$w_{-1} + w_0 + w_1$
$w_0 + \text{POS}$
Features with occurences less than 2 are not taken into account.
Additional feature templates which would improve NER performance are:
$w_0 + \text{suffix and prefix}$
$w_0 + s_0$
$w_0 + s_{-1} + s_0$
$w_{-1} + w_0$
$w_0 + w_1$
$w_0 + \text{CHUNK}$
$w_0 + \text{POS} + \text{CHUNK}$
where s is shape of the word.

.

2 Viterbi Decoding Algorithm

Initialization
$\pi[0, s] = \mathbf{w}.\phi(x, 0, s, s_{-1})$
where $s_{-1} = \text{START}$
Recursion
for $j = 1...m-1, s = 0...m-1$
$\pi[j, s] = \max\limits_{s_t}[\pi[j-1, s_t] + \mathbf{w}.\phi(x, j, s, s_t)]$
where m is the length of the sequence$\mathbf{x}$, $s_{-1} = \text{START}$, $\phi(...)$ is the local feature vector, s is the NER tag
In conclusion, $\max\limits_{s_0,\ldots,s_{m-1}} \sum_{j=0}^{m-1} \mathbf{w}.\phi(\mathbf{x}, j, s_j, s_{j-1}) = \max\limits_{s} \pi[m-1, s]$

3 Ablation Study

Training the dataset is taking about an hour for each iteration. Could not train it with the features in the limited time. But, the framework and code to run the features and include feature variations is all set up and submitted in the code file.

4 Pros and Cons of BIO and IO encoding schemes for NER

IO encoding has no boundary tags and hence if an NER object has two adjacent words, IO encoding fails to identify as one entity and would tag both of them as different entities. BIO encoding as advantages in this case because of its boundary tags. Such case can also lead to better performance for IO encoding because of more training data available for a certain tag compared to BIO encoding.

5 Discussion

Implemented code for NER tagging and took the ideas of feature templates from the reference paper provided in the homework. Unfortunately, due to time constraint, could not complete the run of code and get results.