Relation Extraction

CSE 628 Assignment 2

Due Date: Apr 31st 2015 at 11:59 pm.

In this assignment you will build a relation extractor to identify institution relation instances from Wikipedia sentences. An institution relation institution(x, y) indicates that a person x studied in the institution y. You will be given training and test sentences that are annotated for the relation instances.

Your task is to build a suite of relation extractors as described below and investigate the performance of each set of features.

1) ***Regular Expression Patterns*** – Write **five** regular expression patterns to identify "" relations. Use the training set to develop your patterns and report the performance on the test set. **(20 points)**

2) ***Bag-of-words*** – Every word that occurs in the span between the PERSON mention and the INSTITUTION mention is used as a feature. Veronica has created a simple script that takes in the training and test files to generate BOW feature vectors for each labeled instance. (See generate_arff.py) **(20 points)**

3) ***Clustering Features*** – Using words directly as features often leads to sparsity. A standard technique to overcome sparsity is to use a generalized representation of words via clustering. The idea is to first cluster words based on their usage and represent each word by the cluster to which it belongs. You can use Percy Liang's code here to generate hierarchical word clusters. The algorithm takes in a text file as input and a cluster parameter (c) that controls the number of clusters. The output includes a path file that contains the cluster ID (which is a path on the hierarchical (binary) clustering). Here is a section of the Brown clusters:

Cluster ID	Word	Frequency
11110101	teaches	48
11110101	taught	430
11110101	lectured	9
11110101	speaks	14
11110101	read	80
11110101	practised	18
11110101	practiced	49
11110101	**studied**	1394
11110101	**matriculated**	48
11110101	**enrolled**	132

11110101	flew	9

For each word use its cluster ID as the word id (as opposed to unique feature ID). **(20 points)**

Words that have similar meaning and usages are grouped together (shown in Blue). This is very helpful for relation extraction. However, the clustering is far from perfect and may not care about the similarity dimensions for the task at hand. For instance, *teaches* and *studied* are included in the same cluster. While "studied" is positive evidence for the Institution relation, it may not be evidence

4) ***Dependency Features*** – Mere presence of words that indicate relations isn't adequate.

   e.g.,

   > Mario was born to Roger, **an alumnus of** Harvard, and to Mary **an alumnus of** Yale.

   Even though alumnus is a strong indicator of the studied in relation, the syntactic structure clearly shows that alumnus is not directly related to Mario.

   Process the sentences using Stanford Dependency parser. Code **three** syntactic dependency-based features that ensure that the relation words are connected to the entity mentions.

   (30 points)

5) ***Kitchen Sink*****:** Combine all features including the manual regular expression patterns.

   (10 points)

Classifier

Use a SVM classifier. There is an implementation in Weka (see here). There are also other implementations including libLinear, libSVM and SVMLight.

I recommend using either libLinear or the implementation in Weka with the linear kernel. The SVM's typically require you to specify a regularization parameter (typically called 'C'). You can set this to any fixed value. The goal of this assignment is for you to see how adding the different feature sets affect performance. The ML magic required to get a high performance is not the focus.

Optional for the ML interested folks:

You can try a few values in the range C = {0.001, 0.01, 0.1, 1, 10, 100} on the training dataset to figure out the value that gives you the best performance.

Evaluation

1. For each group of features report the Precision/Recall and F1 on the institution class.

2. For each group of features find the most egregious errors – i.e., instances where there isn't a relation but the classifier assigns a high score to the relation.

   Provide an explanation for why the classifier made these mistakes. You can inspect the features used to form ideas.