Final Project

Zac Branson Jay Kaiser Mike Czerniakowski

April 14, 2016

Introduction

Classifier

- We are using TIMBL because:
- It lets us feed it strings: this makes maintaining word order of the immediate contexts much easier.
- We expect a memory based classifier to perform well for this task because: why?

Preprocessing

Bare Words

- Remove all punctuation that doesn't indicate the end of a sentence (i.e. everything but ". ? !")
- make all capital letters lowercase

Lemmas

- Do the same as above but:
- Use NLTK's Romanian lemmatizer to lemmatize all words.
- Given that Romanian is a highly inflectional language, this might have a profound effect (good and/or bad).
- Good: greatly reduces sparsity of many words. Bad: some of those suffixes might have been useful (case?, number?)
- We'll test that empirically.

Features

bare Words and Lemmas

- Same context features from HW3: add here (lemmas instead of plain words in the case of lemmas)
- Second task is to incorporate PoS features: add tags of the same context words as independent features
- Possibly try it with just PoS tags and no words

PoS Tags

- add tags of the context words as independent features
- possibly try it with just PoS tags and no words
- possibly try it with lemmas instead of words

Parameters

TIMBL Options

- Use a bash script to brute force parameter optimization
- We optimize these parameters: insert

Adding Unlabeled Data

Bootstrapping Approach

- TIMBL outputs distances, which can be used as a measure of prediction confidence
- Take the smallest n/10 of distances per class, add them to the training data, retrain the model
- Iterate:adding the next n/10 smallest distances, potentially until there are none left.
- Profit.

Preliminary Results

Bare Words

table of results

Lemmas

table of results

Words + PoS Tags

table of results