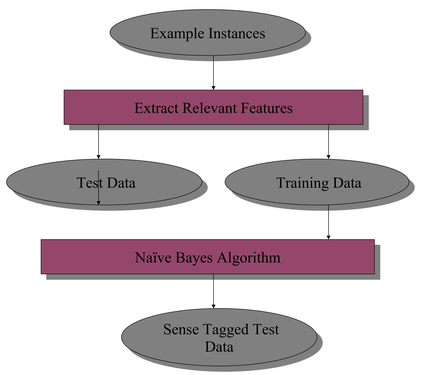
Using Domain Specific Information for Word Sense Disambiguation

Bridget T McInnes[1], Ted Pedersen[2] and John Carlis[1]
University of Minnesota Twin Cities[1] and University of Minnesota Duluth[2]

Background and Introduction

Word Sense Disambiguation is the problem of determining the appropriate sense of a word that has multiple senses. This is a problem for domain specific applications. We explore the question of whether domain specific knowledge sources can be used to help identify the appropriate sense of a word. To do this, we compare the use of biomedical specific features extracted from the Unified Medical Language System with more general English features. We evaluate the features using a supervised learning approach and compare their performance on biomedical and general English text. The package used to conduct these experiments is CuiTools version 0.13 which can be found at http://cuitools.sourceforge.net.

Algorithm



Types of Features

Domain Specific (Biomedical) Feature Set:

- Concept Unique Identifiers (CUI)
- Semantic Types
- Semantic Relation

General English Feature Set:

- Unigrams
- Part-of-speech Information
- Head Information

EXAMPLE: Disambiguating mole using two different feature sets

Instance: He used one **mole** of ethylene oxide in solution to conduct his experiment

Extract Biomedical Features of example instance

C1524062 (CUI: Using)
C0205447 (CUI: One)
C0015087 (CUI: Ethylene Oxide)
C0681814 (CUI: Experiment research)
function (semantic type)
part-of (semantic relation)
....

Extract General English Features of example instance

used (unigram)
one (unigram)
ethylene (unigram)
oxide (unigram)
Noun (part of speech)
Yes (head word)
...

FEATURE	FREQ
C1524062	2
C0205447	1
C0015087	5
C0681814	3
function	0
part-of	1
....	

Search training data for frequency of feature occurring with target word discarding:

a CUI or semantic type feature if it occurs less than two times surrounding the target word for the biomedical feature set

a unigram if it occurs less than three times surrounding the target word for the general English feature set or is a function word.

FEATURE	FREQ
used	1
one	2
ethylene	10
oxide	3
Noun	
yes	
....	

Create a vector of relevant features for the instance

C1524062 C0015087 C0681814 part-of ....

ethylene oxide Noun yes ...

Naïve Bayes Algorithm trained on relevant biomedical features from the training data

Input vector as test data into the Naïve Bayes algorithm trained on the relevant features from the training data.

Naïve Bayes Algorithm trained on relevant general English features from the training data

Mole, unit of measure

The sense is then compared against a gold standard compiled by a human to determine the accuracy of the algorithm

Mole, unit of measure

Data

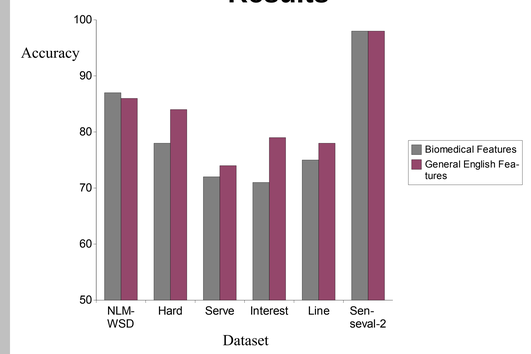
Domain Specific (Biomedical) Dataset:

- National Library of Medicine WSD dataset

General English Dataset:

- Hard, Serve, Interest, Line datasets
- Senseval2 dataset

Results



Conclusions

- Biomedical features performed better than general English features on the biomedical (NLM-WSD) test set
- Biomedical features did not perform as well as general English features on the general English test sets
- Biomedical information is more difficult to obtain general English which brings to question does the increase in accuracy using biomedical features over general English features worth the additional overhead
- The biomedical features provide enough information about general English concepts to accurately differentiate between senses of a specific word