Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

1.1 introduction
1.2 Learning under bias

茨城大学理工学研究科
國井慎也

Introduction

- Many tasks are present in natural language processing(NLP).
  - Documents summarization
  - Documents classification
  - Sentences parsing
  - Word sense disambiguation

  – Language need to be represented in a compact, meaningful way in NLP.

bag-of-words : using arrays of numbers(0 and 1).

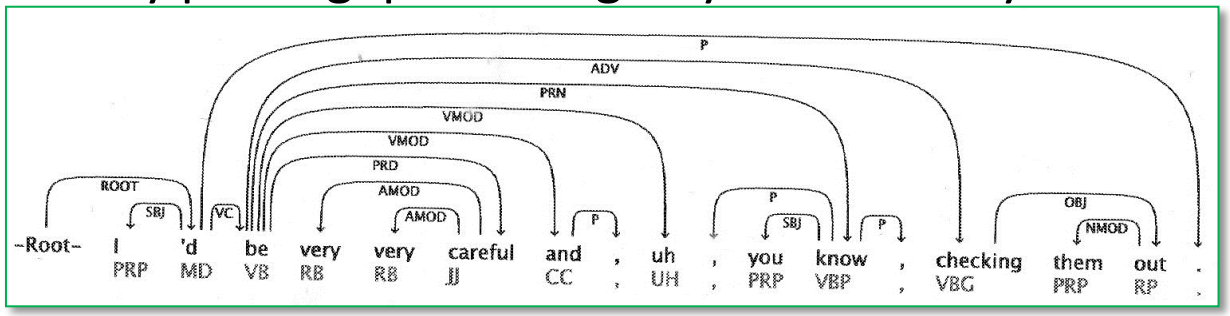
e.g. McCain just gave a cheap plug to Ed Kennedy.

attributes <Obama, McCain, Malcolm, Many , Ed Kennedy>

The above text could be <0,1,0,0,1> in bag-of-words.

Importance of classification and clustering algorithms

- Dependency parsing: predicting a syntactic analysis



- If we think of trees as classes(patterns), and don't use classification and clustering algorithms for dependency parsing, the number of classes would be enormous.

  A sentence with 17 token : $17^{15} \sim 3 * 10^{18}$ classes

It's necessary to combine classification or clustering algorithms with search or parsing algorithms.

Classification and clustering algorithms are important methods and have massive applications in NLP.

Learning under bias

- problems in supervised documents classification
  - The labeled data available is scarce and bias.
  - This holds for the other learning problems too.
  - In NLP dataset with n data points and m features, n << m.

  ⇒ The labeled data is sparse, and vulnerable to data bias.

Enron corpus for spam detection

The dataset was written by computer scientists or employees about work-related matters.

Polarity detection dataset from IMDs and Amazon

A limited set of categories

The Wall Street Journal for POS and dependency parsing

The models made from the Wall Street Journal will be heavily biased when applied to real-world data.

Solutions to bias

- (a)Replace our preferred algorithm with an algorithm that underfits data.

  For example, the algorithm only learns relatively simple, superficial correlations from data, believed to be common to the available data(source) and the data I wish to process(target).

- (b)Use semi-supervised leaning to learn from labeled source and unlabeled target domain.

- (c)Use only some of the data, some of the features, or some of the model parameters in the target data.

  The limited data will be specific to the source data.