Web People Search Task(WePS)

Advaned NLP Project 2

Fan Kai , Wang Chen
10848194 10848263

Peking University

2009.05.12

Major Tasks

- Extract features from web pages.
- Calculate similarity between page pairs.
- Cluster similar web pages.

Thirdparty Tools and Environment

- Jericho HTML Parser
- Stanford Named Entity Recognizer
- Natural Lauguage Toolkit(NLTK)
  - Sentence Tokenizor
  - Word Tokenizor
  - Stop Words
  - Portor Stemmer
- Weka 3: Data Mining Software in Java
- Python 2.5
- Ubuntu 8.04

Text Features

- Full Text Plain text extracted from web page.
- Text Summary Sentences where names occurs.
- Metadata Combination of title, snippet and url.

Named Entity Features

- Person
- Location
- Organization

Other Features

- Link and URL
- Email Address
- Domain
- Number
- Telephone Number
- Year

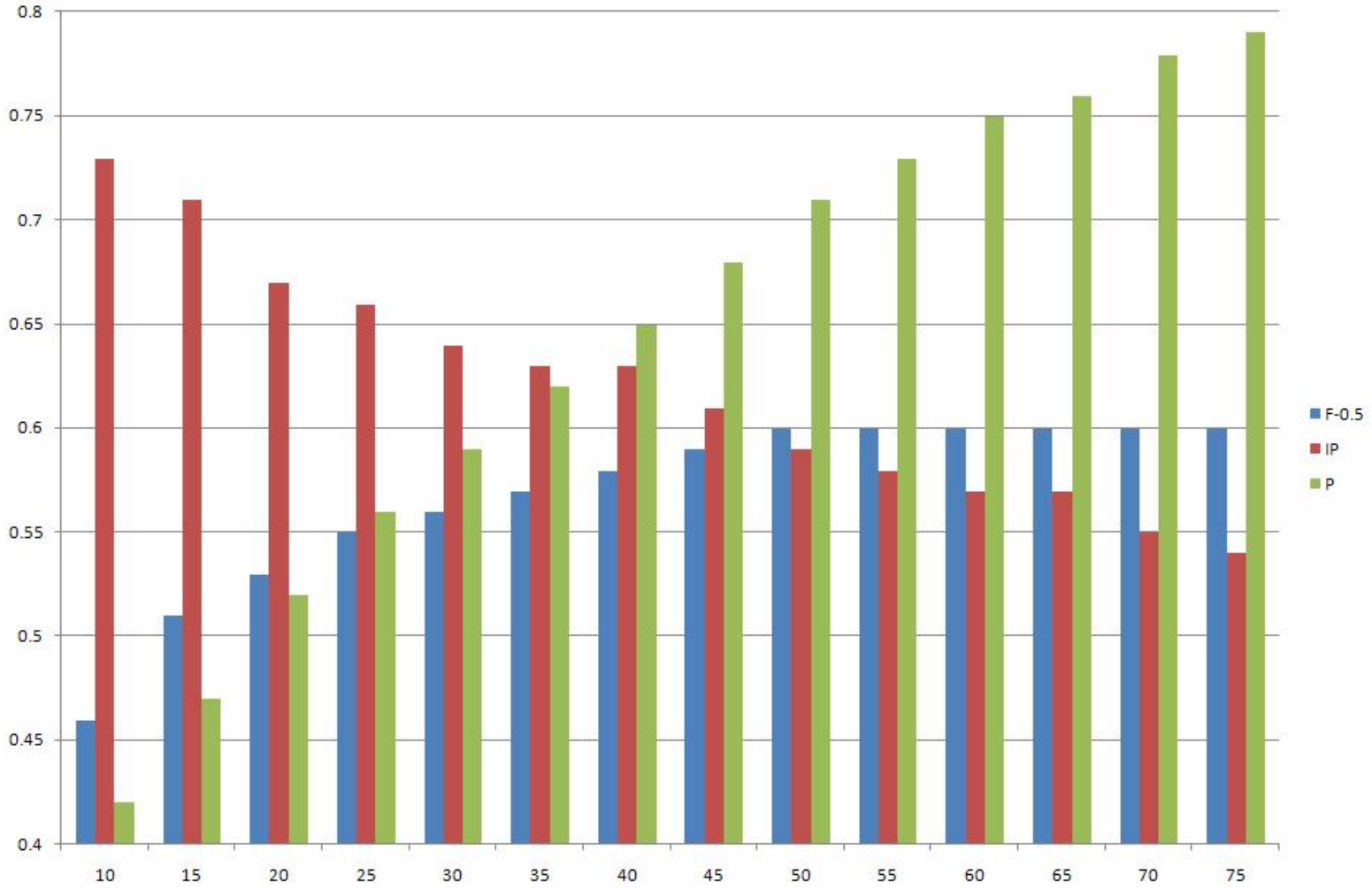
Similarity Calculation

- Vector Space Model
  - Calculate VSM similarity for each feature.
- Machine Learning
  - Learning Features: VSM similarity of each feature
  - Learning Target: Probability two page indicate a single entity, used as similarity
  - Methods
    - Maximum Entropy
    - Naive Bayes
    - Decision Tree
  - Feature Selection
    - All Features
    - Text and Entity Features

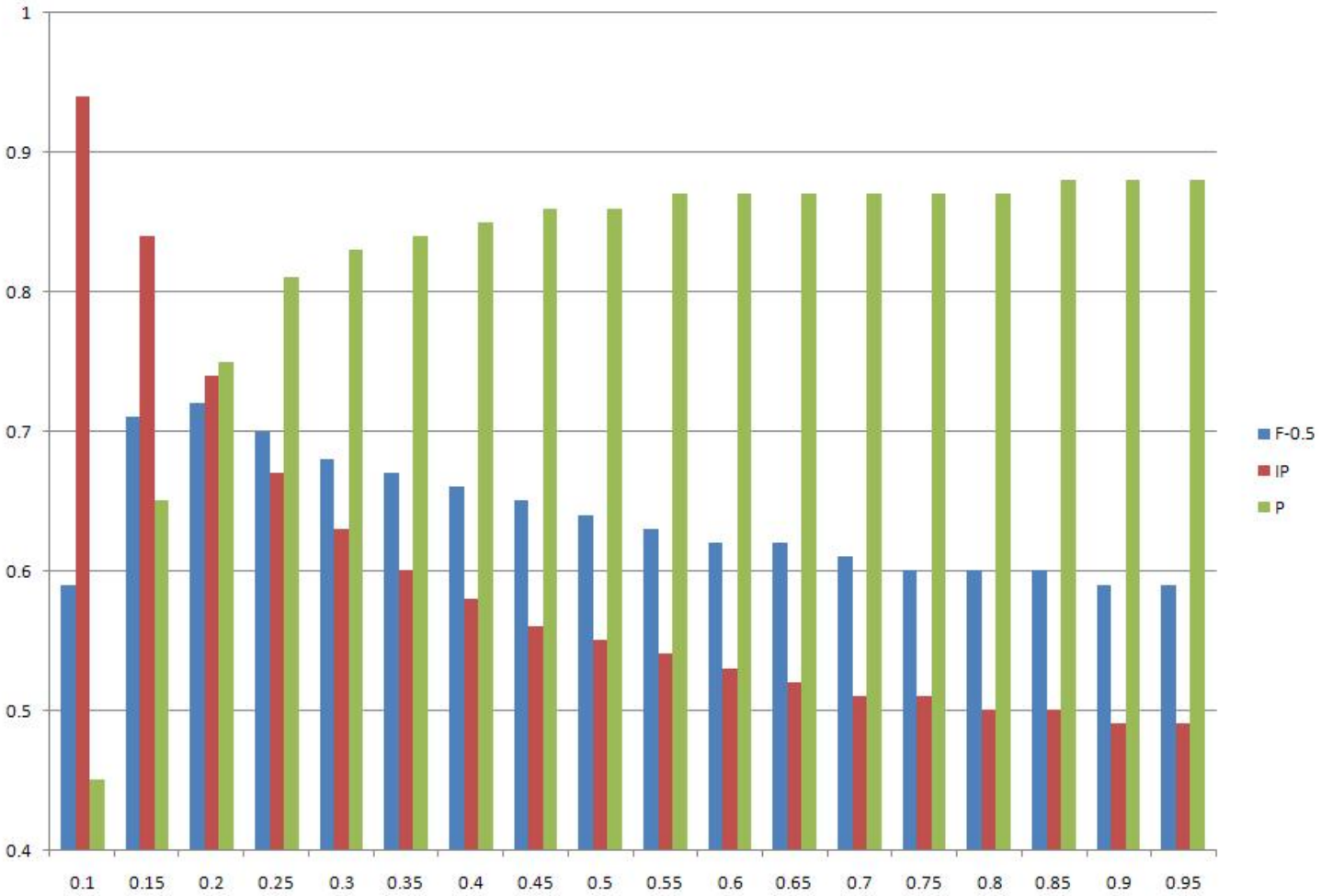
Clustering

- Hierarchical Agglomerative Clustering
  - fixed thresholds
  - try different thresholds
  - try different merge strategy
- K-Means
  - fixed cluster number
  - try different cluster numbers

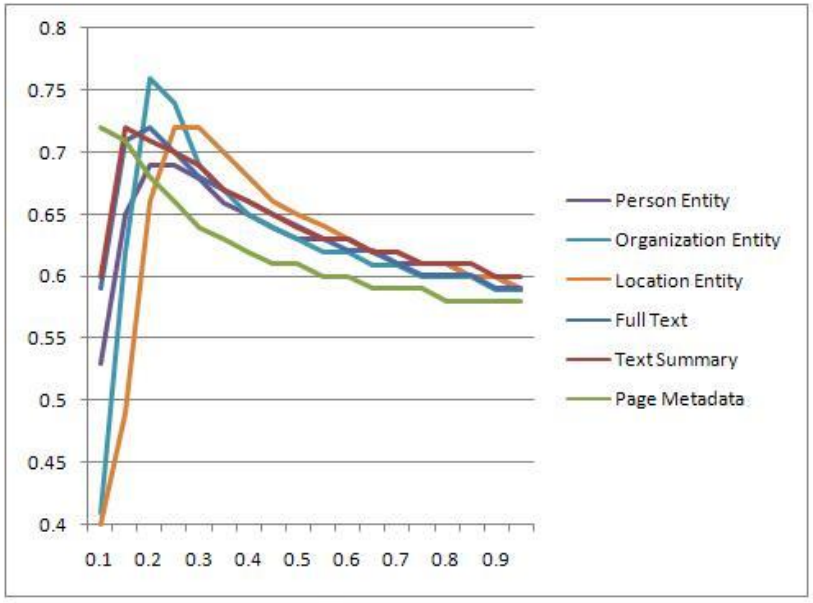
Cluster Fulltext feature using K-Means(Test Set)

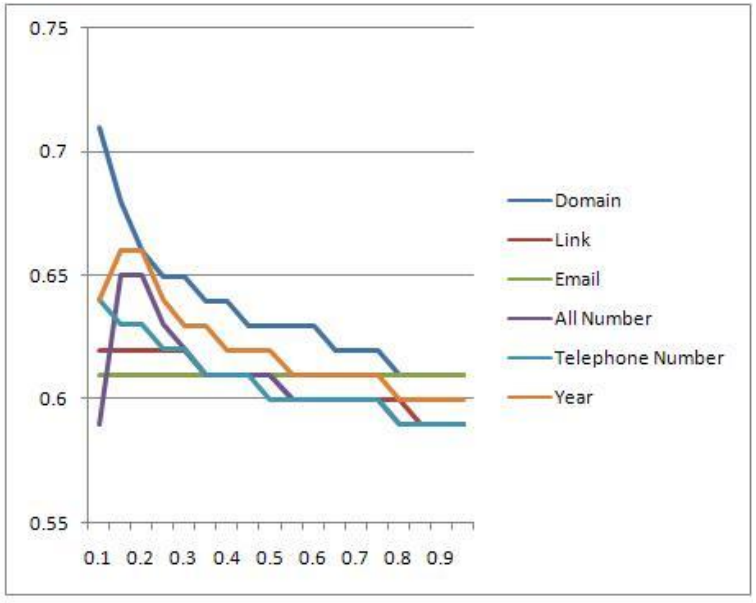


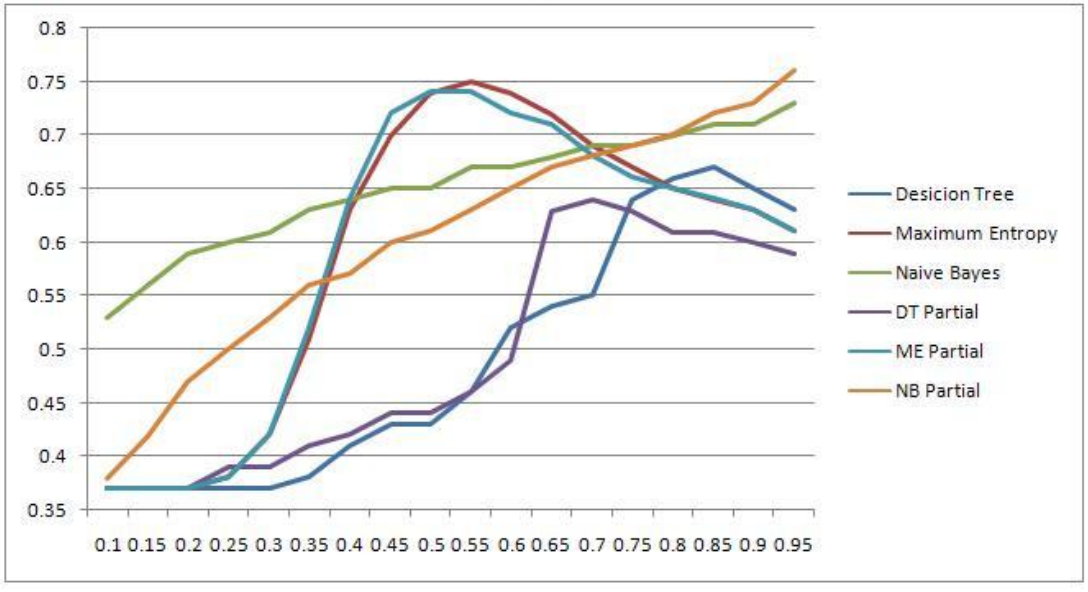
Cluster Fulltext feature using HAC(Test Set)



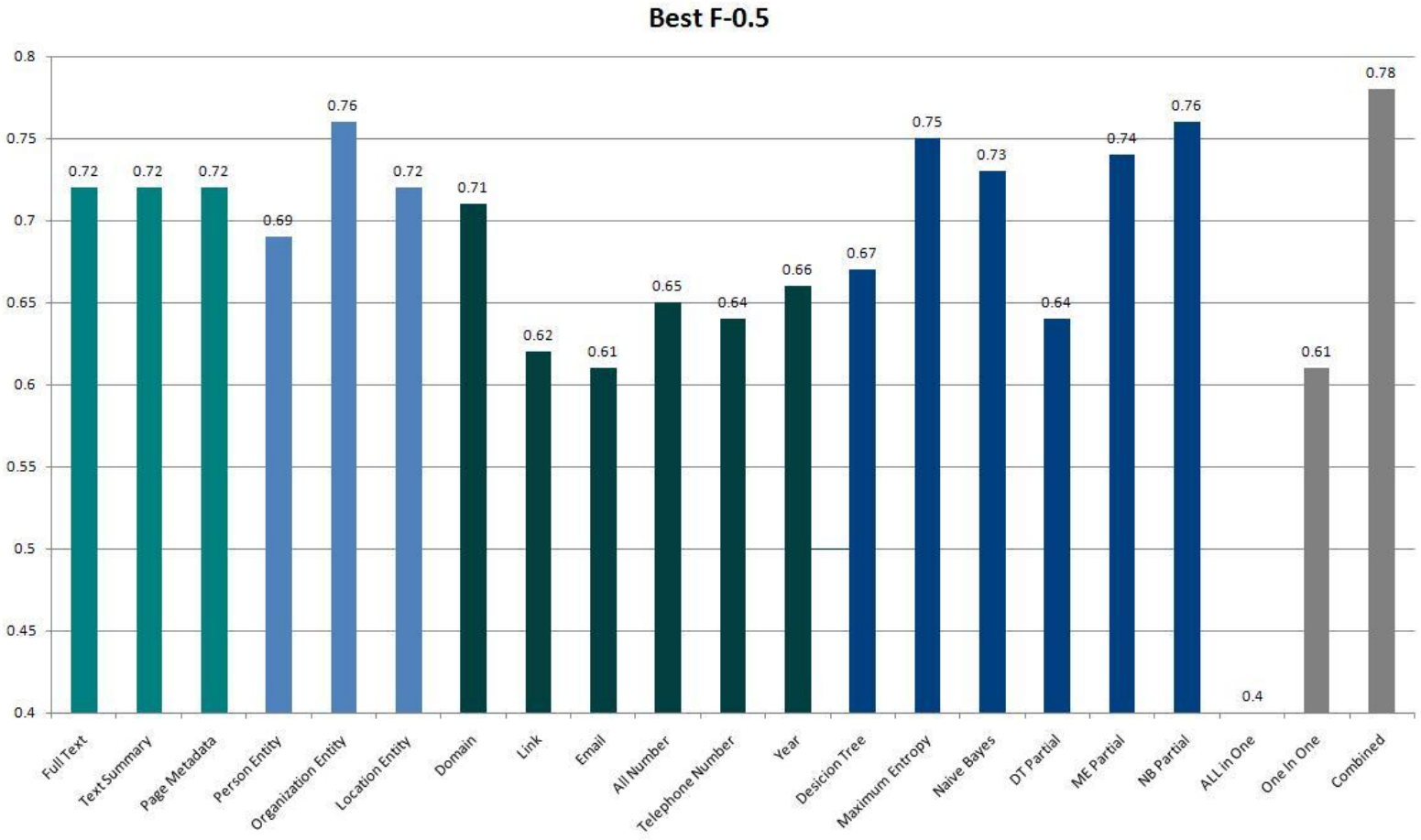
FMeasure-0.5 with different HAC threshold(Test Set)



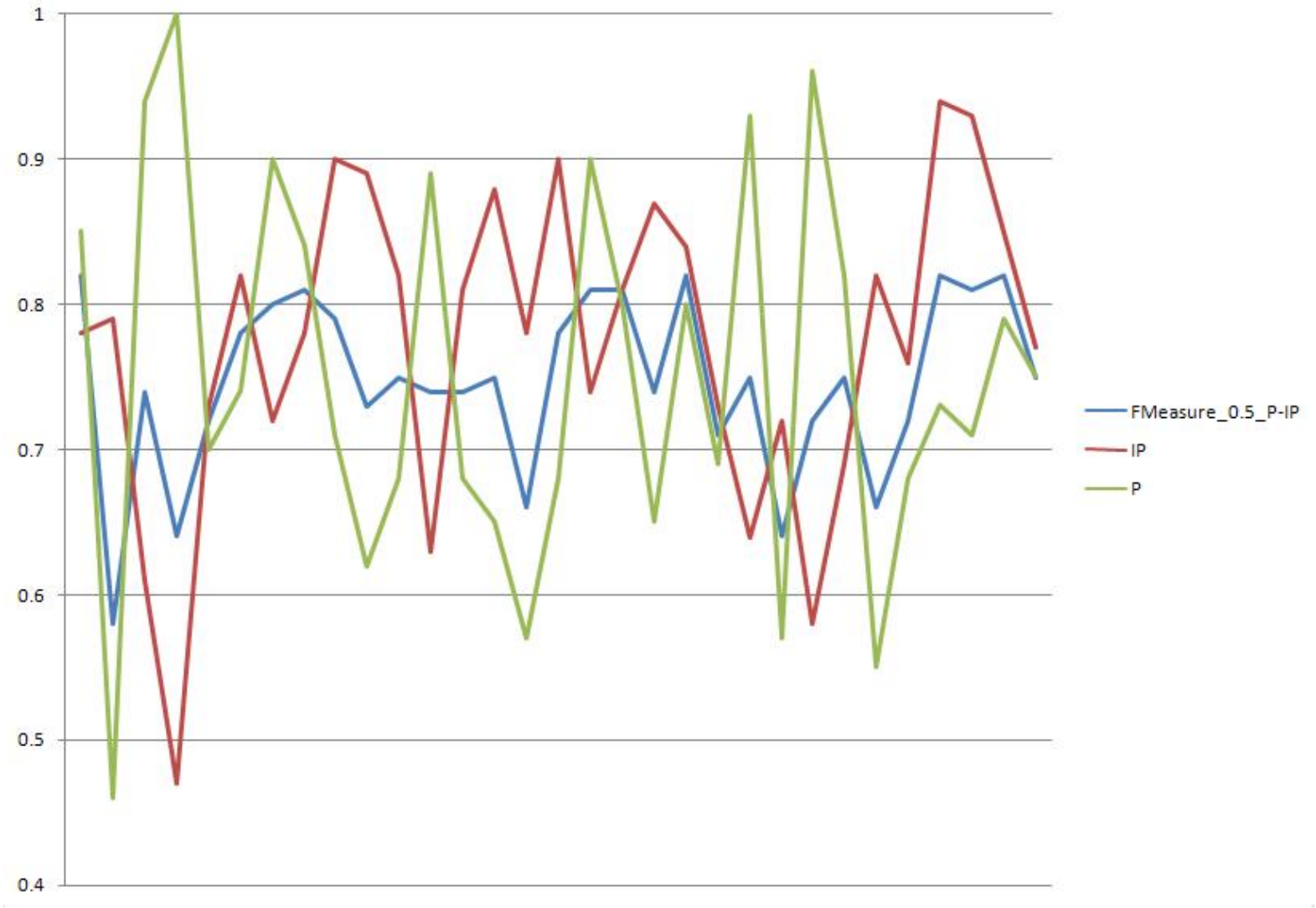




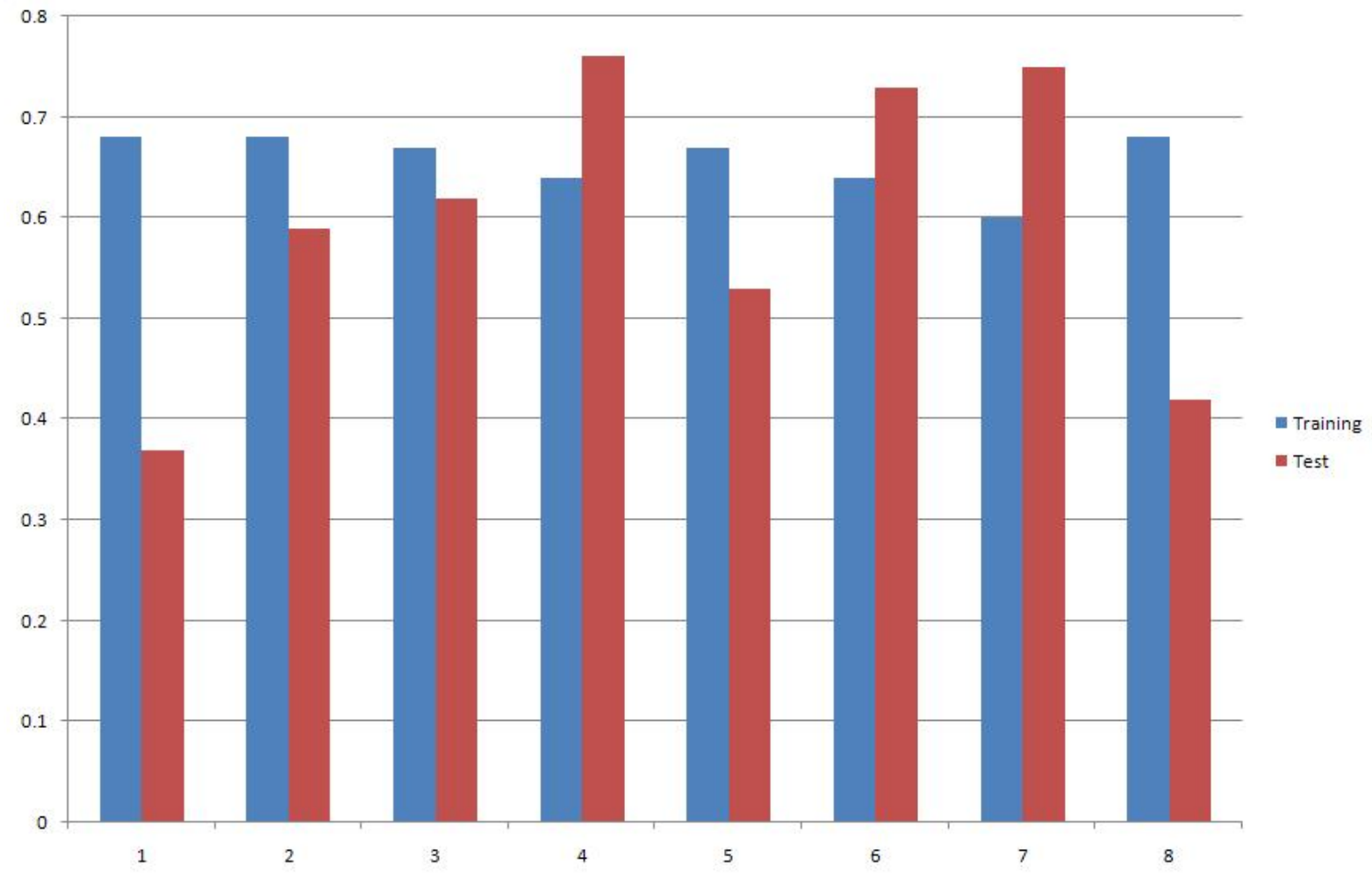
Best FMeasure-0.5(Test Set)



Scores of Different Entities in Test Set(ME-0.55)



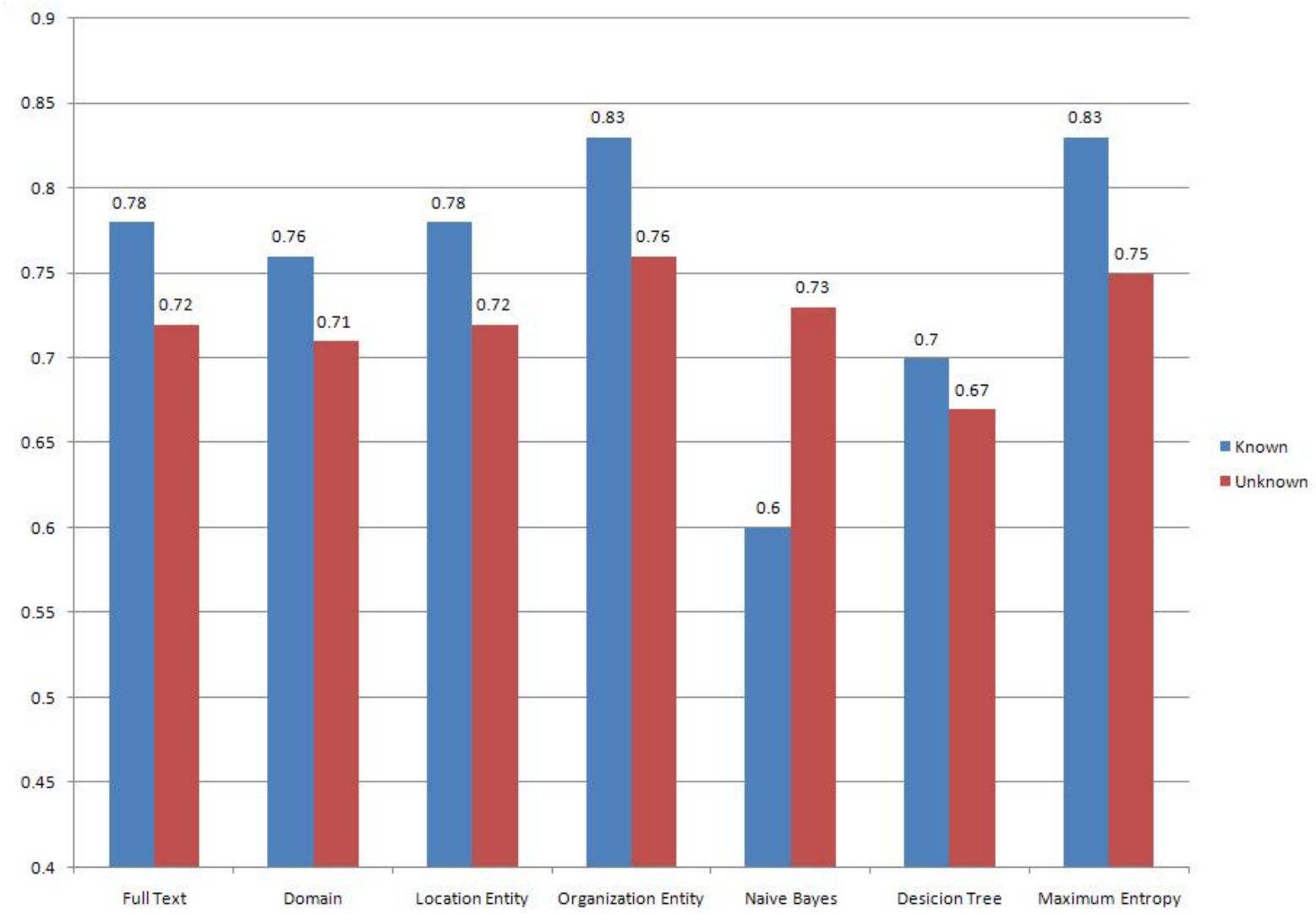
Comparison between Traing and Test Set



Other Issues

- Unknown character encodings.
- How to recognize discarded items?
  - Couldn't find efficient judgement.
  - Current condition: summary contains less than 5 words.

How Discarded Items Matter(Test Set)



Best Scores

- Using best result from traing set(fulltext-0.10):

Set	**F-0.5**	**IP**	**P**
Training	0.68	0.81	0.67
Test	0.59	0.94	0.45

- Discarded items unknown:

Similarity Type	**Threshold**	**F-0.5**	**IP**	**P**
Organization Entity	0.20	0.76	0.83	0.73
Naive Bayes Partial	0.95	0.76	0.84	0.71
Maximum Entropy	0.55	0.75	0.77	0.75

- Discarded items known:

Similarity Type	**Threshold**	**F-0.5**	**IP**	**P**
Organization Entity	0.20	0.83	0.83	0.84
Maximum Entropy	0.95	0.83	0.82	0.86

Future Work

- More features
- More sophisticated feature selection in ML
- Variable HAC threadhold or K-Means cluster number

Thanks!