Text Classification

Mehrdad Bozorg
Student ID: 2911337

universitätbonn

Summery of research

- Topic
  - Hybrid Class Semantics Classifier (HCSC)

- Problem Definition
  - Given: Set of Documents
    - Consists of labeled (D^L) and unlabeled (D^U)
    - Predefined Classes

  - Goal: Finding a classifier to assign label to unlabeled documents

Main challenges

- Sparse data
  - Too many features for each document
    - Decrease efficiency
    - Decrease accuracy
- Semantic mismatching between documents
  - Word with different meaning in different contexts
  - Documents with different word sets, but the same concepts
- Lack of labeled data
  - Affects generality and accuracy

Proposed solution

- Preprocessing
  - To solve sparsity of data

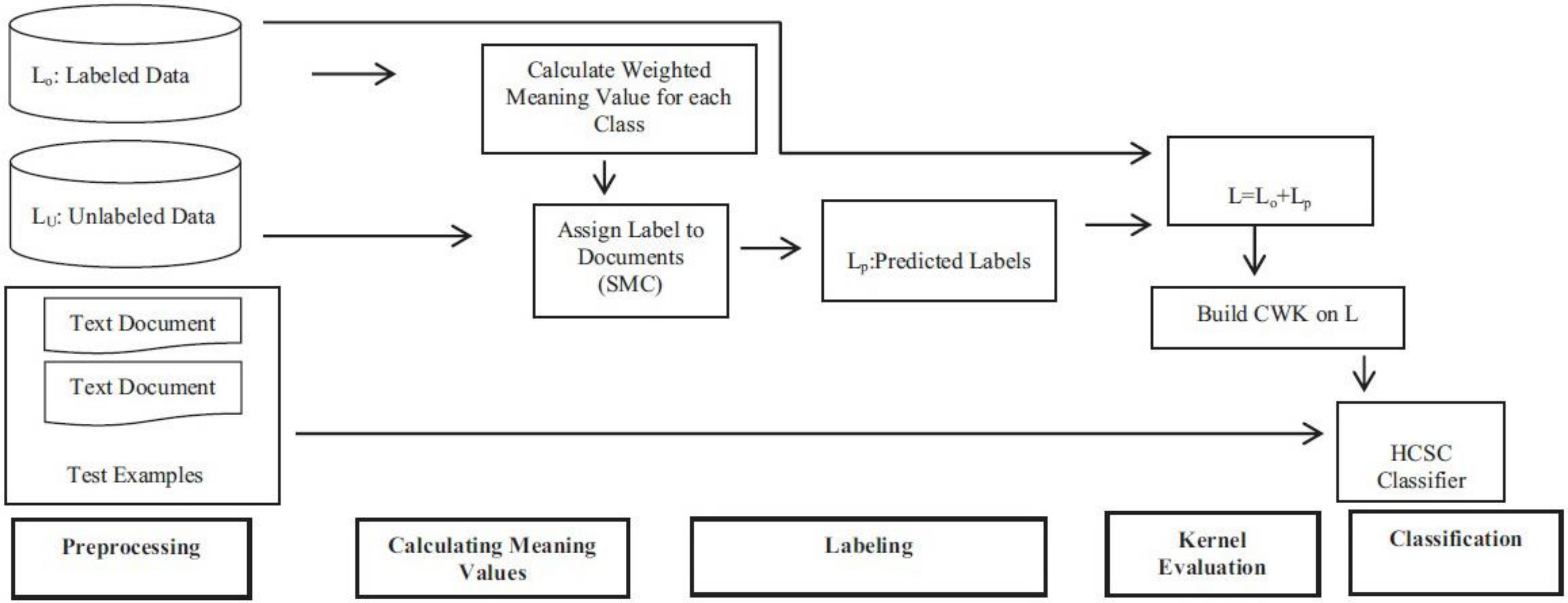
- Semi-supervised learning classification
  - To solve lack of labeled data

- Semantic Kernel Method
  - To solve semantic mismatching

Main steps of algorithm

- Preprocessing
- Calculation of Meaning Values
- Labeling
- Kernel Evaluation
- Classification

Algorithm road map



L_o: Labeled Data

L_U: Unlabeled Data

Text Document

Text Document

Test Examples

Calculate Weighted Meaning Value for each Class

Assign Label to Documents (SMC)

L_p:Predicted Labels

$L=L_o+L_p$

Build CWK on L

HCSC Classifier

Preprocessing

Calculating Meaning Values

Labeling

Kernel Evaluation

Classification

Preprocessing

- Stop-words Elimination
  - nltk.corpus is used

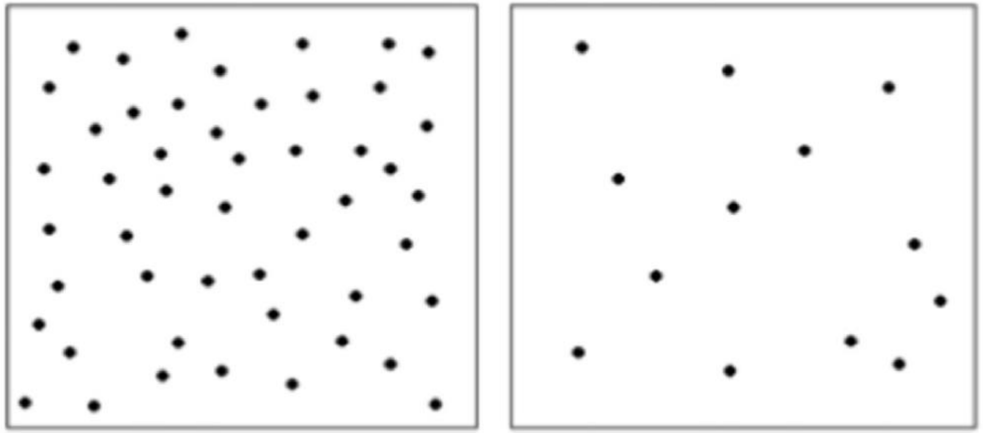
- Stemming
  - nltk.stem is used

- Infrequent words elimination
  - Threshold 3

- Feature Selection
  - Information Gain
  - 200 words selected

Calculation of meaning values

- Semantic kernel method
  - Consider the importance of words in each class
    - Intensify the importance of core words for each class
    - Decrease the importance of general words for each class
- Class Meaning Kernel (CMK)
  - Use Helmholtz principle in Gestalt theory



CMK main steps

- Calculate Number of False Alarms (NFA) of events for words
  - Co-occurrence of *m-tuples* of words together with word *w* in the same document
  - (NFA < 1) indicates meaningless words

- Compute Meaningfulness Matrix M

- $\forall d \in D^U$: $d.M$ indicates meaningfulness of d in each class

- $Argmax(d.M)$ indicates label of d

Kernel evaluation

- Class Weighting Kernel (CWK)
  - Similar to TFIDF Method
  - Compute weight of each word in each class → $Matrix\ W$
  - $d_1 W W^T d_2$ calculates similarity between d_1 and d_2

Classification

- Supervised learning algorithm
- Support Vector Machine Classifier
  - sklearn in python (Nusvc and SVC)
  - One-against-one method

Experiment environment

System Configuration	
OS	Windows 7
CPU	1.8 GHz Intel core i3
RAM	4 GB

- Programming environment: Python 3.5

Data set description

Number of train set items	8158
Number of labeled data items	817
Length of train set	1,883,279 words
Number of unlabeled data items	7342
Number of test Set items	6960
Number of Classes	15
Data distribution	Equal in classes

Testing and empirical results

Proportion of labeled over unlabeled data	1/9
Proportion of Train set over Test set	8/7
Expected precision	0.9
My best precision	0.5

- $\frac{\#Correct\ assigned\ label}{\#Test\ set}$

Future works

- Designing a new algorithm by combining different method in each step of text classification
- Investigation of graph-based document representation

References

[1] B. Altınel, M. C. Ganiz, A new hybrid semi-supervised algorithm for text classification with class-based semantics, J. Knowledge-Based Systems 108 (2016) 50–64.

[2] B. Altınel, M. C. Ganiz, B. Diri, A corpus-based semantic kernel for text classification by using meaning values of terms, J. Engineering Applications of Artificial Intelligence43(2015)54–66.

[3] B. Altınel, B. Diri, M. C. Ganiz, A novel semantic smoothing kernel for text classification with class-based weighting, J. Knowledge-Based Systems 89 (2015) 265–277.

[4] A. K. Uysal, An improved global feature selection scheme for text classification, J. Expert Systems with Applications 43 (2016) 82–92.