WUST at NTCIR-10 RITE-2 Task：Multiple Feature Approach to Chinese Textual Entailment

Maofu Liu, Yue Wang, Yan Li, Huijun Hu
College of Computer Science and Technology, Wuhan University of Science and Technology
liumaofu@wust.edu.cn, wycx121@126.com

Introduction

◆RITE is a generic benchmark task that addresses major text understanding needed in various NLP/Information Access research areas.

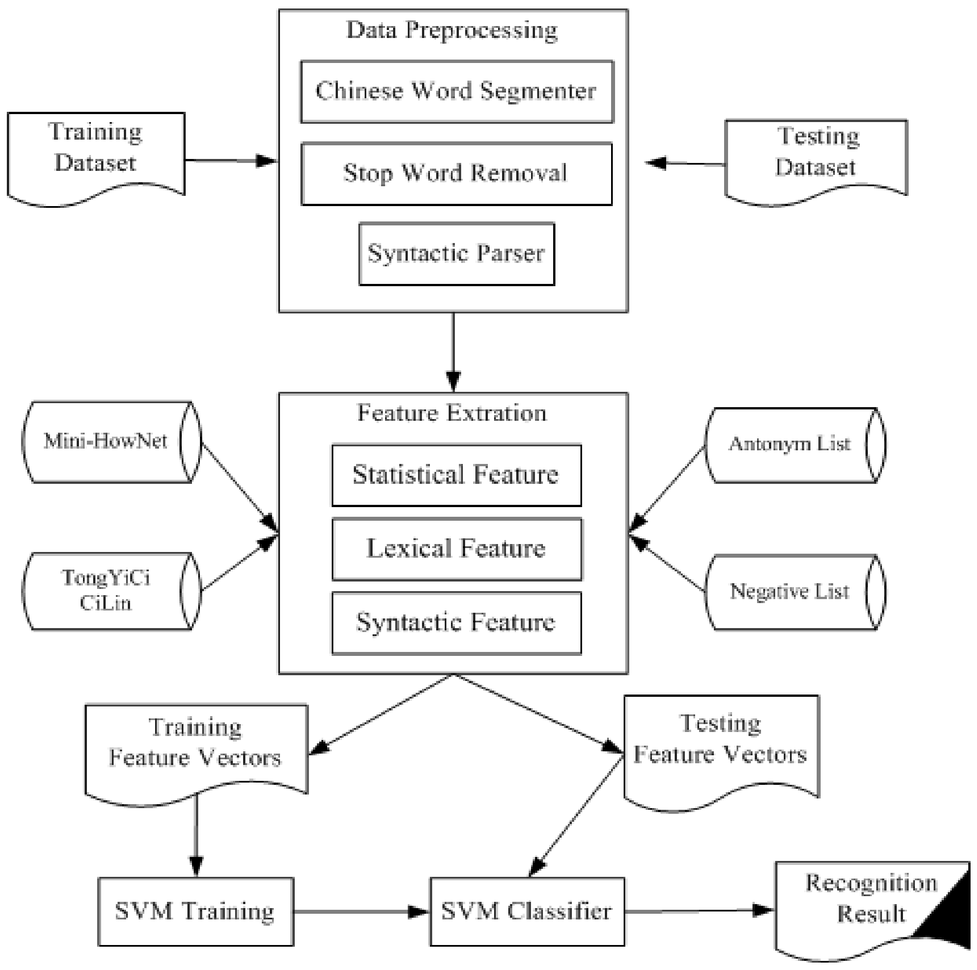
◆We construct the classification model based on support vector machine to recognize semantic inference in Chinese text pair, including entailment and non-entailment for BC subtask and forward entailment, bidirectional entailment, contradiction and independence for MC subtask.

◆The key of the system is the features we choose. We use multiple features including statistical feature, lexical feature and syntactic feature.

◆We choose LIBSVM as the classifier. LIBSVM is a library for support vector classification (SVM) and regression. After preparing and scaling data set in LIBSVM form, our system chooses the RBF kernel function to do the cross-validation.

➢Our system includes three main parts, preprocessing, feature extraction and SVM Classifier.

System Architecture



1. Preprocessing

◆In the data preprocessing, the system mainly implements the Chinese word segmentation and removes the stop words according to the stop word list.

◆We choose Stanford Chinese word segmenter with PKU and stop words list of Harbin Institute of Technology.

2. Feature extraction

(1)Statistical feature

◆Word overlap

◆Length difference

◆Manhattan distance

◆Euclidean distance

◆Jaro-Winkler distance

◆LCS similarity

◆Same words ratio in shorter sentence

(2) Lexical semantic feature

◆Hownet based similarity

◆TongYiCi CiLin based similarity

◆Antonym

◆Negative

(3)Syntactic feature

◆Syntree similarity

3. SVM classifier

◆We choose LIBSVM as the classifier. LIBSVM is a library for support vector classification (SVM) and regression. After preparing and scaling data set in LIBSVM form, our system chooses the RBF kernel function to do the cross-validation.

Experiments

◆The official evaluation results of performance are listed in the Table 1. There is only one type of assessment, automatic assessment by accuracy.

Table1. Formal run experiment official results

Run	Subtask	Accuracy
WUST-CS-BC-01	BC	0.588
WUST-CS-MC-01	MC	0.524

◆In BC subtask, we use the same features with MC subtask. And we have not experiment other features because we have no time. But BC and MC are different, we think if we choose some different features to BC, the accuracy of BC would be higher.

◆In MC subtask , according to the result, we think that the most influence factors of accuracy are the judgment of the C and I mistake. If we consider more features to expand the boundary of C and I, the more improvement we will get.

Conclusions

◆We construct the classification model based on support vector machine to recognize semantic inference in Chinese text pair using multiple features, including statistical, syntactic and lexical semantic ones.

◆Through further analysis, we find that in our system, we use the same features in BC and MC subtasks, but the characteristics of the BC and MC subtasks should be different, which may cause the dissatisfaction of BC result.

◆ Moreover, we mostly consider statistical features in our system, if we add some rule features, the accuracy may be significantly improved.