基于中文维基百科构建的知识库的谓词归一化

韩喆

iampkuhz@gmail.com

2015 年 4 月 2 日

Outline

- 背景
  - 基于维基百科的知识库
- Motivation
- 实验方法
- 特征选取
- 实验效果
  - 分析和改进

Background

基于维基百科的知识库

Background

基于维基百科的知识库

- definition
  - 给定一组句子, 判断其是否是复述
    - binary classification
- Microsoft Research Paraphrase Corpus (MSRP)
  - train: 4,076 sentence pairs (2,753 positive: 67.5 %)
  - test: 1,725 sentence pairs (1,147 positive: 66.5 %)
  - 2 个标注者, 83% 的一致性, 第三个人更正

Sample data

- Sentence 1: Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
- Sentence 2: Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.
- Class: 1 (true paraphrase)

Paraphrase identification

- Common methods
  - lexical features
    - n-gram features, skip-gram fatures, ...
  - semantic features
    - POS tag, wordnet similarity, dependency tree relation, ...
  - classification
    - SVM, voted classifications
- Challenge
  - 没有提取句子的全局信息 (dependency features 利用不足)
  - 对句子涵义的特征提取不足 (没有真正理解句子)