Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection (NIPS2011)

Zhe Han

1401214342

iampkuhz@gmail.com

2014 年 12 月 29 日

Outline

- Paraphrase identification(复述检测)
  - Definition
  - Common methods
- 本文的方法
  - Recursive Autoencoder
  - Dynamic Pooling
  - Experiment Result
- 实验效果
  - 分析, 对比其他任务

Paraphrase identification

- 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)

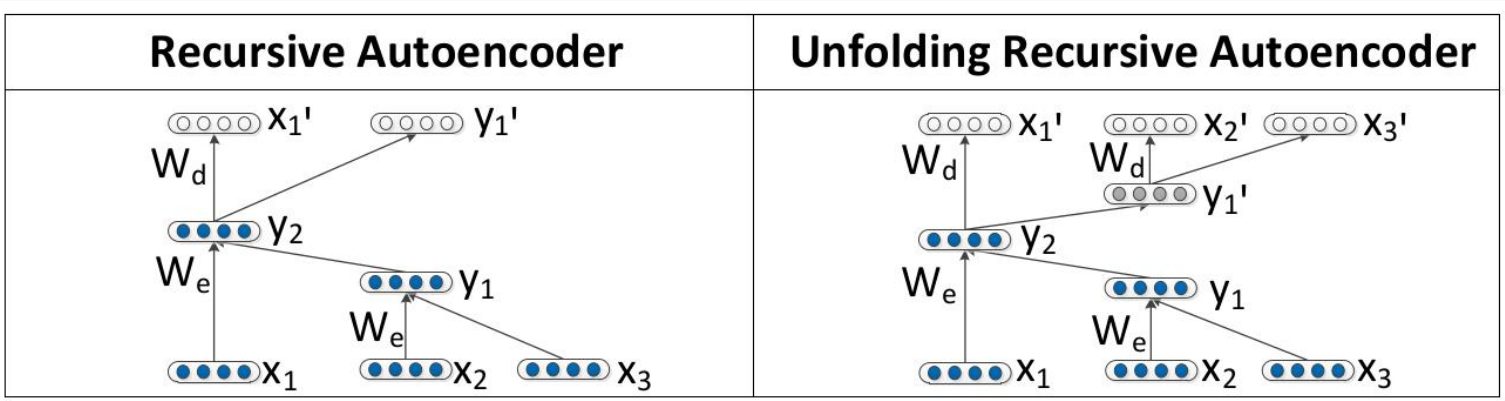
This paper

Main method

- 利用 NYT 新闻训练每个单词的向量 (100 维)
- 对于每个句子 (多个单词向量) 采用训练一个递归的自动编码机, 得到一个句子级别的语义向量.
- 通过判断两个句子的语义向量的相似性得到语义相似性特征

- 递归的自动编码机 (Unfolding Recursive Autoencoder)
  - 抽取句子的语义向量, 得到语法数上每个节点 (单词, 短语) 的向量
- Dynamic Pooling
  - 对于长度变化的两个句子, 抽取固定维数的特征

Unfolding Recursive Autoencoder



- Comparison (on $y_2 -> x_1 y_1$)
  - Autoencoder
    - minimum $\| y_2' - y_2 \|$
  - Recursive Autoencoder
    - minimum $\| [x_1'; y_1'] - [x_1; y_1] \|$
  - Unfolding Recursive Autoencoder
    - minimum $\| [x_1'; x_2'; ...; x_j'] - [x_1; x_2; ...; x_j] \|$

Unfolding Recursive Autoencoder(RAE)

- u RAE 和 RAE 类似, 我们通过解释 RAE 来说明