基于语义的成语检索方法研究

1. 绪论

1.1 研究背景和意义

1.2 研究现状

1.3 本文的工作和创新点

1.4 论文组织结构

1. 原理和知识库

2.1 词的表示学习

2.1.1 词的表示学习的发展

2.1.2 CBOW（Continuous Bag-of-Words Model）

2.1.3 Skip-gram

2.1.4 GloVe

2.2 知识库

2.2.1 BCC语料库

用来构造成语语料库子集、备选集生成

* + 1. 《知网》/ 同义词词林

想用在计算相似度或二次扩展

1. 基于语义的成语检索研究

3.1 基于语义的成语检索研究流程

简述流程，并给出流程图，然后接下来对关键步骤进行阐述

3.2 数据处理

3.2.1 分词与人工校对

对成语解释进行分词，然后人工过滤，主要是去掉不对的说法、把书面语改成口语，比如“貌”改成“的样子”，得到一次扩展集

3.2.2 基于规则的二次扩展

总结归来，比如开心/难过前加副词

3.2.3 同义词词林扩展

本来最后备选集是计算相似度打分排序，同义词可以不扩张，这里待确定

3.3词表示模型的训练

3.3.1主流的词表示训练

3.3.2 忽略虚词的词表示训练

3.4 查询query语法分析

3.4.1分词处理

得到分词词性标注串，用HanLp工具包

3.4.2 关键词提取

语法分析，提取表征句子语义的部分

3.4.3 停用词过滤

特定情况生成特定的过滤词表，包括形容、表示、比喻、用于等

3.5 候选成语的生成

3.6 \*打分排序

1： 读论文，找距离计算理论依据

2： 实验在max、sum、avg选效果最好

3：f值要实验找最优和理论依据

4：各个加分项的权值，实验找最优/找论文理论依据

1. 实验

4.1 \*评价准则

需要两个测试集：

1：100 个query BCC检索结果人工过滤，得到测试集 （仅测试了打分）

2：开放测试，网上搜索成语分类的成语集或者对1中100个query人能想到的成语构造测试集（测试整个成语检索效果）

4.2 实验数据

4.3 模型训练

4.4 备选集生成

BCC检索生成备选集

4.5 实验评价

4.5.1 测试集测试

4.5.2 现有平台互评

100个query和目前主流成语检索平台检索结果比较得到类似下表的评估结果



1. 总结和展望