辅助写作中的成语查询

摘 要

成语是一种相沿习用、形式简洁而意义精辟的固定短语，写作中正确使用成语能收到更好的表达效果，因为成语言简意赅，可使语言表达简洁明快；成语形象生动，可使语言色彩鲜明，增强表现力和感染力；成语形式严整，可使句式匀称，音律和谐，增强节奏感。成语的正确应用和学生写作能力的培养有十分密切关系。现有成语检索系统可通过词形和读音来查询成语的使用、释义等信息。然而，常有这样的情形：用户知道表达意图，却想不出恰当的词语。此时，现有的基于词形或读音或二者结合的方法都无法满足，而需要“基于语义的查询”。

基于关键词的查询方式可以通过关键词的组合来表达简单的语义信息，但主要是对用户输入的查询请求进行字串匹配，并没有真的考虑语义信息，会查询出大量的无用信息，且在大多数情况下用户很难用简单的几个关键词的组合来表达查询需求。所以，要实现通过语义查询词语，基于关键词的查询远远不够。针对这一现状，本文选取成语作为研究对象，探索基于语义的成语查询方法，并融合传统的成语查询，构建一个完善的成语查询系统，使得用户可以在只知道表达意图的时候，检索出对应的成语，也可以在知道具体成语的时候，检索出成语的读音、用法、释义等信息，更好地辅助中文写作。

本文将检索任务进行细化，拆分为“正向查询”、“修饰类反向查询”和“等价类反向查询”三个部分来实现。北京语言大学数据库BCC总字数约 150 亿字，有着庞大的数据资源，本文从其281.8G微博、博客语料中筛选得了到5.86G包含成语的子语料，并对筛选出来的子语料进行了清洗，只保留和成语在一个自然句（以汉语的句号、感叹号等为标准切分）内的部分，作为两种反向查询，也就是基于语义的查询的基础数据，并把《中国成语大辞典》作为正向查询的基础资源。对于修饰类反向类查询，本文将基于BCC查询引擎启动成语查询服务，对用户以自然语言形式提出的查询请求进行语法分析，提取关键词序列，再由关键词序列查询得到成语备选集。相似度计算方面，本文用到了词嵌入技术，利用谷歌开源的word2vec训练词嵌入模型，计算查询请求和备选成语之间的语义相似度，打分排序得到最满足用户需求的成语集。对于等价类反向查询，本文将构建一个成语<=>口语表达的资源库。实验结果表明，本系统实现了通过语义查询成语的功能并融合了传统的成语查询，能很好的满足用户的成语检索需求，更好的辅助中文写作。