语义校对系统中的句子语义骨架模糊匹配算法

郑逢斌[1,2], 陈志国[2], 姜保庆[1,2], 乔保军[2]
(1. 西南交通大学智能控制开发中心, 四川成都 610031; 2. 河南大学计算机科学学院, 河南开封 475001)

摘 要: 本文给出了用句子语义骨架表示句子语义的具体方法和表示形式. 在语义校对系统中建立了这种形式的知识库, 每一个知识条代表描述同一个事件的不同句子的共同特征. 采用模糊匹配方法计算语句的相似程度.

关键词: 术语; 语义骨架; 模糊匹配; 错误隶属度
中图分类号: TP193 文献标识码: A 文章编号: 0372-2112 (2003) 08-1138-03

Fuzzy Matching Technique by Sentence Skeleton in Semantic Collation System

ZHENG Feng-bin[1,2], CHEN Zhi-guo[2], JIANG Bao-qing[1,2], QIAO Bao-jun[2]
(1. Intelligence Control Development Center, Southwest Jiaotong University, Chengdu, Sichuan 610031, China;
2. College of Computer Science, Henan University, Kaifeng, Henan 475001, China)

Abstract: A concrete method and expression form describing the meaning of a sentence by its skeleton is obtained. A knowledge database of this form is established in this system. Each record expresses the common feature of different sentences that describe the same matter. And the similarity degrees between sentences are calculated by fuzzy matching method.

Key words: term; sentence skeleton; fuzzy matching; error membership

1 引言

自然语言是表达人们思想感情, 传递信息的一个特殊的符号系统. 各个符号系统有它自己的类型和结构, 在这个结构中不同成分之间存在着一定的联系, 而各部分的作用合起来又起到整体的效果. 一个语句的语义就是它各部分意义的函数[1]. 因此辨识一个语句的意义首先要了解其中每一个词素或词的语义, 积词成句, 再了解句子的语义, 然后按上下文的语义来理解全篇.

为了使现有的计算机能处理与自然语言理解有关的课题(如语义校对、翻译、自动写作等), 让机器能完全理解和表示自然语言当然是一条求之不得的途径, 但由于自然语言的复杂性, 这条途径困难很大, 近期内没有可能得到完全解决[2]. 为了解决文本的语义校对问题, 我们采取用若干个关键词来描述一个文本句子信息, 即用一个从文本中抽取的各关键词集合组成的句子语义骨架在一定程度上代表文本的语义, 我们籍此来进行语义校对. 下面介绍我们在语义校对系统(YYJDS)中提取关键词集合的方法和语义模糊匹配方法.

2 语义校对系统(YYJDS)简介

YYJDS 是我们于 2001 开发的能完成报刊书籍的政治性错误校对的软件系统, 它能完成报刊上关于台湾问题、领导人姓名、职务、排序问题、民族问题和宗教问题等常出现错误的语义校对. 它主要由词语自动切分、敏感词获取、合并和整理术语、术语校对、知识校对、自学习和关键词自动统计等功能模块组成. 用到的数据库文件有现代汉语语法信息词典、自动分词辅助库、敏感词语库、术语库、校对方式颜色库、校对结果库、语义校对标准知识库和语义校对标准知识库的分类索引库等. 它的工作流程是首先对整个文本进行自动词语切分, 然后在全文中搜索敏感词语并将其位置存入校对结果库, 再合并整理敏感词语为术语, 最后根据术语的不同类型分别进行不同的处理, 有的使用术语校对, 有的使用知识校对, 计算各个术语的错误隶属度并在文本中显示相应的颜色. 在该系统中, 整个句子的错误体现在该句子里面的术语上, 因此用术语的错误隶属度代替它所在句子的错误隶属度.

3 语义知识表示方式

在实际中, 与知识库中关键词或短语完全一样的句子并不多, 类似相近的句子常见, 为了有效区别知识库中的/标准句0与待校对文本中的实际语句之间的语义差别, 我们引进了语义权重和错误隶属度两个概念. 由于关键词集合中不同的关键词在描述相应文本的语义中所起的作用大小是不一样的, 我们给每个关键词一个语义权重来表示其对该语句的重要程度[3].

收稿日期: 2002-06-17; 修回日期: 2003-04-26
基金项目: 国家自然科学基金(No. 60074014)

语义校对标准知识库是用来判断文本中语句的对错程度的基准, 文本中语句是根据与知识库中相关知识进行模糊匹配, 然后计算出该语句的错误程度的. 语义校对知识库由知识条组成, 每个知识条是一个四元组, 用(K, R, O, P) 表示.

K 是一个线性表, K= (K_1, K_2, , , K_n) (本系统中 n [26), 它们的某种线性排列就代表一类句子的语义骨架. 其中 K_i(i= 1, 2, , , n) 仍是一个集合, 在这个集合中可能有一个或多个元素; 若有多个元素, 这些元素之间是同义词或可以互相代替的近义词[4]. K_i 中的元素称为关键词, K_i 称为关键词类.

R= (R_1, R_2, , , R_n), R_i(i= 1, 2, , , n) 表示 K_i 在句子中的权重, 当然 R_1+ R_2+ , + R_n= 1.

O 是一个集合, 其每个元素表示关键词类的一个子集及排列顺序. 用英文小写字母集{ a, b, , , z}的子集的元素之排列表示 O 的元素, 用符号/ | 0 将 O 的元素连接起来表示集合 O. 因此 O 实际上是一个串. 例如: O= abcdef| abdef | fcd 时, 表示三个语义骨架 $K_1K_2K_3K_4K_5K_6$, $K_1K_2K_4K_5K_3K_6$ 和 $K_6K_3K_4$.

P 表示该标准句错误隶属度(即错误的概率) . P= 1 时表示该句绝对有错, P= 0 是表示该句绝对无错, 0[P [1. 0.

例如: K1= {台湾, 台北, 台湾岛};

K_2= {是, 作为, 应为, 代表, 属于};

K_3= {中国, 中华人民共和国, 我国};

K_4= {省, 岛, 领土, 部分, 门户, 地区};

R= {0l3, 0l2, 0l4, 0l1}; O= abcd; P= 0l01.

4　关键词集的抽取及知识库的建立

对用于政治性领域语义校对关键词的抽取, 我们采取人为规定和自动统计相结合的方法.

人工部分的规定如下: (1) 句子的主、谓、宾作为关键词; (2) 否定词是关键词; (3) 出现在我们术语库中的词是关键词.

自动统计关键词模块的工作流程如下:

第一步, 从网上下载大量的有关方面的文章, 通过调用自动切分程序把这些文本变成词的序列, 把没有敏感词的句子全部删掉, 仅对含有敏感词的句子进行统计分析.

第二步, 把文本中的系词、前置词、冠词、代词等词类去掉, 将形容词或副词与其修饰的词结合在一起当作一个复合词[5].

第三步, 依次扫描每个文本, 并按下列方法逐词进行统计[5]:

(a) 每个词在第一次出现时设一个相应的计数器, 并置成 1, 此后该词每出现一次就在其相应的计数器中加 1.

(b) 标题或摘要中出现的词, 除同上面 a 步中的处理外, 再在相应的计数器中外加一个整数 T. YYJDS 系统中取 T 的值为 10.

(c) 在段首或段尾出现的词, 除同上面 a 步中的处理外, 再在相应的计数器中加一个整数 P. YYJDS 系统中取 P 的值为 3.

(d) 在引言和结论段中出现的词, 除同上面 a 和 c 步中的处理外, 再在相应的计数器中加一个整数 Q. YYJDS 系统中取 Q 的值为 3.

第四步, 在选取的文本全部处理完后, 将所有词的计数器的值相加得和数 S, 然后每个计数器的值除以 S 再放回计数器中.

第五步, 截取关键词: 根据需要, 设定一个阈限值 r(0< r < 1), 把该关键词集合中的隶属度小于 r 的关键词滤掉, 剩下的作为我们的部分关键词.

5　模糊匹配算法与语句相似度计算

5l1　模糊匹配算法

根据我们语义校对系统的校对流程, 自动分词和术语整理后, 首先对整个被校对文本进行一遍扫描, 找出文本中的所有敏感词; 其次通过术语校对完成部分敏感词的处理(这部分属于只需要理解词语含义而不需要整个句子语义就可以确定对错的词语) ; 最后要用语义校对标准知识库对剩下的敏感词所在的句子进行处理.

不妨设被校对文本中一个具体的句子 S= W_1W_2, W_m, 其中 W_j(j= 1, 2, , , m) 是一个被我们语义校对系统中的自动分词模块切分后并经过术语整理模块整理好的词或短语, W_q (q= 1, 2, , , m) 是敏感词(如: 台湾). 依次逐条对照知识库的语义分类索引库中分类关键词, 在 S 句子中从 W_q 位置向左右两侧查找分类关键词, 直到找到为止. 若找完了分类索引库仍未找到, 说明现有的知识库中目前还没有要校对的语义知识, 可以通过系统的自学习功能将本语句知识添加到知识库中并在语义分类索引库中添加一个语义类. 当确定了本句所属的语义类之后, 后面只需将本语句与语义校对标准知识库中属于本类的知识条依次进行匹配. 具体的模糊匹配算法如下:

第一步, 从知识库属于本类的知识条中选定一个知识条 (K, R, O, P) , 其中 K= (K_1, K_2, , , K_n) , R= (R_1, R_2, , , R_n) ; 若本类知识条被选完, 取默认值作为该术语的错误隶属度, 跳到第七步. 说明现有的知识库中目前还没有要校对的语义知识, 可以通过系统的自学习功能将本语句知识添加到知识库中但不需要在语义分类索引库中添加一个语义类;

第二步, 从 O 中选取一个语义骨架(不妨设为) ; 若所有语义骨架都处理完, 返回到第一步; 说明 S 与本知识条不匹配;

第三步, 确定 W_q 在K 中的位置 K_{i_a} (S 属于本类知识条, 位置一定存在) ;

第四步, 在 S 中从 W_q 位置向左依次查找与关键词类中相匹配的关键词, 若某个必选关键词未找到, 返回到第二步; 说明 S 与本知识条的该种顺序不匹配;

第五步, 在 S 中从 W_q 位置向右依次查找与关键词类 $K_{i_{a+1}}$ $K_{i_{a+2}}$, +, K_{i_t} 中相匹配的关键词, 若某个必选关键词未找到, 返回到第二步; 说明 S 与本知识条的该种顺序不匹配;

第六步, 至此说明 S 与本知识条的该种语义骨架匹配, 可以进行语句相似度计算并得出该术语的错误隶属度(具体计算方法见下面 5l2).

第七步, 将该术语的错误隶属度(此处用 S 中敏感词语

的错误隶属度代替语句S 的错误隶属度)存入校对结果库, 该句子的语义校对结束.

512 语句相似度计算

在我们的校对系统中, 被校对语句 S 的语义与成功匹配的知识库中知识条(K, R, O, P) 所代表的语义相似度的计算, 就是根据知识条(K, R, O, P) 的错误隶属度 P 计算出语句 S 的错误隶属度Ps 的[6]. 具体计算方法如下:

当 P \ 01 5 时, $P_s = (R_{i_1} + R_{i_2} + \cdots + R_{i_t})P$ (1)

当 P < 0. 5 时, $P_s = 1 - (R_{i_1} + R_{i_2} + \cdots + R_{i_t})(1 - P)$ (2)

由于 P 是校对标准知识库中知识条的错误概率, 选取时都是一些有代表性的错误句和有代表性的正确句, 它们的 P 值都是 0 和 1 或者接近 0 和 1 的小数. 对于 P 值是 1 或接近于 1 的知识条, 我们认为越与它相近的句子越可能错, 能匹配上但又越不相似的句子错误的可能性越小, 为此我们定义了上面的公式 1. 对于 P 值是 0 或接近于 0 的知识条, 我们认为越与它相近的句子越可能正确, 能匹配上但又越不相似的句子不正确的可能性越大, 为此我们定义了上面的公式 2.

6 结束语

中文文本的自动语义校对是自然语言理解中的一项重要研究课题, 本文对语义校对系统中句子语义骨架模糊匹配技术及其应用作了初步探讨, 并在一个特定领域内的语义校对系统(YYJDS) 中得到实现. 当然软件中还有很多问题需要进一步的改进, 我们正在考虑知识库的一般模型, 计划把语义词典用于本系统, 把句子语义骨架中的同义词改进为语义类, 在匹配算法, 处理代词等许多方面还有待进一步完善和改进.

参考文献:

[1] 姚天顺. 一种混合的中文文本校对方法 [J]. 北京: 中文信息学报, 1998, 12(2) : 31- 36.

[2] Christopher D Manning, Hinrich Schutze. Foundations of Statistical Natural Language Processing [M]. Massachusetts: MIT Press, 1999.

[3] Dekang Lin. Extracting Collocations from Text Corpora [D]. Canada: Department of Computer Science University of Manitoba, 1998.

[4] R Garside, G Leech, T McEnery. Corpus Annotation: Linguistic Information from Computer Text Corpora [C]. London: Longman, 1997.

[5] 何新贵. 中文文本的关键词自动抽取和模糊分类 [J]. 中文信息学报, 1999; 13(1) : 9- 15.

[6] 何新贵. 加权模糊及其广泛应用. 北京 [J]. 计算机学报, 1989; 12(6) : 458- 464.

作者简介:

郑逢斌 男, 1963 年出生于河南省新县, 博士生, 副教授, 硕士研究生导师, 主要研究领域为自然语言理解, 软件工程.

陈志国 男, 1955 年出生于河南省开封市, 副教授, 硕士研究生导师, 主要研究领域为人工智能, 计算机网络.

姜保庆 男, 1964 年出生于河南省滑县, 博士生, 副教授, 硕士研究生导师, 主要研究领域为自然语言理解, 模糊逻辑.