中文文本语义错误侦测方法研究

张仰森 郑 佳

(北京信息科技大学智能信息处理研究所 北京 100101)

摘 要 中文文本语义错误侦测一直以来都是中文文本自动查错的难点.该文针对中文文本语义错误,提出了一种基于语义搭配知识库和证据理论的语义错误侦测模型.讨论了三层语义搭配知识库的构建以及基于该知识库和证据理论的语义错误侦测算法.三层语义搭配知识库的构建主要分为两步:(1)根据《现代汉语实词搭配词典》中的实词搭配框架构建词语搭配规则集,从训练语料中抽取词语搭配,并利用互信息和共现频次进行筛选,构建词语搭配知识库;(2)利用《HowNet》抽取词语的义原信息,生成词语-义原和义原-义原搭配知识库,并利用聚合度进行二次筛选.在三层语义搭配知识库的基础上,首先对知识库采用自顶向下的搜索模式确定可能错误的语义搭配,然后使用语义搭配的互信息量 *MI* 和聚合度 *PD* 作为证据,采用统计的方法建立证据信任分配函数,结合证据的冲突处理和加权分配D-S规则进行不确定性推理,获取词语的语义搭配关联强度,以判定是否存在语义错误.实验结果显示,该文所提出的查错模型和算法的F-Score值比其他文献中的最好值提高了14.02%.

关键词 语义错误;知识库;D-S理论;语义搭配;错误侦测算法;自然语言处理;社会媒体

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2017.00911

Study of Semantic Error Detecting Method for Chinese Text

ZHANG Yang-Sen ZHENG Jia

(*Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing* 100101)

Abstract Chinese text semantic error detection is always the difficult point of Chinese text automatic error detection. In this paper, a semantic error detection model is proposed based on semantic knowledge base and D-S theory. We discuss the building method of the three layers semantic collocation knowledge base and the semantic error detection algorithm based on the three layers semantic collocation knowledge base and D-S theory. Construction of three layers semantic collocation knowledge base is divided into two steps: (1) According to the notional collocation frame in *Modern Chinese Dictionary of Notional Words Collocation* to construct words collocation rule set, extract collocations from the training corpus based on the rule set, and building the collocation knowledge base through filtering the some collocations by mutual information and co-occurrence frequency; (2) use *HowNet* to extract the sememe information of word in order to generate the word-sememe and the sememe-sememe knowledge base, and use the polymerization degree model to do the second level filtering. On the basis of the three layers semantic generate knowledge base, a top-down search pattern is used to identify the possible errors firstly, and then the semantic collocation mutual information *MI* and polymerization degree *PD* are used as evidences, adopt statistical method to generate basic probability assignment, combining the evidence conflict resolution and the weighted distribution D-S rules to get the relevancy of semantic collocation to

收稿日期:2016-05-04;在线出版日期:2016-09-04.本课题得到国家自然科学基金(61070119,61370139)和北京市属高等学校创新团队建设与教师职业发展计划(IDHT20130519)资助.**张仰森**,男,1962年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为自然语言处理、人工智能.E-mail:zhangyangsen@163.com.**郑 佳**,男,1991年生,硕士研究生,主要研究方向为自然语言处理.

determine whether there is a semantic error in Chinese text. The experimental result shows that the F-Score values of the error detecting model and algorithm proposed in this paper improved 14.02% than the best values in the literature.

Keywords semantic error; knowledge base; D-S theory; semantic collection; error-detection algorithm; natural language processing; social media

1 引 言

出版业的快速发展和电子书刊的逐步普及对中文文本自动校对技术提出了更高的要求.国外对于英文文本自动校对技术的研究开始较早,在 1960 年,IBM Thomas J. Watson 研究中心实现了一个 TYPO 英文拼写检查器[1],随后,英文文本自动校对技术取得了长足的发展[2-5].相比较于国外,国内对于中文文本校对方面的研究在 20 世纪 90 年代初期才正式开始,但发展得较快,许多大专院校、企业都在这方面投入了大量的财力、物力和人力,并且取得了一定的研究成果.

中文由于自身特点,与英文文本校对不同,中文文本校对是基于错误文本的字词、语法和语义等 3 个层次的分析过程.目前,对字词级别和语法级别的错误侦测研究相对比较充分,也取得了比较好的效果[6-12].但中文文本中还存在许多语义级别的错误,主要是语句中的搭配错误,包括主谓搭配、动宾搭配、形副搭配和量名搭配等,这些错误从句法上看可能并不存在问题,但不符合语义搭配规范,例如:"大雪纷飞,他戴着帽子和皮靴就出门了".其中,谓语动词"戴"和宾语"皮靴"的搭配就是不符合语义搭配规范.本文的目标就是研究如何让计算机能够像人一样来侦测出中文文本中的语义搭配错误.

2 相关工作

语义问题的研究一直是计算语言学和自然语言处理研究中的薄弱环节,也是中文文本校对的难点.对于中文文本中字词表示正确、句法结构完整,而在语义层次上不符合语义搭配规范的错误,到目前为止,都还没有比较成熟的解决方法.词语的语义搭配主要受到语义成分、思维习惯、风俗习惯以及认知习惯等方面[13]的影响,长期以来被学者们广泛研讨.从语义成分的角度看,凡是存在语义组合关系的词语相互之间都有一定的语义要求,并且,这种语义要求都与词的语义成分密切相关.例如,与动词"喝"搭配的受事词语必须包含语义成分:[具体物质]和[液体状态],或者至少不与上述语义成分相冲突,因此"喝水"、"喝茶"等的搭配认为是正确的;然而"饭"的语义特征中虽然有[具体物质],但其[固体状态]的语义特征与[液体状态]是相对的,因此"喝饭"就不符合语义搭配.从思维习惯角度看,语义的搭配还要受到不同的思维习惯的影响.例如,对于"生"和"死"这类特征只有具有生命的事物才具有,因此,人们比较容易接受和理解这样的搭配:"这只小狗死了"、"婴儿出生了",而对于诸如"石头出生了"、"石头死了"之类的搭配难以接受,当然对特定语境或使用修辞手法的情况除外.从风俗文化角度看,不同的民族和方言,其语义搭配是不同的.例如,吴方言中的"吃"与普通话中的"吃"在语义搭配上就有着不同的要求,普通话中的"吃"只能与固体食物搭配,而吴方言中的"吃"还可以与液体食物等搭配,像"吃酒"、"吃茶"、"吃烟"之类的都是可以接受的.这就是不同风俗文化对词语的语义搭配的影响.

清华大学的罗振声教授领导的研究小组[14]综合使用基于实例、基于统计和基于规则的语义搭配关系进行语义错误的检查.通过对待校对文本进行句法分析,得到句子成分,再在语料库中寻找与句子结构相似的实例,计算两者的相似度进行查错;同时他还效仿字词级别错误校对方法,构造词义的 N 元邻接矩阵,计算语义平均概率和语义转移概率进行查错.罗振声教授提出的统计和规则相结合的校对策略,既能检查局部的语义搭配错误,也能检查较长距离的语义搭配错误,收到了较好的效果,为语义错误侦查提供了新思路.但是,其语义规则库的精度和广度对校对结果的影响很大,同时邻接矩阵的数据稀疏问题也会给语义搭配错误的侦测造成阻碍.

西南交通大学控制智能开发中心的郑逢斌博士等人[15]开发出一款针对报刊书籍中的政治性错误的文本校对系统(YYJDS),其校对过程为:首先,对

训练语料抽取关键词集合组成句子语义骨架，建立语义骨架标准知识库；然后，找出待校对文本中所有相关的敏感词，将敏感词所属的语句与知识库中相关的词条采用模糊匹配的方法进行相似度计算. 郑逢斌博士等人提出的方法仅局限于政治领域内的文本语义错误校对，不适用于非限制领域的文本语义搭配错误侦测.

江苏大学的程显毅等人[16]基于 HNC(概念层次网络)理论构建了一个中文文本校对系统模型，该模型将传统查错系统和 HNC 句类分析系统相结合，利用句类知识库和概念关联知识库对校对文本进行句类检查并构成语义块，通过借助概念关联知识库对语义块的语义成分进行分析来侦查语义错误. 该模型对语义方面的错误具有较好的侦测能力，但其语义块的切分与组合处理对句类知识库和概念关联知识库依赖性较强，句类知识库和概念关联知识库的准确性和完备性直接影响其侦测能力.

3　文本语义错误侦测的基本思路

本文经过大量的研究发现，《HowNet》对词语义原的描述形式符合汉语语义搭配查错的需要. 在《HowNet》中，将义项的义原定义为最基本的、不能再分割的意义上的最小单位. 所有的概念都可以表示为各种各样的义原，即用有限的义原集合来描述概念与概念以及概念的属性之间的关系，这一点正好符合语义成分分析的思想，其义原信息能在一定程度上反应词语的语义信息. 利用《HowNet》构建词语的语义搭配知识库是语义错误侦测的基础.

同时，在文本的语义校对过程中，由于词语的语义搭配受到了各种因素的制约，有些搭配的正确与否是很容易判断的，而有一些语义搭配并不容易判断是否正确，处于一种“不知道”的状态，具有一定的模糊性. D-S 证据理论作为一种不确定性推理的理论，恰好能够很好的处理这种由“不知道”而引起的不确定性，具有较大的灵活性，因而受到人们的重视，并在信息融合、故障检测和决策分析等领域得到广泛应用[17-19]. 鉴于此，本文引入改进的 D-S 证据理论，以词语搭配互信息和词语搭配聚合度为证据属性，构建基于多源证据融合的语义关联强度评估模型，以对文本校对过程中的词语语义搭配关系强度进行评估.

本文的基本思路是将基于《HowNet》构建的语义搭配知识库和基于 D-S 证据理论构建的词语语义搭配关联强度评估模型相结合，建立中文文本语义错误侦测模型与算法. 首先，以《人民日报》的标注语料为基础，在语言学知识的指导下，构建搭配规则集进行词语搭配抽取，并以互信息和共现词频进行过滤筛选，构建词语搭配知识库；然后，利用《HowNet》的义原信息以及词语搭配聚合度，构建一个基于实词的三层语义搭配知识库，利用该知识库对文本中的语义搭配错误进行初选；最后，使用词语搭配互信息 *MI* 和词语搭配聚合度 *PD* 作为证据，采用统计方法建立证据信任分配函数，结合证据的冲突处理和加权分配 D-S 规则进行不确定性推理，获取词语的语义搭配关联强度，以此判定文本中是否存在语义搭配错误. 具体的语义错误侦测框架如图 1 所示.

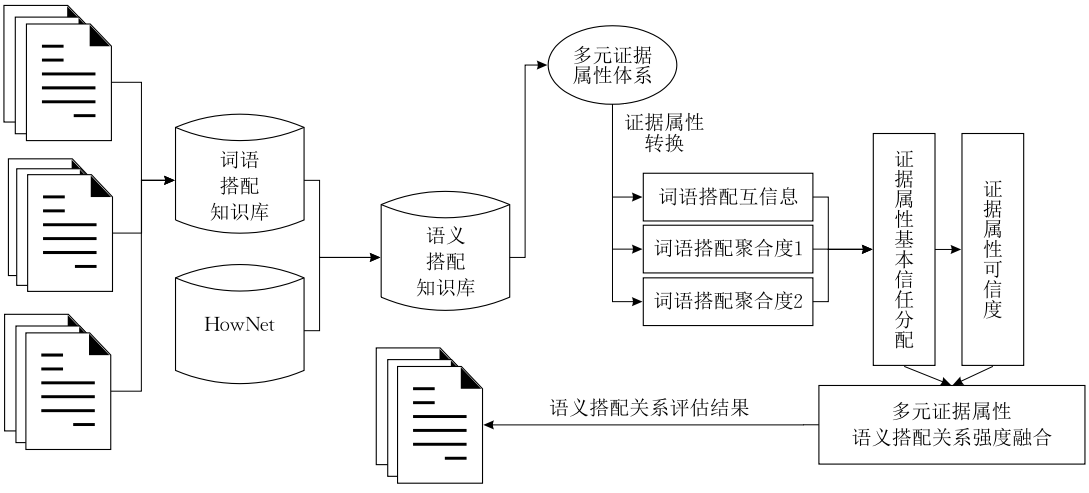

图 1　基于语义知识库和 D-S 证据理论的语义错误侦测框架

4 语义搭配知识库的构建

4.1 基于语言学知识和统计学理论的词语搭配提取模型

词语搭配主要有 3 个特点:任意性、重复性和结构性.针对这 3 个特点进行词语搭配的提取才能获得准确、全面的搭配.目前,对于词语搭配的自动提取方面的技术研究主要分为 3 个方面:(1)采用规则的方法;(2)采用统计学的相关方法;(3)采用规则与统计学相结合的方法.目前的大部分研究都是以统计学的相关方法为主,并结合语法、句法规则或语义等语言学的相关知识,对抽取的词语搭配进行过滤.本文通过研究发现,现有中文词语搭配抽取方法大多存在没有充分考虑词语搭配的结构信息、没有充分利用语言学知识且缺乏通用性等问题.针对以上问题,结合现有的方法和理论,本文拟通过实验对当前中文词语搭配抽取方法做出相应的改进,提出一种基于语言学知识和统计学理论相结合的词语搭配提取模型.

依据语言学的中心词分析方法,两个或者多个词语在构成搭配时,必然有一个词作为主体成分,称为中心词,而另一个或多个词作为修饰成分,称为搭配词.在汉语中,名词、动词和形容词这 3 类实词在表达实际语言意义方面起着关键性的作用,故本文将它们作为中心词构建模型.对于词语搭配,要想其更加符合语言学规律必须引入相关语言学知识,本文通过整合《现代汉语实词搭配词典》中的实词搭配框架,抽取出以名词、动词和形容词为中心词的词语搭配规则如下:

(1)以名词为中心词的搭配词的词性有:名词、动词、形容词、量词;

(2)以动词为中心词的搭配词的词性有:名词、动词、副词;

(3)以形容词为中心词的搭配词的词性有:形容词、副词.

我们依据以上语义搭配规则构建词语搭配规则集 S 如下所示:

$$S \rightarrow N+N \mid V+N \mid A+N \mid Q+N$$
$$S \rightarrow N+V \mid V+V \mid D+V$$
$$S \rightarrow A+A \mid D+A$$

其中:N 表示名词;V 表示动词;A 表示形容词;Q 表示量词;D 表示副词.

通过以上搭配规则集抽取出的词语搭配对,只能说明搭配中的两个词由于某种特殊的原因而相互吸引,经常在一起出现,但其能够提供的信息十分有限,只能起到初步筛选的作用.下面将利用词语的共现词频和互信息进行进一步过滤筛选.

互信息的定义如式(1)所示:

$$MI(w_1,w_2)=\log_2\frac{P(w_1,w_2)}{P(w_1)\times P(w_2)} \tag{1}$$

其中:$P(w_1,w_2)$ 表示词语 w_1,w_2 在训练语料中共同出现的频率;$P(w_1)$、$P(w_2)$分别表示词语 w_1、w_2 在训练语料中各自单独出现的频率,其计算如式(2)、式(3)所示:

$$P(w_1)=\frac{c(w_1)}{N},\ P(w_2)=\frac{c(w_2)}{N} \tag{2}$$

$$P(w_1,w_2)=\frac{c(w_1,w_2)}{N} \tag{3}$$

其中:$c(w_1)$、$c(w_2)$、$c(w_1,w_2)$表示词 w_1、w_2在训练语料中各自单独出现的频次和共同出现的频次;N 为训练语料中出现的字串总数.对于词语互信息 $MI(w_1,w_2)$和词语共现频次$c(w_1,w_2)$的阈值τ_1和 τ_2 的选择方法将在后面的实验部分 7.1.1 节进行讨论.

词语搭配抽取算法的过程描述如算法 1 所示.

算法 1. 词语搭配提取算法.

输入:训练语料 T;规则集 S;τ_1;τ_2

输出:词语搭配知识库

过程:

1. 对 T 进行逐句扫描,抽取两个标点符号之间的短句赋值给 P;

2. 对 P 进行逐词扫描,将 P 当前词语赋值给 W,若 W 是名词跳转至步骤 3,是动词跳转至步骤 4,是形容词跳转至步骤 5,其它则跳转至步骤 6;

3. 对照 S,提取名词相关搭配存入 L,跳转至步骤 6;

4. 对照 S,提取动词相关搭配存入 L,跳转至步骤 6;

5. 对照 S,提取形容词相关搭配存入 L,跳转至步骤 6;

6. 判断 P 的当前扫描位置是否为末尾,如果是则向下执行,否则跳转至步骤 2;

7. 判断 T 的当前扫描位置是否为末尾,如果是则向下执行,否则跳转至步骤 1;

8. 扫描 L,统计各搭配的共现频次,若共现频次小于 τ_2,则删除;

9. 扫描 L,计算各搭配的互信息,若互信息小于 τ_1,则删除;

10. 将 L 存入词语搭配知识库,结束.

4.2 语义搭配知识库的自动生成

在上述算法 1 构建的词语搭配知识库的基础上,将利用《HowNet》中对词语的义原描述体系,将词语搭配知识库中的词语搭配在语义层次上进行扩

充，将两个词语之间的二元搭配组合关系扩展为词语义原之间相互制约的多元组合关系，构建语义搭配知识库.

4.2.1　词语搭配聚合度

目前越来越多的研究表明词语搭配对自然语言处理的显著作用，而词语搭配的语义存储是对词语搭配的最有效存储[20]. 但是，并非所有的词语搭配都适合转换为其语义搭配. 例如："秀丽＋风光"，利用《HowNet》将其转换为义原的搭配为"美-(秀丽的义原)＋景象-无生物-(风光的义原)"，但是，"帅"这个词的义原也为"美-"，显然不能用"帅"与"景象-无生物-"进行搭配. 由此可见，如果将所有的词语都不加限制的转化为其义原，将会导致一些错误的搭配.

本文引入词语搭配聚合度 PD 的概念来衡量词语搭配能转化为其义原搭配的程度. PD 定义如下.

定义 1.　设 x 为中心词，y 为其搭配词，若 se 为 x 的义原，则可得到一个具有相同义原 se 的词集合 SYN_x，该集合中的词语与 y 的搭配程度称为聚合度，以 PD 表示. PD 的计算如式(4)所示：

$$PD=\frac{\sum_{i=1}^{N}C_{w_i,y}}{N} \tag{4}$$

其中：N 为 SYN_x 中词的数量；$C_{w_i,y}$ 的定义如式(5)所示：

$$C_{w_i,y}=\begin{cases}1, & \text{若 } w_i \text{ 与 } y \text{ 搭配，} w_i\in SYN_x \\ & \text{且 } 1\le i\le N \\ 0, & \text{其他}\end{cases} \tag{5}$$

当 PD 越趋于 1 时，则表明在训练语料中与 x 具有相同义原的词与 y 搭配的可能性越大，提取 x 的义原 se 的价值也就越大；当 PD 越趋于 0 时，则表明与 x 具有相同义原的词与 y 搭配的可能性越小，提取 x 的义原 se 的价值也越小. 对于那些义原转换价值不大的词语，我们就不提取其义原.

设 λ 为 PD 的阈值，当 $PD\ge\lambda$ 时，认为与 x 具有相同义原的词语可以与 y 构成正确的搭配，就可以将 x 的义原信息抽取出来以构建语义搭配知识库. 对于阈值 λ 的选择方法将在后面的实验部分 7.1.2 节进行讨论.

4.2.2　语义搭配知识库的生成算法

本文的语义搭配知识库生成算法是以由算法 1 构建的词语搭配知识库为基础数据，利用《HowNet》的义原描述体系，以 PD 为阈值进行构建的. 具体的过程描述如算法 2 所示.

算法 2.　语义搭配知识库的生成算法.

输入：词语搭配知识库；λ

输出：语义搭配知识库

过程：

1. 提取《HowNet》中动词、形容词、副词和名词的义原信息，并统计相同义原词的个数，存入列表 L；

2. 对词语搭配知识库进行逐行扫描，将词语搭配知识库当前词语搭配赋值给 C；

3. 处理 C. 如果 C 是量词＋名词的搭配跳转至步骤 4，否则，中心词转换成义原，匹配中心词-义原结果列表 $hL1$，如果存在，搭配数加 1，如果不存在，存入 $hL1$；然后，搭配词转换成义原，匹配搭配词-义原结果列表 $hL2$，如果存在，搭配数加 1，如果不存在，存入 $hL2$；

4. 对于量词＋名词的搭配，中心词名词转换成义原，匹配量词-义原结果列表 hLq，如果存在，搭配数加 1，如果不存在，存入 hLq；

5. 判断词语搭配知识库的扫描进程，如果已全部扫描完毕则向下执行，否则跳转至步骤 2；

6. 扫描 $hL1$、$hL2$ 和 hLq，计算每个搭配的聚合度赋值给 $PD1$，如果 $PD1$ 小于 λ 则删除该搭配；

7. 将 $hL1$、$hL2$ 和 hLq 存入词语-义原搭配知识库；

8. 对 $hL1$ 和 $hL2$ 进行逐词扫描，将其当前词语搭配分别赋值给 $C1$ 和 $C2$；

9. 处理 $C1$ 和 $C2$. 将 $C1$ 搭配词转换成义原，将 $C2$ 中心词转换成义原，均匹配结果列表 cL，如果存在，搭配数加 1，$PD1$ 相加，如果不存在，连同 $PD1$ 一起存入 cL；

10. 判断 $hL1$ 和 $hL2$ 的当前扫描位置是否为末尾，如果是则向下执行，否则跳转至步骤 8；

11. 扫描 cL，计算每个搭配新的聚合度赋值给 $PD2$，并将 $PD1$ 的平均数赋值给 $PD1$，令 $PD=PD1\times PD2$. 如果 PD 小于 λ^2 删除该搭配；

12. 将 cL 存入义原-义原搭配知识库，结束.

4.3　语义搭配知识库的结构体系

本文设计的语义搭配知识库由 5 个子库构成，分为 3 层. 词语搭配知识库为第 1 层，是从训练语料中抽取的词语搭配，是所有知识的来源；中心词-义原知识库、搭配词-义原知识库和量词-义原知识库构成词语-义原搭配知识库，是知识库的中间层，是构建义原-义原搭配知识库的基础；义原-义原搭配知识库是第 3 层，由词语的义原搭配构成，是知识库的最高层. 整个知识库自顶向下构成了一种层次关系和类属关系，高层知识库都是从低层知识库抽取得来的，能从中找到记录，同时也进行了有限扩充. 低层知识库记录的一部分类属于高层知识库. 这个层次关系和类属关系很好地反映了语义搭配知识对词语搭配的表征程度.

整个语义搭配知识库的体系结构如图 2 所示.

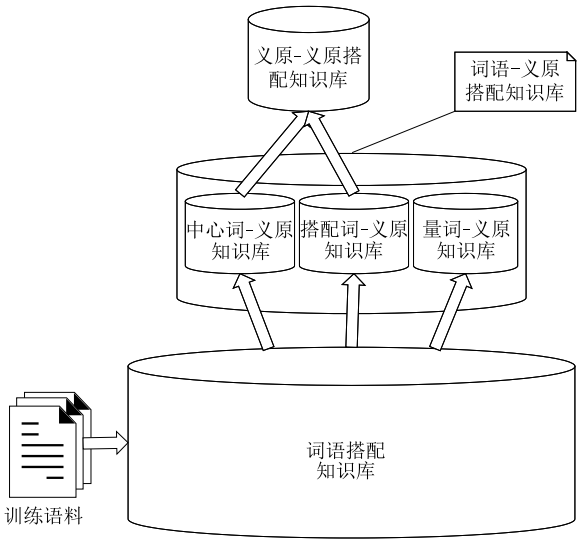

图 2 语义搭配知识库体系结构

下面对各个知识库的结构进行简要的介绍:

(1) 词语搭配知识库

词语搭配知识库以名词、动词和形容词为中心词,抽取了在 4.1 节中构造的规则集 S 中的所有 9 种搭配,其结构如表 1 所示.

表 1 词语搭配知识库结构

搭配词	搭配词词性	中心词	中心词词性	互信息	共现频次

(2) 词语-义原搭配知识库

词语-义原搭配知识库由中心词-义原知识库、搭配词-义原知识库和量词-义原知识库这 3 个子库构成,其包含的搭配类型为规则集 S 中除 $N+N$, $V+V$ 和 $A+A$ 这 3 种类型以外的所有搭配类型.

在《HowNet》中对量词的语义信息表示得很简略,没有什么区分度,例如:

"个"在《HowNet》中的描述为

DEF={ NounUnit|名量,&human|人}

"次"在《HowNet》中的描述为

DEF={ ActUnit|动量,&event|事件}

即量词在《HowNet》的义原结构中只是简单的分为名量、动量等,而在 $Q+N$ 中,中心词名词能提供更多语义信息,因此,对于 $Q+N$ 单独列出,量词保持不变,只将名词提取义原信息,其结构如表 2 所示.

表 2 量词-义原知识库结构

量词	中心词	中心词义原	个数	聚合度

中心词-义原知识库是搭配词保持不变,提取中心词的义原信息构成搭配关系;搭配词-义原知识库是中心词保持不变,提取搭配词的义原信息构成搭配关系,其结构如表 3、表 4 所示.

表 3 中心词-义原知识库结构

搭配词	中心词	中心词义原	个数	聚合度

表 4 搭配词-义原知识库结构

搭配词	搭配词义原	中心词	个数	聚合度

(3) 义原-义原搭配知识库

义原-义原搭配知识库是由中心词-义原知识库提取搭配词的义原信息和搭配词-义原知识库提取中心词的义原信息构成,该知识库结构如表 5 所示.

表 5 义原-义原搭配知识库结构

搭配词	搭配词义原	中心词	中心词义原	个数

5 语义搭配关联强度评估模型

5.1 基于语义搭配知识库的语义搭配错误初选策略

对语义搭配错误的初选我们采用对语义搭配知识库进行检索的策略来确定,对语义搭配知识的检索采用自顶向下的检索模式,从语义搭配知识库的第 3 层逐层向下检索,对文本中的语义错误进行初选,其具体策略如下:

(1) 检查当前词语搭配,如果是 $Q+N$ 搭配,则进入第 2 步检索,否则进入第 3 步检索.

(2) 对于 $Q+N$ 搭配,将搭配的中心词名词转化为其义原,检索量词-义原知识库中的搭配,如果知识库中存在该搭配,则判定该搭配正确,否则进入第 5 步检索.

(3) 将词语搭配的中心词和搭配词都转化为其义原,检索义原-义原搭配知识库,如果知识库中存在该搭配,则判定该搭配正确,否则进入第 4 步检索.

(4) 将词语搭配的中心词转化为其义原,检索中心词-义原知识库,如果知识库中存在该搭配,则判定该搭配正确;否则,将词语搭配的搭配词转化为其义原,检索搭配词-义原知识库,如果知识库中存在该搭配,则判定该搭配正确,否则进入第 5 步检索.

(5) 检索词语搭配知识库,如果知识库中存在该搭配,则判定该搭配正确,否则可以认为该搭配存在语义搭配错误的可能.

5.2 基于证据理论的语义搭配关联强度度量模型

由于训练语料的有限和知识库构建算法的局限,必然会导致语义搭配知识库不够完备,因此,根据语义搭配知识库的检索确定出的错误搭配只能作

为错误的候选对象，其中存在一些“不知道”的情况. 依据第 3 节的论述，为进一步判断这些处于“不知道”状态下的语义搭配的正确与否，本文采用证据理论进行最终的错误判定.

对于语义搭配正确性的判定是建立在语义搭配中两个词语的位置关联信息和语义关联信息的基础之上，因此，我们综合词语搭配的位置关联信息和语义关联信息，使用词语搭配互信息 MI 和词语搭配聚合度 PD 作为证据属性，将语义搭配的正确性判断转化为语义搭配相关和语义搭配无关的问题，定义侦测框架 $\Theta=\{R,N\}$，其中 R 代表“语义相关”，N 代表“语义不相关”.

对于校对文本的词语搭配互信息 MI，采用式(1)计算获取. 对于词语搭配聚合度 PD，我们将搭配的中心词和搭配词分别转换为其对应的义原，构成搭配词-义原和中心词-义原的两组搭配组合，然后采用式(4)分别计算两组搭配的词语搭配聚合度 $PD1$ 和 $PD2$. 在确定好框架和证据属性后，我们需要建立证据信任分配函数(BPA).

5.2.1　证据信任分配函数的确定

利用 D-S 证据理论进行不确定性表示和推理的过程中，不确定性信息的表示是首先需要解决的问题，即构建基于证据理论的不确定推理的初始信度依据，其本质是有关集值随机变量分布建模的相关问题[21]. 这一问题在数学领域仍然是一个未能很好解决的难题，目前虽然存在一些通用的方法[22-23]，但并没有性能优良、实现方便的解决方案，对于一些具体的问题，通常采用专家人工指定的方法来确定. 鉴于此，我们基于语义搭配知识库，采用统计的方法生成面向语义搭配关联强度推理的证据信任分配函数.

设 MI、$PD1$ 和 $PD2$ 为当前证据，构建证据三元组 $E(MI,PD1,PD2)$. 则对于任意词语搭配 x 通过计算 x 每个证据的值，构建 x 的证据三元组为 $E(v(MI),v(PD1),v(PD2))$(其中 $v(*)$ 为 x 的证据属性 $*$ 所对应的值)，对 x 的证据三元组中的任意一个证据分量 E_x($E_x\in\{MI,PD1,PD2\}$)，从语义搭配知识库中抽取所有 $v(p)\geqslant v(E_x)$ 的词语搭配样本，其中 $v(p)$ 为语义搭配知识库中词语搭配的 E_x 属性对应的值，将这些词语搭配样本构成样本集合 $P=\{p_1,p_2,\cdots,p_n\}$，计算样本集合中词语搭配 E_x 属性对应值的均值 $\overline{P}$，如式(6)所示：

$$\overline{P}=\frac{1}{n}\sum_{k=1}^{n}v(p_k) \tag{6}$$

再对 $\overline{P}$ 进行归一化处理，将其转化为 0 到 1 区间的值，如式(7)所示：

$$\widehat{P}=\frac{\overline{P}-\min}{\max-\min} \tag{7}$$

其中，max 和 min 分别为样本集合 P 中词语搭配的 E_x 属性对应值的最大值和最小值. 据此，构建证据 E_x 的初始分配即基本信任分配函数为 $m_{E_x}(S)$，如式(8)所示：

$$m_{E_x}(S)=\begin{cases}\dfrac{\gamma-1}{\gamma}\widehat{P}, & \text{当 } S=R\\ 1-\dfrac{\gamma-1}{\gamma}\widehat{P}, & \text{当 } S=N\\ 0, & \text{其他}\end{cases} \tag{8}$$

其中，γ 为信任参数，其选择方法将在后面的实验部分 7.1.3 节进行讨论.

5.2.2　证据的加权合成规则

对于证据 E_x 的特定词语搭配，我们通过基本信任分配函数获取其初始信度分配. 为了解决证据在融合的过程中出现证据冲突的问题，本文采用证据的可信度充当证据的权值，利用加权分配的证据合成方法对证据进行合成[24]. 具体的合成过程如下.

(1) 构建证据间的相似系数矩阵

两个证据之间的相似系数是用来度量证据之间的相似程度的，定义如下：设 E_1 和 E_2 为侦测框架 Θ 下的两个证据，其初始信任分配分别为 m_{E_1} 和 m_{E_2}，焦元分别为 A_i 和 B_j，则 E_1 和 E_2 的相似系数 d_{12} 计算如式(9)所示：

$$d_{12}=\frac{\sum\limits_{A_i\cap B_j\neq\varnothing}m_{E_1}(A_i)m_{E_2}(B_j)}{\sqrt{\left(\sum m_{E_1}^2(A_i)\right)\left(\sum m_{E_2}^2(B_j)\right)}} \tag{9}$$

证据的相似系数 d_{12} 取值范围为[0,1]，其值越接近于 1，表示两个证据的确定性越好，其值越接近于 0，表示两个证据的冲突越厉害.

在本文中，证据属性有 3 个，由式(9)可以计算出证据三元组 $E(MI,PD1,PD2)$ 中任意两个证据的相似系数，将其表示为相似系数矩阵如式(10)所示：

$$\begin{bmatrix}1 & d_{12} & d_{13}\\ d_{21} & 1 & d_{23}\\ d_{31} & d_{32} & 1\end{bmatrix} \tag{10}$$

在式(10)表示的相似度矩阵中，对角线为 1 表示证据对自身的相似度为 1，而非对角线元素是对称相同的($d_{ij}=d_{ji}$)，表示两个证据的相似度是对称相等的.

(2) 证据的可信度表示

将相似矩阵的每行相加即可得到各个证据对证据 E_x 的支持度，如式(11)所示：

$$Sup(E_x)=\sum_{j=1}^{3}d_{ij},\ i=\begin{cases}1,\text{若证据 }E_x\text{为 }MI\\2,\text{若证据 }E_x\text{为 }PD1\\3,\text{若证据 }E_x\text{为 }PD2\end{cases}\tag{11}$$

支持度 $Sup(E_x)$表示证据 E_x被其它证据所支持的程度.如果其它证据和证据 E_x都比较相似,则认为其它证据和证据 E_x之间相互支持的程度也非常高;相反,如果其它证据和证据 E_x之间的相似程度都比较低,则认为其它证据和证据 E_x之间相互支持程度也比较低.

证据 E_x的可信度定义为将支持度进行归一化处理后的结果,如式(12)所示:

$$Crd(E_x)=\frac{Sup(E_x)}{Sup(E_{MI})+Sup(E_{PD1})+Sup(E_{PD2})}\tag{12}$$

可信度 $Crd(E_x)$是对证据 E_x的可信程度的一种表征.一般来说,如果证据 E_x越可信,其可信度就越大,其被其它证据所支持的程度就越高;如果证据 E_x可信度越低,则该证据就被其它证据所支持的程度就越低.

(3)证据属性的加权

从式(12)可以看出各个证据的可信度之和为 1,即 $Crd(E_{MI})+Crd(E_{PD1})+Crd(E_{PD2})=1$.因此,证据的可信度就可以用来作为证据的权重对证据进行加权平均,如式(13)所示:

$$m_c(S)=Crd(E_{MI})\times m_{E_{MI}}(S)+Crd(E_{PD1})\times m_{E_{PD1}}(S)+Crd(E_{PD2})\times m_{E_{PD2}}(S)\tag{13}$$

$m_c(S)$表示证据三元组 $E(MI,PD1,PD2)$对待评估语义搭配的关联强度为 S 的平均加权证据的基本分配值.

(4)证据属性的加权合成

D-S 合成规则表示的是多个证据联合作用的法则,我们使用 D-S 合成规则对证据三元组 $E(MI,PD1,PD2)$的平均加权证据进行融合.设在侦测框架 Θ 下的证据三元组 $E(MI,PD1,PD2)$的加权分配值为 $m_c(A_i)$和 $m_c(B_j)$,其中 A_i和 B_j为焦元,具体融合规则如式(14)所示:

$$m(S)=\frac{\sum_{A_i\cap B_j\neq\varnothing}m_c(A_i)m_c(B_j)}{1-\sum_{A_i\cap B_j\neq\varnothing}m_c(A_i)m_c(B_j)}\tag{14}$$

其中,$m(S)$为平均加权证据经一次合成得到待评估语义搭配关联强度为 S 的相关程度.由于本文证据有 3 个,所以只需对加权证据进行两次合成,即可得到待评估语义搭配关联强度为 R 和 N 的相关程度 $m(R)$和 $m(N)$,且 $m(R)+m(N)=1$.我们根据 $m(R)$和 $m(N)$的大小即可判断词语的语义搭配是否正确.

6 语义错误侦测算法

语义错误是指那些从语义学角度考查,不符合语义搭配规范的错误.在语义搭配知识库的基础上,可以较好地确立可能的语义错误搭配;对于可能的语义错误搭配,利用证据理论的不确定性推理方法能够很好的确定语义搭配的关联强度,以此判断是否存在语义搭配错误.本文以 3 层语义搭配知识库和证据理论为基础,构建了基于语义搭配知识库和证据理论的语义错误侦测算法.具体的算法过程描述如算法 3 所示.

算法 3. 基于语义搭配知识库和证据理论的语义错误侦测算法.

输入:待校对文本 T;语义搭配知识库

输出:待校对文本校对结果

过程:

1. 将 T 用中科院分词工具进行分词预处理;

2. 对 T 进行逐句扫描,抽取两个标点符号之间的短句赋给 P;

3. 获取 P 的搭配列表 L;

4. 将 L 进行逐词扫描,获取当前词语搭配赋值给 W,若 W 是量词+名词搭配则执行下一步,否则跳转至步骤 6;

5. 对于量词+名词搭配,将名词转换成义原,查询量词-义原知识库,如果存在,判定搭配为正确,跳转至步骤 9,否则,跳转至步骤 8;

6. 将中心词,搭配词分别转换成义原,查询义原-义原搭配知识库,如果存在,判定搭配为正确,跳转至步骤 9,否则执行下一步;

7. 将中心词转换成义原,查询中心词-义原知识库,如果存在,判定搭配为正确,跳转至步骤 9,否则,将搭配词转换成义原,查询搭配词-义原知识库,如果存在,判定搭配为正确,跳转至步骤 9,否则执行下一步;

8. 查询词语搭配知识库,如果存在,判定搭配为正确,否则,将该搭配加入到 wL 并记录相关位置信息;

9. 判断 L 的当前扫描位置是否为末尾,如果是则向下执行,否则跳转至步骤 4;

10. 判断 T 的当前扫描位置是否为末尾,如果是则向下执行,否则跳转至步骤 2;

11. 对 wL 进行逐词扫描,将 wL 当前词语赋值给给 C;

12. 计算 C 的 MI,$PD1$ 和 $PD2$,根据式(8)构造证据信任分配函数,确立证据 MI,$PD1$ 和 $PD2$ 的基本信任分配;

13. 利用式(9)计算证据的相似系数,构建相似系数矩阵,利用式(11)、式(12)计算证据的可信度,利用式(13)对证据的初始信任进行加权平均计算;

14. 使用式(14)对证据 MI，$PD1$ 和 $PD2$ 的平均加权证据进行融合，获取语义搭配相关度为 R 和 N 的相关度大小 mR 和 mN；

15. 比较 mR 和 mN 的大小，若 $mR>mN$，在 wL 中将 C 删除；

16. 判断 wL 的当前扫描位置是否为末尾，如果是则向下执行，否则跳转至步骤 11；

17. 将 wL 依次读出并根据位置信息在校对文本中标红，结束.

7　实　验

为了验证本文构建的语义搭配知识库以及基于该知识库和证据理论的语义错误侦测模型，我们设计了相关实验. 实验主要包括 3 个方面：

(1) 相关参数的确定；

(2) 语义搭配知识库构建评估；

(3) 错误侦测算法评估.

实验的训练语料选择了北京大学计算语言学研究所发布的 1998 年《人民日报》基本标注语料库(约 2600 万字). 该语料属于新闻性质的语言，语句规范、用词标准，而且全部语料已完成了校对、分词和词性标注等基本加工，基本满足词语搭配抽取的需求.

7.1　相关参数的确定

7.1.1　互信息与共现频次的阈值

为了确定互信息和共现频次的阈值，将词语搭配知识库中所有搭配(共计 185 173 个)的互信息和共现频次提取出来，构成了一个 2×185 173 的矩阵，将矩阵中的数据进行区间化处理，根据它们在不同区间的分布密度，来选择互信息和共现频次的阈值. 区间粒度的大小决定了阈值选择的精确度. 经过实验观察，将数据区间均分为 50 等份时，所得到的阈值对词语搭配的正确性判断具有较好的区分效果. 于是我们采用 MATLAB 将 2×185 173 的矩阵归一化为一个 50×50 的矩阵，矩阵中的每个值为互信息和共现频次相对应的区间范围内的词语搭配个数，采用 MATLAB 绘制 50×50 矩阵的密度分布图如图 3 所示，互信息与共现频次的密度值对词语搭配覆盖的趋势图如图 4 所示.

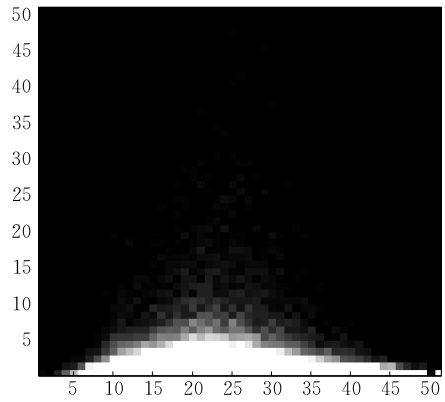

图 3　互信息与共现频次密度矩阵分布图

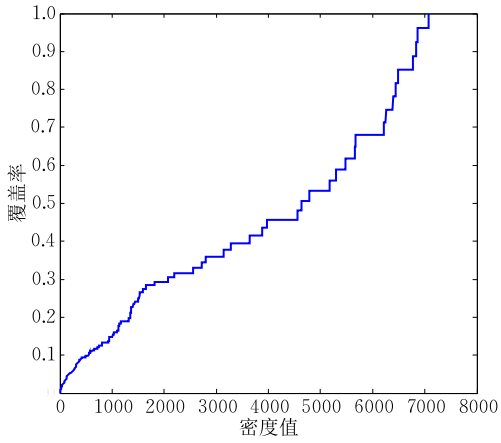

图 4　互信息与共现频次的密度值对词语搭配覆盖的趋势图

通过对图 3、图 4 的分析可得，当密度值为 552 时，其对应密度覆盖率为 9.76%，通过密度矩阵分布图所选择的阈值具有较好的区分度. 我们将密度值 552 转化为互信息和共现频次的对应区间为[2.6，3.4]和[1.2，1.8]. 基于此，我们可以将 4.1 节中的互信息和共现词频的阈值 τ_1，τ_2 分别设置为 3.4 和 2. 经过人工分析发现，采用上述方法所选择的阈值是合理的.

7.1.2　词语搭配聚合度的阈值

为了获取词语搭配聚合度的阈值，我们采用与获取互信息与共现频次阈值相同的方法，将图 2 中的义原-义原知识库中的所有搭配(共 128 211 个)的 PD1 和 PD2 提取出来，采用 MATLAB 绘制其密度矩阵分布图如图 5 所示，聚合度的密度值对义原搭配覆盖的趋势图如图 6 所示.

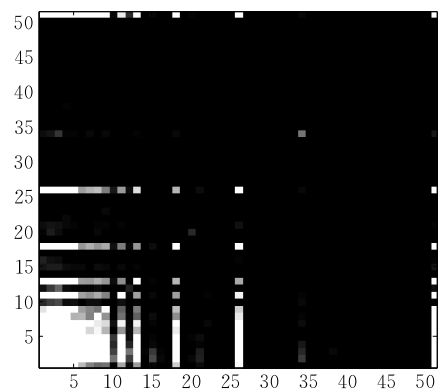

图 5　聚合度密度矩阵分布图

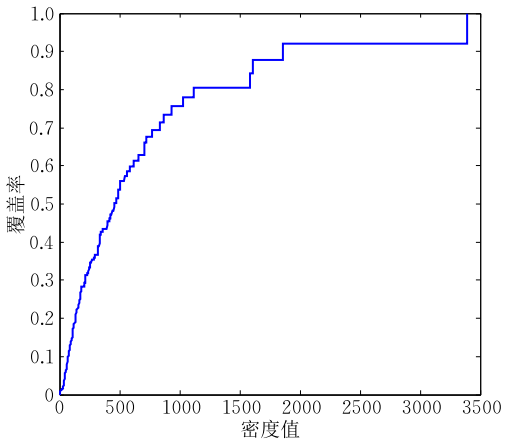

图 6 聚合度的密度值对义原搭配覆盖的趋势图

同样,通过分析图 5 和图 6 可得,当密度值为 87 时,其对应密度覆盖率为 12.52%,通过密度矩阵分布图所选择的阈值具有较好的区分度.我们将密度值 87 转化为聚合度的对应区间为[0.023,0.034],基于此,我们可以将 4.2.1 节中的聚合度的阈值 λ 设置为 0.034.

7.1.3 基本信任分配函数的信任参数

我们通过对基本信任分配函数式(8)分析可知:当互信息和聚合度分别取其对应的阈值 3.4 和 0.034 时,它们的基本信任分配函数应该服从如下关系:$m_{E_x}(R)\approx m_{E_x}(N)\approx 0.5$,通过对上式进行相关计算可得 5.2.1 节的互信息的信任参数 γ 为 8,聚合度的信任参数 γ 为 47.

7.2 语义搭配知识库构建评估

对语义搭配知识库的评估主要包含两个方面:

(1)词语搭配是否正确的转化为语义表示.该指标以语义搭配转化为词语搭配后的准确率 A 来衡量.准确率 A 的定义如式(15)所示:

$$A=\frac{R}{T}\times 100\% \tag{15}$$

其中,R 表示语义搭配转化为词语搭配后正确搭配的个数;T 表示语义搭配转化为词语搭配后的总个数.

(2)对有限的词语搭配在语义层次是否进行了有度量的扩展.该指标以语义扩展的程度即扩展度 E 来衡量.扩展度 E 的定义如式(16)所示:

$$E=\frac{P}{O} \tag{16}$$

其中:P 表示语义搭配转化为词语搭配的个数;O 表示词语搭配知识库中与当前语义搭配相对应的原有搭配的个数.

本文采用算法 1、算法 2 构建语义搭配知识库,各个知识库提取搭配数如表 6 所示.

表 6 知识库记录数详情表

知识库	记录数
词语搭配知识库	185173
词语-义原搭配知识库	260474
义原-义原搭配知识库	128211

由于知识库中记录比较多,篇幅有限,在这里不能将全部搭配进行准确度和扩展度分析,仅以部分搭配为例进行分析,部分分析结果如表 7 所示.

表 7 部分搭配准确度和扩展度分析

词语	搭配词词性	搭配词个数	搭配语义个数	搭配语义正确个数	扩展度	准确度/%
故事	A	6	176	153	29.33	86.93
成就	A	6	286	259	47.67	90.56
现象	A	16	310	279	19.38	90.00
漂亮	D	3	135	119	45.00	88.15
方便	D	6	151	134	25.17	88.74
薄弱	D	4	74	65	18.50	87.84
突破	N	26	176	155	6.77	88.07
继承	N	7	170	151	24.19	88.82
宣布	N	49	673	525	13.73	78.01
发展	D	9	33	30	3.67	90.90
提高	D	13	133	114	10.23	85.71
提醒	D	7	109	101	15.57	92.66
均值					21.60	88.03

相对于基于框架的提取方法[25](准确率为 84.73%)有显著的提高.其中,动词的“突破”,名词的“故事”、形容词的“漂亮”的词语搭配的准确度和扩展度具体分析过程如下:

动词“突破”,知识库中与之搭配的名词共有 26 个,共有 26 个语义信息,如表 8 所示,这些语义信息转化为词语共有 176 个,如表 9 所示,其中有 155 个

表 8 “突破”的名词语义搭配

比率-数量-、场所-商业-总-买-卖-、地方-培养-农-栽植-、多少-场所-商业-卖-、多少-货币-卖-、多少-钱财-赚-、多少-人工物-庄稼-、多少-商业-表演物-卖-票证-、多少-商业-人工物-卖-、多少-设施-停留-交通工具-运送-、货币-捐-、价值-工-人工物-制造-、价值-工-人工物-制造-频度值-时间-年-单-、价值-人工物-庄稼-全额-、境况-群体-人-悲惨-、钱财-金融-赚-、情感-相信-、人-利用-、容积-无生物-、数量-物质-、速度-挖掘-、位置-形成-破开-、位置-中位-、运动器材-圆-、真-、重要-

表 9 “突破”的名词语义搭配转化的词语

千分点、血球容积率、百分比、百分点、百分率、百分数、比、比例、比率、比值、比重、成、成数、反比、反比例、正比、正比例、复比、连比、率、总店、垦区、营业额、销售额、顺差、生产量、产量、上座率、票房、贸易额、贸易量、年销售额、销售量、销量、吞吐量、善款、赠金、赠款、捐款、产值、年产值、总产值、惨境、独木桥、两难的境地、苦海、困处、困境、困局、骑虎之势、穷蹙、水深火热、油水、余利、盈利、盈余、息、收益、贴息、利、利钱、利润、利润回报、利息、复利、回报、赚头、子金、定息、封建迷信、信、信任、信任感、信心、把握、用户、容积、容量、当量、进尺、裂缝、裂痕、裂口、裂纹、裂隙、漏洞、漏缝、漏孔、空洞、空子、孔、孔洞、孔隙、孔穴、孔眼、孔眼线、口、口子、窟、窟窿、窟窿眼、窟窿眼儿、缝、缝隙、缝子、洞、洞窟、洞眼、洞子、夹缝、鸿沟、罅、罅缝、罅隙、堀、窦、壑、朕、扯裂、眼、眼儿、眼孔、砂眼、缺门、溶洞、气孔、堑、窍、隙、隙缝、隙口、隙罅、突破口、塌陷处、无底洞、腰、当中、半道、当腰、半路、半路上、半途、半腰、半中间、半中腰、当央、当间儿、正当中、正中、中、中部、中间、中路、中途、中心、中央、居间、居中、链球、橄榄球、高尔夫球、冰球、保龄球、手球、台球、球、曲棍球、马球、真实、戏剧性

可以与动词“突破”进行搭配.因此：

扩展度为：$E=176/26=6.77$；

准确度为：$A=155/176\times100\%=88.07\%$.

名词“故事”，知识库中与之搭配的形容词共有6个，共有5个语义信息，如表10所示，这5个语义信息转化为词语共有176个，如表11所示，其中有153个可以与名词“故事”进行搭配.因此：

扩展度为：$E=176/6=29.33$；

准确度为：$A=153/176\times100\%=86.93\%$.

表10　“故事”的形容词语义搭配

小-、能-感动-、真-、能-促使-喜欢-、趣-

表11　“故事”的形容词语义转化的词语

美、美好、喜闻乐见、有目共赏、对劲儿、更加喜欢、更为理想、乖巧、好、合意、可人、感人、令人鼓舞、令人兴奋、激动人心、震撼人心、振奋人心、回肠荡气、鼓舞人心、催人奋进、哀壮、缠绵、缠绵哀怨、缠绵悱恻、动人、动人心魄、动人心弦、荡气回肠、深情、声情并茂、生动、生动活泼、生花妙笔、传神、情文并茂、饶有风趣、饶有兴味、饶有兴致、热火、入味、妙趣横生、明快生动、趣、趣味盎然、招笑儿、带修辞色彩、幽默、幽默滑稽、有趣、有声有色、有血有肉、有意思、有滋味、有滋有味、消遣、谐、谐趣横生、兴趣盎然、形象地、动态各异、动听、逗、逗人、成趣、出神入化、百读不厌、别有情趣、跌荡、跌宕、多彩、多姿、多姿多采、发噱、风趣、好看、好玩、好玩儿、好笑、绘声绘色、绘声绘影、绘影绘声、活、活泼、滑稽、作为消遣、俳谐、诙谐、哏、噱头、栩栩、栩栩如活、栩栩如生、龙飞凤舞、可乐、娓娓动听、可笑、扣人心弦、来劲、灵动、脞、蕞、娇小、短小、袖珍、袖珍型、蝇头、小、小号、小件、小型、细、纤、微、微观、微量、微型、迷你、迷你型、毛、属实、实、实际、实际上、实在、实则、实质性、事实上、无可置疑、铁案如山、无容置疑、名不虚传、名副其实、名实相符、千真万确、确、确切、确切无疑、确确实实、确实、确实可靠、确凿、确凿无疑、无疑、现实、响当当、凿、凿凿、有案可稽、有目共睹、道道地地、道地、不容怀疑、不容置疑、地道、毫无疑义、挨边、真、真格的、真确、真实、真实可信、真性、真真、真真实实、真正、活生生、货真价实、洵、笃实、结论性、结论性地、记实、正宗、灼然无疑、可信、肫

形容词“漂亮”，知识库中与之搭配的副词有3个，共有3个语义信息(很-、较-、极-).这3个语义信息转化为词语有135个，如表12所示，其中有119个可以与形容词“漂亮”进行搭配.因此：

扩展度为：$E=135/3=45.00$；

准确度为：$A=119/135\times100\%=88.15\%$.

表12　“美丽”的副词语义搭配转化的词语

大大地、颇、颇为、甚、实在、太、太甚、特、特别、尤、尤其、尤为、尤以、在很大程度上、倍儿、大为、不过、不少、不胜、多、多多、多加、出奇、出奇地、分外、格外、格外地、够、够瞧的、好、好不、很、很是、正经、着实、碜、可、老、老大、良、不得了、酷、截然、綦、剧、惊人地、绝、绝顶、绝对、最为、最最、幡然、之极、之至、至极、极、极大、极大地、极度、极度地、极端、极其、极为、疯狂地、概、非常、非常非常、不亦乐乎、彻底地、不堪、不可开交、到底、倍加、备、贼、异常、逾常、完全、无可估量、万、万般、万分、无以复加、无以伦比、要命、要死、已极、已甚、死、无比、深深、深深地、甚为、十二分、十分、奇、满、满贯、满心、那般、那么、日益、益、益发、愈、愈…愈、愈发、愈加、愈来愈、愈益、远、远远、越、越…越、越发、越加、越来越、越是、大不了、尤甚、逾、比较、多、更、更加、更进一步、更其、更为、还、还要、这样、较、较比、较为、进一步

7.3　错误侦测算法的实现与评估

为了评估本文提出的语义错误侦测算法，我们基于Microsoft Visual Studio 2010和Microsoft SQL Server 2008平台，采用C#编程语言，对所提出的算法进行编程实现，完成了一个面向文本语义错误侦测的智能信息处理平台，并采用一定的测试语料对算法的性能进行评估.所开发的平台如图7所示.

图7　智能信息处理平台语义错误侦测

由于没有语义错误侦测方面的公共测试数据集，目前国内外有关中文文本校对的研究都是基于自己构建的测试数据进行算法测试.基于此，我们整合了200条小学语文的病句修改试题和网上的20篇小学生作文，构建了一个包含有340个语义错误测试点的语义错误测试集.使用所开发的智能信息处理平台对所提出的算法进行测试，同时，为了和相关研究进行比较，我们也对相关文献[14,16]所介绍的方法进行了实现，并用我们构建的语义错误测试集进行测试，发现测试效果均不如文献中所列的实验结果好，故这里采用文献中介绍的实验结果与我们的实验结果进行对比，对比结果如表13所示.

表13　实验结果比较

方法	召回率 R/%	准确率 P/%	F 值/%
罗振声等方法[14]	91.1	62.3	74.00
程显毅等方法[16]	62.3	77.4	69.03
本文方法	93.3	83.3	88.02

罗振声等人的方法召回率偏高主要是由于其采用的测试语料中不仅含有语义错误，而且还含构词错误和句法错误，并且其语义错误在测试语料中所占比例不大，仅为整个错误的10%.程显毅等的方法采用HNC句类分析系统，在命名实体识别方面有一定的优势，对命名实体的有效识别使其准确率偏高.

为了进一步验证本文提出的模型在书刊校对方

面的应用效果，我们收集了本实验室近几年来的15篇硕士论文初稿，对其中的语义搭配错误进行检查，发现了354个语义错误点，通过人工校对，其中真正的语义错误为267个，准确率为75.42%，为进一步确定校对的召回率，我们对其中的两篇论文进行了精读，发现实际存在的语义搭配错误48个，其中，侦测出的语义错误40个，召回率为83.33%.所得准确率和召回率均稍低于表13中的实验数据，是因为所选择的硕士论文涉及到与计算机相关的一些内容，而本文的语义错误侦测模型的训练是以《人民日报》标注语料的词语搭配为基础的.

从表13的实验结果来看，由于引入了证据理论的思想，使得算法的召回率有了明显地提高.同时在提取词语搭配的过程中充分考虑到了连词“和”、“或”以及顿号的影响，使得构建的语义搭配知识库更加完善和准确，这对准确率的提高也做出了不小的贡献.

8 结束语

本文详细介绍了三层语义搭配知识库的构建模型、基于多源证据融合的语义搭配关联强度评估模型以及基于上述两者的中文文本语义错误侦测模型.本文利用《HowNet》对词语搭配进行语义信息的提取，并以词语搭配聚合度进行辅助过滤，同时，使用证据理论的方法进行语义搭配错误的判定.该研究方法在国内相关领域的研究中还鲜有涉及，为语义错误的侦测研究提供了新的思路.

不过，在以上研究过程中，只考虑了简单句型，对于复杂句型所产生的错误考虑不够；也未对文本进行句法结构分析，只是进行了语义搭配的提取；对于语言学知识的运用，只考虑了词语之间的搭配，没有考虑词语之间的语义限制.这些都使得语义搭配知识库中出现了一些错误的搭配，进而对语义错误侦测的准确率造成了影响.

同时，本文在构建语义搭配知识库时，选用了标准化的《人民日报》语料，所构建的语义错误侦测模型和算法对于口语化较严重的社交网络文本来说，侦测的效果并不理想.我们原本认为，社交网络文本中的许多语义搭配错误，大多是由人们为表达某种特殊含义而故意使用错误字词或同音别字(例如，“楼主你杯具了”，“坐看杯具的诞生”)而导致.不过，目前已有人开始对社交网络文本进行校对方面的研究[26]，在今后的工作中，我们将加强这方面的研究.另外，我们也将会在句法分析、词语及短语块的语法限制和多个词语搭配等方面开展研究，对语义搭配的证据理论进行深入的挖掘，继续完善信任分配函数的构造方法以及证据冲突解决策略的研究，进一步提高语义搭配知识库的正确性和语义错误侦测的准确性.

参 考 文 献

[1] Kukich K. Techniques for automatically correcting words in text. ACM Computing Surveys, 1992, 24(4): 377-439

[2] Islam A, Inkpen D. Real-word spelling correction using google web IT 3-grams//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, 2009: 1241-1249

[3] Xu W, Tetreault J, Chodorow M, et al. Exploiting syntactic and distributional information for spelling correction with web-scale *n*-gram models//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011:1291-1300

[4] Sun X, Shrivastava A, Li P. Fast multi-task learning for query spelling correction//Proceedings of the 21st ACM International Conference on Information and Knowledge. New York, USA, 2012: 285-294

[5] Huang Y, Murphey Y L, Ge Y. Intelligent typo correction for text mining through machine learning. International Journal of Knowledge Engineering and Data Mining, 2015, 3(2): 115-142

[6] Yeh J F, Chen W Y, Su M C. Chinese spelling checker based on an inverted index list with a rescoring mechanism. ACM Transactions on Asian and Low-Resource Language Information Processing, 2015, 14(4): 1-28

[7] Luo Wei-Hua, Luo Zhen-Sheng, Gong Xiao-Jin. Study of techniques of automatic proofreading for Chinese texts. Journal of Computer Research and Development, 2004, 41(1): 244-249(in Chinese)
(骆卫华，罗振声，宫小瑾. 中文文本自动校对技术的研究. 计算机研究与发展，2004，41(1)：244-249)

[8] Xiong J H, Zhao Q, Hou J P, et al. Extended HMM and ranking models for Chinese spelling correction//Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing. WuHan, China, 2014: 133-138

[9] Han D X, Chang B B. A maximum entropy approach to Chinese spelling check//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan, 2013: 74-78

[10] Zhang Yang-Sen, Cao Yuan-Da, Yu Shi-Wen. A hybrid model

of combining rule-based and statistical-based approaches for automatic detecting errors in Chinese Text. Journal of Chinese Information Processing, 2006, 20(4): 3-9(in Chinese)
(张仰森，曹元大，俞士汶. 基于规则与统计相结合的中文文本自动查错模型与算法. 中文信息学报，2006，20(4)：3-9)

[11] Liu X D, Cheng F, Duh K, Matsumoto Y. A hybrid ranking approach to Chinese spelling check. ACM Transactions on Asian and Low-Resource Language Information Processing, 2015, 14(4): 1-17

[12] Chang T H, Chen H C, Tseng Y H, Zheng J L. Automatic detection and correction for Chinese misspelled words using phonological and orthographic similarities//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan, 2013: 97-101

[13] Yin Bang-Cai. Study of "the possibility of semantic collocation". Theoretic Observation, 2008(6): 134-135(in Chinese)
(尹邦才. 试论"语义搭配的可能性". 理论观察，2008(6)：134-135)

[14] Luo W H, Luo Z S, Gong X J. Semantic error checking in automatic proofreading for Chinese texts//Proceedings of the 2002 IEEE International Conference on Systems, Man and Cybernetics. Hammamet, Tunisia, 2002: 5-10

[15] Zheng Feng-Bin, Chen Zhi-Guo, Jiang Bao-Qing, Qiao Bao-Jun. Fuzzy matching technique by sentence skeleton in semantic collation system. Acta Electronica Sinica, 2003, 31(8): 1138-1140(in Chinese)
(郑逢斌，陈志国，姜保庆，乔保军. 语义校对系统中的句子语义骨架模糊匹配算法. 电子学报，2003，31(8)：1138-1140)

[16] Cheng Xian-Yi, Sun Ping, Zhu Qian. The research of Chinese text proofreading system model based on HNC. Microelectronics & Computer, 2009(10): 49-52(in Chinese)
(程显毅，孙萍，朱倩. 基于 HNC 的中文文本校对系统模型的研究. 微电子学与计算机，2009(10)：49-52)

[17] Florentin S, Jean D. Advances and Applications of DSmT for Information Fusion. Rehoboth: American Research Press, 2009

[18] Sevastianov P, Dymova L, Bartosiewicz P. A framework for rule-base evidential reasoning in the interval setting applied to diagnosing type 2 diabetes. Expert Systems with Applications, 2012, 39(4): 4190-4200

[19] Mokhtari K, Ren J, Roberts C, Wang J. Decision support framework for risk management on sea ports and terminals using fuzzy set theory and evidential reasoning approach. Expert Systems with Applications, 2012, 39(5): 5087-5103

[20] Quan Chang-Qin, He Ting-Ting, Ji Dong-Hong, Liu Hui. Chinese WSD Based on Selecting the Best Seeds from Collations. Journal of Chinese Information Processing, 2005, 19(1): 30-35(in Chinese)
(全昌勤，何婷婷，姬东鸿，刘辉. 从搭配知识获取最优种子的词义消歧方法. 中文信息学报，2005，19(1)：30-35)

[21] Han Chong-Zhao, Zhu Hong-Yan, Duan Zhan-Sheng. Multi-Source Information Fusion. 2nd Edition. Beijing: Tsinghua University Press, 2010(in Chinese)
(韩崇昭，朱洪艳，段战胜. 多源信息融合. 第 2 版. 北京：清华大学出版社，2010)

[22] Dezert J, Liu Z G, Mercier G. Edge detection in color images based on DSmT//Proceedings of the 2011 International Conference on Information Fusion. Chicago, USA, 2011: 969-976

[23] Han D Q, Deng Y, Han C Z. Novel approaches for the transformation of fuzzy membership function into basic probability assignment based on uncertain optimization. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2013, 21(2): 289-322

[24] Zhao Qiu-Yue, Zuo Wan-Li, Tian Zhong-Sheng, Wang Ying. A method for assessment of trust relationship strength based on the improved D-S evidence theory. Chinese Journal of Computers, 2014, 37(4): 873-883(in Chinese)
(赵秋月，左万利，田中生，王英. 一种基于改进 D-S 证据理论的信任关系强度评估方法研究. 计算机学报，2014，37(4)：873-883)

[25] Qu Wei-Guang, Chen Xiao-He, Ji Gen-Lin. A frame-based approach to Chinese collocation automatic extracting. Computer Engineering, 2004, 30(23): 22-24(in Chinese)
(曲维光，陈小荷，吉根林. 基于框架的词语搭配自动抽取方法. 计算机工程，2004，30(23)：22-24)

[26] Zhang Xin. Research and Implementation of Chinese Text Proofing Method for Social Media[M. S. dissertation]. Heilongjiang University, Harbin, 2015(in Chinese)
(张鑫. 面向社会媒体的中文文本校对方法研究与实现[硕士学位论文]. 黑龙江大学，哈尔滨，2015)

ZHANG Yang-Sen, born in 1962, Ph. D., professor. His major research interests include natural language processing and artificial intelligence.

ZHENG Jia, born in 1991, M. S. candidate. His major research interest is natural language processing.

Background

With the rapid development of the publishing industry and the gradual popularization of electronic books and periodicals, the automatic error detection technology for Chinese text has put forward higher requirements.

Semantic error detecting for Chinese text is a hot topic in research of natural language processing, and is concerned by the industrial and the academia around the world. At present, the research on error detection of the lexical level and syntax level for Chinese text are relatively adequate, and has achieved good results. However, the research on semantic level error detection is not deep enough.

In this paper, we proposed a semantic error detection model based on semantic collocation knowledge base and D-S theory, and obtained good results. The method proposed in this paper for semantic error detection yet few studies to involve. In contrast to other experimental systems, our approach shows significant advantages.

In the last few years, our research team have conducted a thorough research on the automatic error detection technology for Chinese text. In the process of research, we constructed a large scale knowledge base for automatic error detection, and constructed an intelligent information processing platform, which can effectively detect text errors for Chinese text in lexical level, syntax level, semantic level and political field. Some works of the research fields have been published in international journals, domestic journals and the international conferences, and has been used in the Foreign Ministry document processing system.

Our work is supported by the National Natural Science Foundation of China (Grant Nos. 61070119, 61370139) and the Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges under Beijing Municipality (Grant No. IDHT20130519).