

基于语义的英文短语检索与搭配推荐及其在辅助 ESL 学术写作中的应用

朱叶霜¹⁾ 喻 纯¹⁾ 史元春^{1),2)}

¹⁾(清华大学计算机科学与技术系 北京 100084)

²⁾(青海大学计算机技术与应用系 西宁 810016)

摘 要 学术英文写作对于母语非英语的(English as a Second Language, ESL)学者而言是一项挑战. 现有的 ESL 写作辅助系统均基于连续单词(n -gram)在语料库中出现的频率来构建,而忽略了语义信息对于 ESL 写作的作用. 文中提出基于语义的写作辅助方法,支持按照词性、搭配类型和概念扩展等语义条件来检索短语和自动推荐搭配. 此外,为了方便英语知识有限的 ESL 学者利用语义条件进行检索,文中还设计了易于理解的短语检索界面. 该界面提供包含语义修饰符的查询表达式,并可根据用户输入的部分查询表达式在线提示和补全. 最后,为验证语义方法的有效性,文中实现并部署了 ESLWriter 系统. 用户实验表明,ESLWriter 可以有效地推荐搭配并提供短语检索结果,其查询界面直观易用;ESL 学者通过混合使用自动搭配推荐功能和短语检索功能,写作质量和写作信心得到有效提高. 这些结果证明了语义信息在 ESL 写作辅助中的重要作用.

关键词 ESL 写作;语义;查询表达;搭配;语料库;社交网络;慕课;人机交互

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2016.00822

Semantic-Based ESL Phrase Retrieval and Collocation Suggestion

ZHU Ye-Shuang¹⁾ YU Chun¹⁾ SHI Yuan-Chun^{1),2)}

¹⁾(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

²⁾(Department of Computer Technology and Application, Qinghai University, Xining 810016)

Abstract Academic writing in English is a challenge for scholars who use English as a Second Language (ESL). Existing English writing assistant systems are usually based on the frequency of consecutive words (n -gram) in the corpus, but ignore the role of semantic information in ESL writing. In this research, we propose a semantic approach to support ESL writing, which allows users to retrieve corpus phrases and automatically makes collocation suggestions using a number of semantic constraints, including part-of-speech, collocation type, and conceptual expansion. Furthermore, to address the limited linguistic knowledge of general ESL users, we design an easy-to-understand query interface for phrase retrieval, which supports query with semantic modifiers, and can automatically suggest query expressions according to users' input. To examine the validity of our semantic approach, we implement a real running system ESLWriter. User experiments show that ESLWriter is helpful in phrase retrieval and collocation suggestion, and the query interface for end users is intuitive and easy-to-use. Participants can significantly improve their writing quality and confidence by mixed use of the two functions. These results confirm the effectiveness of incorporating semantic information into ESL writing system.

Keywords ESL writing; semantics; query expression; collocation; corpus; social networks; massive open online course (MOOC); human computer interaction

收稿日期:2014-12-06;在线出版日期:2015-11-18. 朱叶霜,男,1990年生,博士研究生,中国计算机学会(CCF)学生会会员,主要研究方向为人机交互. E-mail: yeshuangzhu@gmail.com. 喻 纯,男,1984年生,博士,中国计算机学会(CCF)学生会会员,主要研究方向为图形用户界面、交互式桌面等. 史元春,女,1967年生,博士,教授,博士生导师,中国计算机学会(CCF)常务理事、高级会员,主要研究领域为人机交互、多媒体、普适计算等.

1 引 言

学术英文写作对于母语非英语的(English as a Second/Foreign Language, ESL/EFL, 为简便起见以下统称 ESL)学者而言是一项富有挑战性的任务. ESL 学者往往可以熟练使用自己的母语进行写作, 而对英语的接触相对匮乏导致他们在用英文进行学术写作时效率低, 信心不足, 容易出现词语搭配错误等 ESL 特有的问题. 对使用母语与非母语写作差异的研究表明, ESL 学者与以英语为母语的(native)学者这两类人群的写作错误在类型和类型分布上存在很大差别, 如表 1 所示^[1]. 可以看出, native 学者的写作错误以拼写和语法错误为主, 相对简单和基础, 通常称作 L1(First Language)错误; 而 ESL 学者除了这两类错误外, 还倾向于出现风格错误、搭配错误等独有类型的错误, 通常称作 L2(Second Language)错误.

表 1 ESL 学者与 native 学者部分写作错误比较

类别	native 学者的写作错误 (L1 错误)	ESL 学者的写作错误 (L2 错误)
拼写	近音词或近形词混淆	各种可能, 包括 L1 错误
语法	句子间不加分隔	主谓一致性、冠词介词使用
风格	极少	使用非正式用语
搭配	极少	多种多样

对 ESL 写作错误的进一步研究表明, L2 错误, 尤其是词汇搭配错误, 是 ESL 写作最大的困难之一^[2]. 据统计, 在中国英语学习者语料库 (Chinese Learner English Corpus, CLEC) 中约 30% 的错误与搭配有关^[3]; 并且即使对于有多年英语学习经验的 ESL 学者, 在用英文写作时搭配错误依然十分频繁^[4]. 所谓搭配, 可以从统计角度定义为同时出现的概率远高于各自偶然出现概率的两个词^[2], 它具有与文化背景相关和约定俗成的特性, 并不像语法一样遵循固定的规则. 例如在英语中, “吃药”应为 “take medicine” 而非 “eat medicine”, “浓茶”应当表达为 “strong tea” 而非 “heavy tea” 等等. 因此搭配知识只能通过长时间积累和日常频繁使用才能熟练掌握, 而 ESL 学者恰恰不具备这样的语言学习环境.

除了写作质量低, ESL 学者在写作行为上也体现出独有的特点. 一方面, 由于无法确信自己的用语正确且地道, 他们在写作过程中往往需要借助信息检索工具, 一种常见的行为模式是借助 Google 等

搜索引擎返回结果的多少来判断某种表达是否合理^[5]. 另一方面, 由于难以直接找到最切合的英文表达方式, ESL 学者通常会反复修订写作的内容^[6]. 然而, 目前常见的词典、翻译等语言工具并非专门为帮助 ESL 写作而设计的, 远不能满足 ESL 学者的写作需求, 例如用户难以确定工具给出的用法是否地道、是否合乎目标文体 (例如学术论文) 的用语规范等等. 这些写作过程中大量不确定的语言问题, 导致 ESL 学者的写作思路时常被语言问题打断, 降低了他们的写作效率和写作信心.

以上问题表明, ESL 学者迫切需要一款专用的写作辅助工具来改善写作质量、提升写作效率和信心. 为此, 国内外研究者尝试提出辅助 ESL 写作的方法和工具. 现有工作通常基于由大量地道英语句子构成的语料库, 利用语料库中的连续单词 (n -gram) 的语言模型来检索例句或推荐改错. 然而, 基于 n -gram 的短语检索方式无法按照词性、概念、搭配类型等语义相关的条件进行检索, 并且缺少直观易用的用户查询表达方式, 无法满足 ESL 学者实际写作时的查询需求. 而基于 n -gram 频率统计的写作自动检查和改正方法致力于解决拼写、语法层面的写作问题, 忽略用户文本的语义信息, 针对搭配推荐的准确率不高. 以上问题都大大限制了现有写作辅助系统的有效性.

本研究充分利用语义信息, 从语料库短语检索和搭配自动推荐两个方面入手研究辅助 ESL 写作的方法. 在短语检索方面, 本文设计了一套能够表达词性、搭配类型和概念关联等语义相关条件的机器层查询语言, 并为语言知识有限的 ESL 学者提供直观易用的查询界面. 该界面支持语义修饰的查询表达式以及查询样例的自动提示, 使用户能够表达较为复杂的查询请求. 在搭配推荐方面, 本文以单词的语义依存关系作为搭配推荐的基本单元, 提出将用户写作文本进行概念扩展的搭配推荐方法. 在上述方法的基础上, 本文收集了大量已发表的论文作为语料库, 实现并部署了实际运行的系统 ESLWriter. 为了验证语义方法的有效性, 本文选择辅助 ESL 研究者写作学术文章这一典型的应用场景开展了实地用户实验. 实验结果表明, 基于语义的写作辅助系统 ESLWriter 具有查询表达能力强大、搭配推荐准确率高、界面设计友好等特点, 能够有效提高 ESL 学者的写作质量和写作信心.

本文第 2 节介绍 ESL 写作辅助的相关工作; 第 3 节提出支持语义检索的机器层查询语言 and 用户层

查询表达方式;第4节提出扩展单词概念的搭配推荐方法;第5节介绍 ESLWriter 系统的设计与实现,并通过用户实验评估其有效性;第6节总结全文并给出未来的研究方向。

2 相关工作

目前国内外关于 ESL 写作的研究主要集中在3个方面:ESL 写作困难、自动化查错纠错技术以及语料检索工具。本节将从这3个方面总结相关工作。

2.1 ESL 写作困难研究

通过人工标注和分析 ESL 学生课堂作文语料库^[3,7-8]或者网站中的文本^[5],已有研究总结出常见的 ESL 写作错误集中在冠词、介词和搭配的选取,并分析了这些错误出现的原因。冠词错误占有 ESL 错误的 10%到 26%,因为对于许多非英语母语的学者(如汉语学者),其母语语法中并不存在定冠词和不定冠词,这些学者在用英语写作时会经常错用、漏用冠词^[9-11]。Gamon 等人^[12]对日本英语学者语料库中的写作错误进行分析,指出其中介词错误约占所有错误的 10%^[8]。英语词汇的搭配被认为是区别学者和学者写作质量的最重要的原因之一^[13];对中国英语学者语料库的研究表明,约 30%的写作错误与搭配使用有关^[3];并且即使是对于有多年英语学习经验的学者而言,在用英文写作时搭配错误依然十分频繁^[4]。这是由于英语词汇之间的搭配多是约定俗成的,并不像语法一样遵循固定的规则,只能通过多年的积累和日常的训练才能熟练掌握,而学者恰恰不具备这样的语言学习环境^[1]。

相关工作还研究了 ESL 学者在写作过程中使用现有工具的困难。虽然目前已有诸如 Word 等软件能够提供拼写、语法检查等语言辅助功能,但由于它们的设计主要面向 native 用户,因此并不能处理错误较多的 ESL 文本^[14]。在实际论文写作时,ESL 学者经常将网络文本作为一个大语料库,在搜索引擎上发起一系列带双引号的查询(即 n -gram 查询),通过返回结果的多少来判断已知的 n -gram 是否地道^[5,15-16]。Sha^[17]通过实验对比了写作时检索 Google 语料库与英语国家语料库(British National Corpus, BNC)的优劣,证明了使用 Google 检索可以有效地解决写作中的语言问题。然而,由于网络上的文本包含大量噪声,搜索引擎的查询表达方式以 n -gram 为主,且只能验证已知用法的使用频率而无

法推荐新的表达方式,因此这种检索方式存在很大的局限性^[1]。

2.2 自动化写作查错纠错技术

针对以上 ESL 常见写作错误,研究者们致力于构建自动化机器查错和改错的算法和系统,这些研究运用统计学习方法从包含写作错误及对应更正后文本的平行语料库中提取语言知识,并对用户写作提供预测和改正建议。

上下文拼写技术^[8]利用由 n -gram 训练的分类器来检查和改正用户的冠词和介词错误。短语翻译技术^[14]将写作推荐问题看作从错误 n -gram 到对应正确 n -gram 的统计机器翻译(Statistical Machine Translation, SMT)过程,该技术在检查和改正特定名词的数量词和冠词错误的实验中达到了 61.8%的准确率。Park 等人^[1]提出的 AwkChecker 系统主要针对 ESL 的搭配错误,它以 n -gram 在语料库中的频率分布作为语言模型,以用户写作的 n -gram 与语料库中 n -gram 的编辑距离作为 ESL 用户模型,定义了词汇搭配质量的评价函数,据此标注出用户文本中的潜在错误搭配并给出改正建议。FLOW 系统^[18]利用由中英平行语料库训练的 n -gram 翻译和改写模型,允许中国用户在写作和修改英文句子时输入中英混合的文本,系统根据中英文上下文推荐中文短语翻译或英文短语改写方式,解决英文单词的拼写和选词等问题。

2.3 语料检索工具

除了自动化更正特定写作错误的算法,研究者们也尝试直接通过例句来辅助 ESL 写作,这些语料检索方法同样基于语料库的 n -gram^[17,20-22]。

NETSPEAK^[20]定义了一套类正则表达式的查询语言来让用户检索语料库中的 n -gram 短语及例句。该查询语言支持通配、与或等模式匹配运算符,具备一定的表达能力。PENS^[22]支持用户在英文文档中直接输入汉语拼音,系统将汉语拼音翻译成中文单词,再对中文单词进行同义词扩展来检索双语语料库中的例句。它在检索过程中将目标短语的关键词依次将同义词展开,并利用由编辑距离度量的句式相关性对结果排序。尽管 PENS 系统利用了同义词扩展来检索语义相关的例句,但同义词仅仅是对单词语义的部分描述,实际情况中非同义词之间也可能存在或强或弱的语义关联,且单词与上下文之间还存在搭配关系等其他类型的语义关联,因此 PENS 对语义信息的利用尚不充分。

2.4 现有工作的不足

基于 n -gram 的写作辅助系统由于对语义信息利用不足, 具有以下缺点。

目前, n -gram 语料库检索工具并不支持按词性、概念、搭配关系等语义相关的条件查询, 无法满足用户实际写作时多样化的查询需求。ESL 学者经常需要检索表达某种语义的恰当用词、某个名词的修饰词等, 这些查询请求在有严格位置和顺序关系约束下的 n -gram 检索中难以表达。一方面, 由于现有方法没有利用单词自身的语义信息, 用户无法按照单词词性、单词表示的概念去检索短语。例如, 单词“present”具有多种词性和义项, 当用户搜索 n -gram 短语“present method”时无法明确表达实际的查询意图。另一方面, 现有方法也没有利用单词间的语义关联, 用户无法指明查询请求中单词之间的搭配关系。例如, 短语“propose a method”、“propose this method”、“the method we propose”均能为“提出方法” (“propose ... method”) 这一动宾搭配提供参考, 但它们难以用 n -gram 统一描述。

另外, 在 n -gram 写作推荐系统中, 用户提供短语在语料库中的出现频率是推荐的重要依据。但是, 由于 n -gram 仅表示相邻连续单词间的位序关系, 不直接表示单词本身概念及其相互关联(比如修饰)等语义信息, 因此, 按照 n -gram 频率统计检错和纠错的方法对于以搭配为主的 L2 写作问题作用有限。此外, 语料库中的句子均符合语法规则, 具有重要语义的实词前后往往与虚词(比如冠词“the”)相邻。这些虚词出现频率很高因而经常被推荐, 但是, 用户很多时候并不关心虚词的用法。例如, 在 NETSPEAK 系统中, 与“method”有关的频率最高的 2-gram 为“the method”、“this method”和“a method”, 并不包含具有实质语义的词语(比如用户需求是获得可以修饰“method”的形容词)。最后, 虽然基于错误到正确的平行语料库训练的自动纠错方法针对介词、冠词、数量词等特定问题可以获得较高的准确率, 但受限于平行语料库的规模和错误分布的稀疏性, 仅针对这些方面进行辅助远远不能满足 ESL 学者实际的写作需求。

综上所述, 已有研究受限于所采用的基于 n -gram 的方法, 着重解决拼写、语法等易于用规则描述的问题, 对于解决 ESL 写作中的选词、搭配等更为困难的语义问题效果不理想, 因此在应用到实际写作场景时有效性大大受限。就目前而言尚没有充分利用语义信息辅助 ESL 写作的实用方法和系统。

3 短语检索

3.1 问题定义

语料库短语检索是从用户指定的检索条件出发, 在语料库中检索满足条件的短语及对应的例句。本文将短语检索分为机器层的查询语言(query language)和用户层的查询表达(query expression)两个问题来分别解决:

(1) 机器层查询语言是系统对用户查询意图的精确描述, 它与语料库数据和计算任务紧密相关, 是检索功能的基础, 决定着检索系统的表达能力和检索效率。

(2) 用户层查询表达是指用户将查询意图传递给系统的方式, 如何设计既能满足 ESL 学者多样化的查询需求又能方便用户使用的查询表达方式, 是值得深入研究的问题。

3.2 ESL 写作的查询需求

为进一步了解 ESL 学者写作中的查询行为, 我们收集并分析了他们实际写作时的查询记录。

通过与多名 ESL 学者的交流, 我们了解到 Google (或 Google Scholar) 搜索引擎是他们写作时的首选检索工具, 网络文本中英文短语或搭配的使用情况是他们写文章时的重要依据, 因此他们的查询记录对于分析 ESL 写作查询需求具有重要参考价值。我们共收集了 3 名 ESL 学者的 447 条查询记录进行分析, 所有查询均通过英文表达。

用户通常需要多步检索才能找到目标短语, 即对应同一目标短语(或查询意图), 存在一个查询会话(session)。根据目标短语关键词之间是否存在语义关系, 会话可分为基于 n -gram 的短语查询(简称 n -gram 查询, 例如“so as to”)和基于语义搭配的短语查询(简称搭配查询, 例如“visual angle”)两类。我们排除了仅含单个关键词的查询记录, 因为它们所占比例极小, 且查询意图均为与短语检索无关的拼写检查。经统计, n -gram 与搭配查询会话所占比例分别为 62.7% 和 37.3%, 这表明用户同时存在查询 n -gram 短语和语义搭配短语的需求。

我们采用会话包含的查询次数来表示会话长度。通过进一步考察两类查询会话的平均长度, 我们发现搭配查询相比 n -gram 查询对 Google 用户来说更为困难, 两者的会话平均长度分别为 3.1 和 1.9。这是由于 Google 采用基于关键词和 n -gram 的检索方式, 不支持按语义信息表达查询意图。例如, 当目

标短语为搭配“提出方法”(“present method”)时,用户在 Google 中不得不使用“present a method”、“present the method”等多次查询,否则难以消除“present”可以同时作为动词(“提出”)和形容词(“现在的”)的歧义.

此外,若以每条查询记录包含的单词数作为查询长度,我们发现用户由 2 个或 3 个单词构成的短查询占全部查询请求的 60.6%,这表明用户更加倾向于使用较为简短高效的查询表达方式.

以上发现表明,为了更好地满足用户写作时的查询需求,需要扩展现有查询语言的表达能力,使其能够支持按照词性、搭配等语义层面的条件检索短语,并且查询表达方式也需要精心的设计以降低用户的学习成本,方便普通 ESL 学者使用.

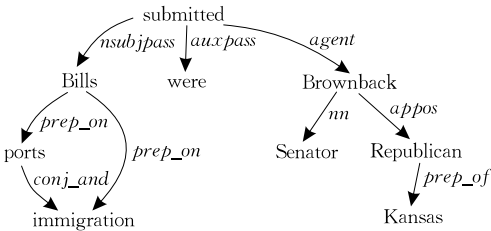
3.3 依存语法与语义搭配提取

本文利用依存语法(dependency grammar)来提取句子中的搭配.依存语法是一套基于单词间语义依存关系的语法规则^[23],依存关系可用三元组 $D=$

NSUBJPASS (submitted, Bills)
AUXPASS (submitted, were)
AGENT (submitted, Brownback)
NN (Brownback, Senator)
APPOS (Brownback, Republican)
PREP_OF (Republican, Kansas)
PREP_ON (Bills, ports)
CONJ_AND (Ports, immigration)
PREP_ON (Bills, immigration)

$S=$ Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas.

图 1 例句的依存关系列表及依存关系图



3.4 PDL 查询语言

为解决语料库短语检索中的机器层查询语言问题,本文提出模式描述语言(Pattern Description Language, PDL).它在满足 n -gram 查询的同时,还提供对语义查询的支持.

PDL 查询的构成如图 2 所示,包括查询单位、距离关系和搭配关系 3 个列表.查询单位对应于句子中的一个单词及其所有变形(复数、过去式、比较

Type(A, B)描述,表示句子中从属词(dependent) B 以 Type 类型的关系依赖于支配词(governor) A .我们借助 Stanford Parser^[24]来提取句子中的依存关系,图 1 给出了例句及其蕴含的所有依存关系.

并非所有的依存关系都表示具有明确语义关联的搭配,如图 1 中的助动词关系 AUXPASS (submitted, were)、连接词关系 CONJ_AND(ports, immigration)等均不能构成语义搭配.为此我们从 53 种依存关系中,选取语义关联较强的修饰、动宾等 7 种关系作为语义搭配.这 7 种关系涉及的单词均为名词、动词、形容词、副词等实词(区别于冠词、介词、连词、助动词等虚词),体现了与写作有关的语义信息.例如对于图 1 中的句子,可以提取出具有语义的搭配:“port(修饰)bill”、“immigration(修饰)bill”、“submit(动宾)bill”.值得注意的是,利用依存语法提取的搭配关系脱离了 n -gram 中单词相邻或者顺序的限制,更加符合用户写作的心理和思维习惯,也极大地方便了短语检索和后面的搭配推荐任务.

级等),可以是中英文单词、单词集合、概念扩展集合或表示任意单词的通配符.其中,通配符还可以通过指明词性使其只匹配特定词性的单词.距离关系描述两个查询单位在句子中相对的左右顺序、距离范围.在无距离关系约束的情况下,系统默认两单位在句子中的位置任意.搭配关系表示两个查询单位构成一个语义搭配,包括搭配的类型以及左右顺序的定义.

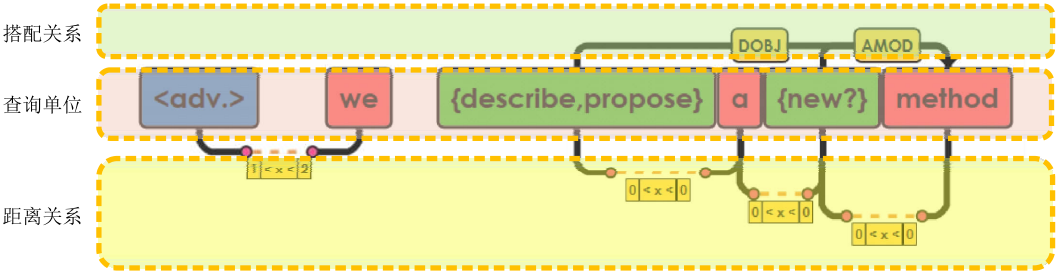


图 2 PDL 查询语言的图形示意

可以看出,PDL 查询语言扩展了现有查询语言的表达能力,不仅包含传统的关键词和 n -gram 检索,还支持按词性、搭配等语义条件的检索.

3.5 PDL 的用户查询表达

上文提出的机器层查询语言 PDL 扩展了现有搜索引擎和语料库检索引擎的表达能力,但语言学知识有限的 ESL 学者并不能迅速理解 PDL 中涉及的每个元素的含义,难以准确地表达包含出语义条件的检索请求.因此,查询系统还需要充分考虑用户的心理模型,设计直观易用的查询表达方式,方便用户使用表达查询意图.

在信息检索领域,典型的查询表达有文本、图形和样例等形式^[21,25-26].对于本节研究的任务——

写作短语检索,文本表达方式与写作任务本身相关性最大,符合用户的心理模型,具有很高的表达效率,是本文首选的表达方式.表 2 列出了所有文本表达方式的功能.但文本表达方式也存在一些缺点,比如:用户难以清楚地理解并表达 PDL 中词性和搭配类型等复杂的语义检索条件.因此,为了兼顾表达能力的丰富性和表达的直观性,本文在文本关键词的基础上,增加了基于样例的查询表达方式,如图 3 所示.在基于样例的表达方式中,用户只需输入单词或通配符“*”,系统自动提示所有可能的选项,并且直接将选项的效果以例句的形式展现,降低了用户理解语言学术语的认知负担.

表 2 PDL 用户查询的文本表达方式

文本符号	含义	查询表达示例	查询结果举例(用法;例句)
“ ”	匹配 n -gram	“new method”	to our knowledge; However, to our knowledge , they have never been available in ...
n. /v. /adj. /adv. /prep. /*	匹配带通配符的 n -gram	“method prep.”	method for; The account provides a ‘bottom up’ method for developing a more ... method of; Thus one method of establishing shared referents is to reformulate ... method in; ... capture people’s activity evolved as a method in photography, film ...
[]	单词集合	[novel,new] method	novel method; In this paper we propose a novel method of human-based electrical ... new method; ... since existing HTTP headers can be applied to any new method, ...
?	概念扩展	good? method	effective method; The most effective method for constructing high quality... expert method; ... heuristic evaluation is an expert method and thus it is easier ... good method; ... it is a good method with regard to computational and hardware ...
(修饰)/(主谓)/(动宾)/(介词)/...	搭配关系	present (修饰) method	present method; ... that the present measurement method is an effective technique ...
-	组合使用概念扩展与搭配关系	present?(动宾) method	present method; We present a method to facilitate parameter tweaking for visual ... introduce method; Then, we introduce our method for controlling the user’s view ... show method; ... we need to show a method for satisfyingthe SQL query ...



图 3 使用例句表达 PDL 查询中的词性约束和搭配关系

由上可见,文本表达方式支持直观的词性通配符查询和概念扩展,相比基于文本关键词的 Google 等检索工具具有更强的查询表达能力;基于样例的表达方式可以在输入关键词时列出相关词性、搭配及例句供用户选择,避免了用户直接以文本指明语义约束条件的困难.通过样例自动提示与文本符号相结合,本文查询表达方式的设计既提供了对语义查询的表达能力,又具有直观易用的特点.从后

文(第 5.4 节)实验的用户反馈中我们可以证明其易用性.

目前用户查询表达的设计并未完全覆盖 PDL 的所有功能,例如从界面上支持用户自定义单词间距离比较困难,在未来工作中用户层查询表达还会根据实际用户的反馈不断迭代完善.同时,这也体现出 PDL 查询语言设计的灵活性和表达潜力.

4 搭配推荐

4.1 问题定义

基于语料库的搭配推荐问题可以用下式描述:

$$t^* = \arg \max \{P(t|s, c)\} \quad (1)$$

其中

$$t \in \{\text{语料库中所有表达式}\}.$$

上式表示根据用户当前的写作内容 s 和上下文信息 c , 在语料库中寻找一个最合适的表达式 t^* 作为推荐结果. 公式中各符号的含义如下:

(1) s 为用户当前写作的内容, 即有待系统检查和推荐的文本.

(2) t 为搭配推荐系统的单元表达式, 通常为单词或短语.

(3) c 为用户当前写作的上下文信息, 即待推荐文本 s 前后已经确定的写作内容, 可以是相邻单词、语法结构、语义搭配等, 不同种类上下文信息的选取反映了不同的语言模型.

根据概率公式, 式(1)可以进一步表示为

$$t^* = \arg \max \{P(t|c)P(s|t, c)\} \quad (2)$$

其中, $P(t|c)$ 称为语言模型, 表示在给定上下文 c 的情况下可能出现与 c 搭配的表达式 t 的分布情况, 可从语料库数据中统计得到; $P(s|t, c)$ 称为 ESL 学者写作的用户模型, 表示学者在希望表达某个确定的表达式 t 时, 由于语言能力的限制误将其表达为 s 的分布情况. 用户模型目前是一个未知的分布, 可以利用语义网或者大规模实际写作数据来建模.

对语言模型的统计和对 ESL 用户模型的估计将极大地影响推荐系统的性能, 本文将针对 s 和 t 均为单词的情况讨论写作搭配推荐的方法.

4.2 语言模型

现有推荐方法中, 上下文信息 c 常为与用户输入 S 相邻的一个或多个单词, 即以 n -gram 作为语言模型^[1]. 而上文已经指出, n -gram 对语义信息利用不充分, 不适合作为搭配推荐的语言模型.

本文采用由语义依存关系提取的语义搭配作为语言模型, 即上下文信息 c 关心的不是语料库中与用户输入 s 相邻的前后单词, 而是前后所有单词中能够与 s 构成语义搭配的单词. 例如与上下文单词“method”存在修饰关系的形容词有“good”、“novel”等, 它们并不一定与 method 相邻, 但均与上下文构成语义搭配关系. 利用语义搭配作为语言模型对上下文的语义信息利用更加直接和充分, 从而能够提

高推荐系统的准确率.

4.3 ESL 用户模型

本文考虑到当用户无法准确表达某个语义时, 会尽可能选取他熟知的语义接近的表达方式, 因此用户输入 s 与实际最合适的表达式 t^* 存在紧密的语义关联. 例如, 当用户需要表达“决定…的质量”这一搭配时, 由于受词汇量限制, 只能联想到“influence…quality”, 但实际更合适的表达应为“determine…quality”. 这里虽然“influence”和“determine”并非同义词关系, 但两词之间存在内在的概念关联. 我们利用 WordNet^[27] 来查找与 s 语义相近的候选单词, WordNet 是由义项集合及其之间的语义、词法关系构成的语义网, 每个义项集合包含若干表示同一语义概念的单词.

本文认为最佳表达式 t^* 只可能落在由与用户输入 s 语义相近的单词构成的集合 $E(s)$ 中, 并称该集合为 s 的概念扩展集. s 的概念扩展集 $E(s)$ 定义为: s 所属的所有义项集合 $A_i \in \text{Syn}(s)$ 中的单词 (即严格意义上 s 的同义词) 以及与这些义项集合存在直接语义概念关联的义项集合 $W_{ij} \in \text{Rel}(A_i)$ 中的单词. 图 4 演示了利用 WordNet 进行概念扩展的过程. 可以看出, 相比仅使用同义词扩展, 概念扩展可以获得更多与用户实际意图相关的候选结果, 且引入了单词之间语义相关度的度量.

此外, 最佳结果 t^* 为 t 的概率应当与 t 到 s 的语义相似度 $\text{Sim}(t, s)$ 成正比. 综上, 我们得到

$$P(s|t, c) \propto \begin{cases} \text{Sim}(t, s), & t \in E(s) \\ 0, & \text{其他} \end{cases} \quad (3)$$

其中

$$\text{Sim}(t, s) = \frac{|\text{Syn}(t) \cap \text{Syn}(s)| + 1}{|\text{Syn}(t) \cup \text{Syn}(s)| + 1},$$

$$E(s) = \bigcup_{\substack{W_{ij} \in \text{Rel}(A_i) \cup \{A_i\} \\ A_i \in \text{Syn}(s)}} W_{ij}.$$

式(3)为基于概念扩展的 ESL 用户模型, 与 t 属于全语料库范围的等概率模型相比, 它充分考虑了用户输入 s 的语义信息. 通过后面的用户实验我们可以证明这种搭配推荐方法具有较高的准确率.

目前本文利用 WordNet 中单词的语义关联来刻画 ESL 写作的用户模型, 是因为它是目前能够得到的较为全面的描述单词间两两相关性的数据 (WordNet 3.0 共包含 117000 个不同的义项集合). 然而, 由于语言内容的丰富性、写作风格的多样性以及作者的个体差异, 仅用 WordNet 统一描述的用户模型存在局限性. 在将来, 通过部署短语检索系统

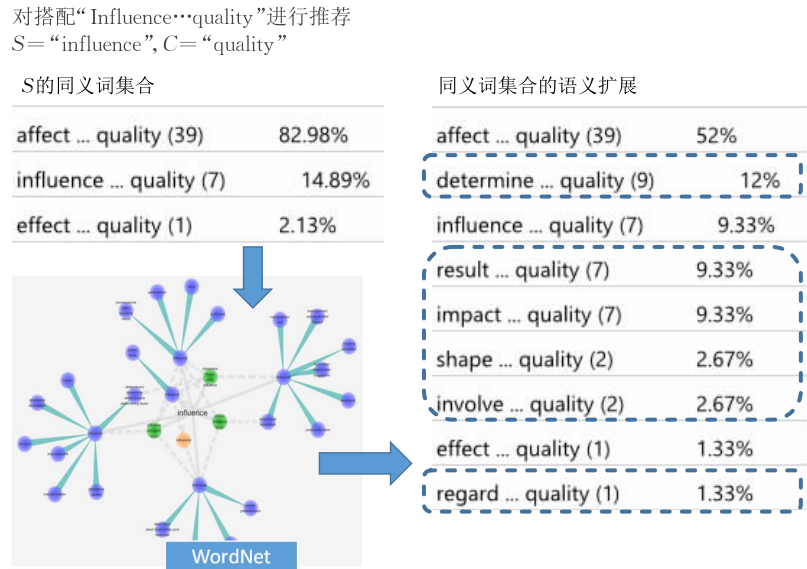


图 4 利用 WordNet 进行概念扩展的示例

(第 3 节)大规模收集和分析 ESL 学者写作时的查询会话,可以为每位 ESL 学者建立更加精准的用户模型,从而进一步提高搭配推荐的精度。

5 ESLWriter 系统实现与评估

5.1 系统实现

在上述研究的基础上,我们选择辅助 ESL 研究者写作学术文章作为一个典型的应用场景,以计算机领域的学术文献作为语料库,实现了辅助 ESL 写作学术文章的 ESLWriter 系统。ESLWriter 为基于

Django 框架的网页端——服务端系统,其语料库数据采用 MongoDB 引擎进行存储和索引,系统架构如图 5 所示。作为实际运行的系统,ESLWriter 长期对外开放,任何用户都可通过浏览器访问^①。

ESLWriter 客户端界面的设计如图 6 所示。在查询页面中,用户在文本框内输入查询表达式,自动弹出的提示框可以以例句的方式提示用户可能的查询表达式。查询结果根据不同用法按使用频率由高到低分组呈现,每组结果还可展开查看详细的例句,用户可以根据自身信息需求查看相应信息。在写作推荐页面中,用户可以在左侧文本框内输入一段文

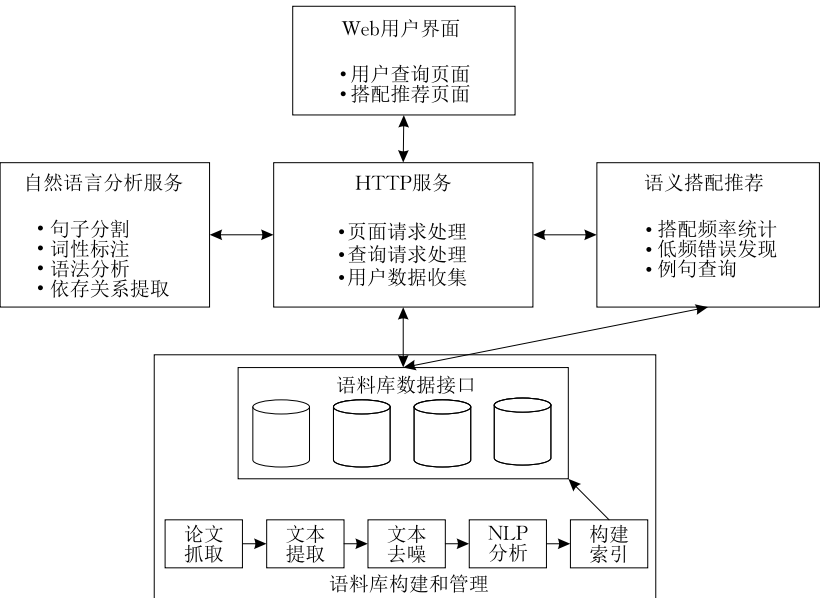


图 5 ESLWriter 系统架构图

① ESLWriter 网址: <http://www.thuesl.org>

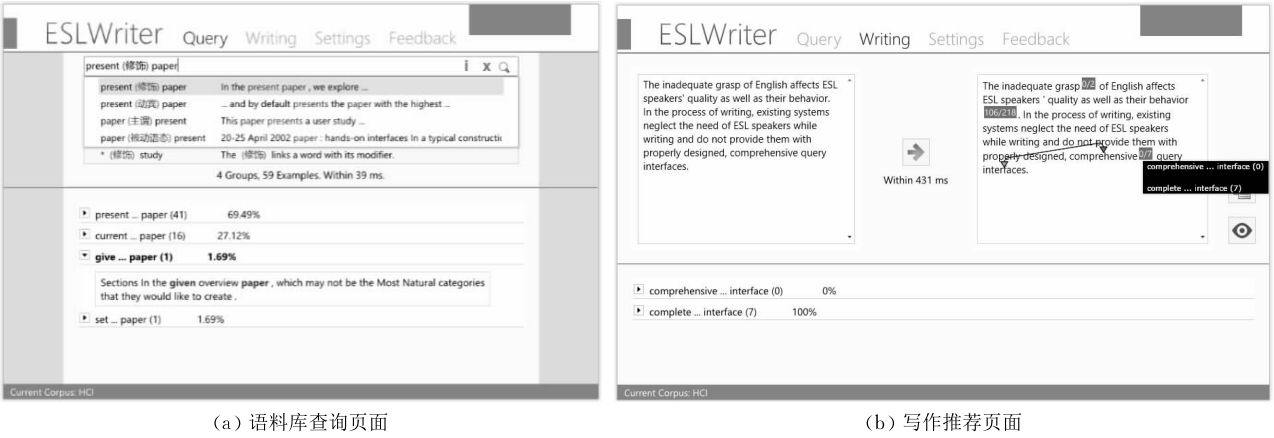


图 6 ESLWriter 的用户界面: 语料库查询页面和写作推荐页面

本, 右侧将在段落相关位置以标签的形式对潜在的
错误搭配给出建议. 此外, 两个页面均提供可分组浏览的
例句列表, 每个例句都可查看其对应的论文信息, 而且例句中的关键字高亮显示, 便于用户在具体
上下文中查看或比较用法.

5.2 领域专用语料库的作用

ESLWriter 以辅助计算机领域的 ESL 研究者
写作学术文章为实际应用场景, 其语料库由该领域
大量已经发表的高质量论文构成, 并且细分为计算
机系统、计算机网络等 9 个不同的子研究领域. 在

实际使用时, 用户可以从若干子领域的语料库中
选择一个或多个同时查询. ESLWriter 语料库中的
论文覆盖了《中国计算机学会推荐国际学术会议和
期刊目录》^①中所列举的计算机学科 9 个研究领域
A 类和 B 类的学术期刊和会议, 论文全文数据来自
Google Scholar 的搜索结果以及各领域允许公开访问的
网上论文集. ESLWriter 语料库的详细信息如
表 3 所示. 领域专用语料库具有其他语料库不可替
代的权威性, 系统记录显示, 用户几乎都会把自己研
究领域的语料库作为首选.

表 3 ESLWriter 语料库信息统计 (计算机领域)

子领域名称	代表会议、期刊	会议、期刊总数	论文总数	句子总数
计算机系统与高性能计算	TOCS, TOC, HPCA	39	31 374	7 019 638
计算机网络	TON, SIGCOMM, INFOCOM	14	11 089	2 754 540
网络与信息安全	TDSC, CCS, S&P	21	11 065	3 418 008
软件工程/系统软件/程序设计语言	TOPLAS, TSE, OSDI	35	18 176	4 975 368
数据库/数据挖掘/内容检索	TODS, SIGMOD, SIGIR	25	23 796	5 516 648
计算机科学理论	SICOMP, STOC	18	19 259	4 320 745
计算机图形学与多媒体	TOG, TVCG, ACM MM	20	13 883	2 911 407
人工智能与模式识别	JMLR, AAAI, ICML	29	48 117	8 374 385
人机交互与普适计算	TOCHI, CHI, UbiComp	12	8 469	1 929 171
总计	—	213	185 228	41 219 910

为进一步分析不同子领域中学术论文用语的差别, 表 4 分别列出了 4 个代表性子领域的论文中出现频率最高的单词和动宾搭配. 从表 4 中可以看出, 即使同属于计算机这一个大领域, 不同子领域论文在用词上也存在明显差异, 主要体现在跟该领域研究问题相关的专业名词以及与这些名词相搭配的动词. 例如计算机科学理论领域的论文以定理证明和公式推导居多, 而人机交互与普适计算领域的论文则包含许多用户研究和实验设计.

领域语料库的用词差异以及用户使用自己研究领域语料库的偏好表明了领域专用语料库对辅助 ESL 学术写作的重要作用. 本文所提出的短语检索

和搭配推荐方法不限定语料库的具体领域, 因此这些方法可以迅速扩展到其他领域的学术写作以及非学术写作的场景. ESLWriter 的语料库管理系统可以根据 PDF 格式的论文集自动构建语料库索引, 方便了我们在将来为系统添加其他领域的语料库. 在未来, 系统还将允许作者上传自己指定的若干论文 (通常作者会选择当前正在写作的论文的参考文献) 以构建个人语料库, 在个人语料库中查询的结果和推荐的搭配将优先展示给作者, 这样使得系统的语言风格更加契合作者当前写作的上下文.

① CCF 推荐排名: <http://www.ccf.org.cn/sites/ccf/paiming.jsp>

表 4 不同领域语料库高频词汇和动宾搭配比较(已滤除停用词)

数据库/数据挖掘/内容检索		计算机科学理论		人工智能与模式识别		人机交互与普适计算	
单词	搭配	单词	搭配	单词	搭配	单词	搭配
datum	solve/problem	algorithm	solve/problem	model	solve/problem	user	perform/task
query	show/result	set	give/algorithm	algorithm	show/result	system	provide/information
set	improve/performance	time	denote/set	image	improve/performance	design	play/role
algorithm	answer/query	problem	present/algorithm	set	present/result	information	solve/problem
result	propose/algorithm	follow	satisfy/condition	method	propose/method	work	complete/task
model	reduce/number	give	prove/result	result	address/problem	participant	answer/question
time		show		show		time	
number		graph		problem		task	
show		result		datum		study	
user		number		feature		datum	

5.3 用户实验

本实验的目的是评估 ESLWriter 对于辅助 ESL 实际写作任务的有效性,需要考察:(1) ESLWriter 查询界面的设计是否直观易用;(2) 基于语义的搭配推荐功能是否能为用户提供有用的搭配建议;(3) 在利用了 ESLWriter 系统的检索和推荐功能后,用户的写作质量是否有提高;(4) 用户在实验和实际中使用系统的主观感受。

特别地,由于本文基于语义的写作辅助方法是对现有基于 n -gram 方法的补充,因此本用户实验旨在验证概念、搭配类型等语义信息对辅助 ESL 写作活动的有效性.同时,现有方法侧重解决拼写、语法等 L1 问题,而本文侧重解决以搭配为主的 L2 问题.故本文现阶段无法与已有系统进行对比实验.

5 名英文写作能力不同的 ESL 学者参加了本实验.他们均为计算机专业的研究生,使用 ESLWriter 在实际中写作或修改自己即将发表的论文(待修改的论文是已经由作者通过其他方式修改,认为没有明显问题的版本),并在 Word 文档中标记出修改信息,包括修改前和修改后的版本.实验结束后,修改结果发送给一名以英语为母语的学者来评价.该评价者根据修改质量对修改进行评分,包括“更好”、“更差”或“均可”(即修改前后的两种表达方式均可接受)3 种情况.

5.4 实验结果

实验结果如表 5 所示.5 名学者在 ESLWriter 系统的帮助下共做出 103 处修改.经评价人打分,78 处(75.7%)为更好,10 处(9.7%)为更差,另外 15 处(14.6%)为均可.以上实验结果表明,ESL 学者在系统帮助下产生的 90%以上的修改结果均可以接受,ESLWriter 对提升写作质量有帮助.值得注意的是,即使对于已经经过作者本人精心修改过的

文章,在 ESLWriter 系统的辅助下,写作质量仍可进一步提高.

表 5 ESLWriter 用户实验结果

用户编号	年龄	发表文章数	实验任务	修改数目	修改质量 (更好/均可/更差)
1	21	0	写作和修改	21	16 / 2 / 3
2	21	1	修改	19	12 / 2 / 5
3	23	2	修改	11	7 / 3 / 1
4	25	5	写作和修改	36	31 / 5 / 1
5	25	10	写作和修改	16	12 / 4 / 0

上述结果也表明,少数情况下系统也可能会提供质量更差的推荐.通过进一步分析这些“更差”推荐,我们发现:缺乏写作经验的用户 1 和 2 共接受了 8 处错误的推荐,而写作经验相对丰富的用户 3、4 和 5 一共接受了 2 处错误推荐(如图 7 所示).用户采访结果表明,用户 1 和 2 由于英语能力不足,难以评价推荐结果的质量,更多地依赖于 ESLWriter 给出的频率分布来判断接受或拒绝.然而,英语经验相对丰富的用户 3、4 和 5 则会凭借自身的英语知识来做出判断,不会盲目采纳系统的推荐.由此我们推断,ESL 学者自身语言能力和写作经验的不同会影响使用写作辅助系统的效果,在未来写作辅助系统需要能够针对不同英语能力的用户自动调整用户模型,从而实现更加个性化的写作推荐.

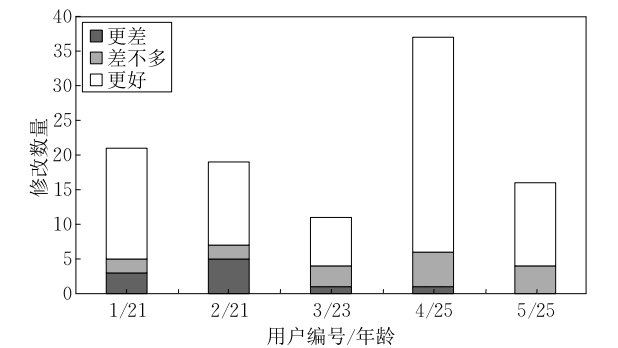


图 7 不同写作经验用户修改质量分布

我们还询问了 5 名实验者使用系统的主观感受. 对于 ESLWriter 提供的短语检索功能, 用户通常只需要 5 分钟的试用即可熟悉系统的使用方式. 用户表示, 基于文本和样例的用户层查询表达式是直观和易于理解的, “很多以前没法查的语言问题都可以在 ESLWriter 里查了”. 对于搭配推荐功能, 用户表示在经过自动推荐后, 他们对自己的写作质量更具信心; 用户 1 甚至表示“如果不把自己文章的所有段落经过 (ESLWriter) 分析一下就感觉不放心”. 以上用户反馈表明, ESLWriter 的查询表达方式易用性好, 推荐结果具有较高的可信度, 而且系统确实能够帮助 ESL 学者提高写作信心.

5.5 分析与讨论

值得说明的是, 以上实验结果是 ESLWriter 查询与推荐功能共同作用的结果, 不应当分开对待. 在实验过程中, 我们观察到用户先使用 ESLWriter 的自动搭配推荐功能发现潜在的搭配错误, 再通过主动查询的结果做出最终修改决定的使用模式. 例如, 用户 2 在修改句子“... improve the gesture recognition capacity of the system and set the alignment of widgets adaptively ...”时, 自动推荐功能给出如下建议: (1) “recognition capacity” → “recognition capability”; (2) “set ... alignment” → “adjust ... alignment”. 用户 2 较为容易地判断并接受了建议 (2), 但由于建议 (1) 在语料库中的相关例句较少, 他难以决定是否接受. 因此用户 2 利用语义检索功能分别检索“capacity”和“capability”可能的名词修饰词, 发现: “capability”主要与“input”、“output”搭配, 意为“能力”, 而“capacity”主要与“storage”、“memory”搭配, 意为“容量”. 因此他最终接受了系统的推荐“recognition capability”. 现有系统均忽略了自动推荐与主动查询两方面功能可以互相补充, 共同为用户写作提供有用的帮助, 这是未来系统设计时需要注意的地方.

此外, 用户实际偏好也进一步证明了本文所提方法的有效性. 自 ESLWriter 开放使用以来, 系统访问量迅速增长. 至本文截稿时, 系统已处理国内外 115 名用户的 PDL 查询请求共计 1.1 万余次, 其中包括 31 名查询请求在 50 次以上的活跃用户; 帮助用户修改段落共计 95 次, 涉及文本总长 1.8 万余单词. 特别地, 由于系统对语义信息的支持, 用户均表示愿意将 ESLWriter 作为未来英文写作的首选辅助工具.

以上实验结果以及 ESLWriter 系统的实际使用情况也为 ESL 写作辅助系统的设计和评估带来了新的启示. 首先, 尽管系统偶尔会产生明显错误的推荐, 但用户自身的语言判断能力以及用户对系统工作原理的认识能够使其避免被误导. 第二, 系统给出的任何建议都应当同时给出例句作为依据, 因为例句能够为用户发现和改正错误额外提供有用的参考信息. 第三, ESL 写作辅助系统的设计并不应以代替用户写作为目标, 而应着眼于提供用户从语料库中获取写作知识的便捷途径, 从辅助 ESL 的实际写作过程出发来达到提高用户写作质量和写作信心的目的. 最后, 评价 ESL 写作辅助系统时除考察系统自身结果的有效性以外, 还应当考虑系统对 ESL 用户写作活动的辅助作用, 即用户使用系统以后写作能力的提升.

6 总结和展望

本文工作旨在研究如何利用语义信息解决 ESL 的写作困难. 一方面, 本文从 ESL 的实际查询需求出发, 设计了同时支持 n -gram 和语义检索的 PDL 查询语言; 另一方面, 本文通过考察搭配推荐时用户输入与实际意图的语义关联, 提出在语义搭配层面上通过扩展用户输入的语义进行搭配推荐的方法. 此外, 为了方便语言知识有限的 ESL 学者表达检索请求, 本文还设计了直观易用的短语检索界面. 最后, 通过实现和评估实际运行的 ESLWriter 系统, 我们验证了基于语义的短语检索和搭配推荐方法的有效性, 达到了辅助 ESL 学者写作的目的.

ESLWriter 作为已经部署和推广的系统, 随着用户的逐渐积累, 不断增长的用户数据为进一步研究 ESL 写作问题提供了广阔的前景. 未来工作可以围绕用户使用 ESLWriter 查询和修改的历史记录展开, 包括进一步研究 ESL 写作的需求和行为模式, 设计更合理易用的查询表达方式, 提出更为精确的 ESL 用户模型等等.

参 考 文 献

- [1] Park T, Lank E, Poupart P, et al. Is the sky pure today?

- AwkChecker: An assistive tool for detecting and correcting collocation errors//Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology. Monterey, USA, 2008; 121-130
- [2] Lin D. Extracting collocations from text corpora//Proceedings of the 1st Workshop on Computational Terminology. Montreal, Canada, 1998; 57-63
- [3] Gui Shi-Chun, Yang Hui-Zhong. Chinese learner English corpus. Shanghai: Shanghai Foreign Language Education Press, 2003(in Chinese)
(桂诗春, 杨惠中. 中国学习者英语语料库. 上海: 上海外语教育出版社, 2003)
- [4] Nesselhauf N. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 2003, 24(2): 223-242
- [5] Yi X, Gao J, Dolan W B. A web-based English proofing system for English as a second language users//Proceedings of the International Joint Conference on Natural Language Processing. Hyderabad, India, 2008; 619-624
- [6] Matsumoto K. Research paper writing strategies of professional Japanese EFLwriters. *TESL Canada Journal*, 1995, 13(1): 17-27
- [7] Wible D, Kuo C H, Tsao N L, et al. Bootstrapping in a language learning environment. *Journal of Computer Assisted Learning*, 2003, 19(1): 90-102
- [8] Gamon M, Gao J, Brockett C, et al. Using contextual speller techniques and language modeling for ESL error correction//Proceedings of the International Joint Conference on Natural Language Processing. Singapore, 2009, 8; 449-456
- [9] Schneider D, McCoy K F. Recognizing syntactic errors in the writing of second language learners//Proceedings of the International Conference on Computational Linguistics. Quebec, Canada, 1998, 8(2): 1198-1204
- [10] Dalgish G. Computer-assisted ESL research and courseware development. *Computers and Composition*, 1985, 2(4): 45-62
- [11] Bitchener J, Young S, Cameron D. The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 2005, 14(3): 191-205
- [12] Izumi E, Uchimoto K, Isahara H. The NICT JLE corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management*, 2004, 12(2): 119-125
- [13] Shei C C, Pain H. An ESL writer's collocational aid. *Computer Assisted Language Learning*, 2000, 13(2): 167-182
- [14] Brockett C, Dolan W B, Gamon M. Correcting ESL errors using phrasal SMT techniques//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual meeting of the Association for Computational Linguistics. Sydney, Australia, 2006; 249-256
- [15] Gamon M, Leacock C, Brockett C, et al. Using statistical techniques and web search to correct ESL errors. *Calico Journal*, 2009, 26(3): 491-511
- [16] Guo S, Zhang G. Building a customized Google-based collocation collector to enhance language learning. *British Journal of Educational Technology*, 2007, 38(4): 747-750
- [17] Sha G. Using Google as a super corpus to drive written language learning: A comparison with the British National Corpus. *Computer Assisted Language Learning*, 2010, 23(5): 377-393
- [18] Chen M H, Huang S T, Hsieh H T, et al. FLOW: A first-language-oriented writing assistant system//Proceedings of the ACL 2012 System Demonstrations. Jeju Island, Korea, 2012; 157-162
- [19] Dahlmeier D, Ng H T. Correcting semantic collocation errors with L1-induced paraphrases//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011; 107-117
- [20] Stein B, Potthast M, Trenkmann M. Retrieving customary Web language to assist writers//Proceedings of the Advances in Information Retrieval. Milton Keynes, UK, 2010; 631-635
- [21] Resnik P, Elkins A. The linguist's search engine: An overview //Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions. Michigan, USA, 2005; 33-36
- [22] Liu T, Zhou M, Gao J, et al. PENS: A machine-aided English writing system for Chinese users//Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Hong Kong, China, 2000; 529-536
- [23] Mel'cuk I. Levels of dependency in linguistic description: Concepts and problems. *Dependency and Valency. An International Handbook of Contemporary Research*, 2003, 1: 188-229
- [24] De Marneffe M C, MacCartney B, Manning C D. Generating typed dependency parses from phrase structure parses//Proceedings of the International Conference on Language Resources and Evaluation. Genoa, Italy, 2006, 6: 449-454
- [25] Glowacka D, Ruotsalo T, Konuyshkova K, et al. Directing exploratory search: Reinforcement learning from user interactions with keywords//Proceedings of the 2013 International Conference on Intelligent User Interfaces. Santa Monica, USA, 2013; 117-128
- [26] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos//Proceedings of the 9th IEEE International Conference on Computer Vision. Beijing, China, 2003; 1470-1477
- [27] Fellbaum C. WordNet. In *Theory and Applications of Ontology: Computer Applications*. Springer Netherlands, Dordrecht, Netherlands, 2010; 231-243



ZHU Ye-Shuang, born in 1990, Ph.D. candidate. His research interests focus on human computer interaction.

YU Chun, born in 1984, Ph. D. , assistant professor. His main research interests include Graphical User Interface (GUI), interactive tabletop.

SHI Yuan-Chun, born in 1967, Ph. D. , professor, Ph. D. supervisor. Her main research interests include human computer interaction, multimedia and pervasive computing.

Background

Academic writing in English has long been a challenge for English as Second Language (ESL) researchers. Typical problems in ESL writing indicate a clear need for useful tools to support effective and fluent English writing. Due to the intrinsic complexity of language, assistive technology for writing involves many previous works in Linguistics and Natural Language Processing. There are many *n*-gram based

works related to professional corpus search engine for linguists and automatic error correction algorithms for writing. However, the use of semantic information and design of a feasible query interface to support users’ writing activity has been overlooked, which is essentially a problem in Human Computer Interaction. Thus, it is a research direction worth a try.