文章编号: 1003-0077(2010)04-0039-05

一种新的面向领域的鲁棒性文本分析算法

陶县俊[1], 邬晓钧[2], 王晓东[1], 郑 方[2]

(1. 河南师范大学 计算机与信息技术学院, 河南 新乡 453007;
2. 清华信息科学技术国家实验室 技术创新与开发部语音和语言技术中心, 北京 100084)

摘 要: 在自然语言处理的应用中,特别是在对口语文本、网络文本的处理中,待分析的文本经常会包含字词和句式上的错误。该文描述了一种基于线图分析方法改进的鲁棒性文本分析算法。该算法利用当前活动弧和规则库中的终结符,对基于领域词表的分词过程无法识别的语句串进行错误推测,将无法识别的语句串纠正为可能的正确文字。实验结果表明,在采用拼音的同音匹配进行推测纠错的情况下,该文所设计的鲁棒性文本分析算法相对于燕方法,分析度提高了 14.78%,而语句平均分析循环次数增长为 9.363%。

关键词: 计算机应用; 中文信息处理; 线图分析方法; 鲁棒性; 错误推测
中图分类号: TP391 文献标识码: A

A New Robust and Domain-Oriented Algorithm of Text Parsing

TAO Xianjun[1], WU Xiaojun[2], WANG Xiaodong[1], ZHENG Fang[2]

(1. College of Computer and Information Technology, Henan Normal University, Xinxiang, Henan 453007, China;
2. Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China)

Abstract: In applications of natural language processing, especially in processing of spoken or web text, errors in word spelling and/or sentence structures are common to be found in the text to be processed. This paper describes a robust parsing algorithm based on the chart parsing method, which can identify the mistakes in the strings unrecognized by the domain vocabulary based word segmentation, and fix them into the correct forms according to the terminal information extracted from the current active arcs and the rule set. The experimental results showed that with error detection and correction by homonymous matching of pinyin syllables, this algorithm improvs the acception rate by 14.78% at the cost of an increase in the average number of loops by 9.363% compared with the robust parsing method of Yan.

Key words: computer application; Chinese information processing; chart parsing methods; robustness; error detection

随着计算机和 Internet 的推广和应用,对自然语言的处理由数据处理、信息处理发展到知识处理,同时对于分析算法的性能和效率的要求也越来越高。分析器在自然语言文本信息处理领域占有十分重要的地位。因为应用环境的影响和自然语言表达的随意性、缩略性,特别是在对口语文本、网络文本的处理中,待分析的文本经常会包含字词和句式上的错误,自然语言分析器的鲁棒性显得尤为重要[1]。一个鲁棒性较高的分析器意味着能够在自然语言处理的过程中对出现的错误进行更为有效地处理[2]。

目前国内外开展了很多针对于鲁棒性文本分析器的研究,不少鲁棒性文本分析方法已在实践中取得了良好的效果。这些方法在鲁棒性激发机制上有较大的共同点: 首先用分析器分析语句,当遇到错误导致分析中断时,调用相应的出错处理机制,使得分析能够正常进行下去。出错处理机制的优劣直接

收稿日期: 2009-07-03 定稿日期: 2009-11-12
基金项目: 河南省重点科技攻关项目资助(082102210007)

作者简介: 陶县俊(1982—),男,硕士,主要研究方向为自然语言处理; 邬晓钧(1976—),男,博士,助研,主要研究方向为自然语言处理; 王晓东(1963—),男,教授,博士,主要研究方向为语义与本体、知识工程。

影响了整个分析方法的鲁棒性。出错处理机制一般从两个方面解决问题: 第一是针对错误类型采用高效的算法进行纠错, 例如部分分析方法、基于编辑距离的最小纠错方法以及短语检出方法等等; 第二是从文法规则方面, 手工构建容错文法。

部分分析方法是现存的使用较为普遍的鲁棒性文本分析方法, 它在解决自发语音中的口语现象和识别错误等方面有较大优势[3]。因为英文语法结构的特点, 在英文的处理中有较为普遍的应用。燕鹏举提出了一种新的部分分析方法[4] (以下简称燕方法), 与一般的部分分析方法不同的是, 该方法可以跳跃待分析文本中的一些文字进行成分的归约, 具有较强的灵活性和鲁棒性。但是燕方法对所分析的文法有较强的依赖性。在部分分析方法的基础上, Boros et al 提出了短语检出(Phrases Spotting)的概念[5]。短语检出的分析方法能够解决对话系统中口语表达的随意性、缩略性的问题, 但在分析的过程中考虑上下文语义的因素较少。针对短语检出分析方法的不足, Ye-Yi Wang 在经典线图分析方法的基础上提出了一种针对分析过程中所产生的成分进行打分的方法[5], 然后根据打分对这些成分剪枝。这种方法不但较好地解决了分析歧义的问题, 而且能够通过剪枝提高分析器的分析效率。在容错文法规则的构建上, Jennifer Foster 和 Carl Vogel 从各种类型的文档里面抽取错误语料[6], 并构建了错误语料库和并行的正确语料库, 然后手工总结语料库里面的错误特征, 针对相应的错误特征手工构建容错文法。在分析器中错误文法的使用虽然提高了分析器的鲁棒性, 但是在实际的应用中, 文法规则的产生以及手工构建容错文法是一个很复杂的过程[7]。

与前述各种鲁棒性文本分析方法不同, 本文在燕方法的基础上提出了一种新的基于线图分析方法的鲁棒性文本分析算法。该算法利用当前活动弧和规则库中的终结符信息对待分析文本中未识别语句串(在基于领域词表的分词过程中无法识别出来语句部分)进行错误推测与纠错处理, 以提高分析器的性能与效率。

本文后面的章节依次介绍论文的相关工作、算法原理以及实验情况, 最后对算法相关问题进行总结和讨论, 提出今后研究的方向。

1 相关工作

线图(Chart)分析算法是一种简单常用的句法分析算法, 是一个由议程表(Agenda)驱动的不断循环的过程(具体算法可参考文献[9])。算法按照初始化策略对议程表进行初始化处理, 如果议程表为空, 那么分析失败, 否则每次按照议程表组织策略, 从议程表中取出一个成分。如果取出的成分覆盖整个句子, 那么返回成功, 否则将取出的成分加入到线图中, 执行规则调用策略和活动弧递进、归约策略将产生的新成分又加入到议程表中。在这个算法流程中, 各项策略均可调整, 通过调整这些策略可以得到改进的线图分析算法。

燕方法是基于线图分析方法的一种改进算法, 本论文所做的相关工作是在燕方法的基础上进行的。燕方法所涉及的文法包含五种不同类型的规则, 分别是苛刻型(up-tying, 即传统规则)、跳跃型(by-passing)、长程型(long-spanning)、无序型(up-messing)以及交叉型(over-crossing)。其中跳跃型规则允许在规则右部各符号之间插入少量的其他符号, 处理口语中的停顿和无意义词现象; 无序型规则在跳跃型规则的基础上还允许规则右部符号组合的出现顺序任意, 解决汉语口语语序随意的问题。燕方法有以下特点:

1) 部分分析的特点: 对句子不作接受或拒绝的简单判断, 而是保留分析过程中得到的所有部分结果, 提供最大信息以便后续处理。

2) 跨成分归约的特点: 不拘泥于传统算法中成分间位置关系的紧密相连性与严格偏序性, 而是根据不同的规则类型采取更为灵活的活动弧递进、归约策略, 容许更为自由的口语通过分析器。

2 算法原理

虽然线图分析方法通过调整规则调用策略和活动弧递进、归约策略能使句子的语法分析具有一定鲁棒性, 但如果不能分析出完整的句子语法树的原因是分词过程无法识别某些语句串, 就不能较好地处理。例如下面的实例:

输入句子: 有房子在上地出租吗?

分词结果: 有、出租—VP; 上地、房子—NP; 在—Prep; 吗—Aux

规则库: $S \rightarrow VP\ NP$; $NP \rightarrow R\ Prep\ NP \mid VP\ Aux \mid R$; $VP \rightarrow NP\ VP$

进行语法分析后, 得到的语法树如图1。

当输入语句中"上地"由于输入时发生同音拼写错误, 整句变为: 有房子在上帝出租吗。因为领域词表中不存在"上帝"这个词汇, 在基于领域词表进

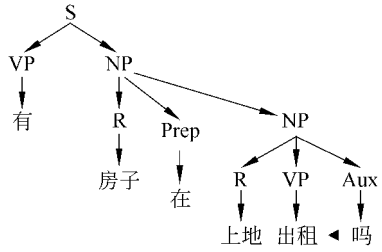

图 1　正确的语法分析树

行分词时无法识别出这个词汇。在运用线图分析方法时, 因为输入语句句法成分地缺失, 无法分析该语句, 分析过程中断。

尽管燕方法能较好地解决口语分析中常见的问题, 但它仍然不能解决由于输入错误改变了输入语句的情况。在上例中如果用燕方法进行分析, “上帝”会被当作垃圾串进行处理, 得到两个分析不完整的语法树, 如图 2 所示。

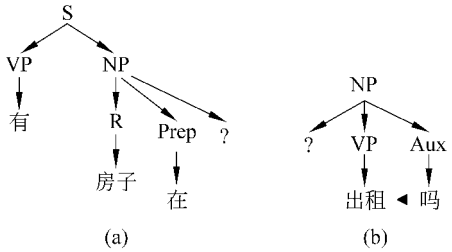

图 2　缺少分析成分的语法树

在上面的例子中线图分析方法和燕方法均不能解决输入错误问题, 即使输入语句只是犯了一个很小的同音拼写错误(这在网络应用背景下是常见的错误) 。本论文的研究工作基于针对错误类型采用高效的算法进行纠错的思想, 利用文法规则和当前活动弧中的终结符信息对分词过程中的未识别语句串进行错误推测, 从而将未识别语句串修正为可能的正确输入, 并相应修改分词结果和分析器的状态, 以使分析能够正常进行。算法的基本原理如图 3 所示。

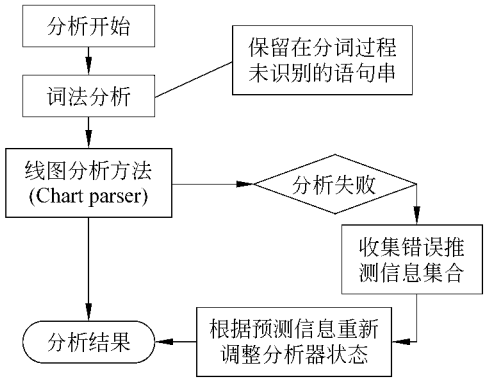

图 3　算法基本原理流程图

2.1　相关术语定义

为了在下面的核心算法流程中描述方便, 对相关术语做以下定义:

定义 **1**　待分析语句: 一个输入语句串经分词之后可看作由词类(未被识别的语句串属于特殊的词类) 组成的一个串, 写成句子 $sent=(K_0, K_1, \cdots, K_{n-1})$, 其中 n 表示句子中的词的个数, K_i $(0 \leqslant i < n)$ 是第 i 个词(包括未被识别的语句串) 的词类。

定义 **2**　未识别语句串: 在句子中会出现单个或者多个连续的无法被分词过程识别的输入文字。在分词的过程中暂时把连续的未识别文字(即未识别语句串) 当成一个词, 并且将这个词归为一个暂时词类 R。

定义 **3**　分析状态: 分析状态指当前分析过程中议程表中的成分和线图中当前活动弧所对应的分析位置以及所包含的具体内容, 可以写为 $T_i=(A_i, C_i)$ 其中 T_i 表示分析器在第 i 个时间点的状态, A_i 表示分析器的议程表在第 i 个时间点的状态, C_i 表示分析器的线图在第 i 个时间点的状态。当前的分析状态决定了下一个分析状态的走向, 若干个 T_i 就组成了整个句子的分析流程。

定义 **4**　规则库推测终结符集: 依据线图分析方法的原理, 分析器从当前状态 T_i 过渡到下一个状态 T_{i+1} 需要从规则库中寻找合适的产生新的活动弧的规则。当分析器面临一个暂时词类 R 时, 规则库中右项第一个为终结符的规则所包含的第一个终结符对这个未识别语句串有推测作用。把规则库中所有这样的终结符集合在一起即为规则库推测终结符集合, 表示为 Rule_T。

定义 **5**　当前活动弧推测终结符集: 当分析器处于某一个分析状态 T_i 时, 所对应的线图中包含一定数量的活动弧, 其中一些活动弧在当前活动位置之后待归约的成分为终结符, 这个终结符对 R 所代表的当前未识别语句串有推测作用。把当前活动弧中所有这样的终结符集合在一起即为当前活动弧推测终结符集合, 表示为 Activate_T。

2.2　核心处理流程

从图 3 中可以看出, 我们所提出的基于线图分析方法改进的鲁棒性文本分析算法的核心处理流程主要包含两个部分: 收集推测未识别语句串终结符集和重新调整分析器的当前分析状态。推测未识别语句串终结符集主要用来对待分析语句中未识别语

句串进行错误推测,以便根据未识别语句串寻找到可能的正确语句串。当通过这样的推测寻找到对下一步的语法分析有利的分析成分时,重新调整分析器的当前分析状态,主要调整的数据结构包括当前分析状态 T_i 下的议程表(例如添加、删除新的成分,或者重新计算当前成分和后续成分的分析位置)和线图(主要是向线图中添加新的对应当前分析状态 T_i 的活动弧)的内容。具体算法如下所示。

输入: 包含未识别语句串的待分析语句。
输出: 所有的部分分析结果。
执行下列过程直到输入为空:
步骤一:
a) 如果议程表为空,查找下一个输入词语的词类,并将它们都加入到议程表中。
b) 从议程表中选择一个成分(假定该成分为 C,其跨度从位置 P_1 到 P_2)。如果该成分为代表未识别语句串的 R,则执行步骤二,否则执行步骤三。
步骤二:
a) 在当前分析状态 T_i 下从规则库中收集规则库推测终结符集 Rule_T 与当前活动弧推测终结符集 Activate_T。
b) 根据集合 Rule_T 与 Activate_T,对未识别语句串进行错误推测,确定未识别语句串对应的可能的正确语句串以及相应的词类 R_i。
c) 对于每一个 R_i,创建新的在当前分析状态 T_i 下的成分 C_R_i,其跨度与 R_i 在输入句子中所对应的跨度一致。
d) 对于文法规则中每条形式为 $X \rightarrow C\_R_iX_1 \ldots X_n$ 的规则,增加一条活动边 $X \rightarrow$ 。$C\_R_iX_1 \ldots X_n$,其跨度与 R_i 在输入句子中所对应的跨度一致。
e) 对于 Chart 中每条形式为 $X \rightarrow X_1 \ldots$ 。$C\_R_i \ldots X_n$ 的活动弧(假设其跨度从位置 P_0 到 P_1),增加一条活动边 $X \rightarrow X_1 \ldots C\_R_i \circ \ldots X_n$,其跨度从位置 P_0 到 P_2。
f) 对于 Chart 中每条形式为 $X \rightarrow X_1 \ldots \ldots X_n \circ$ C_R_i 的活动弧(假设其跨度从位置 P_0 到 P_1),增加一条新的成分 X 到 Agenda 中,其跨度从位置 P_0 到 P_2。返回步骤一。
步骤三:
a) 对于文法规则中每条形式为 $X \rightarrow CX_1 \ldots X_n$ 的规则,增加一条活动边 $X \rightarrow$ 。$CX_1 \ldots X_n$,其跨度从位置 P_1 到位置 P_2。
b) 对于 Chart 中每条形式为 $X \rightarrow X_1 \ldots$ 。$C \ldots X_n$ 的活动弧(假设其跨度从位置 P_0 到 P_1),增加一条活动边 $X \rightarrow X_1 \ldots C \circ \ldots X_n$,其跨度从位置 P_0 到 P_2。
c) 对于 Chart 中每条形式为 $X \rightarrow X_1 \ldots \ldots X_n \circ$ C 的活动弧(假设其跨度从位置 P_0 到 P_1),增加一条新的成分 X 到 Agenda 中,其跨度从位置 P_0 到 P_2。返回步骤一。

在图 4 所示的算法步骤二 b)中,可以根据实际应用中输入语句包含的不同错误类型选择合适的算法推测未识别语句串。在对汉语网络文本的分析中,因为拼音输入方法应用的普遍性,文本错误往往是因为拼音输入的选词错误所导致。针对这类错误,我们可以基于集合 Rule_T 与 Activate_T 所包含的终结符所对应的拼音与未识别语句串的拼音做汉字文本的错误推测,可以采用基于拼音或者基于拼音字符串最小编辑距离的方法寻找可能的正确文本。

3 实验

为了测试所提出算法的有效性和算法效率,我们在燕方法的基础上实现了本文所提出的算法,其中错误推测与纠正部分采用拼音的同音匹配方法,即试图将错误文字纠正为拼音相同的正确文字,然后在特定领域下与燕方法进行实际文本的分析对比。

3.1 评测指标

参考文献[6-8]中所使用的评测方法,我们从分析器的“平均分析循环次数”和“分析度”两个角度展开评测:

1) 平均分析循环次数[6-7]:分析循环(a parse cycle)是专用于衡量线图分析器的一个指标。一次分析循环指的是一个成分从议程表中取出到线图中用来增加新活动弧或者递进、归约出新的成分的过程。输入语句的平均分析循环次数,体现了分析器的效率。在实验中因为基于同音词汇拼音纠错情况地存在,平均分析循环次数会有所增加。

2) 分析度[8]:分析度的计算公式如下:

$$\text{分析度} = \frac{\text{被分析器接受语句的数量}}{\text{分析语句总数量}}$$

当输入的待分析句子都是应接受的领域内文本时,分析度越高说明分析器的鲁棒性越好。

3.2 评测集组织与评测结果

实验使用了北京得意音通技术有限公司提供的1 000 条租房领域句子作为评测数据, 其中正确语句和错误语句各一半, 错误语句中同音拼写错误所占比例约为 80%。我们分别用燕方法和本文方法进行分析, 并计算上述两个评测指标。

从评测集中随机抽取 360 条正确语句和 240 条错误语句进行分析, 作为一次实验, 重复三次, 计算总的分析循环次数和被分析器接受的语句, 计算平均分析循环次数和分析度, 评测结果如表 1 所示。

表 1　平均分析循环次数和分析度

	分析循环次数	接受语句数	输入语句总数	平均分析循环次数	分析度
燕方法	10 515	1 366	1 800	5.842	75.89%
本文方法	11 500	1 568	1 800	6.389	87.11%
相对提高				9.363%	14.78%

4 结论与讨论

从表 1 的实验结果可知, 与燕方法比较, 本文所设计的鲁棒性文本分析算法以增加一定比例的平均分析循环次数为代价, 换取了分析度的更多提升, 实际上将不被分析器接受的领域内句子数相对减少了 46.54%。

本文算法只对输入句子由于错误导致分词过程无法识别部分文字的情况进行处理, 如果错误结果能够被分词过程识别为其他词类(例如采用领域无关的分词算法, 可识别所有汉字单字), 则算法无法启动错误推测和纠正的处理过程。在面向领域的应用中, 如果基于领域词表进行分词和句法分析, 则输入错误导致正确的词变为另一个合法领域词的可能性较小, 算法能比较有效的推测输入错误并纠正为正确的领域词。

另外, 2.2 节所示算法流程主要针对处理汉语的情况进行描述。对于类似英语这样的字母语言, 输入包含一个字母的错误, 经常会导致原单词变为一个错误的单词, 应用上述算法的效果是基于当前分析状态 T_i 下的推测终结符集合, 试图将错误的单词纠正为具有某些词性的单词, 这样要比盲目的基于编辑距离纠正可能会好一些。

本文算法是在基于线图的分析算法基础上附加错误推测与纠正的处理, 以提高实际应用时句子分析的鲁棒性, 对于分析过程所应用的文法规则和活动弧的递进归约策略并没有限制。

未来的工作是在本文的算法框架基础上, 研究更好的错误推测方法, 包括尝试利用拼音音节的最小编辑距离来推测正确的输入语句; 探索鲁棒性分析算法中可能的剪枝策略, 以提高算法的分析效率。

5 参考文献

[1] 冯志伟. 自然语言处理中的概率语法[J]. 当代语言学, 2005, 7(2): 166-178.

[2] 刘智博, Michael Brasser, 郑方, 徐明星. 一个基于文本输入的口语对话系统的新的实现策略[J]. 计算机科学, 2006, 22(11): 205-209.

[3] Pengju Yan, Fang Zheng, Hui Sun, and Mingxing Xu. Spontaneous speech parsing in travel information inquiring and booking system[J]. Journal of Computer Science and Technology, 2002, 17(6): 924-932.

[4] 燕鹏举. 对话系统中自然语言理解研究[D]. 北京: 清华大学, 2002.

[5] Ye-Yi Wang. A Robust Parser For Spoken Language Understanding[C]// Eurospeech, 1999, 5: 2055-2058.

[6] Jennifer Foster and Carl Vogel. Parsing Ill-Formed Text Using an Error Grammar[J]. Artificial Intelligence Review, 2004, 21: 269-291.

[7] Mellish. Some Chart-based Techniques for Parsing Ill-formed Input [C]// Proceedings of the 27th ACL, 1989.

[8] Gertjan van Noord. Error Mining for Wide-Coverage Grammar Engineering[C]// Proceedings of the 42th ACL, 2004.

[9] Kay, M. Algorithm schemata and data structures in syntactic processing[R]. Technical Report CSL. Xerox PARC, 1980.