一种基于实例语境的汉语语音识别后文本检错纠错方法

龙丽霞 李蕾 钟义信

北京邮电大学计算机科学与技术学院智能科学技术研究中心 北京 100876

E-mail:longlixia666@gmail.com {leili,zyx}@bupt.edu.cn

摘 要：为了提高语音识别结果的正确率，依据“信息-知识-智能”转换的思想，本文提出了一种基于实例语境的语音识别后文本检错纠错方法：通过在线查找大量鲜活语料，构建基于实例的语境知识库，并融合语法和语义知识，将识别结果置于特定语境中分析，找出错误点并予以纠正。初步实验结果表明，本文算法具有比较高的检错正确率，并对“字数不变”型错误语句的纠正有较好的表现，使语音识别正确率提高了 20%。

关键词：语音识别 实例集 语境知识 检错 纠错

An Example-Context-Based Approach for Speech Recognition Error Detection and Correction

Lixia Long Lei Li Yixin Zhong

(C IST, S CST, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: In order to improve the accuracy rate of speech recognition, an example-context-based approach for speech recognition error detection and correction was proposed in this paper according to the theory of “Information – Knowledge - Intelligence” transformation. We analyze the recognition results using example-based specific context knowledge base which was constructed by searching great amounts of fresh corpora online, finding out the error and correcting it. Tested with 1017 experimental sentences, the error correction system in this paper has a high accuracy rate of error detection, and well performance in correcting the error sentence typed as “the number of words is not changed”, increasing the accuracy rate of speech recognition by 20%.

Key words: Speech recognition, Example set, Context knowledge, Error detection, Error correction

1 引言

语音识别技术是让机器通过识别和理解过程把人类的语音信号转变为相应的文本或命令的技术[1]。目前，语音识别率并不总是尽如人意，影响了人机对话的顺利进行。本文通过构建和使用丰富的语言学知识，对受到噪声干扰的语音识别结果文本进行深入的分析，达到一定程度的理解，进行一定程度的检错和纠错，来提高语音识别结果的正确率。

2 相关技术研究现状

目前，仅凭单纯的信号处理已经很难提高语音识别系统的正确率，而是越来越取决于语音识别后处理中的基于理解的纠错能力[2]。如李晶娇等利用“词汇语义驱动”分析方法纠错[3]；韦向峰等基于概念层次网络语言模型方法[2]；沈玺等通过语音和概念相似度确定最终查询概念[4]。上述方法或利用文本上下文的局部语言特征，对相邻词间的接续关系进行分析[5, 6]，或利用规则和语言学知识等来分析[7, 8]。它们大多根据语法、语义知识来判断词语使用正确与否，而没有把语句放到大的语言环境中对句子进行整体分析。

语境研究方面，美国 Hiroshi Sekiya 用出现在目标词前面的连续单词序列作为语境[9]；马红妹等提出了用概念关联层次网络的语境知识表示方法[10]；郑杰等提出一种根据单词与语境之间的关系消除单词歧义的模型[11]。张普曾概括介绍了语境研究的应用或应用前景[12]。可见，语境知识通常以目标词周围词语序列来表示，或用目标词所在篇章段落的一些主要词汇来表示。这对确定

词语语义有一定的作用，但没能体现每个语境词与目标词间的强弱联系，而这种联系在理解语音识别结果是至关重要的。

本文提出了一种基于实例语境的汉语语音识别后文本检错与纠错方法，将句子放到具体的语境中，综合利用语法、语义分析，考察词语的语境关联度，找出与语境不和谐的词语并将其纠正。

3 算法整体设计思路

机器检错纠错是人赋予机器的一种智能，我们认为智能是“信息-知识-智能”转换的过程[16]，本文就是据此来设计的。首先，给定一个领域，从网上搜集在线语料（信息获取），然后整理语料，构建知识库（提炼知识），再利用知识库对输入文本进行检错纠错（智能策略）。

知识库是智能系统的核心组成部分。本文综合语法、语义和语境的语言学知识，以语境信息为主导，以语法和语义信息为构成元素，构建正确常识的语境知识库。考虑到实际应用中的语境变化的多样性，我们采用了基于实例的构建方法。系统基本框架如图 1 所示。

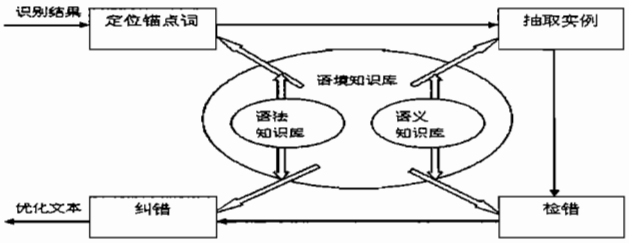

图 1 基于实例语境的文本检错纠错基本框架

相关概念描述：

1）语境知识库：本文所使用的语境是为了表达说话人的目的，在一个语句的上下文之内，各种语法、语义元素正确组合的规律。在特定领域里，有些词的使用频率很高，且能表达某种意义，本文称为核心词。通过在线查找该领域的大量语料，把包含某个核心词的语句搜集起来，形成实例集。所有核心词的实例集收集起来，经分析、计算、整理，就创建了特定领域的语境知识库。

2）语法知识库：本文的语法分析主要是拼音分析，因此，也叫拼音知识库。它为每个词语构建了拼音易混淆词表和词语识别稳定度。

3）语义知识库：提供核心词、普通词在通用语境的语义相似度。

4）锚点词：可信度很高的词，检错纠错的基准点。可信度指的是词语被识别正确的可能性大小。

算法设计思路：从图 1 可见，本文处理都是基于语境知识库进行的。首先是定位锚点词，如果识别结果包含有正确的核心词，就可以利用核心词找到一些匹配的实例，这样就很容易找到识别结果中与正确语境不和谐的词语，即检错。然后再利用拼音易混淆词表提供的纠错建议，对照实例进行纠错。可见，这是以语境信息（或称语用信息）为主导的检错和纠错，而语境信息是建立在语法和语义信息基础之上的。首先，要保证核心词识别正确，需利用词语识别稳定度来定位，即语法分析；其次，搜索最匹配的实例时，需要计算识别结果与每一实例的相关度，即语义分析；最后，确定纠错结果时，

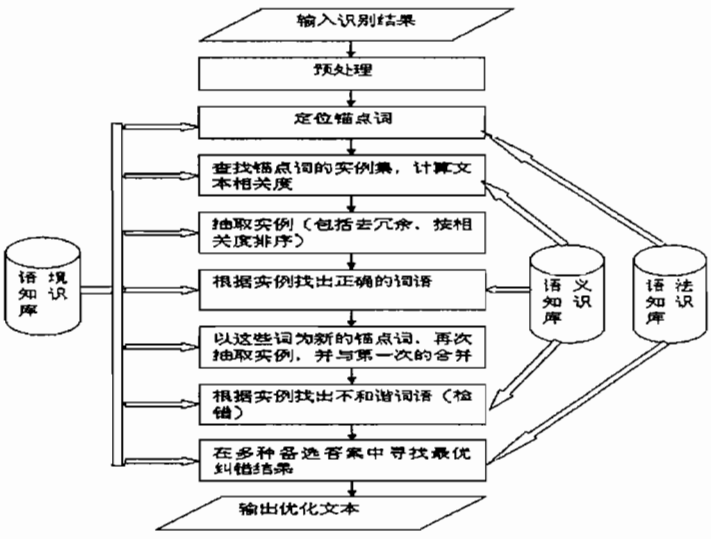

图2 整体检错纠错流程

要综合拼音易混淆词表和正确语境的可能词语计算，这是语境框架下的组合分析。综上可见，本文算法利用了语法、语义和语用信息的全信息知识[17]。整体检错纠错流程如图 2 所示。

4 关键技术介绍

4.1 文本相关度计算

本文的处理对象以句子为单位，所以文本相关度实际上就是句子相关度。我们通过计算两句子词语之间的相关度得出句子相关度。本文提出的词语相关度包括语境关联度（Contextual Correlation， CC）和语义相似度（Semantic Similarity， SS）两部分。由于我们要处理的很可能是一个包含了错误词语的句子，词语之间的语境关联就显得至关重要。例如“你打算什么时候吃饭？”，“什么”和“时候”的关联度要比“吃饭”和“时候”的关联度强。当我们已经确认了“时候”识别正确（锚点词），那么“什么”的可信度就比“吃饭”的可信度要大些。语境关联度通过考察在特定领域里两个词在同一句话中共现的概率而得到。

另外，汉语句子千变万化，我们所建的实例集不可能覆盖所有的句子。如果要求每个文本句都能找到与其正确语句完全相同的实例是不现实的，例如，当文本句为“手足口病是由肠道病毒感染引起的疾病”，而实例集中只存有“手足口病是由肠道病菌感染引起的传染病”这一个实例，我们不能由此就断定“病毒”和“疾病”识别错误，所以语义相似度的计算也必不可少。同时，为了避免出现几个错误词语的相关度之和大于一个正确词语的相关度而影响最佳匹配实例的判断，我们只将语义相似度超过一定阈值 T 的词语加权到句子相关度的计算。本文采用[18]语义相似度算法。句子相关度计算如下：

$$C_{sen} = \sum_{i=1}^{n} C_{Wordi} \tag{4-1}$$

$$C_{Wordi} = SS(wi, wj) + CC(wj, wk)\ (SS(wi,wj) > T,\ wk \text{ 为核心词}) \tag{4-2}$$

在介绍算法前，说明一下本文中实例的表示形式。我们把一个实例分成了三部分：核心词、前向语境词表和后向语境词表，且词表中的每个词具有属性值：与核心词的语境关联度。这样表示可以保证锚点词前面的文本词与其前向语境词匹配，后面的文本词则与后向语境词匹配。这就提高了匹配的效率。算法描述如下：

S1：将文本句中每个词语（文本词）的初始权值置为零；

S2：从第一个词开始，依次与实例中的词语进行匹配。若匹配成功，将该文本词的权值置为：Cword = 1（语义相似度 SS=1）+语境关联度（CC）；

S3：计算文本句总权值 Csen，若为零，返回零，退出程序，否则，转 S4；

S4：依次考察文本句中的每个词，如果权值 Cword 为零，则将该词与它前面文本词已匹配上的语境词之后的词语进行匹配，计算语义相似度 SS，找到一个最大值 Sm，若 Sm 大于设阈值 T，则将其权值置为：Cword = Sm + CC；

S5：计算文本句总权值 Csen，返回 Csen。

4.2 实例抽取

通过计算句子相关度，可以把对应每一个核心词的最大相关度实例抽取出来，抽取的实例有一个或多个。这里的最大相关度是对同一个核心词的实例而言，不计算所有实例的最大相关度是考虑实例抽取的全面性。如文本句“手足口病是由多种肠道病毒感人

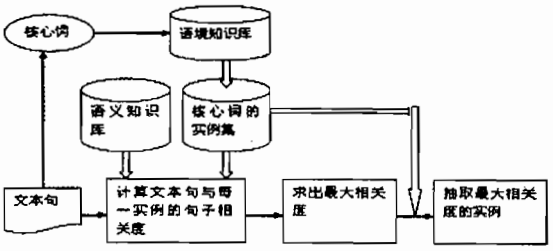

图3 实例抽取流程

引起的常见传染病”。而实例集中，存有以“病毒”为核心词的实例“手足口病是由多种肠道病毒感染引起的一种疾病”，和以“传染病”为核心词的实例“手足口病是一种常见传染病”。显然，第一个实例与文本句的相关度要高于第二个实例，如果只抽取所有实例中相关度最高的实例，那么第二个实例就会被筛掉，“常见”这个词很可能会被认为是识别错误。

此外，为提高效率，我们对抽取出的实例进行了去冗余和排序处理。如果文本句中的锚点词包含多个核心词，就有可能抽取到相同的实例，需去掉。同时，我们应该先处理相关度高的实例。

4.3 检错

检错依据是词语可信度（Confidence Score），由公式（4-2）给出。预先设定一个阈值，达到了就认为识别正确。可信度实际上就是与某一个给定实例的词语相关度。如果某个词语识别错误，它跟一个相关度很高的词语出现在同一语句的概率是很小的。我们把所有已抽取的相关度高的实例处理完毕，找出识别正确的词，剩下的才判为识别错误。经测试，这种方法是可行的。

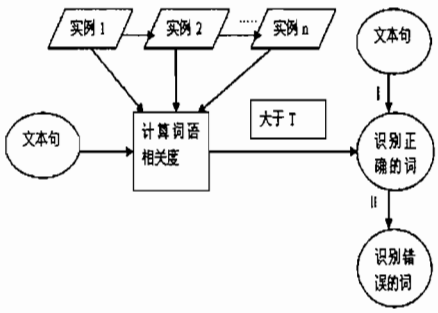

图4 检错流程

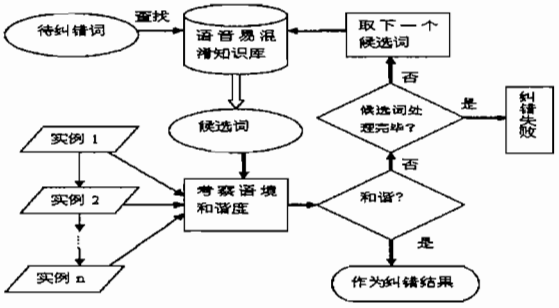

图5 纠错流程

4.4 纠错

对于一个句子，如果句中的某个词被识别为另一个词，这个句子肯定存在某种不合理性。如果将该错误词替换为另一个语音相近的词后能构成一个合理的句子，这个词可能就是正确的。这是本文纠错算法的思想依据。文中，语音相近的候选词由拼音易混淆词表提供。算法流程见图5。

这里的问题是：如果原正确多音节词语被识别为几个单字，或是几个单字被识别为一个词语，如“肠道”识别为“厂豆”，后者判为两个词“厂”、“豆”，它们的候选词都不可能再与“肠道”匹配。为此，本文在第一次纠错后添加了一个优化算法，对未纠正过来的连续错误词语进行重新组合，进行二次纠错。考虑到算法的复杂度，我们只考虑了最普遍的几种情形。具体算法如下：

S1：考察纠错处理后仍未纠正过来的词语，若为空，则跳转至S4，否则，转S2；

S2：若未纠正过来的词语是孤立的双音节词（双），则将其拆成两个单字（单+单）；如果为单+单，则合并双；如果为单+双，则重组为双+单；反之，若为双+单，则重组为单+双：若为单+双+单，则重组为双+双，否则，转S4；

S3：进行第二次纠错（纠错算法同前面）；

S4：输出纠错结果，退出程序。

5 系统测试与结果分析

本文使用真实语音识别文本进行测试，并将测试结果与人工检错纠错结果进行比较。选“手足口病”为测试领域，从网上搜集了近三万字的关于手足口病的文档建立语境知识库。测试分为封闭和开放测试。并对添加优化算法前后的系统分别做了测试。同时，为了更准确地分析测试结果，本文对语音识别错误的语句进行了错误类型归类和统计。

表 1 封闭测试结果统计（添加优化算法前）

测试项目	测试结果	人工判断
总测试句子数	2036	2036
语音识别正确的句子数	615	615
检错正确的句子数	1427	1951
纠错后正确的句子数	867	1649
检错错误但纠错正确的句子数	105	0
检错正确率	70.1%	95.8%
语音识别正确率	30.2%	30.2%
纠错后句子的正确率	42.6%	81.0%
识别正确率提高	12.4%	50.8%

表 2 封闭测试结果统计（添加优化算法后）

测试项目	测试结果	人工判断
总测试句子数	2036	2036
语音识别正确的句子数	615	615
检错正确的句子数	1427	1951
纠错后正确的句子数	1031	1649
检错错误但纠错正确的句子数	105	0
检错正确率	70.1%	95.8%
语音识别正确率	30.2%	30.2%
纠错后句子的正确率	50.6%	81.0%
识别正确率提高	20.4%	50.8%

表 3 开放测试结果统计（添加优化算法前）

测试项目	测试结果	人工判断
总测试句子数	1006	1006
语音识别正确的句子数	414	414
检错正确的句子数	236	984
纠错后正确的句子数	442	911
检错错误但纠错正确的句子数	304	0
检错正确率	23.5%	97.8%
语音识别正确率	41.2%	41.2%
纠错后句子的正确率	44.0%	90.6%
识别正确率提高	2.8%	49.4%

表 4 开放测试结果统计（添加优化算法后）

测试项目	测试结果	人工判断
总测试句子数	1006	1006
语音识别正确的句子数	414	414
检错正确的句子数	236	984
纠错后正确的句子数	460	911
检错错误但纠错正确的句子数	304	0
检错正确率	23.5%	97.8%
语音识别正确率	41.2%	41.2%
纠错后句子的正确率	45.7%	90.6%
识别正确率提高	4.5%	49.4%

表 5 封闭测试识别结果错误类型统计

		识别错误句子总数	错误类型				
			丢字	添字	字数不变(561)		
					词->词	多音节词->单音节词	单音节词->多音节词
		1421	144	154	667	328	128
纠正的句子数	优化算法前	253 (17.8%)	0	0	253 (37.9%)	0	0
	优化算法后	419 (29.5%)	0	0	253 (37.9%)	98 (29.9%)	68 (53.1%)

表 6 开放测试识别结果错误类型统计

		识别错误句子总数	错误类型				
			丢字	添字	字数不变(561)		
					词->词	多音节词->单音节词	单音节词->多音节词
		592	28	47	323	145	49
纠正的句子数	优化算法前	69 (11.7%)	0	0	69 (21.4%)	0	0
	优化算法后	94 (15.9%)	0	0	69 (21.4%)	13 (9.0%)	12 (24.5%)

从封闭测试结果可见，本算法能使语音识别正确率提高约 20%，纠错能力可达到人工水平的一半，且检错的正确率较高（70.1%），但系统支持开放测试的性能较低。下面从语音识别结果和算法两个方面来分析测试结果。

（1）语音识别结果分析

如表 5 所示，根据识别结果的句长变化，可把错误的识别结果分为丢字、添字、字数不变三大类型，其中，“字数不变”型又分为：一个词识别为另一个词（一对一）、一个多音节词识别为多个单音节词（一对多）、连续单音节词识别成一个多音节词（多对一）。由测试结果统计可看出，在加入优化算法前，系统纠正的都是“一对一”错误型的句子。上文已分析过，本文算法以一元词为处理单元，由于每个词的候选词的字数都是相同的，对于字数变化了的词，不可能再找到正确的候选词。“一对多”错误型跟“丢字”实质上是相同的，而“多对一”与“添字”一样。添加优化算法后，部分“一对多” 和“多对一”句子被纠正过来，使纠错效率有了很大提高。

（2）算法分析

本文的算法是综合利用语法、语义、和语境信息的全信息检错纠错算法，下面从这三个方面来分析测试结果。

a)语法层。本文利用的语法知识主要是根据拼音特征分析建立每个词的拼音易混淆词表和词语识别稳定度。因现有的拼音知识库还存在许多缺陷，如，没有考虑吞音、添音、非常用词欠缺等，致使拼音易混淆词表里找不到需要的候选词，即使检出错误也纠正不过来。b)语义层。语义相似度计算不准确会影响检错的输出结果，从而导致纠错失败。如，语音识别结果为：“家长可孩子养成良好的个人卫生习惯和饮食习惯 ”被系统判断为识别正确，原因是实例集中存有实例

“家长和孩子养成良好的个人卫生习惯和饮食习惯”，系统把“可”和“和”的语义相似度计算成了1。所以，语义相似度的计算有待改进。c)语境层。实例覆盖不全面是基于实例方法的最大不足。当语境知识库不存在与输入语句相近的实例时，系统就会把未在实例中出现的正确词也判为错误。这就是开放测试检错率低的主要原因。从测试结果还会发现，有部分句子虽然检错结果错误，但纠错结果却是正确的，尤其在开放测试中，纠错的正确率比检错率还高，这是由于当系统找不到更合适的候选词时，就会保留原词，越改越错的情况很少出现。

在测试中我们还发现，有个别识别错误的句子，人工纠错可能会认为识别正确，但本系统却能检测出错误并纠正过来。比如语句“要对玩具、桌椅等进行修补”无论是语法结构还是语义表达都很合理，但却不符合特定的语境，而系统能利用特定的语境知识将其纠正过来，原正确语句应该是“要对玩具、桌椅等进行消毒”。可见，语境知识的重要性在这里也可得到很好的体现。

6 结论

本文依据“信息-知识-智能”转换思想，综合利用语法、语义、和语境知识，对语音识别后文本进行检错纠错，在一定程度上提高了语音识别的正确率。该方法目前主要针对三种错误类型进行纠错处理，而且支持开放测试性能较低。今后将继续改进算法，让语音识别率得到更大提高。

致谢：

本文得到教育部重点项目（108131）、国家自然科学基金项目（60873001）和国家支撑项目（2007BAH05B02-04）的资助。

参 考 文 献

[1] 高新涛，陈乖丽. 语音识别技术的发展现状及应用前景. 信息技术，2007，36（4）：13

[2] 韦向峰，张全，熊亮. 一种基于语义分析的汉语语音识别纠错方法[J]. 计算机科学，2006，(10)：152-155

[3] 李晶皎， 张王利， 姚天顺. 汉语语音理解中自动纠错系统的研究[J]. 软件学报， 1999，(04)：377-381

[4] 沈玺， 王永成. WEB语音检索中查询概念纠错的研究[J]. 计算机仿真， 2006，(02)：223-226

[5] Andrew R Golding. A Winnow-based Approach to Context-Sensitive Spelling Correction[J]. Machine Learning，1999， 34: 107-130.

[6] Lei Zhang， Ming Zhou， Changning Huang. Multifeature-based Approach to Automatic Error Detection and Correction of Chinese Text[C]. Microsoft Research China Paper Collection， 2000. 193-197.

[7] Golding A R. Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction [C]. Proceeding of the 34th Annual Meeting of the Association for Computational Linguistics，1996. 71-78.

[8] 张仰森， 俞士汶. 文本自动校对技术研究综述. 计算机应用研究，2006， 6（8）：8-12

[9] Hiroshi Sekiya， Takeshi Kondo， Context Representation Using Word Sequences Extracted from a News Corpus.. NAFIPS 2005-2005 Annual Meeting of the North American Fuzzy Information Processing Society，2005，783

[10] Robert Stalnaker, On the Representation of Context. Journal of Logic, Language and Information, 1998,**7:** 3

[11] 马红妹. 汉英机器翻译中语境知识的表示和应用. 第六届计算语言学联合学术会议论文集，2001年，278-284

[12] 郑杰，茅于杭，董清富. 基于语境的语义排歧方法[J]. 中文信息学报，2000，14(5)：1—7，15.

[13] 张普. 汉语信息处理与语境研究[A]. 语境研究论文集[C]，北京语言学院出版社。北京，1992：516—540.

[14] 赵军，金千里，徐波. 面向文本检索的语义计算. 计算机学报，2005，28（12）：2068-2078

[15] 黄元清，刘言生. 语境研究概述. 西华大学学报（哲学社会科学版）增刊，2005，339-340

[16] 钟义信. 机器知行学原理：信息、知识、智能的转换与统一原理. 科学出版社，北京，2007：189

[17] 钟义信. 自然语言理解的全信息方法论. 北京邮电大学学报，2004，27（4）：1-12.

[18] 刘群，李素建. 基于《知网》的词汇语义相似度计算方法. 第三届汉语词汇语义学研讨会，台北，2002. 5