基于单字词转移概率的未登录词识别

何燕

联想研究院 100085

heyanc@legend.com

摘要：未登录词识别是目前自动分词中的主要问题。本文采用排除法，通过排除第一趟分词后形成的分词碎片中的单字词来识别未登录词，取得了一定的效果。以往的未登录词的识别往往需要搜集特定资源，只对某一类型的未登录词进行识别，例如建立中文姓名、外国人名和地名资料库，进行人名、地名的识别。本文未利用任何有关未登录词的资源。实验结果表明，利用单字词转移概率和少量规则来识别未登录词，作为识别未登录词的一种新思路，是行之有效的。

关键词：未登录词 单字词

IDENTIFICATION OF UNLISTED WORDS ON TRANSITIVE PROBABILITY OF MONOSYLLABIC WORDS

Abstract: Identifying unlisted words is a peculiar problem to Chinese segmentation. The variety and vast amount of unlisted words becomes a bottle-neck in processing huge corpora. The paper proposes a simple scheme: identifying monosyllabic words in fragments in raw corpora by calculating the transitive probability of two monosyllabic words linked together in fragments in checked corpora. The contiguous characters that are not monosyllabic words form unlisted words. The result of a preliminary open test is inspiring.

Keywords: unlisted words, monosyllabic words

1、引言

汉语自动分词是汉语自然语言处理的基础。目前汉语自动分词中的主要难题是未登录词的识别问题。未登录词，即分词系统中词表中未收录因而机器不认识的词，包括中国人名、外国地名、中国地名、缩略语、新词等。机器自动分词一般会把未登录词切成一个一个单字。分词碎片是指机器自动分词后形成的若干个连续单字。这些单字或者是单字词，或者是未登录词的一部分。

目前对未登录词的解决方案主要分为两种：

一种是专门解决某一类未登录词，如：人名、地名、机构名等，主要做法为：针对某一类专名建立资料库，例如，中文姓名资料库（台湾张俊盛等 1992，宋柔 1993，孙茂松、黄昌宁 1995）、英美姓名译名用字表（孙茂松等，1995）、中国地名库（沈达阳等，1995），等等。根据这些已有的资源统计出各姓氏、人名、地名用字的概率，在未登录词出现的句子中再以动态规划的方法求出可能最佳的那一类的专名。一般在纯统计的方法中往往还要结合规则，尤其是一些可利用的上下文启发信息。

这些方法都取得了很好的识别效果，其不足之处在于需要建立每一类未登录词的特定资源库，而未登录词的种类可以说是不胜枚举的。有的资源甚至无法得到。如果要尝试寻求一种可以实用的未登录词识别技术，那么一揽子解决方案是更优的选择。

已经被提出的一揽子解决方案有：张普（1988）的有穷多层列举法，王开铸（1995）的无词典自动分词的研究，沈达阳（1997）、刘挺（1998）的局部统计法， 陈小荷（1999）的一揽子解决方案，等等。本文利用单字词共现概率进行未登录词识别，是解决未登录词问题的又一一揽子方案的尝试。

2、未登录词识别的基本思想和从训练语料中获得的资源[1]

2、1 未登录词识别的出发点

在分词碎片中，未登录词和单字词间杂在一起，都以单字的形式出现。如果能首先确定分词碎片中哪些是单字词，那么剩下的相邻的字组合起来就形成了未登录词。如：“已 请 台 湾 的 ”这一分词碎片中，如能确定“已”、“请”、“的”是单字词，剩下的“台湾”就被归为未登录词了。本文的基本算法是：根据单字词成词的概率和单字词共现的概率确定分词碎片中的单字词，从而区别出未登录词。

2、2 关于单字成词和单字词共现频率的研究

据吕叔湘（1963），汉语里的新词都是复合词，单字词集合是基本上不会增长的封闭集合。据苑春法（1995）根据语素库的统计，能单独成词的语素一共有 2878 个。我们对 350 万经过机器分词和人工校对后的训练语料的统计，支持苑春法的结论。2878 个单字成为单字词时，分属各个词类和语义类，两个单字词共现，受到一定限制。我们现在从经过机器分词、人工校对后的语料中获取单字成词和单字词共现的概率，数据稀疏是一个比较大的

[1] 本文所使用的语料为：北京语言文化大学信息处理研究所与香港理工大学合作课题完成的机器自动分词并经人工校对后的全部语料，北京语言文化大学信息处理研究所与清华大学合作课题完成的机器自动分词并经人工校对后的部分语料。感谢上述单位和参加课题以及校对工作的所有老师和同学。

问题，今后的研究会寻求如何从大规模真实文本中扩大现在从训练语料中得到的数据。

2、3 从训练语料中得到的数据

(1) 单字出现概率和单字单独成词概率

(2) 单字词共现对和频度及其单字词的转移概率

如：“建 国 多 年 来 ，我 国 有 了 很 大 发展。”这句话里，一共有 11 个单字词，单字词共现的情况为“建国”、“国多”、“多年”“年来”“我国”“国有”“有了”“了很”“很大”。统计所有在训练语料中出现的单字共现对并累计其出现次数。然后根据转移概率的计算公式

$$P(c2 \mid c1) = P(c1c2) / Pc1 = \frac{F(c1c2)}{Fc1}$$

计算每个单字共现对的转移概率。

(3) 单音词表：包括常用的但不常组词的连词、副词、代词、介词等。

(4) 单音量词表：汇集了常用量词，包括名量词、动量词。

(5) 单音动词表：单音动词表统计了常用单音动词。

3、算法描述

1） 将分词碎片中每相邻的两个字的起始位置及其作为单字词共现的转移概率记入二维数组。

2）从数组的第一行开始，逐一检查每行中的两个字是否能单独成词。将连续同为“是”的字记入一个变量（字与字之间用空格隔开，这是一串单字词），将连续同为“否”的字记入另一个变量（字与字之间无空格，将这些字组合起来视为候选未登录词）。判断这些字为“是”或“否”的标准有：两个字作为单字词的转移概率是否大于指定的阈值；是否是数词+量词的结构；每个字的单字出现概率与成单字词的概率如何。

3）再进一步检查被记入变量中的字的成词情况，区分出其中的单字词和未登录词，记入输出句。

4、对分词碎片中的各种情况的分析及处理策略

4、1 矛盾信息的解决

我们寻找分词碎片中的单字词首先依据的是两个单字词的转移概率，但是从训练语料中的得到的信息会出现矛盾。如：分词碎片“中 国 的”，“中国”的单字词转移概率为 0，“国的”的单字词转移概率为 0.02243，针对这类问题的规则为：

规则 1：如果两个字作为单字词的转移概率为零，单字出现概率都大于 0.002，但其成单字词概率都小于 0.7，则这两个字应该合起来组成一个词。

规则 2：如果两个单字（字 1、字 2），其作为单字词的转移概率大于 0.00004，与字 1（或字 2）相邻的一个字（字 3）作为单字词和字 1（或字 2）作为单字词的转移概率小于 0.00004：如果字 1（或字 2）的成单字词概率大于 0.6，那字 1（或字 2）是单字词，不应该和字 3 组合成词；否则，应该和字 3 组合成词。

规则 3：如果两个单字为数量结构，则这两个单字应切分开。

4、2 检查多字组合而成的可能未登录词的边界

如：“她 朝 达 旺 仓 喊 着”这一分词碎片中，在用单字词转移概率判断之后，形成这样的切分：“她 朝达旺仓喊 着”，很显然，“朝”是一个常见的介词，“喊”是一个常见的动词，在这里很容易将它们和真正的未登录词人名“达旺仓”剥离开。这种检查边界的方法基于对某些常见单字词的归类，如上面提到单音词表、单音动词表，主要就是用于检查边界。由于汉语兼类的现象比较普遍，为了确保检查边界时的正确率，有以下规则：

规则 5：对于二字的候选未登录词，若其中一字是动词而另一个字不是动词，并且这个动词的成单字词概率大于 0.75，而另一字的成单字词概率不小于 0.2，则将这两个字切分开。

规则 6：对于三字的候选未登录词，如果第一字或第三字在单音动词表中出现，第二字没有在单音动词表中出现，那么将动词从这个候选未登录词中切分出来。（如果连续三个字都是动词，那么就不属于超出边界的问题了。）

规则 7：对于三字或三字以上的候选未登录词，如果第一字或最末一字在单音词表中出现，将这个候选未登录词的首字或末字切分出来；如果第一字或最末一字在单音动词表中出现，而且第二字和倒数第二字不是动词，将动词切分出来。

4．3 利用构词法知识

相邻的两个单音动词，有可能形成一个未登录词，如：看到、找出、震颤、评估。但有的动词很少和别的动词组合成词，如：有、爱。有的动词和别的动词组合成词时，它只能出现在后字的位置上，如：来，去。当两个动词的单字相对成词概率都比较高时，这两个动词一般不会组合成一个词，如“想 笑”。检查每相邻的两个单音动词，如果它们符合

动词的构词规则，那么就把这两个字合起来形成为一个未登录词。

5、实验结果报告

我们进行了小规模的开放测试。开放测试的语料是从网上下载的2000年5月3日的北京青年报上的几篇新闻报道，共9546字。

下面就开放测试中的各种情况作详细分析。

开放语料中对专名的识别情况：

实际人名出现123次，我所辨识的人名出现131次，正确 112次，错误19次，人名识别的召回率为91.06%，正确率为85.5%。

实际地名出现91次，我所辨识的地名出现94次，正确74次，错误20次，召回率为81.32%，正确率为78.72%。

实际机构名出现29次，我所辨识的机构名出现26次，正确20次，错误6次，召回率为68.97%，正确率为76.92%。

开放语料中被正确识别出来的人名有：

丘吉尔 / 施罗德 / 福清 / 于红 / 阿琳 / 贾梅沙利 / 艾敏 / 夜船 ……

在开放语料中，被识别错误的人名有：

人史翠 多姆 / 质史蒂 芬 / 克鲁斯提出 / 芬兰人质黎斯托 / 艾迪 多 ……

被正确识别出来的地名有：

伦敦 / 德国 / 柏林 / 汉堡 / 汉诺威 / 瑞士 / 苏黎世 / 西巴丹岛 / 菲律宾……

错误的和未被识别出的地名有：

特 拉法加 / 雅 加达 / 马尼 拉 / 和 鲁岛 /中 村 / 美 国 / 新 加 坡前……

开放语料中对非专名未登录词的识别情况

实际非专名未登录词出现379次，我所辨识的非专名未登录词出现352次，正确253次，错误99次，非专名未登录词识别的召回率为66.75%，正确率为71.86%。

其中被正确识别的非专名未登录词，如：

水瓶 / 雕像 / 英镑 / 本国 / 带来 / 营销 / 高效 / 互联网……

识别错误和未被识别出的未登录词实例：

被 捕者 / 数 以 百计 / 与 会者 / 准 时 / 面 对 / 导 向 / 不 同……

只出现一次的未登录词的识别情况

用局部统计的方法可以识别出现不止一次的未登录词，所以识别只出现一次的未登录词是比较重要的问题。开放测试语料中只出现一次的未登录词233个，被辨识出来的未登录词188个，其中正确的145个，错误的88个，召回率为62.23%，正确率为77.13%。

6、结论

在中文信息处理中，各种积累起来的资源都是很宝贵的。本文尝试在自动分词中利用较少的资源用相对简单的办法来识别未登录词，虽然结果并非完美，但是还是令人鼓舞的。对于本次实验中出现的若干问题，如：训练语料的规模不足带来的数据稀疏问题、规则的有效性和优先级问题以及最后开放测试规模不足等问题，我们将在今后的研究中进一步深入。

参考文献

[1]陈小荷，1999，自动分词中未登录词问题的一揽子解决方案，语言文字应用，3

[7]孙茂松等，1995，中文姓名的自动辨识，中文信息学报，2

[8]孙茂松，1995，中文姓名的自动辨识，中文信息学报，9

[9]沈达阳，1995，中文地名的自动辨识，见：陈力为，袁琦主编，计算语言学进展与应用，北京：清华大学出版社

[10]宋柔，1993，基于语料库和规则库的人名识别法，见：陈力为，袁琦主编，计算语言学研究与应用，北京：北京语言学院出版社

[11]谭红叶等，1999，中国地名自动识别方法研究，见：黄昌宁，董振东主编，计算语言学文集，清华大学出版社

[13]王省等，1999，基于统计和规则相结合的中文姓名识别，见：黄昌宁，董振东主编，计算语言学文集，清华大学出版社

[14]吴立德等，1997，大规模中文文本处理，上海：复旦大学出版社

[15]郑家恒，刘开瑛，1993，自动分词系统中姓氏人名处理策略探讨，见：陈力为，袁琦主编，计算语言学研究与应用，北京：北京语言学院出版社

[16]郑家恒，谭红叶，1998，基于变换的中文姓名识别技术探讨，见：黄昌宁主编，1998 中文信息处理国际会议讨论文集，北京：清华大学出版社

[17]张俊盛等，1992，多语料库作法之中文姓名辨识，中文信息学报，3

[19]张普，张光汉，1998，现代汉语“有穷多层列举”自动分词方法的讨论，见：武汉大学语言自动处理研究组编，汉语自动处理，武汉：武汉大学出版社

[20]张小衡等，1997，中文机构名称的识别与分析，中文信息学报，4

[22]吕叔湘，1963，现代汉语单双音节初探，中国语文，1

[24]孙茂松，1996，Word Segmentation and Part-of-Speech Tagging for Unstricted Chinese Text，2nd，VSMM