

文章编号: 1003-0077(2007)03-0083-09

## 中文网络聊天语言的奇异性与动态性研究

夏云庆<sup>1</sup>, 黄锦辉<sup>2</sup>, 张普<sup>3</sup>

(1. 清华大学 信息技术研究院, 北京 100084; 2. 香港中文大学 系统工程系, 香港;  
3. 北京语言大学 网络教育学院, 北京 100083)

**摘要:** 随着互联网走入社会生活, 网络聊天逐渐成为一种新的沟通渠道, 网络聊天语言便应运而生。这类语言的日益丰富, 给语言信息处理带来了新的挑战。研究发现, 困难主要来自网络聊天语言的奇异性与动态性。本文借助真实网络聊天语言文本, 对网络聊天语言的奇异性与动态性进行详细分析和归纳, 并设计了面向解决奇异性与动态性问题的网络聊天语言文本识别与转换方法。我们先以网络聊天语言语料库为基础建立网络聊天语言模型和语言转换模型, 通过信源<sup>3</sup>信道模型实现网络聊天语言向标准语言的转换。但该方法过于依赖网络聊天语言语料库, 虽然能较好解决奇异性问题, 但不能处理动态性问题。因此, 我们进而以标准汉语语料库为基础建立文字语音映射模型, 对信源<sup>3</sup>信道模型进行改进, 最终有效解决了网络聊天语言的动态性问题。

**关键词:** 计算机应用; 中文信息处理; 网络聊天语言; 奇异性; 动态性; 语言信息处理

**中图分类号:** TP391

**文献标识码:** A

### Toward Anomalous and Dynamic Nature of the Chinese Network Chat Language

XIA Yun-qing<sup>1</sup>, Kam Fai Wong<sup>2</sup>, ZHANG Pu<sup>3</sup>

(1. Research Institute of Information Technology, Tsinghua University, Beijing 100084, China;  
2. Dept. of SEEM, The Chinese University of Hong Kong, Hong Kong, China;  
3. Network Education College, Beijing Language and Culture University, Beijing 100083, China)

**Abstract:** Network chat language becomes ubiquitous due largely to the rapid proliferation of Internet applications. Online chat now acts as an important role in human communication, which in turn makes Network chat language popular. Network chat language processing is important but difficult. The challenges mainly come from the anomalous and dynamic nature of the new text genre. The two distinct features of Chinese Network chat language are investigated and analyzed in this paper. Methods seeking to address the two features in Network chat language processing are also proposed. We first develop a source channel model to convert chat language to standard language. Unfortunately this method relies too heavily on chat language corpus rendering the method poor in addressing the dynamic nature. We propose to introduce phonetic mapping model constructed with standard language corpus to the source channel model. The extended method is proved effective in addressing the dynamic issue by our experiments.

**Key words:** computer application; Chinese information processing; Network chat language; anomalous nature; dynamic nature; language and information processing

### 1 网络聊天语言的现状和挑战

根据中国互联网络信息中心(CNNIC)的统计, 到 2005 年 4 月底, 我国上网用户已达到 1.002 亿

人, 网民数仅次于美国居世界第二位。今天, 每 13 个中国人就有一个与它“亲密接触”, 互联网正在成为各界人士获取信息的主要通道。社会科学院 2005 年互联网报告<sup>[1]</sup>指出, 我国网民平均每天上网的时间是 2.73 小时, 单纯浏览网络论坛而不发言的

收稿日期: 2006-05-16 定稿日期: 2007-03-21

项目基金: 香港中文大学 Direct Grant (2050330); Strategic Grant (4410001)

作者简介: 夏云庆(1972-), 男, 博士, 助理研究员, 主要研究方向为自然语言处理。

网民只占 38.6%。这个比例说明,网民的上网行为不仅仅是寻找信息,还包含了人际交流的活动。报告指出,通常用来双向交流的交流工具有博客(Blog)、论坛(BBS)、微软 MSN、聊天室、ICQ 和电子邮件,有 68.7% 的网民使用聊天室,66.6% 的网民使用 ICQ/OICQ/QQ,44.8% 的网民使用 BBS,43.9% 的网民使用微软 MSN。这些数据表明:随着互联网进入社会生活,网络聊天逐渐成为一种重要的沟通渠道。

网络聊天渠道的发展进一步方便了交流,也给信息技术领域带来机遇。在商业应用中,越来越多的客户服务/呼叫中心/网上教学<sup>[18,19]</sup> 日渐被互联网聊天解决方案取代,聊天室、BBS 张贴,电子邮件和手机短信等方案逐渐被商家采用,甚至在某些应用中取代了电话这个传统交流工具。网络聊天语言应运而生,并已发展成为一种重要的群体语言。这类语言的出现带来了诸多挑战。例如,由于网络聊天渠道大多可以免费使用,信息杂乱无章,因此被色情信息、犯罪信息和恐怖主义传播者所利用,成为他们扰乱社会安定、制造反社会活动的策划与讨论场所<sup>[20,21]</sup>。他们大量采用奇异的网络聊天语言(黑话),混淆安全监控人员的眼睛,这就造成了安全监控任务的难题。再如,商业上提供基于聊天的客户服务已经屡见不鲜,这些聊天记录同传统的电话记录具有同样的价值,网络聊天语言的使用,阻碍了分析研究人员获取重要信息。于是他们带着这些问题求助于自然语言处理工具,希望通过语言分析处理以“解码”这些奇异词汇<sup>[22-25]</sup>。

中文方面,语言学家在中文网络聊天语言研究方面取得了重要进展。文献[2~5]介绍了中文网络语言的基本特征,文献[6~11]对其造词法、语词类型、语用特点及规范进行了探索,文献[12]指出了其谐音现象,文献[13~15]指出了其语言变异现象。文献[16,17]则从交际和哲学高度对中文网络语言进行了深入分析。本文从自然语言处理的角度对中文网络语言进行研究,指出处理难点所在,并提出适当的处理方法。

我们先看下面三个网络聊天语言的例子:

- (1) 有木有[c1] 银[c2]请我 7 饭[c3] (有没有[n1]人[n2]请我 吃饭[n3])
- (2) 偶[c1]稀饭[c2]这样的 GG[c3] (我[n1]喜欢[n2]这样的哥哥[n3])
- (3) 隔 3 差 5[c1]来看你 (隔三差五[n1]来看你)
- (4) 细八细[c1]又要 FB[c2]去 (是不是[n1]又

要腐败[n2]去啊)

这些例子中,括号里面给出的是每个网络聊天语言例子对应的标准语言。我们用[ci]代表网络聊天语言词汇,[ni]代表对应的标准语言词汇。例如,“有木有”对应着标准语言的“有没有”,“银”对应着标准语言的“人”。类似的网络聊天语言词汇很多,在网络聊天室、聊天记录和论坛(BBS)上随处可见。我们知道,传统语言处理工具的对象是标准语言,假定分析对象(文本)符合常规语法。这样,面对网络聊天语言,它们就显得无能为力了。我们用 ICTCLAS<sup>[26]</sup> 处理例(1)的网络聊天气本,分词结果如下:

有/v 木/n 有/v 银/n 请/v 我/r 7/m 饭/n

ICTCLAS 处理不了“有木有”这个网络聊天词汇。当然这并不说明 ICTCLAS 的性能不强,而是因为 ICTCLAS 不包含网络聊天语言的任何信息(词条、规则和统计数据)。我们再看 ICTCLAS 对例(2)进行词性标注的结果:

偶/b 稀饭/n 这样/r 的/u GG/n

ICTCLAS 将“偶”分析为 b(区别词),将“稀饭”分析为 n(名词)。但是实际上,“偶”在这里用作“我”,应该为 r(代词),“稀饭”代表“喜欢”,应为 v(动词)。因之相对于标准词汇的“奇异”效果,我们定义网络聊天语言的该特性为“奇异性”。网络聊天语言的“奇异性”给文本分析和处理带来了困难。对于处理“有木有”这样的奇异词汇,有人建议将它添加到词典里就可以了,在处理“银”时,再将“人”这个义项添加到标准词典里去。我们反对这样做,因为标准汉语基本不会使用“有木有”这个词汇,也不会用到“银”的“人”这个义项,只有在网络聊天环境中才会这样用到。

有人建议将这些奇异的词汇用一个“网络聊天语言词典”收集,通过查询就能够找出对应的标准词汇。这个建议并不能奏效,原因有二:一,网络聊天语言在用作标准词汇时导致歧义。例如“银”可以用作网络聊天语言,代表“人”,也可以用作标准词汇,表示“银”这种金属物质。这时,仅仅通过词典,很难区别网络聊天语言和标准词汇,更不用说去区分网络聊天语言的多种不同用法。二,通过仔细观察研究,我们发现网络聊天语言变化很快,无法用静态的词典去覆盖。典型地,去年使用的一些网络聊天语言,今年就被淘汰了,同时被更多新的网络聊天语言取代。这就是我们所提出的网络聊天语言的“动态性”。虽然不断更新“网络聊天语言词典”是一个解决方法,但网络聊天语言变化快,要做到及时更新非

常费时费力,而且这些花费永无止境。要解决“奇异性”和“动态性”问题,只依赖一个聊天语料库,似乎走进了死胡同。

中国有句俗语:万变不离其宗。我们认为再动态的网络聊天语言也包含着相对静态的因素。我们的细致观察最终证实了这一想法。我们发现,尽管网络聊天语言文本千差万别,但绝大多数(99%以上)中文网络聊天语言的产生都遵循着一个不变的基本原则,即语音映射。网络聊天语言除了表情图标外,极少是从无到有的创造,绝大多数都对应着原始文字模板。例如“偶”对应着“我”,“稀饭”对应着“喜欢”,都是通过方言语音映射得到的,而“隔3差5”则直接对应了同音词“隔三差五”。可见,网络聊天语言的产生具有明显的语音映射基础。

有了这把网络聊天语言处理的钥匙,奇异性 and 动态性问题便迎刃而解。本文借助真实网络聊天语言文本,对网络聊天语言的奇异性 and 动态性进行详细分析和归纳,并初步设计了面向处理奇异性 and 动态性问题的网络聊天语言文本识别与转换方法。我们先以网络聊天语言语料库为基础建立网络聊天语言模型和语言转换模型,通过信源—信道模型(Source Channel Model)实现网络聊天语言向标准语言的转换。但该方法过于依赖网络聊天语言语料库,虽然能较好解决奇异性问题,但不能处理动态性问题。因此,我们进而以标准汉语语料库为基础建

立文字语音映射模型,对信源<sup>3</sup>信道模型进行改进,最终有效解决了网络聊天语言的动态性问题。

## 2 网络聊天语言的奇异性与动态性

我们认为,网络聊天语言具有两个显著特性,即奇异性与动态性。前者从网络聊天语言的表面就能观察得到,是显性的,因而比较容易把握;后者需要经过对不同时间段的网络聊天语言文本进行对比分析才能得知,是隐性的,因而难于驾驭。我们首先通过丰富的实例对网络聊天语言的奇异性进行分析。这些实例均来自NIL语料库<sup>[24]</sup>。

### 2.1 奇异性

网络聊天语言最引人注目的是其奇异性,它看起来奇特怪异,似乎是错别字却被重复使用,似乎是语法错误却频繁出现。奇异性表现在词汇的使用和表达方法两个方面。但篇幅所限,本文重点讨论网络聊天语言在词汇使用上的奇异性。

在词汇使用上,网络聊天语言或者使用奇异词汇,或者使用标准词汇的奇异意义。奇异词汇的使用是网络聊天语言最初的表现形式。通过对网络聊天语言文本语料库的9524个“奇异”网络聊天语言的形态进行观察分析,我们将网络聊天语言划分为六类,如表1所示。

表1 网络聊天语言的六类形态

网络聊天语言形态	出现次数	比例	例子	注释
中文词汇	8 030	61.8 %	稀饭直来直去。	“稀饭”=“喜欢”
中文短语	670	5.2 %	细八细要开个会议?	“细八细”=“是不是”
英文大写字母	2 119	16.3 %	PF 他们的做事态度。	“PF”=“佩服”
阿拉伯数字	1 021	7.9 %	9494,该打。	“94”=“就是”
上述形态的混合形态	1 034	8.0 %	8 错,怎么弄得?	“8 错”=“不错”
表情图标	110	0.8 %	天气真好, -)	“-)”=“愉快”

表1显示,在中文网络聊天语言中,使用频率最高的还是词汇和短语。但是从统计数字来看,英文大写字母也占据了很大比例。这并不是由于中文网络聊天语言使用了英语,这些英文大写字母大都是汉语拼音的声母缩写。例如,“PF”是“佩服”的汉语拼音“pei4 fu2”的声母缩写。少数英文大写字母是来自英语,例如“ING”反映的是英文现

在进行时态在动词后面后缀“ing”,表示“正在”。恰恰相反,许多中文网络聊天语言词汇却借用了英文单词的发音,例如,“粉丝”是借用英文单词“fans”的发音然后通过汉语拼音映射过来的,这种现象被称为“音译”。音译词在中文网络聊天语言中出现频率不高。

我们再对12983个中文网络聊天语言词汇/短

语进一步分析,我们发现奇异词汇的使用与标准词汇奇异意义的使用具有表2所示的分布。

表2 奇异词汇和标准词汇奇异意义的分布

中文网络聊天语言	词汇个数	比例	出现次数	比例
奇异词汇	224	64.4%	4 519	34.8%
标准词汇奇异意义	56	16.1%	7 839	60.4%
其他	68	19.5%	625	4.8%
总计	348	100%	12 983	100%

表2显示,使用标准词汇奇异意义的个数占16.1%的网络聊天语言在聊天语料库中出现了7 839次,占有中文奇异网络聊天语言总数的60.4%。这一现象表明,使用标准词汇奇异意义的网络聊天语言占绝大多数。

我们认为,网络聊天语言的奇异性给网络聊天语言处理带了如下挑战:1)网络聊天语言的使用群体很大,覆盖面很广,想要穷举所有奇异网络聊天语言并非易事。2)网络聊天语言造成了歧义,尤其是同时使用标准词汇奇异意义的网络聊天语言,这给网络聊天语言处理带来巨大困难。

## 2.2 动态性

动态性反映网络聊天语言的变化。例如,去年使用的一些网络聊天语言,今年就被淘汰了,同时又出现了更多新的网络聊天语言。正如张普教授所说,“流行语都有流行周期,流行一过有可能就不使用了。”流行性和动态性实际上反映的是同一个问题。

为了分析网络聊天语言的动态性,我们将两年内的聊天文本语料根据时间划分为4个相等的子集,每半年的聊天文本为一组,然后统计其中网络聊天语言的重复使用状况。统计结果如表3所示。

表3 网络聊天语言的使用重复率

语料组	2004-7	2005-1	2005-7	2006-1	平均
2004-1	0.882	0.823	0.769	0.706	0.795
2004-7	—	0.885	0.805	0.749	0.813
2005-1		—	0.891	0.816	0.854
2005-7			—	0.875	0.875

排除个别例外情况,总的趋势是:越早的子集同越晚的子集重复使用的网络聊天语言越少。从2004年1月到2006年1月间,网络聊天语言改变

了将近30%。从平均使用重复率来看,这个趋势也是明显的。我们完全可以假设,如果语料库能够覆盖五年的网络聊天语言,我们以每半年的网络聊天语言作为语料子集,这种趋势将会更加明显。

我们认为,网络聊天语言的动态性带来如下挑战:1)新的网络聊天语言不断出现,建立在一个静态字典或者一个静态语料库基础上的方法很难识别新出现的网络聊天语言。2)为了能及时捕获新出现的网络聊天语言,需要创建越来越多的语料库,这需要消耗很大的人力物力。这必然给基于语料库的处理技术提出一个难题,即在时间滞后的语料基础上学习,亦要取得一致的处理效果,其中技术难度很大。

## 3 网络聊天语言与语音映射

### 3.1 网络聊天语言的语音映射特点

我们认为动态的网络聊天语言包含着相对静态的基本元素,我们的细致观察最终证实了这一猜想。我们发现,尽管网络聊天语言文本千差万别,但绝大多数(99%以上)中文网络聊天语言的产生都遵循着一个不变的基本原则,即语音映射。网络聊天语言极少是从无到有的创造,绝大多数都对应着原始文字模板。例如“偶”对应着“我”,“稀饭”对应着“喜欢”,都是通过方言语音映射得到的,而“隔3差5”则直接对应了同音词“隔三差五”。可见,网络聊天语言的产生遵循明显的语音映射原则。有了这个语音映射原则,无论网络聊天语言如何千变万化,本质上的语音映射是稳定的、静态的。我们以语音映射方法为尺度,对2.2节所用的观察样本对网络聊天语言的重复使用状况进行再次分析,统计结果(表4)表明,语音映射是动态网络聊天语言处理的钥匙。

表4 网络聊天语言语音映射方法的使用重复率

语料组	2004-7	2005-1	2005-7	2006-1	平均
2004-1	0.987	0.993	0.989	0.993	0.991
2004-7	—	0.993	0.991	0.986	0.990
2005-1		—	0.997	0.992	0.995
2005-7			—	0.995	0.995

### 3.2 语音映射模型形式化

为了便于语音映射模型的形式化描述,我们先

给出字-字映射模型。即三元组：

$$CM = \langle T, C, Pr_{cm}(T|C) \rangle$$

其中,  $CM$  代表字-字映射模型,  $T$  代表网络聊天语言字符,  $C$  代表标准语言字符,  $Pr_{cm}(T|C)$  代表字-字映射的概率。例如网络聊天语言“7”和标准语言“吃”的字-字映射模型为  $\langle 7, 吃, 0.127 \rangle$ 。显然, 由于字-字映射模型只能通过对网络聊天语言语料库的统计获得, 概率参数严重依赖于网络聊天语言语料库。

语音映射模型具有更强更广泛的映射表达能力, 它将语音映射引入字-字映射模型, 即五元组：

$$PM = \langle T, C, pt(T), pt(C), Pr_{pm}(T|C) \rangle$$

其中,  $PM$  代表字-字映射模型,  $pt(T)$  代表网络聊天语言字符对应的语音标记,  $pt(C)$  代表标准语言字符的语音标记,  $Pr_{pm}(T|C)$  代表语音映射模型的概率。我们用汉语拼音表示中文语音标记。例如网络聊天语言“7”和标准语言“吃”的语音映射模型为  $\langle 7, 吃, qi, chi, 0.357 \rangle$ 。语音映射模型不再依赖网络聊天语言语料库, 它可以从标准语言语料库抽取。网络聊天语言语料库的作用只是加强该模型对网络聊天语言的适应性。

### 3.3 语音映射模型参数估计

语音映射模型参数估计主要回答两个问题：一, 字符映射空间从何而来？二, 语音映射概率如何估计？我们从标准汉语语料库抽取所有汉语字符, 同时将这些字符看作网络聊天语言字符的候选对象, 这样我们就获得了两种语言的字符映射空间。由于字符空间的完整性依赖于标准语言语料库的覆盖面, 因此在实验中我们选择了当前覆盖面最大的中文 GIGAWORD (CN GIGA) 语料库。

既然我们认为是语音映射将标准语言字符和网

$$P_{extended}(A, A^{\sim}) = \begin{cases} \frac{sf \times P_{naive}(A, A^{\sim})}{f_{r_{NIL}}(A_i) \times P_{naive}(A, A_i) + \sum_j sf \times P_{naive}(A, A_j)} & \text{if } f_{re_{NIL}}(A^{\sim}) = 0; \\ \frac{f_{r_{NIL}}(A^{\sim}) \times P_{naive}(A, A^{\sim})}{f_{r_{NIL}}(A_i) \times P_{naive}(A, A_i) + \sum_j sf \times P_{naive}(A, A_j)} & \text{otherwise.} \end{cases} \quad (3)$$

这里, 所有字符  $A_j$  在网络聊天语言语料库的标记文本中出现次数为 0。显然, 概率计算的准确度依赖于这个平滑算子。如果平滑算子太大 (超过 1), 就会忽略网络聊天语言语料库对该计算的影响; 如果太小 (等于 0), 语音映射模型就容易过度适应网络聊天语言语料库。我们采用 0.2、0.4、0.6 和 0.8 分别考察它们对实验结果的影响; 同时利用标

络聊天语言字符关联起来的, 那么语音相似度自然是语音映射模型的基本元素。我们开发了汉语拼音相似度计算工具获得此相似度<sup>[25]</sup>。为保证语音映射模型能代表广泛的标准语言统计规律, 我们在语音映射模型的概率估计中考虑字符在语料库里的出现次数。这样我们得到如下语音映射概率计算公式：

$$P_{naive}(A, A^{\sim}) = \frac{(f_{r_{stc}}(A^{\sim}) \times p_{ys}(A, A^{\sim}))}{(f_{r_{stc}}(A_i) \times p_{ys}(A, A_i))} \quad (1)$$

其中,  $\{A_i\}$  是与字符  $A$  在语音上相似的字符集合,  $A^{\sim}$  来自这一集合,  $f_{r_{stc}}(A_i)$  表示字符  $A_i$  在标准语言语料库中的出现次数,  $p_{ys}(A, A_i)$  表示字符  $A$  与  $A_i$  的语音相似度。

为了加强对网络聊天语言的适应性, 我们使用网络聊天语言语料库来调整语音映射模型。于是, 概率计算公式(1)改写为：

$$P_{extended}(A, A^{\sim}) = \frac{f_{r_{NIL}}(A^{\sim}) \times P_{naive}(A, A^{\sim})}{f_{r_{NIL}}(A_i) \times P_{naive}(A, A_i)} \quad (2)$$

这里,  $f_{r_{NIL}}(A_i)$  代表字符  $A_i$  在网络聊天语言语料库的标记文本中的出现次数。这样一来, 如果某些字符在网络聊天语言语料库的标记文本中出现, 相应的语音映射模型概率将会得到提高。

由于使用了网络聊天语言语料库, 必然出现数据稀疏问题, 也就是说, 某些字符可能不出现在网络聊天语言语料库的标记文本中。为此我们引入平滑算子处理数据稀疏问题, 对所有在网络聊天语言语料库标记文本中出现次数为 0 的字符, 用该平滑算子取代其出现次数。这样公式(2)改写为：

标准语言语料库对平滑算子进行评估。计算公式如下：

$$sf(k) = \frac{f_{r_{stc}}(A_k)}{f_{r_{stc}}(A_j)} \quad (4)$$

这里,  $f_{r_{NIL}}(A_j) = 0$  而且  $A_k \in \{A_j\}$ 。可以看出, 在标准语言语料库出现次数越多, 平滑算子值就越大, 反映了广泛的标准语言统计规律。

## 4 网络聊天语言到标准语言的自动转换

网络聊天语言处理的根本目的是实现从网络聊天语言到标准语言的转换。本文描述两个方法:第一个方法只利用网络聊天语言与语料库,通过原始信源-信道模型,实现网络聊天语言的转换;第二个方法引入语音映射模型以扩展原始信源-信道模型,以解决动态性问题。

### 4.1 基于字-字映射的原始信源-信道模型

信源-信道模型是语音识别和机器翻译技术中的常用方法<sup>[27]</sup>,我们采用该方法在字-字映射模型的基础上实现对网络聊天语言的转换。该方法的基本思想是搜索字符映射空间以得到最可能的字符转换结果。根据 Bayes 法则,该条件概率被分解为字符映射模型和语言模型,如公式(5)所示。

$$\hat{C} = \operatorname{argmax}_C p(T|C) = \operatorname{argmax}_C \frac{p(C|T)p(T)}{p(C)} \quad (5)$$

其中  $T = \{t_i\}_{i=1,2,\dots,m}$  代表输入网络聊天语言文本,  $C = \{c_i\}_{i=1,2,\dots,n}$  代表所有可能的标准语言文本映射,  $\hat{C}$  是最优映射转换结果。  $p(C|T)$  代表字符映射模型,  $p(T)$  代表网络聊天语言模型,二者均可从网络聊天语言语料库中训练获得。

该方法采用字-字映射模型的典型方法,其局限性是对网络聊天语言语料库的过度依赖,数据稀疏问题很严重。如果某网络聊天语言不在网络聊天语言语料库中出现,就很难得到正确的转换结果。这导致该方法无法应付网络聊天语言的动态性。

### 4.2 基于语音映射的扩展信源-信道模型

基于语音映射模型的扩展信源-信道模型能够很好解决动态性问题,这得益于语音映射模型的普遍性。我们在公式(4)中插入语音映射模型,得到一扩展的信源-信道模型,如公式(6)所示。

$$\begin{aligned} \hat{C} &= \operatorname{argmax}_C p(T|M,C) \\ &= \operatorname{argmax}_C \frac{p(C|M,T)p(M|T)p(T)}{p(C)} \quad (6) \end{aligned}$$

这里,  $M = \{m_i\}_{i=1,2,\dots,n}$  代表  $T = \{t_i\}_{i=1,2,\dots,m}$  到  $C = \{c_i\}_{i=1,2,\dots,n}$  语音映射集合。  $p(C|M,T)$  即所谓网络聊天语言转换观察模型,  $p(M|T)$  即语音映射模型,  $p(T)$  即网络聊天语言模型。

同基于字-字映射的原始信源-信道模型相比,

基于语音映射的扩展信源-信道模型的搜索空间得到充分扩大。例如,“银”在聊天语料库中仅被标记用作“人”,因此用基于字-字映射的原始信源-信道模型处理“银们散了”,搜索空间是{“人们散了”,“银们散了”};而在基于语音映射的扩展信源-信道模型中,搜索空间扩大为{“因们散了”,“印们散了”,“吟们散了”,“阴们散了”,“人们散了”,“银们散了”}。这显然提高了对网络聊天语言的动态性的处理能力。

## 5 实验与评测

### 5.1 实验数据

实验中我们用到两类训练语料库,即标准语言语料库和网络聊天语言语料库。我们采用中文 GIGA WORD (CN GIGA)<sup>[28]</sup> 作为标准汉语语料库,采用 NIL 语料库<sup>[24]</sup> 作为网络聊天语言语料库。

我们使用了四个测试集 T#1 ~ T#4,均来自太极论坛(bbs.yesky.com)“大嘴区”。每个测试集包含的聊天语句 500 句,时间戳在网络聊天语言语料库中的聊天语句之后,即 2005 年 8 月到 11 月。这样安排测试集,目的是要比较网络聊天语言转换方法在不同时间段测试语料上的性能,从而观察不同方法在处理网络聊天语言的奇异性和动态性上的效果。

### 5.2 评测指标

在网络聊天语言识别上,我们采用同未登录词识别类似的评测指标,即准确率( $p$ )、召回率( $r$ )和  $F1$  指标( $f$ )。这些指标的定义如下:

$$p = \frac{a}{a+b} \quad r = \frac{a}{a+c} \quad f = \frac{2 \times p \times r}{p+r} \quad (7)$$

其中,  $a$  代表正确判断为网络聊天语言的次数,  $b$  代表错误判断为网络聊天语言的次数,  $c$  代表错误判断为非网络聊天语言的次数。

在网络聊天语言的转换上,我们采用类似机器翻译评测的指标,即精确度( $ac$ ),它的定义如下:

$$ac = \frac{\text{正确翻译的句子个数}}{\text{所有测试句子个数}} \quad (8)$$

### 5.3 实验 1: 原始信源-信道模型(SCM)

训练过程是利用 NIL 语料库进行字-字映射模型的参数估计。训练完成后,我们运行原始信源-信道模型方法,分别处理四个测试集。实验结果如表 5 所示。

表 5 原始信源-信道模型的实验结果

测试集	p	r	f	ac
T#1	0.834	0.853	0.843	0.835
T#2	0.801	0.816	0.808	0.802
T#3	0.772	0.782	0.777	0.773
T#4	0.737	0.765	0.751	0.736

表 6 扩展信源-信道模型的实验结果 (固定平滑算子)

测试集	sf = 0.2				sf = 0.4				sf = 0.6				sf = 0.8			
	p	r	f	ac												
T#1	0.865	0.874	0.869	0.864	0.877	0.891	0.884	0.876	0.875	0.890	0.882	0.874	0.858	0.871	0.864	0.857
T#2	0.865	0.872	0.869	0.866	0.882	0.892	0.887	0.885	0.881	0.890	0.886	0.883	0.862	0.872	0.867	0.864
T#3	0.876	0.869	0.872	0.877	0.892	0.884	0.888	0.894	0.891	0.883	0.887	0.892	0.871	0.866	0.868	0.872
T#4	0.868	0.875	0.872	0.868	0.884	0.891	0.887	0.883	0.881	0.888	0.885	0.881	0.866	0.871	0.868	0.865

表 7 扩展信源-信道模型的实验结果 (以语料库评估值为平滑算子)

测试集	p	r	f	ac
T#1	0.891	0.918	0.904	0.890
T#2	0.900	0.911	0.905	0.900
T#3	0.904	0.898	0.901	0.903
T#4	0.898	0.911	0.904	0.895

### 5.5 讨论 1: 方法性能对比

图 1 给出了不同方法在四个测试集上的 F-1 指数对比曲线。总体上看,在四个测试集上,XSCM 方法在使用语料库评估值(XSCM-v)作为平滑算子时,都取得了最好的效果,F-1 指数达到 90%以上,网络聊天语言转化精确度也超过了 89%。SCM 方的 F-1 指数在测试集 T#1 上取得最好效果,即 84.3%,比 XSCM 在测试集 T#3 上取得的最差效果 90.1%低 5.8%。从性能上看,XSCM 方法比 SCM 方法更能准确处理奇异网络聊天语言。

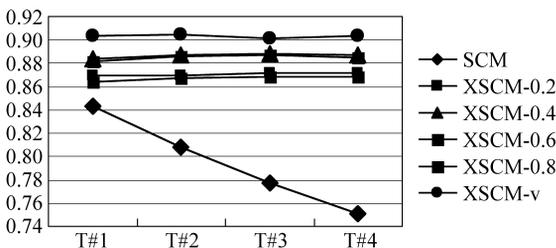


图 1 各种方法性能对比曲线。

### 5.4 实验 2: 扩展信源-信道模型(XSCM)

我们利用 CN GIGA 语料库和 NIL 语料库进行扩展信源-信道模型的参数估计,分别采用 0.2、0.4、0.6、0.8 和语料库评估值作为平滑算子,并分别处理四个测试集。实验结果如表 6、表 7 所示。

### 5.6 讨论 2: 平滑技术性能对比

我们接下来对几种平滑技术进行对比,图 2 给出了在四个测试集上的 F-1 指数对比曲线。我们发现以语料库评估值为平滑算子时 XSCM 取得最好效果,即 F-1 指数最高为 90.7%。在固定平滑算子中,当  $sf = 0.4$  时,XSCM 效果略好于  $sf = 0.6$ ,但超过  $sf = 0.8$  约 2%。总起来看,若采用固定平滑算子,取值在 0.4 和 0.6 之间某值时,XSCM 性能达到最高。

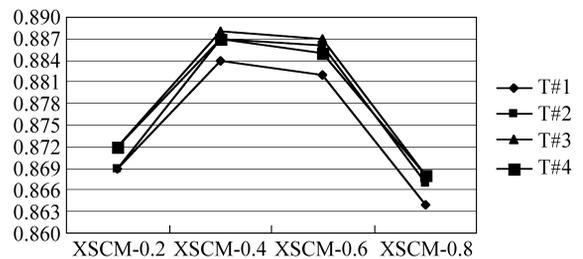


图 2 各种平滑技术对比曲线。

### 5.7 讨论 3: 处理动态聊天文本的健壮性

各种方法在处理动态聊天文本的健壮性可从图 1 看到明显对比。我们发现,使用各种平滑算子的 XSCM 方法都取得了相对平稳的性能。这是因为 XSCM 使用了语音映射模型,该模型在动态网络聊天语言中保持了相对的稳定性。尽管四个测试集的时间戳距离 NIL 语料库的越来越远,XSCM 方法仍能利用稳定的语音映射对动态网络聊天语言进行有效处

理。而 SCM 方法不能适应网络聊天语言的变化,导致其性能急剧下降。这一实验结果有力地证实了语音映射模型在网络聊天语言处理中的重要意义。

## 5.8 错误分析

本部分我们给出两类典型错误,并分析导致错误的主要原因。

### 错误-1: 歧义网络聊天词汇

#### 例 1 我还是 8 米

例 1 中, XSCM 方法没有找到网络聊天词汇,而正确的答案是“我还是不明”。这是由于网络聊天词汇“8”和“米”都包含歧义,当“8”出现在“米”前面时,“8”被识别成数字,而“米”被识别成度量单位。这时,若不通过上下文,很难发现这两个网络聊天词汇。在我们的实验中有 93 个类似错误,这样的错误只有通过基于上下文的话语分析才能得到有效解决。

### 错误-2: 非语音影射网络聊天词汇

#### 例 2 忧虑 ing

XSCM 方法无法识别例 2 中的“ing”,而正确的答案应该是“(正在)忧虑”。这是因为网络聊天词汇“ing”并非产生于语音影射,而是来自英文的现在时态表示。统计发现,大约有 1% 的网络聊天词汇不是通过语音影射创造的,例如表情图标(emoji)就是典型的一种。幸运的是,这些网络聊天词汇通过基于词典的方法就可以处理。因此在实用系统中,我们另外开发一个模块,专门用于处理非语音影射网络聊天词汇。

## 6 相关工作

中文网络聊天语言从 2005 年开始受到自然语言处理研究人员的重视。夏云庆等<sup>[23]</sup>在“NIL Is Not Nothing”项目中对网络聊天语言的奇异性进行了分析和归纳,在小规模网络聊天语言语料库的基础上,设计实现了模式匹配、最大熵和支持向量机方法,以 2004 年 12 月、2005 年 1 月和 2 月的网络聊天语言为训练文本,在处理 2005 年 3 月的网络聊天语言文本时,取得了 87.1% 的 F-1 指数。但是,这些方法在处理更新的网络聊天语言文本时,性能急剧下降。我们认为原因有二:一、现有网络聊天语言语料库规模不足,数据稀疏问题严重,这些方法过度适应网络聊天语言语料库;二、网络聊天语言变化较快,即使有了大规模网络聊天语言语料库,也不能

有效解决动态性问题。

为了建立相当规模的网络聊天语言语料库,夏云庆等<sup>[24]</sup>利用半年时间扩大了 NIL 语料库的规模。为网络聊天语言处理研究提供了更多训练语料。为了解决网络聊天语言动态性问题,夏云庆等<sup>[22]</sup>引入标准语言语料库,利用错误驱动方法,通过计算可信度,来判别输入文本中的奇异网络聊天语言。实验证明,这种方法对动态网络聊天语言文本具有较好的适应性,也取得了同现有最好方法接近的网络聊天语言识别性能。这个方法的问题在于错误驱动机制无法实现对所识别的网络聊天语言进行转换。但是这一实践给我们的宝贵启发是,标准语言语料库对网络聊天语言处理,具有不可忽视的意义。正是从标准语言语料库的相对稳定性,我们发现了语音映射模型。

## 7 结论

本文借助真实网络聊天语言文本,对网络聊天语言的奇异性 and 动态性进行详细分析和归纳,并设计了面向解决奇异性 and 动态性问题的网络聊天语言文本识别与转换方法。我们先以网络聊天语言语料库为基础建立网络聊天语言模型和语言转换模型,通过信源-信道模型实现网络聊天语言向标准语言的转换。但该方法过于依赖网络聊天语言语料库,虽然能较好解决奇异性问题,但不能处理动态性问题。因此,我们进而以标准汉语语料库为基础建立文字语音映射模型,对信源-信道模型进行改进,最终有效解决了网络聊天语言的动态性问题。实验证明,扩展信源-信道模型在引入语音映射模型以后,不但处理网络聊天语言奇异性的能力提高了,还实现了动态网络聊天语言的健壮处理。我们还对解决数据稀疏问题的平滑技术进行了评测,结论是,以语料库评估值为平滑算子时,XSCM 取得了最好效果。

限于现有的网络聊天语言语料库的规模,我们目前还无法完成如下两个工作:一,既然标准语言语料库被引入网络聊天语言处理技术,那么我们将面对如下几个问题:聊天语料库的最小规模多大,才能获得一致的满意性能?标准语言语料库的规模是不是越大越好?当标准语言语料库在规模上同网络聊天语言语料库实现多大的比率时,能够得到最好的训练效果?回答这些问题需要相当规模的网络聊天语言语料库,是我们目前所无法完成的。二,尽管语音映射模型引入后,动态性问题能够得到解决,但

仍然不能忽略网络聊天语言语料库规模提高对网络聊天语言处理的意义。另外,在语音映射模型的假设之外还有 1% 的网络聊天语言需要特殊处理,那么我们会问:大致需要多久以后,XSCM 方法应该在新的网络聊天语言语料库上重新训练一次,才会保持良好的处理性能?这个工作也离不开相当规模的网络聊天语言语料库。这两类问题将在我们未来工作中得到阐述。

### 参考文献:

- [1] 郭良. 05 年中国 5 城市互联网使用现状及影响调查报告[EB]. 社科院社会发展研究中心, 2005.
- [2] 马静. 语言学视野中的网络语言[J]. 西北工业大学学报, 2002, 22(3): 52-56.
- [3] 李雪华. 网络语言初探[J]. 广西社会科学, 2004, (3): 154-155.
- [4] 梁书杰. 对网络语言规范的探讨[J]. 高教论坛, 2005, (6): 191-193.
- [5] 袁星新. 试论网络语言的基本特点[J]. 语言研究, 2005, (12): 20-23.
- [6] 祁伟. 试论社会流行语和网络语言[J]. 语言与翻译, 2002, (3): 18-22.
- [7] 李润生. 网络词汇的造词法探析[J]. 江西教育学院学报, 2003, 24(2): 47-49.
- [8] 李梅. 谈网络语言的语词类型、特点及规范[J]. 语言研究, 2004, (3): 48-50.
- [9] 郭笃凌, 郝怀芳. 网络语言的类型、特点及其语用学意义[J]. 语言应用研究, 2006, (3): 65-67.
- [10] 王登文, 吴晓云. 英汉网络语言语用探析[J]. 外语研究, 2006, (9): 177-178.
- [11] 陈向红, 黎昌抱. 网络聊天中表情达意的非规范手段研究[J]. 广西社会科学, 2006, (3): 190-193.
- [12] 冯念, 冯广艺. 网络词语的谐音及规范问题[J]. 河南师范学院学报, 2005, (1): 138-139.
- [13] 王鸿雁. 汉语网络语言变体探析[J]. 社科纵横, 2005, 20(2): 156-158.
- [14] 李少丹. 谈网络语言的变异现象[J]. 四川理工学院院报, 2006, 21(4): 102-104.
- [15] 赵丽萍. 谈网络语言中的词汇变异现象[J]. 应用语言研究, 2006, (7): 76.
- [16] 李艳, 韩金龙. IRC-聊天室非语言交际研究[J]. 外语电化教学, 2003, (94): 7-11.
- [17] 周卫红. 论网络语言的后现代文化内涵[J]. 哲学研究晋阳学刊, 2006, (2): 76-79.
- [18] Gianforte, G.. 2003. From Call Center to Contact Center: How to Successfully Blend Phone, Email, Web and Chat to Deliver Great Service and Slash Costs[R]. RightNow Technologies.
- [19] Heard-White, M., Gunter Saunders and Anita Pincas. 2004. Report into the use of CHAT in education. Final report for project of Effective use of CHAT in Online Learning[R]. Institute of Education, University of London.
- [20] Finkelhor, D., K. J. Mitchell, and J. Wolak. Online Victimization: A Report on the Nation's Youth[R]. Alexandria, Virginia: National Center for Missing & Exploited Children, 2000, page ix.
- [21] McCullagh, D.. 2004. Security officials to spy on chat rooms. News provided by CNET Networks[R]. November 24, 2004.
- [22] Xia, Y. and K.-F. Wong. 2006a. Anomaly Detecting within Dynamic Chinese Chat Text[A]. In: Proc. of EAACL '06 NEW TEXT workshop[C].
- [23] Xia, Y., K.-F. Wong and W. Gao. 2005. NIL is not Nothing: Recognition of Chinese Network Informal Language Expressions[A]. 4th SIGHAN Workshop at IJCNLP '05[C]: 95-102.
- [24] Xia, Y., K.-F. Wong and W. Li. 2006b. Constructing A Chinese Chat Text Corpus with A Two-Stage Incremental Annotation Approach[A]. In: Proc. of LREC 2006[C].
- [25] Xia, Y., K.-F. Wong and W. Li. 2006c. A Phonetic-Based Approach to Chinese Chat Text Normalization[A]. In: Proc. of ACL '06[C]. 993-1000.
- [26] Zhang, Z., H. Yu, D. Xiong and Q. Liu. HMM-based Chinese Lexical Analyzer ICTCLAS[A]. SIGHAN '03 within ACL '03[C]. 2003. 184-187.
- [27] Epstein, M. E.. 1996. Statistical Source Channel Models for Natural Language Understanding[D]. PhD Thesis. New York University.
- [28] Graf, D., Chen, K., Kong, J., Maeda, K.: Chinese Gigaword Second Edition[DB]. LDC Catalog Number LDC2005T14 (2005).