文章编号：1003-0077（2017）00-0000-00

汉语委婉语语言资源建设

张辰麟[1] 王明文[1] 谭亦鸣[2] 肖文艳[1]
（1.江西师范大学 计算机信息工程学院，江西省 南昌市 330022；
2.东南大学 网络空间安全学院，江苏省 南京市 210000）

摘要：委婉语是语言交流中的不可或缺的交际手段，委婉语研究一直是语言学界的热门话题之一，但在自然语言处理领域，尚未有委婉语相关研究。该文借助现有纸质词典，基于语料库检索和专家人工判别的方式，初步构建了规模为 63000 余条语料的汉语委婉语语言资源；并根据自然语言处理的相关任务需求，结合词典释义对委婉语进行分类。该文提出了利用同类委婉语的上下文语境辅助进行标注的方法。经过实验，对简单语义的委婉语义判别的准确率达 89.71%，对语义复杂的兼类委婉语判别准确率达 74.65%，初步验证了利用计算机辅助人工标注构建委婉语语言资源的可能。

关键词：委婉语；语义辨析；语言资源构建

中图分类号：TP391 **文献标识码：**A

Construction of Chinese Euphemism Language Resources

Zhang Chenlin[1], Wang Mingwen[1], Tan Yiming[2], Xiao Wenyan[1]

(1. School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China;
2. College of Cyberspace Security, Southeast University, Nanjing, Jiangsu 210000, China)

Abstract : Euphemism is an indispensable method of language communication. It has always been one of the hottest issue in linguistics. However, in natural language processing, euphemism has not been discussed yet. In this paper, a language resource of euphemism (about 63,000 sentences) is manually judged and construct, referring to existing paper dictionaries. According to the dictionaries' definition and the requirements of the related natural language processing work, euphemisms are classified at the semantic level. With the collected corpus and classification, we attempted to identify polysemous euphemisms automatically and achieved a fairly accuracy of 89.71% for simple euphemisms and 74.65% for complex ones. This preliminary verifies the possibility of automatically constructing language resources.

Key words: Euphemism; Semantic Discrimination; Language Resource Construction

0 引言

委婉语（Euphemism）是一种普遍的语言现象，在日常交流中不可或缺。在汉语中，“委婉”又称“婉转”、“婉曲”，其定义为：“不直说本意，

收稿收稿日期：201*-**-**；**定稿日期**：201*-**-**
基金项目：基金名（基金号）；基金名(基金号)

而是采用一种委婉曲折的话来表达”。在日常交流过程中，人们经常会遇到不便明说的事物，如死亡、年龄、收入等，需要借用一些“与此事物相关的同义语句婉转曲折地表达”，或者用“与本意相关或相类似的话来代替”，于是便形成了委婉语。委婉语的定义有广义狭义之分。狭义的委婉语主要指包含委婉义的词或者短语，是词汇层面的概念。广义的委婉语又称委婉表达，除了使用词汇手段之外，还利用语音（如轻读、省音）、语法（如否定、省略）、语用手段实现委婉的目的。一般不特殊说明的情况下，委婉语研究指的均是狭义的、词汇范畴下的委婉词语。

自上世纪八十年代以来，委婉语现象逐渐被我国语言学家所重视。束定芳[1] [2]、伍铁平[3]等对委婉语的定义、特点、分类、构造原则和手段、认知理据等问题进行了综合性地阐释[4][5]。谌莉文[6] [7]、邵军航[8] [9]等结合认知语言学的认知空间与范畴理论等，对委婉语中的认知过程和内涵的隐喻进行了剖析[10] [11]。田九胜[12]、梁红梅[13]等从语用学的角度，结合合作原则、自我保护原则等对委婉语现象进行了探讨[14][15]。由于委婉语牵扯到复杂的文化背景知识，也有学者从不同性别、不同语言社团等方面对委婉语进行研究[16]。委婉语的出现，与语言禁忌（Taboo）密不可分，因此一些学者也致力于专门领域的委婉语研究，如宗教[17]、疾病[18]、死亡[19]、外交辞令[20]等的研究。随着委婉语研究的深入，委婉语词典也不断涌现，主要包括国家语言文字工作委员会编写的《通用委婉语词典》[21]、《汉语委婉语词典》[22]、《实用委婉语词典》[23]、《委婉语应用词典》[24]等。

1 委婉语与自然语言处理

从自然语言处理的角度来看，委婉语也同样非常值得研究。

语言禁忌是形成委婉语最主要的原因之一。因此很多委婉语源自人们对事物的避讳，而这种避讳往往与人类的情感相关，如与“死亡”相关的委婉语“百岁”、“不测”、“归西”等往往是与“悲伤”情感有关；与“灾难”相关的委婉语，如“硝烟”、“变故”、“闪失”等往往是由“惊慌”、“恐惧”情感而诞生的语言禁忌；而与“犯罪和惩罚”有关的委婉语如“高墙”、“打秋风”等则往往出现在有厌恶感情的句子当中。委婉语是句子情感的隐性“风向标”，对于情感词典或情感基元建设，辅助情感分类有很大的作用。

“隐喻”是当前自然语言处理热门话题，同时也是构成委婉语的最重要的方式之一。绝大部分委婉语都是通过隐喻的方式达到“迂回曲折地表达意思”的目的。譬如用“拔出萝卜带出泥”来在调查某人或者某事时牵连出其他的人或事；用“吃豆腐”来隐喻占他人便宜；用“分一杯羹”来隐喻参与瓜分一部分利益，用“拈花惹草”来隐喻行为不检点等等。因此，委婉语的研究也可以为“隐喻”挖掘与相关的计算研究提供丰富的参考。

委婉语的语义往往是隐晦的，难以从字面上或者词语已有的义项中得到，因此委婉语的相关研究对自然语言理解也有一定的帮助作用。委婉语依托于语用范畴，是说话人为了遵循会话原则和礼貌原则，刻意对话语的一种改良，主要用于在交际过程中让听话者更容易接受，因此，用委婉语对生成的自然语言句子进行有条件的替换和后编辑，可以使聊天系统的语言显得更生动，更有人情味。

然而从语言学界对委婉语既往研究工作来看，委婉语的研究大多数与语义、语用密不可分。委婉语构成原因和使用的环境复杂多样，语义往往具有隐晦性，在缺乏上下文语境与相关知识的情况下难以对其进行识别或者加以利用，而语用层面则主要依赖于交际的语境、背景知识、身份和人物关系等信息。因此目前自然语言处理领域尚未有委婉语的相关研究。

本文以现有委婉语词典为基础，借助语料库检索手段，旨在为汉语委婉语构建大量的语料资源，并对检索的语料进行人工判别与标注。受汪梦翔[25]等对纸质词典释义的挖掘工作及《义务教育常用词表（草案）》（2019）[26]中义类码的启发，本文还将根据委婉语词典中对委婉语词汇的释义，参照自然语言处理相关的任务需要，对委婉语进行语义层面上的分类，并使用相关技术手段，尝试验证计算机辅助人工标注的可能性，以求获得更大规模的语料资源。

2 委婉语语料的搜集与选取

通过对比四本委婉语词典中的委婉语词汇，我们选择了代表性较强、覆盖面较广、释义编写简明规范的《通用委婉语词典》作为主要参考词典。该词典由国家语言文字工作委员会主编，包括 2064 个委婉语词，每一个委婉语词在词典中有 1-2 个例句作为参考。如：

【放血】迫使某人拿出钱来。【例】我呢，投了赞成票，当然也要放放血。（柳建伟《突出重围》）

【大事】婚姻的事。【例】他……怕误了自己的大事——他不能随便的交女朋友弄坏了名誉。（老舍《文博士》）

【老了】年纪大的人死亡。【例】“祥林嫂？怎么了？”我又赶紧的问。“老了”（鲁迅《祝福》）|原来这铁槛寺原是宁荣二公当日修造，现今还是有香火地亩布施，以备京中老了人口，在此便宜寄故。（[清]曹雪芹，高鹗《红楼梦》）

在纸质词典中，委婉语的例句多数来源于文学作品、白话文或者传统典籍中的句子，这是受词典学的编撰范式决定的。然而这样的例句作为语料不宜进行自动分词，其文体、主题各异，语境上下文信息也较为分散，不适合作为自然语言处理的对象。因此我们以《人民日报》（2014，约 2400 万字）作为语料库，利用检索的方式重新构建委婉语的语料。

我们选择了 Jieba 分词系统对《人民日报》（2014）所有的语料进行分词，为了避免分词系统无法识别未登录的委婉语词或者短语，造成无法成功抽取句子的情况，我们对委婉语词表也同样进行了分词，以方便和分词之后的语料进行匹配。

通过检索，从语料库中成功抽取了包含 923 个委婉语词汇，共 63159 个句子。

然而并不是所有委婉语词汇都只有委婉语这一个义项，根据既往对委婉语构成方法的研究以及语言的经济原则，大部分委婉语往往依附于语言系统中已有的词汇之上，通过添加新的委婉义项构成的，形成委婉语。因此大多数委婉语词都是多义词。例如：

【方便】除了指“省事的、便利的”之外，还能够婉指“上厕所”。

【离去】除了指“离开”，还能够婉指“人的死亡”。

【八卦】除了指《周易》中记载的一套符号系统之外，还能够婉指荒诞低俗的传闻或明星的绯闻。

这些“多义词”在不同词性，不同上下文语境中表现出不同的语义，判别条件较为复杂，目前尚无法通过机器来实现，在缺少前人研究基础的情况下，基础语言资源的前期建设只能使用人工来进行判断。

我们对 63159 个句子中出现的委婉语词是否属于“委婉语”进行人工判别。判别过程采取投票的方式，由五位汉语言文学相关专业的研究生作为专家进行的投票。当投票结果形成 3:2 或者 2:3 时，语料返回专家手中并进行讨论，如有较大争议则详细标注争议原因并进行舍弃。当遇到特殊情况，如句子分词有误、委婉语本身被再次用于隐喻、上下文语境不明、委婉语出现在书名等命名实体当中等，则直接标注问题预料问题产生原因并予以舍弃。通过长期多轮的人工筛选，一共得到了 11624 个正例（即目标词汇以委婉语义出现），50823 个负例（即目标词不以委婉语义出现），并舍弃了 712 个有问题的句子。示例如下：

【下海】婉指改行或开始进入妓女行业。

【正例】①/“/这么/多年/，/就/只/看到/有/老师/主动/辞职/下海/，/还/没/看到/学校/主动/解聘/或/辞聘/老师/的/。（备注：委婉义 1，即改行换业）

②/近年来/“/美/魔女/”/风潮/大为/盛行/，/备受/关注/的/43/岁/美/女主播/桐/嶋/永久/子/决定/挑战/自我/，/宣布/决定/下海/成为/一名/AV/女优/，/理由/竟然/是/“/出自于/生理需求/”/。（备注：委婉义 2，即进入色情行业义）

【负例】①/该/负责人/称/，/球形/棕/囊/藻/繁殖/太/多/，/会/影响/水质/，/对/海洋/养殖业/造成/影响/，/或/对/人体/产生/危害/，/建议/市民/不宜/下海/游泳/。（备注：指下到海里）

②/除了/睡觉/、/下海/作业/，/剩余/的/

时间/只能/看看书/。（备注：指下到海里）

【舍弃】/毛泽东/知道/后/笑/着/对/柯/说/，/上海/这个/地方/原/是/海滩/渔村/，/既有/上海/村/，/也/有/下海/村/！（备注：地名）

一些委婉语词汇在实际的语言中又再次被作为隐喻使用，这样的委婉语往往已经失去了本来的委婉义，表现出新的语义，例如：

①/按说/，/安倍/“/二进宫/”/并/于/去年/7/月/领导/自民党/和/公明党/联盟/在/参议院/选举/中/获得/多数/席位/，/结束/了/日本/6/年/“/扭曲/国会/”/局面/。

②/然而/，/联网/大限/已/过/，/500/个/城市/的/住房/信息/联网/并未/完成/。

“二进宫”本身就是隐喻，婉指“再次入狱”，句①又再次将这个委婉语进行隐喻，表达一种对安倍再次从政的贬低；而“大限”作为委婉语，义为“人的死亡”，这里暗指“联网”这件事情已经结束或者过时，这样的例子委婉语虽然出现，但已经不是单纯的委婉义，因此我们也予以舍弃。

3 委婉语的语义分类

在语言学领域已有不少对委婉语分类的研究，主流分类方法有两种，一种是根据委婉语的语义进行分类，另一种是根据委婉语的构成方法进行分类。鉴于现阶段自然语言处理领域尚无委婉语相关研究，而从语义的层面上进行分类，对自然语言处理中的其他研究更有帮助，我们对检索到的委婉语从语义和感情色彩方面进行了分类。

3.1 结合语义、语境的委婉语分类

结合语言学领域已有的研究成果，以及隐喻、情感分析等自然语言处理任务的要求，我们将委婉语分类如图 1 所示：

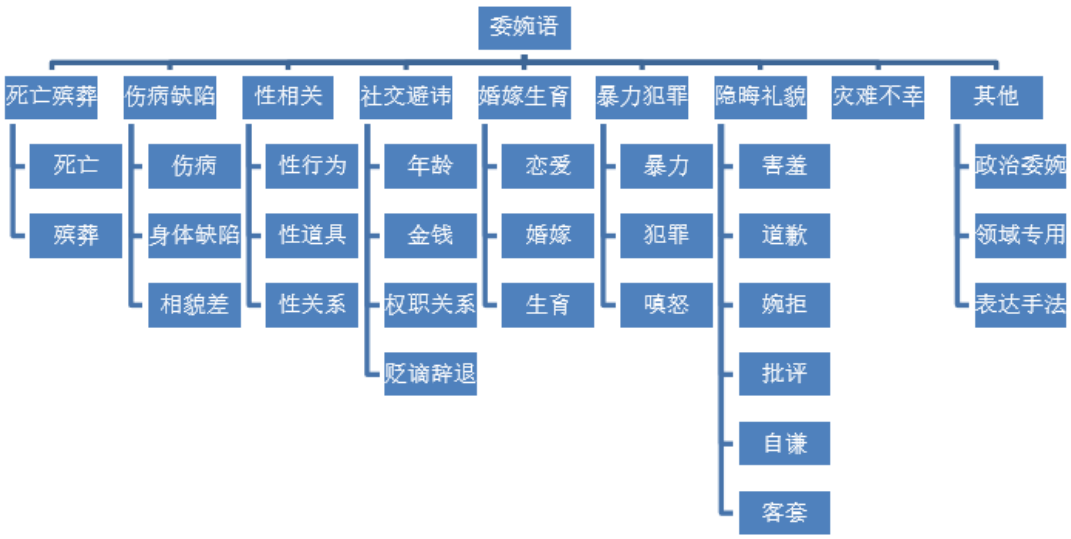

图 1 委婉语的语义分类

表 1 展示了委婉语每个分类的若干例子：

表 1 委婉语分类示例

总分类	次分类	示例	总分类	次分类	示例
死亡殡葬	死亡	百年、入土	暴力犯罪	暴力	动手、回敬
	殡葬相关	西天、送行		犯罪	白粉、强暴
伤病缺陷	伤病	不快、挂彩		嗔怒	爆粗、变脸
	身体缺陷	耳背、失智	隐晦礼貌	害羞	入月、例假

	相貌差	发福、肉感		道歉	不好意思
性相关	性行为	同居、上床		婉拒	敬谢不敏
	性道具	A 片、成人用品		批评	不敢苟同、不知所云
	性关系	出轨、有染		自谦	爱莫能助
社交避讳	年龄	夕阳红、高龄		客套	打扰、失陪
	金钱	份子、囊中羞涩	灾难不幸	无再分类	不测、找不着北
	权职关系	拉关系、有背景	其他	政治委婉	第三世界、超级大国
	贬谪辞退	左迁、另请高明		领域专用	四季青、路路通
婚嫁生育	恋爱	对象、幽会		表达手法	莫非、岂不是
	婚嫁	内助、终身大事			
	生育	大肚子、有喜			

3.2 委外语分类理据

我们对于委婉语的分类主要基于以往语言学家对委婉语在语义层面的分类[1-4] [17-20] [21] [22]，并尽量结合相关的自然语言处理任务目的进行标注。

参照大连理工大学的情感词汇本体库[27]，各个分类委婉语的情感倾向标注如表 2：

表 2 委婉语分类示例

总分类	次分类	情感倾向	总分类	次分类	情感倾向
死亡殡葬	死亡	悲伤、失望（如自杀类）	暴力犯罪	暴力	憎恶
	殡葬相关	悲伤、思		犯罪	憎恶
伤病缺陷	伤病	悲伤、恐惧		嗔怒	憎恶、贬责
	身体缺陷	悲伤、恐惧	隐晦礼貌	害羞	羞
	相貌差	赞扬（把不好的说成好）		道歉	尊敬（受话人）
性相关	性行为	羞		婉拒	部分含尊敬（受话人）
	性道具	羞		批评	贬责
	性关系	憎恶、妒忌*		自谦	尊敬（受话人）
社交避讳	年龄	尊敬		客套	尊敬（受话人）
	金钱	褒贬兼有	灾难不幸	无再分类	慌、恐惧
	权职关系	褒贬兼有	其他	政治委婉	不分类
	贬谪辞退	失望、贬责等		领域专用	不分类
婚嫁生育	恋爱	快乐、安乐		表达手法	不分类
	婚嫁	快乐、安乐、祝愿			
	生育	快乐、安乐、祝愿			*：该类均为负面情感

我们根据以往语言学家对委婉语隐喻的研究[6-9]，将由隐喻，或与隐喻相关的表达手法构成的委婉语做了详细的标注，如：

【吃醋】发酸，比喻嫉妒。

【下身】用整体指代部分，提喻，婉指隐私部位。

“其他”分类中包括了政治委婉语，即“外交辞令”。“领域专用”主要包括方言习语领域的习惯叫法，如蛇常被称为“长虫”；中医药学中对药材的一些特殊叫法如“龙骨”、“守宫”、“龙须”等，根据委婉语词典，它们虽然也属于委婉语，但是由于数量较少，适用范围窄，且与其他语义分类没有太多联系，我们将其统一纳入“领域专用”委婉语，一并归入了“其他”类。

委婉手法是一个特殊分类，这些委婉语虽然也是词汇，但是从功能上来讲却是运用句法或者

语用手段来实现委婉表达，因此严格来讲不属于词汇层面的委婉语。如：

【某】作定语，修饰表示国家、机构、地点、时间、人物、事物、行为等词语，确有所指却不明说或代替人名中不可说的字。

例句：/12/月/31/日/，/记者/走进/湖南/某/省直机关/的/办公楼/，/看到/只有/一楼/贴/有/“/请勿吸烟/”/的/标志/。

【岂敢】哪里敢，婉言不敢。

例句：/假如/是/公卿/子弟/通晓/文墨/的/，/南衙/又/岂敢/随便/侮辱/他/”/，/于是/太平公主/就/趁机/把/自己/的/情人/推荐/给/武则天/。

虽然这类委婉语词看似词汇，但是常与一些句法、构式、隐式语义以及指代相关，因此我们也对其进行了标注，并判别了正负例，以备后续进行广义委婉语相关研究时使用。

一些委婉语比较特殊，其本身具有两个或者两个以上委婉语语义，如：

【红白喜事】指婚丧，男女结婚是喜事，高寿的人去世是喜丧，统称红白喜事。其就语义就同时包含了“死亡”与“婚嫁”两个类别。

【长短】指意外的灾祸，婉指人遇到意外而导致灾祸或者死亡。其语义既可能属于“灾难不幸”也可能属于“死亡”。

对于同时有两个委婉义的委婉语词汇我们会设立两个项目，如“长短$_1$”、“长短$_2$”具体标注每一条语料具体属于哪一个语义，并将语料归入相应类别中。

一些委婉语语义本身就包括了两个语义分类的内容，且不可分割，如：

【强暴】本身除了属于“性行为”、也属于“犯罪”。

【殉情】 婉指因恋爱受挫感到绝望而自杀，其语义同时包含“死亡”和“恋爱婚嫁”。

如出现上述情况，我们将该词和对应语料复制，并同时放入两个语义分类的语料当中。

4 计算机辅助人工标注

4.1 基于 TF-IDF 的计算机辅助人工标注

委婉语多数为多义词，因此判定句子中出现的委婉语是否显示为委婉语语义便成为了一个难题。使用人工进行标注耗时费力代价高昂，并且会涉及标注一致性的问题，因此，为了获得大量的语料以便进行自然语言处理的相关研究，需要考虑如何进行计算机辅助人工标注。

在自然语言处理领域中，通常的思路是使用大量语料组成上下文语境对语义进行消歧。但是由于委婉语词汇本身数量众多，且使用频率多呈现出不均衡的现象，以我们检索到的 923 个委婉语为例，有 475 个在语料中出现次数在 5 以下，占总词汇量的一半以上，这使想要获得大量特定委婉语的语料变得更加困难。

由于委婉语在语义类型上较为有限，且同一类委婉语在语义上往往也比较单一，因此，我们假设同一最底层分类下的所有委婉语共享相同或相似的上下文语境，即同一类委婉语词上下文共现的词语应该相同或相关。基于此假设，我们将同一类的委婉语抽出其中一个多义委婉语及其句子作为验证集，其他的委婉语的句子作为训练集，由于现阶段人工标注的句子数量依然不能满足使用词向量或者神经网络的方法来进行分类的要求，因此我们通过 TF-IDF 来进行简单的辅助标注。

4.2 简单委婉语的自动判别与标注

我们首先从语义比较单一，分类比较明显的委婉语入手，将“死亡殡葬”语义分类下“死亡”这一小类的所有委婉语为对象进行了实验。抽取其中一个多义委婉语“离去”作为测试集，其他“死亡”类委婉语的语料作为训练集。训练集语料包含了 103 个婉指“死亡”的委婉语，共 11401 条句子，其中正例为 1217 条语料，负例为 10159 条语料，另有 25 条因分词错误、委婉语处于命名实体中等原因被排除。测试集“离去”共有 70 条语料，其中其中实际可用语料为 68 条，正例语料 13 条，负例语料 55 条，另有两条语料因长度太短，没有足够上下文语境判别“离去”是否为委婉语，因此予以舍弃。

为了使正负例语料数量接近，我们采取随机抽样的方式，在“死亡”类委婉语语料的 10159 个负例中抽取与正例数量相等的 1217 条负例语料。我们将正例与负例各 1217 条语料里出现的词分别使用 TF-IDF 进行倒排索引，由于语料较

少，使用目标词前后 n 个词作为语境几乎无法获得结果，我们以全句作为目标委婉语的上下文，TF-IDF 公式如式（1）所示：

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

经过 TF-IDF 权重排序后，我们使用哈工大的停用词表对常用词进行停用，并令正负例词表互为停用词表进行第二轮停用，去除二者中相同的词语，以屏蔽由于主题、语体的原因而出现的高频词或相关实体，如人名、地名、“本报讯”等。

由于我们将语料的条数作为 TF-IDF 的文档数进行统计，词频较小的词有可能因句子较短产生较高的权重，形成噪音，因此，我们用分别以词频 0、2、4、6、8、10 为阈值，过滤词频较小的词。得到了 6 个词表 TF-0、TF-2、TF-4、TF-6、TF-8、TF-10。根据 TF-IDF 权重由高到低排列后，又分别保留词表 20%、40%、60%、80%、100%的词，从而生成了 30 个词表。我们为 30 个词表中的词进行赋值，正例的词赋值为 1，负例的词赋值为 -1，并带入验证集“离去”的语料中进行计算，如果结果≥1，则判定该句子中的“离去”为“死亡”义，是委婉语。如结果≤0，则判断该句子中的“离去”不是委婉语。我们将计算的结果与人工标注的结果进行对比，如表 3：

表3 委婉语“离去”的计算机辅助人工标注准确率（正例13负例55）

	TF≥0			TF≥2			TF≥4		
	正例准确	负例错误	准确率	正例准确	负例错误	准确率	正例准确	负例错误	准确率
词表20%	9	7	0.8382	4	5	0.7941	0	5	0.7353
词表40%	11	8	0.8529	7	8	0.7941	5	3	0.8382
词表60%	13	9	0.8676	8	10	0.7794	6	5	0.8235
词表80%	13	10	0.8529	10	10	0.8088	8	6	0.8382
词表100%	13	10	0.8529	11	9	0.8382	8	6	0.8382
	TF≥6			TF≥8			TF≥10		
	正例准确	负例错误	准确率	正例准确	负例错误	准确率	正例准确	负例错误	准确率
词表20%	1	3	0.7794	1	6	0.7353	3	7	0.7500
词表40%	6	4	0.8382	6	5	0.8235	8	8	0.8088
词表60%	7	5	0.8382	8	6	0.8382	10	9	0.8235
词表80%	12	6	0.8971	11	6	0.8824	12	8	0.8676
词表100%	12	8	0.8676	11	7	0.8676	12	8	0.8676

从判断结果中可以看出，对于语义单一分类明显的简单委婉语而言，仅用 TF-IDF 构建权重词表的方式给句子打分，即可最高达到约 90%的分类准确率，但增加语料的数量依然可以显著提高准确率，判断的准确率依旧依赖于语料的规模。通过该步骤，初步证明了简单的委婉语可以单纯依赖上下文语境用同类的其他委婉语判断出来。委婉语的上下文，能够极大地决定其在句子中是否为委婉语义。

由于我们只对正例是否为正例进行判断，即证明一个委婉语词出现在句子中是委婉语，而无法穷举其不是委婉语时的所有语境。因此改实验打分结果≤0 时，均判断为不是委婉语，即当正例、负例的上下文语境词都不能被检索到的情况下，也判断为负例。这等同于将“不知道”也判断为“不是”，因此对于负例较多的委婉语词而言，判断的准确率均较高，因此，该实验只能初步证明根据同类委婉语上下文来进行判断的可行性。

4.3 复杂委婉语的自动判别与标注

如前文 3.2 所示，一些委婉语在语义上不止包含了一类委婉语的语义，同时还兼有另一类委婉语的语义。委婉语“八卦”，在词典上的注释为：“荒诞低俗、没有根据的消息、传闻，通常是明星或者名人的隐私绯闻”，它的释义核心为“名人的绯闻”，我们把它归为“性关系”这一类，但是该词在日常使用的过程中，其语义逐渐发生变化，现也可以指代明星明人正常的“婚恋交往”新闻。因此它同时还兼类了“恋爱”和“婚嫁”类，为了验证这类复杂委婉语是否能够被标

注正确，我们用同样的方法，首先将“八卦”从“性关系”类抽取出来当做验证集，将其他的表示“性关系”的委婉语作为训练集。

在我们的语料中，委婉语“八卦”一共包括72个句子，其中59个正例句子，12个负例句子，有一个句子因为语境不明被舍弃，与4.2中委婉语“离去”不同，该委婉语正例的数量远大于负例。而与“性关系”相关的所有委婉语其语料一共包括5884个句子，其中正例为592个，负例为5245个，舍弃的句子为47个。

表4展示了委婉语通过“性关系”类委婉语来判定“八卦”是否为委婉语的准确率：

表4 委婉语“八卦”的计算机辅助人工标注准确率（正例59负例12）

	TF≥0			TF≥2			TF≥4		
	正例准确	负例错误	准确率	正例准确	负例错误	准确率	正例准确	负例错误	准确率
词表20%	6	0	0.2535	12	0	0.3380	18	0	0.4225
词表40%	15	0	0.3803	20	0	0.4507	22	0	0.4789
词表60%	23	0	0.4930	29	0	0.5775	27	1	0.5352
词表80%	26	0	0.5352	31	1	0.5915	30	2	0.5634
词表100%	30	0	0.5915	35	1	0.6479	34	3	0.6056
	TF≥6			TF≥8			TF≥10		
	正例准确	负例错误	准确率	正例准确	负例错误	准确率	正例准确	负例错误	准确率
词表20%	16	0	0.3944	15	0	0.3803	14	0	0.3662
词表40%	23	1	0.4789	20	0	0.4507	16	0	0.3944
词表60%	26	1	0.5211	21	0	0.4648	22	0	0.4789
词表80%	27	1	0.5352	31	0	0.6056	25	0	0.5211
词表100%	30	1	0.5775	31	0	0.6056	26	0	0.5352

可以看出准确率比较低，这主要是由于“性关系”类的委婉语都为表示“不正当关系”的委婉语，因此不能够完全提供“八卦”做为委婉语的语境。因此我们将“婚嫁”和“恋爱”的委婉语正负例与“性关系”的正负例进行合并，并重复实验过程，准确率见表5：

表5 （加入兼类语料后）委婉语“八卦”的计算机辅助人工标注准确率（正例59负例12）

	TF≥0			TF≥2			TF≥4		
	正例准确	负例错误	准确率	正例准确	负例错误	准确率	正例准确	负例错误	准确率
词表20%	9	0	0.2958	15	0	0.3803	16	0	0.3944
词表40%	19	0	0.4366	24	0	0.5070	27	0	0.5493
词表60%	28	0	0.5634	32	1	0.6056	30	0	0.5915
词表80%	35	0	0.6620	36	1	0.6620	31	0	0.6056
词表100%	37	1	0.6761	41	0	0.7465	35	0	0.6620
	TF≥6			TF≥8			TF≥10		
	正例准确	负例错误	准确率	正例准确	负例错误	准确率	正例准确	负例错误	准确率
词表20%	19	0	0.4366	17	0	0.4085	21	0	0.4648
词表40%	24	0	0.5070	21	0	0.4648	22	0	0.4789
词表60%	28	0	0.5634	29	0	0.5775	24	0	0.5070
词表80%	34	0	0.6479	36	0	0.6761	30	0	0.5915
词表100%	37	0	0.6901	38	0	0.7042	33	0	0.6338

从上表可以看到，最高准确率出现在TF取2以上时，词表取100%，正确率约为74.65%。可见对于语义有一定兼类的复杂委婉语而言，需要更丰富的上下文信息才能保证判断的准确率。最高准确率需要词表的使用率100%，可以明显看出目前仅靠人工标注的委婉语语料依然严重不足，尚需进一步进行扩展。

实验一方面验证了运用自然语言处理方法

研究委婉语的可行性，也验证了人工编纂词典的语义进行分类归纳，可以对委婉语的自动标注起到一定的作用。从上表我们也可以明显看到，在词表规模百分比不变的情况下，限制低频词的出现可以一定程度上提高正确率，但增加正负例词表的词语数量投入却可以显著增加判别的准确率，这也证明了现阶段语料不足依然是委婉语研究的最主要问题。

因此对于丰富委婉语语言资源而言，仍需要从语义简单，语境单一，类型单一，负例较多的委婉语进行入手，在扩展一定数量的语料之后，再对复杂委婉语的语料进行扩展和计算机辅助标注。

5 结语

本文结合语言学中的热门研究方向——委婉语，结合现有的纸质词典，初步构建了一定规模的汉语委婉语语言资源。主要包括从《人民日报》语料中检索并大量进行人工标注的委婉语语料，并进行了委婉语的语义和情感层面的分类、标注等。为验证语料的和分类对自然语言处理的作用，我们用 TF-IDF 辅助词频的筛选，尝试对多义委婉语的语义尝试进行计算机辅助人工标注，获得了一定的成果，证明了计算机辅助人工标注委婉语的可行性。在日后的研究工作中，我们将进一步拓展语料的规模，并将一些没有被检索到的委婉语通过检索其他的语料库，至少每条委婉语增补 5 个正例，以便完善整个语言资源。

委婉语早已不是一个新鲜的话题，但是由于其涉及到很多认知、语境、语用、文化背景等相关知识，在自然语言处理方面撼未有人涉及。做好“委婉语”的研究，可以对自然语言处理中的热点如：隐喻研究、情感计算研究、语义消歧、自然语言理解与生成后编辑提供辅助。汉语是高语境文化[28]，在沟通的过程中，信息多数不会被清晰编码并传递出来，汉语的沟通是含蓄的，内敛的，信息传播时常挂靠在语境而非言语内容上，委婉语作为含蓄沟通的主要手段之一，值得我们进一步关注。

参考文献

[1] 束定芳.委婉语新探[J].外国语，1989 第 3 期：28-34.
[2] 束定芳，徐金元.委婉语研究：回顾与前瞻[J].外国语，1995 第 5 期：17-22.
[3] 伍铁平.从委婉语的机制看模糊理论的解释能力[J].外国语，1989 第 3 期：16-22.
[4] 李国南.英语中的委婉语[J].外国语，1989 第 3 期：23-27.
[5] 孔庆成.委婉语言现象的立体透视[J].外国语，1993 第 2 期：26-30.
[6] 谌莉文.概念隐喻与委婉语隐喻意义构建的认知理据[J].外语与外语教学，2006 第 8 期：17-20.
[7] 谌莉文.英汉委婉语跨空间映射认知对比考察[J].外语教学，2007 第 4 期：39-42.
[8] 邵军航，樊葳葳.委婉机制的认知语言学诠释[J].外语研究，2004 第 4 期：20-25.
[9] 邵军航，樊葳葳.也谈委婉语的构造原则[J].山东师大外国语学院学报，2002 第 2 期：32-34.
[10] 王永忠.从语言模糊性看委婉语的交际功能[J].福建外语，2001 第 4 期：27-30.
[11] 王永忠.范畴理论和委婉语的认知理据[J].外国语言文学，2003 第 2 期：3-5.
[12] 田九胜.委婉语的语用分析[J].福建外语，2001 第 2 期：18-21.
[13] 梁红梅.委婉语的语用分析[J].天津外国语学院学报，2000 第 1 期：30-34.
[14] 徐海铭.委婉语的语用研究[J].外语研究，1996 第 3 期：21-24.
[15] 徐莉娜.跨文化交际中的委婉语解读策略[J].外语与外语教学，2002 第 9 期：6-9.
[16] 彭文钊.委婉语——社会文化域的语言映射[J].外国语，1999 第 1 期：66-70.
[17] 李国南.委婉语与宗教[J].福建外语，2000 第 3 期：1-6.
[18] 伍铁平.病从何来？——委婉语一例[J].当代修辞学，1985 第 2 期：51.
[19] 疏志强.死亡代语及其文化意蕴[J].汉语学习，1998 第 5 期：30-33.
[20] 葛新新.政治委婉语：分类、机制及原则[D].吉林大学，2006.
[21] 许嘉璐，袁贵仁，朱景松等.通用委婉语词典[M].北京：语文出版社，2018.
[22] 张拱贵，王聚元等.汉语委婉语词典[M].北京：北京语言文化大学出版社，1996.
[23] 王雅军.实用委婉语词典[M].上海：上海辞书出版社，2005.
[24] 王雅军.委婉语应用辞典[M].上海：上海辞书出版社，2011.
[25] 汪梦翔，饶琪，顾澄等.汉语名词的隐喻知识表示及获取研究[J].中文信息学报，2017 第 6 期：1-9.
[26] 苏新春.义务教育常用词表（草案）[M].北京：商务印书馆，2019.
[27] 徐琳宏，林鸿飞，潘宇，等.情感词汇本体的构造[J].情报学报，2008，27(2)：180-185.
[28] 赵胤伶，曾绪.高语境文化与低语境文化中的交际差异比较[J].西南科技大学学报（哲学社科版），2009 第 2

期：45-49.