维、哈、柯语舆情标注语料库的构建研究

陈昊[1]，卡哈尔江·阿比的热西提[1]，艾山·吾买尔[1]，吐尔根·依布拉音[1]

（1 新疆大学信息科学与工程学院 新疆 乌鲁木齐 830046）

摘要：针对少数民族舆情标注语料稀缺以及语料标注仅涉及简单的人类智能的事实，提出了一种基于众包的舆情语料标注方法。在制定了维吾尔语、哈萨克语和柯尔克孜语情感标注规范的基础上，建立三层架构的语料标注平台，将语料标注通过互联网推送到大众面前，实现了大众参与舆情语料标注，此外，还提出了纠错机制和质量控制策略用以保证标注质量。维吾尔语、哈萨克语、柯尔克孜语舆情标注语料库的构建能够为少数民族舆情研究提供有力的资源支持。

关键词：众包；语料标注；舆情

中图分类号：TP391　　**文献标识码**：A

Construction of Uygur, Kazak and Kirgiz Public Opinion Tagging Corpus

Chen Hao[1], Kahaerjiang Abiderexiti[1], Aishan Wumaier[1], Tuergen Yibulayin[1]

(School of information science and engineering, Urumqi, Xinjiang 830046,China)

Abstract: Aiming at the lack of public opinion tagging corpus in national minority languages and the fact that corpus annotation only involves a simple human intelligence, this paper proposes a public opinion corpus annotation method based on the crowdsourcing. At first, we formulate the Uygur, Kazak, Kirgiz language emotion tagging specification, then we established a three-layer architecture corpus tagging platform, to realize the public participation in public opinion corpus tagging, and then put forward error correction mechanism and quality control strategies to ensure the quality of tagging. The establish of Uygur, Kazak, Kirgiz language public opinion tagging corpus can provide powerful resources for the national minority public opinion researches.

Key words: crowdsourcing; corpus tagging; public opinion

1 引言

中国少数民族舆情研究是网络舆情研究的重要内容。对新疆地区少数民族广泛使用的维吾尔语、哈萨克语和柯尔克孜语（以下简称维、哈、柯语），舆情研究还处于起步阶段。无论是基于统计的还是基于规则的舆情研究，标注语料都有重要作用。目前，我国少数民族语言网络信息资源数量整体较少，研究的重点是建立一定规模的可用资源，如何在短时间内构建大规模的舆情语料库成为亟待解决的问题。舆情语料资源的标注一直是困扰科研人员的一大难题，因为语料标注工作往往仅涉及简单的人类智能，并没有太多专业性的要求。

目前对舆情语料标注研究主要集中在对情感和事件的标注，在标注过程中为了保

收稿日期：　　**定稿日期**：

基金项目：国家自然科学基金资助项目（61331011，61262060）；国家重点基础研究发展计划（973）资助项目（2014cb340506）

作者简介：陈昊（1998- ），男，硕士研究生，主要研究方向为自然语言处理；卡哈尔江·阿比的热西提（1984- ），男，博士研究生，主要研究方向为自然语言处理；艾山·吾买尔（1981- ），男，副教授，硕士生导师，主要研究方向为自然语言处理；吐尔根·依布拉音（1958- ），男，教授，博士生导师，主要研究方向为自然语言处理、机器翻译、软件工程等。

证质量，科研人员都会设计一个半自动的C/S 模式标注系统。文献[1]讨论了设计和实现中文情感语料库的基本问题，为文本计算提供了资源支持。系统能够实现语料标注功能，且准确率较高。但标注的粒度较粗，且只用于实验室标注。文献[2]提出了维吾尔语情感语料库的构建规范，设计和实现维吾尔语情感语料库辅助整理工具。文献[3]提出基于自动标注的维吾尔语情感词分析句子情感的方法，但由于语料匮乏，只将否定词和连词引入到情感识别中。

针对上述问题，本文提出了一种基于众包的多语舆情语料标注方法。采用众包方法降低语料标注的成本，提高标注效率。提出了情感语料标注规范，并采用三层架构建立标注平台，为以后的扩展提供保证。

2 众包

众包是《连线》杂志记者 Jeff Howe 于2006 年提出的术语,用来描述一种新的商业模式，即企业利用互联网来将工作分配出去、发现创意和解决问题[4]。通俗的说，众包就是使用某种机制使群体共同参与某件事情，达到某个目标。众包为创造性的能量提供了一个出口[5]。但是，众包并不是另一个外包，众包的任务外派给不确定的群体而外包则是外派给确定的个体；此外外包强调的是高度专业化而众包则反其道而行之，更注重自由和创意，跨专业的创新往往蕴含着巨大的潜力。现在，众包已经应用于自然语言处理的众多领域。[6][7]是目前较成功的商业众包系统，它们使用人的智能完成机器很难完成的工作。文献[8][9]利用众包完成语料处理和获取。

Howe 提出用 UGC(User Generate Content)和 Web2.0 分别是众包生产的资源工具和技术工具。UGC 即用户创造内容，它是一种用户使用互联网的新方式，也就是从原来的下载变成下载和上传并重。Web2.0 技术已经是计算机科学上非常成熟的技术，它主张信息是由人贡献出来的,人是 Web2.0 的灵魂。

大众可以参与到舆情语料标注的各个环节，在语料收集阶段，作为语料资源稀缺的维、哈、柯语，可以通过收集网络大众手中语料资源的方法建立生语料库；在语料标注阶段，让大众参与到语料标注当中，用户发挥自己的兴趣爱好并获得一定威望，积极参与语料标注中，建立标注语料库；为大众推送标注好的语料让他们进行纠错，可以获得准确率较高的语料库。

众包有很多的优点[10]，我们使用众包是因为：首先众包能够降低语料标注的成本和时间，通过互联网发布标注任务，不需要招聘专门的人员进行语料标注，另外多人同时标注语料能够在短时间内获得大量的标注语料。其次，众包能够利用集体智慧，对于相同的语料内容，大众可能有不同的认识，但相同学识水平的人总会以比较大的概率偏向于相同的认识，我们取大众的相同认识作为标注的结果，保证了语料标注的准确性。

3 舆情语料的收集

在标注平台构建之前，收集了一部分的舆情语料。作为舆情语料，必须具有平衡性和系统性，能够代表某一范围内的舆情事实。维、哈、柯语属于资源缺乏语言，网络上能找到的资源非常有限，有观点表达的舆情语料少之又少。本实验室一直致力于维、哈、柯语标准化处理的研究，借鉴汉语、英语等资源丰富语言的语料收集方法，我们首先从门户网站、博客等收集数据，标准化处理后，对不同语言的语料进行分类，然后提取有观点表达的语料存入原始语料库，这种方式获取原始语料的流程如图 1 所示。

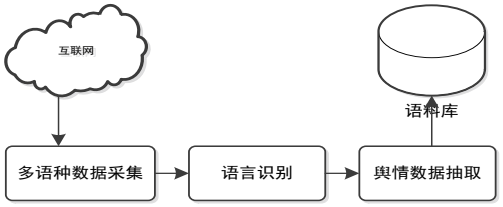

图 1 传统的语料获取方法

使用以上方法，我们初步选取了维吾尔语语料 163 篇，哈萨克语语料 152 篇，以及柯尔克孜语语料 71 篇. 内容包括新闻、博客、微博和评论，三种语言的文字都使用阿拉伯字符,使用UTF_8 编码。从风格上来说，选取的语料涵盖正式、严谨的语料和口语化

语料。

此外，使用用户提供语料的方式收集语找不到语料，所以我们开发了语料上传部分，通过附件上传和用户直接编辑两种方式，让少数民族群众提供原始语料。例如，中小学生作文中包含了大量的人类情感，在舆情情感研究中可以作为原始语料使用，我们希望可以收集到这些语料，标注加工后用于舆情研究。这是众包用在语料收集的体现。

4 语料标注体系

语料标注是将非结构化的文本转化为半结构化文本的过程，是将文本变成知识的过程。为了规范用户的标注，保证标注的一致性，方便后期处理，事先规定好对语料的标注内容，也就是建立语料标注体系。

目前我们只实现对于舆情语料中情感的标注，后期将对语料中的事件等进行标注研究。标注采用文本的情感表达空间模型，把文章情感用空间模型的表达方式表示出来，与文档的自然结构一致。将语料库中篇章文本进行三个层级的标注，即文本层、段落层和句子层的标注；由于微博和评论内容较短，故将语料库中的微博和评论进行两个层级的标注，即文本层和句子层。虽然微博和评论在结构上和句子用词上具有同一性，但为了标注系统的扩展，我们还是将语料分开标注。由于维、哈、柯语在构词和表达习惯上具有相似性，我们将维、哈、柯语的标注进行统一，下面以维吾尔语中情感的标注方法为例介绍维、哈、柯语情感语料标注体系。

一篇文档的情感可以用以下公式来表示：

$$\vec{a} = (e_1, e_2, \ldots, e_i, \ldots, e_n) \quad (1)$$

公式（1）是以每种情感的强度向量来表示篇章的情感。我们将情感分为乐（Joy, خوشال），好（Good, ياخشى），怒（Anger, خاپا），哀（Sorrow, مەيۈسلىنىش），惧（Fear, قورقۇش），恶（Hate, يامان），惊（Surprise, جۆجۈش），所以上式中 n 的值取 7。e_i在{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}中取值，来表示一篇文章中所表达出来的第 i 种情感的强度。此公式亦适用于段落和句子的情感表示：

$$\vec{p} = (e_1, e_2, \ldots, e_i, \ldots, e_n) \quad (2)$$

$$\vec{s} = (e_1, e_2, \ldots, e_i, \ldots, e_n) \quad (3)$$

对于句子中词的标注，我们从表达情感的词或标点（下面作为词来处理）和影响情感表达的词两个方面入手进行标注。对于表达情感的词，在维、哈、柯语中的有些词在某种语境当中可以表达几种情感，这些词语主要是以下几类：第一类是情感关键词。情感关键词是句子中表达情感起决定作用的词。第二类是程度词。程度词表明了它所修饰的情感关键词的程度，它本身就表达情感强度的意思。第三类是标点符号。在情感表达的句子中，使用一些标点符号来加强情感。所以，本文把情感表达的空间模型引入词和标点的情感表达：

$$\vec{w} = (e_1, e_2, \ldots, e_i, \ldots, e_n) \quad (4)$$

影响句子情感表达的词有以下几类：第一类是否定词。句子中出现了一个或者多个否定词，很有可能情感关键词所要表达的意思与整个句子表达的意思不一致。第二类是连词。连词有很丰富的用法，句子中的连词可能是句子中情感的转折点或者升华点。例如维吾尔语句子：

مەن ئۇنىڭ جىشىغا تەگمەيمەن ،چۈنكى ئۇ سىنىپتىكى ئەڭ ئەسكى ئوقۇغۇچى.

我不敢招惹他，因为他是班里最坏的学生。

بۇنداق‌ئەھۋالنى كۆرۈپ ،سىنىڭ يەنە كارىڭ بولمامدۇ؟

看到这样的情况，你还要置之不理吗？

ئالىم بەك سالاپەتلىك،بىراق‌ئۇگىنىشنى ياقتۇرمايدۇ.

阿里木很帅，但是很不爱学习。

第一个句子，程度副词维吾尔语程度词“ئەڭ”（最）表达了说话人对于“ئۇ”（他）的厌恶和恐惧；第二个句子中的用问号强化愤怒与悲哀；第三个句子中使用“بىراق”（但）让情感从褒义变贬义。

由于维、哈、柯语属于黏着语，90%以上的词语是词干加上词缀后形成的，加上词缀的词往往表达了比原词语更丰富的意思。标注过程中，遇到一个词表达多种意思的情况有如下两种：第一种，是“程度词+情感关键词”的组合。第二种，是“否定词+情

感关键词”的组合。对于这两种情况，我们采用不同的处理方法。由于“程度词+情感关键词”是对情感关键词情感的强化或者弱化，可以用情感强度表示出程度词，我们将整个词看成是情感关键词进行标注。而“否定词+情感关键词”是对情感关键词的否定，将其作为关键词和否定词分别进行标注。

表 1 标注体系中各变量说明

类别	变量	说明	取值范围
篇章标注	Title	文章标题	
	Classify	分类	Comment, Microblog, Article
	Language	语言	Uygur, Kazak, Kirgiz
	Emotion	文章情感	Joy，Good，Anger，Sorrow，Fear，Hate，Surprise（其中，每个情感又有 0.0-1.0 的强度值）
	Polarity	情感倾向	Positive，Negative，Neutral
	Topic	主题词	
	Paragraph	段落标注	见段落标注
段落标注	P_no	段落编号	1,2,3……
	Emotion	段落情感	Joy，Good，Anger，Sorrow，Fear，Hate，Surprise（其中，每个情感又有 0.0-1.0 的强度值）
	Polarity	情感倾向	Positive，Negative，Neutral
	Summerise	中心句	
	Fact_opinion	事实，建议	Fact，Opinion
	Sentence	句子标注	见句子标注
句子标注	S_no	句子编号	1,2,3……
	S_length	句子长度	1,2,3……
	Emotion	句子情感	Joy，Good，Anger，Sorrow，Fear，Hate，Surprise（其中，每个情感又有 0.0-1.0 的强度值）
	Polarity	情感倾向	Positive，Negative，Neutral
	E_degree	程度词	包括程度词表达出的情感
	Punctuation	标点	包括标点表达出的情感
	Key_words	关键词	包括关键词表达出的情感
	No_words	否定词	
	Conjunction	连词	
	E_holder	情感发出者	
	E_target	情感接受者	
	Rhetoric	修辞	Analogy， Personification， Exaggeration， Parallelism， Antithesis, Repetition, Rhetorical question, Irony.

综上，本文制定了维吾尔语、哈萨克语和柯尔克孜语的舆情语料情感标注体系：

Document Model=([Title],Classify，Language, Emotion, Polarity, Topic, [Paragraph]) (5)

Paragraph Model=(P_no, Emotion, Polarity, Summerise, Fact_opinion, Sentence) (6)

Sentence Model=(S_no, S_length, Emotion, Fact_opinion， Polarity, [E_degree], [Punctuation], Key_words, [No_words], [Conjunction], [E_holder], [E_target], [Rhetoric]) (7)

公式中，方括号里的标注项是可选择的，其中语篇的标注中，只有文章类的需要标注标题，微博和评论不需要，不在方括号里的为必须标注内容。语料情感标注体系中的变量以及对它们的说明如表 1 所示。

由于程度词、标点能够表达一定情感，且某个句子或者词语所表达出来的情感成分往往不止一种，所以标注中，Emotion、E_degree、Punctuation、Key_words 项中除了要标注表达的情感，还要标注每种情感的强度。另外，否定词、连词对于句子情感有一定影响，系统将其标注出来。修辞中的值包括 Analogy，Personification，Exaggeration，Parallelism，Antithesis，Repetition，Rhetorical question，Irony 即比喻、拟人、夸张、排比、对偶、重复、设问反问和反语。文献[1]也对修辞进行了标注，但是没有标注反语，文献[11]通过对葡萄牙语中反语在政治评论中的作用研究证明，反语对于情感表达有很大的影响，所以，将反语加入修辞的标注内容中。

图 2：生成标注的部分结果

语料标注结果都保存为 XML 格式文档，通过一种半结构化的模型将 XML 文档同数据库对应起来，从而实现对于数据模型的抽取。XML 文档中的标签是根据上面的语料标注体系制定的。图 2 截取了数据库中某篇维吾尔语的部分标注结果。这是对一个句子的标注，句子来自一篇网络博客，意思是：所以女人说闲话，男人也不会抱怨她。整个句子作为一个事实的表达，表达出男人对于女人的态度；“ئەرلەرنۇ”（男人）是意见的发出者；“ئاياللار”（女人）是意见的接受者；句子中的关键词“ئۆسەك”（闲话）表达出了女人的情感，而关键词“ئەيبلەپ”（抱怨）表达了情感发出者对于情感接受者的态度；句子中还有连词“شۇڭا ھازىر”（所以）和否定词“كەتمەيدۇ”（不会）对于情感表达起引导和否定作用。

5 标注平台的设计

5.1 平台架构

使用 ASP.NET 下的 C#语言开发，使用微软 SQL Server 2008 作为存储数据库，开发了多语舆情语料标注平台。平台建设将实际的业务需求同三层架构联系起来，为平台的后期维护以及语言资源扩展提供了方便。平台的架构方式如图 3 所示。

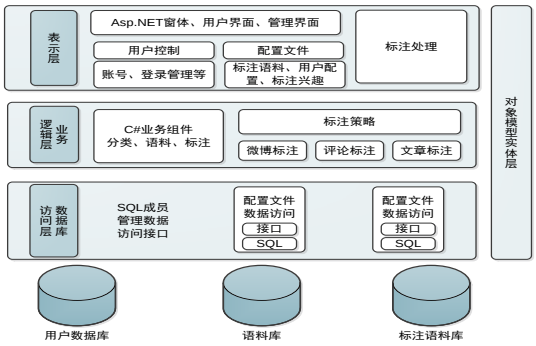

图 3：语料标注平台架构

为保证平台的可维护性和可扩展性，平台建设使用三层架构的方法。标注之前采集到的数据放入语料库，用户加工过的语料通过业务逻辑层和数据访问层保存到标注语料库；数据访问层和业务逻辑层部署在 Web 服务器上，它是平台的核心，平台中标注规范、段落切分、标注评价等算法都在这个层中编写；由于平台是众包的，平台应该对用户非常友好，所以对于表示层的设计非常重要。平台采用 Web 2.0 的 ASP.NET Ajax 技术来减轻服务器的负担、减少用户的等待时间。用户标注非常简单，只需要通过鼠标选

取相应的词点击按钮就可以完成标注，标注后的结果自动显示在文本框中。图 4 是对于句子标注界面的截图。

平台构建过程中，使用音节切分算法对每个句子中的单词进行计数[12]，用启发式搜索算法进行语句切割[13]。通过标注页面的设计可以保证用户标注的准确性和规范性。每个用户完成的标注都会被存进数据库。这是众包在语料标注中的应用。

图 4 情感标注中的句子标注

5.2 质量控制

平台目前采用大多数语料标注系统使用的方法即交叉标注来保证语料标注的准确性，多用户对同一篇语料进行标注，对标注后的结果进行比较，对各用户标注的情感强度值取截尾平均数，保留大多数用户标注相同的结果。但这样做存在弊端，有些标注用户会随意标注或者标注不完全，把这种用户称为欺骗型工作者。欺骗型工作者提交的结果会影响标注结果的判断。所以，本文提出了两种纠错机制。首先，在平台中建立标注修改区，用户可以进入修改区对别的用户的标注进行修改。其次，利用用户评估标注结果，标注结果反映用户信誉并多次迭代的方法进行评估。具体的流程如图 5 所示。

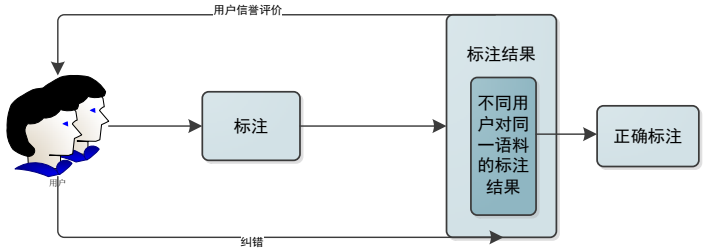

图 5 质量控制流程

文献[14]指出，用户参与众包的动机主要有满足心理需求、金钱以及学习新知识和技能。为满足用户动机平台建立了用户激励机制。我们有一个积分系统对用户的贡献信息进行计算，计算用户获得积分的公式：

$$Score_i = TSN_i \times \mu_i + CSN_i \times \tau_i \quad (8)$$

其中，$Score_i$表示用户i的积分，TSN_i表示用户标注的总的句子数，μ_i表示用户标注的正确率，也就是用户标注正确的句子数与用户标注总句子数的比；CSN_i表示用户修改的总句子数，τ_i表示修改的采纳率，也就是修改正确的句子数与修改总句子数的比值。一定的积分值可以换取相应的金钱或者物品。利用激励机制可以引导用户积极参与到语料标注工作中。

6 结束语

本文提出了一种维吾尔语、哈萨克语和柯尔克孜语舆情语料标注的新方法，即基于众包的舆情语料标注。构建了维吾尔语、哈萨克语和柯尔克孜语语料标注平台，平台使用众包充分利用大众智慧，从语料的收集、语料标注到标注的纠错都使用众包，为语料收集提供了另一种方法，提高了标注的效率和准确性。语料标注平台建设之初制定标注规范，使用三层架构的方法构建标注平台，使用用户评价标注，标注反映用户信用并多次迭代的方法保证标注质量。

本平台还有很大的发展空间。首先，平台目前只用于情感语料库的构建，下一步应该从舆情分析的各个方面入手进行建设，包括事件标注，情感词典建设，维-汉、哈-汉、柯-汉对齐语料建设等。其次，系统的激励机制过于单一，应该设计能够鼓励大众参与标注的激励机制。第三，平台界面现在是汉语的，为让少数民族群众更好参与到标注中来，应开发多语种标注网站。最后，目前的纠错方法只适用于语料规模比较小的情况，在标注的准确率方面有待进一步进行研究。

参考文献

[1] 徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析[J]. 中文信息学报, 2008, 22(1): 116-122.

[2] 冯冠军, 禹龙, 田生伟. 基于 CRFs 自动构建维吾尔语情感词语料库[J]. 现代图书情报技术, 2011, 27: 17-21.

[3] 黄俊、田生伟、禹龙、冯冠军. 基于维吾尔语情感词的句子情感分析. 计算机工程, 2012, 38(9): 183-185.

[4] Howe J. The rise of crowdsourcing[J]. Wired magazine, 2006, 14(6): 1-4.

[5] Savage N. Gaining wisdom from crowds[C]. New York, USA: Communications of the ACM, 2012, 55(3): 13-15.

[6] Kittur A, Chi E H, Suh B. Crowdsourcing user studies with Mechanical Turk[C]. New York, USA: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2008: 453-456.

[7] https://www.duolingo.com/

[8] Mcduff D, el Kalioubyr, Picard R. Taxs, USA: Crowdsourced data collection of facialresponses[C]. Proceedings of the 13th international conference on multimodalinterfaces. ACM, 2011: 11-18.

[9] Castillo C, Mendoza M,Poblete B. Information Credibility on Twitter[C]. Hyderabad, India: WWW 2011-Session:Information Credibility. ACM, 2011:675-684.

[10] 张利斌, 钟复平, 涂慧. 众包问题研究综述[J]. 科技进步与对策, 2012, 29(6): 154-160.

[11] Carvalho P, Sarmento L, Teixeira J, et al. Liars and saviors in a sentiment annotated corpus of comments to political debates[C]. Portland, Oregon, USA :Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011: 564-568.

[12] 阿比达·吾买尔,吐尔根·依布拉音. 维吾尔文音节切分方法的研究与实现[C]. 云南西双版纳:民族语言文字信息技术研究——第十一届全国民族语言文字信息学术研讨会论文集. 2007.

[13] Chris Manning，Hinrich Schütze.统计自然语言处理基础[M].北京：电子工业出版社.2005.

[14] 仲秋雁,王彦杰,裘江南.众包社区用户持续参与行为实证研究[J].大连理工大学学报(社会科学版).2011(3):1-6.

作者联系方式：吐尔根·依布拉音 新疆乌鲁木齐市胜利路14号新疆大学信息学科学与工程学院 830046 13899810133 turgun@xju.edu.cn