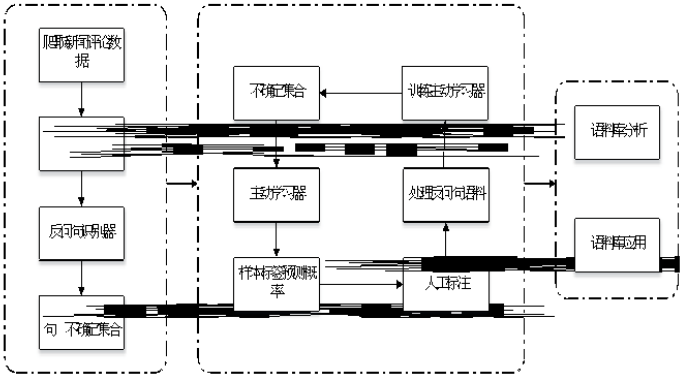
面向新闻评论的汉语反问句语料库构建

李翔，刘承伟，朱晓旭*
苏州大学 计算机科学与技术学院江 苏省苏州市十梓街1号 215006 中国
949460097@qq.com，liuliuchengwei@gmail.com，
xiaoxzhu@suda.edu.cn

- 反问句是汉语中一种常用的、有特色的疑问句，由于其自身的特殊性以及重要的应用价值，一直受到诸多研究者的关注。越来越多的自然语言处理任务都要求对文本进行更细粒度的情感分析，而反问句作为一种带有强烈感情色彩的特殊表达方式，如果能对其进行正确的识别，将会改善情感分析等任务的结果。
- 本文提出一种基于半监督学习和主动学习的半自动反问句语料收集方法，构建了一个6000余句的新闻评论汉语反问句语料库，并将其应用于反问句识别，取得良好效果。

语料库构建流程



- 在新浪体育网中国足球版块获取了大量评论数据，并对数据进行了去重、分句等处理。
- 构建了一个基于特征词典的反问句识别器以筛选出数据中反问特征明显的反问句，并将数据分为反问句集合、不确定集合以及非反问句集合。
- 为了最小化人工标注的工作量，本文提出未标注样本选择策略，构建了基于CNN、LSTM的反问句主动学习器。

基于特征词典的反问句识别器

- 根据反问句特征词及特征词对句子反诘度影响建立特征词典 □□□□□□□□ = {(□□□□□□□□$_1$, □□□□□$_1$), ……(□□□□□□□□$_k$, □□□□□$_k$)}, 0 < □□□□□ < 10，□□□□□值越大，对反诘度影响越大。
- 设句子□ = (□$_1$, □$_2$, □$_3$, …, □$_n$)由n个词语组成，含有m个反问特征词，通过查询反问特征词典可得句子的□□□□□ = $\sum_i^m$ □□□□□(□)。当□□□□□ > □h□□□h□□□1时，□扩充至反问句集合；□□□□□ < □h□□□h□□□2时，□扩充进非反问句集合，其余情况□扩充进不确定集合。通过实验，□h□□□h□□□1取10，□h□□□h□□□2取4时，识别器对反问句的识别精确度为91.6%，对非反问句识别精确率为93%。
- 反问句识别器共计进行了三次迭代，获得3378条反问句，57698条非反问句，不确定集合中含有5802条评论数据。

基于CNN、LSTM的反问句主动学习器

未标注样本选择策略

- 任一学习器的□$_{c0}$、□$_{c1}$的值越接近，学习器对样本属于哪一类标签的不确定度□□□□□□□□□越高，即样本被错误分类的可能性就越大，这一类的样本应由人工标注。因此，□□□□□□□□□大于0.8时，样本被学习器选择

$$□□□□□□□□□ = 1 - |□_{c0} - □_{c1}| \quad (1)$$

- 两个学习器对同一样本数据预测结果是否一致，用该样本的分歧度□□□□□□□□□表示。□□□□□□□□□ < 0,预测结果一致；□□□□□□□□□ > 0，预测结果不一致。

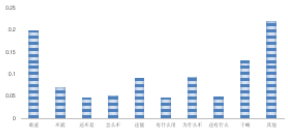
$$□□□□□□□□□ = -(□_{nc0} - 0.5)(□_{lc0} - 0.5) \quad (2)$$

训练好的CNN、LSTM反问句主动学习器按照上述样本选择策略仅筛选不确定度高、分歧度大的样本交由人工标注。本文从不确定集合中人工标注500条反问句训练主动学习器，学习器每次对500条数据进行学习，共计进行10次，获得反问句2670条，其中人工标注数据量为21%。

语料库分析

语料库数据统计

	6548	48.80
	6191	43.80
	2654	49.39
	3894	48.40

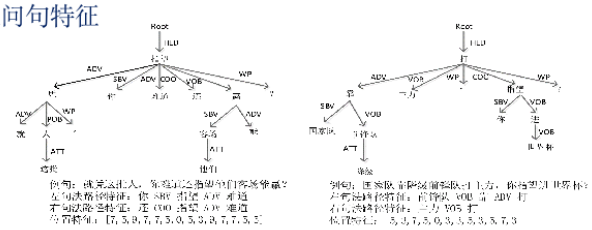


语料库中，反问句数为6548，其中显式反问句占比40.5%，隐式占比59.5%。本文在非反问句集合中抽取6191条非反问句，平均长度为43.8，平均长度为48.8，高于非反问句的平均长度。显式反问句句子平均长度为49.39，隐式反问句句子平均长度为48.40。

左图显示的是在显式反问句中特征词的使用频率。“难道”一词的使用频率最高，在本文的构建的语料库中，“难道”在显式反问句中的占比达到20%，置信度为97.7%。“干嘛”类反问句也是比较典型的反问句，通过句末语气词“嘛”增强反诘句调，语料库中占比13%。

Experiments

- 反问句特征



- 基准系统
  - CNN、LSTM、Transformer、Bert
  - CNN[W+F]表示文治等人设计的模型；Auto-AOA表示李旸等人设计的模型。
- 实验结果

模型	特征	Precision%	Recall%	F1- measure
CNN	Fs	86.35	84.01	85.16
CNN	Fs+Fpo+Fpa	**87.16**	**88.40**	**87.77**
BiLSTM	Fs	86.03	88.71	86.34
BiLSTM	Fs+Fpo+Fpa	**86.91**	**89.34**	**88.01**
Transformer	Fs	87.08	87.62	87.33
Transformer	Fs+Fpo+Fpa	**88.29**	**92.16**	**90.18**
Bert	Fs	90.47	92.16	91.30
Bert	Fs+Fpo+Fpa	**90.79**	**93.57**	**92.15**
CNN[W+F]		89.50	84.20	86.70
Auto-AOA		90.70	92.30	91.40

- 与文治等人的模型相比，CNN使用本文选取的反问句特征时的召回率、F1值要略高于他们的模型。本文考虑到反问句是一种高度依赖语境的表达方式，从句法分析角度提取句子的位置特征，同时考察了的上下文环境，使得召回率提高4.2%，F1值提高1.17%。与李旸等人的模型相比，Bert模型在使用本文构建的语料库时，性能略高于李旸等人的模型。