数据标注研究综述

蔡莉[1,2] 王淑婷[1] 刘俊辉[1] 朱扬勇[2]

1：云南大学软件学院

2：复旦大学计算机科学技术学院

作者介绍

蔡莉：
复旦大学计算机科学技术学院博士
研究领域：
数据挖掘、数据质量、可视化

朱扬勇：
复旦大学计算机软件专业教授、博导
研究领域：
数据学、数据科学

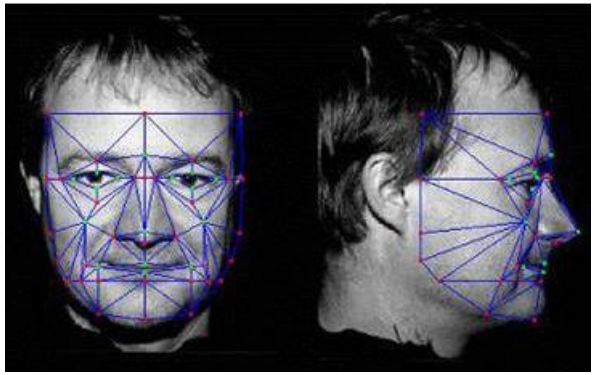
数据标注

数据标注是对未处理的初级数据，包括语音、图片、文本、视频等进行加工处理，并转换为机器可识别信息的过程。

Wang C, Blei DM, Li F, et al. Simultaneous image classification and annotation. Computer vision and pattern recognition, 2009: 1903-1910.

数据标注应用

自动驾驶



智能安防

智慧医疗

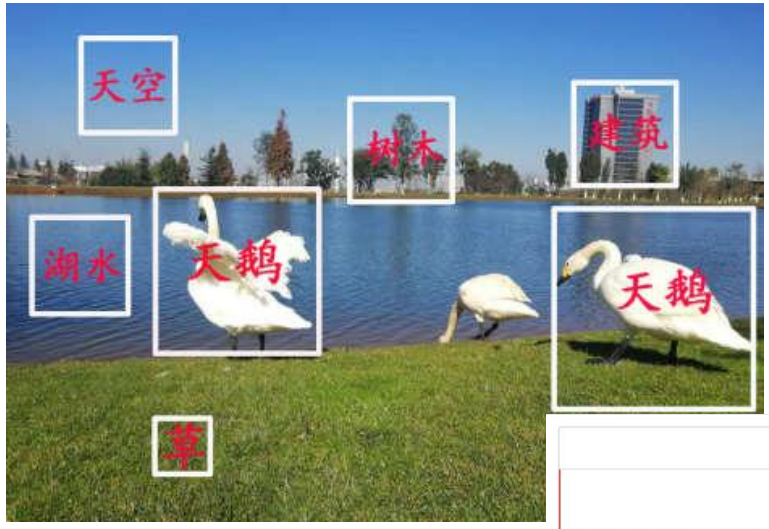
工业4.0

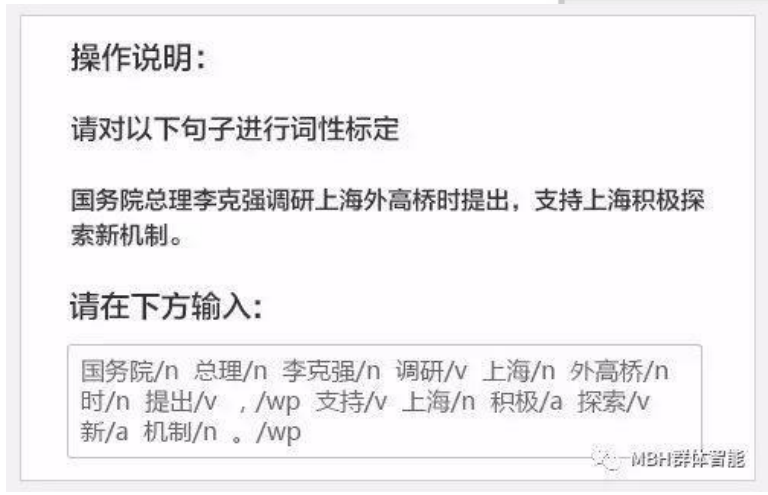
新零售

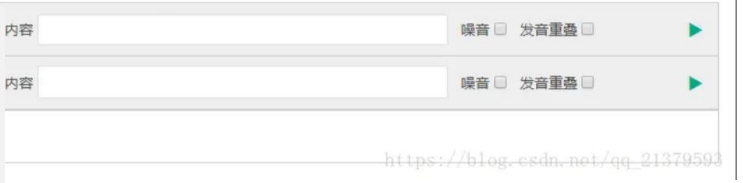
智慧农业

数据标注分类(1/3)

- 标注对象
  - 图像标注
  - 语音标注
  - 文本标注







数据标注分类(2/3)

构成形式

- 结构化标注
  - 标签域确定
  - 标签值选择
- 非结构化标注
  - 标签域自由
  - 标签值自由
  - 带有一定约束
- 半结构化标注
  - 标签域自由
  - 标签值选择

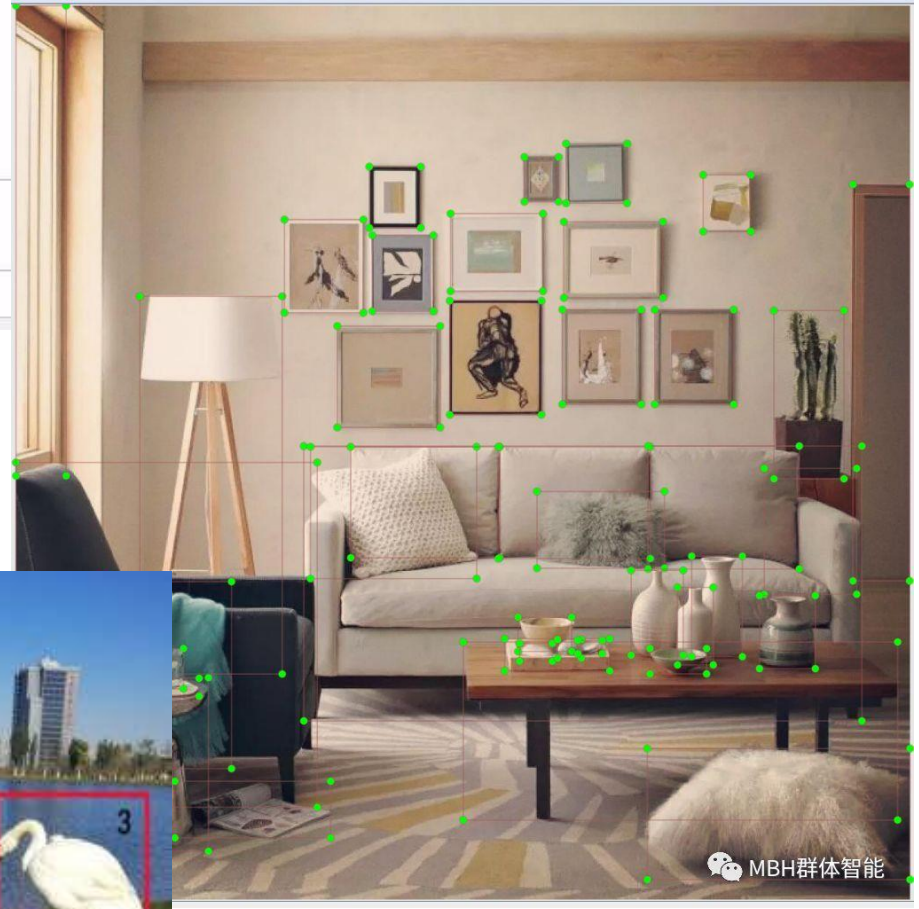
操作说明：

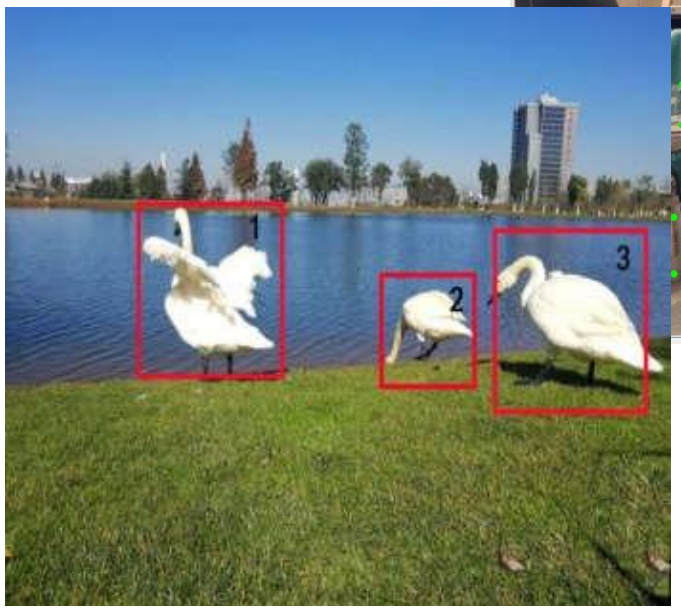
请分析文中内容，并选择其表达的情绪

今天是星期天，可是我们还要加班。

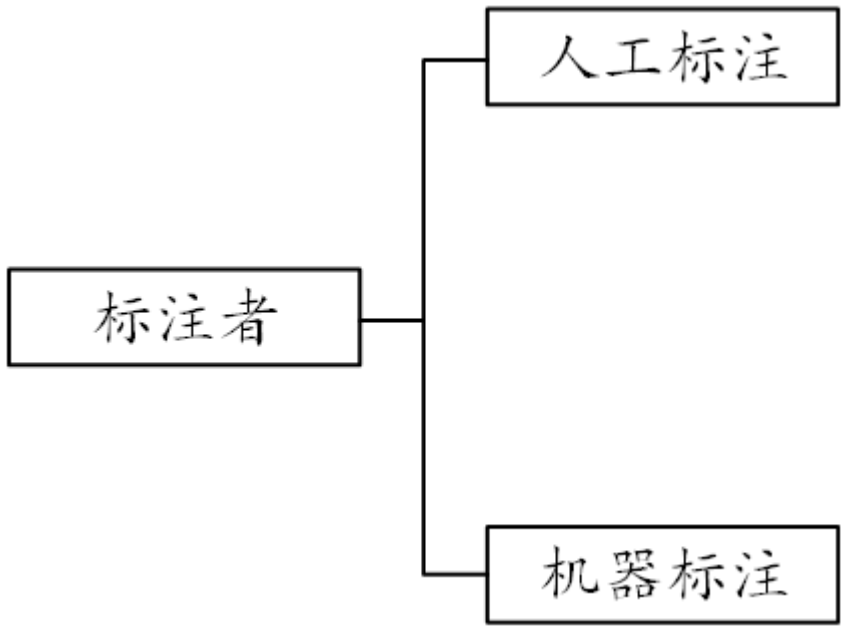
请在下方选择：

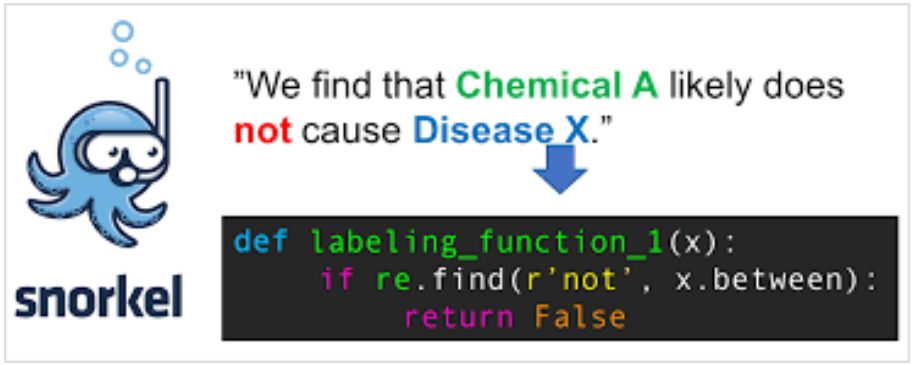
1.开心　2.愤怒　**3.低落**





数据标注分类(3/3)





Snorkel：https://towardsdatascience.com/snorkel-a-weak-supervision-system-a8943c9b639f

数据标注任务

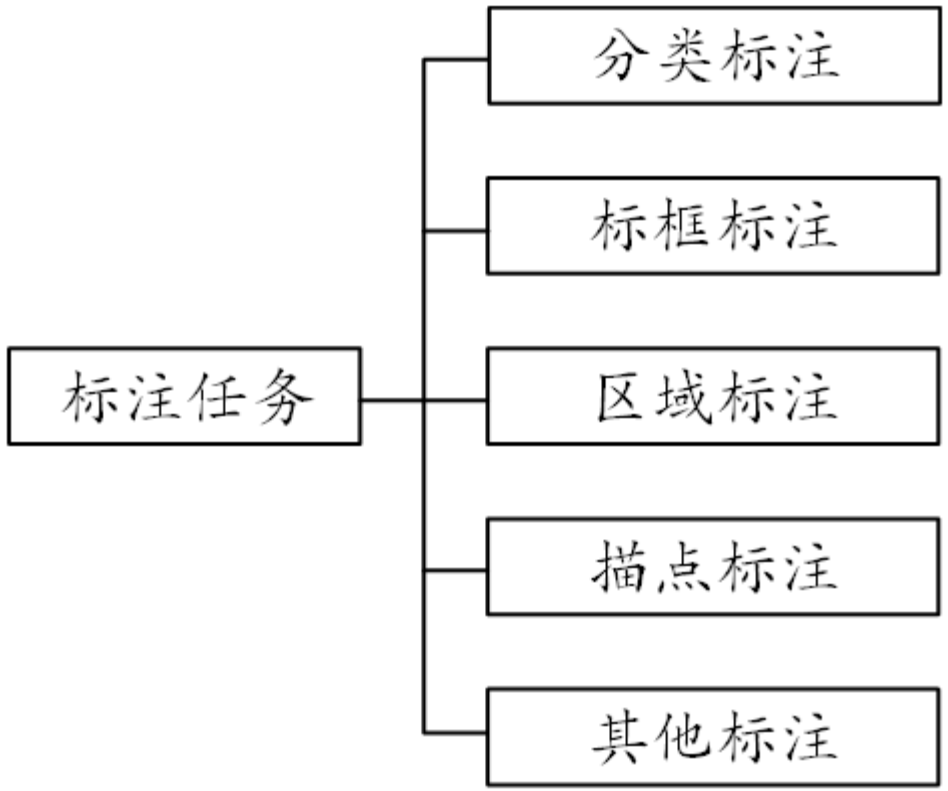


Fig.3 Classification annotation
图 3 分类标注

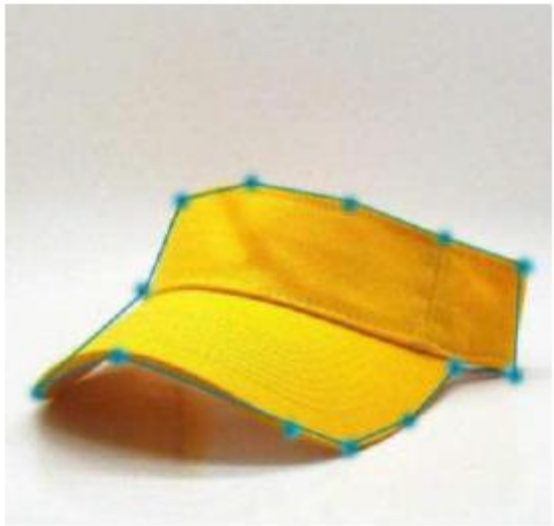


Fig.4 Polygonal frame
图 4 多边形拉框

Fig.6 Region annotation
图 6 区域标注

Fig.7 Point annotation
图 7 描点标注

数据标注数据集

数据标注数据集					
类别	数据集名称	用途	大小	链接	开放
图像标注数据集	ImageNet	图像分类、定位、检测	1TB	http://www.image-net.org/about-stats	是
	COCO	图像识别、分割和图像语义	40G	http://cocodataset.org/#home	是
	PASCAL VOC	图像分类、定位、检测	2GB	http://host.robots.ox.ac.uk/pascal/VOC/	是
	Open Image	图像分类、定位、检测	1.5GB	https://github.com/openimages/dataset	是
	Flickr30k	图片描述	30MB	http://shannon.cs.illinois.edu/DenotationGraph/	是
视频标注数据集	Youtube-8M	理解和识别视频内容	1PB	https://research.google.com/youtube8m/	受限
	Kinetics	动作理解和识别	1.5TB	https://deepmind.com/research/open-source/	是
	AVA	人类动作识别		https://research.google.com/ava/	是
	UCF101	视频分类、动作识别	6.5GB		是
文本标注数据集	Yelp	文本情感分析	2.66GB	https://www.yelp.com/dataset/challenge	是
	IMDB	文本情感分析	80.2MB	http://ai.stanford.edu/~amaas/data/sentiment/	是
	Multi-Domain Sentiment	文本情感分析	52MB	https://www.cs.jhu.edu/~mdredze/datasets/sentiment/	是
	Sentiment 140	文本情感分析	80MB	http://help.sentiment140.com/	是
语音标注数据集	LibriSpeech	训练声学模型	60GB		是
	AudioSet	声学事件检测	80 MB	https://research.google.com/audioset/	是
	FMA	语言识别	1000G	https://github.com/mdeff/fma	是
	VoxCeleb	语音识别、情绪识别	150MB		是

数据标注平台

数据标注平台	
名称	链接
Mechanical Turk	https://www.mturk.com/
Figure-eight	https://www.figure-eight.com/
Mighty AI	被Uber收购，不开放了
数据堂	http://bz.datatang.com/
百度众测	https://test.baidu.com/
阿里众包	https://newjob.taobao.com/
京东微工	https://weigong.jd.com/

数据标注开源工具

数据标注开源工具					
名称	简介	运行平台	标注形式	导出数据格式	链接
LabelImg	著名的图像标注工具	Windows\Linux\Mac	矩形	XML 格式	https://github.com/tzutalin/labelImg
LabelMe	著名的图形界面标注工具,能标注图像和视频	Windows\Linux\Mac	多边形、矩形、圆形、多段线、线段、点	VOC 和COCO 格式	https://github.com/wkentaro/labelme
RectLabel	图像标注	Mac	多边形、矩形、多段线、线段、点	YOLO、KITTI、COCO1与CSV 格式	https://github.com/ryouchinsa/Rectlabel-support
VOTT	微软发布的基于Web 方式本地部署的标注工具,能标注图像和视频	Windows\Linux\Mac	多边形、矩形、点	TFRecord、CSV、VoTT格式	https://github.com/microsoft/VoTT
LabelBox	适用于大型项目的标注工具,基于Web、能标注图像、视频和文本		多边形、矩形、线、点、嵌套分类	JSON 格式	https://github.com/Labelbox/Labelbox
VIA	VGG(Visual Geometry Group)的图像标注工具,也支持视频和音频标注		矩形、圆、椭圆、多边形、点和线	JSON 格式	https://github.com/LakorTi/Via
COCO UI	用于标注 COCO 数据集的工具,基于Web方式		矩形、多边形、点和线	COCO 格式	http://cocodataset.org/#home
Vatic	Vatic 是一个带有目标跟踪的视频标注工具,适合目标检测任务	Linux		VOC 格式	https://github.com/cvondrick/vatic
BRAT	基于Web 的文本标注工具,主要用于对文本的结构化标注	Linux		ANN 格式	https://github.com/nlplab/brat
DeepDive	处理非结构化文本的标注工具	Linux		NLP 格式	https://github.com/HazyResearch/deepdive
Praat	语音标注工具	Windows\Unix\Linux\Mac		JSON 格式	https://github.com/praat/praat
精灵标注助手	多功能标注工具	Windows\Linux\Mac	矩形、多边形和曲线	XML 格式	http://www.jinglingbiaozhu.com/

数据标注规范

图像标注

标注像素点越接近标注物的边缘像素 → 标注质量越高

语音标注

标注与发音时间轴的误差越小（在1个语音帧以内） → 标注质量越高

文本标注

分词不存在歧义

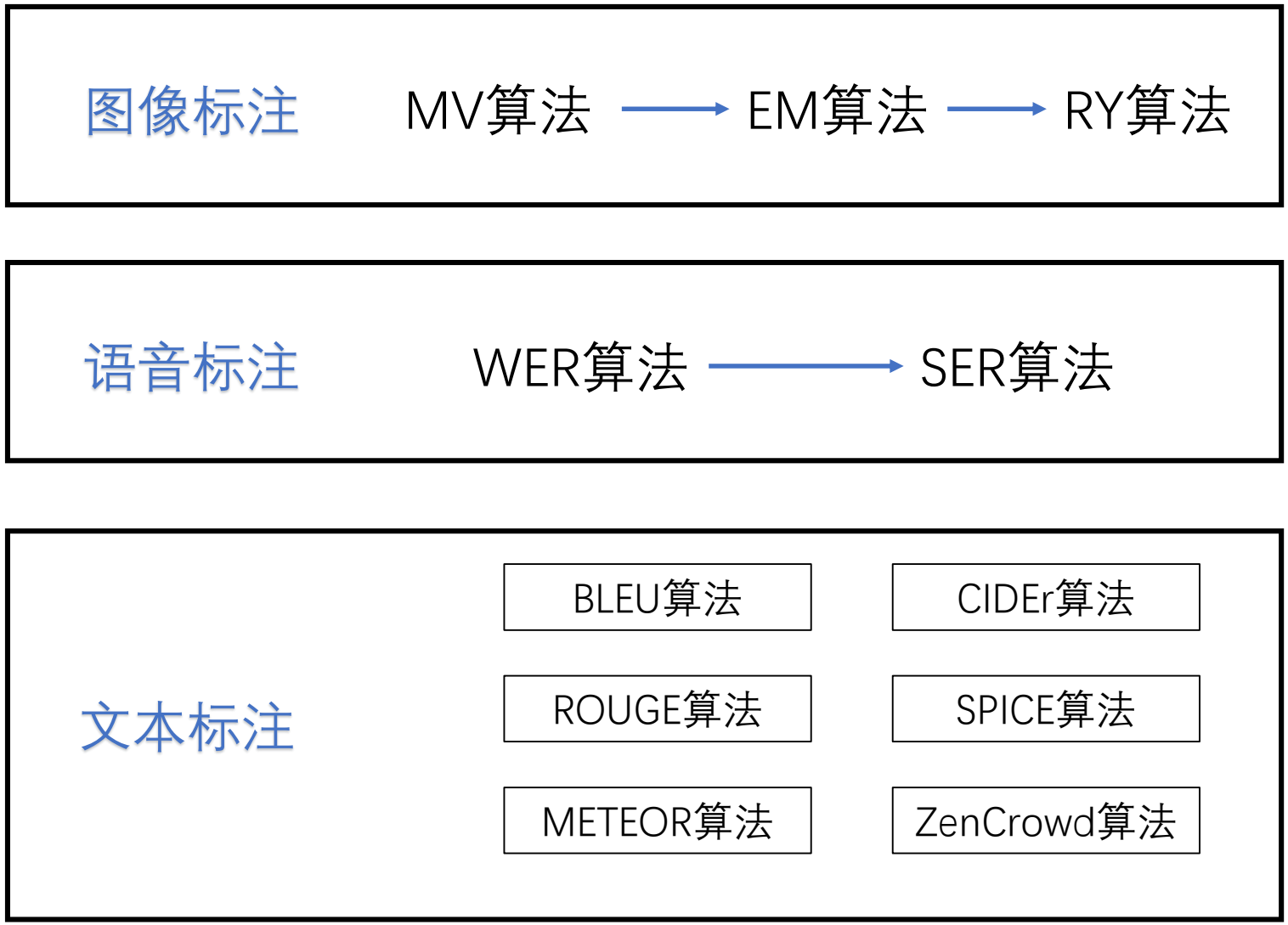
情感分类级别正确

清楚词语或句子的真实语义

→ 标注质量越高

*实际灵活多变，根据任务确定标准

数据标注质量评估


图像标注
MV算法
EM算法
RY算法
语音标注
WER算法
SER算法
文本标注
BLEU算法
CIDEr算法
ROUGE算法
SPICE算法
METEOR算法
ZenCrowd算法

数据标注面临的挑战

挑战1 不同的行业应用对数据标注的任务存在一定的差异性，现有的标注任务还不够细化，无法满足行业的新技术需求。

挑战2 尽管数据标注工具能在一定程度上帮助标注员完成标注任务，但是整体的标注效率仍然较为低下。

挑战3 现有数据标注平台普遍采用众包模式来分配标注任务、造成标注结果的质量层次不齐，影响算法模型的准确性。

挑战4 基于众包模式的数据标注任务会造成用户数据缺乏安全性，并面临隐私泄露的风险。

数据标注发展趋势

发展1 细化数据标注任务

发展2 半自动化数据标注工具的研发

发展3 数据标注质量的改善

发展4 数据标注中的安全性与隐私保护

THANKS!