2019 CCF 大数据与计算智能大赛

金融负面信息及主体判定

队名 zgkxydx

队员 qhd1996 DF1566796370903 DF1537943240664

指导老师 李秋丹

学院 中国科学院大学人工智能学院

中国科学院大学
University of Chinese Academy of Sciences

目录
Contents

01 任务简介

任务简介

- 任务背景

  随着互联网的飞速进步和全球金融的高速发展，网络信息愈发得到人们的重视。获取文对实体的情感倾向对金融决策和商品投资起到重要的影响。

- 任务目标

  给定一条金融文本和文本中出现的金融实体列表，判定该文本是否包含金融实体的负面信息，哪些实体是负面金融实体。

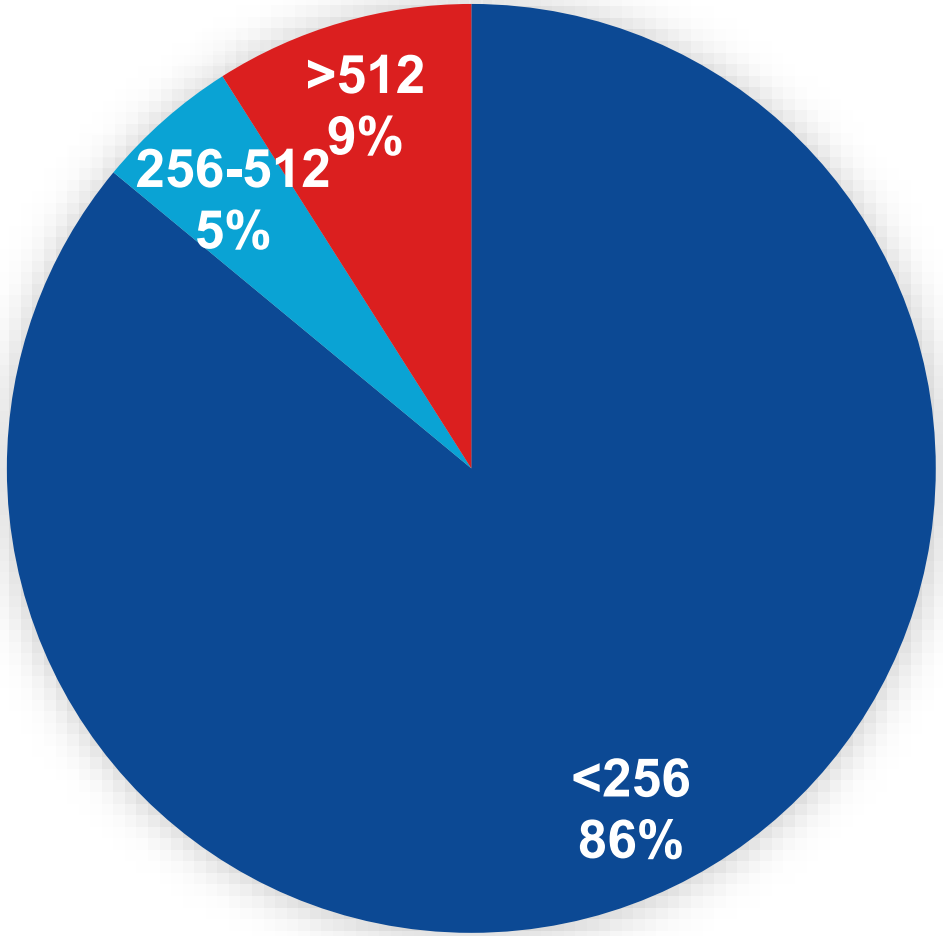
- 数据格式

id	title	text	entity	negative	key entity
1060	天伦金服 2017-12-15 浙江杭州 平台失联	网友爆料天伦金服逾期未回款已失联，大家快去维权吧	天伦金服	1	天伦金服

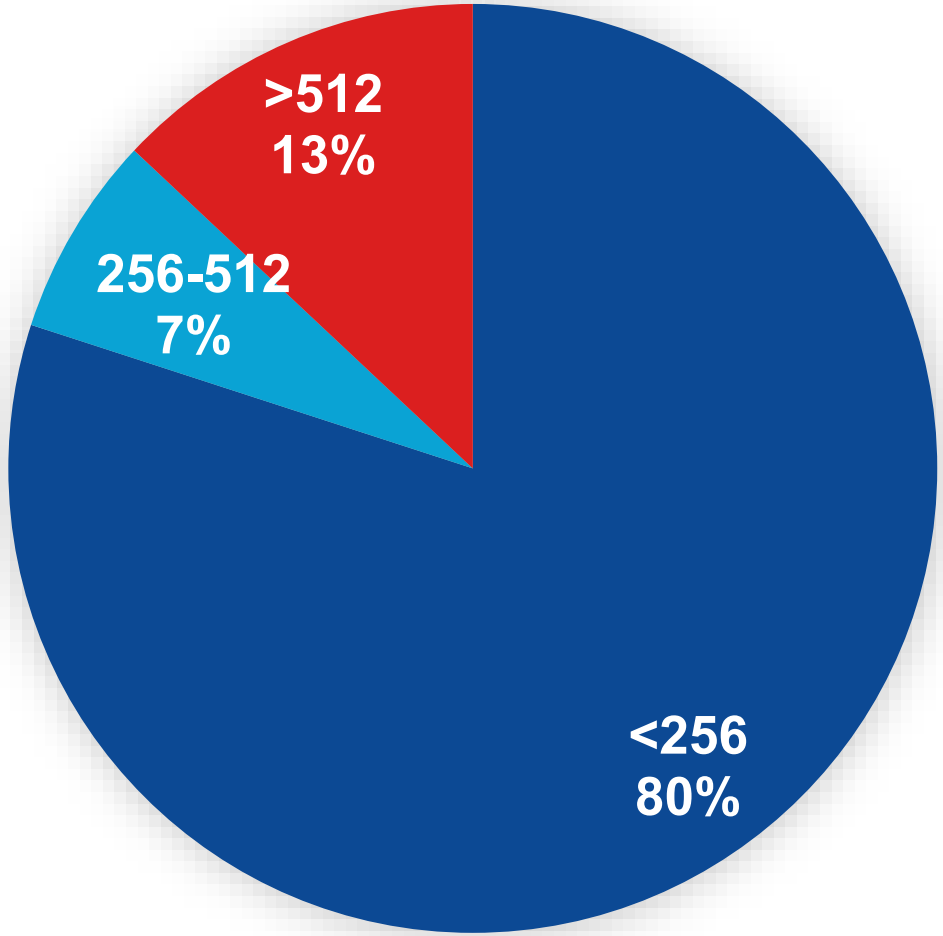
02 数据分析

数据分析

训练集文本长度分布



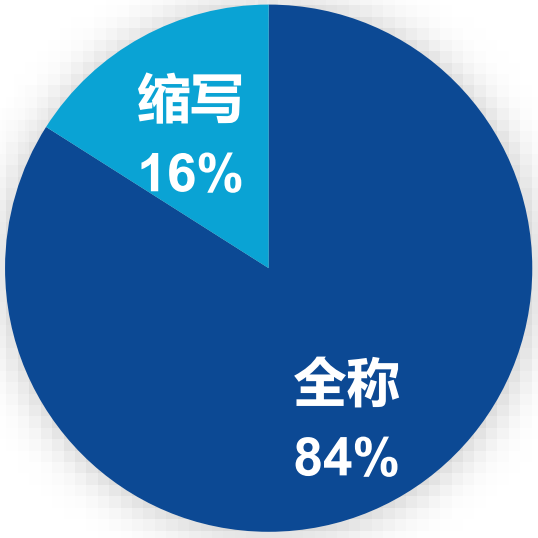
测试集文本长度分布



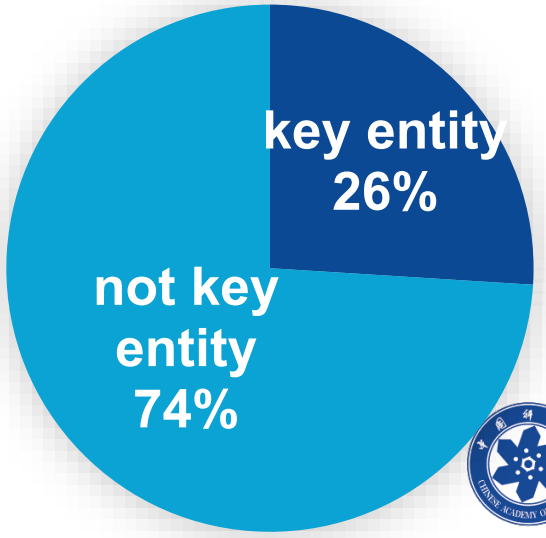
数据分析

id	title	text	entity	negative	key entity
1060	天伦金服 2017-12-15 浙江杭州 平台失联	网友爆料天伦金服逾期未回款已失联，大家快去维权吧	全称 缩写 天伦金服;天伦	1	天伦金服
7360	？？？？ #P2P网络平台暴雷百姓血本无归#诈骗黑窝	？？？？ #P2P网络平台暴雷百姓血本无归#诈骗黑窝	隐式实体 五粮液	1	五粮液

训练集key entity全称与缩率分布

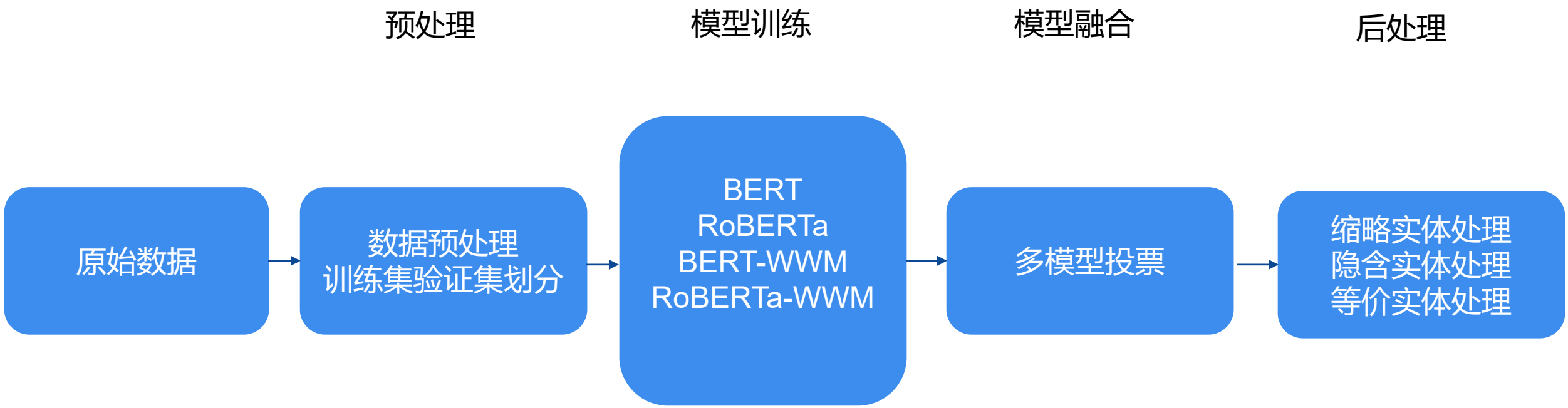


训练集隐式实体出现在key entity的分布



03 方案设计

方案设计 – 整体流程



方案设计 – 数据集构建

将任务转化为句子+entity的二分类任务

text	entity	key entity
2018年7月14日，杭州市公安局对杭州钱内助金融有限公司(平台名:钱内助)涉嫌非法吸收公众存款案进行立案侦查	助金; 杭州钱内助金融有限公司; 钱内助	杭州钱内助金融有限公司; 钱内助

text	entity	label
2018年7月14日，杭州市公安局对杭州钱内助金融有限公司(平台名:钱内助)涉嫌非法吸收公众存款案进行立案侦查	助金	0
2018年7月14日，杭州市公安局对杭州钱内助金融有限公司(平台名:钱内助)涉嫌非法吸收公众存款案进行立案侦查	杭州钱内助金融有限公司	1
2018年7月14日，杭州市公安局对杭州钱内助金融有限公司(平台名:钱内助)涉嫌非法吸收公众存款案进行立案侦查	钱内助	1

方案设计 – 数据预处理

数据预处理

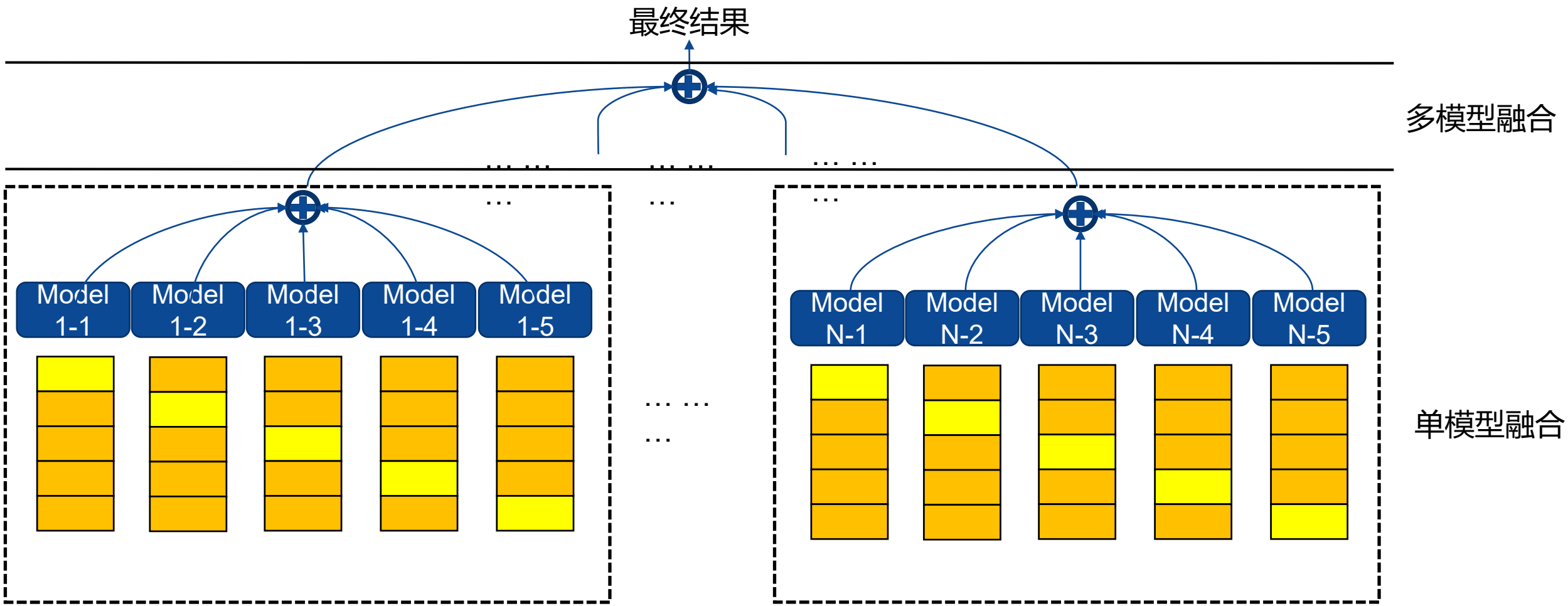
1. 去除不含entity的数据
2. 去除和替换特殊字符
    - 去除超链接
    - 去除非utf-8字符
    - 去除连续出现的标点符号
    - 去除html转义字符
    - 保留#，？
3. 对于长度超过300个文本的文本，截取前300个字符(选择300纯属显存限制)

方案设计 – 模型训练

- 使用预训练模型
- 输入两段文本，第一段是text，第二段是entity
- 初始学习率5e-5,最小学习率1e-5
- 每个模型使用5-fold
- 选择最后一层输出序列cls作为feature

模型	Acc	F1
BERT	0.944	0.927
RoBERTa	0.926	0.914
BERT-WWM	0.932	0.931
RoBERTa-WWM	0.921	0.920

方案设计 – 模型融合


最终结果
多模型融合
Model 1-1
Model 1-2
Model 1-3
Model 1-4
Model 1-5
Model N-1
Model N-2
Model N-3
Model N-4
Model N-5
单模型融合

方案设计 – 后处理

- 针对缩略实体，如果在训练集中出现且判定为1，则在测试集中也标注为1，否则为0;
- 针对隐含实体，由于数量不多，通过人工判定;
- 针对等价类实体，标注为等价类内多数的标签

text	entity	raw label	count	modified label
根据通报，此前出现问题的聚财猫、永利宝、米袋计划、坚果理财、小诸葛平台等均已被立案侦查。	聚财猫	1	label为1: 4	1
	永利宝	1	label为0: 1	1
	米袋计划	1		1
	坚果理财	0		1
	小诸葛平台	1		1

04 结果展示

结果展示

20	-	单走一个6	0.95147389	1	2019-11-17 09:10
21	-	天天喝可乐	0.95139849	0	2019-11-17 09:10
22	↓ 18	zgkxydx	0.95086545	1	2019-11-18 21:53
23	-	realpcy	0.95072275	0	2019-11-17 09:10

University of Chinese Academy of Sciences


中国科学院大学
University of Chinese Academy of Sciences
感谢！