融合领域知识图谱的跨境民族文本聚类方法

CCKS2022

陈春吉、毛存礼✉、 张勇丙 、黄于欣 、高盛祥、 郝鹏鹏
昆明理工大学
云南省人工智能重点实验室

背景

- 跨境民族文化文本间表达差异大，存在**相同实体不同表达**和**缺乏相关背景知识**的问题，如表1所示，导致跨境民族文本聚类困难。因此，对齐实体和融入相关领域知识能够改善聚类。
- 本文提出一种融合领域知识图谱的跨境民族文本聚类方法，使用领域知识图谱使实体对齐，使用构建的文档关联图增加相关背景知识。

表 1 数据示例

文本数据内容	文化类别
丢包，傣语称为“**端麻管**”，是傣族泼水节时民间娱乐的活动……	傣族泼水节
傣族情侣以对歌、踢毽、**丢包**等建立感情……	傣族婚姻
“**端麻管**”是傣族集娱乐与传情求爱于一体的活动……	傣族婚姻

研究方法

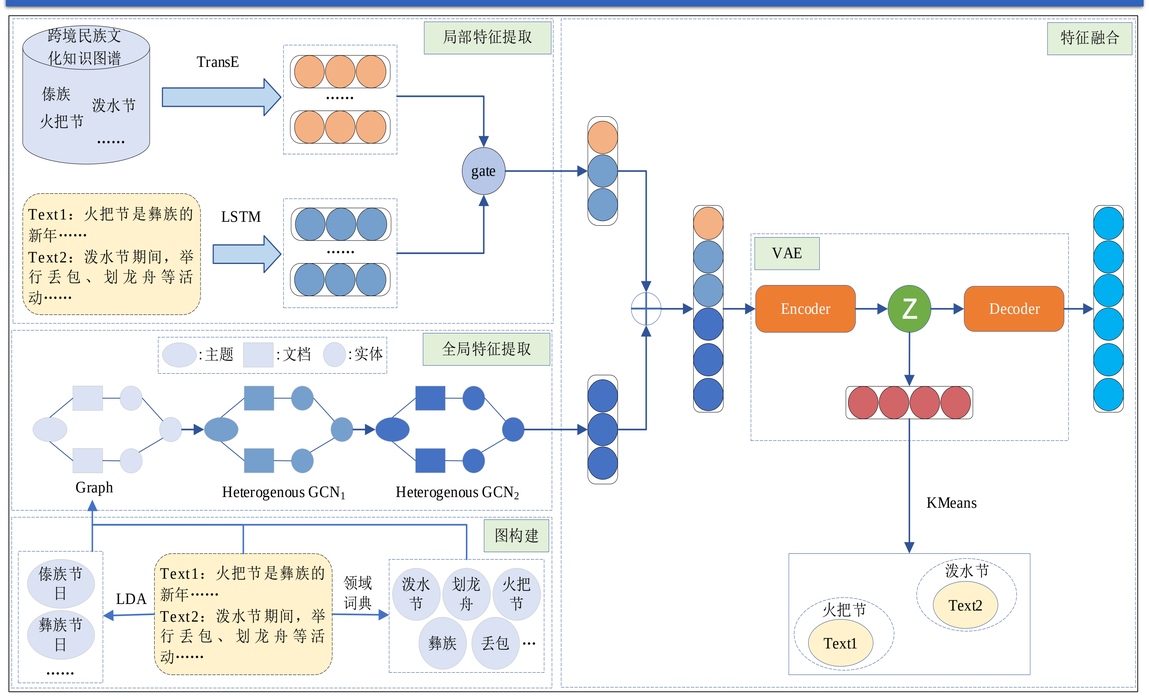

图1 模型架构图

- 文档关联图构建
  - 如图2所示，以文档、主题、实体为节点；文档与对应的概率最大的主题之间构建边（权重为对应主题概率）；文档与其中包含的实体之间构建边（权重为实体的TF-IDF值）。
- 局部特征提取
  - 通过门控机制将跨境民族知识图谱中的实体语义信息和跨境民族文本数据进行融合，将其作为局部特征。

$$s = g_s \odot s_d + (1 - g_s) \odot s_e$$

- 全局特征提取
  - 利用异构图神经网络训练构建好的跨境民族文化文档关联图，提取特征作为全局特征。

$$h^{(l+1)} = \sigma(\sum_{\tau \in T} B_\tau \bullet H_\tau^{(l)} \bullet W_\tau^{(l)})$$

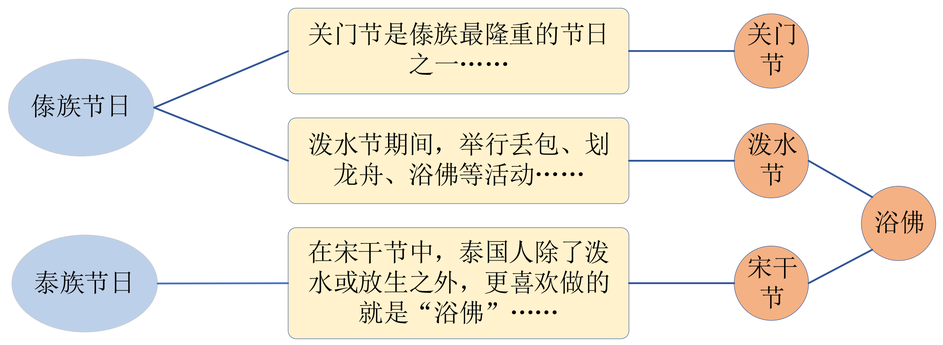

图2 文档关联图示例

- 文本聚类
  - 将全局特征与局部特征拼接，通过变分自编码器进一步提取特征，最后使用KMeans方法，得到聚类结果。

实验

- 数据集介绍

数据详情	数据量大小
类别数	15 个
训练数据规模	18002 条
平均字符长度	145 个字符
跨境民族文化词典规模	33977 个词

表2 实验数据详情

  - 从各个维基百科、含有民族文化相关内容等网站上获取傣族、泰族（泰国）、彝族、倮倮族（越南）文化相关的文本数据，共计15个类别18002条文本数据。
- 文本聚类实验

模型	Acc	NMI	ARI
KM	0.391	0.164	0.132
AE	0.435	0.348	0.237
DCN	0.508	0.379	0.223
SDCN	0.533	0.430	0.341
N2D	0.590	0.432	0.395
SCCL	0.603	0.501	0.411
Ours	**0.717**	**0.511**	**0.505**

表3 文本聚类实验结果

- 实验结果
  - 本文实验方法相比于SCCL等模型在聚类准确率指标上提高了11.4%，有效的提高了跨境民族文化文本聚类的效果。

结论

- 所提出方法融合领域知识图谱并对齐不同实体，有效提高了跨境民族文本聚类效果。
- 未来工作：我们将研究跨语言文本中跨境民族文化之间的关联关系。

通讯作者：毛存礼
(maocunli@163.com)

2022全国知识图谱与语义计算大会
China Conference on Knowledge Graph and Semantic Computing
秦皇岛 8.24-8.27