在不同领域的有效性. 首先使用 SimpleQuestion 数据集与地质领域 300 篇文献标注数据进行对基于字符的神经网络汉语 NER 进行实验识别. 同时我们使用两类数据集进行分类器验证训练, 一类是使用 THUCNews 数据, 每类 6500 条数据; 一类是使用实验室对于地质问答中用户常问问题及问答类型对应通用语句进行标注的数据, 每类平均大约 400 条, 共计 6500 条数据, 按照固定比例划分训练集、测试集、验证集.

实验中使用精确度、召回率和 $F1$ 作为验证评价指标, 对于整体多分类结果使用混淆矩阵.

$$\text{精确率}: P = \frac{TP}{TP + FP} \tag{24}$$

$$\text{召回率}: R = \frac{TP}{TP + FN} \tag{25}$$

$$F1\text{值}: F1 = \frac{2 * P * R}{P + R} \tag{26}$$

混淆矩阵:

$$P_{\text{macro}} = \frac{p_1 + p_2 + \cdots + p_3}{n} \tag{27}$$

$$R_{\text{macro}} = \frac{R_1 + R_2 + \cdots + R_n}{n} \tag{28}$$

4.2 实验过程

实验过程中, 使用 CPU 对实体识别与属性分类进行了训练. 实体识别部分针对双向神经网络使用字符嵌入大小为 100, 单词批量大小为 60, LSTM 单元为 100, 剪枝大小为 5.0, 训练学习速率为 0.001, 与防止过拟合的 dropout 大小为 0.5, 训练内容包括 97 万带 BIOES 标签标注的文本信息, 迭代次数循环 64 次, 直至损失变化幅度稳定结束. 属性分类过程中采用卷积核分别为 3、4、5、 256 个卷积核, 词向量维度为 64, 序列长度为 600, 全连接层为 128 个神经元, 词汇表大小为 500, 迭代总轮次为 10 轮, 每批训练大小 64, 学习率为 0.001, 及 dropout 大小 0.5. 实验结果采用精确率、召回率、$F1$ 值求算数平均值, 作为最后结果.

4.3 结果分析

在实体识别中, 使用地质标注数据集与进行验证, 使用基于模板匹配和基于网格的 LSTM+CRF 的神经网络验证得到结果如表 4.

表 4 基于网格 LSTM+CRF 命名实体识别结果

方法	精确率	召回率	$F1$值
基于模板匹配	0.76	0.68	0.64
基于LSTM+CRF	0.84	0.87	0.86

在用户属性分类中, 使用 THUCNews 数据集对其 10 个类别, 每类 6500 条数据采用基于字符的 CNN、RNN 模型实验结果如表 5、表 6 所示, 通过训练可以发现基于 CNN 的模型较基于 RNN 模型用时较短, 如表 7 所示.

表 5 基于 THUCNews 数据集的字符 RNN 分类模型训练结果

类别	精确率	召回率	$F1$值
体育	0.97	0.99	0.98
财经	0.96	0.98	0.97
房产	0.99	0.99	0.99
家居	0.96	0.82	0.88
教育	0.91	0.94	0.92
科技	0.93	0.98	0.95
时尚	0.94	0.94	0.94
时政	0.96	0.91	0.93
游戏	0.97	0.95	0.96
娱乐	0.92	0.97	0.94

表 6 基于 THUCNews 数据集的字符 CNN 分类模型训练结果

类别	精确率	召回率	$F1$值
体育	1.00	0.99	0.99
财经	0.96	0.98	0.97
房产	1.00	1.00	1.00
家居	0.98	0.84	0.91
教育	0.94	0.98	0.96
科技	0.93	0.98	0.95
时尚	0.93	0.98	0.96
时政	0.95	0.94	0.95
游戏	0.98	0.98	0.98
娱乐	0.98	0.97	0.98

表 7 基于 THUCNews 数据集的字符 CNN 与 RNN 模型对比

方法	平均精确率	运行耗时
基于字符CNN	0.9633	11分15秒
基于字符RNN	0.9545	2时10分37秒

在 THUCNews 的基础上我们可以知基于字符的 CNN 模型不仅运行时间为基于字符的 RNN 模型的 1/13, 且在数据集上得到 96.3% 的精确率, 由此我们使用基于字符的 CNN 模型在我们针对用户一般询问语句人工标注的地质问答数据得到如表 8 所示, 平均精确率达到 96.9%, 使得应用效果超过基线模型.

表 8 基于地质标注数据集的字符 CNN 分类模型结果

方法	平均精确率	运行耗时
基于字符CNN	0.9691	3分44秒
基于字符RNN	0.9556	2时20分20秒

5 结论与展望

本文在地质领域用户意图识别中通过构建地质领域的实体字典, 来源包括地质百科大辞典、搜狗语料等, 在基于字符的网格神经网络上进行专家及用户的询问语句实体识别训练, 采用的是地质文献数据, 在验证集上验证, 采用 Adam 随机梯度下降时, 准确率达到 84.57%、召回率达到 87.12%, *F*1 值更是达到 86.18%, 超过了基于模板匹配与基于 RNN 的现有模型, 可有效地识别特定领域的实体及关系. 同时在短文本信息分类过程中借鉴卷积网络考虑语义信息的优势, 采用基于字符的分类模型达到 96.9% 的精确率, 对于分类结果使用知识图谱分类属性映射得到匹配的知识描述返回用户, 整体实现了在基于地质领域的问答过程中意图识别.

在此基础上, 将来的工作更多的是将用户热点问题及知识意图推理进行深入探索, 通过接下来的实验, 将知识图谱中知识的构建环节引入知识阶层路径, 实现用户复杂文本信息意图的识别.

参考文献

1 罗成, 刘奕群, 张敏, 等. 基于用户意图识别的查询推荐研究. 中文信息学报, 2014, 28(1): 64–72. [doi: 10.3969/j.issn.1003-0077.2014.01.009]

2 赵乐, 张兴旺. 面向 LDA 主题模型的文本分类研究进展与趋势. 计算机系统应用, 2018, 27(8): 10–18. [doi: 10.15888/j.cnki.csa.006456]

3 Li X. Understanding the Semantic Structure of Noun Phrase Queries. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA. 2010. 1337–1345.

4 Ramanand J, Bhavsar K, Pedanekar N. Wishful thinking: Finding suggestions and 'buy' wishes from product reviews. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Stroudsburg, PA, USA. 2010. 54–61.

5 Song HJ, Park SB. Identifying intention posts in discussion forums using multi-instance learning and multiple sources transfer learning. Soft Computing, 2018, 22(24): 8107–8118. [doi: 10.1007/s00500-017-2755-8]

6 孙镇, 王惠临. 命名实体识别研究进展综述. 现代图书情报技术, 2010, (6): 42–47.

7 Florian R, Ittycheriah A, Jing HY, *et al*. Named entity recognition through classifier combination. Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003. Stroudsburg, PA, USA. 2003. 168–171.

8 Borthwick A, Sterling J, Agichtein E, *et al*. NYU: Description of the MENE named entity system as used in MUC-7. Proceedings of the 7th Message Understanding Conference. Fairfax, VA, USA. 1998. 145–150.

9 Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition. Proceedings of the 19th International Conference on Computational linguistics. Stroudsburg, PA, USA. 2002. 1–7.

10 Zhou GD, Su J. Named entity recognition using an HMM-based chunk tagger. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA. 2002. 473–480.

11 张硕果, 汪成亮. 结合 CRFs 的词典分词法. 计算机系统应用, 2010, 19(11): 115–118. [doi: 10.3969/j.issn.1003-3254.2010.11.026]

12 何炎祥, 罗楚威, 胡彬尧. 基于 CRF 和规则相结合的地理命名实体识别方法. 计算机应用与软件, 2015, 32(1): 179–185, 202. [doi: 10.3969/j.issn.1000-386x.2015.01.046]

13 Hu ZT, Ma XZ, Liu Z, *et al*. Harnessing deep neural networks with logic rules. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. 2016. 2410–2420.

14 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv: 1505.01991.

15 Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, CA, USA. 2016. 260–270.

16 杨文明, 褚伟杰. 在线医疗问答文本的命名实体识别. 计算机系统应用, 2019, 28(2): 8–14. [doi: 10.15888/j.cnki.csa.006760]

17 胡泽文, 王效岳, 白如江. 国内外文本分类研究计量分析与综述. 图书情报工作, 2011, 55(6): 78–81, 142.

18 薛春香, 张玉芳. 面向新闻领域的中文文本分类研究综述. 图书情报工作, 2013, 57(14): 134–139. [doi: 10.7536/j.issn.0252-3116.2013.14.022]

19 Yang J, Liang SL, Zhang Y. Design challenges and misconceptions in neural sequence labeling. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, NM, USA. 2018. 3879–3889.

20 Bordes A, Usunier N, Chopra S, *et al*. Large-scale simple question answering with memory networks. arXiv preprint arXiv: 1506.02075.

21 Golub D, He XD. Character-level question answering with attention. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA. 2016. 1598–1607.

22 Yin WP, Yu M, Xiang B, *et al*. Simple question answering by attentive convolutional neural network. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan. 2016. 1746–1756.

23 Lukovnikov D, Fischer A, Lehmann J, *et al*. Neural network-based question answering over knowledge graphs on word and character level. Proceedings of the 26th International Conference on World Wide Web. Perth, Australia. 2017. 1211–1220.

24 Zhang Y, Yang J. Chinese NER using lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, VIC, Australia. 2018. 1554–1564.

25 Kim Y. Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1746–1751.