全连接层。 我们利用全连接层将模型的输出映射为相应意图类别数量，接着使用softmax输出各个意图类别的概率，最终采用概率最高的意图类别作为输出。公式如式(28)-(29)所示。

$$y^* = softmax(W * h^* + b) \tag{28}$$

$$predict = argmax(y^*) \tag{29}$$

3.3 伪标签的挑选

对于单句话语意图模型所带来的意图伪标签，并不是所有的伪标签都能对人机对话段的意图分类产生正面效果。为此，本文的方案中，通过全部伪标签的初步试验，计算出每个伪标签在对话段中的平均覆盖率:

$$pesudo_{cover} = \text{意图伪标签在对话段中的出现次数/对话段的对话轮次} \tag{30}$$

然后，筛选出覆盖率高的伪标签:

$$pseudo_{select} = topk(pesudo_{cover}) \tag{31}$$

在覆盖率分析的基础上，我们进一步从语义角度挑选出高关联度的伪标签嵌入到模型中，去掉了部分关联度不高的标签。在本文的实验中，用以训练单句话语意图识别模型的数据集共有48种单句话语意图类别。经过伪标签挑选，我们的最终方案中，选取了25种与人机对话意图分类任务高关联度的单句话语意图，同时将其它的单句话语意图标签标记为“其它”。不同伪标签集的实验效果见4.5小节。

4 实验

4.1 数据集

本文的人机对话实验数据来自于中国中文信息学会主办的“客服领域用户意图分类评测比赛”[1]，属于客服领域对话文本，视为多轮话的长文本。同时这项比赛还有一个子任务是单句用户话语的自然语言理解(包括意图识别和槽填充)。本文将其中的话语意图识别任务视为本文人机对话任务的辅助任务，用于产生人机对话段中每一句单句话语的意图伪标签。

人机对话数据集人机对话数据集为2万条真实客服对话段标注数据，此数据集中共有35种人机对话意图类别，表2给出了业务类型与用户意图的种类。我们按照8：2的比例分别划分训练集和测试集。训练时，再从训练集中划分出20%作为验证集。

单句话语意图识别数据集 单句话语的数据集为2万条真实的单句话语数据，此数据集中共有48种单句意图类别，如表3所示。经过伪标签优选，选取了与对话段意图分类任务相关度高的25种单句意图类别，其它类别的伪标签标记为“其它”。按照8：2切分训练集和验证集，不设置测试集。

业务类型	对话意图
咨询(含查询)	业务订购信息查询、业务规定、业务订购信息查询、业务资费、产品/业务功能、使用方式、办理方式、号码状态、宽带覆盖范围、工单处理结果、工单处理结果、服务渠道信息、用户资料、电商货品信息、营销活动信息、账户信息
办理	下载/设置、停复机、取消、变更、开通、打印/邮寄、移机/装机/拆机、缴费、补换卡、重置/修改/补发、销户/重开
投诉(含抱怨)	不知情定制问题、业务使用问题、业务办理问题、业务规定不满、信息安全问题、服务问题、网络问题、营销问题、费用问题

Table 2: 业务类型种类与35种对话意图类别

[1] http://www.cips-cl.org/static/CCL2018/call-evaluation.html

单句话语意图标签
查询、查询套餐余量、查询充值缴费记录、查询流量、查询本机号码、查询本机业务、查询余额、查询短信、查询积分、查询语音、查询月初扣费、查询账单、查询宽带、查询手机、咨询宽带、具实帮助、具实返回、具实转人工、具实退出、具实重听、具实业务列表、具实转ivr、办理套餐、办理手机充值、取消流量、预约宽带、修改宽带、开通流量、重置、重置手机、重置宽带、GPRS、拒识、修改、确认、手机、修改手机、短信、宽带、流量、empty、集外说法批评,抱怨,脏话、集外说法机器人、集外说法集外业务、咨询、集外说法结束、集外说法短拒识、集外说法友好问候、集外说法感谢

Table 3: 48种单句话语意图类别

4.2 实验设置

人机对话意图分类实验 batch_size设置为16，epoch设置为20，Dropout设置为0.1，采用Adam优化器，学习率为1e-5，非层级模型设置最大句子长度为512，层级模型设置最大句子数量为25，最大句长为25。

话语伪标签预测实验 batch_size设置为32，epoch设置为100，Dropout设置为0.1，采用Adam优化器，学习率为1e-3，非层级模型设置最大句子长度为30。

4.3 对比方法

本文提出的PLA-HAN模型将与以下代表性的基线方案进行比较，为了公平比较，全部模型都采用了BERT编码：

- BERT FineTune：该方法将BERT fine-tuning (Devlin et al., 2019)应用到分类任务，在BERT分类层增加了一个新的输出层。
- BERT BiLSTM：该方法是文本分类的经典基线 (Vu et al., 2016)，适合于序列问题。
- BERT SoftAtt：该方法采用了Liu等(Liu and Lane, 2016)在ATIS数据集的话语意图识别的BILSTM模型中采用的软注意力。
- BERT HAN：Yang等(Yang et al., 2016)提出的更加适合篇章结构文本的层次注意力模型结构。

4.4 整体性能

我们提出的PLA-HAN模型与几种对比方法进行比较，包括BERT FineTune、BiLSTM、带软注意力的BiLSTM以及层次注意力模型HAN。实验结果如表4所示。

模型类别	模型	意图分类正确率(%)			
		总体	咨询(含查询)	办理	投诉
长文本结构	BERT FineTune	53.11	49.50	66.72	48.14
	BERT BiLSTM	55.75	52.39	69.83	49.35
	BERT SoftAtt	55.87	51.98	71.03	49.91
层次结构	BERT HAN	56.31	52.55	71.08	50.40
	BERT PLA-HAN	**56.94**	**53.33**	**71.26**	**51.35**

Table 4: 不同模型性能对比

从表4的结果可以看到：

(1) 在以整篇长文本直接作为输入的模型中，通过加入编码器BiLSTM和注意力机制的应用，可以在BERT的基础上进一步提升对话段意图分类的性能。

(2) 相比于整篇长文本结构的输入，层次结构模型取得了更好的分类性能。这一方面得益于不受BERT输入长度的限制，另一方面也由于分层次的注意力能更好地建模字到句子再到对话段的语义结构。

(3) 融合了单句话语伪标签注意力的PLA-HAN取得了最好的性能，优于HAN模型。

4.5 进一步分析

通过比较以上结果可以看出PLA-HAN模型取得了良好的性能，我们也想进一步探究模型能有所提升的原因。我们首先分析了不同的伪标签集对模型性能的影响。然后，我们给出了一个不同长度对话段情况下的PLA-HAN模型与基础的HAN模型的性能对比的定量分析。

不同伪标签集对模型性能的影响。为了研究不同伪标签集在PLA-HAN模型中效果，我们对采用三种不同的伪标签集的BERT PLA-HAN的实验结果进行观察。PLA-HAN(All)代表不做选择采用了全部的48种伪标签，PLA-HAN(Fit)代表只选用了和人机对话意图相关的32种伪标签，PLA-HAN(Select)代表进一步优选的25种伪标签。结果如表5所示。

模型	意图分类正确率(%)			
	总体	咨询(含查询)	办理	投诉
BERT PLA-HAN(All)	56.06	52.48	70.41	49.95
BERT PLA-HAN(Fit)	56.59	53.31	70.94	50.72
BERT PLA-HAN(Select)	**56.94**	**53.33**	**71.26**	**51.33**

Table 5: 不同伪标签集的PLA-HAN模型性能对比

结果表明，我们的伪标签集选择策略是必要的，不同的伪标签集对模型整体性能存在明显的影响。详细的分析如下：

- PLA-HAN(All)：在此方案中，我们采用了所有48种伪标签。其中部分伪标签实际上和人机对话意图分类任务中的35种意图的相关性并不强。从实验结果可以看到，分类性能都不够好，甚至都略微低于HAN模型。

- PLA-HAN(Fit)：在此方案中，我们选取与人机对话任务的意图有较高覆盖率的伪标签嵌入到模型中，将基本不相关的标签标记为“其它”。实验中通过验证集选取了32个伪标签。从实验结果可以看到，分类性能相比于PLA-HAN(All)有了明显提高，相同时也优于HAN模型。

- PLA-HAN(Select)：在此方案中，我们在PLA-HAN(Fit)的基础上，进一步从语义角度挑选出高关联度的伪标签嵌入到模型中，去掉了部分关联度不高的标签，一共选取了25种伪标签。从实验结果看，PLA-HAN(Select)取得了最好的性能。

不同长度对话段的模型性能对比。我们进一步对比不同对话段长度情况下的PLA-HAN模型与基础的HAN模型的性能。我们按照对话段的长度分为长(600字以上)、中(301-600字)和短(300字以下)三类进行观察，结果如图2所示。

从图2可以看到：

(1) 长的对话段的意图分类存在较大的挑战，正确率明显低于短的对话段。

(2) 我们的PLA-HAN，在伪标签注意力的帮助下，在不同长度的对话段性能均优于HAN。尤其是长的对话段(超过600字)，意图分类正确率提升较为显著，达到1.56%。

5 结束语

针对现有文本分类方法在人机对话意图分类上存在的挑战，本文提出了一种结合话语伪标签注意力的层次注意力网络模型PLA-HAN。PLA-HAN通过优选伪标签集，设计和计算单句话语意图伪标签注意力，并将其嵌入到HAN的层级结构中，与HAN中的句子级别注意力相融合，提升了人机对话意图分类性能。我们在中国中文信息学会主办的“客服领域用户意图分类评测比赛”的评测语料上进行实验，实验结果证明PLA-HAN模型取得了优于HAN等研究进展文本分类方法的意图分类正确率。

致谢

本文受到国家自然科学基金(项目编号:71472068)的资助。

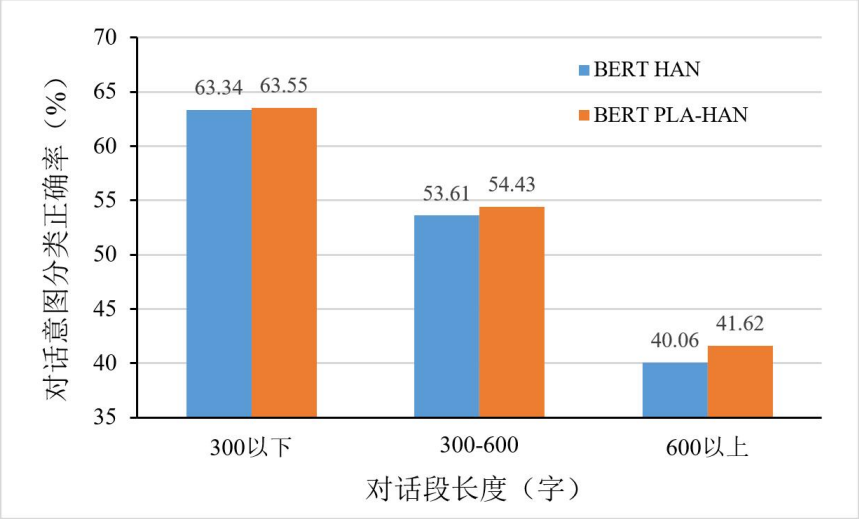

Figure 2: 不同长度对话段的模型性能对比

参考文献

安明慧, 沈忱林, 李寿山, 等. 2019. 基于联合学习的问答情感分类方法. 中文信息学报, 33(10):119-126.

柯子, 黄沛杰, 曾真. 2018. 基于优化“未定义”类话语检测的话语领域分类. 中文信息学报, 32(4):105-113.

俞凯, 陈露, 陈博, 等. 2015. 任务型人机对话系统中的认知技术——概念、进展及其未来. 计算机学报, 38 (12):2333-2348.

Chen, Hongshen and Liu, Xiaorui and Yin, Dawei and Tang, Jiliang. 2017. *A survey on dialogue systems: Recent advances and new frontiers*, volume 19. ACM New York, NY, USA.

J. P. Cheng, L. Dong, and M. Lapata. 2016. Long short-term memory-networks for machine reading. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pp. 551–561.

J. Devlin, M. Chang, K. Lee, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pp. 4171–4186.

A. Graves, N.Jaitly and A. Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. *Proceedings of the 2013 IEEE workshop on automatic speech recognition and understanding (ASRU 2013)*, 273–278.

P. Haffner, G. Tur, and J. H. Wright. 2003. Optimizing SVMs for complex call classification. *Proceedings of the 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, pp. 632-635.

B. Liu and T. Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *Proceedings of the 17th Annual Conference of the Inter-national Speech Communication Association (INTERSPEECH 2016)*, pp. 685-689.

Y. Kim. 2014. Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1746–1751.

Y. Kim, D. Kim, A. Kumar. 2018. Efficient large-scale neural domain classification with personalized attention. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 2214–2224.

X .H. Phan, L. M. Nguyen, S. Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from largescale data collections. *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, pp. 91-100.

S. Ravuri and A. Stolcke. 2016. A comparative study of recurrent neural network models for lexical domain classification. *Proceedings of the 41th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016)*, pp. 6075-6079.

R. Sarikaya, G. E. Hinton, and B. Ramabhadran. 2011. Deep belief nets for natural language call-routing. *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pp. 5680–5683.

N. T. Vu, P. Gupta, H. Adel, et al. 2016. Bi-directional recurrent neural network with ranking loss for spoken language understanding. *Proceedings of the 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, pp. 6060–6064.

P. Y. Xu and R. Sarikaya. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013)*, pp. 78–83.

P. Y. Xu and R. Sarikaya. 2014. Contextual domain classification in spoken language understanding systems using recurrent neural network. *Proceedings of the 39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, pp. 136–140.

K. Xu, J. Ba and R. Kiros,et al. 2015. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the 31st International Conference on Machine Learning (ICML 2015)*, pp. 2048–2057.

Z. C. Yang, D. Y. Yang, C. Dyer, et al. 2016. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, pp. 1480-1489.