scau_SIGSDS 中文人机对话技术用户意图领域分类任务评测报告

唐杰聪，梁泳诗，闫江月，李杨辉，凌大未，曾真，杜泽峰，黄沛杰

（华南农业大学数学与信息学院，广东 广州 510642）

摘要：本文介绍华南农业大学口语对话系统研究室（scau_SIGSDS）开发的话语领域分类系统。我们的系统结合了基于领域关键词的领域识别和多分类器的领域分类。分类模型采用了长短期记忆网络（LSTM），领域关键词表的构建采用了基于数据的提取算法。在开放式评测中，我们对领域关键词表进行了人工扩展，并结合外部信息构建了若干领域知识表，进一步优化话语领域分类结果。我们参加了 2017 中文人机对话技术评测（ECDT）中的用户意图领域分类任务，在封闭式和开放式两项评测中均取得了第一名。

关键词：话语领域分类；人机对话；LSTM；封闭式；开放式

1 引言

近年来，人机对话技术，也称为口语对话系统（spoken dialogue system，SDS）受到了学术界和产业界的广泛关注。学术上，人机对话是人机交互最自然的方式之一，其发展影响及推动着语音识别与合成、口语语言理解、对话管理以及自然语言生成等研究的进展；产业上，众多产业界巨头相继推出了人机对话技术相关产品，如个人事务助理、娱乐型聊天机器人等。以上极大地推动了人机对话技术在学术界和产业界的发展。在第六届全国社会媒体处理大会（SMP 2017）上，哈尔滨工业大学和科大讯飞股份有限公司组织了SMP 2017中文人机对话技术评测（ECDT），为人机对话技术相关的研发人员提供了一个良好的沟通平台。

口语语言理解（spoken language understanding, SLU）是SDS中的重要环节，而话语领域分类（domain classification）则是SLU的关键任务之一[1]。话语领域分类的任务是把话语划分到定义好的不同领域标签[2]，进而将话语正确的分进不同的SLU子系统。如用户提出“帮我写一封邮件”，系统则应该将其划分到“email”领域之中，对该话语进行专门针对“邮件”领域的语言理解。由于口语对话具有长度短小的特点，领域分类通常会被看作是短文本分类。早期的领域分类多采用较为复杂的人工特征，如语法信息、韵律信息、词汇信息等 [3-4]，分类模型采用传统的统计学习模型，如随机森林、隐马尔科夫、条件随机场等。深度学习流行以来，许多研究者开始用深度学习方法解决自然语言处理（natural language processing, NLP）任务，许多任务得到了长足的发展，也包括了领域分类[2, 5-6]。代表性的模型包括了深度置信网络（deep belief network, DBN）、卷积神经网络（convolutional neural networks, CNN）和长短期记忆网络（long and short-term memory, LSTM）等。

本文介绍我们参加 ECDT 评测中用户意图领域分类任务的系统，采用了 LSTM 分类模型，并针对样本训练集数量有限的特点，采用基于数据的领域关键词提取算法以及构建外部领域知识表（在开放式评测中），进一步优化领域分类效果。

2 scau_SIGSDS 话语领域分类系统

2.1 总体技术架构

图 1 是本文提出的方法的总体技术架构。在这个架构中，针对封闭性和开放式的评测，分别分为两个和三个阶段：

（1）基于知识的领域识别。对于封闭式，采用了基于数据的领域关键词提取，得到足够支持率和置信度的领域关键词，用于领域识别；对于开放式，则进一步结合人工知识增加领域关键词表，并对若干个合适的领域构建了领域知识表，如疾病列表，直接识别到对应的领域，或者判定为有限的几个候选领域。

（2）31 领域分类器。对于封闭式和开放式，都通过模型和参数优化，选择和进行 31 分类的多分类器。在开发阶段，采用的是训练集进行 k 折交叉验证；而用于测试模型，则是通过训练集加上开发集一起通过交叉验证训练得到。

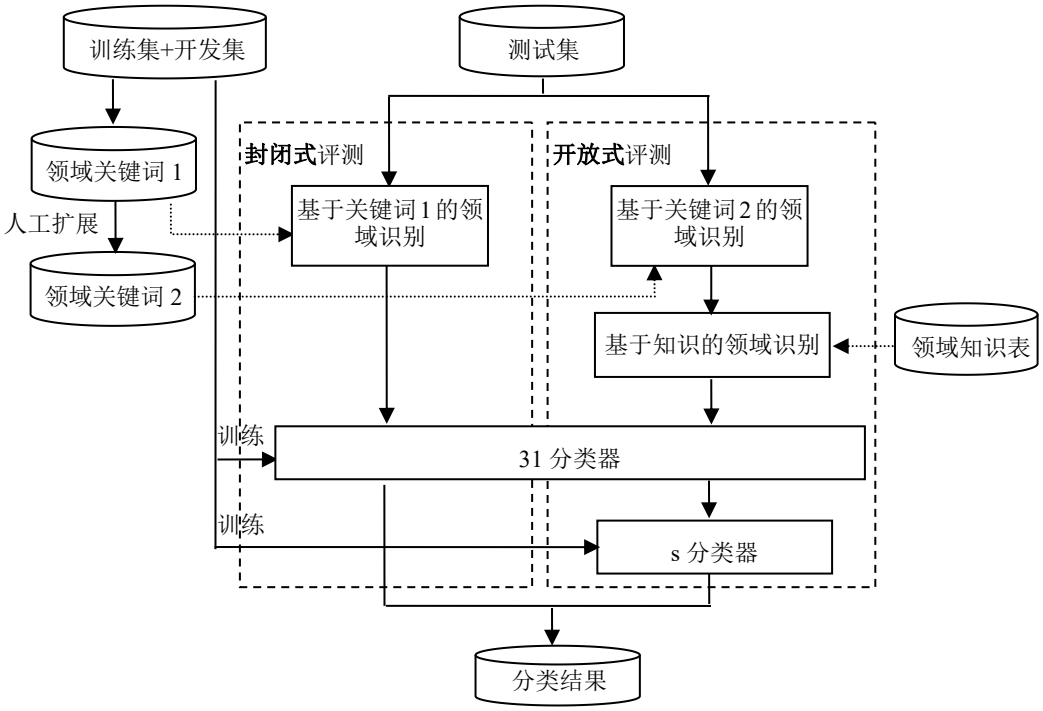

图 1 总体技术架构

（3）s 分类器的结果修正。在开放式测试中，对于在第（1）步中判定为若干个候选领域的样例，则进一步通过训练得到的 s 分类器进行再次分类。s 分类器可以有多个。

2.2 基于领域关键词的领域识别

对于封闭式评测，我们采用了基于数据的领域关键词提取算法，通过对训练集和开发集进行统计，抽取足够置信度和支持率的 2 和 3 字“词”构成的领域关键词表。本系统采用的是置信度=0.95，支持率阈值根据领域类别样本数量分了 0.10、0.15 和 0.18 三个等级。并根据包含关系精简了领域关键词列表。例如，“七乐”和“七乐彩”都在抽取的到的另一关键词集时，保留了“七乐”。这样做可预期具有更好的识别能力，也在一定程度上提高了误判的风险。对于开放式评测，我们进一步结合人工知识扩展了领域关键词表。

2.3 基于 LSTM 的领域分类

相比于CNN，循环神经网络(recurrent neural networks, RNN)有利于学习到句子中字词间的长距离依赖关系，但存在梯度消失/发散问题。目前常用的是RNN的一些变体，如LSTM、GRU(gated recurrent unit)等，他们通过门控机制很大程度上缓解了RNN的梯度消失问题，并防止梯度发散。经典的LSTM模型整体结构如图2所示，LSTM记忆单元细节参阅文献[7]。

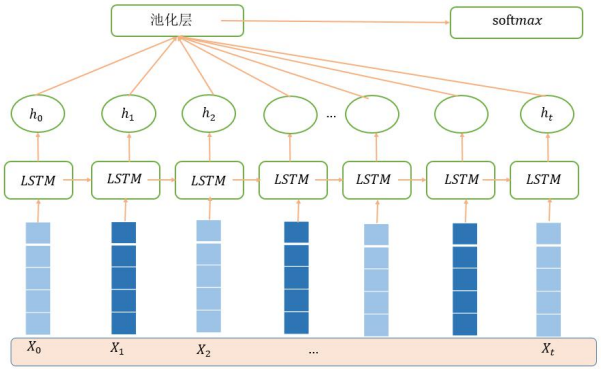

图 2 LSTM 模型

本文在给定数据集上验证了不同RNN变体的领域分类效果，主要包括普通的LSTM、GRU以及带隐层的LSTM。

2.4 基于领域知识表的领域识别

通过分析领域话语特点，我们针对 4 个领域（health、radio、epg 和 tvchannel），结合外部信息构建了 3 个领域知识表（health、radio 和 tv）。其中 tv 知识表用于识别为候选的 epg 或 tvchannel 领域，需要进一步结合 epg-tvchannel 二分类器进行识别。

3 开发与评测

3.1 数据集及任务介绍

SMP 2017意图领域分类任务的数据集共计31个类别，每个类别数据单独为一个文件，具体包括聊天类（chat）和垂类（30个垂直领域）。数据集总体情况如表1所示。

表 1 训练验证以及测试语料的情况

训练集	开发集	测试集
2299	770	681

比赛任务根据是否仅允许使用主办方提供的评测数据进行训练和开发分为封闭式和开放式两项。评价指标采用了 F 值。

3.2 实验结果与分析

3.2.1 验证与开发

为了方便验证和开发采用了正确率的评价指标。我们先用训练集（Train）10 折交叉验证（调节学习率、层数、节点数、卷积核大小、dropout 系数等超参数）进行分类模型和词向量的选择，如表 2 和表 3 所示。LSTM 取得了优于 CNN 的性能，10G 微博数据训练的词向量并结合 Hownet 和同义词词林扩展版进行修正的词语向量表达取得了最好的效果。

表 2 CNN 和 LSTM 的分类效果对比

模型	分类正确率(%)
CNN	88.64
LSTM	91.38

表 3 不同词语向量表达的分类效果对比(分类模型采用 LSTM)

词语向量表达	分类正确率(%)
1.5G微博	91.38
10G微博	91.56
10G微博修正	91.73

我们进一步用训练集（Train）10 折交叉验证对不同的 RNN 变体进行选择，结果如表 4 所示。选择了带隐层的 LSTM 模型。

表 4 不同 RNN 模型的分类效果对比

模型	分类正确率(%)
LSTM	91.73
GRU	91.91
LSTM+隐层	92.04

最后用开发集（Dev）进行预测方式的选择，以及领域关键词和领域知识表的检验，分别如表 5 和 6 所示。采用最佳 10 折交叉验证的 10 个 9 折数据模型进行集成投票的预测方式取得了最好的开发集测试效果。而领域关键词和领域知识表也都进一步提高了分类正确率。

表 5 不同预测方式的分类效果对比

预测方式	分类正确率(%)
最佳9折模型	91.44
10折重训练模型	91.05
集成预测	92.22

表 6 领域关键词和知识表的修正效果

方法	分类正确率(%)
“LSTM+隐层”模型集成预测	92.22
领域关键词(close)修正	92.74
领域关键词(open)修正	92.87
领域知识表(open)进一步修正	93.00

3.3 评测

在最后的评测测试中，我们采用训练集和开发集，重新构建领域关键词表，并采用 10 折交叉验证训练了 LSTM 分类器。表 7 给出了评测结果中单项（封闭式和开放式）前三名的参赛系统以及前十名平均值评测结果。

表 7 评测结果

模型	封闭性评测 F值	开放性评测 F值
scau_SIGSDS	0.9391.	0.9414
义语智能	0.9288	0.9288
SXU_JK	0.9089	0.9123
自动化所-出门问问联合实验室	-	0.9258
Top 10 平均值	0.8993	0.8995

我们的系统取得了封闭式和开放式两项第一名，比第二名的系统的领域分类性能分别高了 1.1%和 1.4%，比 Top10 平均值的领域分类性能高了 4.4%和 4.7%。

4 结束语

本文介绍了我们参加 2017 中文人机对话技术评测（ECDT）中的用户意图领域分类任务的参赛系统的技术方案和评测情况。我们的系统在封闭式和开放式两项评测中均取得了第一名。由于比赛时间有限，开放式版本中，我们仅仅针对少数领域进行了优化，性能提升不够明显。另一方面，随着领域语料的进一步扩展，将有利于进一步提升话语领域分类性能。同时，也有机会揭示出更多的技术挑战。

参考文献

[1] Tur G, Deng L, Hakkani-Tür D, et al. Towards deeper understanding: Deep convex networks for semantic utterance classification[C]// Proceedings of the 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), 2012:5045-5048.

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

[3] Haffner P, Tur G, Wright J H. Optimizing SVMs for complex call classification[C]// Proceedings of the 28th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), 2003: I-632-I-635.

[4] Chelba C, Mahajan M, Acero A. Speech utterance classification[C]// Proceedings of the 28th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), 2003:I-280-I-283.

[5] Sarikaya R, Hinton G E, Deoras A. Application of deep belief networks for natural language understanding[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2014, 22(4):778-784.

[6] Ravuri S, Stoicke A. A comparative study of neural network models for lexical intent classification[C]// Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2015), 2015: 368-374.

[7] Hochreiter S. and Schmidhuber J. Long Short-Term Memory [J]. Neural Computation , 1997, 9(8):1735–1780.