华南农业大学

硕士研究生中期考核与学位论文开题报告

学　　号: 2015307809

姓　　名: 王俊东

学　　院: 数学与信息学院

专业(领域): 计算机技术

研 究 方 向 :自然语言处理

导 师 姓 名: 黄沛杰

攻 读 学 位: 工程专业学位硕士

论文题目	口语对话系统中的对话行为分类研究		
选题来源	省（自治区、直辖市）项目	论文类型	基础研究
开题日期	2016-09-23	涉密	否

一、立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）

1. 研究意义

口语对话系统指的是通过自然语言和人交流的计算机系统，主要研究如何能让计算机理解并生成人们日常所使用的语言，对人向计算机提出的问题，通过对话的方式，用自然语言进行回答。当前，自然语言对话系统已广泛应用于信息查询系统[1-5]、导航系统[6-7]、导游系统[8]和导购系统[9]等自然语言智能助理。口语语言理解（spoken language understand，SLU）是口语对话系统中不可缺少的一部分，是对话系统能提供稳定服务的基础，类似于人的听觉系统，负责将口语化的用户话语转为机器能理解的语言[10]。

口语对话系统中的口语语言属于典型的短文本，其对话行为（dialogue act, DA）识别是SLU 的关键任务。随着互联网的高速发展以及大数据的出现，每天均有成千上万的短文本数据产生，如搜索查询、广告关键词、新闻标题、标签数据、电影评论、微博、微信等[11]。相比传统的文档，短文本有以下特点：

（1）语法更加自由宽松，一方面意味着可利用的语法信息更加缺乏，进一步也导致很多传统的自然语言处理（natural language processing, NLP）工具不能很好的应用到短文本中来，如语法树等；另一方面也意味着文本的样式、语义等变化更复杂，噪音更多。

（2）更加稀疏的上下文，如搜狗实验室提供的中文新闻标题分类数据集，大部分文本数据长度集中在 10-21 字之间[12]，而在限定领域口语对话系统中的超出领域（out-of-domain，OOD）话语更短，话语平均长度只有 3.6 字，集中在 1-8 个字之间[13]，在搜索引擎中，绝大部分的查询语句也少于 5 个词[11]。有限的上下文信息让短文本的特征提取更加困难，理解更加歧义和不明确。

短文本分类（Short Text Categorization）是将短文本归到预定义标签的任务，作为很多应用的第一步骤，如果能够使机器像人一样的理解短文本，将会为很多领域带来巨大的效益，同时也让人们生活变得更加的方便和智能，如垃圾邮件检测[14]、主题标签[15-16]、语义分类[15-19]、DA 识别[20-21]、搜索引擎[11]、自动问答系统、在线广告、个性推荐、情感分析、聊天机器人[13]等。

传统的文本方法主要使用 Bag Of Words 或者 Bag of N-Gram[22]作为特征，用支持向量机（support vector machine，SVM）[23]、逻辑斯蒂回归（logistical regression，LR）、最大熵[12]等线性模型或核技巧作为分类器，但这类方法有个明显的缺陷就是忽略了词序信息，理论上，虽然可以通过增加 N-gram 的长度来缓解，但效果并不明显[24]，且导致特征特别稀疏，在短文本上更加突出。另外一些研究则采用在领域相关的无标签数据集上使用 LDA（latent dirichlet allocation）获取主题特征[25]或者使用神经网络（neural network）训练词向量[12]的方法增加语义特征。近来，不少工作开始探索基于人工神经网路的层次学习方法，如卷积神经网路（convolutional neural networks，CNN）[15,18-20,26-28]和长短期记忆人工神经网络（long-short term memory，LSTM）[20,29]等，这类方法有一定的结构学习能力，通过组合底层特征，可以更好的捕获句子的词序等结构特征。

对话系统中，对话话语作为短文本的类型之一，具有常见短文本的通用特性，但相比微博、评论、新闻标题等短文本信息挑战更大，体现在以下几点：

（1）对话过程中，用户的随意性更大，短文本所固有的噪音多和特征稀疏在口语对话话语中表现得更为突出。另外，口语对话中的话语比电影评论和新闻标题等短文本更为口语化，比微博也多了一些口语化省略的情况。

（2）相比领域内话语，OOD 口语化话语语义更为开放和表达多样，尤其是中文，容易产生超出字典的字词，并且缺少话语携带的相对较为丰富的语义或语法等文本信息，同对话段中的上下文的关联更加稀疏和复杂。

另外，对话语料、中文短文本等语料库相当缺乏，这也是本项目面临的挑战之一。

本课题主要以限定领域的中文口语对话系统为应用背景，领域场景为信息咨询或商品导购。比如，在网上旅游咨询越来越流行的情况下，用户已经不仅仅满足于简单对旅游景点及交通住宿信息的搜索，而希望得到更为智能的出行决策和行程规划的服务，甚至是真人专业的服务。然而 24 小时的在线咨询毕竟很耗费人力，并且，也不可能提供大规模的并行服务。如果能实现一个旅游规划机器人——智能旅游代理，像人一样的理解用户的话语和需求，并进一步为用户提供服务，那么将会大大的满足用户的服务需求。基于实际应用的需求和短文本分类的现状，本课题将结合自然语言处理和机器学习等技术，基于层次学习和领域内外话语处理相结合的基本架构，进一步融合多种不同层次上下文信息的异构特征，从句子内部的理解扩展到不同层次上下文中短文本的理解，并从单标签转移到多标签识别，以改善中文口语对话系统的 DA 识别效果，进而提升对完整对话和用户需求的理解。

2. 国内外研究现状

为了克服短文本具有的噪音多、特征稀疏和主题不明确等特点，基于 Bag of Words、SVM 等传统方法已经不能满足短文本分类的需求，短文本分类方法的研究主要集中在以下几点：

（1）挖掘或利用短文本隐含的语义信息，即描述如何表示输入的问题；

（2）捕获和学习多层次表示特征，即刻画如何从词构造出短语搭配，再由短语结合成句子，进一步扩展到文档层次；

（3）捕获长依赖的上下文关系，即刻画当包含必要信息的上下文距离当前词很远的时候，是否能很好的捕获和联系起来。

传统的文本表示方法是直接使用 Bag of words、Bag of N-gram 等特征表达，这类特征在长文本的应用上效果比较明显，但在短文本应用上，不仅特征特别稀疏，而且没有保存词序信息，同时隔离了词与词之间的语义距离，每个词或短语之间语义距离都是一样的。近年来，深度学习和大规模的无标签数据推动了表示学习的发展，主要代表为分布式表示学习的方法，即词向量，通常也称为“Word Representation”或“Word Embedding”[30-32]，是通过训练无标签语料将每个词映射成低维实数向量的方法，每一维都代表了词的浅层语义特征，通过低维实数向量之间的距离（例如余弦相似度、欧式距离等）来描述字词之间的语义相似度。低维的词向量避免了用传统的稀疏表达在解决某些任务的时候（比如构建语言模型）所造成的维数灾难。另外，由于词向量的相似性本质上只是上下文相似，依然没有彻底解决词的同义问题，因此，结构化语义知识库如 Wikipedia、WordNet 等也常被用于语义相似性计算。

给定词向量等具有语义信息的特征表示之后，很多方法相继提出各种组合方法来学习短语或句子级别的特征表示。直观上，句子的意思可以由词的语义组合而成，而文档的意思可以由句子的语义组合而成[33]。研究方法主要分为三种类型:无序模型（unordered models），序列模型（sequence models）和 CNN 模型。在无序模型中，直接将文本中的所有词的词向量进行相加[18]或平均[34]，这种方法类似于 Bag of words 模型，忽略了词序等结构信息。而递归神经网络（recurrentneural networks，RNN）和 LSTM 等序列模型则是通过不断的迭代来构造句子表示，保存句子的词序信息。此外，序列模型还可以轻易地同语法树转换，使得每个节点都可以组合子节点的信息[35-36]。基于 CNN 的模型已经在句子建模和分类任务中取得了很多标志性的结果。通过卷积操作，CNN 模型通过控制不同长度的卷积核大小来提取变长的区域特征，如 Kalchbrenner[18]的 DCNN 模型通过构造多层的卷积层和动态 k-max 次抽样层一层层的提取句子特征表示。而 Yin[27]等人则进一步利用多种词向量输入和预初始化来进一步提升结果。这些模型也非常的扩展到文档表示的建模，模型基本跟句子级别的模型一样，不同的是以句子级别特征表示为输入[20,,33]。

在口语对话系统的 DA 识别方面，DA 识别通常被当作短文本分类问题[37]，然而，由于天然的口语化特性，DA 识别是短文本分类中更有挑战性的任务，其中 OOD 的特点更加显著，如信息缺失，且更具开放性和表达多样性等。为了解决信息缺失的问题，很多研究提出多种组合上下文信息的方法。如 Dernoncourt[38]等人直接将当前句与上一话语拼接成稀疏的 Bag of N-gram 特征向量，并使用 SVM、LR 等模型进行分类，虽然考虑了上下文信息，但是组合方式比较简单，忽略序列信息，且不符合层次表示学习的思想。而 Lee[20]等人则将对话段转为一个序列预测问题，结合历史对话信息来预测当前话语的 DA，利用 CNN 或 RNN 对用户话语进行建模，进而将当前话语和历史话语的整合到前馈神经网络中进行 DA 识别，虽然一定程度利用历史对话信息改善了识别准确率，但使用的特征比较单一，只用一种词向量对词进行特征表示，仍然没有解决歧义问题，且无法对超出领域话语进行处理。

本课题将通过以下两部分进行 DA 识别的探索：

（1）借助现有短文本分类技术，基于多层次表示学习的基本架构，进一步融合多种异构不同层次的特征，并从单句的理解扩展到某个环境下短文本的理解，以改善中文口语对话系统的 DA 的识别效果，进而提升对完整对话的理解。

（2）领域内话语和 OOD 话语相结合的处理方法。由于相比于领域内话语，OOD 话语语义更为开放和表达多样，容易产生 OOV 字词，并且也缺少领域内话语携带的相对较为丰富的语义或语法等文本信息。本课题将 OOD 话语从对话话语中分离，分别对其采用不同的方法，以期领域内话语的处理更加专业，同时，可以提升 OOD 话语的可通用性和移植性。

参考文献：

[1] Price P J. Evaluation of spoken language systems: the ATIS domain[C]. In Proceedings of DARPA Workshop on Speech and Natural Language, Hidden Valley, PA, 1990.

[2] Gorin A, Riccardi G, Wright J. How may I help you?[J], Speech Communication,1997, 23(1-2): 113–127.

[3] Zue V, Seneff S, Glass J, et al. JUPITER: a telephone-based conversational interface for weather information[J]. IEEE Transactions on Speech and Audio Processing, 2000, 8(1): 85−96.

[4] Durston P, Farrell M, Attwater D, et al. OASIS natural language call steering trial[C]. In Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech 2011), 2001: 1323–1326.

[5] 张琳, 高峰, 郭荣, 等. 汉语股票实时行情查询对话系统[J]. 计算机应用, 2004, 24(7): 61-

63.

[6] 黄寅飞, 郑方, 燕鹏举, 徐明星, 吴文虎. 校园导航系统 EasyNav 的设计与实现[J].中文信息学报, 2001, 15(4): 35-40.

[7] Reichel C S, Sohn J, Ehrlich U, et al. Out-of-domain spoken dialogs in the car: a WoZ study[C]. In Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2014), 2014: 12-21.

[8] Pappu A, Rudnicky A. The structure and generality of spoken route instructions[C]. In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2012), 2012: 99-107.

[9] Huang P J, Lin X M, Lian Z Q, et al. Ch2R: a Chinese chatter robot for online shopping guide[C]. In Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2014), 2014: 26-34.

[10] Wang Y Y, Deng L, Acero A. Spoken language understanding[J]. IEEE Signal Processing Magazine, 2005, 22(5): 16-31.

[11] Wang Z Y, Wang H X. Understanding Short Texts[C]. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), 2016: Tutorial.

[12] 马成龙, 姜亚松, 李艳玲, 等. 基于词矢量相似度的短文本分类[J]. 山东大学学报:理学版, 2014(12):18-22.

[13] 王俊东, 黄沛杰, 林仙茂, 等. 限定领域口语对话系统中超出领域话语的协处理方法[J]. 中文信息学报, 2015, 29(5): 194-203.

[14] Sahami M, Dumais S, Heckerman D, et al.. A bayesian approach to filtering junk e-mail[C]. In Proceedings of AAAI'98 Workshop on Learning for Text Categorization, 1998 (62): 98-105.

[15] Johnson R, Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks[C], Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL (NAACL-HLT 2015), 2015 : 103–112.

[16] Wang S, Manning C D. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification[C]. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers (ACL 2011), 2012: 90-94.

[17] Maas A L, Daly R E, Pham P T, et al.. Learning Word Vectors for Sentiment Analysis[C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011), 2011: 142-150.

[18] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences[C]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics(ACL2014), 2014: 655–665.

[19] Kim Y. Convolutional Neural-Networks for Sentence Classification[C]. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP 2014), 2014: 1746–1751.

[20] Lee J Y, Dernoncourt F. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks[C].In Proceedings of NAACL-HLT 2016, 2016: 515–520.

[21] Kim S, D'Haro L F, Banchs R E, Williams J, et al.. The Fourth Dialog State Tracking Challenge[C]. In Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS 2016).

[22] Louwerse M M, Crossley S A. Dialog act classification using n-gram algorithms[C]. In Proceedings of 19th Florida Artificial Intelligence Research Society Conference (FLAIRS 2006), 2006: 758–763.

[23] Silva J, Coheur L, Mendes A C, et al. From symbolic to sub-symbolic information in question classification[J]. Artificial Intelligence Review, 2011, 35(2):137–154.

[24] Tan C M, Wang Y F, Lee C D. The use of bigrams to enhance text catego-rization[J]. Information Processing and Management, 2002(38):529–546.

[25] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]. In Proceedings of the 17th International World Wide Web Conference (WWW 2008), 2008:91-100.

[26] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[C]. In Proceedings of Advances in Neural Information Processing Systems, 2015: 649-657.

[27] Yin W, Schütze H. Multichannel Variable-Size Convolution for Sentence Classification[C].In Proceedings of the 19th Conference on Computational Language Learning, 2015: 204–214,

[28] Wang P, Xu J, Xu B, et al. Semantic Clustering and Convolutional Neural Network for Short Text Categorization[C]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015(2): 352-357.

[29] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive

matrix-vector spaces. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012: 1201–1211.

[30] Bengio Y, Ducharme R e, Vincent P, et al.. A neural probabilistic lan-guage model[J]. The Journal of Machine Learning Research, 2003(3):1137–1155.

[31] Mikolov T, Sutskever I, Chen K, et al.. Distributed representations of words and phrases and their compositionality[C]. In Proceedings of Advances in neural information processing systems, 2013: 3111–3119.

[32] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP 2014), 2014(12):1532–1543.

[33] Zhang R, Lee H, Radev D. Dependency Sensitive Convolutional Neural Networks for Modeling Sentences and Documents[C]. Human Language Technologies: The 2016 Annual Conference of the North American Chapter of the ACL (NAACL-HLT 2016). 2016: 1512-1521.

[34] Landauer T K, Dumais S T. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge[J]. Psychological review, 1997(104):211.

[35] Tai K S, Socher R, Manning C D. Improved semantic representa-tions from tree-structured long short-term memory networks[J].arXiv preprint,2015:1503.00075.

[36] Socher R, Perelygin A, Wu J Y, et al.. Recursive deep models for semantic compositionality over a sentiment treebank. InProceedings of EMNLP, 2013(1631): 1642.

[37] Novielli N. and Strapparava C. The role of affect analysis in dialogue act identification [J]. IEEE Transactions on Affective Computing, 2013, 6(1): 1-14.

[38] Dernoncourt F, Lee J Y, Bui T H, et al.. Adobe-MIT submission to the DSTC 4 Spoken Language Understanding pilot task[J]. arXiv preprint arXiv:1605.02129, 2016.

二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）

1. 研究内容和目标

本课题主要研究内容分为两部分：

（1）多种不同层次上下文信息的融合，以完整的理解用户话语：

为了避免歧义和主题不明确，可以通过分层次来理解用户话语，从小到大可分别为：词（word）、短语或搭配（phase）、话语（utterance）、对话段（segment）、对话（dialogue）。通常情况下，对话段是指谈论同个属性或话题的对话集。但是对话段的切分通常不明确，用户话语一般需要限定在某个对话段中进行理解，这就可以很自然应用短文本多层次表示。但在实际上，对话段的切分不明确，则就导致模型不能准确的使用上下文信息。通常在电影评论、微博、篇章理解等文本中，不同的句子基本都是围绕同个话题，但在对话段中，由于随着对话进程的推进，会处理不同的事物，所以选取的上文信息很可能跟当前句并不属于同个话题。如果不能正确的选出相关的上文信息，将会引入更大的噪音，造成句子理解的不确定性和偏差。

为了避免引入噪音，另外的一个解决方法则是充分融合各种有用的异构信息，丰富上下文信息，如在完整对话中，可以通过不断的分析用户的情感极性，来进一步确定当前用户的意图，以便正确定位当前用户的态度；在对话段中，可以通过属性抽取、实体识别等技术来进行句子补全和消歧（当用户说“800”的时候，可以通过上文定位谈论的是价格和像素等）；在句子内部，则可以利用 Wikipedia、WordNet 等来进一步确定同义词等，或通过句型理解、语法树、名词关系分析等来捕获句子内部词语的关系等结构信息。当获取多种不同层次的特征输入后，如何有机的融合在一起，是本课题的主要研究内容之一。

（2）领域内外话语相结合的处理方法

研究对话系统的目的是让人同计算机的交流更方便，让计算机具有类似人类的思维从而帮助人们完成更多的工作，从这个角度看，能胜任某一专业领域工作的面向任务（task-oriented）的限定领域（restricted domain）对话系统，比开放领域（open domain），如面向聊天（chat-oriented）的对话系统更有研究和应用价值。

然而，当使用自然语言对话时，即使用户了解某对话系统的限定领域，例如，医疗信息咨询、导航或者导购，用户在对话流程中仍然不可避免会使用一些超出领域话语（utterance），如，问候、个人问题、表达肯定等。尽管这些限定领域对话系统从完成任务角度上看只需要专注于自己预定义的业务功能，但是，如果能较为妥善地处理好 OOD 话语，而不仅仅是提示用户话语超出领域，将会有效地提高用户体验。另外，领域外话语具有通用性，可以轻易的移植到任何限定领域的对话系统中。所以本课题不仅要求模型不仅有处理某个领域内的专业知识，而且可以合理有效的理解 OOD 话语。主要分为三部分：

1）OOD 检测：判断用户话语是否属于 OOD 话语，这是模型的第一步骤，也是本课题的难点之一，如何判断用户话语是否携带特定的领域语义，主要分为两种方法：rule-based 和统计方法。

2）领域内话语的处理：经由 ID 检测，带有领域语义的句子将进入该模块处理，领域内话语相对专业和正式，这部分的处理方法借助领域知识库和句式理解等方法进行有效的理解。另外，还需要借助属性抽取、命名实体识别、名词关系分析等方法进行属性定位、属性值挖掘、句子补充等，属性的正确识别是该部分的难点。

3）领域外话语的处理：相比于领域内话语，OOD 话语语义更为开放和表达多样，容易产生 OOV 字词，并且也缺少领域内话语携带的相对较为丰富的语义或语法等文本信息。此外，由于用户聊天的随意性非常的高，OOD 话语出现比较随意，与对话上下文的关联也远远没有领域内话语高。因此，对于 OOD 话语，可以借助人工智能标记语言 AIML 语料库和一套预定义的 DA 模板对 OOD 话语进行合理的归类，以有效降低语料库建设的难度。相比于每个话语都有专门的回答，虽然在一定程度上降低了系统对用户话语的理解的“准确”度，但并不会影响后续处理和应答。同时由于 DA 模板及后续介绍的模式转换抓住了多样化用户话语的共性，具有一定通用性，也易于扩展到其它限定领域。由于 AIML 语料库不能完整覆盖所有 OOD 话语，对于未匹配到的用户话语，采用基于统计的方法进行分类，比如 CNN、RNN、随机森林、SVM 等方法，同样也需要利用多种不同的上下文信息进行句子信息补充。另外，如何实现对 OOD 话语的多标签标注也是本内容的研究重点。

2. 拟解决的关键问题

1）属性的挖掘和属性值的消歧。对话过程中，用户话语通常更简洁，经常会省略部分关键的信息，如对于价钱，用户可能只会说 800，这就要求从上文中定位具体属性名。

2）不同多层次特征的融合，即如何高效且正确的融合语法树、词向量、知识库、情感极性等语法语义特征。

3）多标签 DA 的标注，对于部分用户话语，用户的意图可能不止一种，即多标签 DA。但目前多数任务都是单标签任务，因此需要解决如何将单标签的任务的经验转移到多标签分类的问题。

三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性）

1.研究方法

在课题的设计最初阶段要通过文献研究法、调查法不断积累相关知识，拓展对整个系统的认知，进而粗略地设计出整个方案的框架图。然后继续通过文献研究法、经验总结法、实验法对方案进行优化、具体化，设计出各个部分的框架图，再继续通过文献研究法、经验总结法、实验法完成各个部分的设计，实现各个模块的有机结合。

2.技术路线

合理吸收、有效利用现有的成熟方法，在短文本和层次学习现有的研究结果之上，加入了以领域本体以及 LDA 主题模型搭建语义场模型，以降低参与聚类的特征维度以及提高聚类的精度等，将数据挖掘、自然语言处理、语义学等技术相融合。

具体分以下三个阶段实现：

1）阶段一 语料库的获取

语料库是实验的基础，由于话语料和中文短文本数据集的缺失，这个阶段主要整理数据，供后面实验使用。

2）阶段二 技术探索。深入研究如下 3 方面：

探索多层次上下文信息对句子理解的影响。

探索如何由单标签转移到多标签的标注问题。

探索如何有效的融合多种特征信息。

3）阶段三 验证和对比

在公共数据集（英文）上验证模型的有效性，进而嵌入到对话系统中进行真实环境下的人机测试，并对测试结果进行分析，对模型不断调优。

四、本研究课题的特色与新颖之处

本课题是研究的是面向中文口语对话系统的短文本分类。与现有技术和研究成果相比，特色先进性体现在以下几方面：

（ 1）考虑多标签问题，进一步细化用户的意图，使得更准确捕获用户需求。

（ 2）结合多层次上下文信息、混合句式理解、属性抽取、情感分析等多种特征融合的技术进行话语的 DA 识别，解决现有研究中口语语言的表达多样性给对话系统语言理解造成的困难。

（ 3） 将 OOD 话语的处理从对话中分出，采用人工智能标记语言（ AIML） 和统计方法相结合的处理机制，克服了现有限定领域口语对话系统不能很好地理解不带有效领域语义信息的用户话语的困难。

五、研究基础与工作条件（1. 与本项目相关的研究工作积累基础 2. 包括已具备的实验条件，尚缺少的实验条件和拟解决途径）

1、工作积累

（1）有限定领域口语对话系统中超出领域话语识别与处理的研究基础。在《中文信息学报》发表（和录用）论文 2 篇，EI 检索国际会议（第 20 届亚洲语言处理国际会议，IALP 2016）论文 1 篇。

（2）有微博立场分析、情感极性分析等短文本分类的相关比赛经验。

（3）精读了较多短文本分类相关的顶级会议论文；熟悉自然语言处理、机器学习、深度学习等算法；熟练使用机器学习的工具包等。

2、实验条件

（1）实验工具：

1）开发工具：Pycharm、Visual Studio 2015 等

2）开发环境：Ubuntu 14.04

3）开发语言：Python 2.7.6

（2）语料库资源：

1）手机导购领域人机中文对话语料：从 2014 年 9 月 26 日到 2016 年 6 月 15 日，Ch2R 手机导购机器人记录了 2716 段原始中文对话语料，参与用户超过 1000 人次，产生 41952 句话语（含机器人和用户话语）。其中，来自原始对话语料的 OOD 话语有 3187 句，去重后有 1045 句，另人工增加 1423 句 OOD 话语，合并共 2468 句，目前已完成 OOD 话语 DA 标签的人工标注（共 25 类）。

2）旅游导购领域英文对话语料：DSTC 4，共 35 段，来自第四届 Dialog State Tracking 比赛。

3）其他部分对话行为资源库：

SwDA（Jurafsky, et al.. 1997）：1134 段对话；

MRDA（Janinet, et al.. ）：73 段对话；

4）其他部分可用短文本公共数据集：

MR（Bo Pang, et al.. ACL 2005）：10662 句电影评论；

SST（Richard Socher et al.. EMNLP 2013）：11,855 句用户电影评论；

TREC（Xin Li, et al.. CL 2002）：5952 句用户问句；

CR（Hu, et al.. KDD 2004）：3775 句用户商品评论；

IDMB（Maas, et al.. ACL 2011）：50,000 有标注用户电影评论；50,000 无标注用户电影评论；

学位论文工作计划

时间	研究内容	预期效果
2016.06——2016.09	前期资料收集	
2016.09——2016.11	阅读文献，分析现有方法局限性，试验各种方法，理解问题	在公共数据集上测试模型的可行性
2016.11——2016.12	设计算法、调整相关参数并试验，找到解决方法	发表一篇核心论文或国际会议论文
2016.12——2017.01	分析实验结果	申请计算机软件著作权
2017.01——2017.03	整理资料，撰写论文初稿	撰写硕士学位论文
2017.03——2017.04	修改论文，完成论文，准备答辩	

经费开支情况表

开支项目	数量	金额(万元)

导师意见（根据该生的政治表现、思想道德、学习成绩、学习态度、科研能力、遵纪守法和执行学校规章制度等提出是否同意继续攻读硕士学位）

导师签名：　　日 期：　　年　　月　　日

中期考核意见（优秀、良好、及格、不及格）

开题报告指导小组意见（包括对以下各参考项目的评价、提出开题报告是否通过的意见。）

请在相应栏目内画“√”

1. 选题与导师研究方向符合度	符合（ ）	一般（ ）	不符合（ ）
2. 选题与专业（领域）符合度	符合（ ）	一般（ ）	不符合（ ）
3. 选题的新颖性	强（ ）	一般（ ）	不明显（ ）
4. 研究的可行性	可行（ ）	一般（ ）	不可行（ ）
5. 研究的工作量	充足（ ）	一般（ ）	不足（ ）
6. 总体评价	优（ ）	良（ ）	中（ ） 差（ ）

对开题报告是否通过的意见：

小组成员签名：

日 期：　　年　　月　　日

学科意见（请对开题报告的选题与专业符合度作出评价，并提出是否同意开题的意见。）

学科带头人签名：　　日 期：　　年　　月　　日

学院意见：

学院负责人签名：　　学院公章：　　日 期：　　年　　月　　日

硕士研究生成绩单

学号	2015307809	姓名	王俊东		性别	男
学院	数学与信息学院	专业	计算机技术		学制	2
导师	黄沛杰			入学日期		2015-09-08
类别	课程名称		学时	学分	学期	成绩	备注
学位课程	研究生综合英语		36	2	1	89	
	公共英语选修课		36	2	2	86	
	中国特色社会主义理论与实践研究		36	2	1	89	
	自然辩证法概论		18	1	2	91	
	算法分析与设计		36	2	1	88	
	高级数据库技术		36	2	1	95	
	面向对象技术		36	2	1	92	
	人工智能原理		18	1	2	91	
非学位课程	最优化方法		36	2	1	93	
	高级数理统计		54	3	1	94	
	数据仓库与数据挖掘		36	2	2	97	
总学分	22			学位课学分		15	

打印日期：2016-09-21