摘　要

人工智能是一门研究如何赋予计算机类人智能的学科，涵盖图像处理、语音处理、自然语言处理等多项技术，其中自然语言处理架起了人类与机器之间语言沟通的桥梁。为了方便人类与机器进行自然且沉浸式的语言交互，科学家们致力于构建诸如 Apple Siri、Google Now、Microsoft Cortana 和 Amazon Alexa 等智能对话系统或社交聊天机器人，这也是人工智能领域长期关注的关键任务之一。构建对话系统的基本问题是如何实现计算机理解对话历史的语义，并预测出下一句合理且自然的回复。其中一条主流技术途径是从给定的回复候选集合中选择最合理的回复，采用该技术途径的对话系统被称为检索式对话系统。检索式对话系统又可以根据应用场景的不同，细分为个性化对话、融合背景知识对话以及多方对话等多个子任务。此外，随着深度学习技术的兴起，构建检索式对话系统的方法也由传统的基于规则和统计的方法向基于神经网络模型学习的方法过渡。然而，现阶段基于深度学习的检索式对话系统往往存在对特定应用场景考虑不足的问题，例如，难以有效地获取对话历史和回复候选之间的复杂语义匹配信息，没有对于对话者的一致性个性进行明确的长期记忆，缺乏对话相应的背景知识依托，难以建模多方对话中复杂的语句和对话者之间的交互等，这些都影响了最终回复选择的准确性。

因此，本文围绕基于深度学习的检索式对话系统，在多轮对话、个性化对话、融合背景知识对话和多方对话等方面开展研究工作，具体包括：

首先，研究了面向多轮对话的检索式对话系统。针对现有研究缺乏对对话历史与回复候选进行句级语义匹配的问题，提出了细粒度句级交互匹配回复选择模型，从对话历史中选择与回复候选最相关的信息，并对回复进行句级分解，将对话历史与回复候选中任意两个语句之间的句间距离先验信息融入到匹配过程；进一步提出了在预训练语言模型中融合说话人表征和领域自适应的回复选择模型预训练方法，实现了在预训练语言模型中体现说话人交替改变的这一对话属性，提升了预训练模型在对话领域的表征能力。以上研究在四个公开的多轮对话回复选择数据集上提升了回复选择的召回率，取得了当时的最优性能。

其次，研究了基于对话者画像的个性化对话系统。针对现有研究只关注回复者自身画像而忽略对话搭档画像这一问题，基于是否考虑画像和语境之间，以及画像和回复之间的交互，提出了四种基于对话者画像的个性化回复选择画像融合策略，并应用于三种不同类型的模型，该研究在一个公开的基于对话者画像的对话回复选择数据集上验证了回复者自身和对话搭档画像对于个性化回复选择的影响。另一方面，针对缺少预先定义的对话者画像所引起的个性化对话系统冷

启动问题，提出了基于早期对话文本搜索近似画像的对话者画像检测方法，构建了一个对话者画像检测数据集，并验证了在对话者画像没有预先给定时，通过细粒度匹配方法检索近似的对话者画像的有效性。

再次，研究了融合背景知识的检索式对话系统。以直接建模背景知识和回复候选之间的语义匹配关系为目标，提出了双向交互匹配回复选择模型，同时进行对话历史和回复候选之间，以及背景知识和回复候选之间语义表征的交互匹配；针对对话历史和背景知识之间相互独立且存在冗余信息，同时单次交互也只能获取浅层匹配特征的问题，提出了基于背景知识过滤的迭代匹配回复选择模型，该模型预先建立对话历史和背景知识之间的交互感知机制，并对背景知识进行有效筛选，随后再通过迭代式交互匹配去获取对话历史和回复候选之间，以及背景知识和回复候选之间的深层匹配信息。以上研究在两个公开的融合背景知识的对话回复选择数据集上提升了回复选择的召回率，取得了当时的最优性能。

最后，研究了面向多方对话的检索式对话系统。多方对话中含有丰富的对话者之间、语句之间，以及对话者与语句之间的交互关系，同时多方对话涉及的说话人识别、接收者预测，以及回复选择等任务间也存在天然的互补关系。因此，本文提出了基于多任务自监督学习的多方对话建模方法，围绕“谁对谁说了什么”这一多方对话中的核心问题，设计了多个自监督学习任务，实现了模型计算出语义更丰富的对话者表征和语句表征，加深了模型对多方对话的理解，增强了模型在多个下游任务的泛化性。以上研究在两个公开的多方对话数据集上提升了说话人识别、回复选择，以及接收者预测任务的准确率和召回率，取得了当时的最优性能。

关键词：自然语言理解，检索式对话系统，回复选择，个性化对话，融合背景知识对话，多方对话

ABSTRACT

Artificial intelligence is a discipline that studies how to endow computers with human-like intelligence, including image processing, speech processing, natural language processing and other technologies. Among them, natural language processing bridges the language communication between humans and machines. Facilitating language interactions between humans and machines in a natural and immersive manner has been one of the key and long-standing tasks in the field of artificial intelligence. Scientists are committed to building intelligent dialogue systems or social chatbots such as Apple Siri, Google Now, Microsoft Cortana, and Amazon Alexa. It is a fundamental problem for dialogue systems to understand the semantics of the conversation context so that it can predict the next response reasonably and naturally. One of the main approaches is to selecting the most reasonable response from a given set of candidates, which is well-known as retrieval-based dialogue systems. Retrieval-based dialogue systems can be categorized into multiple sub-tasks such as personalized conversations, knowledge-grounded conversations and multi-party conversations depending on the application scenarios. In addition, with the rise of deep learning technology, the method of constructing retrieval-based dialogue systems has also transitioned from traditional methods based on rules and statistics to those based on neural network model learning. However, there are still problems in the current deep retrieval-based dialogue systems due to insufficient consideration of specific application scenarios. For example, there is difficulty in capturing the complex semantic matching information between conversation contexts and response candidates effectively, no explicit long-term memory of the consistent personality of an interlocutor, lack of background knowledge that a dialogue is grounded on, and difficulty in modeling complex interactions between utterances and interlocutors in multi-party conversations, which influences the performance of response selection.

Therefore, this thesis focuses on deep retrieval-based dialogue systems, and studies multi-turn conversations, personalized conversations, knowledge-grounded conversations and multi-party conversations respectively. Specifically:

First, this thesis studies multi-turn conversation in retrieval-based dialogue systems. For the lack of utterance-level semantic matching between conversation contexts and response candidates in the existing work, this thesis proposes fine-grained utterance-to-utterance interactive matching network for response selection. This model

selects the most relevant information in conversation contexts for response candidates. Also, a response is decomposed to a set of utterances, and the prior information of the inter-utterance distance between any two utterances in a conversation context and in a response candidate is incorporated into the matching process. Furthermore, this thesis proposes a pre-training method for response selection that integrates speaker representation and domain adaptation into pre-trained language models. This method reflects the dialogue property of speaker alternation in pre-trained language models, and improves the representation ability of pre-training models in dialogue. The above study improves the recall of response selection on four public multi-turn response selection datasets, and achieves the state-of-the-art performance at that time.

Secondly, this thesis studies interlocutor persona-based personalized dialogue systems. As existing research only focuses on the self persona of a respondent and ignores the partner persona, four persona fusion strategies for personalized response selection are proposed based on whether the interaction between personas and contexts, and that between personas and response is considered. These strategies are implemented into three different models to examine the effect of self and partner personas on personalized response selection on a public interlocutor persona-based response selection dataset. On the other hand, to alleviate the cold-start problem in personalized dialogue systems caused by the lack of pre-defined interlocutor personas, a method of speaker persona detection to search for approximate personas based on early conversation contexts is proposed. A dataset for speaker persona detection is constructed, and employed to verify the effectiveness of retrieving approximate interlocutor personas by fine-grained matching when interlocutor personas are not pre-specified.

Thirdly, this thesis studies knowledge-grounded conversation in retrieval-based dialogue systems. To model the semantic matching relationship between background knowledge and response candidates directly, a dually interactive matching network for response selection is proposed to conduct interactive matching of semantic representations between conversation contexts and response candidates, as well as that between background knowledge and response candidates simultaneously. Since there exists independence and redundancy between conversation contexts and background knowledge, and a single time of interaction can only capture shallow matching features, a method of filtering before iteratively referring is proposed for knowledge-grounded response selection. The method pre-establishes the interaction perception mechanism between conversation contexts and background knowledge, and select relevant background knowledge effectively. Then, iteratively interactive matching is conducted to obtain the deep match-

ing information between conversation contexts and response candidates, as well as that between background knowledge and response candidates. The above study improves the recall of response selection on two public knowledge-grounded response selection datasets, and achieves the state-of-the-art performance at that time.

Finally, this thesis studies multi-party conversation in retrieval-based dialogue systems. A multi-party conversation always contains complicated interactions between utterances, between interlocutors, as well as between an interlocutor and an utterance. Meanwhile, the tasks of speaker identification, addressee recognition and response selection in multi-party conversations are complementary among each other. Therefore, this thesis proposes a method for modeling multi-party conversations based on multi-task self-supervised learning. Multiple self-supervised learning tasks are designed to tackle the core issue of "who says what to whom" in multi-party conversations. In this way, models can compute better interlocutor representations and utterance representations containing richer semantics. Furthermore, it can deepen the understanding of multi-party conversations, and enhance the generalization ability across multiple downstream tasks. The above study improves the accuracy and recall of speaker identification, response selection and addressee recognition on two public multi-party conversation datasets, and achieves the state-of-the-art performance at that time.

Key Words: Natural Language Understanding; Retrieval-based Dialogue System; Response Selection; Personalized Conversation; Knowledge-Grounded Conversation; Multi-Party Conversation