智能问答系统研究综述

清华大学语音与语言技术中心 骆天一

1 研究意义

从早期的图书馆检索系统、专家系统到现在的搜索引擎，快速并准确的获取信息一直是人们梦寐以求的追求目标，尤其是在信息浩如烟海的互联网时代。随着时代的发展，一方面数据的量级已从 TB 发展到 PB 乃至 ZB，可以称之为海量数据；另一方面，用户的需求越来越精细化、多样化，基于关键词组合或者基于浅层语义分析的检索系统越来越不能满足用户的需求，智能问答系统通过将数据经过深度加工处理形成具有某种固定结构的知识库，并通过最先进的自然语言处理技术解析用户的需求，从而快速地、准确地为用户提供所需要的信息。

智能问答系统及其相关领域的研究具有重要的研究价值。智能问答系统涉及的领域很广，其中主要关键技术有知识的抽取和表示，用户问句的语义理解和通过知识推理得到答案。这些领域都需要进行深入研究我们才会得到更好的智能问答系统。而无论我们在任一领域取得重大的突破，不仅仅对于智能问答系统，而且对于其它领域，包括文本分类、推荐系统等都会有相当大的促进作用。

另一方面，智能问答系统也具有重大的实际应用价值。能够快速准确地找到信息可以让人们的生活变得非常便利，例如：查询天气，股票价格，飞机航班情况等。而且更让人激动人心的是智能问答系统可以部分地替代人工劳动，例如替代人工客服对客户进行自动回答，可以大大减少企业的人力成本。

2 智能问答系统的发展简史

智能问答系统已经有 70 年的发展历史。早期的智能问答系统大多针对特定的领域而设计，并且数据量也很有限，不容易进行扩展，例如：Baseball[1]和 Lunar[2]，这些诞生在上世纪五六十年代的智能问答系统通常只接受特定形式的自然语言问句，而且可以供智能问答系统进行训练的数据也很少，所以无法进行基于大数据的开放领域的问答从而未被广泛使用。进入九十年代之后，由于互联网的发展，大量可供训练的问答对在网上可以被搜集和找到。尤其是 TREC-QA[3]评测的推出，极大推动促进了智能问答系统的发展，研究人员在该语料库上训练和测试各种问答模型，先后提出了基于逻辑推理的方法[4]，基于模板匹配的方法[5]，基于机器学习的方法[6]和基于数据冗余性的方法[7]等许多领先方法。在此阶段，人们主要利用信息检索或浅层语义理解技术去从大量候选集中寻找答案从而构建智能问答系

统，故检索式问答技术取得了巨大发展。但是检索式问答技术存在一个缺陷，就是答案中一定至少包含一个用户问句中含有的字或者词，但是这在实际情况中往往是不成立的。虽然浅层语义理解技术部分解决了这个问题，但是由于用户问句是自然语言，自然语言有着天然的复杂性，由于存在着以上缺陷，检索式问答技术不能真正很好地解决用户的需求。一直以来，阻碍智能问答系统向前发展的两个最主要因素是缺乏高质量的数据和强大的自然语言处理技术，不过随着维基百科，百度百科，搜狗百科这些基于用户协同生成内容的互联网应用的兴起，越来越多的高质量数据被积累和得到。基于此，大量的被精心设计以自动或半自动方式生成的知识库（例如 Freebase、YAGO、DBpedia 等）被建立起来。至于另一个问题，随着统计机器学习方法的兴起，自然语言处理技术各个子领域都取得了很大的进步，可以说阻碍智能问答系统最大的两个问题正在被科研人员逐步解决。近年来，智能问答系统取得了很大的发展和进步，已经有很多智能问答系统产品问世。例如 IBM 研发的智能问答机器人 Watson 在美国智力竞赛节目《Jeopardy!》中战胜人了选手，其所拥有的 DeepQA 系统集成了统计机器学习、信息抽取、知识库集成和知识推理等深层技术。苹果公司的 Siri 系统和微软公司的 cortana 分别在 iPhone 手机中和 Windows10 操作系统中都取得了很好的效果。在国内，众多企业和研究团体也推出了很多以智能问答技术为核心的机器人，例如：微软公司的“小冰”、百度公司的“度秘”和中科汇联公司的“爱客服”，我们可以看到，这些机器人不仅提供情感聊天的闲聊功能，而且还能提供私人秘书和智能客服这样的专业功能。这些智能系统的出现标志着智能问答技术正在走向成熟，预计未来还会有更多功能的机器人问世和解决用户的各种需求。

3 智能问答系统的研究前沿

尽管已经取得了令人瞩目的成就，但是智能问答系统还远未完美，智能问答系统涉及的领域很广，其中主要关键技术有知识的抽取和表示，用户问句的语义理解和通过知识推理得到答案。这些领域都需要进行深入研究我们才会得到更好的智能问答系统，然而这些领域都各自相对独立的存在并且采用了非常不同的方法，并且这些方法都有着各自的瓶颈。所以智能问答系统的现代研究基本围绕这三方面展开。

（1） 自动信息抽取和构建知识图谱

智能问答系统的实现需要非常强大和全面的知识作为基础，而目前虽然互联网拥有海量的知识资源，这些资源绝大多数都是非结构化的知识，并且存在多种不同的结构，更加大了提取知识的难度，如何将这些大规模的异构数据融合在一起，并且进行理解和萃取，转换为计算机可以处理的形式。事实上，科学家们一直都致力于构建更大更完备的知识资源库。早

期的知识资源库大多是由各个领域的专家去构建的各自领域的专业领域知识资源库。手动由专家构建知识资源库的优点是知识质量高，缺点是进行大规模的知识资源库时费时费力，而且在转换领域的时候，必须手动构建另一个领域的知识资源库，并且各个领域的知识资源库的构建方法和构建完成的知识资源库的结构一般情况下是不一致的。构建方法不具有通用性和知识资源库的结构不具有一致性，不能产生通用领域的智能问答系统。我们可以把智能问答系统中所用到的知识粗略地分为语言知识和世界知识，语言知识指的是语义单元的知识，包括词义信息，上下位关系等等。包含关于这些语言知识的词汇知识库包括英文的 WordNet、FrameNet、中文词汇知识库 HowNet 等。世界知识则是对现实世界中实体以及实体发生的事实的组织和表示，人们很早就以人工方式建立了一个世界知识库 Cyc。该常识知识库涉及了五十万个概念，三万个关系，数百万条事实，是目前为止全世界最大的完全由人工建立的常识数据库。不过即使如此，也远远不能满足开发领域智能问答系统对知识资源的需求。

实际上，现在绝大多数的知识都存在于非结构化的文本数据中。为了克服需要经过人工整理知识资源库费时费力的困难，研究者们希望通过强大的信息抽取技术自动从海量的非结构化文本中获取大规模知识来建立大规模的知识资源库。在这个过程中，我们需要先对文本进行实体识别，然后对实体进行分类和消歧，接着抽取出关系和事件，这样就能构建一个由实体、关系和事件组成的知识资源库，目前科学工作者们一般以 Wikipedia 作为构建知识资源库的一个来源，因为 Wikipedia 集合了群体的智慧，拥有大量的高质量数据资源。所以大量的工作直接或间接地利用 Wikipedia 资源进行知识抽取。德国的马克思普朗克研究院（Max Planck Institute）通过融合 Wikipedia 和 WordNet 构建了一个大规模的知识库类别体系 YAGO，并且定义了几十种关系描述实体之间的关系。其它具有代表性的工作还包括：华盛顿大学图灵实验室的 TextRunner [8]，ReVerb [9]，R2A2 [10]，WOE [11]，OLLIE [12]；德国柏林工业大学 DSIM 组的 Wanderlust [13]，KrakeN [14]等。CMU 的 NELL[15]系统通过持续从互联网上抽取和挖掘知识，构建了一个可以处理多种智能信息需求的海量规模网络知识库，最新的工作包括基于 Wikipedia 自动生成结构化知识资源库的 DBpedia 和 Freebase。目前基于 Wikipedia 自动构建的知识资源库最受人们青睐，因为这些知识资源库可以随着 Wikipedia 的更新而不断自动更新自身的资源库，因而受到了搜索引擎巨头的极大关注。Google 于 2010 年收购了 Freebase 之后一直致力于构建相互关联的实体及其属性的巨大知识图谱（Knowledge Graph），并由此出发加强了 Google 的语义搜索。在国内，各大搜索引擎巨头也纷纷建立了自己的知识图谱，比如百度的知心、搜狗的知立方等。

（2） 问句理解和基于端对端的智能问答系统系统

我们不仅需要构建强大的知识资源库，还需要深入理解人提出的问题。人们的问题都是以自然语言方式呈现的。问句理解要做的就是要将自然语言转化成计算机可以理解的形式化语言。让计算机理解人类语言是一件非常困难的事情，这也是自然语言处理（Natural Language Processing）要解决核心的问题。解决这一问题有两种不同的思路，一种是语义解析方法（Semantic Parsing），另一种是基于信息检索的方法。语义解析方法非常符合人们的直觉，它将一个自然语言形式的问句，按照特定语言的语法规则，解析成语义表达式，在得到语义表达式之后，我们可以非常容易地将其转化为某种数据库的查询语言。

首先对于语义解析方法，研究人员们设计了很多方法进行自然语言问句到语义表达式的转换[15][16][17][18]。其中最常用的方法是利用组合范畴语法 CCG[15][19]，CCG 的核心是词汇，首先自然语言问句中的词汇被映射到语义表达式中的词汇。除了词汇之外，CCG 还按照特定的语法规则将词汇组合起来，进而得到了最终的语义表达式。然而，在 CCG 等这一类方法中起到重要作用的词汇大部分都是由人工生成，这样既限制了领域的转换和扩展，即若发生领域转换必须重新生成一批新领域的特定词汇，并且这类方法也存在人工生成的固有缺点，即在需要生成大规模词汇表的时候需要耗费大量的人力和时间。所以自动学习这种词汇成为了研究人员探索的重点[18][19][20]。

另外一种解决问句理解问题的方法是基于信息检索的方法。首先利用中文分词，命名实体识别等自然语言处理工具找到问句中所涉及到的实体和关键词，然后去知识资源库中去进行检索。举例来说，我们只针对 Freebase 来进行简单的事实类问答，先用命名实体工具识别出问句中的实体，这一步相对比较容易，因为一个实体的表达方式是非常有限的，接着再找出问句中的关系，这一步相对苦难一些，因为自然语言描述同一个关系的方式是多种多样的，比如中文中表达“配偶”关系的就有老婆、老公、妻子、丈夫等等多种说法，不过在 Freebase 里对于某一实体的关系数量是非常有限的，在实际应用中，我们可以利用各种办法很容易地排除掉很多无关的关系，获得实体和关系之后，最后可以很容易地在 Freebase 知识资源库中查到答案。基于信息检索的方法比较成熟且简单实用，而且不需要像 CCG 那样人工生成词汇，所以避免了 CCG 等予以解析方法的领域转换需要生成新词汇和需要人工生成词汇的缺点，但是缺点是基于信息检索的方法要求答案中必须至少包含问句中的一个字或词，所以不如语义解析方法精确。

从上面可以看出，自然语言问句的理解是智能问答系统中最核心也是最困难的一个环节，因为这个环节实际上要解决的问题是如何将自然语言最准确地转化为计算机可以表示和理解的形式。这个是不仅是智能问答系统需要解决的问题，也是人工智能领域所需要解决的最

核心的难题之一。

随着深度学习方法在学术界和工业界不断被验证可以取得最好的效果。人们将端到端的思想应用在了智能问答领域。最新的工作是[22][23]直接将问句和最终的答案做匹配，然后为了解决自然语言问句复杂多样（例如：用形式不同的同义词替换或者经过词语调换顺序后形成的问句事实上还是同样的意思）的特点，我们在深度神经网络中加入了内存的思想，在端对端的基础上可以对知识进行读入和写出[24] [25][26][27]。通过以上端到端的思想，我们就绕开了最困难的问句理解步骤。深度神经网络在其中起到了重要作用，并且这种方法也取得了与传统方法不相上下的效果。

（3） 知识推理

我们不仅需要构建强大的知识资源库，

自然语言转化成计算机可以理解的形式化

在智能问答系统中，不是所有的问题都可以利用现存的知识库直接进行回答，主要也是因为知识覆盖度毕竟还是有限，不过我们可以发现，其实有很多隐含知识我们是可以利用已经抽取到的知识进行推理回答的。比如：知识库中包含了一个人的“出生地”属性信息，但是没有这个人的“国籍”属性信息，但是其实我们可以从出生地推理出推理出这个人的“国籍”属性信息，因为某个地方肯定是属于某个国家的。所以在计算机中，我们需要把类似这样的推理知识进行表示和学习，知识推理任务也就是推理知识得到知识库中没有的隐含知识。

早期的知识推理方法大多对现有知识归纳学习出符号逻辑的推理规则，比如华盛顿大学开发的夏洛克-福尔摩斯系统[28]和 CMU 开发的推理系统 PRA[29]，这两个系统都可以利用已有的知识推理出知识库中不存在的知识。但是这些基于富豪的推理方法未能有效考虑富豪本身的语义，加上推理规则的数量随着关系的数量指数增长，因此很难扩展到大规模知识资源库中。另外，知识推理技术也因为深度学习技术的成熟而拥有了新的思路和方法，大量的工作[30][31]着眼于实体和关系的表示学习。通过在全局条件下对知识资源库的实体和关系进行编码，将实体，概念和关系表示为低维空间中的向量或矩阵，通过在低维空间中的数值计算完成知识推理任务。另外，基于内存的端到端深度神经网络技术也对智能问答系统中涉及到的推理问题进行了研究[25][26]。虽然目前这类推理的效果离实用还有段距离，但是这是一类非常值得继续深入研究的方法，特别是融合符号逻辑，表示学习和基于内存机制的端到端深度神经网络技术的推理技术。

4 总结与展望

我们回顾了语音识别技术的历史并粗略小结了当前研究中的热点问题。需要指出的是，

总的来看，几年来，基于 Wikipedia 这种高质量且会动态更新的开发资源建立起来的知识资源库日趋成熟，包含的知识也越来越多，另一方面，基于统计机器学习的自然语言处理技术和知识推理技术有了极大的发展且日趋成熟，这两方面的进步分别为智能问答系统的发展奠定了资源基础和技术基础。我们相信在不远的将来，越来越多的成熟的智能问答系统必将进入市场为人们提供非常有效的各种智能服务。但是智能问答系统仍然存在着一些亟待解决的问题。第一，网络中充满着大量人们为了某个特定领域建立的大量知识资源库，这些资源库在各自领域发挥了很好的作用，但是未来的趋势是开放领域的多领域覆盖智能问答系统，我们必须要想办法将所有异构的知识源统一起来，形成一个形式统一的知识源，否则各个独立的领域形成了多个信息孤岛，不能满足用户的统一查询需求。第二，我们目前建立的知识资源库中的知识大多是事实性知识，缺乏常识性知识，就是我们刚才在知识推理部分提到的推理知识，人们经常依据常识性知识来生成隐含的知识，常识知识在人类推理过程中具有极其重要的作用，而很多常识性知识难以整理成一个统一的规范化模式，因此，如何将常识性知识融入到智能问答系统中也是一个非常重要的问题。第三，随着深度神经网络技术在各个领域的成功应用，而且结合目前研究人员在应用深度学习于智能问答系统的一些尝试，我们特别希望用基于深度神经网络的智能问答系统能够代替传统的方法（包括基于语义解析和基于信息检索的技术），进行各种问句（包括事实性问题，定义类问题和推理性问题）的回答，尽管目前还难以达到实用的程度，但是我们相信随着深度学习技术的逐步发展和计算机硬件性能的进一步提升，能够自动进行问答的、结构非常统一简洁的深度神经网络技术成为未来智能问答系统的绝对主流技术。

[1] Green Jr, B. F., Wolf, A. K., Chomsky, C., and Laughery, K. Baseball: an automatic question-answer. In Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference (1961), ACM, pp. 219–224.

[2] Woods, W. A. Progress in natural language understanding: an application to lunar geology. In Proceedings of the June 4-8, 1973, national computer conference and exposition (1973), ACM, pp. 441–450.

[3] H. T. Dang, J. Lin, and D. Kelly. Overview of the TREC 2006 question answering track. In 15th Text REtrieval Conference, Gaithersburg, Maryland, 2006.

[4] Moldovan， D. & Rus， V. Logic form transformation of WordNet and its applicability to question answering， in Proceedings of the 39th Annual Meeting on Association for Computational Linguistics， 2001.

[5] M. M. Soubbotin， S. M. Soubbotin. Patterns of Potential Answer Expressions as Clues to the Right Answers. Tenth Text REtrieval Conference (TREC-10). Gaithersburg， MD. November 13-16，2001.

[6] H. Yang， T.-S. Chua. The Integration of Lexical Knowledge and External Resources for Question Answering. Eleventh Text REtrieval Conference (TREC-2002). Gaithersburg， MD. November 2002.

[7] Kwok， Etzioni， Weld: Scaling Question Answering to the Web. Proc. WWW10， Hong Kong.

[8] Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen

Soderland. 2007. Textrunner: Open information extraction on the web. In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, NAACL-Demonstrations ’07, pages 25–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

[9] Fader, A., Soderland, S., and Etzioni, O. Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing(2011), Association for Computational Linguistics, pp. 1535–1545.

[10] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. 2011. Open information extraction: The second generation. In IJCAI, volume 11, pages 3–10.

[11] Fei Wu and Daniel S. Weld. 2010. Open information extraction using wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL ’10, pages 118–127, Stroudsburg, PA, USA. Association for Computational Linguistics.

[12] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open language learning for information extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL ’12, pages 523–534, Stroudsburg, PA, USA. Association for Computational Linguistics.

[13] Alan Akbik and Jügen Bro. 2009. Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In WWW Workshop.

[14] Alan Akbik and Alexander L?ser. 2012. Kraken: Nary facts in open information extraction. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, AKBC-WEKEX ’12, pages 52–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

[15] T. Kwiatkowski, L. Zettlemoyer, S. Goldwater, and M. Steedman, Lexical generalization in ccg grammar induction for semantic parsing, in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1512–1523.

[16] P. Liang, M. I. Jordan, and D. Klein, Learning dependencybased compositional semantics, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011, pp. 590–599.

[17] J. M. Zelle and R. J. Mooney, Learning to parse database queries using inductive logic programming, in Proceedings of the National Conference on Artificial Intelligence, 1996, pp. 1050–1055.

[18] Y. W. Wong and R. J. Mooney, Learning for semantic parsing with statistical machine translation, in Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006, pp. 439–446.

[19] L. S. Zettlemoyer and M. Collins, Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars, in Proceedings of the 21st Uncertainty in Artificial Intelligence, 2005, pp. 658–666.

[20] T. Kwiatkowski, L. Zettlemoyer, S. Goldwater, and M. Steedman, Inducing probabilistic ccg grammars from logical form with higher-order unification, in Proceedings of the 2010 conference on Empirical Methods in Natural Language Processing, 2010, pp. 1223–1233.

[21] S. Clark and J. R. Curran, Log-linear models for widecoverage ccg parsing, in Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003, pp. 97–104.

[22] Bordes, Antoine, Jason Weston, and Nicolas Usunier. Open question answering with weakly supervised embedding models. Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2014. 165-180.

[23] Bordes, Antoine, Sumit Chopra, and Jason Weston. Question Answering with Subgraph Embeddings, in Proceedings of the conference on Empirical methods in natural language processing, 2014.

[24] Graves, Alex, Wayne, Greg, and Danihelka, Ivo. Neural turing machines. arXiv preprint arXiv:1410.5401, 2014.

[25] Weston, Jason, Chopra, Sumit, and Bordes, Antoine. Memory networks. CoRR, abs/1410.3916, 2014.

[26] Sukhbaatar, Sainbayar, Szlam, Arthur, Weston, Jason, and Fergus, Rob. End-to-end memory networks. Proceedings of NIPS, 2015.

[27] Bordes, Antoine, Usunier, Nicolas, Chopra, Sumit, and Weston, Jason. Large-scale simple question answering with memory networks. arXiv preprint arXiv:1506.02075, 2015.

[28] Stefan Schoenmackers, Oren Etzioni, Daniel S. Weld, and Jesse Davis. 2010. Learning first-order horn clauses from web text. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP ’10, pages 1088–1098, Stroudsburg, PA, USA. Association for Computational Linguistics.

[29] N. Lao, T.M. Mitchell, W.W. Cohen. Random Walk Inference and Learning in A Large Scale Knowledge Base. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011.

[30] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In International Conference on Learning Representations (ICLR).

[31] K. Gu, J. Miller, and P. Liang. Traversing knowledge graphs in vector space. In EMNLP, 2015.