

Uma revisão breve sobre perguntas complexas em bases de conhecimento para sistemas de perguntas e respostas

Jorão Gomes Jr.¹, Rômulo Chrispim de Mello², Ana Beatriz Kapps dos Reis²,
Victor Ströele², Jairo Francisco de Souza^{1,2}

¹ Laboratório de Aplicações e Inovação em Computação (LApIC)
Programa de Pós-Graduação em Ciência da Computação – UFJF

² Departamento de Ciência da Computação – UFJF
36.360-900 – Juiz de Fora – MG – Brasil

{joraojunior, romulomello, anabeatrizkapps}@ice.ufjf.br

{victor.stroele, jairo.souza}@ice.ufjf.br

Abstract. *The advance on Question Answering systems has achieved significant results, and new related problems have arisen. Few studies have addressed the Complex Knowledge Base Question Answering (C-KBQA) problem. This work presents an overview of C-KBQA systems. A collection of 54 papers were systematically selected, and a map of the methods, approaches, trends, future directions, and gaps in the C-KBQA research was performed. This study shows that C-KBQA systems need to handle multi-hop and constraint questions, and two approaches are usually used to address this problem.*

Resumo. *O avanço nos sistemas de Question Answering alcançou resultados importantes e novos problemas relacionados, como Complex Question Answering e Knowledge Base Question Answering, surgiram. No entanto, faltam estudos que analisam o problema e abordagens para Complex Knowledge Base Question Answering (C-KBQA). Este trabalho preenche essa lacuna apresentando uma visão geral do C-KBQA. Uma coleção de 54 artigos foi selecionada e um mapa dos métodos, abordagens, tendências e lacunas sobre C-KBQA foi realizado. É mostrado que as questões de múltiplos saltos e restritivas são os dois tipos de questões abordadas na literatura. Três etapas foram identificadas para criar um sistema C-KBQA e duas abordagens são geralmente usadas.*

1. Introdução

Sistemas de Perguntas e Respostas (do inglês *Question Answering* - QA) têm o propósito de recuperar a informação mais relevante (resposta) para uma pesquisa (pergunta) feita por um usuário [Croft et al. 2010]. Diferente dos motores de busca, os sistemas de QA visam encontrar as respostas exatas para uma pergunta em linguagem natural (do inglês *Natural Language Question* - NLQ) [Yin et al. 2014, Rodrigo and Penas 2017]. Para fazer isso, os sistemas de QA precisam reconhecer as informações dentro de uma NLQ. Esta tarefa implica na identificação de objetos relevantes e suas conexões, extraindo as principais descrições ou ideias que estão contidas em uma pergunta. O mapeamento de uma pergunta para seus principais assuntos (conceitos, organizações, pessoas, etc) é uma tarefa que tem sido explorada para sistemas de QA [Hao et al. 2019, Hua et al. 2020a].

Bases de conhecimento (do inglês *Knowledge Base* - KB) é um modelo de dados baseado em uma rede semântica que, usualmente, utiliza um formato de triplas para representar e relacionar as informações contidas em um domínio de dados [Ji et al. 2020]. Os sistemas de QA que fazem uso de KBs são chamados *Knowledge Base Question Answering* (KBQA). Estes sistemas usam essas estruturas semânticas, por exemplo, Freebase ou Wikidata para responder diretamente uma pergunta. Portanto, os sistemas de KBQA extraem os principais recursos do texto e os mapeiam em uma KB para responder a uma pergunta. O uso de KBs fornece um resultado mais preciso e conciso, uma vez que uma NLQ pode ser mapeada para consultas estruturadas em uma KB [Cui et al. 2019].

Mesmo com o uso de KBs, os sistemas de QA ainda precisam lidar com diferentes tipos de perguntas. Pode-se dividir esses tipos de perguntas em dois grupos: simples e complexas. Perguntas simples são aquelas que contêm respostas diretas e apenas um fato precisa ser detectado para obter a resposta [Bordes et al. 2015]. Já perguntas complexas precisam de mais informações do que as explícitas que podem ser extraídas diretamente. Nesse caso, é necessário realizar consultas avançadas para coletar a resposta dos KBs, como a exploração de relações indiretas entre entidades, multi-relações, restrições qualitativas e quantitativas, entre outras [Bao et al. 2016]. Atualmente, os sistemas de QA obtêm melhores resultados ao responder a perguntas simples e, por conta disso, os sistemas de QA para perguntas complexas estão recebendo atenção [Ding et al. 2019, Bhutani et al. 2020].

Assim, este trabalho visa apresentar uma visão geral de sistemas *Complex Knowledge Base Question Answering* (C-KBQA). A compreensão das soluções para C-KBQA inclui a investigação das técnicas mais utilizadas, as principais soluções atuais, onde essas soluções são aplicadas, e a identificação dos principais desafios desse campo de estudo. As principais contribuições deste trabalho são: (i) uma visão geral do estado da arte em C-KBQA; (ii) uma coleção de 54 artigos selecionados sistematicamente a partir de 898 artigos; (iii) um mapeamento de métodos e abordagens usadas no cenário de perguntas e respostas complexas. É mostrado que os sistemas de C-KBQA tentam resolver dois tipos de perguntas complexas: perguntas com múltiplos saltos e perguntas restritivas. Além disso, existem três etapas para construir sistemas de C-KBQA e duas abordagens são usadas nesse processo. Por fim, são discutidos os desafios e tendências para C-KBQA.

Este trabalho é assim estruturado: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve o protocolo do mapeamento sistemático. A Seção 4 apresenta as discussões do mapeamento. A Seção 5 apresenta as tendências e desafios futuros para os sistemas de C-KBQA. A Seção 6 apresenta as considerações finais do trabalho.

2. Trabalhos Relacionados

De acordo com [Dimitrakis et al. 2019], o problema de responder perguntas complexas foi reconhecido como um desafio para sistema de QA sobre dados ligados em [Höffner et al. 2017, Rodrigo and Penas 2017]. No entanto, faltam trabalhos que apresentem um mapeamento para esse problema no campo KBQA. Trabalhos anteriores focam em sistemas de perguntas e respostas simples e apresentaram uma breve introdução ao assunto de perguntas complexas, mesmo em revisões recentes de QA [Höffner et al. 2017, Wu et al. 2019b, da Silva et al. 2020]. Esses trabalhos não deixaram claro o problema do C-KBQA, quando surgiu e como ele vem sendo resolvido na

literatura.

Em [Höffner et al. 2017], os autores realizaram um levantamento sobre os desafios de responder a perguntas na Web semântica. Os autores reúnem um total de 62 sistemas desenvolvidos entre 2010 e 2015. No entanto, o problema da pergunta complexa não é discutido em detalhes. Os autores têm apenas uma seção que apresenta uma visão geral de alto nível do problema de responder a perguntas complexas. Isso era esperado uma vez que os artigos analisados pelos autores foram publicados até o início de 2015 e o assunto de pergunta complexa ainda não estava profundamente enunciada para ser pesquisada. [Höffner et al. 2017, Rodrigo and Penas 2017] discutem que os sistemas ainda não estão preparados para ter um bom desempenho dentro do cenário de perguntas e respostas complexas como já fazem no cenário de perguntas simples. No entanto, com o crescimento das pesquisas na área de perguntas complexas, fica evidente o contrário.

Além disso, em [Dimitrakis et al. 2019] os autores apresentam um mapeamento geral do cenário KBQA. No entanto, eles discutem apenas pequenas lacunas no cenário de perguntas complexas mostrando também que as pesquisas em perguntas complexas precisam de melhorias. Isso confirma a necessidade de estudar e compreender este problema. O principal objetivo deste trabalho é ajudar outros pesquisadores(as) da área e permitir que eles(as) entendam os principais conceitos e desafios do C-KBQA.

Este trabalho se diferencia dos demais por realizar uma análise dos métodos propostos para os sistemas de C-KBQA. Além disso, realizamos uma abordagem sistemática com o objetivo de encontrar os trabalhos de forma mais precisa e imparcial, dando a outros pesquisadores a opção de reproduzir os processos que foram feitos. O trabalho apresenta um mapeamento quantitativo e qualitativo de diversos aspectos do CKBQA e permite uma visão geral do que tem sido feito na área. Portanto, este estudo facilita o primeiro contato de novos pesquisadores com o tema.

3. Mapeamento Sistemático

O processo utilizado neste trabalho foi baseado no protocolo apresentado por [Neiva et al. 2016]. O processo consiste nas seguintes etapas: (i) definição das questões de pesquisa, (ii) seleção dos termos de pesquisa relevantes, (iii) definição dos critérios de exclusão e (iv) seleção dos repositórios de pesquisa.

As questões de pesquisa objetivam categorizar e criar uma visão geral da literatura, descobrindo tópicos cobertos na área de pesquisa [Petersen et al. 2015]. Foram definidas duas questões de pesquisa (QP): “Que tipos de perguntas são definidas como perguntas complexas para os sistemas de C-KBQA na literatura e como elas tem sido resolvidas?”(QP1) e “Quais são os recursos e métodos usados para resolver o problema C-KBQA?”(QP2). O método PICOC (*Population, Intervention, Comparison, Outcome, e Context*) foi usado para definir o escopo da pesquisa [Petticrew and Roberts 2006]. A Tabela 1 descreve os elementos PICOC e os termos de pesquisa para cada elemento. A *string* de busca foi criada usando os termos de pesquisa, onde cada elemento foi separado por um *AND* e seus sinônimos foram separados por *OR*.

Foram definidos 6 critérios de exclusão (CE) para que trabalhos não relacionados fossem excluídos da avaliação: Duplicado (CE1); Não apresenta um sistema de KBQA (CE2); Não trata o problema de pergunta complexa (CE3); Não escrito em inglês (CE4);

Tabela 1. Definição do PICOC e os termos de pesquisa.

Elemento	Descrição	Termos e Sinônimos
Population (P)	Artigos que apresentam sistemas de QA	<i>question answering, qa, semantic search, search engine, answering engine</i>
Intervention (I)	Abordagens para resolver perguntas complexas	<i>complex question, complex information, complex queries, complex query, complex answer</i>
Comparison (C)	-	-
Outcome (O)	As soluções para responder a perguntas complexas	<i>method, technique, algorithm, approach, application, system</i>
Context (O)	Sistemas de QA que fazem uso de bases de conhecimento	<i>knowledge base, knowledge graph, kb, kg, linked data, linked open data, lod, semantic web, semantic data</i>

Literatura cinzenta¹ (CE5); Indisponíveis na íntegra (CE6). As bases *Scopus*, *Google Scholar*, *ISI Web of Science*, *IEEE Digital Library* foram selecionados para a execução da *string* de busca. A coleta dos artigos foi finalizada em 17 de novembro de 2020 e 898 artigos foram obtidos. Os artigos duplicados foram excluídos e a leitura do título e do resumo foi realizada. Os artigos restantes tiveram suas introduções e conclusões lidas. Foram selecionados 45 artigos por responderem as QP. Finalmente, 9 artigos foram adicionados após a etapa de *snowballing*, resultando nos 54 artigos mapeados. Esta etapa tem como objetivo encontrar artigos relevantes que não foram retornados pela *string* de busca, observando os trabalhos citados nas referências dos artigos aceitos.

4. Relatório do mapeamento sistemático

Esta seção discute os dados coletados para responder às perguntas da pesquisa. Os artigos analisados estão disponíveis e informações adicionais podem ser encontradas aqui².

4.1. Resultados relativos a QP1

Foram identificados dois subtipos de perguntas complexas: perguntas com múltiplos saltos e perguntas com restrições. Foram encontrados 52 trabalhos abordando perguntas com múltiplos saltos e 17 trabalhos abordando perguntas com restrições. Alguns trabalhos tentam resolver os dois tipos de perguntas complexas simultaneamente e que o termo “perguntas com multi-restrições” se refere a união das categorias listadas anteriormente. Ao abordar perguntas com múltiplos saltos, um sistema C-KBQA tem que lidar com várias entidades que podem ser extraídas de uma pergunta. As entidades detectadas nessas perguntas precisam ser vinculadas e tratar relações indiretas, ao contrário de perguntas simples que podem ser respondidas diretamente. As triplas (sujeito, predicado, objeto) são exploradas dentro de uma KB, e os sistemas fazem saltos entre os objetos detectados na pergunta e as entidades/predicados do KB para obter a resposta. Quando relacionado a questões com restrições, a pergunta inclui algumas restrições que limitam as opções de resposta para uma determinada pergunta [Shin and Lee 2020]. Essas restrições podem ser de vários tipos, por exemplo, temporal (“... antes de 2000”), ordinal (“O primeiro que ...”), quantitativa (“... tendo mais de 5 ...”), dentre outras. Essas restrições podem modificar o assunto principal e conseqüentemente, alterar a resposta a ser obtida.

¹ Artigos sem revisão dos pares, como, por exemplo, *pre-prints*, relatórios técnicos, patentes, etc.

² A lista de artigos e informações adicionais podem ser acessadas em <https://github.com/lapic-ufjf/CKBQA-systematic-mapping-2021>

Os artigos da categoria múltiplos saltos tentam resolver este problema detectando as entidades e relações, criando uma lista de possíveis candidatos para realizar os saltos entre as relações e predicados. Para as perguntas com restrições, a criação de um modelo de perguntas, regras de restrições e decomposição de perguntas é o caminho mais utilizado. O custo computacional é um dos principais problemas que artigos procuram resolver em C-KBQA. Primeiro, uma pergunta pode precisar de muitos saltos para obter a resposta. Em segundo lugar, os sistemas de C-KBQA podem gerar um alto nível de lista de candidatos, uma vez que é necessário certo tempo para processar várias conexões de triplas do KB. Finalmente, uma vez que os dois foram resolvidos, é necessário limitar a lista de candidatos apenas dentro das restrições. A maioria dos artigos tem um módulo apenas para podar e classificar os melhores candidatos para esses problemas.

4.2. Resultados relativos a QP2

Em geral, pode-se dividir o pipeline C-KBQA em três etapas: Análise da pergunta, Representação da pergunta e Classificação de candidatos. Na etapa de análise da pergunta, é realizada a seleção do tipo de pergunta e a identificação dos principais temas. Primeiro, é preciso encontrar os tipos de perguntas que um NLQ possa corresponder, como “quando”, “o quê”, “como”, entre outros. Esses tipos de perguntas são chamados de *wh-questions*. A marcação Part-of-speech (POS) e as árvores de dependência são geralmente usadas para extrair a semântica da frase e para entender qual é o tipo da pergunta. Além disso, o Reconhecimento de Entidade Nomeada e os métodos de reconhecimento de relação são realizados para extrair as principais informações da frase. O DBpedia Spotlight [Mendes et al. 2011], S-MART [Yang and Chang 2015] e Stanford Named Entity Recognizer³ são exemplos de ferramentas usadas nesta etapa.

Na etapa de representação da pergunta é realizado o mapeamento semântico. Após a análise da pergunta, o sistema C-KBQA possui as entidades e relações extraídas de uma NLQ. No entanto, é necessário mapear e conectar as entidades e relações identificadas para corresponder à estrutura da KB. As pesquisas seguem dois caminhos: abordagens baseadas em análise semântica e abordagens baseadas em redes neurais. Abordagens baseadas em análise semântica (ou baseadas em regras) mapeiam as questões e as informações extraídas em um conjunto de formulários lógicos para serem posteriormente transformadas em modelos de consulta de tripla do KB. Abordagens baseadas em rede neural (ou livre de regras) usam redes neurais para identificar automaticamente os tipos de perguntas e quais são os padrões de consulta mais apropriados para obter a resposta. Ambas as abordagens podem criar um conjunto de candidatos que podem ser considerados como a resposta final. A Seção 4.2.1 apresenta os detalhes sobre essas duas abordagens.

Após a representação semântica e a geração dos candidatos, é realizada a classificação dos candidatos. O objetivo desta etapa é remover as respostas incorretas com base no tipo e na semântica da NLQ original. O melhor candidato é selecionado com base em uma função de avaliação. Para isso, alguns trabalhos utilizam métricas de avaliação como similaridade por cosseno ou função de log-verossimilhança. Em outros casos, um modelo de aprendizado de máquina (ex. SVM) é treinado para coletar os padrões e classificar os candidatos em uma lista dos melhores resultados.

Por fim, as etapas acima são realizadas em duas fases: experimentos *offline* e

³<https://nlp.stanford.edu/software/CRF-NER.html>

experimentos *online*. Os experimentos *offline* têm como foco a geração de materiais a serem utilizados nos experimentos *online* e não possuem interação com os usuários. A etapa de criação do candidato é criada na etapa *offline*, pois precisa predefinir o conjunto de regras ou treinar um modelo de rede neural. Os modelos de aprendizado de máquina também precisam ser treinados na etapa *offline* para serem usados na etapa de classificação de candidatos. Nos experimentos *online*, todas as etapas do pipeline são executadas para obter a resposta, utilizando os modelos gerados na etapa *offline*.

4.2.1. Representação das perguntas e geração de candidatos

As abordagens de representação de perguntas e geração de candidatos para C-KBQA podem ser divididas em dois tipos: abordagens baseadas em análise semântica e baseadas em redes neurais. Foram encontrados 47 trabalhos abordando abordagens baseadas em análise semântica e 31 trabalhos abordando abordagens baseadas em redes neurais.

A abordagem baseada em análise semântica (baseada em regras) é o mapeamento de uma NLQ para sua representação de significado [Tong et al. 2019]. A NLQ é transformada em uma representação intermediária que pode ser posteriormente representada como uma forma lógica [Trivedi et al. 2017]. As abordagens nesta categoria criam uma lista de regras predefinidas como expressão lógica, *tempalte questions*, aproximação de sub-grafos, regras gramaticais ou outra estrutura semântica equivalente que representa a semântica da NLQ [Wu et al. 2019a]. Quando a NLQ corresponde a esses padrões, é mais fácil consultá-lo em uma KB. Sistemas de C-KBQA fazem uso da análise semântica para mapear os assuntos da NLQ em dados semanticamente estruturados, mapear os dados estruturados dentro da representação tripla KB e, finalmente, responder à NLQ de forma concisa. Inicialmente, é realizada a quebra da NLQ em sub perguntas. Após a extração do assunto principal da NLQ, cria-se a representação intermediária de cada sub pergunta. Em seguida, a união das formas lógicas de cada NLQ na representação da KB. Por fim, a tripla da KB é criada e pode ser executada posteriormente no esquema da KB.

Esse mapeamento facilita o processo de extração dos termos relevantes de uma pergunta e o processo de conexão com as triplas do KB. No entanto, por ser dependente de regra, às vezes essas abordagens podem não funcionar bem para NLQ que não tem um formato de decomposição e recomposição no conjunto de regras. Essas abordagens também podem ter problemas com a escalabilidade da consulta ao usar muitos relacionamentos, devido a grande quantidade de regras candidatas que, conseqüentemente, aumentam o número de triplas do KB [Agarwal et al. 2019].

Abordagens baseadas em redes neurais (livres de regras) tentam usar arquiteturas de redes neurais para codificar perguntas e respostas em um modelo de espaço vetorial e selecionar os padrões de consulta mais apropriados para obter a resposta [Luo et al. 2018]. Com isso, é possível identificar tipos de questões e padrões comuns para responder a um determinado tipo de pergunta [Luo et al. 2018]. Os trabalhos nesta categoria geralmente são compostos de uma camada para coletar os *embeddings* e uma camada de rede neural.

Primeiramente, a camada de *embeddings* é usada para transformar a NLQ em uma sequência de vetores de palavras ou vetores de frases. *Embeddings* de palavras reduzem a complexidade computacional, já que as operações de vetores e matrizes são rápidos

de calcular. Word2vec [Mikolov et al. 2013], GloVe [Pennington et al. 2014] e FastText [Bojanowski et al. 2017] são exemplos de arquiteturas de *embeddings* de palavras usadas na literatura por sistemas de C-KBQA. Em seguida, uma rede neural profunda é usada. Nesta etapa, Redes Neurais Recorrentes (RNNs) é a arquitetura mais comum. A RNN é usada para transmitir uma cadeia de informações históricas por meio de uma sequência de unidades de rede neural. A RNN funciona como uma arquitetura de rede em cadeia e analisa a entrada atual e a saída anterior em cada etapa de tempo durante o processamento de dados sequenciais. Assim, a RNN pode extrair a propagação do contexto de informação de uma NLQ. Além disso, a RNN tem sido usada como arquiteturas de codificação e decodificação (*seq-2-seq*). Nesse processo, uma unidade RNN é usada para codificar a NLQ e outra RNN é usada para coletar as informações históricas da NLQ e decodificá-las na sequência de resposta. *Gated Recurrent Unit* (GRU), *Long Short-Term Memory* (LSTM) e suas variantes são as mais usadas para realizar esta etapa, uma vez que essas RNNs podem lidar melhor com o problema da dissipação do gradiente.

Um dos principais problemas da RNN é a queda no desempenho para sequências mais longas e complexas. Para resolvê-lo, trabalhos recentes utilizam o mecanismo de atenção para enfatizar as partes mais relevantes de uma NLQ e preservar o contexto das sentenças [Bhutani et al. 2019, Ding et al. 2019, Tong et al. 2019, Bhutani et al. 2020]. Embora RNNs sejam amplamente utilizadas, Redes de Memória [Miller et al. 2016, Hao et al. 2019, Saha et al. 2018, Hua et al. 2020b, Hua et al. 2020a] e Redes Neurais Convolucionais [Hu et al. 2018, Bao et al. 2016] podem ser usadas nesta etapa. Independente da rede utilizada, esta etapa de treinamento depende de recursos computacionais e pode levar muito tempo para obter resultados satisfatórios, além de depender de conjuntos de dados suficientemente grandes e diversificados para evitar problemas de *overfitting*. Entretanto, bases de dados para sistemas de C-KBQA ainda são um problema em aberto para o cenário, por ser uma campo de pesquisa recente.

As abordagens de análise semântica baseada em redes neurais tentam resolver perguntas complexas usando uma combinação de análise semântica e arquiteturas de redes neurais e estão se tornando o estado da arte [Luo et al. 2018, Ding et al. 2019]. Essa abordagem consiste em treinar uma rede neural para corresponder a um conjunto de regras de análise semântica, em vez de apenas a resposta final. Assim, o modelo aprende a semântica por trás de um NLQ em vez de apenas aprender os padrões de consulta mais apropriados para obter a resposta. Essas abordagens tendem a ser mais generalistas, pois aprendem o passo a passo para responder uma pergunta.

5. Tendências e desafios futuros

Redes Neurais Recorrentes têm se tornado bastante utilizadas neste campo, porém trabalhos recentes mostram que o uso de modelos pré-treinados e transferência de aprendizagem podem ser uma nova opção para treinar mais rapidamente novas soluções. [Lukovnikov et al. 2019] mostrou que modelos pré-treinados como o BERT obtêm bons resultados para responder a perguntas simples. Os autores afirmam que esses modelos podem ter um impacto maior em sistemas de C-KBQA. Avanços em C-KBQA e Deep learning podem criar arquiteturas que demandam menos memória e têm menor custo de treinamento, permitindo a popularização dos modelos KBQA para domínios específicos.

Os sistemas de KBQA são dependentes das informações contidas nas KB e não

é trivial criar sistemas que utilizem informações atualizadas em suas respostas, uma vez que o sistema aprende a responder às perguntas relacionadas ao conjunto de dados de treinamento e teste. É necessário comparar esses sistemas com usuários reais e também avaliá-los em cenários reais. Uma opção para fazer isso é testá-los usando plataformas de *crowdsourcing* e avaliar sua usabilidade. Os sistemas podem fazer uso de abordagens *cross-language* aplicadas ao KB [Schumacher et al. 2020], desambiguação [Kartsaklis et al. 2018], busca em grafos [Namaki et al. 2017], *embedding* [Dettmers et al. 2018] e *reasoning* [Chen et al. 2020]. Além disso, novas técnicas de compreensão de linguagem natural podem melhorar sistemas de perguntas e respostas, como avanços no tratamento de perguntas com ruído [Zhang et al. 2018], novos analisadores de dependência e métodos de rotulagem de função semântica [Zheng and Kordjamshidi 2020], e raciocínio de senso comum [Liu et al. 2020].

Por fim, esses sistemas podem ser frágeis e espúrios. Frágeis porque ainda não são robustos o suficiente e podem falhar em responder uma pergunta quando apenas algumas partes da pergunta são um pouco modificadas, mesmo se o significado principal for preservado [Jia and Liang 2017]. Além disso, são espúrios porque, mesmo se a mesma arquitetura for treinada várias vezes no mesmo conjunto de dados, cada modelo pode memorizar artefatos e vieses diferentes em vez de realmente aprender e, portanto, falhar no desempenho de generalização [McCoy et al. 2019]. É necessário lidar cuidadosamente com a etapa de pré-processamento do conjunto de dados, separá-la com classes de treinamento balanceadas e também tentar explorar algumas paráfrases de perguntas para garantir que a maioria dos tipos de perguntas está sendo explorada.

6. Considerações finais

Este artigo apresentou um mapeamento sistemático sobre *Complex Knowledge Base Question Answering Systems* (C-KBQA). Sistemas de C-KBQA tentam lidar com dois tipos de perguntas complexas: perguntas com múltiplos saltos e perguntas com restrições.

Além disso, foi apresentada uma visão geral do processo de construção de sistemas de C-KBQA e como as principais abordagens são realizadas. Os trabalhos usam duas abordagens principais: Análise Semântica e Redes Neurais. No entanto, nos últimos anos, a combinação dessas duas abordagens tornou-se o estado da arte (chamada de análise semântica baseada em redes neurais). Por fim, foi discutido como a área C-KBQA está em crescimento e que novos sistemas de C-KBQA ou módulos que tentam melhorar partes das etapas de C-KBQA são prováveis de serem criados nos próximos anos.

Como trabalhos futuros estão a análise detalhada de conjuntos de dados para C-KBQA e as métricas de avaliação mais usadas nesse cenário, que ainda são desafios em aberto. Por fim, um panorama das publicações e dos locais mais relevantes através dos anos mostrará como a área do C-KBQA vem crescendo e recebendo mais atenção.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Código de Financiamento 001

Referências

Agarwal, P., Ramanath, M., and Shroff, G. (2019). Retrieving relationships from a knowledge graph for question answering. In Azzopardi, L., Stein, B., Fuhr, N., Mayr, P.,

- Hauff, C., and Hiemstra, D., editors, *Advances in Information Retrieval*, pages 35–50, Cham. Springer International Publishing.
- Bao, J., Duan, N., Yan, Z., Zhou, M., and Zhao, T. (2016). Constraint-based question answering with knowledge graph. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2503–2514, Osaka, Japan. The COLING 2016 Organizing Committee.
- Bhutani, N., Zheng, X., and Jagadish, H. V. (2019). Learning to answer complex questions over knowledge bases with query composition. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19*, page 739–748, New York, NY, USA. Association for Computing Machinery.
- Bhutani, N., Zheng, X., Qian, K., Li, Y., and Jagadish, H. (2020). Answering complex questions by combining information from curated and extracted knowledge bases. In *Proceedings of the First Workshop on Natural Language Interfaces*, pages 1–10, Online. Association for Computational Linguistics.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Bordes, A., Usunier, N., Chopra, S., and Weston, J. (2015). Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*, abs/1506.02075.
- Chen, X., Jia, S., and Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948.
- Croft, W. B., Metzler, D., and Strohman, T. (2010). *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Publishing, USA.
- Cui, W., Xiao, Y., Wang, H., Song, Y., Hwang, S., and Wang, W. (2019). KBQA: learning question answering over QA corpora and knowledge bases. *CoRR*, abs/1903.02419:565–576.
- da Silva, J. W. F., Venceslau, A. D. P., Sales, J. E., Maia, J. G. R., Pinheiro, V. C. M., and Vidal, V. M. P. (2020). A short survey on end-to-end simple question answering systems. *Artificial Intelligence Review*, 53(7):5429–5453.
- Dettmers, T., Minervini, P., Stenetorp, P., and Riedel, S. (2018). Convolutional 2d knowledge graph embeddings. In McIlraith, S. A. and Weinberger, K. Q., editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1811–1818, New Orleans, Louisiana, USA. AAAI Press.
- Dimitrakis, E., Sgontzos, K., and Tzitzikas, Y. (2019). A survey on question answering systems over linked data and documents. *Journal of Intelligent Information Systems*, 55:1–27.
- Ding, J., Hu, W., Xu, Q., and Qu, Y. (2019). Leveraging frequent query substructures to generate formal queries for complex question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

- International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2614–2622, Hong Kong, China. Association for Computational Linguistics.
- Hao, Z., Wu, B., Wen, W., and Cai, R. (2019). A subgraph-representation-based method for answering complex questions over knowledge bases. *Neural Networks*, 119:57–65.
- Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., and Ngonga Ngomo, A.-C. (2017). Survey on challenges of question answering in the semantic web. *Semantic Web*, 8(6):895–920.
- Hu, S., Zou, L., and Zhang, X. (2018). A state-transition framework to answer complex questions over knowledge base. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2098–2108, Brussels, Belgium. Association for Computational Linguistics.
- Hua, Y., Li, Y.-F., Haffari, G., Qi, G., and Wu, T. (2020a). Few-shot complex knowledge base question answering via meta reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5827–5837, Online. Association for Computational Linguistics.
- Hua, Y., Li, Y.-F., Haffari, G., Qi, G., and Wu, W. (2020b). Retrieve, program, repeat: Complex knowledge base question answering via alternate meta-learning. In Bes-siere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3679–3686, Virtual, Japan. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Ji, S., Pan, S., Cambria, E., Marttinen, P., and Yu, P. S. (2020). A survey on knowledge graphs: Representation, acquisition and applications. *arXiv preprint arXiv:2002.00388*, abs/2002.00388.
- Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Vancouver, Canada. Association for Computational Linguistics.
- Kartsaklis, D., Pilehvar, M. T., and Collier, N. (2018). Mapping text to knowledge graph entities using multi-sense LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1959–1970, Brussels, Belgium. Association for Computational Linguistics.
- Liu, Y., Wan, Y., He, L., Peng, H., and Yu, P. S. (2020). Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning.
- Lukovnikov, D., Fischer, A., and Lehmann, J. (2019). Pretrained transformers for simple question answering over knowledge graphs. In *International Semantic Web Conference*, volume abs/2001.11985, pages 470–486. Springer.
- Luo, K., Lin, F., Luo, X., and Zhu, K. (2018). Knowledge base question answering via encoding of complex query graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2185–2194, Brussels, Belgium. Association for Computational Linguistics.

- McCoy, R. T., Min, J., and Linzen, T. (2019). Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*, abs/1911.02969:217–227.
- Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8, New York, NY, USA. Association for Computing Machinery.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- Miller, A., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A., and Weston, J. (2016). Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.
- Namaki, M. H., Chowdhury, F. A., Islam, M., Doppa, J., and Wu, Y. (2017). Learning to speed up query planning in graph databases. *Proceedings of the International Conference on Automated Planning and Scheduling*, 27(1):9.
- Neiva, F. W., David, J. M. N., Braga, R., and Campos, F. (2016). Towards pragmatic interoperability to support collaboration: A systematic review and mapping of the literature. *Information and Software Technology*, 72:137–150.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Petersen, K., Vakkalanka, S., and Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18.
- Petticrew, M. and Roberts, H. (2006). Systematic reviews in the social sciences: a practical guide. 2006. *Malden USA: Blackwell Publishing CrossRef Google Scholar*, 6:304–305.
- Rodrigo, A. and Penas, A. (2017). A study about the future evaluation of question-answering systems. *Knowledge-Based Systems*, 137:83–93.
- Saha, A., Pahuja, V., Khapra, M., Sankaranarayanan, K., and Chandar, S. (2018). Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):9.
- Schumacher, E., Mayfield, J., and Dredze, M. (2020). Cross-lingual transfer in zero-shot cross-language entity linking.
- Shin, S. and Lee, K.-H. (2020). Processing knowledge graph-based complex questions through question decomposition and recomposition. *Information Sciences*, 523:234–244.
- Tong, P., Zhang, Q., and Yao, J. (2019). Leveraging domain context for question answering over knowledge graph. *Data Science and Engineering*, 4(4):323–335.

- Trivedi, P., Maheshwari, G., Dubey, M., and Lehmann, J. (2017). Lc-quad: A corpus for complex question answering over knowledge graphs. In *International Semantic Web Conference*, pages 210–218, Cham. Springer, Springer International Publishing.
- Wu, L., Wu, P., and Zhang, X. (2019a). A seq2seq-based approach to question answering over knowledge bases. In *Joint International Semantic Technology Conference*, pages 170–181, Singapore. Springer, Springer Singapore.
- Wu, P., Zhang, X., and Feng, Z. (2019b). A survey of question answering over knowledge base. In *China Conference on Knowledge Graph and Semantic Computing*, pages 86–97, Singapore. Springer, Springer Singapore.
- Yang, Y. and Chang, M.-W. (2015). S-MART: Novel tree-based structured learning algorithms applied to tweet entity linking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 504–513, Beijing, China. Association for Computational Linguistics.
- Yin, W., Ge, W., and Wang, H. (2014). Cdqa: An ontology-based question answering system for chinese delicacy. In *2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, pages 1–7, Shenzhen, China. IEEE.
- Zhang, Y., Dai, H., Kozareva, Z., Smola, A., and Song, L. (2018). Variational reasoning for question answering with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):6069–6076.
- Zheng, C. and Kordjamshidi, P. (2020). SRLGRN: Semantic role labeling graph reasoning network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8881–8891, Online. Association for Computational Linguistics.