知識グラフを用いた対話型質問応答における推論方法の検討

A Survey of Inference Methods in Conversational Question Answering using Knowledge Graph

吉兼拓生† 谷津元樹‡ 森田武史‡
Takumi Yoshikane† Motoki Yatsu‡ Takeshi Morita‡

† 青山学院大学大学院 理工学研究科
‡ 青山学院大学 理工学部
† Graduate School of Science and Engineering, Aoyama Gakuin University.
‡ College of Science and Engineering, Aoyama Gakuin University.

要旨
近年, 様々な質問応答データセットが構築され, それに答えられる質問応答システムの研究が盛んに行われている. 多くの質問応答システムではテキストや，大規模知識グラフを知識源として用いている．一問一答のような質問応答もあれば，対話形式で質問を重ねていく質問応答もある．知識源の表現をそのまま用いて回答を行う質問応答もあれば，知識源の表現から推論を行う必要のあるものも存在する．本稿では，現在利用可能な質問応答データセットを知識源，質問の連続性，推論の必要性による3つの分類軸でまとめる．また，質問応答データセットに基づいて最近の質問応答システムの研究動向について調査し，それらの精度と課題について考察する．

1. はじめに

近年，様々な質問応答データセットが構築され，それに答えられる質問応答システムの研究が盛んに行われている．本稿では質問応答システムにおける知識源，質問の連続性，推論の必要性の 3 つの観点に着目する. 多くの質問応答システムではWikipedia の記事やニュース記事のようなテキストや, Wikidata や Freebase のような大規模知識グラフを知識源として用いている. 一問一答のような質問応答もあれば，対話形式で質問を重ねていく質問応答もある．知識源の表現をそのまま用いて回答を行う質問応答もあれば，知識源の表現から四則演算やカウント，真偽判定や集合演算などの推論を行う必要のあるものも存在する．本稿では，現在利用可能な質問応答データセットを知識源，質問の連続性，推論の必要性による 3 つの分類軸でまとめる．また，質問応答データセットに基づいて最近の質問応答システムの研究動向について調査し，それらの精度と課題について考察する.

2. 質問応答の分類

知識源による分類，質問の連続性による分類，推論の必要性による分類の3つの分類軸により質問応答タスクを見ていく．表1に各分類における質問と応答の例を示す.

2.1. 知識源による分類

知識源による分類は，質問に回答する際に使用するデータの種類による分類である．SQuAD2.0[1]のように Wikipedia の記事のような自然文のテキストが知識として与えられるもの，SimpleQuestions[2]のように知識グラフのようなテキスト以外のものが知識として与えられるものがある．質問応答データセットによってはデータセットが与えるデータの使用を想定する場合もあれば，一般公開されているデータの使用を想定する場合もある.

2.2. 質問の連続性による分類

質問の連続性による分類は, 複数の質問間の関連性による分類である. SQuAD 2.0 や SimpleQuestions, DROP[3]のように各質問が独立しているものや，CoQA[4]や CSQA[5]のように対話形式で質問と応答が行われ，以前の質問と回答に関連した質問がされるようなものが存在する.

2.3. 推論の必要性による分類

推論の必要性による分類は，知識源の表現を用いて回答するのではなく推論を行う必要があるかという観点での分類である．SQuAD 2.0 や SimpleQuestions のように，知識源から抽出してきたものをそのまま回答とすればよいタイプと，CoQA や DROP，CSQA のような知識源中のデータから数値演算や集合演算などの推論を行う必要があるタイプに分けることができる.

表 1 各分類における質問と応答の例

分類方法		質問の例	応答の例
知識源	テキスト	In what country is Normandy located?	France
	知識グラフ	which country locates hayashima?	www.freebase.com/m/03_3d （日本を示す Freebase の URI）
質問の連続性		Q1: Who had a birthday? Q2: How old would she be?	A1: Jessica A2: 80
推論の必要性		How many countries are diplomatically related to Italy?	74

表 2 各質問応答データセットの 3 つの分類軸における位置付けおよび質問数

データセット	知識源	連続性	推論の必要性	質問数
SQuAD	テキスト			15 万
SimpleQuestions	知識グラフ			10 万
CoQA	テキスト	✓	✓	12 万 7 千
DROP	テキスト		✓	9 万 6 千
CSQA	知識グラフ	✓	✓	160 万

3. 質問応答データセット

本章では公開されていて利用可能な質問応答データセットについて述べる．表 2 に各質問応答データセットの 3 つの分類軸における位置付けおよび質問数を示す.

3.1. SQuAD 2.0

SQuAD 2.0 (https://rajpurkar.github.io/SQuAD-explorer/)は Wikipedia の記事の一部がテキストで知識源として与えられ，質問の連続性はなく，テキストの一部分を抽出してきたもので回答を行うデータセットである．前身の SQuAD 1.1[6]では答えがテキスト中にあることが保証されている 10 万以上の質問が含まれていたが，SQuAD 2.0 では回答不可能な質問が追加され，15 万以上の質問が含まれている.

3.2. SimpleQuestions

SimpleQuestions (https://research.fb.com/downloads/babi/)は, 知識グラフデータベース Freebase から抽出を行って作成した FB2M を知識源として用い，質問の連続性はなく，知識グラフデータベースから抽出したデータをそのまま回答するデータセットである．FB2M には 200 万種類以上のエンティティと 6 千種類以上の関係が存在している．SimpleQuestions には単一のデータを参照して回答する単純な形式の質問のみが 10 万以上含まれている.

3.3. CoQA

CoQA (https://stanfordnlp.github.io/coqa/)は Wikipedia や CNN のニュース記事など 7 箇所から取得されたテキストが知識源として与えられ，質問応答を行うため質問に連続性があり，Yes/No やカウント，選択など抽出したままでない形で回答する質問も含まれるデータセットである．抽出したままでない形で回答する質問は全体の約 1/3 で，そのうちの 3/4 以上が Yes/No で答えるものである．自然な応答になるようにテキストの表現に単語を挿入・削除する場合もある．回答の他にテキストの根拠となる部分も提示する．8 千を超える対話に，12 万 7 千以上の質問が含まれている.

3.4. DROP

DROP (https://allennlp.org/drop)は Wikipedia の記事の一部がテキストで知識源として与えられ，質問の連続性はなく，数値演算やカウントなどの推論を行う必要があるデータセットである．テキストの一部分または複数の部分を抽出する必要のある質問も存在する．9 万 6 千を超える質問が含まれている.

3.5. CSQA

CSQA (https://amritasaha1812.github.io/CSQA/)は Wikidata の知識グラフを知識源として用い，対話形式で質問応答を行うため質問に連続性があり，集合演算・比較などの推論が必要となる質問も含まれるデ

ータセットである．約20万の対話に約160万の質問が含まれている．

4. 質問応答システムの研究

4.1. ALBERT

ALBERT[7]は，ラベルなしのテキストデータで事前学習したモデルに対して，個別のタスクで転移学習を行うことで様々な言語タスクに汎用的に使用できる BERT[8]を改良したものである．パラメータの共有によるモデルサイズの削減や学習方法の一部変更などが行われている．モデルサイズが減少し，それに伴い学習時間を短くできるにも関わらず, 複数のタスクで性能が向上していることが示されている．SQuAD 2.0 では単体で人の F1 スコア 89.45 を上回る 92.2 を達成しているほか，より高いスコアを出しているシステムに組み込まれてもいる．

4.2. BiLSTM + CRF

このシステムは，SimpleQuestions データセットに対する調査[9]の中で提案されたシステムである．BiLSTM（双方向LSTM）とCRF（条件付確率場）を用いて回答を行う．このSimpleQuestionsデータセットに対する調査ではデータセットの曖昧性などにより最大でも83.4%しか確実には正解できないとしながらも，提案されたシステムでは78.1%の精度を達成している．

4.3. RoBERTa + AT + KD

RoBERTa＋AT＋KD はBERT の改良である RoBERTa[10]と，敵対的学習(Adversarial Training, AT)，知識の蒸留(Knowledge Distillation, KD)を用いたシステムである．アンサンブル学習のために遺伝的アルゴリズムも用いている．CoQA において，人のスコア 88.8 を上回る 90.7 を達成している．代名詞・時制などの言い換えやカウントが必要な質問における間違いが典型的な間違いとして挙げられている．

4.4. QDGAT

QDGAT[11]は質問と与えられたテキストから含まれている数値やエンティティの間の関係を表すグラフを作成し，数値・エンティティの種類や関係を用いて推論を行うシステムである．QDGAT は DROP データセットで 86.38 の F1 スコアを達成している．質問を日付に関する質問，その他の数値に関する質問，テキストの部分を抽出して答える質問に分類すると，日付に関する質問でのスコアが他と比べて20ポイント程度低くなっている．

4.5. CIPITR

CIPITR[12]は回答を求める手続きを表現するプログラムを質問から生成する方法を学習することで，知識グラフデータベースに基づいた質問に推論を伴い回答を行うシステムである．生成するプログラムでは知識グラフのエンティティや関係，数値，集合などの 9 種類の変数の型と，データベースへのクエリや集合演算，比較などの20種類の演算子（関数）を使うことができる．他のシステムとの比較のために CSQA データセットの質問の連続性を取り除いて行われた実験では，(1) 抽出したものをそのまま回答する質問，(2) 集合演算を必要とする質問，(3) Yes か No で答えられる質問，(4) 量の推論を必要とする質問，(5) 量の推論とカウントを必要とする質問，(6) 比較を必要とする質問，(7) 比較とカウントを必要とする質問の 7 種類に質問を分類して F1 スコアを求めている．(1), (2), (3)の 3 つは 85 以上のスコアに到達しているが，量に関する(4), (5)では 60 に届かず，比較に関する(6), (7)では 20 に到達しないという結果になっている．スコアが低くなっている種類の質問になるにつれて必要となるプログラムの平均行数が長くなるという分析がされている．

5. 考察

知識源の表現で回答を行う質問システムでは知識源がテキストと知識グラフのどちらでも，テキストの方では回答が可能かどうかも含めて，現状で高い精度を実現できている．SimpleQuestions データセットに対する調査で曖昧性が解決できないとされているが，もしこれが知識グラフデータセットに起因するものであれば，ほかの知識グラフデータセットと組み合わせることで解決できる可能性がある．対話型の質問応答では，RoBERTa＋AT＋KD において言い換えの間違いが典型的にみられることと，通常の

テキストのみから言い換えの学習を行うことは難しいのではないかと考えられることから，言い換えの学習方法や言い換えに対応または特化したデータを作成することで精度を上げられるのではないかと考えられる．推論を伴う質問応答では，QDGAT において日付に関する質問の精度が他の質問に比べて20ポイント程度低いことや，RoBERTa＋AT＋KD においてカウントの質問の間違いが典型的にみられること，CIPITR において量や比較に関する質問の精度が低くなっている．現状では推論の種類ごとに異なる回答の求め方をうまく分けて学習することができていないことが考えられ，これを改善する方法を探す必要があると考えられる．CSQA のような知識グラフを用いる対話型で推論を伴う質問応答に対するシステムでは，上記のすべてを考慮する必要性が出てくることが考えられる．

6. まとめ

本稿では，現在利用可能な質問応答データセットを知識源，質問の連続性，推論の必要性による3つの分類軸でまとめた．また，質問応答データセットに基づいて最近の質問応答システムの研究動向について調査し，それらの精度と課題について考察した．今後はここで得られた知見を活かし，知識グラフを用いた推論を伴う対話型の質問応答システムの研究開発を行う予定である．

参考文献

[1] Rajpurkar, P., Jia, R., & Liang, P., Know What You Don't Know: Unanswerable Questions for SQuAD. Preprint, http://arxiv.org/abs/1806.03822, 2018.
[2] Bordes, A., Usunier, N., Chopra, S., & Weston, J., Large-scale Simple Question Answering with Memory Networks, http://arxiv.org/abs/1506.02075, 2015.
[3] Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., & Gardner, M., DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs, http://arxiv.org/abs/1903.00161, 2019.
[4] Reddy, S., Chen, D., & Manning, C. D., CoQA: A Conversational Question Answering Challenge, Transactions of the Association for Computational Linguistics, 7, 2019, 249–266.
[5] Saha, A., Pahuja, V., Khapra, M. M., Sankaranarayanan, K., & Chandar, S., Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph, 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 2018, 705–713.
[6] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P., SQuAD: 100,000+ Questions for Machine Comprehension of Text, http://arxiv.org/abs/1606.05250, 2016.
[7] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R., ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, http://arxiv.org/abs/1909.11942, 2019.
[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1, 2019, 4171–4186.
[9] Petrochuk, M., & Zettlemoyer, L., SimpleQuestions Nearly Solved: A New Upperbound and Baseline Approach, http://arxiv.org/abs/1804.08798, 2018.
[10] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V., RoBERTa: A Robustly Optimized BERT Pretraining Approach, http://arxiv.org/abs/1907.11692, 2019.
[11] Chen, K., Xu, W., Cheng, X., Xiaochuan, Z., Zhang, Y., Song, L., Wang, T., Qi, Y., & Chu, W., Question Directed Graph Attention Network for Numerical Reasoning over Text, http://arxiv.org/abs/2009.07448, 2020.
[12] Saha, A., Ansari, G. A., Laddha, A., Sankaranarayanan, K., & Chakrabarti, S., Complex Program Induction for Querying Knowledge Bases in the Absence of Gold Programs, Transactions of the Association for Computational Linguistics, 7, 2019, 185–200.