Multilingual Knowledge-Based Question Answering

Mengshi Ma

Institute of Computer Science

Abstract. The desire for knowledge is one of the most important characteristics of humans. Since knowledge graphs store knowledge in a graph structure rather than plain text, logical query can be applied to obtain implicit information, which is a more efficient and effective way. Many researchers published approaches for converting English natural language questions to SPARQL queries, which are used to query Wikidata and DBpedia. However, multilingual question answering systems so far perform well on English, but not on other languages, especially on languages with missing training data. In this paper, we develop a knowledge-graph multilingual question-answering system using the mt5 model from google, a pre-trained transformer model, which supports 101 languages. Our system treats SPARQL queries as another language and converts natural language questions to SPARQL queries. We explore zero-shot along with different fine-tuning strategies to reach the best performance. Our experiment results show the state-of-the-art performance of the system not only on English, but also on other languages for QALD-9-plus dataset.

Keywords: Transformers · KGQA · Wikidata

1 Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et

magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

The remainder of the paper starts with a presentation of related work (Sect. 2). It is followed by a presentation of hints on LaTeX (**??**). Finally, a conclusion is drawn and outlook on future work is made (Sect. 3).

2   Related Work

Language model

In Natural Language Processing (NLP), a probability distribution of words is defined as a Language model. Based on previous tokens and task, a language model predicts the next token continuously to generate a sequence. There are various approaches in language model. In feature-based approaches, linguistic features are extracted to understand a text. In transformer-based approaches, self-attention plays an significant part.

Feature-based Approach

Feature-based approaches, e.g. parsing, POS tagging and Term Frequency-Inverse Document Frequency, embed input text into a numerical representation by analysing linguistic features, which are essential for computer to understand languages. POS tagging assigns part-of-speech or grammatical tagging to each word in a text with e.g. nouns, verbs, adjectives and more. Term Frequency-Inverse Document Frequency (TF-IDF) weights a word based on how often it appears in a text and how many documents include it. If a word appears rarely, it contains more information, therefore has a higher weight and vice versa. Depending on the task, features are extracted from different aspects. Frequently, these features are combined with transformer-based approach to achieve better performance.

Transformer-based Approach

Transformers is another popular approach in Natural Language Processing using self-attention. A model is pre-trained in unsupervised or semi-supervised manner

on a large data set. For downstream tasks, the pre-trained model only needs to be fine-tuned on a few data for a short time, which saves a lot of time and compute resources. In 2017, Vaswani et al. introduced the first Transformers [32], which allows all input tokens to be processed in parallel via self-attention. In recent years, many transformer models are presented based on this paper, including new pre-training methods, different architectures, etc.

Bidirectional Encoder Representations from Transformers [7], referred to as BERT, is an encoder-based transformer, which is pre-trained with two tasks: masked language model and next sentence prediction on BooksCorpus [38] and English Wikipedia datasets. mBERT [6] is a variant of BERT which is pre-trained on 104 languages with Wikipedia data. After fine-tuning, it could be used for multilingual tasks e.g. translation. Rebalanced multilingual BERT (RemBert) [4] improves mBERT by reducing input embedding and enlarging output embedding, which shows that tweaking parameters is important in transformers and can lead to state-of-the-art performance.

GPT-2 [25] is a decoder only transformer, which functions like a language model: given all previous tokens to predict the next token. GPT-2 is pre-trained in unsupervised manner with 40 GB WebText and can be used for tasks, including question answering, summarization, and translation, even without fine-tuning.

Compare to BERT and GPT-2, Text-To-Text Transfer Transformer (T5) [26] combines encoder and decoder to convert NLP tasks into a text-to-text format. It is pre-trained on multi-tasking unsupervised and supervised. In unsupervised training, a BERT style approach is applied, namely mask tokens and predict next sentence. In supervised training, prefixes are given in input texts to define the tasks to be done. Thus, pre-trained T5 model can process several tasks well via adding the specific prefix to the text, e.g. "cola sentence: ..."for Sentence acceptability judgment (CoLA [34]), "summarize: ..."for summarization, "mrpc sentence1: ... sentence2: ..."for Paraphrasing/sentence similarity (MRPC [9]), and more. Xue et al. [35] released mT5 which utilizes the methodology as T5 for 101 languages. mT5 is trained on multilingual Colossal Clean Crawled Corpus (mC4) unsupervisedly, therefore, if must be fine-tuned and prefixes are not necessary, unless a multi-task fine-tuning is required. Similar to T5 model, BART [15] is also a sequence-to-sequence model, which consists of a encoder and a decoder and pre-trained with a combination of span masking and sentence shuffling objectives.

mBART [16] is a cross-lingual version of BART, pre-trained with denoising full texts in 25 languages. In multilingual transformers, XLM [13], XLM-RoBERTa [5] also achieves state-of-the-art performance and support 100 languages. Both of them optimize BERT architecture and pre-trained on 100 languages.

MARGE [14] provides another pre-training approach, reconstruct a document in one language by retrieving documents in other languages. MARGE model supports 26 languages and performs well even merely pre-training.

Knowledge Graph

In contrast to many information in the Internet are stored in plain text, Knowledge graphs manages in a structural fashion. Relations between entities connects them

together as edges in a graph, so that knowledge is linked instead of a lot of isolated island and implicit knowledge can be simply retrieved. Wikidata [33] is currently the largest knowledge graph, which contains about 100 Million instances in over 400 languages. It can be applied in e.g. enriching informations for applications, deriving new insights from stated data, etc.

DBpedia [1] consists of structured infobox information extracted from Wikipedia. It contains about 900 million triples by January 2021 and provides free APIs for online accessing and querying.

YAGO [29] is also one of the large knowledge graph, which combines information from Wikidata with WordNet [10]. In the newest YAGO 4 [20], they combine Wikidata with more user-friendly taxonomy from schema.org [11] while containing more than 64 million entities and 2 billion facts.

NELL [3] is short for Never-Ending Language Learner developed by Carlson et al., which performs a reading task and a learning task each day to extract information from Internet while learning to categorise noun phrases and find relations to pairs of noun phrases. NELLs expands itself rapidly, for example, it contains more than 242,000 new facts after running it 67 days.

OpenCyc [17] is a subset of Cyc knowledge base, a knowledge graph with about forty years history. OpenCyc contains 239,000 concepts in its 4.0 version, accessible via a interpreter on Linux and Microsoft Windows. However, it is not available after 2017.

conceptnet [28] represents facts in sentences in knowledge graph with terms and assertions, including knowledge from Open Mind Common Sense (OMCS), Wiktionary, Open Multilingual WordNet [2], etc. There are roughly 34 million statements in conceptnet and 304 language are supported in total, 10 of them have core support.

Wikidata, DBpedia and YAGO have SPARQL endpoint while NELL and conceptnet doesn't support SPARQL query. OpenCyc is currently not accessible. Wikidata has the most language supports in these knowledge graphs. Moreover, entities and relations are represented and stored with ids in Wikidata while these are all represented explicitly in DBpedia.

Knowledge Graph Question Answering (KGQA)

KGQA-Leaderboard [23] is a open source platform to record and compare scores of QA systems for different datasets. Presently, SGPT [27] is leading in English QALD-9 [31] dataset and QAnswer in the multilingual QALD-9-PLUS [22] dataset.

SGPT [27] combines feature-based approach and transformer-based approach. Linguistic features, e.g. POS-tag, dependency relation embedding, are extracted from the question, then fed into transformer encoders, lastly GPT-2 generates a SPARQL query. This approach avoids high cost of templates and provides deeper understanding of questions using lingustic features.

To et al. [30] derived a Linguistic Term Question Answering (LingTeQA), in which a natural language question is analysed with dependency tree and

converted to a question template, then a SPARQL template is retrieved from pre-generated template set to create SPARQL query.

qaSQP [36] is proposed by Zheng et al. Instead of detecting the template to use for a question, they use a structural query pattern framework: build a sketch of the query, then extend entities according to the question and lastly add constraints defined in the question.

Panchbhai et al. [18] used a completely different approach. They consider SPARQL as another language and use Neural Machine Translation to transform natural language questions to SPARQL queries. A Neural SPARQL Machine is built to process the Sequence-to-Sequence task, including three components, namely generator, learner and interpreter.

Multilingual KGQA

To the best of our knowledge, there are only a few researches focusing on multilingual KGQA.

Veyseh [24] proposed a four-stage cross-lingual approach: keyword extraction, keyword classification, entity linking and ontology type extraction and answer extraction using unified semantic representation of concepts, which is evaluated on Persian and Spanish.

Hakimov et al. also used semantic parsing mapping natural language input into logical form and developed AMUSE [12], which works for English, German ans Spanish.

Platypus [19] is presented by Tanon et al., integrating two analyzers: grammatical and template-based analyzer, for queuing Wikidata in 3 languages.

Diefenbach et al. [8] provided an approach over the Semantic Web, which works on 5 knowledge-graphs with 5 different languages.

Recently, transformer-based approach is applied to multilingual KGQA.

Zhou et al. [37] derived an approach using zero-shot transfer setting, given only a high-resource language (e.g. English) and exploit unsupervised bilingual lexicon induction (BLI) to generate augmented training data. The effectiveness is proved by their experiment in 11 zero-resource languages.

Perevalov et al. [21] experience with machine translation to extend supported languages of existing multilingual KGQA systems. They demonstrate that translated to English provides the best performance in most cases and machine translation quality only has small or moderate positive correlation with the question answering score.

3   Conclusion and Outlook

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla

et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: Aberer, K., Choi, K.S., Noy, N., Allemang, D., Lee, K.I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) The Semantic Web, pp. 722–735, Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
2. Bond, F., Foster, R.: Linking and extending an open multilingual Wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1352–1362, Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013), URL `https://aclanthology.org/P13-1133`
3. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, p. 1306–1313, AAAI'10, AAAI Press (2010)
4. Chung, H.W., Févry, T., Tsai, H., Johnson, M., Ruder, S.: Rethinking embedding coupling in pre-trained language models (2020), `https://doi.org/10.48550/ARXIV.2010.12821`, URL `https://arxiv.org/abs/2010.12821`
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. CoRR **abs/1911.02116** (2019), URL `http://arxiv.org/abs/1911.02116`
6. Devlin, J.: Multilingual bert (2018), URL `https://github.com/google-research/bert/blob/master/multilingual.md`.
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018), `https://doi.org/10.48550/ARXIV.1810.04805`, URL `https://arxiv.org/abs/1810.04805`

8. Diefenbach, D., Both, A., Singh, K., Maret, P.: Towards a question answering system over the semantic web (2018), `https://doi.org/10.48550/ARXIV.1803.00832`, URL `https://arxiv.org/abs/1803.00832`

9. Dolan, W.B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the Third International Workshop on Paraphrasing (IWP2005) (2005), URL `https://aclanthology.org/I05-5002`

10. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, MA (1998), ISBN 978-0-262-06197-1

11. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: Evolution of structured data on the web. Commun. ACM **59**(2), 44–51 (jan 2016), ISSN 0001-0782, `https://doi.org/10.1145/2844544`, URL `https://doi.org/10.1145/2844544`

12. Hakimov, S., Jebbara, S., Cimiano, P.: Amuse: Multilingual semantic parsing for question answering over linked data (2018), `https://doi.org/10.48550/ARXIV.1802.09296`, URL `https://arxiv.org/abs/1802.09296`

13. Lample, G., Conneau, A.: Cross-lingual language model pretraining. CoRR **abs/1901.07291** (2019), URL `http://arxiv.org/abs/1901.07291`

14. Lewis, M., Ghazvininejad, M., Ghosh, G., Aghajanyan, A., Wang, S.I., Zettlemoyer, L.: Pre-training via paraphrasing. CoRR **abs/2006.15020** (2020), URL `https://arxiv.org/abs/2006.15020`

15. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. CoRR **abs/1910.13461** (2019), URL `http://arxiv.org/abs/1910.13461`

16. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. CoRR **abs/2001.08210** (2020), URL `https://arxiv.org/abs/2001.08210`

17. Matuszek, C., Cabral, J., Witbrock, M., DeOliveira, J.: An introduction to the syntax and content of cyc. pp. 44–49 (01 2006)

18. Panchbhai, A., Soru, T., Marx, E.: Exploring sequence-to-sequence models for SPARQL pattern composition (2020), URL `https://arxiv.org/abs/2010.10900`

19. Pellissier Tanon, T., de Assunção, M.D., Caron, E., Suchanek, F.M.: Demoing platypus – a multilingual question answering platform for wikidata. In: Gangemi, A., Gentile, A.L., Nuzzolese, A.G., Rudolph, S., Maleshkova, M., Paulheim, H., Pan, J.Z., Alam, M. (eds.) The Semantic Web: ESWC 2018 Satellite Events, pp. 111–116, Springer International Publishing, Cham (2018), ISBN 978-3-319-98192-5

20. Pellissier Tanon, T., Weikum, G., Suchanek, F.: Yago 4: A reason-able knowledge base. In: Harth, A., Kirrane, S., Ngonga Ngomo, A.C., Paulheim, H., Rula, A., Gentile, A.L., Haase, P., Cochez, M. (eds.) The Semantic Web, pp. 583–596, Springer International Publishing, Cham (2020)

21. Perevalov, A., Both, A., Diefenbach, D., Ngonga Ngomo, A.C.: Can machine translation be a reasonable alternative for multilingual question answering systems over knowledge graphs? In: Proceedings of the ACM Web Conference 2022, p. 977–986, WWW '22, Association for Computing Machinery, New York, NY, USA (2022), ISBN 9781450390965, `https://doi.org/10.1145/3485447.3511940`, URL `https://doi.org/10.1145/3485447.3511940`

22. Perevalov, A., Diefenbach, D., Usbeck, R., Both, A.: Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers. In: 2022 IEEE 16th International Conference on Semantic Computing (ICSC), pp. 229–234 (2022), `https://doi.org/10.1109/ICSC52841.2022.00045`

23. Perevalov, A., Yan, X., Kovriguina, L., Jiang, L., Both, A., Usbeck, R.: Knowledge graph question answering leaderboard: A community resource to prevent a replication crisis. CoRR **abs/2201.08174** (2022), URL `https://arxiv.org/abs/2201.08174`

24. Pouran Ben Veyseh, A.: Cross-lingual question answering using common semantic space. In: Proceedings of TextGraphs-10: the Workshop on Graph-based Methods for Natural Language Processing, pp. 15–19, Association for Computational Linguistics, San Diego, CA, USA (Jun 2016), `https://doi.org/10.18653/v1/W16-1403`, URL `https://aclanthology.org/W16-1403`

25. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)

26. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR **abs/1910.10683** (2019), URL `http://arxiv.org/abs/1910.10683`

27. Rony, M.R.A.H., Kumar, U., Teucher, R., Kovriguina, L., Lehmann, J.: Sgpt: A generative approach for sparql query generation from natural language questions. IEEE Access **10**, 70712–70723 (2022), `https://doi.org/10.1109/ACCESS.2022.3188714`

28. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge (2017), URL `http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972`

29. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, p. 697–706, WWW '07, Association for Computing Machinery, New York, NY, USA (2007), ISBN 9781595936547, `https://doi.org/10.1145/1242572.1242667`, URL `https://doi.org/10.1145/1242572.1242667`

30. To, N.D., Reformat, M.: Question-answering system with linguistic terms over rdf knowledge graphs. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4236–4243 (2020), `https://doi.org/10.1109/SMC42975.2020.9282949`

31. Usbeck, R., Gusmita, R., Saleem, M., Ngonga Ngomo, A.C.: 9th challenge on question answering over linked data (qald-9) (11 2018)

32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017), URL `https://arxiv.org/pdf/1706.03762.pdf`

33. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (sep 2014), ISSN 0001-0782, `https://doi.org/10.1145/2629489`, URL `https://doi.org/10.1145/2629489`

34. Warstadt, A., Singh, A., Bowman, S.R.: Neural network acceptability judgments. CoRR **abs/1805.12471** (2018), URL `http://arxiv.org/abs/1805.12471`

35. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. CoRR **abs/2010.11934** (2020), URL `https://arxiv.org/abs/2010.11934`

36. Zheng, W., Zhang, M.: Question answering over knowledge graphs via structural query patterns (2019), `https://doi.org/10.48550/ARXIV.1910.09760`, URL `https://arxiv.org/abs/1910.09760`

37. Zhou, Y., Geng, X., Shen, T., Zhang, W., Jiang, D.: Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5822–5834, Association for Computational Linguistics, Online (Jun 2021), `https://doi.org/10.`

18653/v1/2021.naacl-main.465, URL https://aclanthology.org/2021.naacl-main.465

38. Zhu, Y., Kiros, R., Zemel, R.S., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. CoRR **abs/1506.06724** (2015), URL http://arxiv.org/abs/1506.06724