Towards Constructing Multi-hop Reasoning Chains using Local Cohesion

Sabine Ullrich[1] and Michaela Geierhos[1]

[1] Research Institute Cyber Defense; Universität der Bundeswehr, Munich, Germany
{sabine.ullrich,michaela.geierhos}@unibw.de

Abstract. For complex questions in question-answering systems, simple information extraction techniques are not sufficient to provide a satisfactory answer. To explain why an answer is correct and to provide a cohesive line of argumentation, reasoning chains can be useful. For more challenging questions, multi-hop reasoning chains help to connect consecutive sentences. Currently, however, multi-hop reasoning can only be detected when key entities overlap between two connected sentences. To address this problem, we will study the linguistic features of highly cohesive argumentation chains in order to apply them in more general models. Local cohesion is defined by the connectedness of sequences at the sentence level. While this measure is currently mainly used for automated essay scoring, we propose to use local cohesion to detect connections between sentences when none or not all crucial words overlap. Instead, cohesive lexical and structural features such as synonyms, paraphrases, and hypernyms should be considered. After analyzing multi-hop reasoning chains, new delexicalized chain representations are abstracted to construct generalized reasoning chains.

Keywords: Question Answering · Reasoning Chain · Cohesion

1 Introduction

Question Answering (QA) systems are widely used today, replacing simple information retrieval systems for user queries. Even virtual assistants such as Amazon's Alexa, Apple's Siri, Google Assistant, and Microsoft's Cortana are no longer limited to online shopping assistance, music playback, and news. However, most QA systems are still restricted to simple fact-based questions, such as when Alexa searches a catalogue using standard knowledge graph-based techniques [1]. Simple product-related questions, such as "Does Kindle support Japanese?", can be easily answered. When extracting interpretative questions that require logical thinking, current state-of-the-art systems are unable to solve these problems [2]. Imagine a complex question such as "How is the current situation in Syria?". Answering this question is not easy and cannot be done by a simple knowledge graph or ontology. Most QA systems trust that answers can be extracted online, so rarely more than a phrase or sentence is returned as an answer.

Complex questions require sophisticated answers. Thus, if a user asks "How is the current situation in Syria?", a simple answer like "critical" is not satisfactory. For complex questions, the derived answer should be comprehensive and reasoning steps should be given. A sample answer might be as follows:

> The situation in Syria is still *critical and unclear*. There is an ongoing civil war since 2011 between the Syrian Arab Republic led by Syrian president Bashar al-Assad and various domestic and foreign forces. Syria's economy is in its worst state since the start of the conflict. (...) More than half of the country's population was displaced. Even as the armed conflict winds down, it is unclear when or if they will be able to return.[1]

In order to identify highly connected sentences that form a valid reasoning chain (RC), connectives need to be examined. Currently, overlapping words and entities are considered as connectives between sentences [3]. However, when RCs are created from heterogeneous sources, the wording can vary significantly. For example, while one article refers to "civil war" and another source uses the term "armed conflict", both could refer to the confrontation between al-Assad and the domestic forces in Syria. Detecting these hidden overlaps between sentences can be achieved using local cohesion. Two phrases show high cohesion when overlapping words, synonymous expressions, or paraphrases are used. Our hypothesis is that local cohesion can be used to identify relevant RCs in QA.

2 Related Work

QA is widely used, especially when it comes to simple, factoid questions. There is less research in the field of multi-hop QA. Presented datasets support multi-hop question answering [3], explainable reasoning [4–6], natural language inference [7, 8], and multi-hop reading comprehension across documents [9]. Research related to QA tasks includes reading comprehension [10, 11], commonsense reasoning [12, 13], fact-checking [14], and compositional explanation [15]. Most tasks rely on overlapping words, but there are few approaches that address the problem of semantic relations between sentences when no entities overlap. Tsuchida et al. [16] present a natural language inference method using auto-discovered rules.

Currently, cohesion is measured in educational studies for automated essay scoring [17]. It is also used to calculate the connectivity of sentences comparing the explanations of experts and intermediates [18]. Lachner et al. [17] find that the higher the cohesion, the better an explanation is understandable.

[1] The phrases that contribute to this sample answer are taken from https://www.reuters.com/world/middle-east/cost-ten-years-devastating-war-syria-2021-05-26/ and from https://en.wikipedia.org/wiki/Syrian_civil_war.

3 Challenges

In our planned research, we aim to answer complex questions that require reasoning. To achieve this goal, three major challenges have to be tackled.

The first challenge is to define cohesive features for simple sentences. Currently, cohesion is used to measure the connectedness between phrases, especially in longer texts. However, shorter phrases and paragraphs have not yet been considered. We therefore plan to calculate cohesion for fewer and shorter phrases in RCs. While global cohesion is mainly used to measure cohesion in longer essays [19], local cohesion could capture the connectedness between shorter chain segments.

The second challenge is to create a natural language dataset for training and evaluation that comprises naturally occurring RCs. Datasets available for RC generation and extraction tend to include artificially connected chain links with overlapping entities at all times. However, this is not realistic when applied to real-world problems, especially when chains are extracted from different sources with different writing styles. While one author may use one term, another could use a synonymous expression, a paraphrase, or a technical term.

The third challenge is to extract delexicalized RCs from the created dataset. Currently, only generalized RCs with overlap exist. The generation of delexicalized multi-hop RCs is particularly challenging because

1. short phrases have to be extracted from different sources,
2. cohesive links between the phrases have to be determined,
3. a possible pool of delexicalized candidates has to be generated, and
4. without an available gold standard, the chains must be ranked and evaluated.

4 Approach

In the following, we present our approach to tackle the challenges above mentioned. Since answering complex questions cannot be done by extracting single, isolated phrases or sentences, multi-hop RCs are considered. Our goal is to generate general templates for retrieving valid RCs with non-overlapping chain links. These links should be determined by measuring the local cohesion between phrases. For this purpose, the pipeline shown in Figure 1 is implemented.

4.1 Measuring Local Cohesion in Reasoning Chains

The first idea is to transfer the knowledge from automated essay scoring to automated evaluation of RCs. While in essay scoring, cohesion has been helpful in the automated assessment of text quality, it could also be applied to determine high quality RCs that are connected. Connectedness in essays is measured using global cohesion, which focuses on causal relationships throughout the text, and semantic similarity between paragraphs in the text, examines the structure as a whole. For RCs that are smaller in scope, our idea is to measure connectedness

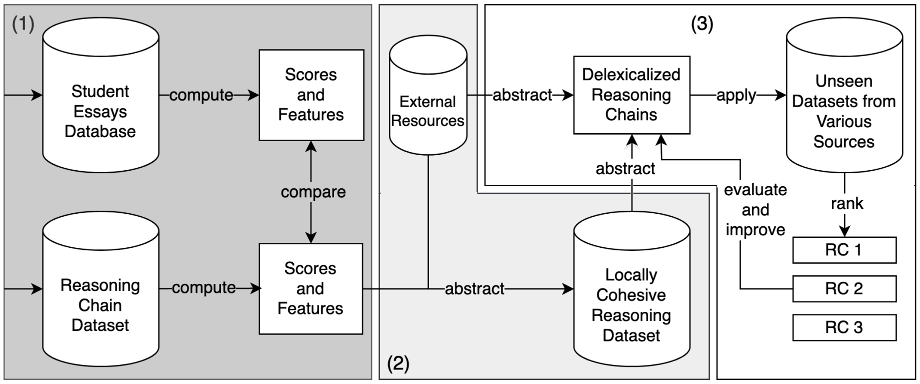

Fig. 1. The pipeline for complex RC extraction. Firstly, cohesion for RCs and essays is calculated and compared (1). Using the computed scores and relevant features, a cohesive dataset is created (2). From this dataset, delexicalized RCs are generated and applied to unseen data sources. The retrieved chains are ranked and scored by calculating their cohesion. If the results still need to be improved, the RCs are adjusted accordingly and new chains are extracted from the dataset. (3)

using locally cohesive features. Local cohesion refers to cohesion at the sentence level, taken into account connectives between sentences, semantic similarity between sentences, and noun overlap. Cohesive features include

- connectives
- type-token-ratio (TTR)
- lexical overlap
- synonym overlap

The scores for each feature can be measured using the Tool for the Automatic analysis of Cohesion (TAACO)[2] [19]. In our experiments, the cohesion of both student essays and reasoning datasets is calculated. Both scores are compared at both local and global levels in the next step. Next, features that mainly contribute to the cohesion score are extracted and stored.

4.2 Creating a Non-Overlapping Reasoning Dataset

The second challenge to be overcome is the creation of a cohesive dataset that can be used for generalized RC construction. Since existing datasets are based on word overlaps, they have to be adapted so that different cohesive features connect the chain links. An extensive list of logically connected reasoning datasets has been collected by Wiegreffe and Marasović [4]. Possible candidates for our dataset adaptation are *eQASC* [3] or *HotpotQA* [6], as both work with multi-hop RCs for QA.

[2] https://www.linguisticanalysistools.org/taaco.html

Replacing words with synonymous expressions can be done with the Python Natural Language Toolkit (NLTK) and WordNet [20]. WordNet is a lexical database for the English language in which content words are grouped into sets of cognitive synonyms, each expressing a particular concept.

There are several approaches for generating paraphrases. NLPAug[3], for example, uses the PPDB[4] paraphrase database, which contains over 220 million paraphrase pairs in English [21]. HuggingFace already lists 16 paraphrasers[5] that can be used to perturb the dataset. A new promising model presented on GitHub is PARROT[6], which generates paraphrases based on the T5 transformer model.

4.3 Generating Delexicalized Reasoning Chains

The new locally cohesive dataset serves as basis for the next step. The aim here is to generate delexicalized templates from the database. Delexicalization [22] involves replacing repeated noun phrases with variables. To find the noun phrases, we will perform part-of-speech tagging with NLTK and replace the candidates with a predefined set of special tokens. To create even more general RCs, delexicalization is not enough. Graph-based local grammars like Unitex[7] could help to capture repetitive structures in the RCs. If the size of the database allows it, deep learning approaches could also be considered. This could be done, for example, by using a pre-trained BERT model for encoding and a two-layer feed-forward neural network with ReLU to predict valid RCs [3].

After extracting valid RCs, the templates should be applied to unseen sources to check whether our approach can expand the candidate pool of RCs and answer complex questions. If the list still needs improvement after evaluation, the delexicalized reasoning chains should be re-evaluated and adjusted accordingly.

References

1. D. Carmel, L. Lewin-Eytan, and Y. Maarek, "Product question answering using customer generated content-research challenges," in *The 41st International ACM SIGIR*, pp. 1349–1350, 2018.
2. E. Dimitrakis, K. Sgontzos, and Y. Tzitzikas, "A survey on question answering systems over linked data and documents," *Journal of Intelligent Information Systems*, vol. 55, no. 2, pp. 233–259, 2020.
3. H. Jhamtani and P. Clark, "Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering," in *Proc. of the 2020 EMNLP*, pp. 137–150, 2020.
4. S. Wiegreffe and A. Marasović, "Teach me to explain: A review of datasets for explainable NLP," *arXiv preprint arXiv:2102.12060*, 2021.

[3] https://github.com/makcedward/nlpaug

[4] http://paraphrase.org/#/download

[5] https://huggingface.co/models?pipeline_tag=text2text-generation&search=paraphrase

[6] https://github.com/PrithivirajDamodaran/Parrot_Paraphraser

[7] https://unitexgramlab.org/

5. P. Jansen, E. Wainwright, S. Marmorstein, and C. Morrison, “Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference,” in *Proc. of the Eleventh International LREC 2018*, 2018.
6. Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, “Hotpotqa: A dataset for diverse, explainable multi-hop question answering,” in *Proc. of the 2018 EMNLP*, pp. 2369–2380, 2018.
7. S. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” in *Proc. of the 2015 Conf. on EMNLP*, pp. 632–642, 2015.
8. D. Demszky, K. Guu, and P. Liang, “Transforming question answering datasets into natural language inference datasets,” *arXiv preprint arXiv:1809.02922*, 2018.
9. J. Welbl, P. Stenetorp, and S. Riedel, “Constructing datasets for multi-hop reading comprehension across documents,” *Transactions of the ACL*, vol. 6, pp. 287–302, 2018.
10. C. Clark and M. Gardner, “Simple and effective multi-paragraph reading comprehension,” in *Proc. of the 56th Annual Meeting of the ACL (Volume 1: Long Papers)*, pp. 845–855, 2018.
11. Y. Jiang, N. Joshi, Y.-C. Chen, and M. Bansal, “Explore, propose, and assemble: An interpretable model for multi-hop reading comprehension,” in *Proceedings of the 57th Annual Meeting of the ACL*, pp. 2714–2725, 2019.
12. A. Talmor, J. Herzig, N. Lourie, and J. Berant, “Commonsenseqa: A question answering challenge targeting commonsense knowledge,” in *Proc. of the 2019 Conf. of the NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.
13. N. F. Rajani, B. McCann, C. Xiong, and R. Socher, “Explain yourself! leveraging language models for commonsense reasoning,” in *Proc. of the 57th Annual Meeting of the ACL*, pp. 4932–4942, 2019.
14. N. Kotonya and F. Toni, “Explainable automated fact-checking for public health claims,” in *Proc. of the 2020 EMNLP*, pp. 7740–7754, 2020.
15. Q. Ye, X. Huang, E. Boschee, and X. Ren, “Teaching machine comprehension with compositional explanations,” in *Proc. of the 2020 Conf. on EMNLP: Findings*, pp. 1599–1615, 2020.
16. M. Tsuchida, K. Torisawa, S. De Saeger, J.-H. Oh, C. Hashimoto, H. Ohwada, *et al.*, “Toward finding semantic relations not written in a single sentence: An inference method using auto-discovered rules,” in *Proc. of 5th IJCNLP*, pp. 902–910, 2011.
17. A. Lachner and M. Nuckles, “Experts’ explanations engage novices in deep-processing,” in *Proc. of the Annual Meeting of the CogSci*, vol. 35, 2013.
18. A. Lachner, J. Gurlitt, and M. Nuckles, “A graph-oriented approach to measuring expertise-detecting structural differences between experts and intermediates,” in *Proc. of the Annual Meeting of the CogSci*, vol. 34, 2012.
19. S. A. Crossley, K. Kyle, and D. S. McNamara, “The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion,” *Behavior research methods*, vol. 48, no. 4, pp. 1227–1237, 2016.
20. G. A. Miller, “WordNet: a lexical database for English,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
21. J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, “PPDB: The paraphrase database,” in *Proc. of the 2013 Conference of the NAACL: Human Language Technologies*, pp. 758–764, 2013.
22. S. Suntwal, M. Paul, R. Sharp, and M. Surdeanu, “On the Importance of Delexicalization for Fact Verification,” in *Proceedings of the 2019 EMNLP-IJCNLP*, pp. 3404–3409, 2019.