HAL open science

Calor-Dial : a corpus for Conversational Question Answering on French encyclopedic documents

Frédéric Béchet, Ludivine Robert, Lina Rojas-Barahona, Géraldine Damnati

To cite this version:

Frédéric Béchet, Ludivine Robert, Lina Rojas-Barahona, Géraldine Damnati. Calor-Dial : a corpus for Conversational Question Answering on French encyclopedic documents. CIRCLE (Joint Conference of the Information Retrieval Communities in Europe), Jul 2022, Samatan, France. hal-03714189

HAL Id: hal-03714189

https://hal.science/hal-03714189v1

Submitted on 5 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Calor-Dial : a corpus for Conversational Question Answering on French encyclopedic documents

Frédéric Béchet[1,*], Ludivine Robert[1], Lina Rojas-Barahona[2] and Géraldine Damnati[2]

[1]Aix-Marseille University - CNRS, Marseille, France
[2]Orange Innovation, DATAAI/AITT, Lannion, France

Abstract

Calor-Dial is an enriched version of the **Calor** corpus, collected from French encyclopedic data in order to study Information Extraction on domain specific data. The corpus was initially annotated in semantic Frames (**Calor-Frame**) and enriched with a first set of questions for Machine Reading Question Answering (**Calor-Quest**). The new **Calor-Dial** version presented here addresses the scope of conversational Question Answering. The main originality is that different types of questions are annotated, including more challenging configurations than in classical QA corpora. This paper describes the corpus and proposes some baseline results obtained with models trained on the **FQuAD** corpus.

Keywords

datasets, conversational question answering, multihop question answering

1. Introduction

Machine Reading Question Answering is an Information Retrieval task consisting in retrieving from a document a word span corresponding to the answer to a question on the document content. This task became very popular with the availability of large benchmark datasets such as SQuAD [1] containing 100K triplets *(document,question,answer)*. Current end-to-end models based on pretrained language models such as BERT obtain almost perfect results on SQuAD [1] as it contains single questions with answers consisting of only one word span in the document. Moreover most of the questions are rather literal with respect to the sentence containing the answer, making this task an easy task for powerful Information Retrieval model based on pretrained representation.

Two kinds of extension have been proposed to make this task more challenging: adding context with Conversational Question Answering [2] and having answers based on several word spans in Multi-Hop Question Answering [3].

Unlike single question answering tasks, Conversational Question Answering (CQA) involves a sequence of questions and answers. The answers can be found either in a paragraph [2, 4, 5] or in a knowledge-base [5]. In conversational question answering, the system faces the additional

CIRCLE'22: Joint Conference of the Information Retrieval Communities in Europe, July 04–07, 2022, Samatan, France
*Corresponding author.
frederic.bechet@univ-amu.fr (F. Béchet); ludivine.robert3@gmail.com (L. Robert); linamaria.rojasbarahona@orange.com (L. Rojas-Barahona); geraldine.damnati@orange.com (G. Damnati)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)
[1]https://rajpurkar.github.io/SQuAD-explorer

difficulty that questions may contain linguistic phenomena such as coreferences and ellipsis or implicit references to past turns. Existing corpora are available in English and usually their conversations refer to short and simple paragraphs such as excerpts of Wikipedia, children stories, web search or news. [2, 4, 5]. Moreover, answers correspond to single spans in the paragraph. Recently these datasets have been enriched with paraphrases of questions: question rewriting[6, 7]and paraphrases of answers [8]. Question rewriting refers to paraphrasing in-context questions with out-of-context questions.

In Multi-Hop Question Answering [3] the task consists in identifying several word spans in a document that has to be taken together in order to form the answer to a question. This is much more challenging than single QA as the simple similarity between a question and a sentence won't be sufficient to localize their answers.

Conversational and Multi-hop corpora are an opportunity to challenge current Machine Reading Question Answering (MRQA) models in order to check their ability to handle linguistic phenomenon such as coreference resolution, ellipsis or paraphrase.

In this context this paper will present the **Calor-Dial** corpus which is a Conversational Question Answering for French. This corpus contains encyclopedic documents with manually written questions where the answers can be contained in distinct spans of the document (i.e Multihop QA). In other words, answers can gather disjoint evidence sentences. Besides annotating the spans containing the answer, **Calor-Dial** also provides annotations on question rewriting and answer paraphrasing. **Calor-Dial** contains 234 dialogues and 1663 questions with their answers.

To the best of our knowledge this is the first conversational corpus on rich encyclopedic documents for multihop conversational QA, question rewriting and answer paraphrasing. The corpus is publicly available on the following archive: https://gitlab.lis-lab.fr/calor/calor-dial-public

2. The Calor corpus

Calor is a corpus collected for Information Extraction studies[2] and regularly enriched with annotations at various levels. It gathers French encyclopedic documents annotated with semantic information (**Calor-Frame**) following the Berkeley Framenet paradigm described in [9], questions on semantic roles for Machine Reading Question Answering (MRQA) (**Calor-Quest**) [10] and now a new set of questions for Conversation Question Answering) (**Calor-Dial**). The **Calor-Frame** corpus was initially built in order to alleviate Semantic Frame detection for the French language with two main purposes. The first one was to have a large amount of annotated examples for each Frame with all their possible Frame Elements, with the deliberate choice to annotate only the most frequent Frames. As a result, the corpus contains 53 different Frames but around 26k occurrences of them along with around 57k Frame Element occurrences. The second purpose was to study the impact of domain change and style change. To this end the corpus was built by gathering encyclopedic articles from two thematic domains (`WW1` for First World War and `Arch` for Archeology and Ancient History) and 3 sources (`WP` Wikipedia, `V` for

[2]https://gitlab.lis-lab.fr/alexis.nasr/calor-public, the annotations presented here will be added to the repository by the time the paper will be published if it is accepted.

the Vikidia encyclopedia for children and CT for the Cliotext collection of historical documents), resulting in the 4 subcorpora that will be further described in Table 1.

3. Calor-Dial Annotation Process

For building the **Calor-Dial** corpus annotators were asked to write a sequence of questions on a document, each question containing a reference to a previous question in the sequence. The main originality of **Calor-Dial** is the labels attached to each question. The annotators had to qualify every question they wrote according to 4 dimensions:

1. *in-context vs. out-of-context* → does the question need to access to the conversational context in order to be found?
2. *literal vs. paraphrase* → is the question very literal with respect to the sentence containing the answer or is it more abstract?
3. *self vs elliptical vs coreference* → is the question elliptical?, does it contains co-references? or is it self sufficient?
4. *simple vs multihop* → Is it necessary to access to distinct spans in the document to answer the question?

In addition, annotators were asked :

- to write an out-of-context version for each in-context question
- to write two versions of each answer, a *short* one containing the smallest word sequence containing the answer and a *long* version containing the context of the question.

4. Example

An example of a sequence of 6 questions from the *WP_arch* collection is presented below.

- *Q0* : **Quels sont les trois noms d'Hammourabi?**
  Q0 : *What are the three names of Hammourabi?*
  - *type:* `paraphrase-self-simple`
  - *short answer:* `Hammourabi, Hammurabi ou Hammurapi.`
  - *answer with context:* `Les trois noms d'Hammourabi sont : Hammourabi, Hammurabi ou Hammurapi.`
  - *word span supporting answer in document:* `Hammourabi , ou Hammurabi ou encore Hammurapi`
- *Q1* : **Qui est-il?**
  Q1 : *Who is he?*
  - *out-of-context question:* `Qui est Hammourabi?` *Who is Hammourabi?*
  - *type:* `litteral-coreference-multihop`
  - *short answer:* `Le vrai fondateur du premier empire de Babylone et créateur du code d'Hammurabi.`

  - *answer with context:* Hammourabi est le vrai fondateur du premier empire de Babylone et créateur du code d'Hammurabi.
  - *word span supporting answer in document:* le vrai fondateur du premier empire de Babylone [. . .] célèbre pour le code d' Hammurabi

- *Q2* : **Qu'est-ce que ce code?**
  Q2 : *What is this code?*
  - *out-of-context question:* Qu'est-ce que le code d'Hammurabi? *What is the Hammurabi code?*
  - *type:* litteral-coreference-simple
  - *short answer:* Un recueil de lois.
  - *answer with context:* Le code d'Hammurabi est un recueil de lois.
  - *word span supporting answer in document:*

- *Q3* : **Sur quel support a-t-il été écrit?**
  Q3 : *On which support was it written?*
  - *out-of-context question:* Sur quel support a été écrit le code d'Hammurabi? *On which support was the Hammurabi code written?*
  - *type:* paraphrase-coreference-simple
  - *short answer:* Sur une stèle.
  - *answer with context:* Le code d'Hammurabi a été écrit sur une stèle.
  - *word span supporting answer in document:* sur une stèle

- *Q4* : **Où a-t-elle été découverte ?**
  Q4 : *Where was it discovered?*
  - *out-of-context question:* Où a été découverte la stèle supportant le code d'Hammurabi? *Where was the stele supporting Hammurabi code discovered?*
  - *type:* paraphrase-ellipse-simple
  - *short answer:* À Suse.
  - *answer with context:* La stèle supportant le code d'Hammurabi a été retrouvée à Suse.
  - *word span supporting answer in document:* à Suse

- *Q5* : **Où est-elle exposée aujourd'hui ?**
  Q5 : *Where is it exhibited nowadays?*
  - *out-of-context question:* Où est aujourd'hui exposée la stèle supportant le code d'Hammurabi? *Where is the stele supporting the Hammurabi code exhibited nowadays?*
  - *type:* paraphrase-coreference-simple
  - *short answer:* Au musée du Louvre à Paris.
  - *answer with context:* La stèle supportant le code d'Hammurabi est aujourd'hui exposée au musée du Louvre à Paris.
  - *word span supporting answer in document:* au musée du Louvre à Paris

5. Statistical description of the annotated corpus

After the annotation process of the **Calor** corpus we obtained the following statistics: 234 conversations have been annotated for a total amount of 1663 questions. The questions are spread in the four subcorpora as described in Table 1.

Table 1
Annotated questions in the 4 **Calor** corpus collections

collection	domain	source	#docs	#questions
V_antiq	`Arch`	Vikidia	61	630
WP_arch	`Arch`	Wikipedia	96	497
CT_WW1	`WW1`	ClioText	16	103
WP_WW1	`WW1`	Wikipedia	123	488

Sequences of questions have variable length with an average of 7.3 questions per dialogue. The distribution is provided in Table 2.

The distribution of questions according to the different dimensions listed above can be found in table 3.

6. Baseline Reading Comprehension experiments

The **Calor-Dial** corpus can be used with different experimental settings. Traditional MRQA experiments can be run by using only out-of-context questions (including the first questions of each conversation, as well as the following questions in their full out-of -context reformulation). For this experimental setup, it is possible to analyse the results along with 4 levels of difficulty refering to both the formal similarity between the question and the paragraph (Question: `literal` *vs* `paraphrase`) and to the level of analysis that must be performed within the paragraph to retrieve the answer (Paragraph: `simple` *vs* `multihop`):

1. literal-simple (611 questions)
2. paraphrase-simple (369 questions)
3. literal-multihop (336 questions)
4. paraphrase-multihop (379 questions)

Obviously, conversational MRQA experiments can also be run by taking into account successive questions, with potential coreferences and ellipses. In this configuration 12 levels of difficulty can be defined to better analyse the results.

It can also be used for language generation tasks such as full answer generation (from *short answer* to *answer with context*) or question rephrasing (from *in-context question* to *out-of-context question*).

In this paper we provide baseline MRQA results for the first out-of-context experimental setup. To this purpose we fine-tune the transformer model *CamemBERT* (large version, 335M parameters) [3] on the *FQuAD* corpus [11].

[3] https://huggingface.co/camembert/camembert-large

Table 2
Dialogue length distribution

dialogue length	2	3	4	5	6	7	8	9	10	11	12	13	14	17
nb. dialogues	3	6	18	36	42	35	23	17	16	22	7	7	1	1

Table 3
Question distribution

Type of question	#questions
literal-self-simple	249
literal-self-multihop	75
literal-coreference-simple	275
literal-coreference-multihop	218
literal-ellipse-simple	87
literal-ellipse-multihop	43
paraphrase-self-simple	91
paraphrase-self-multihop	59
paraphrase-coreference-simple	222
paraphrase-coreference-multihop	280
paraphrase-ellipse-simple	56
paraphrase-ellipse-multihop	40

Table 4
MRQA results on out-of-context questions for a model trained on FQuad

Difficulty level	Question	Paragraph	# questions	EM	F1	Precision	Recall
1	literal	simple	611	33.94	70.12	81.71	61.41
2	paraphrase	simple	369	27.62	59.84	74.69	49.91
3	literal	multihop	336	10.54	47.68	62.94	38.38
4	paraphrase	multihop	379	5.71	36.43	49.65	28.77

The results obtained on the **Calor-Dial** corpus are given in table 4. We use the following metrics: exact-match and F-score between the word spans expected in the reference annotations and the prediction by the MRQA model. As we can see the results obtained are much lower that those that can be obtained on the FQuAD or the SQuAD test corpora. This can be explained by the fact that the specific topics in the **Calor-Dial** corpus are quite different from those in FQuAD. We can also verify that paraphrase and multihop are two complexity factors that affect greatly the performance of the MRQA model. Each level of difficulty has an impact of roughly 10 points of F-measure compared to the previous one. This advocates the need for more sophisticated model to be able to handle properly difficult phenomena such as paraphrases and multihop.

7. Conclusion

In its current form the **Calor-Dial** corpus can be used as an evaluation corpus for Machine Reading Question Answering models in order to check their ability to handle different linguistic

difficulties corresponding to the different dimensions characterizing each questions. It can also be used for evaluation text generation models such as Answer Generation models (from in-context to out-of-context answers), Question Rewriting models (paraphrasing in-context question into out-of-context questions) and Question generation. The corpus is publicly available on the following archive: https://gitlab.lis-lab.fr/calor/calor-dial-public

References

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 2383–2392. URL: http://aclweb.org/anthology/D16-1264. doi:10.18653/v1/D16-1264.

[2] S. Reddy, D. Chen, C. Manning, CoQA: A Conversational Question Answering Challenge, Transactions of the Association for Computational Linguistics 7 (2019) 249–266. URL: https://doi.org/10.1162/tacl_a_00266. doi:10.1162/tacl_a_00266.

[3] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, Hotpotqa: A dataset for diverse, explainable multi-hop question answering, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2369–2380.

[4] E. Choi, H. He, M. Iyyer, M. Yatskar, W. Yih, Y. Choi, P. Liang, L. Zettlemoyer, QuAC: Question Answering in Context, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2174–2184. URL: https://aclanthology.org/D18-1241. doi:10.18653/v1/D18-1241.

[5] A. Saha, V. Pahuja, M. Khapra, K. Sankaranarayanan, S. Chandar, Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[6] G. Kim, H. Kim, J. Park, J. Kang, Learn to Resolve Conversational Dependency: A Consistency Training Framework for Conversational Question Answering, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 6130–6141. URL: https://aclanthology.org/2021.acl-long.478. doi:10.18653/v1/2021.acl-long.478.

[7] Q. Brabant, G. Lecorve, L. M. Rojas-Barahona, Coqar: Question rewriting on coqa, in: 13th International Conference on Language Resources and Evaluation, 2022.

[8] A. Baheti, A. Ritter, K. Small, Fluent response generation for conversational question answering, arXiv preprint arXiv:2005.10464 (2020).

[9] G. Marzinotto, J. Auguste, F. Bechet, G. Damnati, A. Nasr, Semantic frame parsing for information extraction : the calor corpus, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), European Language Resource Association, 2018. URL: http://aclweb.org/anthology/L18-1159.

[10] F. Béchet, C. Aloui, D. Charlet, G. Damnati, J. Heinecke, A. Nasr, F. Herledan, Calor-quest: generating a training corpus for machine reading comprehension models from shallow

semantic annotations, in: MRQA: Machine Reading for Question Answering-Workshop at EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing, 2019.

[11] M. d'Hoffschmidt, W. Belblidia, Q. Heinrich, T. Brendlé, M. Vidal, FQuAD: French question answering dataset, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1193–1208. URL: https://aclanthology.org/2020.findings-emnlp.107. doi:10.18653/v1/2020.findings-emnlp.107.