**M1 Computer Science – UE Project**

**Logbook : behind the scenes of documentary research**

*The items you entered in this booklet will be scored*

**Full name and specialty :**

| LE Hai Nam DAC |
| --- |
| Vincent MARILLER DAC |
|  |
|  |

**Topic :**

| Conversational systems for Information retrieval / Campagne d’évaluation TREC |
| --- |

**Instruction :**

1. **Introduction (5-10 lines):** Briefly describe your research topic, its various aspects and issues, and the angle from which you have decided to address it.

1. **Selected keywords:** List the keywords you used for your bibliographic research. Organize the in the form of a mindmap.
2. **Description of the documentary research (10-15 lines):** Describe your use of the different research tools (search engines, database, catalogs, bounce search, etc.). Then compare the tools to each other. What sources did they allow you to access ? What are their specificities ? What is thei level of specialization?
3. **Produced bibliography within the framework of the project:** Use the ACM standard.
4. **Evaluation of sources (5 lines minimum per source)**: Choose 3 sources from your bibliography, describe how you found them, and critically evaluate them using the criteria seen on the TD materials.

You logbook must be sent at:

[**thomas.antignac@sorbonne-universite.fr**](mailto:thomas.antignac@sorbonne-universite.fr)

Reminder : the TD materials are available at the following address:

**http://www.pearltrees.com/formationbsu/master-info/id23514400**

1. **Introduction**

Conversational information retrieval is the task of information retrieval where users seek information via multi-turn conversations of natural language. The TREC Conversational Assistant Track (TREC CAsT) focuses on this task.

Conversational information retrieval involves two aspects: conversational query reformulation and ad hoc IR retrieval. The former rewrites user's utterance into a query best describing user's information need, based on historical context of the conversation. The latter, traditionally divided into first stage ranking and reranking, involves retrieving documents from a corpus in response to such information need.

We incorporate ideas from past researches to construct one such system in order to participate in TREC CAsT.

1. **Keywords**

I present a tree structure.

PLDAC

NLP basics

token embeddings

recurrent neural networks

Information retrieval

BM25

Conversational information retrieval

Conversational query reformulation

First stage retrieval

Document expansion

(Re-ranker) Models

Transformers

Bert

T5

Colbert

Tools

PyTerrier

TREC CAsT

Overview

Datasets

MARCO

KILT

Washington Post

CANARD

1. **Documentary research**

We used the Google search engine and Google Scholar search engine. As almost all of our sources are scientific papers referenced with standard references (ACM, IEEE) and are available for free on arxiv.org, these search engines proved to be sufficient.

We also relied on sources provided by our encadrant, namely theses on Information Retrieval and Natural Language Processing.

We use search engines to find an initial set of resources on principal aspects of the task and of TREC CAsT.

We then seek further information using two strategies:

* Examine past participants’ contributions within TREC CAsT. The overview of the track contains references to participants’ papers, which prove to be a valuable source of information.
* Investigate recent approaches in the research field using Google Scholar. This gives us access to up-to-date techniques with competitive performances. This method converges with the previous, as a research team behind this line of research participated and achieved the highest result in TREC CAsT 2020.

We then completed our document research with papers referenced by previously retrieved papers that are relevant to the task.

1. **Bibliography**

[1] Nawel Astaouti, Thomas Gerald, Maya Touzari, Jian-Yun Nie, and Laure Soulier. MLIA-LIP6@TREC-CAST2021: Feature augmentation for query recontextualization and passage ranking. 7.

[2] Nawel Astouati, Thomas Gerald, Maya Touzari, Jian-Yun Nie, and Laure Soulier. Feature augmentation for query recontextualization and passage ranking. 12.

[3] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs] (October 2018). Retrieved February 20, 2022 from <http://arxiv.org/abs/1611.09268>

[4] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview. (2020), 10.

[5] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. arXiv:2003.13624 [cs] (March 2020). Retrieved February 20, 2022 from <http://arxiv.org/abs/2003.13624>

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] (May 2019). Retrieved February 20, 2022 from <http://arxiv.org/abs/1810.04805>

[7] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. 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), Association for Computational Linguistics, Hong Kong, China, 5917–5923. DOI:<https://doi.org/10.18653/v1/D19-1605>

[8] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural Approaches to Conversational Information Retrieval. arXiv:2201.05176 [cs] (January 2022). Retrieved February 20, 2022 from <http://arxiv.org/abs/2201.05176>

[9] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. arXiv:2004.12832 [cs] (June 2020). Retrieved February 20, 2022 from <http://arxiv.org/abs/2004.12832>

[10] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling Dense Representations for Ranking using Tightly-Coupled Teachers. arXiv:2010.11386 [cs] (October 2020). Retrieved February 20, 2022 from <http://arxiv.org/abs/2010.11386>

[11] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), Association for Computational Linguistics, Online, 163–173. DOI:<https://doi.org/10.18653/v1/2021.repl4nlp-1.17>

[12] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized Query Embeddings for Conversational Search. arXiv:2104.08707 [cs] (November 2021). Retrieved February 20, 2022 from <http://arxiv.org/abs/2104.08707>

[13] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. TREC 2020 Notebook: CAsT Track. 6.

[14] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021. Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. arXiv:2005.02230 [cs] (March 2021). Retrieved February 20, 2022 from <http://arxiv.org/abs/2005.02230>

[15] Gia-Hung Nguyen. Modèles neuronaux pour la recherche d’information: approches dirigées par les ressources sémantiques. 212.

[16] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 708–718. DOI:<https://doi.org/10.18653/v1/2020.findings-emnlp.63>

[17] Rodrigo Nogueira and Jimmy Lin. From doc2query to docTTTTTquery. 3.

[18] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. arXiv:1904.08375 [cs] (September 2019). Retrieved February 20, 2022 from <http://arxiv.org/abs/1904.08375>

[19] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. arXiv:2009.02252 [cs] (May 2021). Retrieved February 20, 2022 from <http://arxiv.org/abs/2009.02252>

[20] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. arXiv:2101.05667 [cs] (January 2021). Retrieved February 20, 2022 from <http://arxiv.org/abs/2101.05667>

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs, stat] (July 2020). Retrieved February 20, 2022 from <http://arxiv.org/abs/1910.10683>

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs] (December 2017). Retrieved February 20, 2022 from <http://arxiv.org/abs/1706.03762>

[23] Laure Soulier. Définition et évaluation de modèles de recherche d’information collaborative basés sur les compétences de domaine et les rôles des utilisateurs. Retrieved February 20, 2022 from <http://thesesups.ups-tlse.fr/2735/1/2014TOU30190.pdf>

[24] TREC Conversational Assistance Track (CAsT). The TREC Conversational Assistance Track (CAsT). Retrieved February 20, 2022 from <https://www.treccast.ai/>

1. Evaluation of sources
2. Modèles neuronaux pour la recherche d’information: approches dirigées par les ressources sémantiques

Date: 18/12/2018

Relevance: This thesis provides the necessary basis for our task at hand, namely concepts of neural networks, text representations and their applications on Information Retrieval

Provenance: Gia-Hung Nguyen. Doctorat de l'Université de Toulouse. IRIT - University of Toulouse 3

Content rigor: We read through and understood the key concepts in sections relevant to our project.

Aim: To advance scientific research

1. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Date: 2020-07-28

Relevance: The model T5 defined in this paper is crucial to the neural reranking task, the document expansion task, and optionally, certain implementation choices of the query reformulation task

Provenance: Raffel Colin et al. Google, Mountain View, CA 94043, USA. They are researchers at Google.

Content rigor: The paper is of quality.

Aim: To advance scientific research

1. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models

Date: 2021-01-14

Relevance: This paper defines the expando-mono-duo design pattern, which is to be applied to our task (specifically, the ad hoc IR subtask).

Provenance: Ronak Pradeep et al. David R. Cheriton School of Computer Science, University of Waterloo. The research team at David R. Cheriton School of Computer Science achieved best result in TREC CAsT 2020.

Content rigor: The paper is of quality.

Aim: To advance scientific research