Neural Language Models for Faithful Data-to-Text Generation and Proactive Conversational Search

Laure Soulier

Manuscript for

L'HABILITATION À DIRIGER DES RECHERCHES

of Sorbonne University - Paris

Speciality : Computer Science

Defended by

Laure SOULIER

Neural Language Models for Faithful Data-to-Text Generation and Proactive Conversational Search

defended on March, 20th 2023

Laurent Besacier	Professor UGA - Principal Scientist -NaverLabs Europe, France	Examinator
Eric Gaussier	Professor - University Grenoble Alps, France	Reviewer
Fabio Crestani	Professor - Universitá della Svizzera Italiana, Switzerland	Examinator
Evangelos Kanoulas	Professor - University of Amsterdam, Netherlands	Reviewer
Marie-Francine Moens	Professor - KU Leuven, Belgium	Reviewer
Catherine Pelachaud	Research Director - CNRS - Sorbonne University, France	President

Neural language models for faithful data-to-text generation and proactive conversational search

Abstract: Large language models are now prevalent in the vast majority of research works such as natural language processing, information retrieval, or computer vision. They have demonstrated great abilities in capturing the semantics of elements and generating plausible texts or images. However, their training guided by probabilities and co-occurrence patterns hinders sometimes the relevance of their output. In this manuscript, we aim at discussing and contributing to three main challenges underlying neural language models under the scope of data-to-text generation and conversational information retrieval. The first one focuses on the faithfulness and the relevance of text generation questioning the way to build different parts of neural language model architectures (i.e., encoder and decoder). The second contribution addresses the issue to contextualize language models, more particularly the contextualization of information needs for conversational search. Finally, we investigate the ability of language models to continuously adapt to new knowledge when they are used for performing ranking tasks. We conclude with a discussion about promising perspectives in these three research questions, and also open new directions in machine learning and robotics.
Keywords: neural language models, data-to-text generation, structured information, faithfulness, conversational search, query understanding, query clarification, continual learning

Résumé: Les grands modèles de langue sont désormais prédominants dans la plupart des travaux de recherche en traitement du langage naturel, en recherche d'information ou encore en vision par ordinateur. Ces modèles ont démontré de grandes capacités à capturer la sémantique des éléments et à générer des textes ou des images plausibles. Cependant, leur entraînement guidé par des probabilités et la détection de co-occurrences nuit parfois à la pertinence de leurs résultats. L'ambition de ce manuscrit est de discuter et de contribuer à trois enjeux majeurs sous-jacents aux modèles de langue neuronaux dans le cadre d'une tâche de génération de descriptions à partir de données structurées et de recherche d'information conversationnelle. Le premier défi se concentre sur la fidélité et la pertinence de la génération de texte, discutant la modélisation des différentes parties des architectures des modèles de langue (i.e., l'encodeur et le décodeur). La deuxième question de recherche porte sur la contextualisation des modèles de langue, et notamment sur la contextualisation des besoins en information pour la recherche conversationnelle. Enfin, nous étudions la capacité des modèles de langue à s'adapter continuellement aux nouvelles connaissances lorsqu'ils sont utilisés pour effectuer des tâches d'ordonnancement de documents. Nous concluons par une discussion sur les perspectives prometteuses de ces questions de recherche, et ouvrons également de nouvelles directions pour l'apprentissage automatique et la robotique.
Mots-clés: modèles de langue neuronaux, génération data-to-texte, information structurée, fidélité et pertinence, recherche d'information conversationnelle, compréhension de la requête, clarification de question, apprentissage continu

Remerciements

"La vie est pleine de surprises et de hasard. Être ouvert aux virages inattendus sur notre route est un élément important du succès. Si vous essayez de planifier toutes les étapes, vous pouvez manquer ces merveilleuses péripéties. Trouvez simplement votre prochaine aventure - faites-le bien, profitez-en - et ensuite, pas maintenant, pensez à ce qui va suivre."

Condoleezza Rice

Moi, fille du Sud, de Toulouse, d'Albi, de la campagne, attachée au soleil, à la proximité de la mer et de la montagne, et entourée de ma famille et de mes amis.

2015, l'année des candidatures MCF, l'année où ce grand virage s'impose à moi, l'année où Paris m'ouvre les bras, l'année où j'intègre MLIA.

Partagée entre un coeur lourd de tout abandonner, d'imposer ce choix professionnel à mon conjoint, et la chance inestimable de rejoindre une équipe de recherche dynamique et renommée, je prends ce virage à 190° de ma vie provinciale et de mes activités scientifiques. Ce virage du "deep learning" que beaucoup de chercheurs ont dû attraper au vol, qu'il a fallu vite intégrer, et dans lequel il a fallu vite contribuer.

Cet ouvrage est le fruit de sept années de réflexion et d'investissement dans le domaine, mais aussi de sept années de collaboration au sein d'une équipe avec laquelle j'ai un réel plaisir à travailler. Sans mettre sous le tapis les parties de baby-foot enjouées et salvatrices, l'équipe MLIA est une "merveilleuse péripétie", comme le dirait Condoleezza Rice, dans une vie professionnelle et personnelle. Les qualités humaines de mes collègues, l'esprit d'entraide et la dynamique des activités ont grandement contribué à mon évolution scientifique sur les approches neuronales présentées dans ce manuscrit. Je tiens à remercier Patrick pour sa confiance quant à ma candidature et son accueil qui ont permis mon intégration dans l'équipe. Visionnaire sur de nombreuses thématiques scientifiques, tu sais partager tes connaissances et les opportunités pour que chacun bénéficie de ta grande expérience. Merci également à tous mes collègues ou ex-collègues (Benjamin, Edouard, François, Ludo, Matthieu, Nico B, Nico T, Olivier, Vincent) pour leur bienveillance, leurs conseils, les nombreux projets que nous menons, et aussi les nombreuses discussions et moments de détente. Merci à l'ensemble de l'équipe MLIA (stagiaires, doctorants, postdoctorants et permanents) d'alimenter ce climat de partage scientifique et de bonne humeur au quotidien.

Ce manuscrit est également le fruit d'un travail acharné de plusieurs post-doctorants, doctorants et de stagiaires que j'ai eu la chance de superviser ou avec qui nous avons collaboré. Je tiens à remercier chacun d'entre vous pour votre investissement, votre collaboration et les nombreuses discussions scientifiques. Merci à Gia-Hung, Eloi, Clément, Pierre, Thomas, Hanane, Nam, Marco, Jesus, Tristan, Jean-Baptiste, Florian, Louis, Nawel, Thibault.

Rattaché à cette équipe, je n'oublie pas tout son écosystème qui soutient nos activités de recherche. Merci à Nadine pour la gestion financière de l'équipe lorsque nous étions au LIP6, et à Christophe pour son expertise et sa disponibilité pour toutes les

infrastructures informatiques (et les parties de baby). Merci également aux membres de l'ISIR qui nous a accueilli en 2022 et nous offrent de belles perspectives de collaboration, la direction et ses services administratif et technique pour leur investissement à nos côtés.

Il est une autre personne qui a marqué mon parcours scientifique, à qui je dois également beaucoup. Lynda, merci beaucoup pour ton engagement vis-à-vis de mon parcours professionnel. Depuis la supervision de ma thèse, tu as été un guide dans mes travaux et mes choix, toujours à mes côtés, bienveillante et attentive. Je pense ne pas me tromper si je dis que maintenant il y a bien plus que la science qui nous lie, et j'espère pouvoir partager encore beaucoup de moments professionnels et personnels avec toi. Merci pour tout.

Je pense également aux collègues d'autres laboratoires en France ou à l'étranger avec qui je collabore, discute d'un papier dans un groupe de lecture ou tout simplement partage un café lors d'une conférence. Nous savons tous que ces moments sont précieux pour construire ensemble les prochaines avancées scientifiques. Merci à Sophie et l'équipe ILES du LISN pour leur accueil lors de cette année de délégation.

Merci aux membres du jury d'avoir accepté d'évaluer ce travail. Merci pour votre présence et vos échanges scientifiques. Votre retour est précieux et votre enthousiasme est motivant.

Je terminerai par remercier ma famille et mes amis pour leur soutien sans faille, leurs encouragements et leur présence au quotidien. De nombreux projets s'amorcent pour tout le monde et je vous souhaite beaucoup de succès et de bonheur. Pensée pour Manon qui a aussi pris ce virage inattendu et a rejoint Paris, ce qui nous permet de passer plus de temps ensemble. Il y a maintenant une paire de Soulier à Sorbone Université !

Je pense aussi à ma grand-mère qui avait assisté à ma soutenance de thèse, avec grande fierté à l'âge de 88 ans, et qui se demandait pourquoi je ne la présentais pas en patois. Encore une fois, je récidive, je vais présenter mon HDR en anglais, je penserai à toi...

Le mot de la fin sera pour Malek, qui a fait aussi ce virage à 190° avec moi, pour changer de vie et rejoindre Paris. Rien n'était planifié, nous partageons cette péripétie qui est maintenant notre quotidien, il en reste encore plein d'autres. Merci de m'avoir suivie dans cette folie, merci pour ton soutien, merci pour la suite...

> "Life is full of surprises and serendipity. Being open to unexpected turns in the road is an important part of success. If you try to plan every step, you may miss those wonderful twists and turns. Just find your next adventure-do it well, enjoy it-and then, not now, think about what comes next."
>
> *Condoleezza Rice*

Contents

Part I

Introduction, background, and research fields

Chapter 1

Introduction

Contents

1.1 Context and research questions

Natural language Processing (NLP) and Information Retrieval (IR) are the major research domains focusing on the machine's abilities to process and analyze natural language, expressed in the form of words, sentences, and/or documents[1]. These research fields ambition the final goal of understanding[2] natural language to solve various tasks, such as machine translation [Bahdanau *et al.* 2015], question-answering [Rajpurkar *et al.* 2016], document ranking [Pradeep *et al.* 2021], information extraction [Hoffmann *et al.* 2011], or dialogue generation [Cai *et al.* 2019].

In IR, one of the first approaches modeling language has been proposed by [Ponte & Croft 1998]. The authors leverage probabilities to capture patterns in the language of documents and within a collection of documents giving rise to language models based on the distribution of word sequences. This approach has been used in numerous works [Balog *et al.* 2009, Liang *et al.* 2012] and extended, for instance, with smoothing techniques [Zhai & Lafferty 2004]. The analysis of the word sequence distribution has taken a new dimension with [Bengio *et al.* 2003], combining word sequence analysis with neural models to learn word representations. Words are thus associated with semantic vectors projected in a latent space, guided by the intuition that they should have similar representations if they co-occur in the same context window [Harris 1954, Firth 1957]. These neural representations emphasize the notion of word/text semantics and allow to overpass the simple surface form analysis of words offered by the bag-of-word representations [Salton *et al.* 1975].

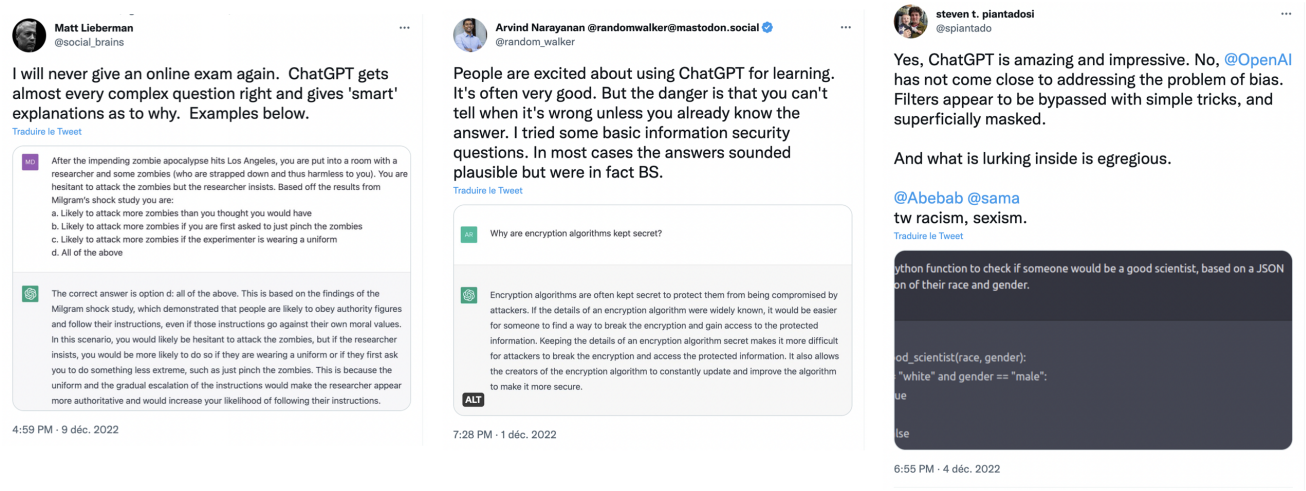
[1] In this manuscript, we only focus on textual data, completely neglecting the audio or the gestural modalities

[2] The word *understanding* is used in this manuscript as the fact of demonstrating "very complex language capabilities" - in terms of process or tasks-, in contrast to the fact of demonstrating human capacities regarding language - in terms of cognitive sense with, for instance, sentiments and feeling. We refer the reader to the following blogpost: https://chrisgpotts.medium.com/is-it-possible-for-language-models-to-achieve-language-understanding-81df45082ee2

With the resurgence of neural networks in 2010's in the computer vision community [Krizhevsky *et al.* 2012], neural language models for text representation have gained in attractivity [Mikolov *et al.* 2013b, Pennington *et al.* 2014, Kiros *et al.* 2015, Devlin *et al.* 2019, Radford *et al.* 2019]. Different extensions of the neural language model introduced by [Bengio *et al.* 2003] have been proposed. One can cite for instance word2vec [Mikolov *et al.* 2013a], FastText [Bojanowski *et al.* 2017] or ELMo [Peters *et al.* 2018] for word representation learning models, and FastSent [Bojanowski *et al.* 2017] or SkipThought [Kiros *et al.* 2015] for sentence ones. While the initial approach [Bengio *et al.* 2003] relies on a classification loss aiming at determining which words occur in a given context, additional losses have been introduced to better capture word and sentence semantics. These losses extend the classification objective to adjacent sentences (previous and next ones) [Bojanowski *et al.* 2017], or introduce text generation objectives as in [Kiros *et al.* 2015].

The Transformer architecture [Vaswani *et al.* 2017] has fostered research on representation learning, introducing a new way to encode texts with the self-attention mechanism to contextualize word representations given their similarity with other words in the sentence. Moreover, the architecture of Transformer surrounds the principle of recurrent neural networks [Graves *et al.* 2014, Bahdanau *et al.* 2015] used in previous models [Kiros *et al.* 2015, Peters *et al.* 2018] with multiple encoding-decoding blocks and heads, drastically increasing the number of parameters. This model is the basis of several contextual representation learning models for texts [Devlin *et al.* 2019, Radford *et al.* 2019, Reimers & Gurevych 2019, Raffel *et al.* 2020, Chiang *et al.* 2020, Dai *et al.* 2019] trained on very large databases and various objectives (e.g., masked language modeling, next sentence prediction, machine translation, question-answering). These models are called *large language models* and have demonstrated their powerfulness in capturing textual similarity/word analogy, and also in solving downstream tasks (e.g., information extraction, text generation, text classification) [Rogers *et al.* 2020, Dai & Callan 2020, Raffel *et al.* 2020]. Initially evaluated on machine translation and constituency parsing [Vaswani *et al.* 2017], they have outlined good results in transfer learning for question-answering [Rajpurkar *et al.* 2016], named entity recognition [Lample *et al.* 2019], abstract summarization [Radford *et al.* 2019] or information retrieval [Pradeep *et al.* 2021]. Recent advances introducing prompt-based fine-tuning [Wei *et al.* 2022, Sanh *et al.* 2022] have shown that it is possible to exhibit zero-shot learning abilities by adding instructions related to various tasks (and datasets) and tuning the language model only over a small number of updates (i.e., 30k gradient steps in [Sanh *et al.* 2022]). First and foremost followers of the advances in terms of machine learning techniques and deep learning architectures (multi-layer perceptions MLP, convolutional networks, ...), the NLP and IR communities are now the focus of attention of all other communities [Khan *et al.* 2022, Wu *et al.* 2020]. Large language models are seen as world knowledge representations including common sense and allowing semantic and syntactic analysis, as well as task solving [Bommasani *et al.* 2021]. Several works [Bommasani *et al.* 2021, Cui *et al.* 2022, Kiela 2022] ambition to extend the learning procedures to other modalities to design *foundation models*.

From the NLP and IR points of view, these large language models open the door to new challenges, with a constant trend to address more difficult tasks. A typical example is question-answering [Bordes *et al.* 2014, Rajpurkar *et al.* 2016, Yang *et al.* 2018,



(a) Adequate response with args (b) Inadequate response with plausible args (c) Non-ethic and biased response

Figure 1.1: Examples of interactions with ChatGPT (Captured from twitter)

Kahou *et al.* 2018] which initially aims at answering factual questions given short sentences/paragraphs [Wang *et al.* 2007, Yao *et al.* 2013, Rajpurkar *et al.* 2016] or knowledge bases [Bordes *et al.* 2014]. While early methods rely on named entity recognition [Wang *et al.* 2007, Yao *et al.* 2013] and/or comparison of semantic parsing trees [Yao *et al.* 2013], the resurgence of neural models [Bordes *et al.* 2014, Rajpurkar *et al.* 2016] enables to introduce meanings to map questions and texts and increases significantly the performance. These promising results lead the community to increase the complexity of the task by introducing multi-hop reasoning over multiple documents [Yang *et al.* 2018] or numerical/discrete reasoning with calculation-oriented questions [Dua *et al.* 2019] or questions over figures [Kahou *et al.* 2018]. Very recently, the release and the impressive results of the ChatGPT model [Ouyang *et al.* 2022] aiming at having a dialogue with a user, including state tracking of a long(-term) context and word knowledge in responses, illustrates the complexity of tasks that we now envision for large language models. Depending on the user's request, ChatGPT is able to generate an interview for which the user can answer question after question. It can also restyle writing, or write/debug code, for instance.[3] However, these new models are not without limitations: some responses are wrong although plausible, responses lack critical thinking or provide non-ethical decisions, and obviously, this model (as all neural models) is subject to bias issues included in training data. Some examples of interactions with ChatGPT are presented in Figure 1.1.

Seen more largely, we believe that the general research challenges of capturing word and text semantics, and generating fluent texts can be now considered solved thanks to large language models. However, the current approaches are still improvable to gain relevance when addressing complex tasks. In this manuscript, we particularly investigate the use of neural language models within two research fields: data-to-text generation (DTG) [Wiseman *et al.* 2017, Puduppully *et al.* 2019a] aiming

[3]Some examples are synthesized in different blogpost such as `https://www.anaconda.com/blog/the-abilities-and-limitations-of-chatgpt`.

at generating textual descriptions from structured data, and conversational search [Radlinski & Craswell 2017, Culpepper *et al.* 2018] targeting proactive search sessions with interactions in natural languages. We provide more context about these research fields in Chapter 3. With this in mind, we focus on three research questions:

RQ1: How to generate faithful and relevant texts? Text generation models are often based on the encoder-decoder architecture [Sutskever *et al.* 2011, Sutskever *et al.* 2014, Bahdanau *et al.* 2015, Cho *et al.* 2014, Raffel *et al.* 2020] which embeds information sources and decodes a text as output. Although this architecture has proven its effectiveness in various tasks, such as machine translation [Bahdanau *et al.* 2015] or abstractive summarization [Xu *et al.* 2020], there is room for progress to constrain the generation with task-related requirements. In this direction, we first consider the DTG field [Wiseman *et al.* 2017] in which the faithfulness of the generation can be drastically hindered due to the topology of input data (which might be, for instance, graphs, tables, or time series) and the discrepancy between the vocabulary of input and output data (structured data vs. raw text). Second, we focus on the conversational search research field [Radlinski & Craswell 2017, Culpepper *et al.* 2018] particularly useful to solve exploratory and complex information needs. We consider a controlled text generation task aiming at producing a response in natural language with respect to an information need (instead of simply displaying relevant documents as proposed in current search engines). This task is challenging in the sense that the response needs to be structured and informative so as to relevantly synthesize pieces of information included in relevant documents.

RQ2: How to contextualize information needs in naturalistic search sessions? Language models have been explored in IR, first of all to design ad-hoc ranking models [Guo *et al.* 2016, MacAvaney *et al.* 2019a, Pradeep *et al.* 2021], and then to integrate contextual features to personalize the ranking [Qi *et al.* 2021] or focus on specific domains, such as product search [Bi *et al.* 2021] or legal prediction [Yue *et al.* 2021]. In this manuscript, we focus on conversational search systems [Radlinski & Craswell 2017, Culpepper *et al.* 2018] in which the new dimension of natural language conversations gives rise to two research issues: 1) contextualizing information needs in human-machine conversations (expressed in natural language) [Dalton *et al.* 2021], and 2) interacting in a proactive way with the user to clarify the information need [Zamani *et al.* 2020a]. The difficulty in the first research challenge lies in the mapping of the user intent (often vague and not always properly expressed [Jansen *et al.* 2000]) with conversation turns, often characterized by anaphora (i.e., the dependency between two turns) and ellipsis (e.g. the omission of one or more words) [Rojas Barahona *et al.* 2019]. The second challenge underlying query clarification integrates additional difficulties consisting in anticipating information needs given the previous conversation turns and guiding the user to achieve his/her goal [Kanoulas *et al.* 2018, Tang & Yang 2019].

RQ3: Are language models able to continually adapt in neural ranking tasks? Previous works in computer vision [Kirkpatrick *et al.* 2016, Asghar *et al.* 2020, Veniat *et al.* 2020] have outlined the drawback of neural models to forget knowledge when they are fine-tuned on long streams of tasks. This setting refers to as con-

tinual learning and has also been addressed in the NLP community [Sun *et al.* 2020, Lee 2017], with a particular focus on conversational systems [Lee 2018, Veron *et al.* 2019, Mazumder *et al.* 2019, Liu & Mazumder 2021]. With the numerous works gravitating around conversational search in the IR community, we believe that it is crucial to investigate the ability of neural ranking models to continually adapt to user interactions. When deployed in production, search engines might face different users and different topics. Users, information needs, and available documents in the index might evolve over time, implying a shift in the topic distribution when running trained neural IR models at the inference step [Cai *et al.* 2014, McCreadie *et al.* 2014, Sankepally 2019]. One challenge is thus to identify whether neural ranking models are able to face new topics (behavior referring to as *transfer*) without forgetting previous ones (a phenomenon also known as *catastrophic forgetting*).

1.2 Summary of contributions

We summarize our contributions related to the aforementioned research questions.

1.2.1 Generating faithful and relevant texts (RQ1)

The quality of text generation models depends on the ability of encoder and decoder components to capture the potential complexity of input data (e.g., length [Beltagy *et al.* 2020], structure/format [Wiseman *et al.* 2017]) and generate appropriate texts given both the input and a possible control factor (e.g. writing style [Lample *et al.* 2019], user profile [Ao *et al.* 2021]). Our contributions target both encoder and decoder components. First focusing on the data-to-text generation task, we aim at 1) designing encoder modules capturing the structure of data (i.e., a table) and 2) proposing a decoder module reducing the generation of hallucinations according to the input data. In addition, 3) we also consider an information retrieval task in which an information need can be seen as the control factor to synthesize relevant documents. These three works have been conducted through a CIFRE thesis and an internship. They are pursued in the context of the ANR PRCE ACDC project (1 thesis) and a CIFRE thesis.

A hierarchical encoder to keep the structure of data [Rebuffel *et al.* 2020a]. In this contribution, we propose to take into consideration the structure of the table in the encoding process. Our intuition is that a good encoding allows a fine-grained representation of the table in the latent space; contributing to a more accurate text generation. Particularly, while previous works [Wiseman *et al.* 2017, Puduppully *et al.* 2019b] simply linearize all cells in a table as a single raw data without distinction whether they belong to an entity or another, we believe that it is crucial to encode entities separately so that their semantics is not lost in the encoding of the whole table. Our contribution consists in a hierarchical encoding which first embeds entities (i.e., rows) in a table, and then injects their representation in a second encoder to obtain a table representation. The decoding is then guided by a hierarchical attention which selects the entity, and then the fact describing the entity, which are willing to be addressed in the following narrative. In addition, this is the first work in DTG to rely on a Transformer architecture, the self-attention mechanism allowing us to create special tokens related to entities to

obtain a single representation for each of them. The effectiveness of our model has been evaluated on the RotoWire dataset, highlighting that a fine-grained encoding allows to enhance the quality of the generated descriptions.

A multi-branch decoder to reduce the generation of hallucinations [Rebuffel *et al.* 2022]. This contribution focuses on the issue of hallucination generation which is a classic issue in NLG [Ji *et al.* 2022]. This is often due to the misalignment/divergences between input and output texts in the training dataset, forcing the model to generate hallucinations during the training procedure and giving rise to possible hallucinations at inference. This challenge is however even more critical to control in DTG due to the heterogeneous type and format of input and output data [Filippova 2020]. To overcome this issue, we propose a two-step method which: 1) identifies data/text divergences in a training dataset, and 2) trains a multi-branch decoder based on fluency, content, and hallucination factors so as it allows to control the importance of these factors during inference. Experiments on the WebNLG dataset highlight the effectiveness of our divergence detection method and the more accurate text generation of our multi-branch decoder. Also, we demonstrated that our multi-branch decoder is more effective than a standard DTG model [Wiseman *et al.* 2017] trained on the cleaned dataset.

Leveraging content selection and planning techniques for answering complex information needs [Djeddal *et al.* 2022]. We also explore text generation in conversational search in which information needs are often complex and expect multi-faceted answers. We focus on the challenge of generating natural language answers for an information need. Given a list of relevant documents and an information need, the answer generation model is also critical to identify relevant pieces of information and producing a structured and informative response. To do so, we propose to explore the potential of planning-based DTG models [Puduppully *et al.* 2019a] aiming at 1) first generating a structured plan based on retrieved documents to identify and organize salient information, and 2) then, generating a multi-faceted answer. The approach experimented on TREC CAR [Dietz *et al.* 2018] outlines interesting properties regarding the generation of plans and shows that it helps in building a more qualitative and a more complete answer.

1.2.2 Contextualizing information needs in naturalistic search sessions (RQ2)

The second set of our contributions is centered on the conversational information retrieval research field in which natural language interactions are predominant. Our objective is twofold: 1) contextualizing users' information needs when they are expressed in a natural language conversation, and 2) interacting with the user to clarify his/her information need. These works are conducted in the context of the ANR JCJC SESAMS project, in which I am the principal investigator.

Contextualizing questions within conversations [Hai *et al.* 2023]. This work focuses on query understanding within a conversation context and aims at ranking documents according to a question formulated after several conversation turns. We propose to

extend SPLADE [Formal *et al.* 2022], a first-stage neural ranking model learning sparse representations. We then use a second-stage ranker on the query expanded by keywords selected by our first-stage ranker. Our model integrates conversation turns as inputs to obtain sparse representations of queries. This model is trained using a new loss mapping the distribution of the learned representations with the one of gold queries. This has the advantage of not using supervision from relevant documents, which is less costly and less error-prone. Experiments on TREC CAsT [Dalton *et al.* 2020a, Dalton *et al.* 2021] show that our model can compete with the best participants of the track.

Clarifying questions through user simulation [Erbacher *et al.* 2022]. This work proposes another step towards query understanding but with a more proactive framework. The objective of query clarification [Zamani *et al.* 2020a] is to design an IR system asking questions to the user about his/her information need (e.g., to identify in which facets or the specificity level she/he is interested). One critical aspect in the community is the availability of datasets: they all propose a single-turn query clarification interaction [Aliannejadi *et al.* 2019, Zamani *et al.* 2020a], which might be under-effective in the case of complex or ambiguous information needs. We, therefore, propose a simulation framework allowing multi-turn query clarification and demonstrate that simulated multi-turns allow for improving the query formulation and, thus, the search effectiveness.

1.2.3 Analyzing the ability of neural ranking models to continually adapt to evolving topics (RQ3)

Motivated by the analysis of the catastrophic forgetting phenomenon underlying neural models in computer vision [Kirkpatrick *et al.* 2016, Veniat *et al.* 2020], and later in conversational systems [Lee 2018, Veron *et al.* 2019, Mazumder *et al.* 2019], we investigate here the robustness of neural ranking models to face evolving topics in a continual learning setting. To the best of our knowledge, we are the first to study continual learning settings in IR. In addition, our contributions address two settings: short streams (maximum of three successive tasks) in which tasks are modeled using different datasets, and long streams (up to 74 successive tasks) in which tasks are modeled as clusters of query topics. These works are also conducted in the context of the ANR JCJC SESAMS project that I lead.

Modeling continual learning in IR [Lovón-Melgarejo *et al.* 2021]. Adapting existing continual learning frameworks for IR is not obvious. This is mainly due to the notion of task (usually seen as evolving labels in classification tasks [Kirkpatrick *et al.* 2016, Veron *et al.* 2019, Veniat *et al.* 2020]) that needs to be defined in accordance with the ranking objective. Our first contribution is therefore to formalize a continual learning framework for IR and instantiate the notion of Forward Transfer and Backward Transfer in IR.

Investigating catastrophic forgetting phenomenon in short streams [Lovón-Melgarejo *et al.* 2021]. We design here short streams of tasks through datasets of different domains. Our empirical analyses exhibit the behavior of neural ranking models regarding the catastrophic forgetting phenomenon. Neural ranking models being often used as second-stage rankers, we aim at measuring the addi-

tional knowledge they capture in a continual learning framework regarding a first-stage ranker relying on exact-matching signals (e.g., BM25) and whether it impacts catastrophic forgetting. Finally, we explore the gain of a well-known lifelong-learning strategy [Kirkpatrick *et al.* 2016] when applied to neural ranking models.

Investigating catastrophic forgetting phenomenon in long streams and controlled IR scenarios [Gerald & Soulier 2022]. In this contribution, we aim at designing and investigating longer continual scenarios, guided by the intuition that the behavior regarding knowledge acquisition and forgetting might be more pronounced. We, therefore, propose and validate a continual learning dataset based on the MSMarco one including three scenarios of topic streams of different sizes (19, 27, and 74 topic sequences). Then we analyze the behavior of neural ranking models in these scenarios and investigate the correlation between task similarity and catastrophic forgetting. Finally, we design and explore controlled IR settings modeling direct transfer, information update, and language drift.

Chapter 2

Background: Neural textual representations

Contents

In this chapter, we provide an overview of neural language models. We refer the reader to different surveys [Naseem *et al.* 2021, Gruetzemacher & Paradice 2022] or tutorials [Meng *et al.* 2021, Flanigan *et al.* 2022] for a complete overview of text-based representation models.

2.1 Language models, word and sentence embeddings

Word representations. Initially used for text generation, language models are derived from Markov chain conditioning the probability of appearance of word w_t to its previous words $w_1, ..., w_{t-1}$ (i.e., $P(w_t|w_1, ..., w_{t-1})$). Bengio *et al.* 2003 have revisited language models to learn for each word a continuous representation in the latent space by maximizing the log-likelihood of the word sequence $w_1, ..., w_T$ through a language model $f()$ as follows:

$$L = \frac{1}{T}\sum_{t=1}^{T} \log f(w_1, ..., w_t; \theta) + R(\theta) \tag{2.1}$$

where $f(w_1, ..., w_t; \theta)$ is the language model of the sequence (with parameters θ) and $R(\theta)$ is a regularization term. The language model estimates the probability of a word w_t given its previous ones $w_1, ..., w_{t-1}$ using neural networks:

$$f(w_1, ..., w_t; \theta) = \prod_t P(w_t|w_1, ..., w_{t-1}; \theta) \tag{2.2}$$

$$with\ P(w_t|w_1, ..., w_{t-1}; \theta) = g(w_t, \mathbf{w_1}, ..., \mathbf{w_{t-1}}; \theta) \tag{2.3}$$

where $g(;\theta)$ is a neural network of parameter θ and $\mathbf{w_t}$ is the embedding of word w_t into the latent space of dimension d. Although pioneering neural language models, this model might be computationally expensive depending on the form of $g(;\theta)$ and intractable for learning embedding through large datasets. The n larger is, the larger the training corpus should be to obtain good estimates (e.g., $10^{4\times2}$ for bi-grams, $10^{4\times3}$ for tri-grams, ...). The model size increases exponentially with n.

Collobert *et al.* 2011 and Mikolov *et al.* 2013a have contributed to reduce the cost of the pretraining word embeddings by 1) using negative sampling and 2) defining the

concept of the window around a central word. The most famous model is the Skip-Gram model [Mikolov *et al.* 2013a] aiming at predicting, through the whole word sequence $w_1, ..., w_T$, the surrounding contextual words given the central word. The loss function is expressed as below:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log P\left(w_t \mid w_{t+j}\right) \tag{2.4}$$

where c is the size of the context. Each word has two representations depending on whether it is a context word or a central word (respectively called output and input vectors $\mathbf{w_O}$ and $\mathbf{w_I}$). The probability $P\left(w_{t+j} \mid w_t\right)$ is estimated as the following softmax function:

$$p\left(w_O \mid w_I\right) = \frac{\exp\left(\mathbf{w_O}^\top \mathbf{w_I}\right)}{\sum_{i=1}^{N}\exp\left(\mathbf{w_i}^\top \mathbf{w_I}\right)} \tag{2.5}$$

where N is the number of terms in the vocabulary, making the sum impractical in practice. Using negative sampling and approximation function, Equation 2.5 is estimated as follows:

$$\log\sigma\left(\mathbf{w_O}^\top \mathbf{w_I}\right) + \sum_{i=1}^{k}\mathbb{E}_{w_i\sim P_n(w)}\left[\log\sigma\left(-\mathbf{w_i}^\top \mathbf{w_I}\right)\right] \tag{2.6}$$

where $\sigma(\mathbf{x}) = 1/(1+\exp(-\mathbf{x}))$ and w_i is a negative context sampled from the noise distribution $P_n(w)$. A similar model, called C-BOW, learns word representations by reversing the sequence modeling: they predict the central word given its context. A competitive model is Glove [Pennington *et al.* 2014] which relies on both Global Matrix Factorization as done in LSA and local context window as Skip-Gram. Although these models have demonstrated great performances in word similarity or word analogy tasks, they suffer from out-of-vocabulary limitations, since they rely on a predefined vocabulary, and are not able to generate word representations for new words.

Text units. To tackle this issue, several works have rethought language models by focusing on sub-word units, instead of words. These units can take different forms:

- Byte-pair encoding (BPE) [Bojanowski *et al.* 2017] in which words are obtained using a tokenizer and split into Unicode characters. The latter are merged depending on their n-gram frequency to form a new symbol. The merging of symbols can be iterated until reaching the word level.

- Wordpieces [Wu *et al.* 2016] which follow the same principle as BPE but instead of merging the most frequent bigrams, Wordpieces merge the symbol pair that maximizes the likelihood of a unigram language model and the mutual information between these two symbols.

- SentencePiece [Kudo & Richardson 2018] applying BPE or Wordpieces algorithms but at the sentence level without applying any tokenizer and by including space and separation characters.

For convenience, we use in the remainder of this chapter the term *"word"* to mention the different units encoded in the text (i.e., word, tokens, BPE, Wordpieces).

Sentence representations. Learning the representation of sentences or longer texts is a difficult challenge since it raises the question of aggregating semantics over all words in the text. Early works [Yin *et al.* 2016b, Vulić & Moens 2015] have proposed to simply combine linearly word embeddings without considering word order. But, this approach is less prevalent due to its low effectiveness results when used in downstream NLP and IR tasks. Other works focused on extending language models until then applied at the word level to consider the sentence level [Dai *et al.* 2015]. This imposed to modify the input granularity level by averaging/summing words in the sentence to predict, for instance, a central word given the whole sentence [Dai *et al.* 2015]. Other losses have been designed to predict words in previous/next sentences [Hill *et al.* 2014] or also to predict contextual sentences [Kenter & De Rijke 2015].

The document representation approaches have evolved with the development of RNN. This architecture offers a relevant alternative with respect to MLP for encoding text in which the sequence of words is prevalent. Language models are thus estimated by recursively encoding previous words in the sequence as follows:

$$f(w_1, ..., w_t; \theta) = P(w_t | w_1, ..., w_{t-1}) \quad (2.7)$$

$$\approx g(V\mathbf{s_t}; \theta_g) \quad (2.8)$$

$$with\ s_t = h(W\mathbf{s_{t-1}} + U\mathbf{w_t}; \theta_h) \quad (2.9)$$

with $g(;\theta_g)$ and $h(;\theta_h)$ are usually non linear and linear functions, respectively. U, V and W are weighting matrices and $\mathbf{s_t}$ is the hidden state of the t^{th} word. To learn text representations with RNN, different losses might be used, generally varying between word/sentence classification [Logeswaran & Lee 2018] or generation [Kiros *et al.* 2015] tasks. RNN is the backbone of auto-encoder or encoder-decoder architectures which have been largely used and are still used in NLP and IR tasks [Wang *et al.* 2016, Cho *et al.* 2014]. While auto-encoders aim at reconstructing the input based on its projection in the latent space, encoder-decoder (and particularly seq2seq) aims at mapping an input sentence to a different output sentence. While training these architectures using the maximum likelihood criterion (i.e., teacher forcing [Lamb *et al.* 2016]), the distribution of words in the sequence generation might differ from the original ground truth labels. Indeed, teacher forcing consists in re-injecting the correct word as input for the next word generation process. This consequently restrains the word distribution seen during training, which is critical in case of discrepancy between training and inference. Erroneous words generated might lead to an inconsistent subsequent generation. This problem refers to as exposure bias and is generally addressed by using domain adaptation [Goyal *et al.* 2016], reinforcement learning [Ranzato *et al.* 2016], adversarial training [Scialom *et al.* 2020], or learning to search [Wiseman *et al.* 2018].

These architectures have also been improved with an attention mechanism allowing to learn a linear combination of the representation of words/parts of the sentence to build the context vector. The weights of the linear combination are called attention weights and denote the importance of specific parts in the sentence. When sentences are long, the multiplication of gradients in an RNN might lead to a vanishing gradient. To overcome this limitation, memory-based RNN networks, such as LSTM [Bahdanau *et al.* 2015] or GRU [Cho *et al.* 2014], have been proposed. Also, bi-directional encoding has been used [Schuster & Paliwal 1997] to take into account the whole sentence input (rather than

only previous words). It is based on two RNNs, respectively moving forward from the beginning and backward from the end of the text. The final decision is then taken on the concatenation of both hidden states. At the inference step, the network generates a single word/token given the current hidden state (we call this technique *greedy decoding*). This might hinder the intelligibility of the generated sentence since there is nothing to avoid repetitions or non-fluent sentences. To tackle this issue, beam search [Huang *et al.* 2018] or sampling strategies [Holtzman *et al.* 2019] are generally used to maximize the text likelihood.

2.2 Contextual embeddings, large language models, and foundation models

One limitation of the language models aiming at representing words presented above is that we assume that a word must be represented by the same vector regardless of the context in which the word occurs. It is not coherent with the different meanings a word can have depending on its context (i.e., polysemy) or the different entities a surface form can refer to (e.g., does the term "Washington" refer to the city or the politician?).

One of the first attempts to solve word polysemy has been proposed by [Iacobacci *et al.* 2015] with their SENSEMBED model by leveraging word senses inventoried in the BabelNet resource. Later, Peters *et al.* 2017 address this issue for a Named Entity Recognition task. In an end-to-end fashion, the authors propose to combine word embeddings and recurrent language models with the objective that the latter contextualizes the word embedding. Said otherwise, instead of simply using language models to map a word to a predefined vector, the recurrent language model refines word embeddings according to the sequence of words (e.g., the sequence "Washington eats" suggests that "Washington" is more likely the politician rather than the city). In addition, McCann *et al.* 2017 propose to leverage the natural machine translation task to learn contextual embeddings. Their intuition is that the objective of machine translation is to preserve word meaning even if the input/output languages are not the same. To do so, they rely on a two-layer bi-LSTM encoding English texts to decode in an attention-driven seq2seq architecture.

The real first breakthrough contribution for learning contextual word embeddings has been proposed by [Peters *et al.* 2018] with the ELMo embeddings (Embeddings from Language Models). Their strength relies on several aspects: 1) they learn word embeddings using long contexts (long sentences/paragraphs) instead of context window, 2) they learn a bi-directional neural language model and use all the network layers in the prediction, 3) the importance of each layer depends on the targeted task. Experiments have shown that layers encode different aspects of text understanding: lower layers are generally related to syntax analysis (part-of-speech tagging, NER, ...) while higher layers capture high-level semantics (question answering, sentiment, ...).

In parallel, Vaswani *et al.* 2017 have introduced the self-attention mechanism within the powerful model called Transformer. Self-attention allows to compare a sequence with itself using three vectors (query, key, values), and therefore identifies which part(s) of the sequence is(are) important for each word in the sequence. In addition,

Transformer is an encoder-decoder architecture composed of multiple heads, including themselves several blocks of self-attention and feed-forward networks, surrounded by skip-connections. Starting from these statements, several models [Devlin *et al.* 2019, Radford & Narasimhan 2018, Raffel *et al.* 2020] have been developed, guided by the intuition that a neural language model should be able to encode low-level and high-level semantic and syntax information. Consequently, a standard learning scheme has been established: neural language models should be trained simultaneously on different tasks to capture all this information. The number of parameters has been exponentially increased with respect to previous language models, switching training time from a few GPU hours to several GPU days. This is the beginning of Large Language Models with three main lines of models, depending on which part of the Transformer they use: BERT (encoder), GPT (decoder), and T5 (encoder-decoder).

- BERT [Devlin *et al.* 2019] is an encoder-only model: it relies on the encoder part of the Transformer. Its encoder is bi-directional and is trained using two unsupervised tasks: 1) masked language modeling aiming at recovering words removed from the text, and 2) next sentence prediction. This pre-trained language model can be fine-tuned on a given task by simply adding a task-based classifier on top of BERT. This model has attracted a lot of attention in the community due to its effectiveness to capture word semantics and to solve NLP tasks. This led to several BERT-based models and also to a research field, called Bertology, [Rogers *et al.* 2020, Dai *et al.* 2022] aiming at explaining the signals captured by the language models through, among other strategies, probing tasks, and locating which level of the architecture is concerned with.

- GPT [Radford & Narasimhan 2018] is a decoder-only model, relying on the decoder module of the Transformer and processing inputs in a uni-directional manner. The training loss is similar to the one of neural language models, i.e. autoregressive text generation. In contrast to BERT models which require fine-tuning, GPT has shown great abilities in zero-shot / few-shot learning settings. Two variants were derived from GPT (GPT-2 [Radford *et al.* 2019] and GPT-3 [Brown *et al.* 2020]), different in terms of number parameters (from 117 M in GPT to 1.5 B in GPT-2 to 175B in GPT-3). Also, GPT-4 is on the way to be presented to the community. Recently, the extended version of GPT-3 (i.e. GPT-3.5) was used as a basis for the smashing ChatGPT model (inspired by [Ouyang *et al.* 2022]) exhibiting powerful ability for dialogue and text generation.

- T5 [Raffel *et al.* 2020] follows the Transformer encoder-decoder architecture but is trained on five unsupervised and supervised tasks: masked language modeling and next sentence prediction as in BERT, and translation, question answering, and classification tasks. These are all framed as text-to-text tasks. T5 is effective on other NLP tasks if we prefix a short instruction to the input (e.g., for translation: "translate English to German: ...").

These models are now well-established and largely used as standard models for NLP and IR tasks, fostering the extension of these architectures for devoted tasks or also improving the language models according to different aspects. We can cite some examples through the following (non-exhaustive) list:

- RoBERTa [Liu *et al.* 2019d] which removes the next sentence prediction loss and trains the masked language-modeling task on more data.
- Distillation models, such as DistillBert [Sanh *et al.* 2019], leading to a smaller and faster Transformer model.
- Models reducing the number of parameters. For instance, ALBERT [Chiang *et al.* 2020] splits the embedding matrix into two smaller matrices, and, thus, leads to lower memory consumption and faster training.
- Adversarial models, such as Electra [Clark *et al.* 2020] which relies on adversarial training to distinguish real and fake texts.
- Sparse models: Transformer-XL [Dai *et al.* 2019] and XLNet [Yang *et al.* 2019c] using sparse attention. A version for long text encoding, such as LongFormer [Beltagy *et al.* 2020], is also available.
- Specialized models: for scientific texts (SciBert [Beltagy *et al.* 2019]), for French (Camembert [Martin *et al.* 2019] or FlauBert [Le *et al.* 2019]), for multiple languages [Pires *et al.* 2019, Scao *et al.* 2022a, Chowdhery *et al.* 2022a].

All these variants have contributed to the success of Transformer-based models in NLP but also in other domains, such as vision [Khan *et al.* 2022] or time series [Wu *et al.* 2020].

Beyond that, the different lessons learned from model training (effective knowledge/semantics/syntax captured through self-attention and multi-task learning) and, accordingly, the breakout step in the effectiveness of unsupervised learning with no necessary fine-tuning on the targeted task have opened tremendous perspectives towards a new line of models called *foundation models* [Bommasani *et al.* 2021, Cui *et al.* 2022, Kiela 2022]. The objective of foundation models is to pre-train models on a sufficiently diverse set of modalities and tasks, as well as on large datasets, so that they can learn world knowledge at different granularity levels, and therefore be effective on a new task in a few-shot or zero-shot setting [Cui *et al.* 2022]. Said otherwise, foundation models are a generalization of large language models and change the paradigm of task-focused models towards large knowledge models able to solve any task without fine-tuning. It is worth noting that prompt-based large language models, such as the FLAN model [Wei *et al.* 2022] or the T0 from BigScience [Sanh *et al.* 2022], are the premises of those foundation models, but they only consider the textual modality. One critical issue of those models is the computational cost to train those models [Bender *et al.* 2021]. Also, although the perspective of having a single model to perform various tasks (or fine-tuning it with a small amount of data) is quite exciting, ethical questions remain (which are also valid for previous models)[Bender *et al.* 2021]: what is the dataset used for training? Should we be aware of possible bias in the model decision? Even if we fine-tune the foundation model on our small dataset, do we have control over the model output since the training was performed on a large variety of datasets/tasks?

Chapter 3

Research fields: data-to-text generation, conversational search, and continual learning

Contents

3.1 Data-to-text generation

3.1.1 General overview

Data-to-text generation (DTG) aims at understanding structured data (e.g., key-value pairs, graphs or RDF triplets, tables, charts, figures, temporal data) and describing it with natural language descriptions [Reiter & Dale 2000, Reiter 2007]. This field is relevant for several application domains (such as journalism [Oremus 2014] or medical diagnosis [Pauws *et al.* 2019]) or in wide-audience applications (such as financial [Plachouras *et al.* 2016], weather reports [Reiter *et al.* 2005], or sport broadcasting [Chen & Mooney 2008, Wiseman *et al.* 2017]). This is a subfield of Natural Language Generation (NLG) in which the objective is to generate new texts, often conditioned by an input.

Given an input x, the objective of conditioned text generation is to generate a text y. This generation is mainly performed through auto-regressive learning that aims a maximizing at a given step t, the probability $p(\hat{y}_t|\hat{y}_1, \ldots, \hat{y}_{t-1}, x)$ to generate word $\hat{y}_t$ given input x and all previously generated words $\hat{y}_1, \ldots, \hat{y}_{t-1}$. This research field has been boosted by the recent advances in deep learning,

TEAM	H/V	WINS	LOSSES	PTS	REB	AST	...
Hawks	H	46	12	95	42	27	...
Magic	V	19	41	88	40	22	...

PLAYER	PTS	REB	AST	STL	BLK	CITY	...
Al Horford	17	13	4	2	0	Atlanta	...
Kyle Korver	8	3	2	1	2	Atlanta	...
Jeff Teague	17	0	7	2	0	Atlanta	...
N. Vucevic	21	15	3	1	1	Orlando	...
Tobias Harris	15	4	1	2	1	Orlando	...
...	...	...	...	...	...	...	...

H/V: home or visiting; PTS: points; REB: rebounds; AST: assists; STL: steals; BLK: blocks

The **Atlanta Hawks** (**46-12**) beat the **Orlando Magic** (**19-41**) **95-88** on Friday. **Al Horford** had a good all-around game, putting up **17 points, 13 rebounds, four assists and two steals** in a tough matchup against **Nikola Vucevic**. **Kyle Korver** was the lone Atlanta starter not to reach double figures in points. **Jeff Teague** bounced back from an illness, he scored **17 points** to go along with **seven assists and two steals**. After a rough start to the month, the **Hawks** have won three straight and sit atop the Eastern Conference with a nine game lead on the second place Toronto Raptors. The **Magic** lost in devastating fashion to the Miami Heat in overtime Wednesday. They blew a seven point lead with 43 seconds remaining and they might have carried that with them into Friday's contest against the **Hawks**. **Vucevic** led the **Magic** with **21 points and 15 rebounds**. **Aaron Gordon** (ankle) and **Evan Fournier** (hip) were unable to play due to injury. The **Magic** have four teams between them and the eighth and final playoff spot in the Eastern Conference. The **Magic** will host the Charlotte Hornets on Sunday, and the **Hawks** with take on the Heat in Miami on Saturday.

Figure 3.1: Example of a training instance in the RotoWire dataset [Wiseman *et al.* 2017].

proposing more and more sophisticated architectures (from multi-layer perceptron-MLP [Rosenblatt 1958], Recurrent Neural Network-RNN [Graves *et al.* 2014], Long-Short Term Memory networks-LSTM [Bahdanau *et al.* 2015] to large language models [Vaswani *et al.* 2017, Devlin *et al.* 2019]) which are more and more effective to capture the semantics. The backbone architecture in NLG is an encoder-decoder that projects the input data into a latent semantic space (the encoder part) and then estimates a probability distribution over the vocabulary to build a sequence of words as output (the decoder part). The parameters of the language generation model are learned by maximizing the cross entropy loss between the generated text $\hat{y}$ and the gold reference text y given input x:

$$\arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{(x,y)\in\mathcal{D}} \log P(\hat{y} = y|x;\theta) \tag{3.1}$$

where $\mathcal{D}$ is the training dataset including pairs of input-output texts. During inference, the sequence $\hat{y}$ is generated using a greedy decoding aiming at approximating the maximum likelihood probability conditioned on the input data x:

$$\hat{y}^{*}_{1:T} = \arg\max_{\hat{y}_{1:T}} \prod_{t=1}^{T} P(\hat{y}_t|\hat{y}_{1:t-1}, x;\theta) \tag{3.2}$$

where $\hat{y}_{1:T}$ corresponds to a generated sequence of T words.

DTG follows this task formalization, except that the input data x is a structured data and not a free-form text. As discussed earlier, the main characteristic of DTG is that input and output data (x and y) might be of different formats and different types. This makes even harder the alignment between elements in a structured data and their translation into a natural language text. For instance, in the RotoWire dataset [Wiseman *et al.* 2017] in Figure 3.1, the input data is a table about statistics of basketball players and the output data is a summary of the basketball game. During training, it is not obvious at all that a statistical model can identify that the term "assists" in the gold reference text refers to the column "AST" or that the expression "lost devastating fashion to" requires comparing the columns "WINS" in the first table and to reason that the line with the lower score refers to the looser team.

3.1.2 From rule-based to deep learning models

Until recently, efforts to bring out semantics from structured data relied heavily on expert knowledge [Deng *et al.* 2013, Reiter *et al.* 2005]. For example, to better transcribe

numerical time series of weather data to a textual forecast, [Reiter *et al.* 2005] implement complex template schemes in collaboration with weather experts to build a consistent set of data-to-word rules. From a general point of view, expert-based systems follow a pipeline of three main steps: *content selection* aiming at identifying salient information (also called *macro-planning*), *micro-planning* focusing on the content ordering to build a plan of the textual description, and *surface realization* generate the sentences in natural language. Although accurate for the devoted domain and efficient at inference time, one drawback of expert-based systems is that they are human costly to adapt to new use cases.

With the rise of neural networks, research turned towards the use of neural text generation models. Encoder-decoder networks augmented with attention [Bahdanau *et al.* 2015] and copy [Gulcehre *et al.* 2016, See *et al.* 2017] mechanisms are rapidly adopted as backbone models. The different subtasks in expert-based systems disappear with end-to-end training on aligned pairs of data and texts [Gatt & Krahmer 2018], framed as text-to-text generation models. To fit with these architectures, the authors represent the data as a single sequence of facts (pairs of key-value possibly associated with an entity) to be entirely translated into natural language. For example, the table from Figure 3.1 is linearized to [(Hawks, H/V, H), ..., (Magic, H/V, V), ...], effectively leading to losing the distinction between rows, and therefore entities. Moreover, Wiseman *et al.* 2017 show the limits of traditional encoder-decoder models on larger structured-data, since they fail to accurately extract salient elements.

A line of works proposes to leverage text generation based on macro-planning [Kondadadi *et al.* 2013] to design neural decoders based on planning and templates. The intuition of such a module is to ensure factual and coherent mentions of input records in generated descriptions. For example, Puduppully *et al.* 2019a propose a two-step decoder which 1) first identifies "what to say" - i.e. an ordered plan of salient information that should be included in the summary - (referring to as *content selection and planning*), and 2) then focuses on the "how to say" by generating fluent sentences by following the plan built in the previous step (referring to as *text generation*). The joint probability of generating a text y given a data structure s is thus decomposed as the product, over all possible plans z, of probabilities $p(z \mid s)$ and $p(y \mid s, z)$, respectively denoting the content selection and planning step, and the text generation step:

$$p(y \mid s) = \sum_{z} p(z \mid s) p(y \mid s, z) \tag{3.3}$$

From a reverse point of view, Li & Wan 2018 propose a delayed copy mechanism for which their decoder acts in two steps: 1) using a classical LSTM decoder to generate a fill-in-the-blank text and 2) using a pointer network [Vinyals *et al.* 2015] to replace placeholders by records from the input data.

3.1.3 Ensuring faithful generation...

... by taking into account the structure. Aware of the limitation underlying the linearization of input data, some works [Liu *et al.* 2018, Liu *et al.* 2019a, Puduppully *et al.* 2019b] propose to take into account the data structure. In TAPAS [Herzig *et al.* 2020], the authors propose to add embeddings contextualizing tokens in the table (e.g., column and row IDs, order of magnitude for numerical variables). A

further step is proposed by Liu *et al.* 2018 and Liu *et al.* 2019a with their dual encoder [Liu *et al.* 2019a] which encodes separately the sequence of element names/labels and the sequence of element values. These approaches are however designed for single-entity data structures and do not account for delimitation between entities. Taking into account multiple entities, Iida *et al.* 2021 propose to encode the table using two transformers, respectively for rows and columns. Each cell is then contextualized over these two dimensions.

Considering entities as a whole, Puduppully *et al.* 2019b follow entity-centric theories [Grosz *et al.* 1995, Mann & Thompson 1988] and propose a model based on dynamic entity representation at decoding time. It consists in conditioning the decoder on entity representations that are updated during inference at each decoding step. For instance, Puduppully *et al.* 2019b introduce dynamic encoding updating, where the model updates part of the source data encoding at each decoding step to accurately guide the decoder throughout the generation. Recently, Wang *et al.* 2022 leverage transformation invariance and structure awareness through attention flow to understand cell relations and reinforce the model robustness regarding the data structure.

In parallel, recent works investigate on answering questions on tables [Pasupat & Liang 2015, Yin *et al.* 2016a, Sun *et al.* 2016, Yin *et al.* 2020, Chen *et al.* 2021]. Early works propose to leverage semantic parsing and build knowledge graphs from the table [Pasupat & Liang 2015, Sun *et al.* 2016] or to simply encode each cell in the semantic space [Yin *et al.* 2016a]. Other approaches encode tables in a very similar way to those described above. More particularly, they generally follow the linearization principle over all cells [Chen *et al.* 2021] or at the row level [Yin *et al.* 2020] but integrate an attention module guided by the question or retrieval techniques to identify relevant information [Chen *et al.* 2021].

... by controlling hallucinations. Another drawback of previous models (and text generation models in general) is that they are subject to over-generation [Elsahar *et al.* 2021], i.e., hallucinations. Most of the available corpora are often constructed from internet sources, which, while easy to access and aggregate, do not consist of perfectly aligned source-target pairs [Perez-Beltrachini & Gardent 2017, Dhingra *et al.* 2019]. Therefore, misaligned fragments from training instances, namely *divergences*, can induce similarly misaligned outputs during inference, the so-called *hallucinations*. This problem arises both from the training procedure (training via maximum likelihood leads to language models strongly mimicking human behaviors), and from the testing protocols. Indeed, standard metrics (e.g., BLEU [Papineni *et al.* 2002], ROUGE [Lin 2004], METEOR [Banerjee & Lavie 2005]) only measure similarity to ground truth reference texts and do not fully capture relevance to the source data. Thus, there is no distinction between a mismatch caused by a paraphrase, poor lexicalization of content, or made-up/incorrect statement, leading to imperfect model selection.

When corpora include a mild amount of noise, as in handcrafted ones (e.g. E2E, WebNLG), dataset regularization techniques [Nie *et al.* 2019, Dusek *et al.* 2019] or handcrafted rules [Juraska *et al.* 2018] can help to reduce hallucinations. For instance, Juraska *et al.* 2018 leverage templating and hand-crafted rules to re-rank the top outputs of a model decoding via beam search. However, beyond the significant annotation labor, all proposed neural approaches still suffer from *exposure bias* underlying teacher

forcing training. To overcome these limitations, a strategy [Shen *et al.* 2020] consists in increasing the coverage of neural outputs, by constraining the decoder to focus its attention exclusively on each table cell sequentially until the whole table was discussed in the narrative. Similarly, Wiseman *et al.* 2017 propose to include a reconstruction loss aiming at reconstructing the source table from the hidden states of the decoder. In another direction, Perez-Beltrachini & Lapata 2018 introduce a classifying neural network, trained (via multi-instance training) to label text tokens depending on their alignment with the associated table. They use these labels in a reinforcement learning framework to generate sentences with a maximum of aligned tokens. Also, Liu *et al.* 2019c propose a reward based on document frequency to favor words from the source table more than rare words.

Leveraging controlled text generation [Li *et al.* 2016, Sennrich *et al.* 2016, Lample *et al.* 2019], Filippova 2020 introduces an *hallucination score* simply attached as an additional attribute that reflects the degree of hallucinated content in the associated target description. During inference, this attribute acts as an *hallucination handle* used to produce more or less factual text. However, this approach is not without limitations since it requires a strict alignment at the instance level, namely between control factors and the whole target text.

3.2 Conversational search

3.2.1 General overview

Search-oriented conversational systems are characterized by a heterogeneous context involving: 1) an IR system retrieving documents according to an information need and/or collecting users' clicks, and 2) a dialogue system interacting with the user in natural language to improve the search experience. The purpose of conversational IR systems is thus to replace or augment IR systems to support users during their search session [Culpepper *et al.* 2018]. Depending on the interaction mode, users might interact with both the search engine and the dialogue system, or simply with the dialogue system which also displays information snippets or documents throughout the conversation. This setting is relevant for complex and/or exploratory information needs that require multiple steps or document recommendations. In addition, Radlinski & Craswell 2017 and Culpepper *et al.* 2018 define such so system as a pro-active system in which the collaboration is jointly conducted by the user and the system (we call these interactions "mixed-initiative" ones). Conversational search systems have also the role of 1) eliciting information needs by asking clarifying questions [Zamani *et al.* 2020a] and 2) maintaining the conversation awareness in order to avoid repeated questions from the system side and provide the user reminder from previous sessions or previous search interests. Radlinski & Craswell 2017 characterize conversational search by two learning processes: user revealment in which the system helps the user to clarify and learns about his/her need, and system revealment in which the user leverages the system's abilities to increase his/her knowledge. An interesting synthesis of all these notions has been proposed at the devoted Dagstuhl seminar [Anand *et al.* 2020] and is illustrated in Figure 3.2. Particularly, the report highlights the dimension of conversational search based on: 1) user and system engagement toward the conversation, 2) the concurrency of the different interactions which should be immediate but with a possible delayed task achievement, 3) the

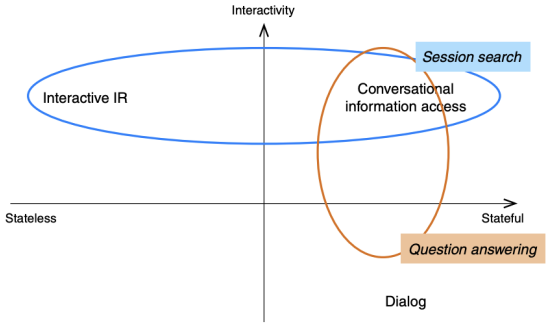


Figure 3.2: Dimension of conversational search regarding other fields in IR - image from [Anand *et al.* 2020].

naturalness of interactions, 4) the interactivity level, and 5) the state of the conversation to ensure the session awareness.

It is worth noting that conversational IR has a strong relation with general dialogue systems [Roller *et al.* 2020], both characterized by a multi-turn conversation between the user and the system. However, in contrast to chitchat conversational systems [Ritter *et al.* 2011, Li *et al.* 2016] that just aim to keep the conversation going, the purpose of introducing conversational systems in IR is to use natural language interactions to find the desired relevant pieces of information over large document collections. It is also different from a task-oriented conversation (e.g., restaurant booking [Bordes & Weston 2016]) evolving in a closed world [Seneff & Polifroni 1996, Wang & Lemon 2013]. It is worth noting that conversational IR is also different from question-answering (QA) [Bordes *et al.* 2014, Haug *et al.* 2018] according to the final goal. Indeed, conversational search aims at solving information needs that are often under-specified and complex to explicit [Jansen *et al.* 2000], in contrast to question-answering which often focuses on a fact or an entity. For instance, in question-answering, typical questions might be "When was Franklin D. Roosevelt born?" [Roberts *et al.* 2020] or also "What does the zip in zip code stand for?" [Lee *et al.* 2019] while information needs in conversational search were initially defined by keywords (e.g. "dinosaur" or "south Africa" [Aliannejadi *et al.* 2019]). With the impressive results of large language models [Devlin *et al.* 2019, Radford *et al.* 2019], the frontier between question-answering and conversational search is dissolving: question-answering tends to address more complex questions requiring multi-hop reasoning over different documents, and conversational search limits the usage of keyword queries for the benefit of natural language questions which might be more explicit.

3.2.2 Contextualizing information needs in conversations

Understanding an information need formulated in natural language is a central issue for conversational systems [Mikolov *et al.* 2015] and a longstanding goal in IR [Jansen *et al.* 2000, Cronen-Townsend & Croft 2002, Sanderson 2008].

One first line of works relies on query reformulation [Rocchio 1971, Lavrenko & Croft 2001, Amati & Van Rijsbergen 2002, Zukerman & Raskutti 2002] where the objective is to rewrite the query. A lot of effort has been provided to design models based on either (pseudo-)relevance feedback

[Rocchio 1971, Lavrenko & Croft 2001, Amati & Van Rijsbergen 2002] or external knowledge resources [Zukerman & Raskutti 2002].

Another category of works focuses on search/query diversification [Carbonell & Goldstein 1998, Agrawal *et al.* 2009, Cai *et al.* 2016, Nogueira *et al.* 2019a, MacAvaney *et al.* 2021] to increase the query coverage, particularly when the query is multi-faceted. Recently, MacAvaney *et al.* 2021 proposed to focus on query diversification by generating queries by designing a Distributional Causal Language Modeling. However, for all these diversification techniques, the issued document list might include some top-ranked documents that do not match the user's intent [Wang & Zhu 2009].

The keen interest in conversational search has shown that it is possible to better understand queries by taking into account the session context that is the different utterances of the conversation. While a few works have proposed to model IR sessions as sequential actions, and thus, using agents [Nogueira *et al.* 2019a, Tang & Yang 2019, Chen *et al.* 2020c], most prior works rely on a Historical Query Expansion step [Lin *et al.* 2020b, Zamani *et al.* 2022b]. Inspired by previous work modeling users based on their search logs to infer their search intent [Xiang *et al.* 2010, Matthijs & Radlinski 2011, Bennett *et al.* 2012, Harvey *et al.* 2013, Kong *et al.* 2015], this approach consists of a query expansion mechanism that takes into account all past queries and their associated answers. Such query expansion model is learned on the CANARD dataset [Elgohary *et al.* 2019], which is composed of a series of questions and their associated answers, together with a disambiguated query, i.e. *a gold query*. However, relying on a reformulation step is computationally costly and might be sub-optimal as in [Lin *et al.* 2021b, Krasakis *et al.* 2022]. [Krasakis *et al.* 2022] propose to use ColBERT [Khattab & Zaharia 2020] in a zero-shot manner, considering as input a sequence of queries (instead of a single query), without any training of the model. Lin *et al.* 2021b propose to learn a dense *contextualized* representation of the query history, optimizing a learning-to-rank loss over a dataset composed of weak labels.

A promising approach has been proposed in [Aliannejadi *et al.* 2019, Krasakis *et al.* 2020, Aliannejadi *et al.* 2021, Sekulic *et al.* 2021a, Tavakoli *et al.* 2022] to clarify information needs by proactively interacting with the user. Inspired by previous work in voice queries [Kiesel *et al.* 2018] and dialogue systems [Stoyanchev *et al.* 2014], Aliannejadi *et al.* 2019 propose a conversation framework that consists in generating clarifying questions when the query is ambiguous. Clarifying questions might be query reformulations (e.g., "Would you like to know how to care for your dog during heat?" for the initial query "dog heat" as in [Aliannejadi *et al.* 2019]) or questions with possible options (e.g., "what do you want to know about this British mathematician? Options: movie, suicide note, quotes, biography" for the initial query "alan turing" as in [Zamani *et al.* 2020a]). With this in mind, the classic workflow for asking clarifications is based on three main steps [Aliannejadi *et al.* 2019]: 1) the IR system produces a clarifying question for the user, 2) the latter provides an answer or selects an option, and 3) the IR system ranks documents according to the user's feedback. The pioneering work [Aliannejadi *et al.* 2019] aims at generating clarifying questions by 1) retrieving a predefined set of questions using a BERT-based model and 2) at each turn, selecting the best query through a conversation history-driven model. One drawback of this approach, due to the cost of using real user interactions, is that the multi-turn conversation is

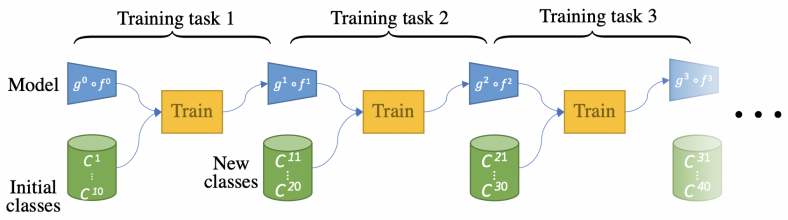


Figure 3.3: Training procedure in a continual learning setting [Douillard 2022].

log-based, interactively simulated using predefined logs of conversation history (i.e., sequence of questions/answers obtained by HITS). This simulated conversation defined a priori without interaction with the proposed question selection model might hinder the evaluation performance in the sense that we are not sure about the soundness of the conversation flow. Zamani *et al.* 2020a and Sekulic *et al.* 2021a tackle this issue by proposing generative models, that create clarification questions or query suggestions. But they do not address the multi-turn framework, stopping the clarification process at the first interaction.

3.3 Continual learning

3.3.1 General overview

Continual learning generally defines the setting in which a model is trained consecutively on a sequence of tasks and needs to adapt itself to newly encountered tasks [Lomonaco & Maltoni 2017]. In Figure 3.3, [Douillard 2022] illustrates the training procedure underlying continual learning settings. Formally, let's consider a sequence of classification tasks $\mathcal{T}_1 \rightarrow \mathcal{T}_2 \rightarrow \cdots \rightarrow \mathcal{T}_n$ which respectively aims at classifying data on ten different labels. Said otherwise, each timestamp t of the sequence is associated to a classification dataset $\mathcal{D}_t$ based on ten different classes: classes $C^1, \ldots, C^{10}$ for task $\mathcal{T}_1$, classes $C^{11}, \ldots, C^{20}$ for task $\mathcal{T}_2$, and so on. Training a neural model $\mathcal{M}_0 = g^0 \circ f^0$ on this sequence consists in building incrementally the model $\mathcal{M}_t = g^t \circ f^t$ at each timestamp t on the basis of the previous model $\mathcal{M}_{t-1}$ and the dataset $\mathcal{D}_t$. This setting is also named class-incremental learning setting.

To enhance the transferability of neural models from a source domain to a target domain, transfer learning strategies such as fine-tuning [Yang *et al.* 2019b], multi-tasking [Liu *et al.* 2015], domain adaptation [Pan & Yang 2010], and more recently adversarial learning [Cohen *et al.* 2018], have been widely used. However, these strategies have in essence two critical limitations reported in the machine learning literature [Chen & Liu 2018, Kirkpatrick *et al.* 2016]. The first one, which is also acknowledged in the NLP and IR communities [Cohen *et al.* 2018, Liu *et al.* 2015], is that they require all the domains to be available simultaneously at the learning stage (except the fine-tuning). The second limitation is that the model leans to *catastrophically forget* existing knowledge (source domain) when the learning is transferred to new knowledge (target domain), leading to a significant drop in performance on the source domain [Kirkpatrick *et al.* 2016]. Investigating catastrophic forgetting is addressed as a research field in its own right called *lifelong learning*. It has been particularly studied in neural-network-based classification

tasks in computer vision [Kirkpatrick *et al.* 2016, Li & Hoiem 2018, Veniat *et al.* 2020, Douillard *et al.* 2020b] and more recently in NLP [de Masson d'Autume *et al.* 2019, Mosbach *et al.* , Thompson *et al.* 2019, Wiese *et al.* 2017, Lee 2017, Veron *et al.* 2019, Liu & Mazumder 2021].

3.3.2 Lifelong learning strategies

To solve the catastrophic forgetting issue, three main categories of works can be outlined [Lange *et al.* 2019].

Regularization approaches continually learn to address new tasks using soft or hard preservation of weights [Kirkpatrick *et al.* 2016, Wiese *et al.* 2017, Zenke *et al.* 2017, Li & Hoiem 2018]. For instance, the *Elastic Weight Consolidation* model [Kirkpatrick *et al.* 2016] softly updates weights for a new task according to their importance in the previous one. The intuition is to leverage the diagonal Fisher information matrix to model importance factors and identify which parameters are important for a task. Li & Hoiem 2018 and [Rebuffi *et al.* 2017] propose to constraint weights in the network on the basis of its output through knowledge distillation techniques.

Rehearsal approaches replay examples of previous tasks while training the model on a new one [Rebuffi *et al.* 2017, Asghar *et al.* 2020, de Masson d'Autume *et al.* 2019]. The number of previous instances might be limited to respect the continual learning setting. Different strategies are used to choose examples: random sampling [Castro *et al.* 2018], nearest-neighbor sampling in the latent space [Castro *et al.* 2018], uniform sampling over all classes [Chaudhry *et al.* 2019], or sampling regarding the loss criteria [Aljundi *et al.* 2019]. Another work [Lesort *et al.* 2019] proposes to generate pseudo-samples for rehearsal by leveraging Generative Adversarial Networks or auto-encoders. However, this last strategy suffers from catastrophic forgetting and is not always able to generate instances (i.e., images) of adapted sizes.

Architecture-based approaches rely on a dynamic strategy to adapt the network architecture for each task [Fernando *et al.* 2017, Cai *et al.* 2019, Li *et al.* 2019, Veniat *et al.* 2020, Yan *et al.* 2021]. The first line of works proposes to adapt by activating/deactivating parts of the network as done in [Fernando *et al.* 2017, Cai *et al.* 2019, Li *et al.* 2019] while other works [Veniat *et al.* 2020, Yan *et al.* 2021] investigate a strategy consisting in expanding the network through neural architecture search.

Main investigation in NLP. There is a recent research trend in NLP toward lifelong learning of neural networks, particularly in machine translation [Thompson *et al.* 2019, Garcia *et al.* 2021], language understanding tasks [Mosbach *et al.* , Wiese *et al.* 2017, Xu *et al.* 2018a, Sun *et al.* 2020], and for conversational systems [Lee 2017, Veron *et al.* 2019, Liu & Mazumder 2021]. For instance, Xu *et al.* 2018a have recently revisited the domain transferability of traditional word embeddings [Mikolov *et al.* 2013a] and proposed *lifelong domain embeddings* using a meta-learning approach. The meta-learner is fine-tuned to identify similar contexts of the same word in both past domains and the new observed domain. In LAMOL [Sun *et al.* 2020], the authors introduce a language model that jointly learns a new task and generates

training instances for previous tasks. These generated instances are integrated into the training procedure of the new task, consisting thus of a rehearsal strategy. In IR, despite the existence of ranking approaches able to perform well on different domains (e.g., batch-balanced topics [Hofstätter *et al.* 2021]), lifelong learning is still under-studied.

Part II

Contributions

Chapter 4

Generating faithful textual and relevant texts

Contents

In this chapter, we introduce a summary of our works aiming at ensuring faithful and relevant text generation (RQ1). Depending on the contribution, our works are related to the research fields of DTG or conversational search.

4.1 Preliminary

In our contributions focusing on the data-to-text generation task, we mainly consider structured data as tables or key-value elements, which can easily be modeled similarly: the row and column labels in a table can be seen as an element-key pair, and the corresponding cell as a value. The adaptation to other data structures might not be straightforward depending on their complexity, but the main principle of models can serve as a basis. This choice towards tables or key-value elements was mainly guided by the available datasets in the research community and also the application domain of industrial collaborations. Specifically, our objective was to propose models adaptable to descriptive tables in the financial domain[1]. We, therefore, introduce the formalism with this

[1]The financial domain is the application domain of the thesis on DTG I co-supervised.

perspective. With this in mind, let's introduce the following notations that we use in the next sections.

We consider a DTG task, in which the dataset $\mathcal{D}$ is composed of a set of N data structure-description pairs, (s, y). A data structure s is an unordered set of I entities e_i. We denote $s \coloneqq \{e_1, ..., e_i, ..., e_I\}$. An entity e_i is a variable-sized set of J key-value pairs of key k_{ij} - values v_{ij}: $x_{ij} \coloneqq (k_{ij}, v_{ij})$. Please note that the number J of pairs might vary across entities. A *description* $y \coloneqq y_{1:Y}$ is a sequence of Y tokens representing the (target) natural language description of the data structure s. We refer to the tokens spanning from indices t to t' of a description y as $y_{t:t'}$.

The objective of a DTG model is thus to propose a model that produces a textual description y given a data structure s.

4.2 Leveraging the structure for data-to-text generation

In this section, we introduce our proposed hierarchical model [Rebuffel *et al.* 2020a] taking into account the data structure with the assumption that a good encoding will help to reduce erroneously generated texts. We focus here on tables which might include several lines and columns. Lines often refer to the different entities that are studied, and columns express the different analyzed features. More particularly, our contribution is threefold:

- **Encoding the structure of data:** instead of flatly concatenating elements from the data structure to encode them as a single sequence [Liu *et al.* 2018, Puduppully *et al.* 2019a, Wiseman *et al.* 2017], we propose a hierarchical modeling so that the delimitation between entities remains clear.

- **Using Transformers to account for the arbitrary order of elements:** We believe that RNNs are not well-fitted for encoding some structures, particularly tables in which the order of columns is not particularly relevant. We thus exploit the Transformer architecture [Vaswani *et al.* 2017] and leverage its self-attention to directly compare elements with each other, avoiding arbitrary assumptions on their ordering. We do not use any positional embedding to discard the sequence order.

- **Leveraging hierarchical attention mechanism:** we adapt the attention mechanism to the hierarchical modeling to guide the decoding process.

4.2.1 Model formalization

Our model follows the encoder-decoder architecture [Bahdanau *et al.* 2015] in which we integrate a hierarchical encoder. The latter aims at representing first entities e_i (low-level encoder in Figure 4.1) and then the whole data structure s (high-level encoder). Both the low-level and high-level encoders consider their input elements as unordered and rely on the Transformer architecture. For the decoding module, we used the same as in [Puduppully *et al.* 2019a, Wiseman *et al.* 2017]: a two-layers LSTM network with a copy mechanism.

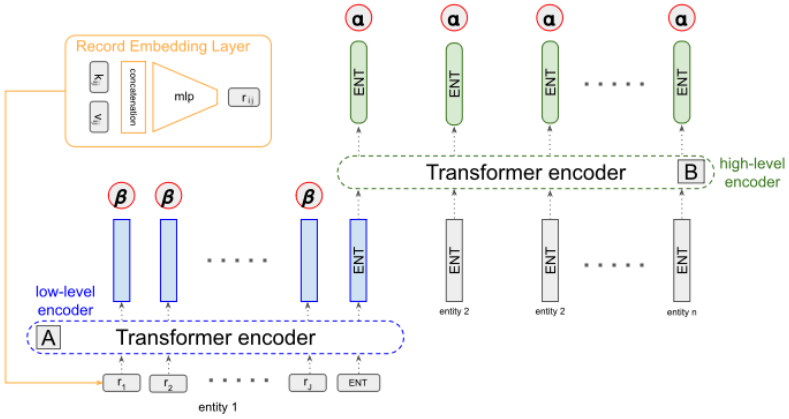


Figure 4.1: Diagram of the proposed hierarchical encoder. Once the records are embedded, the low-level encoder works on each entity independently (A); then the high-level encoder encodes the collection of entities (B). In circles, we represent the hierarchical attention scores: the α scores at the entity level and the β scores at the record level.

Low-level encoder. It encodes each entity e_i on the basis of its record embeddings $\mathbf{x}_{ij}$ obtained from its record x_{ij}. Each record embedding $\mathbf{x}_{ij}$ is compared to other record embeddings using the self-attention mechanism of Transformers to learn its final hidden representation $\mathbf{h}_{ij}$. Furthermore, we add a special record `[ENT]` for each entity, illustrated in Figure 4.1 as the last record. Since entities might have a variable number of records, this token allows to aggregate final hidden record representations $\{\mathbf{h}_{ij}\}_{j=1}^{J}$ in a fixed-sized representation vector $\mathbf{h}_i$. The representation of an entity e_i is thus estimated as follows:

$$\mathbf{h}_i = transformer_{low}([ENT], (\mathbf{x}_{i1}, ..., \mathbf{x}_{i,J}, [ENT])) \quad (4.1)$$

$$with\ \mathbf{x}_{ij} = \text{ReLU}(\mathbf{W}_{kv}[\mathbf{k}_{ij};\ \mathbf{v}_{ij}] + \mathbf{b}_x) \quad (4.2)$$

where $transformer_{low}(x, seq)$ allows to obtain the representation of the token x given the sequence seq using a transformer network. $\mathbf{W}_x \in \mathbb{R}^{2d \times d}$ and $\mathbf{b}_x \in \mathbb{R}^d$ are learned parameters. d is the dimension of the representation space. Each pair is embedded through a linear projection on the concatenation of the embeddings of its key and value: $[\mathbf{k}_{ij}; \mathbf{v}_{i,j}]$.

High-level encoder It encodes the data structure based on entities' representations $\mathbf{h}_i$. Similarly to the **Low-level encoder**, the final hidden state $\mathbf{e_i}$ of an entity is computed by comparing the entity representations $\mathbf{h}_i$. The data-structure representation $\mathbf{z}$ is computed as the mean of the entity representations $\mathbf{e_i}$, and is used for the decoder initialization:

$$\mathbf{z} = \frac{1}{I}\sum_{i=1}^{I}\mathbf{e_i} \quad (4.3)$$

$$with\ \mathbf{e_i} = transformer_{high}(h_i, \{\mathbf{h}_1, ..., \mathbf{h}_I\}) \quad (4.4)$$

$$(4.5)$$

where $transformer_{high}(x, seq)$ allows to obtain the representation of the token x given the sequence seq using a transformer network.

Hierarchical attention To fully leverage the hierarchical structure of our encoder, we adapt the attention mechanism to compute the context fed to the decoder module. Two different approaches are described below:

• *Traditional Hierarchical Attention.* As in [Puduppully *et al.* 2019b], we hypothesize that a dynamic context should be computed in two steps: first attending to entities, then to records corresponding to these entities. To implement this hierarchical attention, at each decoding step t, the model learns the first set of attention scores $\alpha_{i,t}$ over entities e_i and the second set of attention scores $\beta_{ij,t}$ over key-value pairs x_{ij} associated with entity e_i. The $\alpha_{i,t}$ scores are normalized to form a distribution over all entities e_i, and $\beta_{ij,t}$ scores are normalized to form a distribution over pairs x_{ij} of entity e_i. Each entity is then represented as a weighted sum of its record embeddings, and the entire data structure is represented as a weighted sum of the entity representations. The dynamic context is computed as:

$$\mathbf{c_t} = \sum_{i=1}^{I}(\alpha_{i,t}(\sum_j \beta_{ij,t}\mathbf{x}_{ij})) \tag{4.6}$$

$$where \quad \alpha_{i,t} \quad \propto exp(\mathbf{d}_t\mathbf{W}_\alpha\mathbf{e}_i) \quad and \quad \beta_{ij,t} \propto exp(\mathbf{d}_t\mathbf{W}_\beta\mathbf{h}_{ij}) \tag{4.7}$$

where $\mathbf{d_t}$ is the decoder hidden state at time step t, $\mathbf{W}_\alpha \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_\beta \in \mathbb{R}^{d\times d}$ are learned parameters, $\sum_i \alpha_{i,t} = 1$, and for all $i \in \{1, ..., I\}$ $\sum_j \beta_{i,j,t} = 1$.

• *Key-guided Hierarchical Attention.* This variant is motivated by the intuition that once an entity is chosen to be mentioned (thanks to $\alpha_{i,t}$), only the type of records is important to determine the content of the description. For example, when deciding to mention a player, all experts automatically report his score without consideration of its specific value. The attention scores are thus modeled by computing the $\beta_{ij,t}$ scores from Equation 4.7 solely on the embedding of the *key* rather than on the full record representation $\mathbf{h}_{ij}$:

$$\hat{\beta}_{ij,t} \propto exp(\mathbf{d}_t\mathbf{W}_{a_2}\mathbf{k}_{ij}) \tag{4.8}$$

4.2.2 Experiments

This model has been evaluated on the RotoWire dataset [Wiseman *et al.* 2017] using the BLEU metric [Papineni *et al.* 2002] and Information-extraction ones (RG, CS, CO) [Wiseman *et al.* 2017]. These last three metrics respectively estimate how well the system can generate text containing factual (i.e., correct) records, how well the generated document matches the gold document in terms of mentioned records, and how well the system orders the records discussed in the description.

We compare our hierarchical model against four systems:

- *Wiseman* [Wiseman *et al.* 2017] is a standard encoder-decoder system with copy mechanism.
- *Li* [Li & Wan 2018] is a standard encoder-decoder with a delayed copy mechanism: the text is first generated with placeholders, which are replaced by salient records extracted from the table by a pointer network.
- *Puduppully-plan* [Puduppully *et al.* 2019a] acts in two steps: a first standard encoder-decoder generates a plan, *i.e.* a list of salient records from the table; a second standard encoder-decoder generates text from this plan.

	BLEU	RG		CS			CO	Nb
		P%	#	P%	R%	F1		Params
Gold descriptions	100	96.11	17.31	100	100	100	100	
Wiseman	14.5	75.62	**16.83**	32.80	39.93	36.2	15.62	45M
Li	16.19	84.86	19.31	30.81	38.79	34.34	16.34	-
Pudupully-plan	16.5	87.47	34.28	34.18	51.22	41	18.58	35M
Puduppully-updt	16.2	**92.69**	30.11	38.64	48.51	43.01	**20.17**	23M
Flat	$16.7_{.2}$	76.62_{1}	$18.54_{.6}$	$31.67_{.7}$	42.9_{1}	$36.42_{.4}$	$14.64_{.3}$	14M
Hierarchical-kv	$17_{.3}$	89.04_{1}	$21.46_{.9}$	$38.57_{1.2}$	$51.50_{.9}$	$44.19_{.7}$	$18.70_{.7}$	14M
Hierarchical-k	$\mathbf{17.5}_{.3}$	$89.46_{1.4}$	$21.17_{1.4}$	$\mathbf{39.47}_{1.4}$	$\mathbf{51.64}_{1}$	$\mathbf{44.7}_{.6}$	$18.90_{.7}$	14M

Table 4.1: Evaluation on the RotoWire test set using relation generation (RG) count (#) and precision (P%), content selection (CS) precision (P%) and recall (R%), content ordering (CO), and BLEU. -: the number of parameters unavailable. For each proposed variant of our architecture, we report the mean score over ten runs, as well as the standard deviation in subscript.

- *Puduppully-updt* [Puduppully *et al.* 2019b]. It consists of a standard encoder-decoder, with an added module aimed at updating record representations during the generation process. At each decoding step, a gated recurrent network computes which records should be updated and what should be their new representation.

We also test the importance of the input structure by training different variants of the proposed architecture: i) *Flat*, where we feed the input sequentially to the encoder, losing all notion of hierarchy. As a consequence, the model uses standard attention. This variant is closest to *Wiseman*, with the exception that we use a Transformer to encode the input sequence instead of an RNN. ii) *Hierarchical-kv* is our full hierarchical model, with traditional hierarchical attention, *i.e.* where attention over records is computed on the full record encoding, as in Equation 4.7. iii) *Hierarchical-k* is our full hierarchical model, with key-guided hierarchical attention, *i.e.* where attention over records is computed only on the record key representations, as in Equation 4.8.

Our results on the RotoWire test set are summarized in Table 4.1. From a general point of view, we can see from Table 4.1 that our scenarios obtain significantly higher results in terms of BLEU over all models; our best model *Hierarchical-k* reaching 17.5 vs. 16.5 against the best baseline. We would like to draw attention to the number of parameters used by those architectures. We note that our scenarios rely on a lower number of parameters (14 million) compared to all baselines (ranging from 23 to 45 million). This outlines the effectiveness of the design of our model relying on a structured encoding, in contrast to other approaches that try to learn the structure of data/descriptions from a linearized encoding. Besides, our in-depth insights about our model are reported below.

Hierarchical encoding of entities is better than linearized inputs. Our hierarchical models achieve significantly better scores on most metrics when compared to the flat architecture *Wiseman* and our *Flat scenario*, reinforcing the crucial role of structure in data semantics and saliency. Results show that our *Flat* scenario obtains a significantly higher BLEU score than *Wiseman* (16.7 vs. 14.5) and generates fluent descriptions with accurate mentions (RG-P%) that are also included in the gold descriptions (CS-

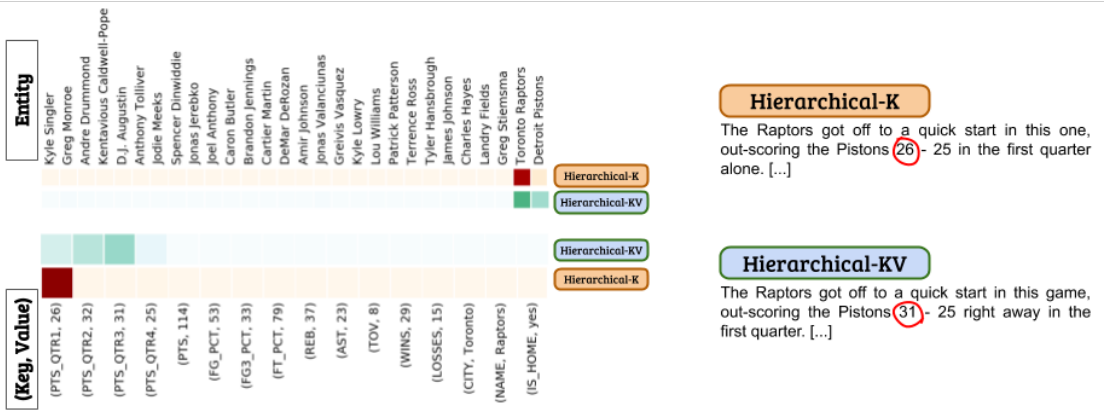


Figure 4.2: Right: Comparison of a generated sentence from *Hierarchical-k* and *Hierarchical-kv*. Left: Attention scores over entities (top) and over records inside the selected entity (bottom) for both variants, during the decoding of respectively 26 or 31 (circled in red).

R%). This suggests that introducing the Transformer architecture is a promising way to implicitly account for the data structure.

Hierarchical attention on high-level information of the structure is sufficient. The comparison between scenarios *Hierarchical-kv* and *Hierarchical-k* shows that omitting entirely the influence of the record values in the attention mechanism is more effective: this last variant performs slightly better in all metrics except CS-R%, reinforcing our intuition that focusing on the structure modeling is an important part of data encoding. To illustrate this intuition, we depict in Figure 4.2 attention scores (recall $\alpha_{i,t}$ and $\beta_{i,j,t}$ from Equations 4.7 and 4.8) for both variants *Hierarchical-kv* and *Hierarchical-k*. We particularly focus on the timestamp where the models should mention the number of points scored during the first quarter of the game. Scores of *Hierarchical-k* are sharp, with all of the weight on the correct record (PTS_QTR1, 26) whereas scores of *Hierarchical-kv* are more distributed over all PTS_QTR records, ultimately failing to retrieve the correct one.

Incorporating the structure into the encoder is more effective than in the decoder. Our hierarchical models outperform the two-step decoders of *Li* and *Puduppully-plan* on both BLEU and all qualitative metrics. For a reminder, these models impose a structure within the decoder through planning or templating intermediary steps. Interestingly, the baseline *Puduppully-plan* reaches 34.28 mentions on average, showing that incorporating modules dedicated to entity extraction leads to over-focusing on entities; contrasting with our models that learn to generate more balanced descriptions.

The way to encode the structure in the encoder matters. The comparison with *Puduppully-updt* shows that dynamically updating the encoding across the generation process can lead to better Content Ordering (CO) and RG-P%. However, this does not help with Content Selection (CS) since our best model *Hierarchical-k* obtains slightly better scores.

4.2.3 Conclusion

In this work we have proposed a hierarchical encoder for structured data, which 1) leverages the structure to form an efficient representation of its input; 2) has strong synergy with the hierarchical attention of its associated decoder. This results in an effective and more light-weight model[2]. Qualitative analyses on the RotoWire benchmark shows that our approach can still lead to erroneous facts or even hallucinations. This challenge is addressed in the next section to prevent inaccurate descriptions.

4.3 Handling hallucinations in data-to-text generation

As explained in Section 3.1.3, text generation models might lead to over-generation issues such as hallucinations. This might be because models are trained on non-aligned datasets in which, in the case of data-to-text generation, the textual description diverges from the structured data. It is thus critical to design models that generate faithful descriptions in accordance with the input data. Based on a literature review, we aim here at bridging two lines of work: 1) text generation models which integrate regularization into the loss to constrain the model by lack of control [Wang 2019, Liu *et al.* 2019b, Rebuffel *et al.* 2020b], and 2) controlled text generation models which enable choosing the defined features of generated texts [Filippova 2020]. Moreover, unlike previous CTG approaches [Li *et al.* 2016, Sennrich *et al.* 2016, Ficler & Goldberg 2017, Filippova 2020] which propose instance-level control factors, we propose an original approach [Rebuffel *et al.* 2022] in which the word-level information is integrated at all phases:

- we propose a **word-level labeling procedure**, which makes the correspondence between the input table and the text explicit, based on token co-occurrences and sentence structure through dependency parsing. This mitigates the failure of the strict word-matching procedure, while still producing relevant labels in complex settings.

- we introduce a **weighted multi-branch neural decoder**, guided by the proposed alignment labels acting as word-level control factors. During training, the model is able to distinguish between aligned and unaligned words and learns to generate accurate descriptions without being misled by un-factual reference information.

4.3.1 Model formalization

Word-level Alignment Labels. Our word-level alignment labels are driven by two intuitive constraints: (1) important words (names, adjectives, and numbers) should be labeled depending on their alignment with the data structure, and (2) words from the same statement should have the same label. We define a statement in the textual description as text spans expressing one single idea (obtained using dependency relations on the basis of part-of-speech).

We, therefore, estimate an alignment score between important words (i.e., nouns, adjectives, or verbs) and the data structure using occurrences and co-occurrences statistics. Then, given a statement, we normalize the score of composing words so that they obtain

[2]The code is available at https://github.com/KaijuML/data-to-text-hierarchical

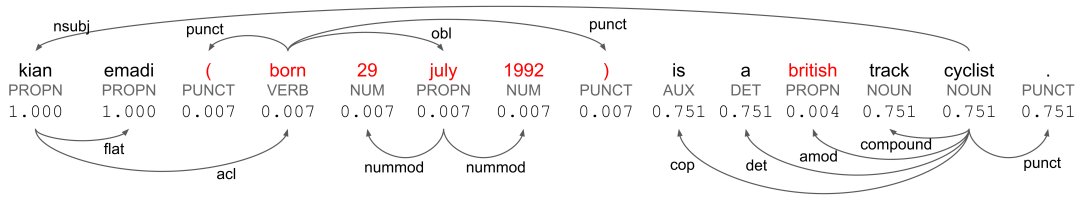


Figure 4.3: Word-level alignment labeling procedure. Every token is associated with its Part-of-Speech tag and its alignment score a_t. Words in red denote $a_t < \tau$, i.e., divergent words. The dependency parsing is represented by labeled arrows that flow from parents to children. Important words are *kian*, *emadi*, *29*, *july*, *1992*, *british*, *track*, and *cyclist*.

all the same score. This will ensure that, if we consider that a statement is not aligned with the structure data, we can remove it from the description without impacting text fluency. The *alignment score* a_t for a given token y_t and a data structure s is estimated as follows:

$$a_t \coloneqq norm(\max_{x_{ij} \in s} align(y_t, x_{ij}), \quad y) \tag{4.9}$$

where:

- the function *align()* estimates the alignment between important words y_t and a key-value pair x_{ij} from the input data s based on occurrences and co-occurrences statistics. If the word y_t appears in the key-value pair x_{ij}, $align(y_t, x_{ij})$ outputs 1; otherwise, the output is obtained scaling the number of occurrences $co_{y_t,x}$ between y_t and x through the dataset:

$$align(y_t, x) \coloneqq \begin{cases} 1 & \text{if} \quad y_t \in x \\ a \cdot (co_{y_t,x} - m)^2 & \text{if} \quad m \leq co_{y_t,x} \leq M \\ 0 & \text{if} \quad 0 \leq co_{y_t,x} \leq m \end{cases} \tag{4.10}$$

  where M is the maximum number of word co-occurrences in the dataset vocabulary and the row x, m is a threshold value, and $a \coloneqq \frac{1}{(M-m)^2}$.

- *norm()* is a normalization function based on the dependency structure of the description y constraining all words in a statement to be assigned to the same alignment score a_t). We first split the sentence y into statements $y_{t_i:t_{i+1}-1}$, via dependency parsing and its rule-based conversion to constituency trees. Given a word y_t associated with the score a_t and belonging to statement $y_{t_i:t_{i+1}-1}$, its normalized score corresponds to the average score of all important words in this statement:

$$norm(a_t, y) = \frac{1}{t_{i+1} - t_i} \sum_{j=t_i}^{t_{i+1}-1} a_j \tag{4.11}$$

This in-statement average depends on both the specific word and its context, leading to coherent hallucination scores which can be thresholded without affecting the syntactical sentence structure, as shown in Fig. 4.3. Words are colored in red if this score is lower than a threshold τ, denoting an alignment label equal to 0.

Multi-Branch Architecture. The proposed Multi-Branch Decoder (MBD) architecture aims at separating co-dependent factors during generation. We build upon the

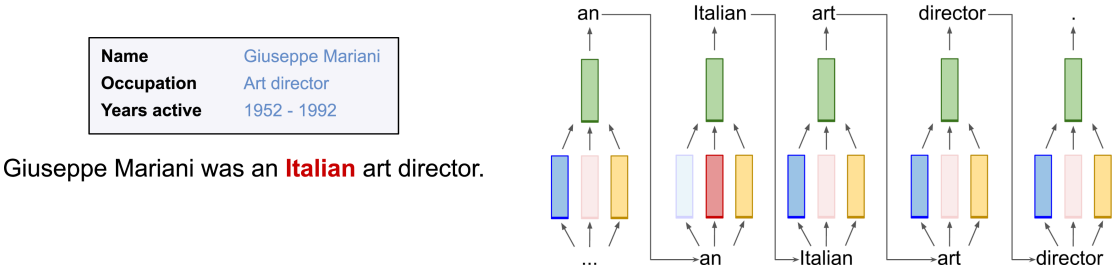


Figure 4.4: Our proposed decoder with three branches associated with content (in blue – left), hallucination (in red – middle), and fluency (in yellow – right). Semi-transparent branches are assigned the weight 0.

standard DTG architecture, an encoder-decoder with attention and copy mechanisms, which we modify by duplicating the decoder module into three distinct parallel modules. The decoding modules' actual architecture may vary, as we framed the MBD model from a high-level perspective. Therefore, all types of decoder can be used, such as Recurrent Neural Networks (RNNs) [Rumelhart *et al.* 1986], Transformers [Vaswani *et al.* 2017], and Convolutional Neural Networks [Gehring *et al.* 2017].

Our objective is to enrich the decoder to be able to tune the content/hallucination ratio during generation, aiming at enabling the generation of hallucination-free text when needed. Our key assumption is that the decoder's generation is conditioned by three co-dependent factors:

- *A content factor* constrains the generation to transcribe only the information included in the input;
- *An hallucinating factor* favors lexically richer and more diverse text, but may lead to hallucinations not grounded by the input;
- *A fluency factor*[3] conditions the generated sentences toward global syntactic correctness, regardless of the relevance.

Each control factor (i.e. content, hallucination, or fluency) is modeled via a single decoding module, also called a branch, whose output representation can be weighted according to its desired importance.

Our network has a single encoder and $F = 3$ distinct decoding RNNs, noted RNN^f respectively, one for each factor. During each decoding step, the embedding $\mathbf{y_{t-1}}$ previously decoded word is fed to all RNNs, and a final decoder state $\mathbf{d_t}$ is computed using a weighted sum of all the corresponding hidden states,

$$\mathbf{d_t^f} \coloneqq \text{RNN}^f(\mathbf{d_{t-1}^f}, [\mathbf{y_{t-1}}, \mathbf{c_t}]) \tag{4.12}$$

$$\mathbf{d_t} \coloneqq \sum_{f=1}^{F} \omega_t^f \mathbf{d_t^f} \tag{4.13}$$

[3] [Wiseman *et al.* 2018] showed that the explicit modeling of a fluency latent factor improves performance.

KEY	VALUE
name	ryan moore
spouse	nichole olson -lrb- m. 2011 -rrb-
children	tucker
college	unlv
yearpro	2005
tour	pga tour
prowins	4
pgawins	4
masters	t12 2015
usopen	t10 2009
open	t10 2009
pga	t9 2006
article_title	ryan moore -lrb- golfer -rrb-

Ref.: ryan david moore -lrb- born december 5 , 1982 -rrb- is an american professional golfer , currently playing on the pga tour .
PB&L: ryan david moore -lrb- born december 5 , 1982 -rrb- is an american professional golfer , currently playing on the pga tour .
Ours: ryan david moore -lrb- born december 5 , 1982 -rrb- is an american professional golfer , currently playing on the pga tour .

Figure 4.5: WikiBio instances' hallucinated words according to either our alignment scoring procedure or to the method proposed by [Perez-Beltrachini & Lapata 2018]. *PB&L* labels words incoherently and sometimes the whole reference text (as in the example). In comparison, our approach leads to a fluent breakdown of the sentences in hallucinated/factual statements.

where $\mathbf{d_t^f}$ and ω_t^f are resp. the hidden state and the weight of the f^{th} RNN at time t. Weights are used to constrain the decoder branches to the desired control factors ($\omega_t^0, \omega_t^1, \omega_t^2$ for the content, hallucination, and fluency factors resp.) and sum to one.

During training, the weights are dynamically set depending on the *alignment score* $a_t \in \{0, 1\}$ of the target token y_t. During inference, the weights of the decoder's branches are set manually by a user, according to the desired trade-off between information reliability, sentence diversity, and global fluency. Text generation is then controllable and consistent with the control factors.

Figure 4.4 illustrates a training step over the sentence "*Giuseppe Mariani was an Italian art director*", in which *Italian* is a divergent statement (i.e. is not supported by the source table). While decoding factual words, the weight associated with the content (resp. hallucination) branch is set to 0.5 (resp. 0) while during the decoding of *Italian*, the weight associated with the content (resp. hallucination) branch is set to 0 (resp. 0.5). Note that the weight associated with the fluency branch is always set to 0.5, as fluency does not depend on factualness.

4.3.2 Experiments

We evaluated the model on two representative large-size datasets: WikiBio and ToTTo. Both have been collected automatically and present a significant amount of table-text divergences for training (ToTTo being noisier). To evaluate the generated text, we use the BLEU [Papineni *et al.* 2002] and PARENT [Dhingra *et al.* 2019] metrics. We also consider: 1) the *hallucination rate* which computes the percentage of tokens labeled as hallucinations, 2) the average generated sentence length in the number of words, and 3)

Model	BLEU↑	PARENT↑			Halluc. rate↓	Mean sent. length	Flesch↓
		Precision	Recall	F-measure			
Gold	-	-	-	-	23.82%	19.20	**53.80%**
`stnd`	41.77%	79.75%	45.02%	55.28%	4.20%	13.80	58.90%
`stnd_filtered`	34.66%	**80.90%**	42.48%	53.27%	**0.74%**	12.00	62.10%
`hsmm`	35.17%	71.72%	39.84%	48.32%	7.98%	14.80	58.60%
$\texttt{hal}_{\texttt{WO}}$	36.50%	79.50%	40.50%	51.70%	-	-	-
`MBD`	41.56%	79.00%	**46.40%**	**56.16%**	1.43%	14.60	58.80%

Table 4.2: Comparison results on WikiBio. ↑ (resp. ↓) means higher (resp. lower) is better. "Gold" refers to the gold reference texts included in the dataset.

the classic readability Flesch index [Flesch 1962], which is based on words per sentence and syllables per word, and is still used as a standard metric [Kosmajac & Keselj 2019, Smeuninx *et al.* 2020, Stajner & Hulpus 2020, Stajner *et al.* 2020]. We also perform various human evaluations to obtain word-level hallucination labels of gold descriptions and an evaluation of generated descriptions.

Evaluation of alignment scores To assess the effectiveness of our alignment labels, we first compare the alignment labels to human judgment on 300 instances, and then explore their impact on a DTG task. As a baseline for comparison, we report performances of *PB&L* [Perez-Beltrachini & Lapata 2018], which is, to the best of our knowledge, the only work proposing such a fine-grained alignment labeling.

Our scoring procedure significantly improves over *PB&L*: the latter only achieves 46.9% accuracy and 29.7% F-measure, against 87.5% and 68.7% respectively for our proposed procedure. Figure 4.5 depicts an example of this phenomenon. Words labeled as hallucinated by each respective method are outlined in red, and we can see that the method proposed in [Perez-Beltrachini & Lapata 2018] over-labels words as hallucinated, leading to information loss. In contrast, our method is able to detect hallucinated statements inside a sentence, without incorrectly labeling the whole sentence as hallucinated.

Evaluation of our Multi-Branch Decoder To evaluate our Multi-Branch Decoder (*MBD*), we consider different baselines: i) *stnd* [See *et al.* 2017] and *stnd_filtered*, LSTM based encoder-decoder models with attention and copy mechanisms. *stnd_filtered* has been trained on a filtered version of the training set: tokens deemed hallucinated according to their hallucination scores, are removed from target sentences. ii) *hsmm* [Wiseman *et al.* 2018], an encoder-decoder model with a multi-branch decoder. The branches are not constrained by explicit control factors. iii) hal_{WO} [Filippova 2020], a *stnd*-like model trained by augmenting each source table with an additional instance-level attribute (*hallucination ratio*, *value*).

Table 4.2 shows the performances of our model and all baselines on the WikiBio dataset. The comparison of generated texts over different baselines is presented in Figure 4.6. The result analysis, combined with a human evaluation based on fluency, factualness, and coverage criteria, allows to outline the following main statements:

- **Reducing hallucinations is reached with success**, as highlighted by the hallucination rate (1.43% vs. 4.20% for a standard encoder-decoder and 10.10% for the best SOTA model on BLEU).

	name zack lee birth_name zack lee jowono nationality indonesian occupation actor , boxer , model birth_date 15 august 1984 birth_place liverpool , merseyside , england , uk years_active 2003 - present parents hendra and ayu jowono spouse nafa urbach (2007 - present) article_title zack lee
Gold	zack lee (born 15 august 1984) is an indonesian actor , model and boxer british descent .
stnd	zack lee jowono (born 15 august 1984) is an indonesian actor and model .
stnd_filtered	zack lee (born zack lee jowono ; 15 august 1984) is an indonesian actor .
hsmm	zack lee jowono (born 15 august 1984) is an indonesian actor who has appeared in tamil films .
MBD[.4, .1, .5]	zack lee (born zack lee jowono ; 15 august 1984) is an indonesian actor , boxer and model .
	name wayne r. dynes birth_date 23 august 1934 occupation professor , historian , and encyclopedist article_title wayne r. dynes
Gold	wayne r. dynes (born august 23 , 1934) is an american art historian , encyclopedist , and bibliographer .
stnd	wayne r. dynes (born august 23 , 1934) is an american historian and encyclopedist .
stnd_filtered	wayne r. dynes is a professor .
hsmm	wayne r. dynes (born august 23 , 1934) is an american historian , historian and encyclopedist .
hier	wayne r. dynes (born august 23 , 1934) is an american professor of history at the university of texas at austin .
MBD[.4, .1, .5]	wayne r. dynes (born august 23 , 1934) is an american professor , historian , and encyclopedist .

Figure 4.6: Qualitative examples of our model and baselines on the WikiBio test set. Note that: (1) *gold* references may contain divergences; (2) *stnd* and *hsmm* seem to perform well superficially, but often hallucinate; (3) *stnd_filtered* doesn't hallucinate but struggles with fluency; (4) *MBD* sticks to the fact contained by the table, in concise and fluent sentences.

- **Training standard generation model on a cleaned dataset is not sufficient** regarding PARENT and BLEU metrics (*MBD* > *stnd_filtered*, except for the Precision metric). Sentences are shorter and naive in terms of the Flesch readability index.

- **Multibranch decoder models help**, but adding a controlled factor is more effective (*hsmm* < *MDB*).

- **Word-level is better than sentence-level:** finer-grain annotation of hallucination (at the word-level for *MBD* vs. at the instance level for *hal_wo*) increases the recall in the text generation.

- **All factors are co-dependent:** an additional analysis (not presented here) of the impact of different weight combinations outlines, as expected, that changing weights in favor of the hallucination factor leads to decreases in both precision and recall (from 80.37% to 57.88% and 44.96% 4.82% respectively). It is interesting to

note that strictly constraining on content (i.e., removing the hallucination branch) yields sensibly more factual outputs, at the cost of constraining the model's generation creativity. The best combination of weights is [0.4 0.1 0.5], for content, hallucination, and fluency. It has more "freedom of speech" and sticks more faithfully to domain lingo (recall and BLEU), without compromising too much in terms of content.

A similar analysis on the noisy ToTTo dataset outlines that all models show significantly decreased scores. They struggle at generating syntactically correct sentences but, at the same time, they have still learned to leverage their copy mechanism and to stick to the input. Our proposed finer-grained approach proves helpful in this setting: sentences generated by *MBD* are more fluent and more factual. The multi-branch design enables the model to leverage the most of each training instance, leading to better performances overall. We acknowledge that despite over-performing other models, *MBD* obtains only 55.9% of *factual* sentences. The difficulty of current models to learn on very noisy and diverse datasets shows that there is still room for improvement in hallucination reduction in DTG.

4.3.3 Conclusion

We proposed a Multi-Branch decoder, able to leverage word-level alignment labels to produce factual and coherent outputs. Our proposed labeling procedure is more accurate than previous work, and outputs from our model are estimated, by automatic metrics and human judgment alike, more fluent, factual, and relevant. Experiments on ToTTo outline the narrow exposure to language of current models when used on very noisy datasets. The *naive* failure of our model on the noisy version of ToTTo could be attributed to its narrow exposure to language. We believe that large pre-trained language models, which have seen significantly more varied texts, may attenuate this problem.

4.4 Generating relevant answers in natural language in response to complex information needs

In contrast to our previous contributions [Rebuffel *et al.* 2020a, Rebuffel *et al.* 2022] which consider structured data as input, we tackle here unstructured data, namely documents, that need to be synthesized for an information retrieval task. The objective of the generation slightly slides from a faithfulness constraint to a relevance one, implying to identify information for a decision process objective [Park 1993, Florance & Marchionini 1995].

In this objective, we focus on complex search tasks which aim at generating a complete and structured answer on the basis of retrieved documents and an information need. This gives rise to the challenges of 1) considering all the retrieved documents both as pieces of evidence and sources to generate the answer leading to difficulties in discriminating between relevance and salience of the spans, and 2) building a multiple-span answer from these documents. We basically assume that the list of documents covers the different query facets. A naive approach would be to exploit text-to-text models [Devlin *et al.* 2019, Radford & Narasimhan 2018]. However, we believe that answering

multi-faceted queries would require the modeling of the structure prior to generating the answer's content [Culpepper *et al.* 2018].

With this in mind, we propose to leverage one category of works within the literature on data-to-text generation models which focuses on content planning [Puduppully *et al.* 2019a, Li & Wan 2018]. This technique, beyond encoding structured data (in our case a list of retrieved documents), allows to integrate the structure into the decoding part to produce a structured text. To do so, the encoder-decoder architecture is complemented by an intermediary step which determines *what to say* (called the *content selection/planning* step) [Puduppully *et al.* 2019a]. Based on the planning step, the model then defines how to say it (called the *surface realization* step). This intermediary step reinforces the factualness and the coverage of the generated text since 1) it organizes the data structure in a latent form to better fit with the generated output, and 2) on the reverse side, it provides a structure to the generated text based on the elements of the initial structured data.

Our contribution [Djeddal *et al.* 2022], therefore, bridges the gap between information retrieval and data-to-text generation to provide relevant natural language answers in response to complex queries. We, therefore, frame the answer generation task as a data-to-text task in which documents can be seen as entities and the list of the documents as a table. The objective is thus to generate a query-driven answer guided by the content of the document list.

4.4.1 Model formalization

Model overview. The designed model is driven by the intuition that the response should be surrounded by a plan to cover most of the query facets. Therefore, the decoding phase follows the principles proposed in [Puduppully *et al.* 2019a] and processes in two steps: decoding a plan aiming at structuring the answer and then generating an answer by leveraging both the generated plan and the context embedding. Figure 4.7 presents an example of a query from TREC Complex Answer Retrieval (CAR) dataset [Dietz *et al.* 2017] and the two variants of answers (*plain answers*, *structured answers*) generated by our proposed model.

Notations. We consider a document collection $\mathcal{D}$ and a set $Q \times A \times P$ of query-answer-plan triplets, where $q \in Q$ refer to queries, answers $a \in A$ to the final response in natural language provided to the user and plans $p \in P$ to the hierarchical structure of answers a. All documents d, queries q, and answers a are represented by lists of tokens. For modeling the structure of plans p, we use $p = \{h_1, ..., h_i, ..., h_{|p|}\}$ where h_i represents an item in the plan and is modeled as a heading (e.g., title, subtitles, etc.).

Given a query q and a document collection $\mathcal{D}$, our objective is to generate an answer a. To do so, we follow the "Retriever Generator" framework [Lewis *et al.* 2020, Nakatsuji & Okui 2020, Song *et al.* 2018] in which: 1) a ranking model $\mathcal{M}_{ret}$ retrieves a ranked list $\mathcal{D}_q$ of documents in response to query q, where $\mathcal{D}_q = \{d_q^1, \ldots, d_q^n\}$ and 2) a text generation model $\mathcal{M}_{gen}$ generates the answer a given the retrieved list $\mathcal{D}_q$ and query q. As outlined earlier, the challenges of our task mainly rely on aggregating information over the ranked list of documents and generating a structured answer in natural language. Thus, we use a pre-trained retrieval model $\mathcal{M}_{ret}$ and focus on the generation

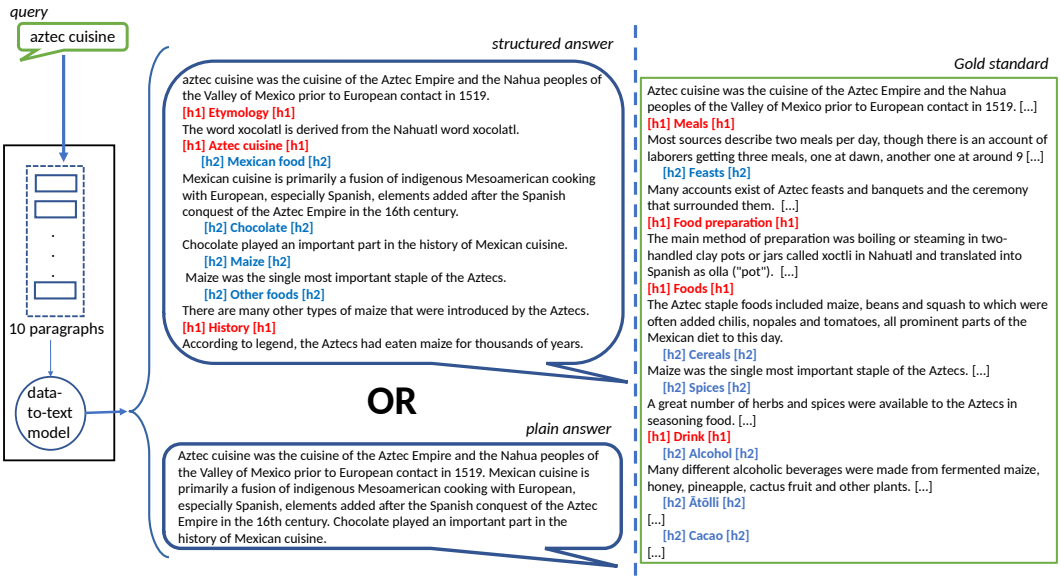


Figure 4.7: Example of a query from the CAR dataset [Dietz *et al.* 2017] and variants of outputs (structured or plain answers) obtained using a sequential DTT planning-based model.

model $\mathcal{M}_{gen}$. The latter exploits the DTT generation model based on content selection and planning [Puduppully *et al.* 2019a].

Decoding plans and answers with the generation model $\mathcal{M}_{gen}$.
To generate the intermediary plan p and the answer a, we rely on two successive encoder-decoders (based on T5 [Devlin *et al.* 2019] as the building-box model):

- The **planning encoder-decoder** encodes the list $\mathcal{D}_q$ of documents d_q and the query to guide the generation of plan p. The training of such a network is guided by the auto-regressive generation loss:

$$\mathcal{L}_{planning}(q,p) = P(p|q,\mathcal{D}_q) = \prod_{j=1}^{|p|}\prod_{k=1}^{|h_j|} P(h_{jk}|h_{j,<k},q,\mathcal{D}_q) \tag{4.14}$$

  where j and k point out respectively to the heading h_j and the k^{th} token h_{jk} in heading h_j. $h_{j,<k}$ corresponds to the token sequence in heading h_j before the k^{th} token.

  In practice, we use a pre-trained text-to-text model (i.e., T5) which encodes the following input:

$$[Query:]q[Documents:][Document:]d_1[Document:]d_2[Document:]d_3\ldots \tag{4.15}$$

  where $[Query:]$, $[Documents:]$ and $[Document:]$ are separator tokens, trained with loss defined in Equation 4.14.

- The **content generation encoder-decoder** encodes the generated plan p and the query q and the document list $\mathcal{D}_q$ to generate an answer a. The model input is a concatenation of associated embeddings for each component: for the plan,

we used the last layer of the planning encoder-decoder model; for the query q and the document list $\mathcal{D}_q$, we used the embeddings obtained from a pre-trained T5 model. The training of the content generation model is guided by the auto-regressive generation loss:

$$\mathcal{L}_{answer}(q,a,p) = P(a|q,p,\mathcal{D}_q) = \prod_{k=1}^{|a|} P(a_k|a_{<k},q,p,\mathcal{D}_q) \tag{4.16}$$

where a_k and $a_{<k}$ resp. express the k^{th} token in answer a and the token sequence of answer a before the k^{th} token.

- The final loss is a combination of both losses:

$$\mathcal{L} = \sum_{\{q,a,p\}\in Q\times A\times P} \mathcal{L}_{planning}(q,p) + \mathcal{L}_{answer}(q,a,p) \tag{4.17}$$

4.4.2 Experiments

Dataset and metrics. We selected the TREC CAR (Complex Answer Retrieval) 2017 corpus [Dietz *et al.* 2017]. This dataset includes: (1) queries - denoting complex search tasks with multiple facets, (2) plans - expressing the different expected facets, and (3) paragraphs extracted from English Wikipedia - corresponding to texts associated with plan sections. The TREC CAR task consists of retrieving the paragraphs associated with each plan section to build a structured answer combining both plan sections and paragraphs. We used these *structured answers* as the final objective of our generation model given the queries; and the plans as the structure prior. Due to the structure prior constraint, we removed from the training set answers without any plans.

To compare the models' abilities to generate *structured answers*, we also evaluate a new form of expected answer (*plain answers*) where the structure is not taken into account. For this aim, we built a new dataset upon the initial TREC CAR dataset but only considered the paragraphs (without plans). Thus, we obtain two versions of datasets (for *structured answers* and *plain answers*) which both follow the original split of the TREC CAR dataset[4] (see an example in Figure 4.7).

For our model, we implemented two versions: 1) **Planning-seq**: a sequential model where the planning module (Equation 3.13) and the content generation module (Equation 3.15) are trained separately. For this setting, the input embeddings of the content generation module are obtained using a pre-trained T5 model. 2) **Planning-e2e**: the version of our model in which the planning module and the content generation are trained in an end-to-end manner. We compare them with two baselines: 1) the **T5** model [Devlin *et al.* 2019] which is fine-tuned on each dataset (for *structured answers* and *plain answers*), and 2) **Ext**, an extractive method where we extract, for each sentence in the ground truth, a sentence in the input supporting documents that maximizes the F1 score of BERTScore.

We evaluate the quality of the generation using three well-known metrics: 1) the ROUGE-L mid metric (Rouge-P, Rouge-R, Rouge-F) [Lin 2004], 2) the BERTScore [Zhang *et al.* 2019a] (the F1 score is reported), and 3) the QuestEval [Scialom *et al.* 2021]. To evaluate the model's ability to generate structure (namely the plans), we use the METEOR score [Banerjee & Lavie 2005] to capture how well-ordered the output words are.

[4]The large train set for training, and the Y1 benchmark test set for testing.

			# tokens	Rouge-P	Rouge-R	Rouge-F	BERTScore	QuestEval
structured	answers	EXT	898.22	36.50	26.99	29.86	85.50	41.99
		T5	126.25	**76.19**	08.41	14.25	**84.95**	39.06
		Planning-seq	181.39	62.94	09.57	15.36	84.44	37.47
		Planning-e2e	203.48	63.4	**10.21**	**16.09**	84.91	**39.31**
plain	answers	EXT	885.35	34.35	26.73	28.99	86.30	42.34
		T5	110.62	**78.05**	09.24	15.48	85.51	39.89
		Planning-seq	163.58	65.73	10.34	16.27	84.29	38.46
		Planning-e2e	126.91	75.92	**10.34**	**17.05**	**85.67**	**40.78**

Table 4.3: Effectiveness of the answer generation. In bold are the highest metric value among the generation models (T5, Planning-seq, Planning-e2e).

Results Results are presented in Table 4.3, outlining the following statements:

- **Planning-based generation models are competitive regarding the T5 generation baseline.** Our models allow for generating longer answers (avg. 200 tokens), thus increasing the recall metric (Rouge-R). The smaller precision does not hinder the semantic content of the answer (see BERTScore and QuestEval values which are very close to the EXT metrics).
- **Generating structured texts is more difficult than plain texts, but intermediary plans are useful in both cases.** One can see the general trend towards higher metrics for all models in the *plain answers* setting compared to the *structured answers* setting over all models. In the *plain answers* setting, our models are more effective (with an advantage for Planning-e2e). Even if the *plain answers* setting does not expect plans in the final answer, our models generate an intermediary plan that guides the answer generation.
- **End-to-end learning is better than sequential learning.** Our end-to-end model seems more effective than the sequential one, suggesting the relevance of guiding the learning of the planning encoder-decoder by the answer generation task.
- **The analysis of the plans** in the final structured answers outlines that: 1) our plans are longer and more complex than the one generated by the T5 model (more tokens by plan section - up to 1.88 on average vs. 1.4 for the T5, more and deeper headings - up to 4/5 headings on average vs. 3 for the T5), 2) our plans generally cover more facets (higher recall), in the correct order (higher Meteor) with a better relevant semantics (higher BERTScore).

4.4.3 Conclusion

Traditionally, IR approaches solving complex information needs focus on leveraging multi-turn interactions to provide optimal rankings of candidate documents at each turn. In this work, we have suggested alternative retrieval models that do not rely on the interactive updating of queries and document rankings as answers. We suggest that data-to-text generation is an alternative way to generate both natural language and structured answers. Experimental evaluation of a planning-based DTT model using the TREC CAR dataset shows the potential of our intuition. The discussion on answer effectiveness (and the higher performance of our models regarding T5) suggests that there

is a balance to reach between raw text and plan generation and that the structure prior is however highly beneficial for generating a good answer.

4.5 Discussion and achievements

In this chapter, we presented our works dealing with faithful and relevant text generation in data-to-text generation and conversational search. From a general point of view, our research corroborates the literature review exposed in several surveys [Yu *et al.* 2020, Li *et al.* 2021b, Li *et al.* 2022, Zhang *et al.* 2022] in the sense that text-to-text models show great abilities to generate fluent and coherent sentences, but that it is more challenging to ensure faithfulness regarding input and relevance regarding world knowledge or a user's intent. Fine-tuning those models is often the first strategy used to fit with the task objective [Devlin *et al.* 2019, Radford *et al.* 2019, Raffel *et al.* 2020], but it might be under-effective [Yu *et al.* 2020, Li *et al.* 2021b, Li *et al.* 2022, Zhang *et al.* 2022], as also shown in our experiments. Indeed, we exhibit that our models including task peculiarities are more efficient than simple text-to-text models fine-tuned on the training dataset (e.g. the Flat transformer vs. our hierarchical model in Table 4.1, or also T5 vs. our Planning-e2e model in Table 4.3). More particularly, the main conclusions that we can draw from this line of research are directed toward the properties of encoder and decoder modules:

The way to encode the input is important: toward preserving the structure of data. We have demonstrated that simply concatenating the data input as done in most text-to-text generation models is under-effective for encoding complex data, hindering the faithfulness and relevance of the generated text. For instance, in the data-to-text application domain, the flat scenario in Table 4.1 obtains the lowest values for the RG, CS, and CO metrics, measuring how well the generated text includes elements from the input data. Indeed, linearizing each cell in structured data might suit when data describe a single element. However, when data concerns multiple entities or heterogeneous semantic information, it is necessary to better leverage the data structure, as we have done with our hierarchical encoder [Rebuffel *et al.* 2020a]. This need for structure in the encoding process is corroborated with other works addressing different data structures, such as graphs [Ribeiro *et al.* 2019, Ribeiro *et al.* 2020] in which local and global contexts of nodes matter in the encoding process or SQL queries [Xu *et al.* 2018b] which are transformed into directed graphs to preserve their structure. From a larger point of view, not limited to data-to-text generation, Li *et al.* 2022 reinforce this intuition in their overview of different strategies used in the literature to encode input data, ranging from hierarchical encoding [Li *et al.* 2021c, Gu *et al.* 2021a], inter-sentential semantics modeling [Liu & Lapata 2019, Zhang *et al.* 2019b] to structural encoding module [Ribeiro *et al.* 2021, Li *et al.* 2021a].

Not all the input data are relevant: forcing the encoder to identify what is relevant. We have also outlined that, even though we encode the structure of data, it might be interesting to identify what is relevant in the structure depending on the task objective. For instance, in our first contribution (Section 4.2), we have shown that considering our data-to-text generation task, our model is more effective when it

focuses on keys rather than key-value pairs, reflecting the need to identify first the fact related to an entity, instead of the value associated with the fact. It is worth noting that the strategy might be totally different for question-answering tasks on tables [Chen *et al.* 2021, Yin *et al.* 2020] in which the value might be necessary to map the semantics of the question with the table. To identify what is relevant, a promising approach relies on prompt-tuning [Wei *et al.* 2022, Sanh *et al.* 2022] (or prefix-tuning [Li & Liang 2021]) in which the input of large language models includes continuous token embeddings related to the task and concatenated to the input data (resp. key and value vectors at each attention layer, for each prepended token in the input). The objective of such a technique is twofold: 1) fine-tuning large language models in a lightweight strategy: the large language model is frozen, and only prompt or prefix vectors are learned, and 2) guiding the encoder in identifying what is relevant in the input data. Early experiments in [Li & Liang 2021] have shown that this strategy is well-adapted for data-to-text generation, for instance.

The decoder needs to be controlled to ensure faithfulness and relevance. First, guided by the statement that the training procedure of text generation models might lead to mimic divergences contained in training data and therefore generate inconsistent sentences [Elsahar *et al.* 2021], we have shown that generating descriptions both relevant and grounded in the data is not obvious. Handling hallucinations in data-to-text generation models might be different from text-to-text generation models due to the different natures of data in input and output. It seems that there is a clear balance between precision and recall metrics regarded the mentioned facts. For instance, in Table 4.2, the backbone model trained on a cleaned dataset (without divergences) is the best model to limit hallucination mentions, but at the cost of recall metrics, denoting an incomplete generation of facts. This trend has also been seen in various summarization tasks [Ji *et al.* 2022]. Another explanation for hallucinations might also come from large language models which have been pre-trained on several NLP tasks [Radford *et al.* 2019, Raffel *et al.* 2020, Sanh *et al.* 2022], and therefore include large knowledge. These models have demonstrated their effectiveness for all NLP tasks. However, this over-generation behavior is critical for some tasks, such as data-to-text generation in which we need very accurate reports of data [Wiseman *et al.* 2017]. Different strategies might be used to limit over-generation (including hallucinations) [Ji *et al.* 2022], including controlled generation [Lample *et al.* 2019, Filippova 2020] as we have proposed in [Rebuffel *et al.* 2022], reinforcement learning [Paulus *et al.* 2018, Rebuffel *et al.* 2020b], or planning-driven generation [Puduppully *et al.* 2019a, Puduppully *et al.* 2019b, Moryossef *et al.* 2019, Shao *et al.* 2019, Djeddal *et al.* 2022].

Planning-based strategies are promising in NLG, but also in other research fields. Considering the last strategy based on planning, we have shown in our third contribution (Section 4.4) the interesting properties of generating intermediate plans regarding the relevance of the generated text. By introducing a structure in the output to ensure coverage in terms of information, we believe that this strategy also allows to better focus on what is important in the input data and therefore reduces the generation of hallucinations. In the same mind, Shao *et al.* 2019 have been observed that, for other NLG tasks, such as advertising and recipe text generation, planing-based models obtain

the highest coverage metrics and limit the redundancy of information. To complement and make a parallel with another research field, it is worth noting that planning-based models have also shown interesting properties in robotics to optimize and control actions for a robot [Andreas *et al.* 2018, Sharma *et al.* 2022], reinforcing our intuition that the control of text generation is crucial for ensuring faithfulness and relevance.

However, our works are not without limitations. We have addressed the faithfulness and relevance in general, without a throughout error analysis to identify the topology of hallucinations. We discuss in what follows some possible improvements we plan for future work.

- **How to better introduce relevance signals in text generation?** In our third work dealing with conversational search, we focused only on the decoding part using a planning strategy and evaluated the impact of an intermediary plan on the final text generation. However, we do not have a clear overview of how the input documents have been considered: whether the pre-trained language model generates a text from its own knowledge or heavily relied on inputs, and whether the documents are all considered as relevant or not. We believe that we can enhance the faithfulness and relevance of the text generation by better forcing the network to rely on provided evidence sources, namely the relevant documents, for instance with copy-mechanisms [Gu *et al.* 2016] or prompt-tuning [Li & Liang 2021, Weizenbaum 1966] as discussed earlier.

- **How to constrain the text generation with the generated plan?** One drawback of the classic planning-based strategy proposed by [Puduppully *et al.* 2019a] and used in our work [Djeddal *et al.* 2022] is that there is no guarantee that the paragraph is semantically related to its associated headline (title/subtitle/...), and the sequence of headlines is coherent. In practice, the structured text (including a sequence of headlines and their associated paragraphs) is generated on the flow given a generated intermediary plan. It is however difficult for language models to align headlines by headlines the intermediary plans and the structured answer: intermediary plans are encoded as a whole and are fed as input to the final generator. We believe that the decoding process can be enhanced by leveraging variational sequential planning [Shao *et al.* 2019, Ye *et al.* 2020, Puduppully *et al.* 2022], conditioning 1) the text generation to a specific part of the plan, and 2) the headline generation to the text generated for all previously generated headlines.

Outcomes. I briefly describe my supervision activity regarding the topic:

- The data-to-text generation topic has been initiated in the team by Patrick Gallinari with whom I co-supervised the thesis of Clément Rebuffel (September 2018 - July 2022) addressing the issues of encoding the data structure and handling hallucinations.

- I co-supervised a master student (Hanane Djeddal) through a collaboration with IRIT (Lynda Tamine-Lechani and Karen Pinel-Sauvagnat) on the third contribution of using data-to-text generation models for complex search tasks.

- We (Lynda Tamine, Karen Pinel-Sauvagnat, and myself) have submitted an application for a CIFRE thesis with ECOVADIS to the ANRT. This thesis will be the continuation of the intern topic we have supervised (leveraging data-to-text generation for generating responses in an information retrieval setting).
- I also co-supervise a thesis, started in December 2022 with Vincent Guigue and Alexandre Allauzen in the context of the ANR PRCE ACDC - "Apprentissage Contrefactuel pour Data-to-text Contrôlé" (PI: Sylvain Lamprier).

You can find below a list of related international publications[5]:

- Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, Patrick Gallinari: A Hierarchical Model for Data-to-Text Generation. ECIR 2020: 65-80
  Code: https://github.com/KaijuML/data-to-text-hierarchical.
- Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, Patrick Gallinari: PARENTing via Model-Agnostic Reinforcement Learning to Correct Pathological Behaviors in Data-to-Text Generation. INLG 2020: 120-130
  Code: https://github.com/KaijuML/PARENTing-rl
- Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, Patrick Gallinari: Data-QuestEval: A Referenceless Metric for Data-to-Text Semantic Evaluation. EMNLP 2021: 8029-8036
  Code: https://github.com/KaijuML/QuestEval
- Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, Patrick Gallinari: Controlling hallucinations at word level in data-to-text generation. Data Min. Knowl. Discov. 36(1): 318-354 (2022)
  Code: https://github.com/KaijuML/dtt-multi-branch
- Hanane Djeddal, Thomas Gerald, Laure Soulier, Karen Pinel-Sauvagnat, Lynda Tamine: Does Structure Matter? Leveraging Data-to-Text Generation for Answering Complex Information Needs. ECIR 2022: 93-101
  Code: https://github.com/hanane-djeddal/Complex-Answer-Generation

[5]National publications are not mentioned since they are simple translations of international publications.

Chapter 5

Contextualizing information needs in conversational IR

Contents

In this chapter, we address the research challenge of contextualizing information needs in conversational search (RQ2) according to two strategies. We first focus on the understanding of information needs given a natural language conversation between a dialogue system and the user. Then, we target a more proactive setting aiming, from the dialogue system side, to clarify the user intent by interacting with him/her in natural language.

5.1 CoSPLADE: Contextualizing SPLADE for Conversational IR

The first step toward the understanding of natural language questions expressing complex information needs consists of modeling conversation turns and integrating them into a query reformulation or ranking model. We report here our participation to the TREC CAsT Track [Dalton *et al.* 2020a, Dalton *et al.* 2021] and the extended version presented in [Hai *et al.* 2023]. With this in mind, the TREC CAsT Track focuses on conversational retrieval sessions containing around 10 turns of exchange. Each turn corresponds to a query and its associated canonical answer[1] is provided as context for future queries. For each turn $n \leq N$, where N is the last turn of the conversation, we denote by q_n and a_n respectively the corresponding query and its canonical response. The context of a query q_n at turn n corresponds to all the previous queries and answers,

[1] Selected by the organizer as the most relevant answer of a baseline system.

Title: Steroid use in US sports **Description**: The history of steroid use in US sports.	
Turn	**Conversation Utterances**
1	What's the history of steroid use in sports in the US?
2	What were Ziegler's improvements?
3	Why are they banned?
4	Are there visible signs?
5	That sounds easy to spot. How do they get away with it?
6	What is the NFL policy?
7	Isn't that speed?
8	What is the difference between the two policies?
9	I heard it even affects card players. Didn't bridge also have a problem?
10	I know what bridge is. I heard there was a drug scandal recently.
11	Does the article have more about it?

Figure 5.1: Example of TREC CAsT conversation.

i.e. q_1, a_1, q_2, a_2, ..., q_{n-1}, a_{n-1}. The main objective of the *TREC CAsT* challenge is to retrieve, for each query q_n and its context (i.e., the conversation turns), the relevant passages d within a passage collection $\mathcal{D}$. An example of conversation is presented in Figure 5.1.

Most of the previous methods have focused on a multi-stage ranking approach relying on query reformulation with query expansion systems trained with the CANARD dataset [Zamani *et al.* 2022b], a critical intermediate step that might lead to a sub-optimal retrieval. Other approaches have tried to use a fully neural IR first-stage [Krasakis *et al.* 2022, Lin *et al.* 2021b], but are respectively designed as a zero-shot setting or as a full learning-to-rank based on a dataset with pseudo-labels.

In this contribution, we aim at bridging these two directions and propose a much lighter training process for the first-stage ranker, where we focus on queries and do not make use of any passage (and thus of a learning-to-rank training). It moreover sidesteps the problem of having to derive weak labels from the CANARD dataset. More particularly, we propose to leverage the sparse representation of queries and documents provided by the SPLADE model [Formal *et al.* 2022] with a new loss that optimizes first-stage ranker in lightweight training. Shortly, we require that the representation of the query matches that of the disambiguated query (i.e. the *gold query*). We then train a second-stage ranker (i.e. re-ranker). Leveraging the fact that our first-stage ranker outputs weights over the (BERT) vocabulary, we propose a simple mechanism that provides a conversational context to the re-ranker in the form of keywords selected by SPLADE.

5.1.1 Model formalization

5.1.1.1 Background: the SPLADE model

In the following, we present our first-stage ranker and second-stage re-ranker, along with their training procedure, both based, directly or indirectly, on the SPLADE (v2) model described in [Formal *et al.* 2022] (which is an extension of SPLADE [Formal *et al.* 2021]. SPLADE has shown results on par with dense approaches on in-domain collections while exhibiting stronger abilities to generalize in a zero-shot setting. It outputs a sparse rep-

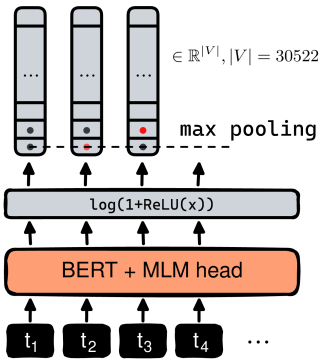


Figure 5.2: Overview of the Splade model [Formal *et al.* 2022].

resentation of a document or a query in the BERT vocabulary, which is key to our model during training and inference. The SPLADE model we use (v2) [Formal *et al.* 2022] includes a contextual encoding function, followed by some aggregation steps: ReLU, log saturation, and max pooling over each token in the text. Thanks to the FLOPS regularizer [Paria *et al.* 2020], it learns a sparse vector for queries and documents with only positive or zero components in the BERT vocabulary space $\mathbb{R}^{|V|}$. This allows to easily perform term weighting and query expansion for queries and documents as shown in Figure 5.2. In addition, the SPLADE model [Formal *et al.* 2022] scores a document using the dot product between the sparse representation of a document ($\hat{d}$) and of a query ($\hat{q}$): $s(\hat{q}, \hat{d}) = \hat{q} \cdot \hat{d}$. In this work, we use several sets of parameters for the same SPLADE architecture and distinguish each version by its parameters θ, and the corresponding model noted $SPLADE(\ldots;\theta)$.

5.1.1.2 First stage ranking.

The first-stage ranking performs a cosine similarity between query and document embeddings. Similarly to [Lin *et al.* 2021b], we suppose that the document representation has been sufficiently well-tuned on the standard ad-hoc IR task. The document embedding $\hat{d}$ is thus obtained using the pre-trained SPLADE model, i.e. $\hat{d} = SPLADE(\texttt{[CLS]}\ d;\ \theta_{SPLADE})$ where θ_{SPLADE} are the original SPLADE parameters obtained from HuggingFace[2]. These parameters are not fine-tuned during the training process. In the following, we present how to contextualize the query representation using the conversation history. Then, we detail the training loss of the extended SPLADE model aiming at reducing the gap between the representation of the gold query and the contextualized representation.

Query representation. Like state-of-the-art approaches for first-stage conversational ranking [Lin *et al.* 2021b, Krasakis *et al.* 2022], we contextualize the query with the previous ones. Going further, we propose to include the answers in the query representation process, which is easier to do thanks to our lightweight training. An overview of our approach is presented in Figure 5.3.

[2]The weights can be found at https://huggingface.co/naver/splade-cocondenser-ensembledistil

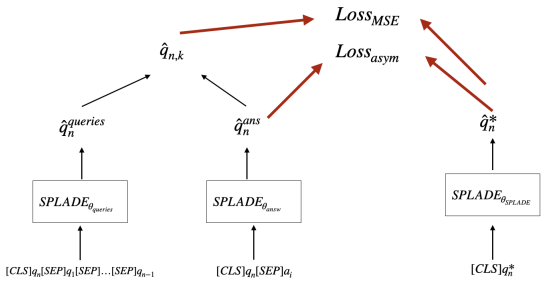


Figure 5.3: Learning query representations with CoSPLADE.

To leverage both contexts, we use a simple model where the contextual query representation at turn n, denoted by $\hat{q}_{n,k}$, is the combination of two representations, $\hat{q}_n^{queries}$ which encodes the current query in the context of all the previous queries, and $\hat{q}_{n,k}^{ans}$ which encodes the current query in the context of k the past answers[3]. Formally, the contextualized query representation $\hat{q}_{n,k}$ is:

$$\hat{q}_{n,k} = \hat{q}_n^{queries} + \hat{q}_{n,k}^{ans} \tag{5.1}$$

where we use two versions of SPLADE parameterized by $\theta_{queries}$ for the full query history and $\theta_{answers,k}$ for the answers.

Following [Lin *et al.* 2021b], we define $\hat{q}_n^{queries}$ to be the query representation produced by encoding the concatenation of the current query and all the previous ones:

$$\hat{q}_n^{queries} = SPLADE(\texttt{[CLS]}\ q_n\ \texttt{[SEP]}\ q_1\ \texttt{[SEP]}\ \ldots\ \texttt{[SEP]}\ q_{n-1}; \theta_{queries}) \tag{5.2}$$

To take into account the answers that the user had access to, we need to include them in the representation. Following prior work [Arabzadeh & Clarke 2020], we can consider various numbers of answers k, and in particular, we can either choose $k = 1$ (the last answer) or $k = n - 1$ (all the previous answers). Formally, the representation $\hat{q}_{n,k}^{ans}$ is computed as:

$$\hat{q}_{n,k}^{ans} = \frac{1}{k} \sum_{i=n-k}^{n-1} SPLADE(\texttt{[CLS]}\ q_n\ \texttt{[SEP]}\ a_i; \theta_{answers,k}) \tag{5.3}$$

Training The goal of the training is to reduce the difference between the gold query representation $\hat{q}_n^*$ and the representation $\hat{q}_{n,k}$ computed by our model.

To do so, we can leverage the gold query q_n^*, that is, a (hopefully) contextualized and unambiguous query. We can compute the representation $\hat{q}_n^*$ of this query by using the original SPLADE model, i.e.

$$\hat{q}_n^* = SPLADE(\texttt{[CLS]}\ q_n^*; \theta_{SPLADE}) \tag{5.4}$$

For example, for a query "How old is he?" the matching gold query could be "How old is Obama?". The representation of the latter given by SPLADE would be as follows:

$$[(\text{"}Obama\text{"}, 1.5), (\text{"}Barack\text{"}, 1.2), (\text{"}age\text{"}, 1.2), (\text{"}old\text{"}, 1.0), (\text{"}president\text{"}, 0.8), ...]$$

[3] In the experiments, we also explore an alternative model where answers and queries are considered at once. See results in Section 5.1.2.2

where the terms "Obama" and "Barack" clearly appear alongside other words related to the current query ("old" and the semantically related "age").

An obvious choice of a loss function is to match the predicted and gold representations using cosine loss (since the ranking is invariant when scaling the query). However, we observed in our preliminary experiments that models trained with the direct MSE do not capture well words from the context, especially for words from the answers. The reason is that the manually reformulated gold query usually only contains a few additional words from the previous turns that are directly implied by the last query. Other potentially useful words from the answers may not be included. This is a conservative expansion strategy which may not be the best example to follow by an automatic query rewriting process. We, therefore, design an asymmetric loss $Loss_{asym}()$ designed to encourage term expansion from past answers, but which avoids introducing noise by restricting the terms to those present in the gold query q_n^*. The final loss is a combination of the MSE and asymmetric losses:

$$Loss(\hat{q}_{n,k}, \hat{q}_n^*) = Loss_{MSE}(\hat{q}_{n,k}, \hat{q}_n^*) + Loss_{asym}(\hat{q}_{n,k}^{ans}, \hat{q}_n^*) \quad (5.5)$$

$$with\ Loss_{MSE}(\hat{q}_{n,k}, \hat{q}_n^*) = MSE(\hat{q}_{n,k}, \hat{q}_n^*) \quad (5.6)$$

$$and\ Loss_{asym}(\hat{q}_{n,k}^{ans}, \hat{q}_n^*) = \left(\max(\hat{q}_n^* - \hat{q}_{n,k}^{ans}, 0)\right)^2 \quad (5.7)$$

with $MSE()$ is the standard MSE loss, the maximum is component-wise. $Loss_{asym}$ pushes the $\hat{q}_{n,k}^{ans}$ representation to match the golden query representation $\hat{q}_n^*$ *if it can*, and $Loss_{MSE}$ pushes the queries-biased representation $\hat{q}_{n,k}$ to compensate *if not*. It thus puts a strong focus on extracting information from past answers. As a reminder, the parameters $\theta_{queries}$ and $\theta_{answers,k}$ used to obtain the different query representations are learned by optimizing the loss defined in Eq. (5.5).

5.1.1.3 Reranking

We perform reranking using a T5Mono [Nogueira *et al.* 2020] approach, where we enrich the raw query q_n with keywords identified by the first-stage ranker. Our motivation is that these words should capture the information needed to contextualize the raw query. The enriched query q_n^+ for conversational turn n is as follows:

$$q_n^+ = q_n.\ Context: q_1\ q_2\ \ldots\ q_{n-1}.\ Keywords: w_1, w_2, ..., w_K \quad (5.8)$$

where the w_i are the top-K most important words that we select by leveraging the first-stage ranker as follows. First, to reduce noise, we only consider words that appear either in any query q_i or in the associated answers a_i (for $i \leq n-1$). Second, we order words by using the maximum SPLADE weight over tokens that compose the word.[4]

We denote the T5 model fine-tuned for this input as $T5^+$. As in the original paper [Nogueira *et al.* 2020], the relevance score of a document d for the query q_n is the probability of generating the token "`true`" given a prompt $\text{pt}(q_n^+, d)$ = "`Query:` q_n^+. `Document:` d. `Relevant:`":

$$score(q_n^+, d; \theta) = \frac{p_{T5}(\texttt{true}|\text{pt}(q_n^+, d); \theta)}{p_{T5}(\texttt{true}|\text{pt}(q_n^+, d); \theta) + p_{T5}(\texttt{false}|\text{pt}(q_n^+, d); \theta)} \quad (5.9)$$

[4] To improve coherence, we chose to make keywords follow their order of appearance in the context, but did not vary this experimental setting.

where θ are the parameters of the T5Mono model.

Differently from the first-stage training, we fine-tune the ranker by aligning the scores of the documents, and not the weight of a query (which is obviously not possible with the T5 model). Here the "gold" score of a document is computed using the original T5Mono with the gold query q_n^*. The T5 model is initialized with weights made public by the original authors[5], denoted as θ_{T5}. More precisely, we finetune the pre-trained T5Mono model using the MSE-Margin loss [Hofstätter *et al.* 2020]. The loss function for the re-ranker (at conversation turn n, given documents d_1 and d_2, with d_1 more relevant than d_2) is calculated as follows:

$$\mathcal{L}_R = \left[(s(q_n^+, d_1; \theta_{T5+}) - s(q_n^+, d_2; \theta_{T5+})) - (s(q_n^*, d_1; \theta_{T5}) - s(q_n^*, d_2; \theta_{T5}))\right]^2$$

We optimize the θ_{T5+} parameters by keeping the original θ_{T5} to evaluate the score of gold queries.

5.1.2 Experimental evaluation

5.1.2.1 Protocol

To train our model, we used the CANARD corpus[6], a conversational dataset focusing on context-based query rewriting. More specifically, the CANARD dataset is a list of conversation histories, each being composed of a series of queries, short answers (human-written) and reformulated queries (contextualized). The training, development, and test sets include respectively 31.538, 3.418, and 5.571 contextual and reformulated queries.

To evaluate our model, we used the TREC CAsT 2020 and 2021 datasets which include respectively 25 and 26 information needs (topics) and a document collection composed of the MS MARCO dataset, an updated dump of Wikipedia from the KILT benchmark, and the Washington Post V4 collection. For each topic, a conversation is available, alternating questions and responses (manually selected passages from the collection, aka canonical answers). For each question (216 and 239 in total), the dataset provides its manually rewritten form as well as a set of about 20 relevant documents.

Metrics and baselines We used the official evaluation metrics considered in the TREC CAsT 2020 and 2021, namely nDCG@3, MRR, Recall@X, MAP@X, nDCG@X, where the cut-off is set to 1000 for the CAsT 2020 and 500 for the CAsT 2021. For each metric, we calculate the mean and variance of performance across the different queries in the dataset. With this in mind, we present below the different baselines and scenarios used to evaluate each component of our model.

- **First-stage ranking scenarios.** To evaluate the effectiveness of our first-stage ranking model (Section 5.1.1.2), we compare our approach CoSPLADE, based on the query representation of Eq. (5.1) with different variants (the document encoder is set to the original SPLADE encoder throughout our experiments): **SPLADE_rawQuery** (lower bound): SPLADE [Formal *et al.* 2021] using only the original and ambiguous user queries q_n; **SPLADE_goldQuery** (kind of upper bound):

[5] We used the Huggingface checkpoint `https://huggingface.co/castorini/monot5-base-msmarco`
[6] `https://sites.google.com/view/qanta/projects/canard`

SPLADE using the manually rewritten query q_n^*; **CQE** [Lin *et al.* 2021b], a state-of-the-art dense contextualized query representation learned using learning-to-rank on a dataset with pseudo-labels.

To model answers when representing the query using $\hat{q}_{n,k}^{ans}$, we design variants of our CoSPLADE model (first-stage ranking model learning queries representation with MSE and asymmetric losses) by using two historical ranges ("**All**" with $k = n-1$ answers and "**Last**" where we use only the last one, i.e. $k = 1$) and three types of answer inputs: **Answer** in which answers are the canonical answers; **Answer-Short** in which sentences are filtered as in the best performing TREC CAsT approach [Lin *et al.* 2021d]. This allows for consistent input length, at the expense of losing information; **Answer-Long** : as answers from CANARD are short (a few sentences extracted from Wikipedia – contrarily to CAsT ones), we expand them to reduce the discrepancy between training and inference. For each sentence, we find the Wikipedia passage it appears in (if it exists in ORConvQA [Qu *et al.* 2020]), and sample a short snippet of 3 adjacent sentences.

Finally, we also conducted ablation studies (on the best of the above variants) by modifying either the way to use the historical context or the training loss: **flat-Context** a one-encoder version of our SPLADE approach in which we concatenate all information of the context to apply SPLADE to obtain a single representation of the query (instead of two representations $\hat{q}_n^{queries}$ and $\hat{q}_{n,k}^{ans}$ as in Equations 5.2 and 5.3) trained using a MSE loss function (Equation 5.6) since there is no more two representations. **MSE** the version of our SPLADE approach trained with the MSE loss (Equation 5.6) instead of the proposed one (Equation 5.5); **cosine** the version of our SPLADE approach trained with a cosine loss instead of the proposed loss (Equation 5.5).

• **Second-stage ranking scenarios.** We have compared our model with several baselines (variants of a T5Mono ranker and of our model). Please note that we will not present the results of all these baselines but will directly present the final result of our model with respect to the score obtained by TREC participants.

5.1.2.2 First-stage ranking effectiveness

In our experiments, we focus on the first-stage ranking component of our CoSPLADE model. Results of the different baselines and scenarios on the TREC CAsT 2021 dataset are provided in Table 5.1[7]

In general, one can see that all variants of our approach (CoSPLADE_* models) outperform the scenario applying the initial version of SPLADE on raw and, more importantly, gold queries. This is very encouraging since this latter scenario might be considered as an oracle, i.e. the query is manually disambiguated. Finally, we improve the results over CQE [Lin *et al.* 2021b] for all the metrics – showing that our simple learning mechanism, combined with SPLADE, allows for achieving SOTA performance. More specifically, we can outline the following statements.

Leveraging queries and answers history better contextualizes the current query. The results of the flatContext scenario with respect to the SPLADE_goldQuery allow for comparing the impact of evidence sources related to the conversation since

[7]Similar trends are observed on CAsT 2020, but are not reported.

	Recall@500	MAP@500	MRR	nDCG@500	nDCG@3
Baselines					
SPLADE_rawQuery	30.8±2.7	5.5±0.9	21.3±2.9	17.8±1.8	13.1±2.1
SPLADE_goldQuery	68.8±2.0	16.1±1.2	55.5±3.3	42.8±1.7	38.3±2.8
CQE [Lin *et al.* 2021c]	79.1	28.9	60.3	55.7	43.8
Effect of answer processing: CoSPLADE_...					
AllAnswers	79.5±2.2	28.8±1.7	61.7±3.1	55.3±2.0	46.5±2.9
AllAnswers-short	72.8±2.6	25.7±1.9	54.4±3.3	49.5±2.3	40.1±3.0
AllAnswers-long	80.4±2.1	29.3±1.8	62.0±3.2	55.6±2.1	**48.9±3.0**
LastAnswer	83.4±2.0	31.2±1.8	61.8±3.1	58.1±2.0	47.4±3.0
LastAnswer-short	79.2±2.2	28.1±1.8	61.4±3.3	54.3±2.1	46.4±3.0
LastAnswer-long	**85.2±1.8**	**32.0±1.7**	**64.3±03.0**	**59.4±1.9**	48.6±3.0
CoSPLADE_LastAnswer-long variants					
flatContext	77.0±2.0	26.0±2.0	55.0±3.0	52.0±2.0	42.0±3.0
MSE loss	70.9±2.4	21.6±1.7	48.7±3.4	45.2±2.3	39.6±3.1
cosine loss	70.4±2.5	22.6±1.7	52.5±3.3	46.9±2.2	39.0±3.0

Table 5.1: Effectiveness of different scenarios of our first-stage ranking model on the TREC CAsT 2021.

they both use the same architecture (SPLADE). We can observe the usefulness of context to better understand the information need.

More detailed answers perform better. Since answers are more verbose than questions, including them is more complex, and we need to study the different possibilities (CoSPLADE_AllAnswers* and CoSPLADE_LastAnswer*). One can see that: 1) trimming answers (*-short) into a few keywords is less effective than considering canonical answers, but 2) it might be somehow effective when combined with the associated Wikipedia passage (*-long). Moreover, it seems more effective to consider only the last answer rather than the whole set of answers in the conversation history. Taking all together, these observations highlight the importance of the way to incorporate information from answers into the reformulation process.

Dual query representation with asymmetric loss leverages sparse query representations. The results of the flatContext scenario show that considering at once past queries and answers perform better (compared to the MSE loss scenario which is directly comparable). However, if we separate the representations *and* use an asymmetric loss function (AllAnswers* and LastAnswer* lines in Table 5.1), the conclusion changes. Moreover, the comparison of our best scenario CoSPLADE_LastAnswer-long with a similar scenario trained by simply using MSE or cosine losses reveals the effectiveness of our asymmetric MSE (Equation 5.7). Remember that this asymmetric loss encourages the consideration of previous answers in the query encoding. This reinforces our intuition that the conversation context, and particularly verbose answers, is important for the conversational search task. It also reveals that the context should be included at different levels in the architecture (input and loss).

TREC CAsT 2020	Recall@1000	MAP@1000	MRR	nDCG@1000	nDCG@3
TREC Participant (best)	63.3	30.2	59.3	52.6	45.8
TREC Participant (median)	52.1	15.1	42.2	36.4	30.4
TREC Participant (low)	27.9	1.0	5.9	11.1	2.2
CoSPLADE	82.4±2.0	26.9±1.5	58.1±2.9	54.2±1.8	44.0±2.7
TREC CAsT 2021	Recall@500	MAP@500	MRR	nDCG@500	nDCG@3
TREC Participants 1 (best)	85.0	37.6	67.9	63.6	52.6
TREC Participants 2 (median)	36.4	17.6	53.4	33.6	37.7
TREC Participants 3 (low)	58.9	7.6	27.0	31.4	15.4
CoSPLADE	84.9±1.7	35.5±1.8	69.8±3	62.2±1.9	54.4±2.9

Table 5.2: TREC CAsT 2020 and 2021 performances regarding participants.

5.1.2.3 Effectiveness compared to TREC CAsT participants

We finally compare our approach (first-stage + second-stage rankings) with TREC CAsT participants for the 2020 and 2021 evaluation campaigns. For each evaluation campaign, we report in Table 5.2 the best, the median, and the lowest TREC CAsT participants according to the nDCG@3 metric from the two overviews [Dalton *et al.* 2020a, Dalton *et al.* 2021]. For both years, we can see that we obtain effectiveness metrics that are very close or higher than the ones reached by the best participants. Indeed, CoSPLADE surpasses the best TREC participant for the 2020 evaluation campaign regarding Recall@1000 and nDCG@1000. For 2021, our model obtains better results than the best one for the MRR and nDCG@3 metrics. For both years, the best participant is the h2oloo team [Lin *et al.* 2021d, Dalton *et al.* 2021] which uses query reformulation techniques, either using AllenAI or T5. Our results suggest that our approach leveraging the SPLADE model trained using a ranking loss and fine-tuned on the conversation context using a query-driven loss allows combining the benefit of query expansion and document ranking in a single model that eventually helps the final reranking step. In other words, simply rewriting the query without performing a joint learning of document ranking can hinder the overall performance of the search task.

We also outline that our CoSPLADE model based on sparse representation obtains better results than the dense retrieval model T5Mono applied on queries reformulated with a T5 model (e.g., 84.9 vs. 80.4, respectively, for the Recall@500 for TREC CAsT 2021 - results not presented in this manuscript). This reinforces our intuition that sparse retrieval models, although being more sensitive to information loss because of the use of sparse representation to focus on terms, are well adapted to contextualize information needs.

5.1.3 Conclusion

In this contribution, we have shown how a sparse retrieval neural IR model, namely SPLADE [Formal *et al.* 2022], could be leveraged together with a lightweight learning process to obtain a state-of-the-art first-stage ranker. We further showed that this first-stage ranker could be used to provide context to the second-stage ranker, leading to results comparable with the best-performing systems. However, this setting is limited in the sense that it considers a passive IR system, i.e. simply performing ad-hoc IR. Current approaches are more willing to investigate mixed-initiative [Aliannejadi *et al.* 2021], en-

gaging the system in proactive interactions. Indeed, the TREC CAsT Track evolves into a mixed-initiative-oriented Track called IKAT[8]. This paradigm is addressed by another contribution presented in the following section.

5.2 User simulation for query clarification

In this contribution [Erbacher *et al.* 2022], we focus on query clarification which consists of a mixed-initiative between users and conversational search systems to solve an IR task. The objective of the IR system is thus to propose to the user a clarification of his/her information need and to interact with him/her to better understand his/her intent.

Unlike previous query clarification work based on single-turn interactions [Rao 2017, Rao & Daumé III 2018, Zamani *et al.* 2020a, Sekulic *et al.* 2021a, Sekulic *et al.* 2021b, Tavakoli *et al.* 2022] or simulated session logs in which successive actions are independent, except for the common interest towards the global topic [Aliannejadi *et al.* 2019], we propose here to build a fully simulated query clarification framework allowing multi-turn interactions between IR and user agents. Following [Aliannejadi *et al.* 2019], the IR agent identifies candidate queries and ranks them in the context of the user-system interactions to clarify the initial query issued by the user (agent). We target simple information needs, leaving multi-faceted information needs for future work since they might impact the modeling of the query ranking function. Our framework can be seen as a proof-of-concept for future work willing to integrate sequential models (namely reinforcement learning models) for question clarification. It is worth noting that large language models relying on attention mechanisms (transformers) are not yet well suited to handle sequential interactions and long-term planning, as current models are hardly trainable with current reinforcement learning algorithms [Chen *et al.* 2020b]. Thus, all agent components in our framework are based on continuous embeddings and simple models.

5.2.1 Question Clarification Simulation Framework

5.2.1.1 Overview and Research Hypotheses

Our query clarification simulation framework is inspired by [Aliannejadi *et al.* 2019], but provides the possibility of leveraging user and system agents' interactions sequentially. More particularly, our framework is detailed in Algorithm 1 and illustrated in Figure 5.4. The design of this evaluation framework is guided by some choices/hypotheses.

- First, following [Aliannejadi *et al.* 2019], we consider a fixed set of candidate queries $Q = \{q_1, q_2, ..., q_m\}$ constituting the reformulation of the initial query q_0. All the interactions are leveraged to improve step by step the ranking of this candidate query set so that, at the end of the session, the final query used for retrieving documents is a good clarification of its initial one. Obviously, this means that the set of candidate queries includes a large variety of queries which, for some of them, improve the search performance.

- Second, following [Zamani *et al.* 2020a], we propose to model question clarification as a possible option between two reformulated queries. In other words, expressed

[8] https://www.trecikat.com/

Algorithm 1 Our simulation framework for query clarification

- A) The user issues an initial query q_0 associated to her/his information need i.
- B) The IR system generates a set $Q = \{q_1, q_2, ..., q_m\}$ of candidate queries which might express different query reformulations or diversified queries to better explore the information need i.
- C) The IR system selects N queries to display to the user. To do so, we propose to follow [Aliannejadi *et al.* 2019] and design a model ranking the candidate query set Q to identify the top N queries.
- D) The user selects one of the N queries, enabling to extract positive and negative feedback, resp. noted (q^+, q^-).
- Steps C) and D) can be repeated several times to model multi-turn interactions. The query set ranking function (step C) integrates the user's sequential feedback (q^+, q^-) to improve the query ranking along with the interaction simulation.
- E) After T turns, the IR system considers the best-ranked query as the optimal query reformulation and runs a ranking model to retrieve documents.

in natural language, the IR system agent would ask the user agent the following question: "Which reformulated query do you prefer? A or B". This implies that the user is willing to judge queries A or B regarding its information need.

- Third, guided by the motivation to propose a framework for future work on sequential models, we consider here that each agent component is modeled at the embedding level. Indeed, leveraging large language models for generating/ranking questions is very effective, but integrating them into reinforcement learning models is still challenging (one main reason being the computational cost). This means that we processed *a priori* all queries and documents to represent them using text embeddings. This processing is done offline, alleviating the sequential modeling of the text encoding.

In what follows, we present the different components behind the IR system and user agents.

5.2.1.2 The IR System Agent

The IR agent has three objectives in our framework: 1) generating the set of candidate reformulated queries willing to be presented to the user, 2) ranking this set to identify the most relevant queries according to the interaction history, 3) ranking documents using the best-ranked query (ending the interactive session).

Generation of the candidate reformulated query set. The objective here is to instantiate various and diverse reformulations covering a wide range of relevant topics for the initial query q_0. Different techniques might be used, leveraging large language models [Nogueira *et al.* 2019b, Raffel *et al.* 2020, Rao & III 2019], query diversification [Cai *et al.* 2016, MacAvaney *et al.* 2021, Ye *et al.* 2021] or query expansion [Pal *et al.* 2013]. We propose here to use the T5 model [Raffel *et al.* 2020] which is designed to translate token sequences into other token sequences. It has already been used

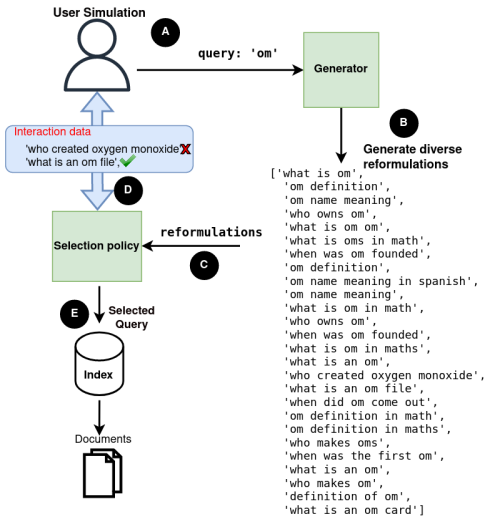


Figure 5.4: Query clarification simulation framework

for query reformulation tasks [Chen *et al.* 2020b, Raffel *et al.* 2020, Lin *et al.* 2020a]. On the top of that model, the generation process is driven by the diversity beam mechanism [Vijayakumar *et al.* 2016] which aims at generating a set Q of diversified query reformulation, $Q = \{q_1, q_2, ..., q_m\}$.

Ranking of queries based on the interaction history. The role of the selection policy is to select queries used to interact with the user agent. Following [Aliannejadi *et al.* 2019] which proposes to rank queries according to both performance criteria and the interaction context, we compute a pairwise score between two candidate queries q_i and q_j given the context, i.e., the initial query q_0 and the additional information provided by interaction $feedback_{t-1}, \ldots, feedback_1$ with the user. Formally, the ranking model relies on the probability that a query q_i obtains better IR performances (y_i) than query q_j (retrieval performance y_j) given the initial query q and the feedback obtained on queries displayed at previous utterances:

$$P(y_i > y_j | q_0, q_i, q_j, feedback_{t-1}, ..., feedback_1) \tag{5.10}$$

In practice, the model architecture is a siamese network that estimates the score y_i of a query q_i given the context and trained using a Lambda loss [Wang *et al.* 2018]. Each query score is computed as follows:

$$y_i = RNN_{score}(\mathbf{q_0}, \mathbf{q_i}, \mathbf{feedback_{t-1}}, \ldots, \mathbf{feedback_1}; \theta_{score}) \tag{5.11}$$

$$with \quad \mathbf{feedback_t} = RNN_{feedb}(cos(\mathbf{q}^+), sin(\mathbf{q}^-); \theta_{feedb}) \tag{5.12}$$

where RNN_{score} and $RNN_{feedb}()$ are two different recurrent neural networks with their own parameters θ_{score} and θ_{feedb}. $\mathbf{q_0}$ and $\mathbf{q_i}$ are embeddings of queries q_0 and q_i. $\mathbf{feedback_t}$ is the embedding of the user's feedback $feedback_t$, corresponding to the action of selecting or not the queries displayed at interaction turn t. We note q^+ and q^- those selected or non-selected queries, and $\mathbf{q}^+$ and $\mathbf{q}^-$ their associated embeddings. To capture the positive and negative feedback, we encode queries differently using the *cosine* and the *sine* functions, respectively.

At inference, queries are ranked according to their score estimated using Equation 5.11.

Final ranking of documents Documents are retrieved with the top-ranked query using a Dense Retriever model [Hofstätter *et al.* 2021].

5.2.1.3 The User Agent

After issuing the initial query q_0, the user agent interacts with the IR system agent to refine her/his information need. With this in mind, we hypothesize that the user is greedy toward her/his intent and fully cooperative. Greedy means that the user always selects the query which is the most similar to the overall intent. Despite being unrealistic, we ignore the click bias problem for the clarification questions presented in [Zamani *et al.* 2020a, Zamani *et al.* 2020b] which relies on position, presentation, and trust dimensions. Other choices for user simulation could be done, as experimented in [Câmara *et al.* 2022], but we let these variations for future work.

In practice, let $\mathbf{d}$ be the vector representing a user intent, q_i and q_j the clarification queries presented to the user agent. The user agent selects the best query (noted q^+ for highlighting positive feedback from the user) according to a similarity metric (in our case, the dot product) between the representation of the proposed queries q_i and q_j and intent d:

$$q^+ = argmax_{q_i}(\langle \mathbf{q_i}, \mathbf{d} \rangle) \tag{5.13}$$

5.2.2 Experimental evaluation

5.2.2.1 Protocol

Evaluating our simulation framework consists in measuring the effectiveness of the final ranking after T clarification interactions. Since the user behavior is greedy and follows a simple behavior dependent on the query selection process, the effectiveness results mainly denote the quality of this query ranking component. Other components (candidate set generation and final document ranking) do not depend on the interaction feedback, so we mainly focus on understanding whether the selection policy integrates users' feedback and takes good decisions to select the N clarification questions.

We carry out our experiments on the MS Marco 2020 passages dataset [Nguyen *et al.* 2016a] which regroups 8.8M passages and more than 500K query-passage relevance pairs. We evaluate our model on 2 sets: the small test set (43 queries) and a subset of the dev set (1000 queries sampled from 59 000). One motivation to consider these two datasets is their difficulty level: in the dev set, only one passage per query is labeled relevant in the ground truth, while several passages are considered as relevant in the test set.

Baselines and Scenarios To evaluate the effectiveness of our selection policy component, we compare with:

- Non-interactive settings. We measure the ranking effectiveness of the user's initial user query (noted **User Query**) and the **Best Reformulation** in the candidate query set - which can be seen as an oracle.
- **Naive interactive selection**: At each step, we select the 2 top ranked queries from

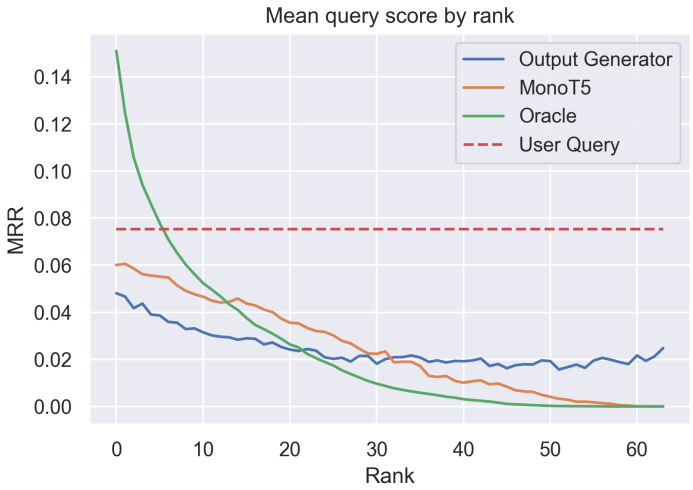


Figure 5.5: Effectiveness score of query reformulation by rank. The order of the query is determined either by the diversity beam search generation process (Output Generator), the query IR performances predicted by a pre-trained MonoT5 model (MonoT5), and the optimal ranking of queries according to their performances obtained by a Dense retrieval model [Hofstätter *et al.* 2021].

the current query rank and then remove the query which has not been selected by the user agent. The re-ranking of the candidate query set is only carried out once, at the beginning of the session, and the size of this list decreases with the interaction number.

To instantiate the selection policy after each interaction-driven query ranking step (step C in Figure 5.4), we consider these scenarios:

(1) **Interact. + Random Sample**: we sample 2 queries from the ranked candidate query set to constitute the interaction pair.

(2) **Interact. + Top 2**: we select the top 2 query reformulations at each turn.

(3) **Interact. + random sample@5**: we randomly select 2 queries among the top 5 query reformulations at each turn.

(4) **Interact. + Kmeans selection**: At each turn, queries in the candidate set are clustered in 2 groups using K-means. Queries from each cluster are ranked by the model. The best-ranked query within each cluster is selected. The cluster of the query not selected by the user is removed for the next turn from the set of candidate queries. This strategy corresponds to a refinement strategy as suggested in [Mustar *et al.* 2022], removing a group of semantically similar queries that have not been chosen by the user and going deeper into the other cluster.

5.2.2.2 Preliminary analysis: measuring the potential of ranking the query set

Our model introduces user-system interactions through query clarification to identify its information need and therefore enhance the retrieval process. To do so, a predefined query set is generated, assuming to cover a large diversity of information needs related to the initial query, and we propose to re-rank this query set to identify the most relevant

		No interaction	1	2	3	4	5
Best Reformulation (oracle)	MRR@10	0.872	-	-	-	-	-
User Query	MRR@10	**0.455**	-	-	-	-	-
Naive selection	MRR@10	0.213	0.327	0.359	0.403	0.419	0.427
Interact. + random sample	MRR@10	0.403	0.478	**0.481**	**0.490**	0.481	0.501
Interact. + Top 2	MRR@10	0.403	0.474	0.469	**0.490**	0.478	0.501
Interact. + random sample@5	MRR@10	0.403	0.473	0.467	**0.490**	0.479	0.501
Interact. + Kmean	MRR@10	0.403	**0.523**	0.465	0.469	**0.486**	**0.551**

Table 5.3: Effectiveness results on the Test set of MS Marco passage 2020 (43 queries - multiple relevant documents per query)

queries according to the initial need and the interactions. To test our hypothesis that it is possible to automatically identify the most relevant queries within a predefined query set without the supervision of relevant documents, we perform here a preliminary analysis (Figure 5.5) to quantify the potential retrieval performance gain of the candidate query set when they are ranked according to different criteria. More particularly, given the candidate query set generated by the T5 model (first paragraph in section 5.2.1.2), we compare the retrieval performance using the Mean Marginal Rank metric) according to different ordering within the query set: 1) the initial order provided by the Diversity Beam Search (called **Output Generator**), 2) the **Oracle** order in which queries are ranked in decreasing order according to their performance obtained through a Dense retrieval model [Hofstätter *et al.* 2021] regarding the Mean Marginal Rank metric, 3) the **MonoT5** order in which queries are ranked according to the performance score predicted by the pre-trained MonoT5 retrieval model [Pradeep *et al.* 2021].

We can see that predicting the query performance with MonoT5 allows to improve the performance for the top k queries regarding the query order provided by the Output Generator. This hurts the end of the list, but it is not critical in our case, since we consider the selection policy regarding the top query list. Moreover, one can notice that, although performance is increased, there is still a gap between the curve of the MonoT5 ranked list and the Oracle curve (order defined according to the real performance of queries using a Dense retriever [Hofstätter *et al.* 2021]). Our intuition is that leveraging users' interactions will lower this gap, which leads to the evaluation we performed in what follows.

5.2.3 Measuring the retrieval effectiveness after multi-turn query clarification

Second, we analyze the performance of the query ranker at different interaction turns using MRR@10. Tables 5.3 and 5.4 resp. show the results on the MS Marco passage 2020 test set and dev set. From a general point of view, we can see that performance metrics are lower for the dev set (Table 5.4) than for the test set (Table 5.3). This can be explained by the task difficulty, which is higher for the dev set in which only one document per query is assessed as relevant. By comparing all baselines and scenarios, we can outline the following trends.

- The first candidate query ranking within our interactive models (No interaction columns) provides lower performance than non-interactive baselines. For instance,

		No interaction	1	2	3	4	5
Best Reformulation (oracle)	MRR@10	0.411	-	-	-	-	-
User Query	MRR@10	**0.209**	-	-	-	-	-
Naive selection	MRR@10	0.122	0.151	0.165	0.176	0.186	0.191
Interact. + random sampl	MRR@10	0.171	0.201	**0.199**	0.195	0.200	0.201
Interact. + Top 2	MRR@10	0.171	**0.202**	0.198	0.197	0.201	0.199
Interact. + random sampl@5	MRR@10	0.171	**0.202**	0.198	0.196	0.200	0.200
Interact. + Kmean	MRR@10	0.171	0.174	0.198	**0.201**	**0.215**	**0.222**

Table 5.4: Effectiveness results on the subset of MS Marco passage 2020 dev set (1000 queries - 1 relevant document per query)

the **Interact. + Top2** scenario observes a decrease of 12% in terms of MRR@10 for the test set w.r.t. the initial user query.

- This trend is reversed with each interaction turn to obtain for certain scenarios performance higher than baseline ones (see all interaction models in the test set, and the **Interact + Kmeans** for the dev set).
- The interaction model with K-mean strategy looks to be the best selection policy for question clarification since it obtains the highest MRR@10 for both datasets. This is somehow intuitive because this strategy might correspond to a refinement strategy, going deeper and deeper into clusters. This is also connected with the dataset peculiarity since MS Marco is mainly composed mono-faceted questions in natural language.
- It is moreover worth noting that performances increase with interaction turns but additional exploratory experiments highlight a saturation point after 5/6 interaction turns. Our setting, therefore, allows to interact with the user to clarify his/her needs without overloading the search session.

5.2.4 Conclusion

This exploratory work focuses on sequential click-based interaction with a user simulation for clarifying queries. We provide a simple and easily reproducible framework simulating multi-turn interactions between a user and an IR system agent. The advantage of our framework is the simplicity of interactions, as there is no need for a dataset of real and annotated user-system interactions. Experiments highlight performance gain in terms of document retrieval through the multi-turn query clarification process and provide a comparative analysis of selection strategies. This framework can be improved in terms of naturalness to better fit with conversational search. In practice, this implies learning to generate natural language interactions for both the IR system and user agents.

5.3 Discussion and achievements

Understanding information needs is a longstanding issue [Jansen *et al.* 2000, Moshfeghi *et al.* 2016] which has gained in maturity with the modeling [Kuhlthau 1991, Azzopardi 2014] and the leveraging of users' interactions [Lavrenko & Croft 2001,

Agichtein *et al.* 2006]. It has been addressed through different evaluation campaigns, such as TREC Interactive [Hersh & Over 2001], TREC Contextual Suggestion [Dean-Hall *et al.* 2014], or TREC Session Search [Carterette *et al.* 2014], and more recently TREC CAsT [Dalton *et al.* 2020b] which has introduced interactions in natural language for solving an IR task. In this chapter, we presented works dealing with the contextualization of information needs in conversational search, implying to leverage of users' interactions to explicit their intent and improve the search effectiveness. Our objectives were twofold: 1) investigating different conversational search settings: ad-hoc search given a conversation context and proactive search with query clarification interactions, and 2) exploring the potential of existing retrieval or question generation strategies based on large language models. We draw the three following main conclusions from our works.

Sparse retrieval models have good transfer properties to contextualize information needs when they are fine-tuned in a lightweight fashion. Sparse neural retrieval models are known as models focusing on term unit, either for both sparse indexing [Zamani *et al.* 2018] or term expansion [Bai *et al.* 2020, MacAvaney *et al.* 2020, Formal *et al.* 2021]. All these models rely on sparse representations and, compared to dense approaches [Guo *et al.* 2016, Pradeep *et al.* 2021], have the advantages to be efficiently used for indexation, to explicit lexical matching, and being interpretable. They provide also good generalization performances on out-of-domain benchmarks [Formal *et al.* 2022]. Guided by this last statement, our CoSPLADE model introduces a lightweight fine-tuning strategy to contextualize queries according to conversations. More particularly, we have not considered relevance signals as evidence sources of the fine-tuning, but we have rather focused on query intent and matched the representations of queries with the ones of gold queries. The results obtained by our CoSPLADE model highlight that it is possible to improve ranking performances in a target task (here conversational search) without requiring supervision of relevant documents. We believe that this outcome is promising for future works in IR investigating new research fields which might lack datasets with complete supervision data. More explicitly, by leveraging the transfer abilities of those sparse neural retrieval models (in our case SPLADE), it seems that it is possible to integrate new search dimensions (here, the conversation) without having the supervision of relevant documents in the (new) targeted IR task. Obviously, more experiments are needed to evaluate the generalization of this statement to other sparse neural retrieval models [Bai *et al.* 2020, MacAvaney *et al.* 2020] and other emerging tasks (e.g., FACT-IR or Personal Information Access [Culpepper *et al.* 2018]).

The world knowledge captured by language models does not capture well matching signals: toward the combination of IR signals and language models. (Large) language models are often seen as world knowledge since they are learned on large datasets such as Wikipedia enclosing a wide range of knowledge and they capture several language cues thanks to various learning objectives [Devlin *et al.* 2019, Raffel *et al.* 2020, Dai *et al.* 2022, Kiela 2022, Bommasani *et al.* 2021]. They have also shown great transfer abilities over various tasks, either with fine-tuning or in zero-shot settings. However, we believe that IR requires specific signals that might not be learned in standard language models. In 2016, Guo *et al.* 2016 already discussed the

difference between semantic matching and relevance matching. The literature review [Mitra 2021, Lin *et al.* 2021a, Fan *et al.* 2022] highlights the need to integrate IR features in neural ranking models or to pre-train language models by integrating the IR objective. Indeed, framing IR tasks as pointwise document classification as in the monoBERT model [Nogueira & Cho 2019] provides an efficient strategy but has quickly been improved by integrating IR techniques such as interaction matrix [Hofstätter *et al.* 2020], ranking losses [Pradeep *et al.* 2021], or leveraging prompt-tuning [Fan *et al.* 2022].

Throughout this chapter, we have confirmed this need in our both settings. The experimentation of our CoSPLADE model highlights the synergic effect of addressing both query reformulation and relevance score prediction to contextualize information needs and obtain promising retrieval performances. Similarly, the preliminary analysis of our simulation framework shows that simply generating queries with a T5 model is not sufficient, and that the decoding should be guided by the relevance signal.

When interactivity can complement language models to perform IR tasks. The birth of large language models has enabled tremendous advances in numerous research fields in NLP, particularly in IR allowing to complement matching signals with world knowledge [Nogueira & Cho 2019, Arabzadeh *et al.* 2021]. The recent large language models released to the community, such as ChatGPT [Ouyang *et al.* 2022], T0 [Sanh *et al.* 2022], BLOOM [Scao *et al.* 2022b], or PALM [Chowdhery *et al.* 2022b], are able to answer a wide range of questions, questioning the place of search engines in our society. However, in a recent article in the press (`https://iai.tv/articles/all-knowing-machines-are-a-fantasy-auid-2334`), Emily M. Bender and Chirag Shah explain that a search engine, and more largely information seeking, is more than a tool providing answers to questions. Large language models are effective tools for fact-based questions, but there is a crucial need to solve search sessions with human engagement, particularly for complex intents that require cognitive efforts. As already outlined in early information-seeking models (e.g., the Information Search Process (ISP) [Kuhlthau 1991], Anomalous States of Knowledge (ASK) [Belkin *et al.* 1982], or the Ellis model [Ellis 1989]), a search session is characterized by multiple sense-making actions relying on the formulation of the search intent after different observation/exploration/organization phases, and in the end on the assessment of the displayed information to ensure its truthiness/relevance regarding the search intent. As suggested in [Culpepper *et al.* 2018], we believe that conversational search is not a way to erase all these steps, but should rather support them. Interacting with users in natural language offers exciting abilities to allow them to explicit their intent, and early works [Aliannejadi *et al.* 2019, Zamani *et al.* 2020a] on clarifying questions for a mixed-initiative system are the proof that language models are used in the right way. Our second contribution in this chapter highlights the retrieval effectiveness improvement of such interactive settings compared to an ad-hoc search setting. Also, the next focus of the TREC CAsT Track giving up ranking in conversation context at the benefit of mixed-initiative corroborates the intuition of leveraging large language models to support search. However, additional steps should be addressed in the future by the IR community, such as ensuring the truthiness of answers by complementing natural language answers with relevant pieces of documents [Culpepper *et al.* 2018] or providing an overview of sense-making process (e.g., of the

search process pathway) towards explainability [Gu *et al.* 2021b].

It is worth noting that the results obtained for our contributions are limited to the experimental design. We mention below some limitations of our works, particularly for the work dealing with query clarification which reflects the current research trends toward mixed-initiative, and the associated research perspectives.

- **Designing accurate evaluation tasks for conversational search.** Although mentioning that conversational search can be particularly relevant for complex and multi-faceted queries, we have considered the MSMarco dataset [Nguyen *et al.* 2016b], which rather refers to simple queries for a few reasoning. This choice is motivated by the desire to build a proof of concept with simple elements (from both the IR and user agents) experimented on simple datasets. Considering a multi-faceted query implies therefore a more complex IR and system model balancing relevance and coverage. We think that it would be beneficial to combine our framework with existing models focusing on the modeling of multi-faceted queries [MacAvaney *et al.* 2021] or interactive refinement strategies [Mustar *et al.* 2022].

  In addition, there is a crucial need in the community to design adapted tasks (and datasets) in which conversational search will be useful. As already discussed, large language models are able to answer factual questions but we believe that conversational search can bring more sense-making and more truthiness in the search process, allowing to solve complex information-seeking tasks [Shah & Bender 2022]. The current datasets on query clarification [Aliannejadi *et al.* 2019, Tavakoli *et al.* 2022] and the exploratory analyses of user engagements [Sekulic *et al.* 2021b] introduce the first step in this direction and the efforts need to be pursued. Having in mind that Wizard of Woz evaluations [Sun *et al.* 2021] might be costly and time-consuming, one strategy could be to leverage simulations to build new datasets and allowing to vary users' behaviors as done in [Câmara & Hauff 2020].

- **Towards more naturalistic mixed-initiative with multi-turn interactions in natural language.** One limitation of our framework is that it relies on interactions consisting in displaying two queries and letting users choose between one of them, without natural language interactions, and especially without discussion. If the displayed queries are not relevant, the user agent always chooses one of them, which might hinder the intent clarification process. We believe that the next step for future work is to enhance our framework with natural language interactions. To do so, we propose to leverage actual query clarification datasets, such as QuLac [Aliannejadi *et al.* 2019] to build system and user agents, respectively aiming at generating natural language query clarifications and providing answers in response to those clarifications. Large language models might appear as basic tools to learn these interactions, but they critically miss the integration of IR task signals to generate queries that are semantically relevant to the search intent and enhance the retrieval effectiveness of the conversational search session. We, therefore, envision using IR techniques, such as pseudo-relevance feedback, to enhance the language generation underlying query clarification.

Outcomes All these works are conducted in the context of the ANR JCJC SESAMS for which I am the principal investigator. I briefly describe my supervision activity regarding the topic:

- The query clarification topic is addressed by a PhD student, Pierre Erbacher, co-supervised with Ludovic Denoyer.
- I have also co-supervised two master students Nawel Astouati and Nam Le Hai who have participated in the TREC CAsT evaluation track, resp. in 2021 and 2022. A research paper has also been submitted to ECIR 2023 to discuss our model proposed for TREC CAsT 2022. This work is done in collaboration with Jian-Yun Nie, Thomas Gerald, Thibaut Formal, and Benjamin Piwowarski.

You can find below a list of related publications[9]:

- Pierre Erbacher, Laure Soulier: État de l'art des approches de modélisation et de simulation utilisateur pour la recherche d'information conversationnelle. CORIA 2021
- Pierre Erbacher, Ludovic Denoyer, Laure Soulier: Interactive Query Clarification and Refinement via User Simulation. SIGIR 2022: 2420-2425
- Nawel Astaouti, Thomas Gerald, Maya Touzari, Jian-Yun Nie et Laure Soulier. "MLIA- LIP6@TREC-CAST2021 : Feature augmentation for query recontextualization and passage ranking". In : Working Notes of TREC CAST 2021. 2021.
- Le Hai Nam, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarski, Laure Soulier. "MLIA-DAC@TREC CAsT 2022: Sparse Contextualized Query Embedding". In : Working Notes of TREC CAST 2022. 2022.
- Le Hai Nam, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarski, Laure Soulier. CoSPLADE: Contextualizing SPLADE for Conversational Information Retrieval. In: ECIR 2023

[9] National publications are not mentioned if they are simple translations of international publications.

Chapter 6

Investigating neural ranking model behaviors in continual learning

Contents

In this section, we focus on the understanding of neural ranking model behaviors in terms of knowledge transfer in a continual learning scenario (RQ3).

6.1 Continual learning framework for neural IR

The large majority of works in computer vision [Kirkpatrick *et al.* 2016, Douillard *et al.* 2020b] and NLP [Chen *et al.* 2015, Veron *et al.* 2019] address continual learning through classification tasks (e.g., labeling objects or pixels for object segmentation). The usual framework consists of a sequence of classification tasks $\mathcal{T}_1 \rightarrow \mathcal{T}_2 \rightarrow \cdots \rightarrow \mathcal{T}_n$ in which labels evolve with tasks. Given a task $\mathcal{T}_i$ decomposed as a tuple of an input set $\mathcal{X}_i$ and their associated label set $\mathcal{Y}_i$, the classification function is formalized as: $F : \mathcal{X}_i \rightarrow \mathcal{Y}_i$. The labels $\mathcal{Y}_i$ belong to a set of predefined classes $\mathcal{C}_i$ that evolve with tasks, implying for the classification model to learn an output distribution that differs from one task to another one. For instance, a model can be trained on classifying images of dogs and cats (task 1), and then used for classifying cars and boats (task2), and so on.

In neural IR, the setting is somehow different. Given a set of queries $\mathcal{Q}$ and a document collection $\mathcal{D}$, a neural ranking model M aims at predicting a score y for a pair of query-document $(q, d) \in \mathcal{Q} \times \mathcal{D}$; therefore, $M : \mathcal{Q} \times \mathcal{D} \rightarrow \mathcal{R}$. If we try to apply this setting in a continual learning framework, there is no sense to vary the distribution of the

output (i.e., the scores of the documents). What is more willing to evolve in IR are the users' intents and the document collection, namely the input distribution of the neural model. Therefore, a continual learning stream for IR can be formalized as a sequence of tasks $\mathcal{T}_i$ which denotes a tuple $(\mathcal{Q}_i, \mathcal{D}_i)$ of query-document sets:

$$\mathcal{T}_1 \rightarrow \mathcal{T}_2 \rightarrow \cdots \rightarrow \mathcal{T}_n \tag{6.1}$$

$$with\ \mathcal{T}_i = (\mathcal{Q}_i, \mathcal{D}_i) \tag{6.2}$$

For experimental reasons, each query set $\mathcal{Q}_i$ in the stream is composed of train and test sets. Depending on the continual learning setting, either both the query and document sets might evolve with tasks $\mathcal{T}_i$ or only one of them (i.e., only the queries or only the document collection).

With this in mind, the neural ranking model M is trained sequentially using an adaptation method (e.g., fine-tuning or continual learning techniques) on each training set of the query set $\mathcal{Q}_i$ and the associated document collection $\mathcal{D}_i$, one by one, to obtain at each training step a model M_i with parameters $\hat{\theta}^i$. For instance, the model M_2 with parameters $\hat{\theta}^2$ is obtained by training model M on the training set of $\mathcal{Q}_1$ applied on the document collection $\mathcal{D}_1$ to initialize model M_1 with parameters $\hat{\theta}^1$. Then, this last model is trained using an adaptation method on the training set $\mathcal{Q}_2$ associated with the document collection $\mathcal{D}_2$.

For each trained model M_i with parameters $\hat{\theta}^i$, we can estimate its retrieval performance $R_{i,i}$ on the test set of each query set $\mathcal{Q}_i$ given the document collection $\mathcal{D}_i$. In addition, we can measure the model's ability to accumulate/forget knowledge through the stream:

- The Forward Transfer (FT) which estimates the influence that learning a task $\mathcal{T}_i$ has on the performance on a future task $\mathcal{T}_j$, with $j > i$. When the forward transfer is positive, the model M_i is able to accumulate knowledge from previous tasks and is, therefore, addressing the zero-shot learning problem.

- The Backward Transfer (BT) which measures the impact of learning a task $\mathcal{T}_i$ on the performance on a previous task $\mathcal{T}_j$, with $j < i$. When the performance on previous tasks is lowered, we call this phenomenon *(catastrophic) forgetting*.

We, therefore, denote $R_{i,j}$ the performance of model M_i trained on the task stream up to $\mathcal{T}_i$ and evaluated on task $\mathcal{T}_j$. Depending on whether $j < i$ or $j > i$, $R_{i,j}$ is included in the estimation of the Forward and Backward Transfer. At a high level, a continual learning framework in IR can be illustrated in Figure 6.1 and in Algorithm 2.

All the difficulty in designing continual learning scenarios for IR lies in the design of the stream. One can think that each query in a dataset can be equated to a single task, and thus the continual learning setting is built as a stream of successive single queries $(\mathcal{Q}_i = q_i)$ and a fixed document collection $(\mathcal{D}_i = \mathcal{D}; \forall i = 1, \ldots, n)$. One drawback of this modeling is inherent to the continual learning framework: the additional signals captured at an iteration while training the model on each task will be very small. Indeed, we are not sure that the relevance signals of a single query are sufficient to measure knowledge drift. We, therefore, believe that a task $\mathcal{T}_i$ in an IR continual learning framework might include a group of queries (and a possibly evolving collection of documents), characterized by similar properties to have enough similar knowledge to impact the training step.

Algorithm 2 A continual learning framework for IR

Set up an ordered task stream **setting** $\mathcal{T}_1 \rightarrow \ldots \mathcal{T}_{n-1} \rightarrow \mathcal{T}_n$ with $\mathcal{T}_i = (\mathcal{Q}_i, \mathcal{D}_i)$
Initialize a model M_0 with random parameters or use a pre-trained model
for *k=1* **to** n **do**
  Train model M_i on the training query set $\mathcal{Q}_i$ given the document collection $\mathcal{D}_i$.
  Measure the retrieval performance $R_{i,i}$ of model M_i on test query set $\mathcal{Q}_i$.
  Measure the retrieval performance $R_{i,j}$ $(j > i)$ of model M_i on the testing instances of next query set $\mathcal{Q}_j$ (forward transfer).
  Measure the retrieval performance $R_{i,j}$ $(j < i)$ of model M_i on the testing instances of previous query sets $\mathcal{Q}_j$ (backward transfer)

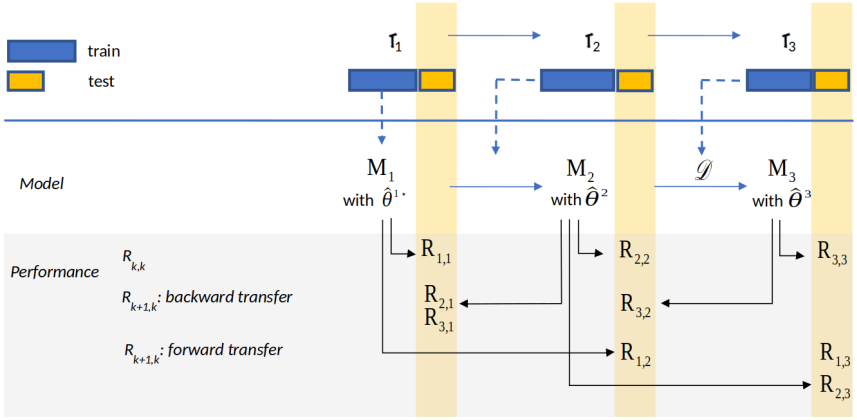


Figure 6.1: Illustration of the continual learning framework in IR using a 3-task stream setting for a given model M

The next sections instantiate the continual learning framework in IR in which task streams are designed using different assumptions:

- A short stream in which tasks are delimitated by the application domain: we consider different document collections (and their associated queries) dealing with the generic or medical domain, as well as documents from microblogs. In this setting, both query and document sets evolve with tasks.
- A long stream in which tasks are delimitated by topics: queries of a single dataset are clusterized so as to build sets of queries belonging to the same topic/subtopic (e.g., cooking with barbecue, salad cooking, gardening, etc...). In this setting, the dataset is fixed throughout all tasks $\mathcal{T}_i$ and only the query sets $\mathcal{Q}_i$ evolve.

We investigate the behavior of neural ranking models regarding the catastrophic forgetting issue, measured using the backward transfer.

6.2 Analyzing catastrophic forgetting in short streams

Our objective is twofold: 1) evaluating different neural ranking models on a short stream of successive tasks $\mathcal{T}_i$ delimited by different domains and 2) investigating their behavior regarding the catastrophic forgetting issue.

6.2.1 Experimental setting

We use three datasets chosen to fit with the requirement of cross-domain adaptation [Pan & Yang 2010]: 1) MS MARCO (***ms***) [Nguyen *et al.* 2016a] a passage ranking dataset built using the Bing search logs; 2) TREC Microblog (***mb***) [Lin & Efron 2013], an ad-hoc search dataset from TREC Microblog 2013 and 2014, which contains a public Twitter sample stream; 3) TREC CORD19 (***c19***) [Wang *et al.* 2020] an ad-hoc document search dataset including research articles dealing with SARS-CoV-2 or COVID-19 topics.
Besides, we consider four settings (See Table 6.1, column **"Setting"**) among which three 2-dataset (n = 2) and one 3-dataset (n = 3) settings. As done in previous work [Li & Hoiem 2018, Asghar *et al.* 2020], these settings follow the patterns ($task_1 \rightarrow \mathcal{Q}_2$) or ($\mathcal{T}_1 \rightarrow \mathcal{T}_2 \rightarrow \mathcal{T}_3$) where query set orders (i.e., dataset orders) are based on the decreasing sizes of the training sets assuming that larger datasets allow starting with well-trained networks.

Neural ranking models. We consider five state-of-the-art models [Yang *et al.* 2019a]: 1) interaction-based models: DRMM [Guo *et al.* 2016], PACRR [Hui *et al.* 2017] and KNRM [Xiong *et al.* 2017]; 2) BERT-based models: Vanilla BERT [Devlin *et al.* 2019] and CEDR-KNRM [MacAvaney *et al.* 2019b]. We use the OpenNIR framework [MacAvaney 2020] that provides a complete neural ad-hoc document ranking pipeline (a first-stage ranking with BM25 followed by a second-stage ranking with the mentioned models). Note that in this framework, the neural models are trained by linearly combining their own neural score (S_{NN}) with a BM25 score (S_{BM25}). We call the final score the *global relevance score*.

Domain adaptation and lifelong learning methods. We adopt the standard fine-tuning strategy (training on one domain and fine-tuning on the other) as the representative domain adaptation method. Additionally, we investigate the Elastic Weight Consolidation (EWC) [Kirkpatrick *et al.* 2016] as the lifelong learning method $\mathcal{L}$ and analyze its potential in IR.

Measures. To measure the knowledge acquired by the model during the re-ranking step, we measure the relative improvement achieved with the ranking based on the global relevance score (resp. the neural score) trained and tested on the previous dataset over the performance of the BM25 ranking obtained on the same testing dataset. We note this metric MAP@100 Δ_{MAP} (resp. Δ_{MAPN}). The objective is therefore to estimate how much knowledge is captured by neural ranking models given the first stage.

Concerning the catastrophic forgetting measure, we use the *remembering* measure (REM) derived from the *backward transfer measure* (BWT) proposed in [Rodríguez *et al.* 2018].

- **BWT**: measures the intrinsic effect (either positive or negative) that learning a model M on a new task $\mathcal{T}_i$ has on the model performance obtained on an old task $\mathcal{T}_j$ with $j < i$, referred as *backward transfer*. Practically, in line with a lifelong learning perspective, this measure averages, in the task stream, the differences between the performances of the model obtained on the previous task and the

performances of the oracle model trained and tested on the same previous task. Thus, while positive values represent positive backward transfer, negative values express catastrophic forgetting. Formally, the BWT measure is computed as:

$$BWT = \frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} (R_{i,j} - R^*_{j,j})}{\frac{n(n-1)}{2}} \tag{6.3}$$

$R_{i,j}$ is the performance measure of model M_i obtained right after learning on task $\mathcal{T}_j$. $R^*_{j,j}$ is the performance of the oracle model M^*_j trained on task $\mathcal{T}_j$ and tested on the same task. To make fair comparisons between the different studied neural models, we normalize the differences in performance $(R_{i,j} - R^*_{j,j})$ on model agnostic performances obtained using $BM25$ model on each previous task $\mathcal{T}_j$. Formally, we estimate $R_{ij} = \frac{MAP(M_i, \mathcal{T}_j)}{MAP(BM25, \mathcal{T}_j)}$ where $MAP(M_i, \mathcal{T}_j)$ is the effectiveness of model M_i on the task $\mathcal{T}_j$. In our work, we only report the REM values computed using the MAP measure (we observe similar trends for NDCG@20 and P@20).

- **REM**: because the BWT measure has a bivalent meaning, i.e. positive values for positive backward transfer and negative values for catastrophic forgetting, we report the REM metric that is only concerned about forgetting. Formally, it is estimated as follows:

$$REM = 1 - |min(BWT, 0)| \tag{6.4}$$

A REM value equal to 1 means that the model does not catastrophically forget. We denote REM and REMN, the remembering metric applied on the ranked list obtained using, respectively, 1) a linear combination of BM25 and neural scores (also called global relevance score), and 2) solely the neural score.

6.2.2 Results

Table 6.1 reports all the metric values for each model/setting pairwise. Regarding the **"Fine-tuning"** adaptation technique aiming at measuring the catastrophic forgetting (RQ1), we can outline the following statements.

Catastrophic forgetting in short IR streams is not as clear as in Computer Vision. While previous works have shown that neural models suffer from catastrophic forgetting in large proportion [Kirkpatrick *et al.* 2016], the REM and REMN metrics in IR are in general close to 1, with small variation. This suggests that catastrophic forgetting is not as strong as in computer vision, and that neural ranking models are more driven by relevance matching signals during the learning process than the application domain or the topic of queries.

Bert-based models are able to bring effectiveness gains additively to those brought by the exact-based matching signals in BM25. Only CEDR and VBERT models achieve positive improvements w.r.t to both the global ranking (Δ_{MAP} : +19.6%, +17.4% resp.) and the neural ranking (Δ_{MAP}: +29.2%, +25.8% resp.), particularly under the setting where mb is the previous dataset ($mb \rightarrow c19$). These effectiveness gains can be viewed as new knowledge in terms of semantic matching.

Model	Setting	Fine-tuning		EWC-based lifelong learning	
		REM(REMN)	$\Delta_{MAP(MAPN)}$	*REM(REMN)*	$\Delta_{REM(REMN)}$
DRMM	$ms \rightarrow c19$	1.000(1.000)	+2.2(-73.6)	1.000(1.000)	0(0)
	$ms \rightarrow mb$	0.962(0.943)	+2.2(-73.6)	0.971(0.974)	+0.9(+**3.3**)
	$mb \rightarrow c19$	1.000(0.965)	-1.7(-7.7)	1.000(0.662)	0(-31.4)
	$ms \rightarrow mb \rightarrow c19$	0.976(0.938)	+2(-73.6)	0.979(1.000)	+**0.3**(+**6.6**)
PACRR	$ms \rightarrow c19$	1.000(0.760)	+2.5(-30.1)	1.000(0.756)	0(-0.5)
	$ms \rightarrow mb$	1.000(1.000)	+2.5(-30.1)	1.000(1.000)	0(0)
	$mb \rightarrow c19$	1.000(0.523)	0(+10)	1.000(0.940)	0(+**79.7**)
	$ms \rightarrow mb \rightarrow c19$	1.000(0.759)	+2.5(-30)	1.000(0.874)	0(+**15.2**)
KNRM	$ms \rightarrow c19$	1.000(1.000)	-12.1(-89)	1.000(1.000)	0(0)
	$ms \rightarrow mb$	1.000(1.000)	-12.1(-89)	1.000(1.000)	0(0)
	$mb \rightarrow c19$	1.000(0.810)	-2(-13.8)	1.000(0.902)	0(+**11.4**)
	$ms \rightarrow mb \rightarrow c19$	1.000(1.000)	-12.1(-89)	1.000(0.963)	0(-3.7)
VBERT	$ms \rightarrow c19$	0.930(1.000)	-10.6(0)	1.000(1.000)	+**7.5**(0)
	$ms \rightarrow mb$	1.000(0.883)	-10.6(0)	1.000(1.000)	0(+**13.3**)
	$mb \rightarrow c19$	0.913(1.000)	+17.4(+25.8)	1.000(1.000)	+**9.5**(0)
	$ms \rightarrow mb \rightarrow c19$	0.989(0.922)	-10.6(0)	1.000(1.000)	+**1.1**(+**8.5**)
CEDR	$ms \rightarrow c19$	0.826(1.000)	+2.6(+14.2)	1.000(1.000)	+**21.1**(0)
	$ms \rightarrow mb$	0.510(0.920)	+2.6(+14.2)	1.000(1.000)	+**96.1**(+**8.7**)
	$mb \rightarrow c19$	0.940(1.000)	+19.6(+29.2)	1.000(1.000)	+**6.4**(0)
	$ms \rightarrow mb \rightarrow c19$	0.771(0.946)	+2.6(+14.2)	0.891(1.000)	+**15.6**(+**5.7**)

Table 6.1: Per model-setting results in our fine-tuning and EWC-based lifelong learning experiments. All the measures are based on the MAP@100 metric. The improvements $\Delta_{MAP(MAPN)}$ and $\Delta_{REM(REMN)}$ are reported in percent (%).

Capturing additional knowledge w.r.t exact-matching signals does not avoid catastrophic forgetting. While some models-settings pairs are able to capture a large amount of additional knowledge (e.g., VBERT and CEDR in the $mb \rightarrow c19$ setting) without forgetting information on previous tasks (REM and REMN metrics close to 1), this trend is not obvious in other cases when looking at the correlation between REM and Δ_{MAP} metrics. For instance, DRMM generally does not forget but the accumulation of knowledge regarding exact-matching signals is very diverse, ranging from negative to positive values. Interestingly, the KNRM does not accumulate knowledge regarding exact-matching signals nor forgets knowledge during the fine-tuning.

We turn now our attention to the **"EWC-based lifelong learning"** columns in Table 6.1 to investigate the gain of lifelong learning strategies [Kirkpatrick *et al.* 2016] (RQ3). Our experiment results show that among the 9 (resp. 11) settings that exhibit catastrophic forgetting in the combined model (resp. neural model), the **EWC strategy allows to improve** 9/9 i.e., 100% (resp. 9/11 i.e., 88%) of them in the range $[+0.3\%, +96.1\%]$ (resp.$[+3.3\%, +79.7\%]$). Given, on the one hand, the high variability of the settings derived from the samples, and on the other hand, the very low number of settings (10% i.e., 2/20) where a performance decrease is observed in the previous dataset, we could argue that the **EWC-based lifelong learning is not inherently impacted by dataset order leading to a general effectiveness gain over the models**.

6.3 Designing long topic streams and analyzing pathological IR behaviors

One drawback of the previous contribution is that it focuses on very short and synthetic streams, which can be limited to infer robust behavior in real continual learning settings characterized by an infinite timeline. Our objective here is to build a long stream for evaluating the behavior of neural ranking models in a continual learning framework. Due to computational reasons, we limit this analysis to two neural ranking models: 1) the vanilla Bert [Devlin *et al.* 2019] (noted **VBert**) and 2) the *Mono-T5-Ranker* [Nogueira *et al.* 2020] (noted **MonoT5**).

6.3.1 Building a dataset with long topic sequences.

One main difficulty is to create this sequence considering the availability of IR datasets. In contrast to our previous work based on a sequence of datasets of different domains [Lovón-Melgarejo *et al.* 2021], we propose to model the task at a lower granularity level, namely topics, instead of the dataset granularity[1]. To create the long sequence, we consider a fixed dataset $\mathcal{D}$, namely the MSMarco dataset [Nguyen *et al.* 2016c], assuming that several queries might deal with the same user's interest (e.g., "what is the largest source of freshwater on earth?" or "what is water shortage mitigation"). These groups of queries $\mathcal{Q}_i$ denote what we call *topics* and each task $\mathcal{T}_i$ in the stream is thus built of a query set $\mathcal{Q}_i$ and the dataset $\mathcal{D}$.

To extract topics $\mathcal{Q}_i$, we propose a clustering-based method consisting in extracting clusters from randomly sampled queries using a sentence-BERT clustering[2] and populating those clusters with queries from the whole dataset. Finally, the sequence of topics is produced by randomly rearranging clusters to avoid bias of cluster size. Depending on the value of clustering hyper-parameters, we obtain three datasets of topic sequences of different sizes (19, 27, and 74), resp. called *MS-TS*, *MS-TM* and *MS-TL* (for small, medium, large). To evaluate our topic sequence methodology, for each of the three datasets we create a long topic sequence baseline in which clusters are randomly built. We obtain three randomized datasets denoted *MS-RS*, *MS-RM*, and *MS-RL*.

To verify the relevance of the clusters, we measure retrieval similarity within and between clusters (i.e., queries within clusters might have similar retrieval evidence and queries between clusters might have different ones). As retrieval similarity between query clusters, we use the retrieved documents for each query using the *BM25* model[3]. Our intuition is that similar queries should have more commonly retrieved documents (and vice versa). For this, we denote the $c-score$ which measures the ratio of common documents between two topics $\mathcal{Q}_i$ and $\mathcal{Q}_j$. Statistics of these three topic sequences and the relevance of clusters through intra and inter c-score are described in Table 6.2. Moreover, Figure 6.2 depicts the $c-score$ matrix for all couples $(i,j) \in \{1,2,\ldots,|S|\}^2$ for a subset of 8 topics (for more clarity in the figures) of the $MS-S$ and $MS-RS$ corpora. We observe that for the randomized matrix (Figure 6.2 right), the metric value

[1] Please note that the number of datasets adapted to neural *IR* with a sufficiently large number of queries and relevance judgments is not sufficient to build a long sequence of datasets as we envision.

[2] https://www.sbert.net/examples/applications/clustering (fast clustering)

[3] Implemented in pyserini: https://github.com/castorini/pyserini

Name	$\lvert\#topics\rvert$	#queries by topics	inter	intra
MS-TS	19	$3,650 \pm 1,812$	3.8%	31.4%
MS-TM	27	$3,030 \pm 1,723$	4.1%	32.1%
MS-TL	74	$1,260 \pm 633$	3.3%	34.6%
MS-RS	19	$3,650 \pm 1,812$	10.3%	10.2%
MS-RM	27	$3,030 \pm 1,723$	9.9%	9.8%
MS-RL	74	$1,260 \pm 633$	8.7%	8.8%

Table 6.2: Parameters and statistics of the generated dataset and their inter/intra topic similarity metric ($c-score$). The intra-score is the mean $c-score$ when comparing a topic with itself, and the inter-score when comparing different topics.

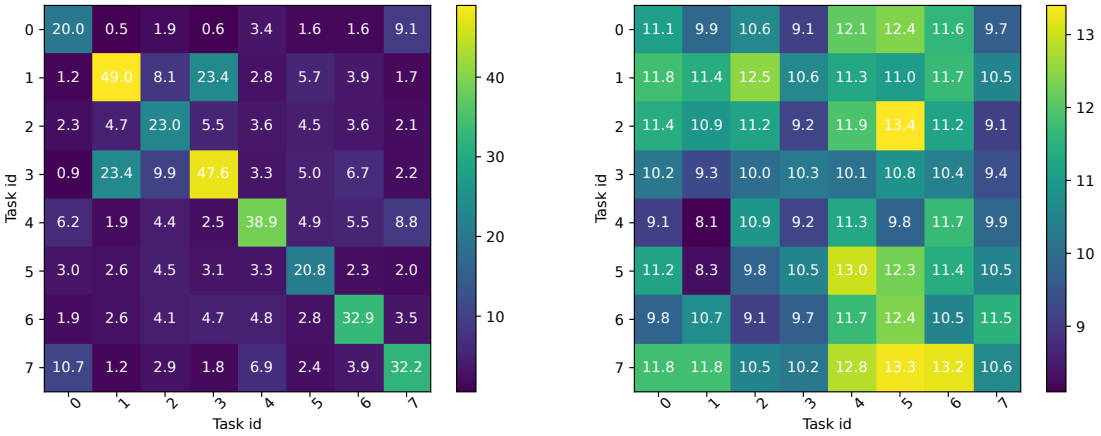


Figure 6.2: Matrix of similarities between topics for 8 topics of MS-S (left) and MS-RL (right) datasets. The c-score (×100) is processed on all topic pairs, a high value (yellow) denotes the level of retrieved document overlap between queries of topics.

is relatively uniform. In contrast, in the matrix obtained from our long topic sequence based on clustering (Figure 6.2 left), the c-score is very small when computed for different topic clusters (low inter similarity) and higher in the diagonal line (high intra similarity).

6.3.2 Analyzing the behavior of neural ranking models on long topic sequences

We investigate now the global performance of neural ranking models after having successively been fine-tuned on topics in our MSMarco-based long sequence setting (Table 6.3). For comparison, we report results for the multi-task baseline in which models are trained on all the topics of the sequence jointly (without sequence consideration). At first glance, we can remark that, in a large majority, neural models after fine-tuning on random sequences or multi-task learning obtain better results than after the fine-tuning on our long topic sequences. This can be explained by the fact that, within our setting, **the topic-driven sequence impacts the learning performance: a supplementary effort is needed by the model to adapt to new domains**, which is not the case in the random setting. In this latter, the diversity is at the instance level. This trend is depicted in Figure 6.3, highlighting peaks in the clustering-based setting (blue line) referring to topic/cluster changes. This result confirms that catastrophic forgetting might occur with neural ranking models.

Model	Dataset	Learning protocol		
		Random	clustering	Multi-task
VBert	SMALL	18.4/19.6	16.3/17.5	**18.5/19.7**
	MEDIUM	**17.9/19.0**	17.8/18.9	17.5/18.7
	LARGE	**18.8/19.9**	17.3/18.5	18.5/19.7
MonoT5	SMALL	**16.1/17.3**	13.1/14.4	15.5/16.8
	MEDIUM	15.4/16.7	13.4/14.7	**15.7/17.1**
	LARGE	13.9/15.1	13.8/15.1	**15.7/17.0**
BM25	SMALL		10.8/11.7	
	MEDIUM		10.5/11.4	
	LARGE		11.7/12.7	

Table 6.3: General performance of neural ranking models on long topic sequences. Mean performances on all the topic sequences reporting mrr@10/mrr@100 for the different models.

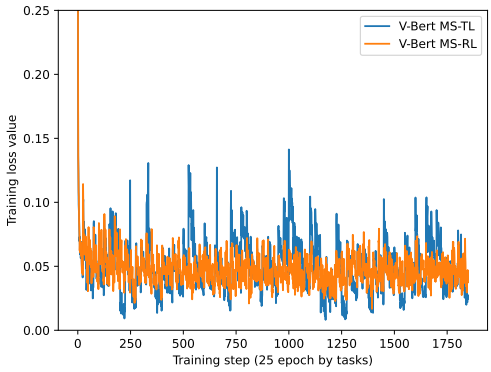


Figure 6.3: General performance of neural ranking models on long topic sequences. VBert loss values for both random and clustering-based large corpus.

Through our different analyzes (see the paper [Gerald & Soulier 2022] for complete experiments), we highlighted the following general trends:

- **Long stream of tasks implies noticeable catastrophic forgetting.** The comparison of neural ranking model performances when trained on task sequences (random or clustering-based) w.r.t multi-task learning highlights that continual learning leads to lower effectiveness results. It is more prevalent in our setting, i.e., when tasks are split by topic. Combined with the previous analysis highlighting small catastrophic behaviors on short streams (although using different metrics), this suggests that neural ranking models are more prone to forget when trained on long topic streams. This result is consistent with previous work in computer vision [Kirkpatrick *et al.* 2016, Douillard *et al.* 2020b].

- **Ranking models behave differently in terms of catastrophic forgetting.** We notice that catastrophic forgetting occurs more the MonoT5 model being more sensitive to new domains than the VBert model. This can also explain by the difference in the way of updating weights (suggested in the original papers

[Devlin *et al.* 2019, Nogueira *et al.* 2020]). In VBert, two learning rates are used: a small one for the Bert model and a larger one for the scorer layer; implying that the gradient descent mainly impacts the scorer. This intuition needs more investigation since we use the second-order gradient descent of ADAM. In contrast, the MonoT5 is learned using a single learning rate leading to modify the whole model. For a reminder, the previous analysis on short stream highlighted forgetting behavior for the CEDR model, which is also a joint model introducing the representation learning of contextual embedding into neural ranking models such as PACRR, DRMM, and KRNM. We can therefore infer that multiple objective functions might hinder knowledge retention in continual learning settings.

- **The more topics are similar, the less neural ranking models forget.** In contrast to continual learning in other application domains [Kirkpatrick *et al.* 2016, Rebuffi *et al.* 2017] in which fine-tuning models on other topics always deteriorates previous topic performance, our analysis suggests that topics might help each other (particularly when they are relatively similar), at least in lowering the catastrophic forgetting. Moreover, as discussed in [Guo *et al.* 2016], relevance matching signals play an important role in model performance, often more than semantic signals. The topic sequence may lead to a synergic effect to perceive these relevance signals.

In brief, continual learning in IR differs from the usual classification/generation lifelong learning setting. It is more likely to have different topics allowing to "help" each other, either by having closely related topics or by focusing on query-document matching signals.

6.3.3 Analyzing pathological behaviors using IR-driven controlled stream-based scenarios

Having in mind that a task $\mathcal{T}_i$ is built of a tuple $(\mathcal{Q}_i, \mathcal{D}_i)$ of query and document sets, we have seen in our two previous analyses that query and documents sets might evolve simultaneously with the stream (as in our short stream scenario) or not (as in our long stream scenario). Guided by IR-driven use cases, we aim here to further our investigation regarding the typology of evolving data (documents and/or queries). Typically, the available documents may change over time, and some might become outdated (for instance documents relevant at a certain point in time). Also, queries evolve, either because of new trends, the emergence of new domains, or shifts in language formulation. To model those scenarios, we propose three different short task streams designed as IR-controlled scenarios. Tasks are based on our long topic sequence $S = \{\mathcal{T}_1, \ldots, \mathcal{T}_i, \ldots, \mathcal{T}_n\}$ built on MSMarco. For each scenario, we consider an initial setting $\mathcal{T}_{init}$ modeling the general knowledge before analyzing a particular setting. In other words, $\mathcal{T}_{init}$ constitutes the data used for the pre-training of neural ranking models before fine-tuning on a specific sequence. The proposed controlled settings are presented in what follows and Tables 6.4 and 6.5 present the obtained results for all settings according to the different scenario configurations.

- **Direct Transfer scenario [Veniat *et al.* 2020]**: The task sequence is $(\mathcal{T}_{init}, \mathcal{T}_i^+, \mathcal{T}_j, \mathcal{T}_i^-)$ where tasks $\mathcal{T}_i^+$ and $\mathcal{T}_i^-$ belong to the task $\mathcal{T}_i$ and have different sizes ($|\mathcal{T}_i^-| \ll |\mathcal{T}_i^+|$). This setting refers to when the same task comes back

	DT scenario			B
	$\mathcal{T}_i^+$	$\mathcal{T}_j$	$\mathcal{T}_i^-$	
MonoT5	26.6	24.9	26.6	27.2
VBert	28.5	26.7	27.3	28.9

Table 6.4: Model performances using MRR@10 on the Direct Transfer (DT) *IR*-driven controlled setting. B stands for the baseline model: training on both $\mathcal{T}_{init}$ and $\mathcal{T}_i$.

		IU			LD		
		$\mathcal{T}_i'$	$\mathcal{T}_i''$	B	$\mathcal{T}_i^*$	$\mathcal{T}_i^{**}$	B
MonoT5	$Q_{i1}D_{i1}$	28.15	29.6	-	15.6	23.0	-
	$Q_{i2}D_{i2}$	7.75	26.0	-	16.8	26.5	-
	$Q_{i1}D_{i1} \cup Q_{i2}D_{i2}$	18.2	27.8	27.2	15.6	23.8	27.2
VBert	$Q_{i1}D_{i1}$	23.7	30.2	-	28.2	30.1	-
	$Q_{i2}D_{i2}$	14.5	31.4	-	25.5	25.5	-
	$Q_{i1}D_{i1} \cup Q_{i2}D_{i2}$	19.1	30.9	28.9	26.6	27.0	28.9

Table 6.5: Model performances using MRR@10 on *IR*-driven controlled settings: Information Update (IU) and Language Drift (LD). B stands for the baseline models: fine-tuning on $\mathcal{T}_i$ for IU and LD scenarios.

in the stream with newly available data (new queries and new relevant documents). As shown in the DT scenario column in Table 6.4, the performance of both models on task $\mathcal{T}_i$ drops after fine-tuning on a foreign topic (i.e., on task $\mathcal{T}_2$). This highlights a catastrophic forgetting behavior. However, both models are able to slightly adapt their retrieval performance after fine-tuning on task $\mathcal{T}_i^-$. This final performance is however lower than the baseline model (training on both $\mathcal{T}_{init}$ and $\mathcal{T}_i$) and for the VBert model lower than its initial performance at the beginning of the learning sequence. These two last statements suggest the ability of neural models to quickly reinject a part of the retrained knowledge learned in the early sequence to adapt to new query/document distributions on the same topic.

- **Information Update scenario**: The task sequence is $(\mathcal{T}_{init}, \mathcal{T}_i', \mathcal{T}_i'')$ where $\mathcal{T}_i'$ and $\mathcal{T}_i''$ have dissimilar document distributions and a similar query distribution. Intuitively, it can be interpreted as a shift in the required documents, such as new trends concerning a topic or an update of the document collection. The IU columns in Table 6.5 highlight that evaluation performances increase throughout the fine-tuning process over the sequence. This denotes the ability of models to adapt to new document distributions (i.e., new information in documents). The adaptation is more important for the MonoT5 model, probably explained by its better adaptability to new topics. Interestingly, the performance at the end of the learning sequence overpasses the result of the baseline (fine-tuning on $\mathcal{T}_i$): this can be explained by the methodology used to create this setting, associating pseudo-relevant documents to existing queries to simulate the information update (more details in the paper [Gerald & Soulier 2022]). Our intuition is that the introduced pseudo-relevant documents in task $\mathcal{T}_i'$ might help in perceiving relevance signals.

- **Language Drift scenario**: The topic sequence is $(\mathcal{T}_{init}, \mathcal{T}_i^*, \mathcal{T}_i^{**})$ where $\mathcal{T}_i^*$ and

$\mathcal{T}_i^{**}$ have similar document distributions and a dissimilar query distribution. This can correspond to a change of query formulation or a focus on the same topic. As outlined in the LD columns in Table 6.5, the behavior is relatively similar to IU in terms of adaptation: performances increase throughout the sequence. Note that MonoT5 seems more flexible in terms of adaptation. However, it seems difficult to sufficiently acquire enough knowledge to reach the baseline performance (although pseudo-relevant documents have also been introduced as in the IU scenario). This might be due to the length of queries concerned by the distribution drift: when the vocabulary changes in a short text (i.e., queries), it is more difficult to capture the semantics for the model and to adapt itself in terms of knowledge retention than when the change is carried out on long texts (i.e., documents as in the information update).

6.3.4 Conclusion

In this work, we have designed a continual learning dataset for IR including long topic sequences and controlled IR sequences. Our investigation aims at observing a catastrophic forgetting metric for different models, and also in regard to topic similarity. Our analysis suggests different design implications for future work: 1) catastrophic forgetting in IR exists but is low compared to other domains [Kirkpatrick *et al.* 2016, Veniat *et al.* 2020], 2) when designing lifelong learning strategy, it is important to care for topic similarity, the position of the topic in the learning process and for the type of the distribution that needs to be transferred (short vs. long texts).

6.4 Discussion and achievements

In these works [Lovón-Melgarejo *et al.* 2021, Gerald & Soulier 2022], we have defined a continual learning framework for IR and investigated the catastrophic forgetting behavior of neural ranking models in short and long settings. We have carried out a fined-grained evaluation, observing a catastrophic forgetting metric for different models, and also in regards to topic similarity. The main conclusions that we draw from this line of research are the followings.

Neural ranking models have generally good properties regarding catastrophic forgetting. Previous works in computer vision [Kirkpatrick *et al.* 2016, Davidson & Mozer 2020, Douillard *et al.* 2020a, Ramasesh *et al.* 2021] highlight a clear trend toward catastrophic forgetting whether measuring performance on short [Kirkpatrick *et al.* 2016, Ramasesh *et al.* 2021] or long [Douillard *et al.* 2020a, Davidson & Mozer 2020] streams. We outlined in our experimental evaluation that this trend is not as strong. Neural ranking models do not seem to forget a lot throughout the learning process, and the size of the stream (2 or 3) is not correlated to the amount of forgotten information. In the long-stream setting, neural ranking models seem to forget knowledge when compared with multi-task learning. To get a better understanding of the phenomenon, we analyze the performance according to two additional dimensions. First, by putting in the abyss the forgetting phenomenon with the similarity of successive tasks, we show that the greater the similarity, the less the neural ranking

model forgets. It is worth noting that this trend toward task similarity has also been demonstrated in natural language understanding [Cattan *et al.* 2022]. Second, inspired by [Veniat *et al.* 2020], we evaluate three controlled IR-driven scenarios, highlighting generally good properties in terms of knowledge retention. Our intuition underlying this lower propensity to catastrophic forgetting is that neural ranking models are, by nature, designed for capturing relevance matching signals beforehand semantic matching signals [Guo *et al.* 2016, Lin *et al.* 2021a]. Consequently, switching topics/domains in IR is less critical than continually learning over different images to identify their labels that directly map with the semantics of images [Davidson & Mozer 2020, Douillard 2022].

The model architecture matters. As previously observed in [Arabzadeh *et al.* 2021], we have shown that neural ranking models are able to capture additional knowledge than the one captured by first-stage ranking models based on exact matching, such as BM25. However, this ability to capture additional knowledge does not bring necessarily catastrophic forgetting behaviors. This suggests that this additional knowledge does not obviously refer to new semantics but can rather complement relevance-matching signals already captured by exact-matching models. We thus believe that the model architecture impacts its behavior toward catastrophic forgetting. Depending on the architecture (i.e., transformer [Devlin *et al.* 2019, MacAvaney *et al.* 2019b] or not [Guo *et al.* 2016, Hui *et al.* 2017, Xiong *et al.* 2017], based on semantic clustering [Xiong *et al.* 2017, MacAvaney *et al.* 2019b] or not) and the losses used (classification [Nogueira & Cho 2019] or ranking loss [Devlin *et al.* 2019]), we notice different behaviors in terms of knowledge acquisition and retention. The more the features are semantically oriented, the more the model will tend to forget. This statement related to the topology of neural architecture has already been observed in computer vision by Huo & Zyl 2020.

The type of evolving data matters. The typology of data evolving in the continual learning scenario (e.g., documents or queries) impacts the learning behaviors regarding the evolving knowledge. In our IR-driven scenarios, neural ranking models outline good properties to face direct transfer or information update. In contrast, language drift in the query vocabulary remains a difficult task, probably due to the small expressiveness of queries (due to their size). This fine-grained analysis of input-output distribution has been initiated by Veniat *et al.* 2020 who have also noticed that a neural model can exhibit different behaviors regarding different controlled settings (analyzing for instance the transfer to similar input/output distributions, the knowledge update, the direct transfer, or the scalability). These experiments highlight the importance of analyzing different dimensions of the catastrophic forgetting issue. Combined with our investigations highlighting different behaviors regarding the typology of streams, we believe that a relevant strategy for designing neural ranking models robust to continual learning settings is to modularize the learning strategies according to the properties of the evolving data.

We are aware that obtained results are limited to the experimented models and settings, although we have considered various evaluation scenarios over different dataset peculiarities (variation in terms of domain, stream size, and controlled settings). We believe that much remains to be accomplished for more generalizable results, particularly

at the model level. For instance, it would be interesting to experiment with sparse neural ranking models [Bai *et al.* 2020, MacAvaney *et al.* 2020, Formal *et al.* 2021] to identify whether their zero-shot learning abilities are robust to a continual learning setting. However, we hope that our exploratory analysis is a step forward in the understanding of continual *IR* model learning and the design of more robust neural ranking models. More particularly, we believe that, although less characterized by catastrophic forgetting issues than neural models in computer vision, neural ranking models can gain robustness if they are able to identify critical evolution in the training data and alleviate this forgetting phenomenon. One promising strategy emerging from the Machine Learning community arises from mode connectivity [Kuditipudi *et al.* 2019, Benton *et al.* 2021, Wortsman *et al.* 2021] aiming at connecting different regions within the parameter space to leverage various signals. We believe that this principle could be used to design models modularizing their parameters depending on the typology of evolving data.

Outcomes These works are conducted in the context of the ANR JCJC SESAMS for which I am the principal investigator. I briefly describe my supervision activity regarding the topic:

- I initiated the framework of continual learning in IR (domain adaptation in short dataset streams). I collaborated on this topic with Lynda Tamine-lechani and Karen Pinel-Sauvagnat from the IRIT laboratory through the co-supervision of a master student (Jesús Lovón-Melgarejo).
- This work has been pursued with a one-year postdoctoral researcher I supervised, Thomas Gerald.

You can find below a list of related publications[4]:

- Jesús Lovón-Melgarejo, Laure Soulier, Karen Pinel-Sauvagnat, Lynda Tamine: Studying Catastrophic Forgetting in Neural Ranking Models. ECIR 2021: 375-390
  Code: `https://github.com/jeslev/OpenNIR-Lifelong`
- Thomas Gerald, Laure Soulier: Continual Learning of Long Topic Sequences in Neural Information Retrieval. ECIR 2022: 244-259
  Code: `https://github.com/tgeral68/continual_learning_of_long_topic`

[4]National publications are not mentioned if they are simple translations of international publications.

Part III

Other contributions and conclusion

Chapter 7

Other contributions

Contents

In this chapter, I briefly introduce other contributions (past and ongoing works) not discussed in the manuscript.

7.1 Past work: grounding textual representations

A research area that I have worked on between 2015 and 2019 is that of grounding textual representation through external knowledge. With early text embedding strategies (e.g., Gloves, word2vec, etc...) before large language models, several works have shown that textual embeddings do not capture all the semantics [Petroni *et al.* 2019]. One explanation of this limitation is the human reporting bias, i.e. we report in texts only key facts and not basic world knowledge acquired otherwise, leading to perception bias [Gordon & Van Durme 2013].

Hill *et al.* 2015 have shown that co-occurrence extraction leads to confusion between semantic similarity and conceptual relationships. For instance, the terms "bike" and "tire" will be close to the term "car" since they co-occur frequently although they are related differently to the term "car". "bike" is similar to "car" since they have the same functionality while "car" and "tire" have a functional relationship. With the same state of mind, Mrkšić *et al.* 2016 and Iacobacci *et al.* 2015 have also outlined that embeddings are not able to distinguish synonyms and antonyms.

We have therefore explored the potential of text grounding, aiming at anchoring textual representation in complementary resources. We have considered two types of resources: **knowledge resources** and **visual ones**. To do so, we have designed multi-modal representation learning models (often mid-fusion) aiming at leveraging the knowledge available either in images or knowledge bases to improve the semantics of text embeddings. Two types of evaluation have been conducted: intrinsic evaluation checking the quality of embeddings, and extrinsic evaluation analyzing the impact of such representations on NLP and IR tasks.

Outcomes

- 2 defended theses: Gia-Hung Nguyen (collaboration with Lynda Tamine and Nathalie Souf at IRIT) and Eloi Zablocki (collaboration with Benjamin Piwowarski and Patrick Gallinari).
- Participation to the MUSTER CHIST-ERA project (MUltimodal processing of Spatial and TEmporal expRessions)
- Publications on textual grounding with knowledge bases
  - Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, Nathalie Bricon-Souf: Toward a Deep Neural Approach for Knowledge-Based IR. Neur4IR workshop at SIGIR (2016)
  - Gia-Hung Nguyen, Laure Soulier, Lynda Tamine, Nathalie Bricon-Souf: DSRIM: A Deep Neural Information Retrieval Model Enhanced by a Knowledge Resource Driven Representation of Documents. ICTIR 2017
    Code: https://github.com/giahung24/dsrim
  - Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, Nathalie Souf: Learning Concept-Driven Document Embeddings for Medical Information Search. AIME 2017
  - Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, Nathalie Souf: A Tri-Partite Neural Document Language Model for Semantic Information Retrieval. ESWC 2018
  - Lynda Tamine, Laure Soulier, Gia-Hung Nguyen, Nathalie Souf: Offline versus Online Representation Learning of Documents Using External Knowledge. ACM Trans. Inf. Syst. 37(4): 42:1-42:34 (2019)
- Publication on textual grounding with images
  - Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, Patrick Gallinari: LIP6@CLEF2017: Multi-Modal Spatial Role Labeling using Word Embeddings. CLEF (Working Notes) 2017
  - Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, Patrick Gallinari: Learning Multi-Modal Word Representation Grounded in Visual Context. AAAI 2018
    Code: https://github.com/EloiZ/embedding_evaluation
  - Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, Patrick Gallinari: Context-Aware Zero-Shot Learning for Object Recognition. ICML 2019
  - Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, Patrick Gallinari: Incorporating Visual Semantics into Sentence Representations within a Grounded Space. EMNLP/IJCNLP 2019
    Code: https://github.com/pbordes/multimodal_sentence_rep
  - Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, Matthieu Cord: Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings. SIGIR 2018
    Code: https://github.com/Cadene/recipe1m.bootstrap.pytorch

7.2 On-going works

7.2.1 Domain adaptation and continual learning

As discussed in the previous chapter, domain adaptation is crucial for interactive neural models that face evolving trends and different users. From a more general point of view, human-machine collaboration settings in which an agent and a user interact to solve a particular task together are also constrained by adaptation issues, whether in terms of new environments or new interactions. For both neural models or reinforcement learning approaches, training an optimal model or an optimal policy able to generalize for all types of interactions/environments is complex. To tackle this issue, we are exploring the potential of **mode connectivity** [Benton *et al.* 2021, Kuditipudi *et al.* 2019] and more particularly the characteristics of neural subspaces [Wortsman *et al.* 2021], to exhibit interesting properties towards the generalization setting. These methods analyze the shape of the parameter space to build neural network subspaces. The latter contain diverse solutions (i.e., a set of model parameters) that process information differently. The intuition is that neural models in a subspace can be ensembled at the inference step, and having access to it instead of a single policy facilitates the adaptation without any cost of additional training. We propose two main works in this direction under the scope of reinforcement learning. First, we have addressed **neural subspace for reinforcement learning** switching the paradigm to subspaces of policies, instead of subspaces of neural networks. We have also demonstrated in a second contribution that subspaces of policies are well adapted **for continual reinforcement learning**.

We are now exploring **neural subspaces for information retrieval**, but the complexity is even harder since state-of-the-art models are all based on large language models. Given their large number of parameters, it is not reasonable to build a neural subspace including the set of all parameters in the encoder-decoder architecture. We are therefore exploring which parts of large language models can be considered for being included in the subspace and conducting experiments regarding zero-shot learning.

Outcomes These works have been initiated during the Ph.D. of Jean-Baptiste Gaya (Facebook CIFRE) and the NLP extension is done in collaboration with Thomas Gerald (now a postdoctoral researcher at LISN) and Pierre Erbacher (Ph.D. on the ANR JCJC SESAMS). Works focusing on reinforcement learning have led to two publications:

- Jean-Baptiste Gaya, Laure Soulier, Ludovic Denoyer: Learning a subspace of policies for online adaptation in Reinforcement Learning. ICLR 2022

- Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, Roberta Raileanu. Building a Subspace of Policies for Scalable Continual Learning. Deep RL workshop @NeurIPS 2023. Also under review for an international conference.

7.2.2 Contextual information extraction

Named Entity recognition (NER) and Relation Extraction (RE) can be seen as the reverse side of data-to-text generation with the objective to extract entities and their

relationships within a text in natural language (see the WebNLG challenge[1] addressing the data-to-text and text-to-data task). We consider two types of contexts:

- **The textual context**, i.e. the paragraph in which a entity occurs, under the assumption that the entity class can vary according to the context. Indeed, all current approaches [Liu *et al.* 2011, Liang *et al.* 2020, Souza *et al.* 2019] have a major drawback: they all consider an entity as a universal concept, linked to a single class, even if it may appear in different surface forms and contexts. This limits the potential of the information extracted which could be useful for more elaborated downstream tasks. As an example, *Amazon* will always be classified as a *company*, regardless of the context in which it is mentioned. But viewing this entity through the concepts of *seller*/*buyer* implies great differences in the way we perceive it and treat it. *Amazon* is likely to *sell* a product to an *individual person* but *buy* from another *company*. We therefore propose the Dynamic NER task in which the label of entity varies depending on the context. We define two datasets and an evaluation benchmark.
- **The multi-modal context**, i.e. the whole document, following the line of work combining textual and visual modalities so as to leverage document layout [Xu *et al.* 2019]. Our objective is to improve the multimodal fusion which is generally performed either at an early or a late stage, hindering their interaction throughout the learning process. We believe that it is crucial to jointly keep modality independent (to avoid error propagation that can be related, for instance, to the OCR) and let the possibility for the network to merge them when necessary.

Outcomes This work is conducted with Tristan Luiggi, a PhD student of CIFRE, co-supervised with Vincent Guigue.

- Tristan Luiggi, Laure Soulier, Vincent Guigue. Dynamic Named Entity Recognition. In: SAC 2023.

[1] https://webnlg-challenge.loria.fr/challenge_2020/

Chapter 8

Conclusion

Contents

8.1 Contributions and perspectives

In this manuscript, I have introduced our contributions focused on three main research axes dealing with relevance and faithfulness in text generation, contextualization of information needs, and neural ranking models adaptability in continual learning settings. In what follows, I sum up these contributions and the associated perspectives.

8.1.1 Towards faithful and relevant text generation

To ensure faithfulness and relevance in text generation, we have addressed two main challenges.

The first challenge focused on input encoding to capture data peculiarities related to the structure of the data. To do so, we focused on the data-to-text generation research domain and proposed a hierarchical data encoding aiming at representing entities separately before embedding the data as a whole. This encoding is surrounded by a hierarchical attention mechanism identifying first which entity is needed to be discussed and then which element is interesting for this entity. This model has a twofold contribution: it was the first work to both explicitly encode the data structure and use a transformer network for data-to-text generation.

The second challenge focused on the decoding process for which we have studied two use cases:

- In the data-to-text generation research domain, we addressed the pathological behavior of generation models that produce hallucinations due to the misalignment

of training data. This problem is also encountered in standard text-to-text generation tasks. However, it is even more challenging in data-to-text generation for several reasons: 1) the nature of input and output elements is different (e.g., numerical data vs. generated text), 2) the textual description might include reasoning over data (e.g., "player A has mastered the game" means that he scored the most points, implying a maximum calculus). Altogether, these task peculiarities hinder the semantic matching of the data input and the textual output, and accordingly highlight the difficulty to build a relevant semantic space bridging both modalities and decoding a faithful text. To tackle this issue, we have proposed two models based either on reinforcement learning (not introduced in the manuscript) or on a multi-branch decoder. The latter aims at separating during the decoding stage relevant and divergent textual information with respect to the input data constraint. Our motivation was to learn different decoder modules regarding three factors (fluency, content, and hallucination) to control the importance of each of them during the inference step.

- In the conversation search research field, we addressed the issue of query-driven text generation in which the difficulty relies on generating texts that are both faithful regarding the data input (in our case a list of documents) and the query (i.e., the information need). Guided by the constraint of the query to solve a complex information need, we have shown that the planning-based models are useful to guide the text generation process and produce structured and relevant texts.

Perspectives.
While several steps forwards have been done these last years toward faithful and relevant text generation, generative models can largely be improved. We present in what follows the different research directions, particularly related to data-to-text generation and conversational search, we envision for the future.

Numerical reasoning. In the data-to-text research field, one critical error that stands out is about numbers [Ji *et al.* 2022]. Current approaches [Puduppully *et al.* 2019b, Wiseman *et al.* 2017, Rebuffel *et al.* 2022] are generally effective in reporting values of tables and paraphrasing them. However, the data-to-text generation task is more complex: it often requires comparing values between them (e.g., identifying the best player or counting the difference in terms of point numbers between two teams in a basketball game) or to perform operations with abstractive concepts (e.g., estimating the number of days between a date and an event such as Christmas). While several works have been addressed in the Machine Reading Comprehension task [Dua *et al.* 2019, Herzig *et al.* 2020] or numerical reasoning tasks [Trask *et al.* 2018, d'Ascoli *et al.* 2022], they are often limited to simple numerical operations (such as sum or difference) and have not been envisioned in the perspective of data-to-text generation. Our ambition is to address numerical reasoning through different NLP tasks, such as question-answering [Dua *et al.* 2019, Herzig *et al.* 2020] or data-to-text generation [Puduppully *et al.* 2019a] including more complex data, e.g., time series of sensor data for weather presentation [Reiter *et al.* 2005]. We plan to integrate numerical executors [Andor *et al.* 2019, Pi *et al.* 2022] into language models so as to identify which parts of the input are relevant and how to combine them to build new knowledge in the generated output. Combined with the ability of large language models, we believe that it should allow to reason over (like-wise) numerical

information to improve the faithfulness of textual descriptions.

Personnalization of text generation. One way to ensure the relevance of text generation is to adapt the generation to the user and his/her intent. While several works have focused on style transfer [Ao *et al.* 2021, Chawla & Yang 2020, Malmi *et al.* 2020], personalizing data-to-text generation is not obvious. We believe that personalization is not only an issue of style but also a challenge of selecting the relevant content for the user. To the best of our knowledge, there is no work and no available dataset in this direction. To tackle these issues, we are currently building a dataset on movies including structured meta-data, reviews modeling users' interests, and personalized textual descriptions. We then plan to explore different techniques to personalize the data-to-text generation process by either using prompt-based language models [Yao *et al.* 2022] or injecting user profiles during decoding [Ao *et al.* 2021].

Model transferability. For specific text generation tasks, such as data-to-text generation, models are trained or fine-tuned on specific datasets (restaurants [Dušek *et al.* 2020], basketball games [Wiseman *et al.* 2017], ...). Their transferability to real use cases that can be encountered by company needs (e.g., summarizing financial information) is therefore limited. Although large language models have demonstrated great ability towards zero-shot adaptation [Devlin *et al.* 2019, Wei *et al.* 2022, Cui *et al.* 2022], we believe that they can reach some limitations in the data-to-text generation task. Indeed, depending on the structure/format of the input data, it can be difficult to understand their semantics from a zero-shot setting, and even more, identify salient information that should be decoded.

8.1.2 Contextualizing information needs expressed in natural language

Another research topic presented in this manuscript is the understanding of information needs in conversational search systems which are a core topic in the IR community since 2018 [Culpepper *et al.* 2018]. We focused on the conversation flow underlying the understanding the information need, either at a given conversation turn (as designed in TREC CAsT) or through a proactive interaction (as the query clarification task).

Through our participation in TREC CAsT, we have designed query reformulation models and contextual ranking models able to take into consideration the conversation to better represent the query. We have proposed the CoSPLADE model, a contextualized first-stage ranking model trained without the supervision of documents relevant in the conversation context. While this step is crucial, we also have addressed the query understanding issue as a proactive setting in which the system interacts with the user to understand and anticipate his/her information need. While several works [Aliannejadi *et al.* 2019, Zamani *et al.* 2020a] have addressed this task through a one-turn interaction, we proposed a framework simulating user-system interactions aiming at suggesting a set of query clarifications to the user who identifies the best one given his/her initial topic. The query clarification model is based on a diverse set of queries related to the initial topic which is re-ranked according to user's interactions. Experiments have shown the benefit of such a clarification process in the retrieval process.

Perspectives.
Having in mind that mixed-initiative are prevalent for conversational search, our perspectives focus on the query clarification task. One underlying challenge relies on the

fact that there is no dataset with jointly long-term interactions in natural language and a large amount of supervised data related to the IR tasks. We, therefore, envision two main challenges for query clarification.

Toward multi-turn query clarification simulation framework with interactions in natural language. As discussed in section 5.3, the next research issue concerns the interaction mode, through natural language allowing a more natural framework in which the user and the system interact. We, therefore, envision extending our simulation framework by integrating interactions in natural language instead of simply displaying queries for the IR agent and clicking on the best query for the user agent.

Lightweight domain adaptation of our query clarification simulation framework. Having in mind the deployment in production, we believe that a second challenge could be the transferability of our query clarification simulation framework to other domains/datasets. Guided by our previous simulation framework demonstrating the potential of query clarification to enhance ad-hoc IR settings and constrained by the fact that it exists only a single query clarification dataset [Aliannejadi *et al.* 2019], we plan to work on unsupervised domain adaptation strategies. For instance, by adapting the language model of our query clarification components through masked language modeling, we hope that after a few simulated interactions, the IR system would benefit from a clearer vision of the information needed to perform the retrieval step.

8.1.3 Investigating the ability of neural ranking models to continually adapt to evolving topics.

Assuming that IR models need to adapt to evolving users and/or topics, we investigated the continual learning research field and proposed a continual learning framework for iR modeling short and long topic sequences. We also analyzed the behavior of neural ranking models while fine-tuning successive tasks. We have compared transformer-based models with interaction-based models, highlighting different transferability levels and different abilities to face catastrophic forgetting. These works are the first ones to envision continual learning in IR and can serve as an evaluation framework for future works.

We plan two main future directions: pursuing our effort toward continual learning in IR and extending this work to intent detection in conversational systems.

Toward lifelong learning strategies adapted to ranking tasks. For IR, we plan to focus on the long topic sequence scenario, which is the most realistic one, and explore continual learning techniques for neural ranking models. In contrast to previous works in vision which mainly address classification tasks and in which catastrophic forgetting is highly noticeable, we are aware of the possible difficulty to adapt continual learning techniques for document scoring and acknowledge the unusual behavior of neural ranking models that show a small catastrophic forgetting in specific settings. Therefore, as shown in our preliminary experiments using EWC [Lovón-Melgarejo *et al.* 2021], continual learning techniques are promising and we envision adapting other strategies for ranking models. For instance, architecture-based approaches in computer vision [Cai *et al.* 2019, Veniat *et al.* 2020] propose to extend the network by integrating additional classes at the output level. In IR, the task is different, and the additional knowledge should rather be extended in the intermediate layers, focusing more on the learned embedding space than on the output. Having also in mind the outcomes of our exploratory analysis, we believe that such lifelong techniques must be integrated into

neural ranking models with an awareness level regarding the properties of evolving data. Said otherwise, neural ranking models robust to continual learning settings should include a modularization component tracking the critical changes in the training data and adapting accordingly the learning strategy.

Investigating continual learning for cross-lingual intent detection. For intent detection, we will address the limitation due to the language specificity for intent detection. If we desire to deploy virtual assistants all over the world, it is therefore important to design models able to address a large number of languages. Although multilingual models are a solution, it can be difficult to design a model trained simultaneously on all languages, particularly for under-resourced ones. We can therefore assume that the deployment of virtual assistants can be done step by step over different countries in the world and, thus, that virtual assistants will face different languages at different times. This assumption implies that, when designing/training a model for this task, languages can be incrementally added to the training procedure. In our case, we propose to explore a continual learning setting in which the task is fixed, but the stream is based on different languages. The model, therefore, learns the knowledge of language peculiarities. To satisfy the initial condition of virtual assistants to address different languages, we need to ensure that our task-based model does not forget previous languages while training on new ones.

8.2 Future research directions

8.2.1 Retrieval-augmented Machine Learning

The majority of neural models for NLP or Machine Learning are based on the assumption that all knowledge and reasoning required for the task are captured by parameters. Large language models have demonstrated that increasing the number of parameters generally leads to performance increases. However, this strategy focused on parameter size is not scalable, and accordingly not desired in terms of computational cost. A recent research paper [Zamani *et al.* 2022a] has discussed the potential of enhancing neural models with IR systems. The intuition is to couple neural models with IR systems to access and reason over large text corpora and knowledge stores with the final objective to reduce the number of parameters in neural models and improve their scalability. This strategy has already been used for pseudo-relevance feedback in IR [Croft & Harper 1997], question-answering [Gao *et al.* 2022, Hsu *et al.* 2021] or, more recently, to train language models [Guu *et al.* 2020]. Beyond scalability, the authors also argue that accessing external knowledge through IR systems has several merits: 1) improving the generalization performance of the model, 2) being more robust to information updates and temporal changes, and 3) grounding model decisions with external knowledge leading to more interpretability and explainability.

We believe that this paradigm deserves attention and is interesting to revisit the different research fields addressed in this manuscript.

- For the data-to-text generation task, the IR system can thus serve as anchor sources for knowledge grounding for both in-domain and out-of-domain datasets. For in-domain, it is worth reminding that the data structure is not fully explicit in terms of semantics (e.g., for a table, columns can be abbreviated, and values are of different

formats...). Therefore, understanding the semantics of the data structure might be difficult. In addition, in many situations, the expected decision relies on implicit information directly related to the application. For example, in NBA games, the action "passing" implies that passes are only made between players of the same team. Therefore, when player A is identified as belonging to the winning team and passes to player B, B is de facto the winner. Conversely, intercepting a pass from A would assign B to the losing team. For the out-of-domain case, the difficulty is greater because, beyond remaining in-domain challenges, the model needs to capture the semantics of another domain than the one used for its training. We can believe that using a retrieval module can help to face a larger knowledge over different application domains and, thus, improve the generalization performance.

- For the interactive information retrieval task, both query clarification and continual learning settings can benefit from IR-augmented models, but for different reasons. For query clarification, the difficulty lies in the diversification of suggestions. Therefore, retrieval-augmented models might help in grounding the initial query and therefore in suggesting different facets or orthogonal topics to improve the search process. For the continual learning setting, a retrieval-augmented system might be beneficial to identify samples characterizing information updates or temporal changes or to interpolate new knowledge. The component can therefore serve as a replay buffer for instance for rehearsal strategies.

8.2.2 Language-augmented Robotics

With the recent affiliation of the MLIA team to the Robotic Laboratory (ISIR) of Sorbonne University, new research axes have been discussed and promise tremendous collaborations. One of the research challenges we are planning to address in the following years is to improve reinforcement learning models for robotics with natural language processing. Autonomous agents require reasoning and planning strategies for performing tasks. We, therefore, believe that the semantics captured by large language models can enhance the decision process at different levels.

First, it can allow grounding object representations with common sense to identify their intrinsic and actionable properties. Large language models and also common sense knowledge bases, such as ConceptNet[1], can be used as complementary information sources, implying to design representation models leveraging multi-modal information. The difficulty would be to identify which properties are relevant for objects and how to fuse them into a single representation. Another strategy can be to encode objects differently according to each modality and then use self-attention to learn the possible interactions that are relevant for the task solving. Object grounding has been addressed in [Sridharan & Mota 2022, Tsiami *et al.* 2018], but we believe that the grounding needs to be extended to the scene to better model the context and object properties.

Second, natural language can serve for building and clarifying the planning strategy, and therefore the actions done by a robot. Several works have addressed instruction identification as abstract representation [Andreas *et al.* 2018, Jacob *et al.* 2021] or natural language expression [Sharma *et al.* 2022], but the limited data supervision is often

[1] https://conceptnet.io/

a challenge [Chen *et al.* 2020a, Sharma *et al.* 2022]. To tackle this issue, we envision interactive training processes, implying asking humans to label situations with sentences, with strong care on limiting interactions to a few relevant situations, to reduce human effort. The challenge consists in defining when to interact with real users in the planning and which information asking to increase the supervision data. One can imagine a policy combined with a language model to 1) identify whether to generate the following instructions or to ask humans about the next instruction. This decision can be taken by evaluating whether the language model has sufficient knowledge to capture all the semantics of the current scene (e.g., through a task-guided masked language modeling loss), 2) leverage the language model to interact with humans if necessary, and 3) generate the next instruction according to the scene, the state of the policy and the interaction with humans.

8.3 My last words

Writing this manuscript allowed me to gather different research fields (data-to-text generation, conversational search, continual learning) that are, on a daily basis, addressed independently. They all rely on the design of neural models or leverage language models. Beyond this, I think that there is a synergy in all my research. While NLP techniques and language models are able to exploit a large amount of human knowledge to capture the meaning of textual data, IR models allow the retrieval of knowledge from large databases, and planning techniques are well known to consider complex behavior and reasoning at the machine level. The crossroad of these research fields conducts me to design step-by-step components of relevant and robust human-machine collaboration systems.

8.4 Acknowledgements

The work presented in this manuscript is founded thanks to different grants:

- ANR JCJC SESAMS (ANR-18-CE23-0001) - conversational search and continual learning
- CIFRE BNP Parisbas - data-to-text generation

Bibliography

[Agichtein *et al.* 2006] Eugene Agichtein, Eric Brill and Susan Dumais. *Improving Web Search Ranking by Incorporating User Behavior Information.* In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, pages 19–26, 2006.

[Agrawal *et al.* 2009] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson and Samuel Ieong. *Diversifying Search Results.* In Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09, page 5–14, New York, NY, USA, 2009. Association for Computing Machinery.

[Aliannejadi *et al.* 2019] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani and W. Bruce Croft. *Asking Clarifying Questions in Open-Domain Information-Seeking Conversations.* In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, page 475–484, New York, NY, USA, 2019. Association for Computing Machinery.

[Aliannejadi *et al.* 2021] Mohammad Aliannejadi, Leif Azzopardi, Hamed Zamani, Evangelos Kanoulas, Paul Thomas and Nick Craswell. *Analysing Mixed Initiatives and Search Strategies during Conversational Search.* In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang and Hanghang Tong, editors, CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, pages 16–26. ACM, 2021.

[Aljundi *et al.* 2019] Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin and Lucas Page-Caccia. *Online Continual Learning with Maximal Interfered Retrieval.* In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 11849–11860, 2019.

[Amati & Van Rijsbergen 2002] Gianni Amati and Cornelis Joost Van Rijsbergen. *Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness.* ACM Trans. Inf. Syst., vol. 20, no. 4, page 357–389, oct 2002.

[Anand *et al.* 2020] Avishek Anand, Lawrence Cavedon, Matthias Hagen, Hideo Joho, Mark Sanderson and Benno Stein. *Conversational Search - A Report from Dagstuhl Seminar 19461.* CoRR, vol. abs/2005.08658, 2020.

[Andor *et al.* 2019] Daniel Andor, Luheng He, Kenton Lee and Emily Pitler. *Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension.* In Kentaro Inui, Jing Jiang, Vincent Ng and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 5946–5951. Association for Computational Linguistics, 2019.

[Andreas *et al.* 2018] Jacob Andreas, Dan Klein and Sergey Levine. *Learning with Latent Language*. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2166–2179, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Ao *et al.* 2021] Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He and Xing Xie. *PENS: A Dataset and Generic Framework for Personalized News Headline Generation*. 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), pages 82–92, Online, August 2021. Association for Computational Linguistics.

[Arabzadeh & Clarke 2020] Negar Arabzadeh and Charles L. A. Clarke. *WaterlooClarke at the Trec 2020 Conversational Assistant Track*. 2020.

[Arabzadeh *et al.* 2021] Negar Arabzadeh, Xinyi Yan and Charles L. A. Clarke. *Predicting Efficiency/Effectiveness Trade-Offs for Dense vs. Sparse Retrieval Strategy Selection*. In Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management, CIKM '21, page 2862–2866, New York, NY, USA, 2021. Association for Computing Machinery.

[Asghar *et al.* 2020] Nabiha Asghar, Lili Mou, Kira A. Selby, Kevin D. Pantasdo, Pascal Poupart and Xin Jiang. *Progressive Memory Banks for Incremental Domain Adaptation*. In ICLR, volume abs/1811.00239, 2020.

[Azzopardi 2014] Leif Azzopardi. *Modelling interaction with economic models of search*. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke and Kalervo Järvelin, editors, The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014, pages 3–12. ACM, 2014.

[Bahdanau *et al.* 2015] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[Bai *et al.* 2020] Yang Bai, Xiaoguang Li, Gang Wang, Chaoliang Zhang, Lifeng Shang, Jun Xu, Zhaowei Wang, Fangshan Wang and Qun Liu. *SparTerm: Learning Term-based Sparse Representation for Fast Text Retrieval*. 2020.

[Balog *et al.* 2009] Krisztian Balog, Leif Azzopardi and Maarten de Rijke. *A language modeling framework for expert finding*. Inf. Process. Manag., vol. 45, no. 1, pages 1–19, 2009.

[Banerjee & Lavie 2005] Satanjeev Banerjee and Alon Lavie. *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments*. pages 65–72. Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005.

[Belkin *et al.* 1982] Nicholas J Belkin, Robert N Oddy and Helen M Brooks. *ASK for information retrieval: Part I. Background and theory.* Journal of documentation, 1982.

[Beltagy *et al.* 2019] Iz Beltagy, Kyle Lo and Arman Cohan. *SciBERT: A pretrained language model for scientific text.* arXiv preprint arXiv:1903.10676, 2019.

[Beltagy *et al.* 2020] Iz Beltagy, Matthew E. Peters and Arman Cohan. *Longformer: The Long-Document Transformer.* CoRR, vol. abs/2004.05150, 2020.

[Bender *et al.* 2021] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* In Madeleine Clare Elish, William Isaac and Richard S. Zemel, editors, FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, pages 610–623. ACM, 2021.

[Bengio *et al.* 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent and Christian Janvin. *A Neural Probabilistic Language Model.* J. Mach. Learn. Res., vol. 3, pages 1137–1155, March 2003.

[Bennett *et al.* 2012] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk and Xiaoyuan Cui. *Modeling the impact of short- and long-term behavior on search personalization.* In SIGIR '12, 2012.

[Benton *et al.* 2021] Gregory W. Benton, Wesley Maddox, Sanae Lotfi and Andrew Gordon Wilson. *Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling.* In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, pages 769–779. PMLR, 2021.

[Bi *et al.* 2021] Keping Bi, Qingyao Ai and W. Bruce Croft. *Learning a Fine-Grained Review-based Transformer Model for Personalized Product Search.* In SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, pages 123–132. ACM, 2021.

[Bojanowski *et al.* 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov. *Enriching Word Vectors with Subword Information.* Transactions of the Association for Computational Linguistics, vol. 5, pages 135–146, 2017.

[Bommasani *et al.* 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill*et al. On the opportunities and risks of foundation models.* arXiv preprint arXiv:2108.07258, 2021.

[Bordes & Weston 2016] Antoine Bordes and Jason Weston. *Learning End-to-End Goal-Oriented Dialog.* CoRR, vol. abs/1605.07683, 2016.

[Bordes *et al.* 2014] Antoine Bordes, Sumit Chopra and Jason Weston. *Question Answering with Subgraph Embeddings.* In EMNLP, pages 615–620, 2014.

[Brown *et al.* 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell*et al. Language models are few-shot learners.* Advances in neural information processing systems, vol. 33, pages 1877–1901, 2020.

[Cai *et al.* 2014] Fei Cai, Shangsong Liang and Maarten de Rijke. *Time-Sensitive Personalized Query Auto-Completion.* In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14, page 1599–1608, New York, NY, USA, 2014. Association for Computing Machinery.

[Cai *et al.* 2016] Fei Cai, Ridho Reinanda and Maarten De Rijke. *Diversifying Query Auto-Completion.* ACM Trans. Inf. Syst., vol. 34, no. 4, jun 2016.

[Cai *et al.* 2019] Hengyi Cai, Hongshen Chen, Cheng Zhang, Yonghao Song, Xiaofang Zhao and Dawei Yin. *Adaptive Parameterization for Neural Dialogue Generation.* In EMNLP-IJCNLP, pages 1793–1802, November 2019.

[Câmara & Hauff 2020] Arthur Câmara and Claudia Hauff. *Diagnosing BERT with Retrieval Heuristics.* In ECIR 2020 Part I, pages 605–618, 2020.

[Câmara *et al.* 2022] Arthur Câmara, David Maxwell and Claudia Hauff. *Searching, Learning, and Subtopic Ordering: A Simulation-based Analysis.* CoRR, vol. abs/2201.11181, 2022.

[Carbonell & Goldstein 1998] Jaime Carbonell and Jade Goldstein. *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries.* In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, page 335–336, New York, NY, USA, 1998. Association for Computing Machinery.

[Carterette *et al.* 2014] Ben Carterette, Evangelos Kanoulas, Mark M. Hall and Paul D. Clough. *Overview of the TREC 2014 Session Track.* In Ellen M. Voorhees and Angela Ellis, editors, Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014, volume 500-308 of *NIST Special Publication.* National Institute of Standards and Technology (NIST), 2014.

[Castro *et al.* 2018] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid and Karteek Alahari. *End-to-End Incremental Learning.* In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu and Yair Weiss, editors, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII, volume 11216 of *Lecture Notes in Computer Science*, pages 241–257. Springer, 2018.

[Cattan *et al.* 2022] Oralie Cattan, Christophe Servan and Sophie Rosset. *On the cross-lingual transferability of multilingual prototypical models across NLU tasks.* CoRR, vol. abs/2207.09157, 2022.

[Chaudhry *et al.* 2019] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr and Marc'Aurelio Ranzato. *Continual Learning with Tiny Episodic Memories.* CoRR, vol. abs/1902.10486, 2019.

[Chawla & Yang 2020] Kunal Chawla and Diyi Yang. *Semi-supervised Formality Style Transfer using Language Model Discriminator and Mutual Information Maximization.* In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2340–2354, Online, November 2020. Association for Computational Linguistics.

[Chen & Liu 2018] Zhiyuan Chen and Bing Liu. Lifelong machine learning, second edition. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2018.

[Chen & Mooney 2008] David L. Chen and Raymond J. Mooney. *Learning to Sportscast: A Test of Grounded Language Acquisition.* In Proceedings of the 25th International Conference on Machine Learning, ICML '08, pages 128–135, New York, NY, USA, 2008. ACM.

[Chen *et al.* 2015] Zhiyuan Chen, Nianzu Ma and Bing Liu. *Lifelong Learning for Sentiment Classification.* In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 750–756, Beijing, China, July 2015. Association for Computational Linguistics.

[Chen *et al.* 2020a] Haonan Chen, Hao Tan, Alan Kuntz, Mohit Bansal and Ron Alterovitz. *Enabling Robots to Understand Incomplete Natural Language Instructions Using Commonsense Reasoning.* In 2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020, pages 1963–1969. IEEE, 2020.

[Chen *et al.* 2020b] Jerry Zikun Chen, Shih Yuan Yu and Haoran Wang. *Exploring Fluent Query Reformulations with Text-to-Text Transformers and Reinforcement Learning.* ArXiv, vol. abs/2012.10033, 2020.

[Chen *et al.* 2020c] Limin Chen, Zhiwen Tang and Grace Hui Yang. *Balancing Reinforcement Learning Training Experiences in Interactive Information Retrieval.* In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen and Yiqun Liu, editors, Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pages 1525–1528. ACM, 2020.

[Chen *et al.* 2021] Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang and William W. Cohen. *Open Question Answering over Tables and Text.* In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[Chiang *et al.* 2020] Cheng-Han Chiang, Sung-Feng Huang and Hung-yi Lee. *Pretrained language model embryology: The birth of ALBERT.* arXiv preprint arXiv:2010.02480, 2020.

[Cho *et al.* 2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.* In EMNLP, pages 1724–1734, 2014.

[Chowdhery *et al.* 2022a] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann*et al. Palm: Scaling language modeling with pathways.* arXiv preprint arXiv:2204.02311, 2022.

[Chowdhery *et al.* 2022b] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov and Noah Fiedel. *PaLM: Scaling Language Modeling with Pathways.* CoRR, vol. abs/2204.02311, 2022.

[Clark *et al.* 2020] Kevin Clark, Minh-Thang Luong, Quoc V Le and Christopher D Manning. *Pre-training transformers as energy-based cloze models.* arXiv preprint arXiv:2012.08561, 2020.

[Cohen *et al.* 2018] Daniel Cohen, Bhaskar Mitra, Katja Hofmann and W. Bruce Croft. *Cross Domain Regularization for Neural Ranking Models Using Adversarial Learning.* In ACM SIGIR, May 2018.

[Collobert *et al.* 2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu and Pavel Kuksa. *Natural language processing (almost) from scratch.* Journal of machine learning research, vol. 12, no. ARTICLE, pages 2493–2537, 2011.

[Croft & Harper 1997] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information, page 339–344. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[Cronen-Townsend & Croft 2002] Steve Cronen-Townsend and W. Bruce Croft. *Quantifying Query Ambiguity.* In Proceedings of the Second International Conference on Human Language Technology Research, HLT '02, page 104–109, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[Cui *et al.* 2022] Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar and Aravind Rajeswaran. *Can Foundation Models Perform Zero-Shot Task Specification For Robot Manipulation?* In Learning for Dynamics and Control Conference, L4DC 2022, 23-24 June 2022, Stanford University, Stanford, CA, USA, volume 168 of *Proceedings of Machine Learning Research*, pages 893–905. PMLR, 2022.

[Culpepper *et al.* 2018] J. Shane Culpepper, Fernando Diaz and Mark D. Smucker. *Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018).* SIGIR Forum, no. 1, pages 34–90, 2018.

[Dai & Callan 2020] Zhuyun Dai and Jamie Callan. *Context-Aware Term Weighting For First Stage Passage Retrieval.* In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1533–1536. ACM, 2020.

[Dai *et al.* 2015] Andrew M Dai, Christopher Olah and Quoc V Le. *Document embedding with paragraph vectors.* arXiv preprint arXiv:1507.07998, 2015.

[Dai *et al.* 2019] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le and Ruslan Salakhutdinov. *Transformer-xl: Attentive language models beyond a fixed-length context.* arXiv preprint arXiv:1901.02860, 2019.

[Dai *et al.* 2022] Yuqian Dai, Marc de Kamps and Serge Sharoff. *BERTology for Machine Translation: What BERT Knows about Linguistic Difficulties for Translation.* In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 6674–6690, 2022.

[Dalton *et al.* 2020a] Jeffrey Dalton, Chenyan Xiong and Jamie Callan. *CAsT 2020: The Conversational Assistance Track Overview.* vol. 1266, 2020.

[Dalton *et al.* 2020b] Jeffrey Dalton, Chenyan Xiong and Jamie Callan. *TREC CAsT 2019: The Conversational Assistance Track Overview.* CoRR, vol. abs/2003.13624, 2020.

[Dalton *et al.* 2021] Jeffrey Dalton, Chenyan Xiong and Jamie Callan. *TREC CAsT 2021: The Conversational Assistance Track Overview.* page 7, 2021.

[d'Ascoli *et al.* 2022] Stéphane d'Ascoli, Pierre-Alexandre Kamienny, Guillaume Lample and François Charton. *Deep symbolic regression for recurrence prediction.* In International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, pages 4520–4536, 2022.

[Davidson & Mozer 2020] Guy Davidson and Michael C. Mozer. *Sequential Mastery of Multiple Visual Tasks: Networks Naturally Learn to Learn and Forget to Forget.* In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 9279–9290. Computer Vision Foundation / IEEE, 2020.

[de Masson d'Autume *et al.* 2019] Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong and Dani Yogatama. *Episodic Memory in Lifelong Language Learning.* CoRR, vol. abs/1906.01076, 2019.

[Dean-Hall *et al.* 2014] Adriel Dean-Hall, Charles L. A. Clarke, Jaap Kamps, Paul Thomas and Ellen M. Voorhees. *Overview of the TREC 2014 Contextual Suggestion Track.* In Ellen M. Voorhees and Angela Ellis, editors, Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014, volume 500-308 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2014.

[Deng *et al.* 2013] Dong Deng, Yu Jiang, Guoliang Li, Jian Li and Cong Yu. *Scalable column concept determination for web tables using large knowledge bases.* Proceedings of the VLDB Endowment, vol. 6, no. 13, pages 1606–1617, August 2013.

[Devlin *et al.* 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, June 2019.

[Dhingra *et al.* 2019] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das and William Cohen. *Handling Divergent Reference Texts when Evaluating Table-to-Text Generation.* In ACL, 2019.

[Dietz *et al.* 2017] Laura Dietz, Manisha Verma, Filip Radlinski and Nick Craswell. *TREC Complex Answer Retrieval Overview.* TREC, 2017.

[Dietz *et al.* 2018] Laura Dietz, Manisha Verma, Filip Radlinski and Nick Craswell. *TREC Complex Answer Retrieval Overview.* In TREC, 2018.

[Djeddal *et al.* 2022] Hanane Djeddal, Thomas Gerald, Laure Soulier, Karen Pinel-Sauvagnat and Lynda Tamine. *Does Structure Matter? Leveraging Data-to-Text Generation for Answering Complex Information Needs.* In Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II, volume 13186 of *Lecture Notes in Computer Science*, pages 93–101. Springer, 2022.

[Douillard *et al.* 2020a] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert and Eduardo Valle. *PODNet: Pooled Outputs Distillation for Small-Tasks Incremental Learning.* In Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX, volume 12365 of *Lecture Notes in Computer Science*, pages 86–102. Springer, 2020.

[Douillard *et al.* 2020b] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert and Eduardo Valle. *Small-Task Incremental Learning.* In ECCV, 2020.

[Douillard 2022] Arthur Douillard. *Continual Learning for Computer Vision. (Apprentissage continu pour la vision par ordinateur).* PhD thesis, Sorbonne University, Paris, France, 2022.

[Dua *et al.* 2019] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh and Matt Gardner. *DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs.* In Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Dusek *et al.* 2019] Ondrej Dusek, David M. Howcroft and Verena Rieser. *Semantic Noise Matters for Neural Natural Language Generation.* In INLG, 2019.

[Dušek *et al.* 2020] Ondřej Dušek, Jekaterina Novikova and Verena Rieser. *Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge.* Computer Speech & Language, vol. 59, pages 123–156, January 2020.

[Elgohary *et al.* 2019] Ahmed Elgohary, Denis Peskov and Jordan Boyd-Graber. *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), pages 5918–5924, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Ellis 1989] David Ellis. *A behavioural model for information retrieval system design.* Journal of Information Science, vol. 15, no. 4-5, pages 237–247, 1989.

[Elsahar *et al.* 2021] Hady Elsahar, Maximin Coavoux, Jos Rozen and Matthias Gallé. *Self-Supervised and Controlled Multi-Document Opinion Summarization.* In Paola Merlo, Jörg Tiedemann and Reut Tsarfaty, editors, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, pages 1646–1662. Association for Computational Linguistics, 2021.

[Erbacher *et al.* 2022] Pierre Erbacher, Ludovic Denoyer and Laure Soulier. *Interactive Query Clarification and Refinement via User Simulation.* In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper and Gabriella Kazai, editors, SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, pages 2420–2425. ACM, 2022.

[Fan *et al.* 2022] Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang and Jiafeng Guo. *Pre-training Methods in Information Retrieval.* Found. Trends Inf. Retr., vol. 16, no. 3, pages 178–317, 2022.

[Fernando *et al.* 2017] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel and Daan Wierstra. *PathNet: Evolution Channels Gradient Descent in Super Neural Networks.* CoRR, vol. abs/1701.08734, 2017.

[Ficler & Goldberg 2017] Jessica Ficler and Yoav Goldberg. *Controlling Linguistic Style Aspects in Neural Language Generation.* In Workshop on Stylistic Variation @ ACL, 2017.

[Filippova 2020] Katja Filippova. *Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data.* In Findings of EMNLP, 2020.

[Firth 1957] J. R. Firth. *A synopsis of linguistic theory 1930–55. In Studies in linguistic analysis*, 1957.

[Flanigan *et al.* 2022] Jeffrey Flanigan, Ishan Jindal, Yunyao Li, Tim O'Gorman and Martha Palmer. *Meaning Representations for Natural Languages: Design, Models and Applications.* In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022. ACL, 2022.

[Flesch 1962] R. Flesch. The art of readable writing. Wiley, 1962.

[Florance & Marchionini 1995] Valerie Florance and Gary Marchionini. *Information Processing in the Context of Medical Care.* In Edward A. Fox, Peter Ingwersen and Raya Fidel, editors, SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum), pages 158–163. ACM Press, 1995.

[Formal *et al.* 2021] Thibault Formal, Benjamin Piwowarski and Stéphane Clinchant. *SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking.* In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, pages 2288–2292, New York, NY, USA, July 2021. Association for Computing Machinery.

[Formal *et al.* 2022] Thibault Formal, Carlos Lassance, Benjamin Piwowarski and Stéphane Clinchant. *From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective.* In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, pages 2353–2359, New York, NY, USA, July 2022. Association for Computing Machinery.

[Gao *et al.* 2022] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu and Prem Natarajan. *Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering.* In CVPR 2022, 2022.

[Garcia *et al.* 2021] Xavier Garcia, Noah Constant, Ankur P. Parikh and Orhan Firat. *Towards Continual Learning for Multilingual Machine Translation via Vocabulary Substitution.* In NAACL-HLT, pages 1184–1192, 2021.

[Gatt & Krahmer 2018] Albert Gatt and Emiel Krahmer. *Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation.* J. Artif. Int. Res., vol. 61, no. 1, page 65–170, jan 2018.

[Gehring *et al.* 2017] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats and Yann N. Dauphin. *Convolutional Sequence to Sequence Learning.* In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 1243–1252. JMLR.org, 2017.

[Gerald & Soulier 2022] Thomas Gerald and Laure Soulier. *Continual Learning of Long Topic Sequences in Neural Information Retrieval.* In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg and

Vinay Setty, editors, Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I, volume 13185 of *Lecture Notes in Computer Science*, pages 244–259. Springer, 2022.

[Gordon & Van Durme 2013] Jonathan Gordon and Benjamin Van Durme. *Reporting Bias and Knowledge Acquisition.* In Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, AKBC '13, pages 25–30, New York, NY, USA, 2013. ACM.

[Goyal *et al.* 2016] Raghav Goyal, Marc Dymetman and Éric Gaussier. *Natural Language Generation through Character-based RNNs with Finite-state Prior Knowledge.* In Nicoletta Calzolari, Yuji Matsumoto and Rashmi Prasad, editors, COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, pages 1083–1092. ACL, 2016.

[Graves *et al.* 2014] Alex Graves, Greg Wayne and Ivo Danihelka. *Neural turing machines.* arXiv preprint arXiv:1410.5401, 2014.

[Grosz *et al.* 1995] Barbara Grosz, Aravind Joshi and Scott Weinstein. *Centering: A Framework for Modelling the Coherence of Discourse.* Technical Reports (CIS), 01 1995.

[Gruetzemacher & Paradice 2022] Ross Gruetzemacher and David Paradice. *Deep Transfer Learning amp; Beyond: Transformer Language Models in Information Systems Research.* ACM Comput. Surv., vol. 54, no. 10s, sep 2022.

[Gu *et al.* 2016] Jiatao Gu, Zhengdong Lu, Hang Li and Victor O. K. Li. *Incorporating Copying Mechanism in Sequence-to-Sequence Learning.* In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics, 2016.

[Gu *et al.* 2021a] Xiaodong Gu, Kang Min Yoo and Jung-Woo Ha. *DialogBERT: Discourse-Aware Response Generation via Learning to Recover and Rank Utterances.* In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 12911–12919. AAAI Press, 2021.

[Gu *et al.* 2021b] Ziwei Gu, Jing Nathan Yan and Jeffrey M. Rzeszotarski. *Understanding User Sensemaking in Machine Learning Fairness Assessment Systems.* In Proceedings of the Web Conference 2021, WWW '21, page 658–668, New York, NY, USA, 2021. Association for Computing Machinery.

[Gulcehre *et al.* 2016] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou and Yoshua Bengio. *Pointing the Unknown Words.* In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1:

Long Papers), pages 140–149, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Guo *et al.* 2016] Jiafeng Guo, Yixing Fan, Qingyao Ai and W. Bruce Croft. *A Deep Relevance Matching Model for Ad-hoc Retrieval.* In Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016, pages 55–64. ACM, 2016.

[Guu *et al.* 2020] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat and Ming-Wei Chang. *Retrieval Augmented Language Model Pre-Training.* In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR, 2020.

[Hai *et al.* 2023] Nam Le Hai, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarski and Laure Soulier. *CoSPLADE: Contextualizing SPLADE for Conversational Information Retrieval.* In Advances in Information Retrieval - 45nd European Conference on IR Research, ECIR 2023, Lecture Notes in Computer Science. Springer, 2023.

[Harris 1954] Zellig S Harris. *Distributional structure.* Word, vol. 10, no. 2-3, pages 146–162, 1954.

[Harvey *et al.* 2013] Morgan Harvey, Fabio A. Crestani and Mark James Carman. *Building user profiles from topic models for personalised search.* Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 2013.

[Haug *et al.* 2018] Till Haug, Octavian-Eugen Ganea and Paulina Grnarova. *Neural Multi-step Reasoning for Question Answering on Semi-structured Tables.* In Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings, pages 611–617, 2018.

[Hersh & Over 2001] William R. Hersh and Paul Over. *The TREC 2001 Interactive Track Report.* In Ellen M. Voorhees and Donna K. Harman, editors, Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001, volume 500-250 of *NIST Special Publication.* National Institute of Standards and Technology (NIST), 2001.

[Herzig *et al.* 2020] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno and Julian Eisenschlos. *TaPas: Weakly Supervised Table Parsing via Pre-training.* In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4320–4333, Online, July 2020. Association for Computational Linguistics.

[Hill *et al.* 2014] Felix Hill, Roi Reichart and Anna Korhonen. *Multi-modal models for concrete and abstract concept meaning.* Transactions of the Association for Computational Linguistics, vol. 2, pages 285–296, 2014.

[Hill *et al.* 2015] Felix Hill, Roi Reichart and Anna Korhonen. *SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation.* Computational Linguistics, vol. 41, no. 4, pages 665–695, 2015.

[Hoffmann *et al.* 2011] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer and Daniel S. Weld. *Knowledge-based Weak Supervision for Information Extraction of Overlapping Relations.* In Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, pages 541–550, 2011.

[Hofstätter *et al.* 2020] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan and Allan Hanbury. *Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation.* ArXiv, vol. abs/2010.02666, 2020.

[Hofstätter *et al.* 2021] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin and Allan Hanbury. *Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling.* In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones and Tetsuya Sakai, editors, SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, pages 113–122. ACM, 2021.

[Holtzman *et al.* 2019] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes and Yejin Choi. *The curious case of neural text degeneration.* arXiv preprint arXiv:1904.09751, 2019.

[Hsu *et al.* 2021] Chao-Chun Hsu, Eric Lind, Luca Soldaini and Alessandro Moschitti. *Answer Generation for Retrieval-based Question Answering Systems.* In Chengqing Zong, Fei Xia, Wenjie Li and Roberto Navigli, editors, Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4276–4282. Association for Computational Linguistics, 2021.

[Huang *et al.* 2018] Liang Huang, Kai Zhao and Mingbo Ma. *When to finish? optimal beam search for neural text generation (modulo beam size).* arXiv preprint arXiv:1809.00069, 2018.

[Hui *et al.* 2017] Kai Hui, Andrew Yates, Klaus Berberich and Gerard de Melo. *Pacrr: A position-aware neural ir model for relevance matching.* arXiv preprint arXiv:1704.03940, 2017.

[Huo & Zyl 2020] Jiahao Huo and Terence L van Zyl. *Comparative Analysis of Catastrophic Forgetting in Metric Learning.* In 2020 7th International Conference on Soft Computing Machine Intelligence (ISCMI), pages 68–72, 2020.

[Iacobacci *et al.* 2015] Ignacio Iacobacci, Mohammad Taher Pilehvar and Roberto Navigli. *SensEmbed: Learning Sense Embeddings for Word and Relational Similarity.* In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 95–105, Beijing, China, July 2015. Association for Computational Linguistics.

[Iida *et al.* 2021] Hiroshi Iida, Dung Thai, Varun Manjunatha and Mohit Iyyer. *TABBIE: Pretrained Representations of Tabular Data.* In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pages 3446–3456. Association for Computational Linguistics, 2021.

[Jacob *et al.* 2021] Athul Paul Jacob, Mike Lewis and Jacob Andreas. *Multitasking Inhibits Semantic Drift.* In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5351–5366, Online, June 2021. Association for Computational Linguistics.

[Jansen *et al.* 2000] Bernard J. Jansen, Amanda Spink and Tefko Saracevic. *Real life, real users, and real needs: A study and analysis of user queries on the Web.* Information Processing and Management, vol. 36, no. 2, pages 207–227, March 2000.

[Ji *et al.* 2022] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto and Pascale Fung. *Survey of Hallucination in Natural Language Generation.* CoRR, vol. abs/2202.03629, 2022.

[Juraska *et al.* 2018] Juraj Juraska, Panagiotis Karagiannis, Kevin K. Bowden and Marilyn A. Walker. *A deep ensemble model with slot alignment for sequence-To-sequence natural language generation.* NAACL-HLT, 2018.

[Kahou *et al.* 2018] S. Kahou, A. Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler and Yoshua Bengio. *FigureQA: An Annotated Figure Dataset for Visual Reasoning.* ArXiv, vol. abs/1710.07300, 2018.

[Kanoulas *et al.* 2018] Evangelos Kanoulas, Leif Azzopardi and Grace Hui Yang. *Overview of the CLEF Dynamic Search Evaluation Lab 2018.* In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian-Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato and Nicola Ferro, editors, Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings, volume 11018 of *Lecture Notes in Computer Science*, pages 362–371. Springer, 2018.

[Kenter & De Rijke 2015] Tom Kenter and Maarten De Rijke. *Short text similarity with word embeddings.* In Proceedings of the 24th ACM international on conference on information and knowledge management, pages 1411–1420, 2015.

[Khan *et al.* 2022] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan and Mubarak Shah. *Transformers in vision: A survey.* ACM computing surveys (CSUR), vol. 54, no. 10s, pages 1–41, 2022.

[Khattab & Zaharia 2020] Omar Khattab and Matei Zaharia. *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT.* In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pages 39–48. ACM, 2020.

[Kiela 2022] Douwe Kiela. *Grounding, Meaning and Foundation Models: Adventures in Multimodal Machine Learning.* In MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022, page 5. ACM, 2022.

[Kiesel *et al.* 2018] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand and Matthias Hagen. *Toward Voice Query Clarification.* In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu and Emine Yilmaz, editors, The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, pages 1257–1260. ACM, 2018.

[Kirkpatrick *et al.* 2016] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran and Raia Hadsell. *Overcoming catastrophic forgetting in neural networks.* CoRR, vol. abs/1612.00796, 2016.

[Kiros *et al.* 2015] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba and Sanja Fidler. *Skip-thought vectors.* Advances in neural information processing systems, vol. 28, 2015.

[Kondadadi *et al.* 2013] Ravi Kondadadi, Blake Howald and Frank Schilder. *A Statistical NLG Framework for Aggregated Planning and Realization.* In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1406–1415, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[Kong *et al.* 2015] Weize Kong, Rui Li, Jie Luo, Aston Zhang, Yi Chang and James Allan. *Predicting Search Intent Based on Pre-Search Context.* In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, page 503–512, New York, NY, USA, 2015. Association for Computing Machinery.

[Kosmajac & Keselj 2019] Dijana Kosmajac and Vlado Keselj. *Twitter User Profiling: Bot and Gender Identification.* In CLEF, 2019.

[Krasakis *et al.* 2020] Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides and Evangelos Kanoulas. *Analysing the Effect of Clarifying Questions on Document Ranking in Conversational Search.* In Krisztian Balog, Vinay Setty, Christina Lioma, Yiqun Liu, Min Zhang and Klaus Berberich, editors, ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020, pages 129–132. ACM, 2020.

[Krasakis *et al.* 2022] Antonios Minas Krasakis, Andrew Yates and Evangelos Kanoulas. *Zero-shot Query Contextualization for Conversational Search.* In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, pages 1880–1884, New York, NY, USA, July 2022. Association for Computing Machinery.

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. *ImageNet Classification with Deep Convolutional Neural Networks.* In F. Pereira, C.J. Burges, L. Bottou and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 25. Curran Associates, Inc., 2012.

[Kuditipudi *et al.* 2019] Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Rong Ge and Sanjeev Arora. *Explaining Landscape Connectivity of Low-cost Solutions for Multilayer Nets.* In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox and Roman Garnett, editors, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 14574–14583, 2019.

[Kudo & Richardson 2018] Taku Kudo and John Richardson. *Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.* arXiv preprint arXiv:1808.06226, 2018.

[Kuhlthau 1991] Carol Collier Kuhlthau. *Inside the search process: Information seeking from the user's perspective.* J. Am. Soc. Inf. Sci., vol. 42, no. 5, pages 361–371, 1991.

[Lamb *et al.* 2016] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville and Yoshua Bengio. *Professor forcing: A new algorithm for training recurrent networks.* Advances in neural information processing systems, vol. 29, 2016.

[Lample *et al.* 2019] Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato and Y-Lan Boureau. *Multiple-Attribute Text Rewriting.* In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.

[Lange *et al.* 2019] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory G. Slabaugh and Tinne Tuytelaars. *Continual learning: A comparative study on how to defy forgetting in classification tasks.* CoRR, vol. abs/1909.08383, 2019.

[Lavrenko & Croft 2001] Victor Lavrenko and W. Bruce Croft. *Relevance Based Language Models.* In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, page 120–127, New York, NY, USA, 2001. Association for Computing Machinery.

[Le *et al.* 2019] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier and Didier Schwab. *Flaubert: Unsupervised language model pre-training for french.* arXiv preprint arXiv:1912.05372, 2019.

[Lee *et al.* 2019] Kenton Lee, Ming-Wei Chang and Kristina Toutanova. *Latent Retrieval for Weakly Supervised Open Domain Question Answering.* In Anna Korhonen, David R. Traum and Lluís Màrquez, editors, Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July

28- August 2, 2019, Volume 1: Long Papers, pages 6086–6096. Association for Computational Linguistics, 2019.

[Lee 2017] Sungjin Lee. *Toward Continual Learning for Conversational Agents.* CoRR, vol. abs/1712.09943, 2017.

[Lee 2018] Sungjin Lee. *Accumulating Conversational Skills Using Continual Learning.* In 2018 IEEE Spoken Language Technology Workshop (SLT), pages 862–867, 2018.

[Lesort *et al.* 2019] Timothée Lesort, Hugo Caselles-Dupré, Michaël Garcia Ortiz, Andrei Stoian and David Filliat. *Generative Models from the perspective of Continual Learning.* In International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019, pages 1–8. IEEE, 2019.

[Lewis *et al.* 2020] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan and Hsuan-Tien Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[Li & Hoiem 2018] Z. Li and D. Hoiem. *Learning without Forgetting.* IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 12, pages 2935–2947, 2018.

[Li & Liang 2021] Xiang Lisa Li and Percy Liang. *Prefix-Tuning: Optimizing Continuous Prompts for Generation.* In Chengqing Zong, Fei Xia, Wenjie Li and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 4582–4597. Association for Computational Linguistics, 2021.

[Li & Wan 2018] Liunian Li and Xiaojun Wan. *Point Precisely: Towards Ensuring the Precision of Data in Generated Texts Using Delayed Copy Mechanism.* In Proceedings of the 27th International Conference on Computational Linguistics, pages 1044–1055, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[Li *et al.* 2016] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao and Bill Dolan. *A persona-based neural conversation model.* In ACL, 2016.

[Li *et al.* 2019] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher and Caiming Xiong. *Learn to Grow: A Continual Structure Learning Framework for Overcoming Catastrophic Forgetting.* In ICML, volume 97, pages 3925–3934, 2019.

[Li *et al.* 2021a] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan and Ji-Rong Wen. *Few-shot Knowledge Graph-to-Text Generation with Pretrained Language Models.* In Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1558–1568. Association for Computational Linguistics, 2021.

[Li *et al.* 2021b] Junyi Li, Tianyi Tang, Wayne Xin Zhao and Ji-Rong Wen. *Pretrained Language Model for Text Generation: A Survey.* In Zhi-Hua Zhou, editor, Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, pages 4492–4499. ijcai.org, 2021.

[Li *et al.* 2021c] Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng and Jie Zhou. *Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances.* In Chengqing Zong, Fei Xia, Wenjie Li and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 128–138. Association for Computational Linguistics, 2021.

[Li *et al.* 2022] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie and Ji-Rong Wen. *Pretrained Language Models for Text Generation: A Survey*, 2022.

[Liang *et al.* 2012] Feng Liang, Runwei Qiang and Jianwu Yang. *Exploiting real-time information retrieval in the microblogosphere.* In Karim B. Boughida, Barrie Howard, Michael L. Nelson, Herbert Van de Sompel and Ingeborg Sølvberg, editors, Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, Washington, DC, USA, June 10-14, 2012, pages 267–276. ACM, 2012.

[Liang *et al.* 2020] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao and Chao Zhang. *Bond: Bert-assisted open-domain named entity recognition with distant supervision.* In SIGKDD, 2020.

[Lin & Efron 2013] Jimmy Lin and Miles Efron. *Overview of the TREC-2013 Microblog Track.* In Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA, 2013.

[Lin *et al.* 2020a] Sheng-Chieh Lin, Jheng-Hong Yang and Jimmy J. Lin. *TREC 2020 Notebook: CAsT Track.* In TREC, 2020.

[Lin *et al.* 2020b] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang and Jimmy Lin. *Conversational Question Reformulation via Sequence-to-Sequence Architectures and Pretrained Language Models.* CoRR, vol. abs/2004.01909, 2020.

[Lin *et al.* 2021a] Jimmy Lin, Rodrigo Nogueira and Andrew Yates. Pretrained transformers for text ranking: BERT and beyond. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2021.

[Lin *et al.* 2021b] Sheng-Chieh Lin, Jheng-Hong Yang and Jimmy Lin. *Contextualized Query Embeddings for Conversational Search.* pages 1004–1015, 2021.

[Lin *et al.* 2021c] Sheng-Chieh Lin, Jheng-Hong Yang and Jimmy Lin. *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), pages 163–173. Association for Computational Linguistics, 2021.

[Lin *et al.* 2021d] Sheng-Chieh Lin, Jheng-Hong Yang and Jimmy Lin. *TREC 2020 Notebook: CAsT Track.* Technical report, TREC, December 2021.

[Lin 2004] Chin-Yew Lin. *Rouge: A package for automatic evaluation of summaries.* pages 74–81. Text summarization branches out, 2004.

[Liu & Lapata 2019] Yang Liu and Mirella Lapata. *Text Summarization with Pretrained Encoders.* In Kentaro Inui, Jing Jiang, Vincent Ng and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3728–3738. Association for Computational Linguistics, 2019.

[Liu & Mazumder 2021] Bing Liu and Sahisnu Mazumder. *Lifelong and Continual Learning Dialogue Systems: Learning during Conversation.* In AAAI, pages 15058–15063, 2021.

[Liu *et al.* 2011] Xiaohua Liu, Shaodian Zhang, Furu Wei and Ming Zhou. *Recognizing named entities in tweets.* In ACL, 2011.

[Liu *et al.* 2015] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh and Ye-Yi Wang. *Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval.* In NAACL HLT 2015, pages 912–921, 2015.

[Liu *et al.* 2018] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang and Zhifang Sui. *Table-to-text Generation by Structure-aware Seq2seq Learning.* In AAAI, 2018.

[Liu *et al.* 2019a] Tianyu Liu, Fuli Luo, Qiaolin Xia, Shuming Ma, Baobao Chang and Zhifang Sui. *Hierarchical Encoder with Auxiliary Supervision for Neural Table-to-Text Generation: Learning Better Representation for Tables.* Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pages 6786–6793, 07 2019.

[Liu *et al.* 2019b] Tianyu Liu, Fuli Luo, Qiaolin Xia, Shuming Ma, Baobao Chang and Zhifang Sui. *Hierarchical Encoder with Auxiliary Supervision for Neural Table-to-Text Generation: Learning Better Representation for Tables.* In AAAI, 2019.

[Liu *et al.* 2019c] Tianyu Liu, Fuli Luo, Pengcheng Yang, Wei Wu, Baobao Chang and Zhifang Sui. *Towards Comprehensive Description Generation from Factual Attribute-value Tables.* In ACLs, 2019.

[Liu *et al.* 2019d] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692, 2019.

[Logeswaran & Lee 2018] Lajanugen Logeswaran and Honglak Lee. *An efficient framework for learning sentence representations.* arXiv preprint arXiv:1803.02893, 2018.

[Lomonaco & Maltoni 2017] Vincenzo Lomonaco and Davide Maltoni. *CORe50: a New Dataset and Benchmark for Continuous Object Recognition.* In 1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings, volume 78 of *Proceedings of Machine Learning Research*, pages 17–26. PMLR, 2017.

[Lovón-Melgarejo *et al.* 2021] Jesús Lovón-Melgarejo, Laure Soulier, Karen Pinel-Sauvagnat and Lynda Tamine. *Studying Catastrophic Forgetting in Neural Ranking Models.* In Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I, volume 12656 of *Lecture Notes in Computer Science*, pages 375–390. Springer, 2021.

[MacAvaney *et al.* 2019a] Sean MacAvaney, Andrew Yates, Arman Cohan and Nazli Goharian. *CEDR: Contextualized Embeddings for Document Ranking.* In SIGIR, pages 1101–1104, 2019.

[MacAvaney *et al.* 2019b] Sean MacAvaney, Andrew Yates, Arman Cohan and Nazli Goharian. *CEDR: Contextualized embeddings for document ranking.* In ACM SIGIR, pages 1101–1104, 2019.

[MacAvaney *et al.* 2020] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian and Ophir Frieder. *Expansion via Prediction of Importance with Contextualization.* pages 1573–1576, 2020.

[MacAvaney *et al.* 2021] Sean MacAvaney, Craig Macdonald, Roderick Murray-Smith and Iadh Ounis. *IntenT5: Search Result Diversification using Causal Language Models.* CoRR, vol. abs/2108.04026, 2021.

[MacAvaney 2020] Sean MacAvaney. *OpenNIR: A Complete Neural Ad-Hoc Ranking Pipeline.* In WSDM 2020, 2020.

[Malmi *et al.* 2020] Eric Malmi, Aliaksei Severyn and Sascha Rothe. *Unsupervised Text Style Transfer with Padded Masked Language Models.* In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8671–8680, Online, November 2020. Association for Computational Linguistics.

[Mann & Thompson 1988] William C. Mann and Sandra A. Thompson. *Rhetorical Structure Theory: Toward a functional theory of text organization.* Text - Interdisciplinary Journal for the Study of Discourse, 1988.

[Martin *et al.* 2019] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah and Benoît Sagot. *CamemBERT: a tasty French language model.* arXiv preprint arXiv:1911.03894, 2019.

[Matthijs & Radlinski 2011] Nicolaas Matthijs and Filip Radlinski. *Personalizing Web Search Using Long Term Browsing History.* WSDM '11, page 25–34, New York, NY, USA, 2011. Association for Computing Machinery.

[Mazumder *et al.* 2019] Sahisnu Mazumder, Bing Liu, Shuai Wang and Nianzu Ma. *Lifelong and Interactive Learning of Factual Knowledge in Dialogues.* In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019, pages 21–31. Association for Computational Linguistics, 2019.

[McCann *et al.* 2017] Bryan McCann, James Bradbury, Caiming Xiong and Richard Socher. *Learned in translation: Contextualized word vectors.* Advances in neural information processing systems, vol. 30, 2017.

[McCreadie *et al.* 2014] Richard McCreadie, Romain Deveaud, M-Dyaa Albakour, Stuart Mackie, Nut Limsopatham, Craig Macdonald, Iadh Ounis, Thibaut Thonet and Bekir Taner Dinçer. *University of Glasgow at TREC 2014: Experiments with Terrier in Contextual Suggestion, Temporal Summarisation and Web Tracks.* In Ellen M. Voorhees and Angela Ellis, editors, Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014, volume 500-308 of *NIST Special Publication.* National Institute of Standards and Technology (NIST), 2014.

[Meng *et al.* 2021] Yu Meng, Jiaxin Huang, Yu Zhang and Jiawei Han. *On the Power of Pre-Trained Text Representations: Models and Applications in Text Mining.* In KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021, pages 4052–4053. ACM, 2021.

[Mikolov *et al.* 2013a] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space.* In ICLR 2013, 2013.

[Mikolov *et al.* 2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. *Distributed Representations of Words and Phrases and their Compositionality.* In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc., 2013.

[Mikolov *et al.* 2015] Tomas Mikolov, Armand Joulin and Marco Baroni. *A Roadmap towards Machine Intelligence.* CoRR, vol. abs/1511.08130, 2015.

[Mitra 2021] Bhaskar Mitra. *Neural methods for effective, efficient, and exposure-aware information retrieval.* SIGIR Forum, vol. 55, no. 1, pages 19:1–19:2, 2021.

[Moryossef *et al.* 2019] Amit Moryossef, Yoav Goldberg and Ido Dagan. *Step-by-Step: Separating Planning from Realization in Neural Data-to-Text Generation.* In Jill Burstein, Christy Doran and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 2267–2277. Association for Computational Linguistics, 2019.

[Mosbach *et al.*] Marius Mosbach, Maksym Andriushchenko and Dietrich Klakow. *On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines.* arXiv e-prints.

[Moshfeghi *et al.* 2016] Yashar Moshfeghi, Peter Triantafillou and Frank E. Pollick. *Understanding Information Need: An FMRI Study.* In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, page 335–344, New York, NY, USA, 2016. Association for Computing Machinery.

[Mrkšić *et al.* 2016] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen and Steve Young. *Counter-fitting Word Vectors to Linguistic Constraints.* In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 142–148, San Diego, California, June 2016. Association for Computational Linguistics.

[Mustar *et al.* 2022] Agnès Mustar, Sylvain Lamprier and Benjamin Piwowarski. *IRnator: A Framework for Discovering Users Needs from Sets of Suggestions.* In ICTIR '22: The 2022 ACM SIGIR International Conference on the Theory of Information Retrieval, Madrid, Spain, July 11 - 12, 2022, pages 138–143. ACM, 2022.

[Nakatsuji & Okui 2020] Makoto Nakatsuji and Sohei Okui. *Answer Generation through Unified Memories over Multiple Passages.* In Christian Bessiere, editor, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pages 3823–3829. ijcai.org, 2020.

[Naseem *et al.* 2021] Usman Naseem, Imran Razzak, Shah Khalid Khan and Mukesh Prasad. *A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models.* ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 20, no. 5, jun 2021.

[Nguyen *et al.* 2016a] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder and Li Deng. *MS MARCO: A Human Generated MAchine Reading COmprehension Dataset.* vol. 1773, 2016.

[Nguyen *et al.* 2016b] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder and Li Deng. *MS MARCO: A Human Generated MAchine Reading COmprehension Dataset.* In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez and Greg Wayne, editors, Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches

2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[Nguyen *et al.* 2016c] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder and Li Deng. *MS MARCO: A Human Generated MAchine Reading COmprehension Dataset*. In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez and Greg Wayne, editors, Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[Nie *et al.* 2019] Feng Nie, Jinpeng Wang, Rong Pan and Chin-Yew Lin. *An Encoder with non-Sequential Dependency for Neural Data-to-Text Generation*. In INLG, 2019.

[Nogueira & Cho 2019] Rodrigo Frassetto Nogueira and Kyunghyun Cho. *Passage Reranking with BERT*. CoRR, vol. abs/1901.04085, 2019.

[Nogueira *et al.* 2019a] Rodrigo Nogueira, Jannis Bulian and Massimiliano Ciaramita. *Multi-agent query reformulation: Challenges and the role of diversity*. In Deep Reinforcement Learning Meets Structured Prediction, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019. OpenReview.net, 2019.

[Nogueira *et al.* 2019b] Rodrigo Nogueira, Wei Yang, Jimmy J. Lin and Kyunghyun Cho. *Document Expansion by Query Prediction*. ArXiv, vol. abs/1904.08375, 2019.

[Nogueira *et al.* 2020] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep and Jimmy Lin. *Document Ranking with a Pretrained Sequence-to-Sequence Model*. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 708–718. Association for Computational Linguistics, 2020.

[Oremus 2014] Will Oremus. *The First News Report on the L.A. Earthquake Was Written by a Robot*, 2014.

[Ouyang *et al.* 2022] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike and Ryan Lowe. *Training language models to follow instructions with human feedback*. CoRR, vol. abs/2203.02155, 2022.

[Pal *et al.* 2013] Dipasree Pal, Mandar Mitra and Kalyankumar Datta. *Query Expansion Using Term Distribution and Term Association*. CoRR, vol. abs/1303.0667, 2013.

[Pan & Yang 2010] Sinno Jialin Pan and Qiang Yang. *A Survey on Transfer Learning*. 2010.

[Papineni *et al.* 2002] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. *BLEU: A Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[Paria *et al.* 2020] Biswajit Paria, Chih-Kuan Yeh, Ian En-Hsu Yen, Ning Xu, Pradeep Ravikumar and Barnabás Póczos. *Minimizing FLOPs to Learn Efficient Sparse Representations.* In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

[Park 1993] Taemin Kim Park. *The nature of relevance in information retrieval: An empirical study.* The library quarterly, vol. 63, no. 3, pages 318–351, 1993.

[Pasupat & Liang 2015] Panupong Pasupat and Percy Liang. *Compositional Semantic Parsing on Semi-Structured Tables.* In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics.

[Paulus *et al.* 2018] Romain Paulus, Caiming Xiong and Richard Socher. *A Deep Reinforced Model for Abstractive Summarization.* In 6th International Conference on Learning Representations, ICLR 2018, 2018.

[Pauws *et al.* 2019] Steffen Pauws, Albert Gatt, Emiel Krahmer and Ehud Reiter. Making effective use of healthcare data using data-to-text technology: Methodologies and applications, pages 119–145. 01 2019.

[Pennington *et al.* 2014] Jeffrey Pennington, Richard Socher and Christopher Manning. *Glove: Global Vectors for Word Representation.* In EMNLP, pages 1532–1543, 2014.

[Perez-Beltrachini & Gardent 2017] Laura Perez-Beltrachini and Claire Gardent. *Analysing data-to-text generation benchmarks.* INLG, 2017.

[Perez-Beltrachini & Lapata 2018] Laura Perez-Beltrachini and Mirella Lapata. *Bootstrapping Generators from Noisy Data.* In NAACL-HLT, 2018.

[Peters *et al.* 2017] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula and Russell Power. *Semi-supervised sequence tagging with bidirectional language models.* arXiv preprint arXiv:1705.00108, 2017.

[Peters *et al.* 2018] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer. *Deep contextualized word representations.* arXiv preprint arXiv:1802.05365, 2018.

[Petroni *et al.* 2019] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu and Alexander H. Miller. *Language Models as Knowledge Bases?* In Kentaro Inui, Jing Jiang, Vincent Ng and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural

Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 2463–2473. Association for Computational Linguistics, 2019.

[Pi *et al.* 2022] Xinyu Pi, Qian Liu, Bei Chen, Morteza Ziyadi, Zeqi Lin, Yan Gao, Qiang Fu, Jian-Guang Lou and Weizhu Chen. *Reasoning Like Program Executors.* CoRR, vol. abs/2201.11473, 2022.

[Pires *et al.* 2019] Telmo Pires, Eva Schlinger and Dan Garrette. *How multilingual is multilingual BERT?* arXiv preprint arXiv:1906.01502, 2019.

[Plachouras *et al.* 2016] Vassilis Plachouras, Charese Smiley, Hiroko Bretz, Ola Taylor, Jochen L. Leidner, Dezhao Song and Frank Schilder. *Interacting with Financial Data Using Natural Language.* In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, pages 1121–1124, New York, NY, USA, 2016. ACM.

[Ponte & Croft 1998] Jay M. Ponte and W. Bruce Croft. *A Language Modeling Approach to Information Retrieval.* In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson and Justin Zobel, editors, SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, pages 275–281. ACM, 1998.

[Pradeep *et al.* 2021] Ronak Pradeep, Rodrigo Nogueira and Jimmy Lin. *The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models.* CoRR, vol. abs/2101.05667, 2021.

[Puduppully *et al.* 2019a] Ratish Puduppully, Li Dong and Mirella Lapata. *Data-to-Text Generation with Content Selection and Planning.* In AAAI, 2019.

[Puduppully *et al.* 2019b] Ratish Puduppully, Li Dong and Mirella Lapata. *Data-to-text Generation with Entity Modeling.* In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2023–2035, Florence, Italy, July 2019. Association for Computational Linguistics.

[Puduppully *et al.* 2022] Ratish Puduppully, Yao Fu and Mirella Lapata. *Data-to-text Generation with Variational Sequential Planning.* Trans. Assoc. Comput. Linguistics, vol. 10, pages 697–715, 2022.

[Qi *et al.* 2021] Tao Qi, Fangzhao Wu, Chuhan Wu and Yongfeng Huang. *Personalized News Recommendation with Knowledge-aware Interactive Matching.* In SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, pages 61–70. ACM, 2021.

[Qu *et al.* 2020] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft and Mohit Iyyer. *Open-Retrieval Conversational Question Answering.* pages 539–548, 2020.

[Radford & Narasimhan 2018] Alec Radford and Karthik Narasimhan. *Improving Language Understanding by Generative Pre-Training.* 2018.

[Radford *et al.* 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. *Language Models are Unsupervised Multitask Learners*. 2019.

[Radlinski & Craswell 2017] Filip Radlinski and Nick Craswell. *A Theoretical Framework for Conversational Search*. In Ragnar Nordlie, Nils Pharo, Luanne Freund, Birger Larsen and Dan Russel, editors, Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017, pages 117–126. ACM, 2017.

[Raffel *et al.* 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J. Liu. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. J. Mach. Learn. Res., vol. 21, pages 140:1–140:67, 2020.

[Rajpurkar *et al.* 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev and Percy Liang. *SQuAD: 100, 000+ Questions for Machine Comprehension of Text*. CoRR, vol. abs/1606.05250, 2016.

[Ramasesh *et al.* 2021] Vinay Venkatesh Ramasesh, Ethan Dyer and Maithra Raghu. *Anatomy of Catastrophic Forgetting: Hidden Representations and Task Semantics*. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[Ranzato *et al.* 2016] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli and Wojciech Zaremba. *Sequence Level Training with Recurrent Neural Networks*. In 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016.

[Rao & Daumé III 2018] Sudha Rao and Hal Daumé III. *Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information*. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2737–2746, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Rao & III 2019] Sudha Rao and Hal Daumé III. *Answer-based Adversarial Training for Generating Clarification Questions*. In Jill Burstein, Christy Doran and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 143–155. Association for Computational Linguistics, 2019.

[Rao 2017] Sudha Rao. *Are You Asking the Right Questions? Teaching Machines to Ask Clarification Questions*. In Proceedings of ACL 2017, Student Research Workshop, pages 30–35, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[Rebuffel *et al.* 2020a] Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten and Patrick Gallinari. *A Hierarchical Model for Data-to-Text Generation*. In Advances in Information Retrieval - 42nd European Conference on IR Research,

ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I, volume 12035 of *Lecture Notes in Computer Science*, pages 65–80. Springer, 2020.

[Rebuffel *et al.* 2020b] Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten and Patrick Gallinari. *PARENTing via Model-Agnostic Reinforcement Learning to Correct Pathological Behaviors in Data-to-Text Generation.* In INLG, 2020.

[Rebuffel *et al.* 2022] Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere and Patrick Gallinari. *Controlling hallucinations at word level in data-to-text generation.* Data Min. Knowl. Discov., vol. 36, no. 1, pages 318–354, 2022.

[Rebuffi *et al.* 2017] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl and Christoph H. Lampert. *iCaRL: Incremental Classifier and Representation Learning.* In CVPR, pages 5533–5542, 2017.

[Reimers & Gurevych 2019] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* In EMNLP, 11 2019.

[Reiter & Dale 2000] Ehud Reiter and Robert Dale. Building natural language generation systems. Studies in Natural Language Processing. Cambridge University Press, 2000.

[Reiter *et al.* 2005] Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu and Ian Davy. *Choosing Words in Computer-generated Weather Forecasts.* Artif. Intell., vol. 167, no. 1-2, pages 137–169, September 2005.

[Reiter 2007] Ehud Reiter. *An architecture for data-to-text systems.* In proceedings of the eleventh European workshop on natural language generation (ENLG 07), pages 97–104, 2007.

[Ribeiro *et al.* 2019] Leonardo F. R. Ribeiro, Claire Gardent and Iryna Gurevych. *Enhancing AMR-to-Text Generation with Dual Graph Representations.* In Kentaro Inui, Jing Jiang, Vincent Ng and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3181–3192. Association for Computational Linguistics, 2019.

[Ribeiro *et al.* 2020] Leonardo F. R. Ribeiro, Yue Zhang, Claire Gardent and Iryna Gurevych. *Modeling Global and Local Node Contexts for Text Generation from Knowledge Graphs.* Trans. Assoc. Comput. Linguistics, vol. 8, pages 589–604, 2020.

[Ribeiro *et al.* 2021] Leonardo F. R. Ribeiro, Yue Zhang and Iryna Gurevych. *Structural Adapters in Pretrained Language Models for AMR-to-Text Generation.* In Marie-Francine Moens, Xuanjing Huang, Lucia Specia and Scott Wen-tau Yih, editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 4269–4282. Association for Computational Linguistics, 2021.

[Ritter *et al.* 2011] Alan Ritter, Colin Cherry and William B. Dolan. *Data-driven Response Generation in Social Media.* In Conference on Empirical Methods in Natural Language Processing, EMNLP '11, 2011.

[Roberts *et al.* 2020] Adam Roberts, Colin Raffel and Noam Shazeer. *How Much Knowledge Can You Pack Into the Parameters of a Language Model?* In Bonnie Webber, Trevor Cohn, Yulan He and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 5418–5426. Association for Computational Linguistics, 2020.

[Rocchio 1971] Joseph John Rocchio. *Relevance feedback in information retrieval.* Gerard Salton, editor, The SMART Retrieval System - Experiments in Automatic Document Processing, pages 313–323, 1971.

[Rodríguez *et al.* 2018] Natalia Díaz Rodríguez, Vincenzo Lomonaco, David Filliat and Davide Maltoni. *Don't forget, there is more than forgetting: new metrics for Continual Learning.* CoRR, vol. abs/1810.13166, 2018.

[Rogers *et al.* 2020] Anna Rogers, Olga Kovaleva and Anna Rumshisky. *A primer in bertology: What we know about how bert works.* Transactions of the Association for Computational Linguistics, vol. 8, pages 842–866, 2020.

[Rojas Barahona *et al.* 2019] Lina M. Rojas Barahona, Pascal Bellec, Benoit Besset, Martinho Dossantos, Johannes Heinecke, Munshi Asadullah, Olivier Leblouch, Jeanyves. Lancien, Geraldine Damnati, Emmanuel Mory and Frederic Herledan. *Spoken Conversational Search for General Knowledge.* In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, pages 110–113, Stockholm, Sweden, September 2019. Association for Computational Linguistics.

[Roller *et al.* 2020] Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff*et al. Open-Domain Conversational Agents: Current Progress, Open Problems, and Future Directions.* arXiv preprint arXiv:2006.12442, 2020.

[Rosenblatt 1958] Frank Rosenblatt. *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychological review, vol. 65, no. 6, page 386, 1958.

[Rumelhart *et al.* 1986] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. *Learning representations by back-propagating errors.* Nature, 1986.

[Salton *et al.* 1975] Gerard Salton, Anita Wong and Chung-Shu Yang. *A vector space model for automatic indexing.* Communications of the ACM, vol. 18, no. 11, pages 613–620, 1975.

[Sanderson 2008] Mark Sanderson. *Ambiguous Queries: Test Collections Need More Sense.* In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, page 499–506, New York, NY, USA, 2008. Association for Computing Machinery.

[Sanh *et al.* 2019] Victor Sanh, Lysandre Debut, Julien Chaumond and Thomas Wolf. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* arXiv preprint arXiv:1910.01108, 2019.

[Sanh *et al.* 2022] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf and Alexander M. Rush. *Multitask Prompted Training Enables Zero-Shot Task Generalization.* In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.

[Sankepally 2019] Rashmi Sankepally. *Event Information Retrieval from Text.* In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, page 1447, New York, NY, USA, 2019. Association for Computing Machinery.

[Scao *et al.* 2022a] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé*et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.* arXiv preprint arXiv:2211.05100, 2022.

[Scao *et al.* 2022b] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani and et al. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.* CoRR, vol. abs/2211.05100, 2022.

[Schuster & Paliwal 1997] Mike Schuster and Kuldip K Paliwal. *Bidirectional recurrent neural networks.* IEEE transactions on Signal Processing, vol. 45, no. 11, pages 2673–2681, 1997.

[Scialom *et al.* 2020] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski and Jacopo Staiano. *Discriminative Adversarial Search for Abstractive Summarization.* In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, pages 8555–8564. PMLR, 2020.

[Scialom *et al.* 2021] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang and Patrick Gallinari. *QuestEval: Summarization Asks for Fact-based Evaluation.* In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6594–6604, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[See *et al.* 2017] Abigail See, Peter J. Liu and Christopher D. Manning. *Get To The Point: Summarization with Pointer-Generator Networks.* In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[Sekulic *et al.* 2021a] Ivan Sekulic, Mohammad Aliannejadi and Fabio Crestani. *Towards Facet-Driven Generation of Clarifying Questions for Conversational Search.* In Faegheh Hasibi, Yi Fang and Akiko Aizawa, editors, ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021, pages 167–175. ACM, 2021.

[Sekulic *et al.* 2021b] Ivan Sekulic, Mohammad Aliannejadi and Fabio Crestani. *User Engagement Prediction for Clarification in Search.* In Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I, volume 12656 of *Lecture Notes in Computer Science*, pages 619–633. Springer, 2021.

[Seneff & Polifroni 1996] Stephanie Seneff and Joseph Polifroni. *A new restaurant guide conversational system: Issues in rapid prototyping for specialized domains.* In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, volume 2, pages 665–668. IEEE, 1996.

[Sennrich *et al.* 2016] Rico Sennrich, Barry Haddow and Alexandra Birch. *Controlling politeness in neural machine translation via side constraints.* In NAACL-HLT, 2016.

[Shah & Bender 2022] Chirag Shah and Emily M. Bender. *Situating Search.* In David Elsweiler, editor, CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022, pages 221–232. ACM, 2022.

[Shao *et al.* 2019] Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu and Xiaoyan Zhu. *Long and Diverse Text Generation with Planning-based Hierarchical Variational Model.* In Kentaro Inui, Jing Jiang, Vincent Ng and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3255–3266. Association for Computational Linguistics, 2019.

[Sharma *et al.* 2022] Pratyusha Sharma, Antonio Torralba and Jacob Andreas. *Skill Induction and Planning with Latent Language.* In Smaranda Muresan, Preslav

Nakov and Aline Villavicencio, editors, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 1713–1726. Association for Computational Linguistics, 2022.

[Shen *et al.* 2020] Xiaoyu Shen, Ernie Chang, Hui Su, Jie Zhou and Dietrich Klakow. *Neural Data-to-Text Generation via Jointly Learning the Segmentation and Correspondence.* In ACL, 2020.

[Smeuninx *et al.* 2020] Nils Smeuninx, Bernard De Clerck and Walter Aerts. *Measuring the Readability of Sustainability Reports: A Corpus-Based Analysis Through Standard Formulae and NLP.* International Journal of Business Communication, 2020.

[Song *et al.* 2018] Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang and Daniel Gildea. *Leveraging Context Information for Natural Question Generation.* In Marilyn A. Walker, Heng Ji and Amanda Stent, editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 569–574. Association for Computational Linguistics, 2018.

[Souza *et al.* 2019] Fábio Souza, Rodrigo Nogueira and Roberto Lotufo. *Portuguese named entity recognition using BERT-CRF.* CoRR, 2019.

[Sridharan & Mota 2022] Mohan Sridharan and Tiago Mota. *Combining Commonsense Reasoning and Knowledge Acquisition to Guide Deep Learning in Robotics.* CoRR, vol. abs/2201.10266, 2022.

[Stajner & Hulpus 2020] Sanja Stajner and Ioana Hulpus. *When Shallow is Good Enough: Automatic Assessment of Conceptual Text Complexity using Shallow Semantic Features.* In LREC, 2020.

[Stajner *et al.* 2020] Sanja Stajner, Sergiu Nisioi and Ioana Hulpus. *CoCo: A Tool for Automatically Assessing Conceptual Complexity of Texts.* In LREC, 2020.

[Stoyanchev *et al.* 2014] Svetlana Stoyanchev, A. Liu and Julia Hirschberg. *Towards natural clarification questions in dialogue systems.* AISB 2014 - 50th Annual Convention of the AISB, 01 2014.

[Sun *et al.* 2016] Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su and Xifeng Yan. *Table Cell Search for Question Answering.* In Proceedings of the 25th International Conference on World Wide Web - WWW '16, pages 771–782. ACM Press, 2016.

[Sun *et al.* 2020] Fan-Keng Sun, Cheng-Hao Ho and Hung-Yi Lee. *LAMOL: LAnguage MOdeling for Lifelong Language Learning.* In ICLR, 2020.

[Sun *et al.* 2021] Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen and Maarten de Rijke. *Simulating User Satisfaction for the Evaluation of Task-oriented Dialogue Systems.* In SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, pages 2499–2506. ACM, 2021.

[Sutskever *et al.* 2011] Ilya Sutskever, James Martens and Geoffrey Hinton. *Generating Text with Recurrent Neural Networks.* In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 1017–1024, USA, 2011. Omnipress.

[Sutskever *et al.* 2014] Ilya Sutskever, Oriol Vinyals and Quoc V Le. *Sequence to Sequence Learning with Neural Networks.* In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 3104–3112. Curran Associates, Inc., 2014.

[Tang & Yang 2019] Zhiwen Tang and Grace Hui Yang. *Dynamic Search - Optimizing the Game of Information Seeking.* CoRR, vol. abs/1909.12425, 2019.

[Tavakoli *et al.* 2022] Leila Tavakoli, Johanne R. Trippas, Hamed Zamani, Falk Scholer and Mark Sanderson. *MIMICS-Duo: Offline & Online Evaluation of Search Clarification.* In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper and Gabriella Kazai, editors, SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, pages 3198–3208. ACM, 2022.

[Thompson *et al.* 2019] Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh and Philipp Koehn. *Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation.* In NAACL, pages 2062–2068, 2019.

[Trask *et al.* 2018] Andrew Trask, Felix Hill, Scott E. Reed, Jack W. Rae, Chris Dyer and Phil Blunsom. *Neural Arithmetic Logic Units.* CoRR, vol. abs/1808.00508, 2018.

[Tsiami *et al.* 2018] Antigoni Tsiami, Petros Koutras, Niki Efthymiou, Panagiotis Paraskevas Filntisis, Gerasimos Potamianos and Petros Maragos. *Multi3: Multi-Sensory Perception System for Multi-Modal Child Interaction with Multiple Robots.* In 2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018, pages 1–8. IEEE, 2018.

[Vaswani *et al.* 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. *Attention is All you Need.* In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 5998–6008, 2017.

[Veniat *et al.* 2020] Tom Veniat, Ludovic Denoyer and Marc'Aurelio Ranzato. *Efficient Continual Learning with Modular Networks and Task-Driven Priors.* CoRR, vol. abs/2012.12631, 2020.

[Veron *et al.* 2019] Mathilde Veron, Sahar Ghannay, Anne-Laure Ligozat and Sophie Rosset. *Lifelong Learning and Task-Oriented Dialogue System: What Does It Mean?* In Erik Marchi, Sabato Marco Siniscalchi, Sandro Cumani, Valerio Mario Salerno and Haizhou Li, editors, Increasing Naturalness and Flexibility in Spoken

Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019, volume 714 of *Lecture Notes in Electrical Engineering*, pages 347–356. Springer, 2019.

[Vijayakumar *et al.* 2016] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall and Dhruv Batra. *Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models*. CoRR, vol. abs/1610.02424, 2016.

[Vinyals *et al.* 2015] Oriol Vinyals, Meire Fortunato and Navdeep Jaitly. *Pointer Networks*. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 2692–2700. Curran Associates, Inc., 2015.

[Vulić & Moens 2015] Ivan Vulić and Marie-Francine Moens. *Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings*. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pages 363–372, 2015.

[Wang & Lemon 2013] Zhuoran Wang and Oliver Lemon. *A Simple and Generic Belief Tracking Mechanism for the Dialog State Tracking Challenge: On the believability of observed information*. In SIGDIAL 2013 Conference, page 423–432, 2013.

[Wang & Zhu 2009] Jun Wang and Jianhan Zhu. *Portfolio Theory of Information Retrieval*. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, page 115–122, New York, NY, USA, 2009. Association for Computing Machinery.

[Wang *et al.* 2007] Mengqiu Wang, Noah A. Smith and Teruko Mitamura. *What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA*. In Jason Eisner, editor, EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic, pages 22–32. ACL, 2007.

[Wang *et al.* 2016] Daixin Wang, Peng Cui and Wenwu Zhu. *Structural deep network embedding*. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1225–1234, 2016.

[Wang *et al.* 2018] Xuanhui Wang, Cheng Li, Nadav Golbandi, Mike Bendersky and Marc Najork. *The LambdaLoss Framework for Ranking Metric Optimization*. In Proceedings of The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), pages 1313–1322, 2018.

[Wang *et al.* 2020] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill*et al. CORD-19: The Covid-19 Open Research Dataset*. ArXiv, 2020.

[Wang *et al.* 2022] Fei Wang, Zhewei Xu, Pedro Szekely and Muhao Chen. *Robust (Controlled) Table-to-Text Generation with Structure-Aware Equivariance Learning*. arXiv preprint arXiv:2205.03972, 2022.

[Wang 2019] Hongmin Wang. *Revisiting Challenges in Data-to-Text Generation with Fact Grounding.* In INLG, 2019.

[Wei *et al.* 2022] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai and Quoc V. Le. *Finetuned Language Models are Zero-Shot Learners.* In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.

[Weizenbaum 1966] Joseph Weizenbaum. *ELIZA—a computer program for the study of natural language communication between man and machine.* Communications of the ACM, vol. 9, no. 1, pages 36–45, 1966.

[Wiese *et al.* 2017] Georg Wiese, Dirk Weissenborn and Mariana Neves. *Neural Domain Adaptation for Biomedical Question Answering.* In CoNLL 2017, pages "281–289", August 2017.

[Wiseman *et al.* 2017] Sam Wiseman, Stuart Shieber and Alexander Rush. *Challenges in Data-to-Document Generation.* In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2253–2263, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[Wiseman *et al.* 2018] Sam Wiseman, Stuart Shieber and Alexander Rush. *Learning Neural Templates for Text Generation.* In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3174–3187, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Wortsman *et al.* 2021] Mitchell Wortsman, Maxwell Horton, Carlos Guestrin, Ali Farhadi and Mohammad Rastegari. *Learning Neural Network Subspaces.* In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, pages 11217–11227. PMLR, 2021.

[Wu *et al.* 2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey*et al. Google's neural machine translation system: Bridging the gap between human and machine translation.* arXiv preprint arXiv:1609.08144, 2016.

[Wu *et al.* 2020] Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, Ying Wei and Junzhou Huang. *Adversarial sparse transformer for time series forecasting.* Advances in neural information processing systems, vol. 33, pages 17105–17115, 2020.

[Xiang *et al.* 2010] Biao Xiang, Daxin Jiang, Jian Pei, Xiaohui Sun, Enhong Chen and Hang Li. *Context-aware ranking in web search.* Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010.

[Xiong *et al.* 2017] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu and Russell Power. *End-to-end neural ad-hoc ranking with kernel pooling.* In ACM SIGIR, pages 55–64, 2017.

[Xu *et al.* 2018a] Hu Xu, Bing Liu, Lei Shu and Philip S. Yu. *Lifelong Domain Word Embedding via Meta-Learning.* In IJCAI-18, pages 4510–4516. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[Xu *et al.* 2018b] Kun Xu, Lingfei Wu, Zhiguo Wang, Yansong Feng and Vadim Sheinin. *SQL-to-Text Generation with Graph-to-Sequence Model.* In Ellen Riloff, David Chiang, Julia Hockenmaier and Jun'ichi Tsujii, editors, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 931–936. Association for Computational Linguistics, 2018.

[Xu *et al.* 2019] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei and Ming Zhou. *LayoutLM: Pre-training of Text and Layout for Document Image Understanding.* CoRR, vol. abs/1912.13318, 2019.

[Xu *et al.* 2020] Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He and Bowen Zhou. *Self-Attention Guided Copy Mechanism for Abstractive Summarization.* In Dan Jurafsky, Joyce Chai, Natalie Schluter and Joel R. Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pages 1355–1362. Association for Computational Linguistics, 2020.

[Yan *et al.* 2021] Shipeng Yan, Jiangwei Xie and Xuming He. *DER: Dynamically Expandable Representation for Class Incremental Learning.* In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 3014–3023. Computer Vision Foundation / IEEE, 2021.

[Yang *et al.* 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov and Christopher D. Manning. *HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering.* pages 2369–2380. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing ACL, October-November 2018.

[Yang *et al.* 2019a] Wei Yang, Kuang Lu, Peilin Yang and Jimmy Lin. *Critically Examining the" Neural Hype" Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models.* In ACM SIGIR, pages 1129–1132, 2019.

[Yang *et al.* 2019b] Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li and Jimmy Lin. *Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering.* CoRR, vol. abs/1904.06652, 2019.

[Yang *et al.* 2019c] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov and Quoc V Le. *Xlnet: Generalized autoregressive pretraining for language understanding.* Advances in neural information processing systems, vol. 32, 2019.

[Yao *et al.* 2013] Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch and Peter Clark. *Answer Extraction as Sequence Tagging with Tree Edit Distance.* In Lucy Vanderwende, Hal Daumé III and Katrin Kirchhoff, editors, Human Language Technologies: Conference of the North American Chapter of the Association of

Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, pages 858–867. The Association for Computational Linguistics, 2013.

[Yao *et al.* 2022] Yuan Yao, Bowen Dong, Ao Zhang, Zhengyan Zhang, Ruobing Xie, Zhiyuan Liu, Leyu Lin, Maosong Sun and Jianyong Wang. *Prompt Tuning for Discriminative Pre-trained Language Models.* In Findings of the Association for Computational Linguistics: ACL 2022, pages 3468–3473, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Ye *et al.* 2020] Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei and Lei Li. *Variational Template Machine for Data-to-Text Generation.* In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

[Ye *et al.* 2021] Jiacheng Ye, Tao Gui, Yichao Luo, Yige Xu and Qi Zhang. *One2Set: Generating Diverse Keyphrases as a Set.* 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), pages 4598–4608, Online, August 2021. Association for Computational Linguistics.

[Yin *et al.* 2016a] Pengcheng Yin, Zhengdong Lu, Hang Li and Ben Kao. *Neural Enquirer: Learning to Query Tables in Natural Language.* In Subbarao Kambhampati, editor, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, pages 2308–2314. IJCAI/AAAI Press, 2016.

[Yin *et al.* 2016b] Wenpeng Yin, Hinrich Schütze, Bing Xiang and Bowen Zhou. *Abcnn: Attention-based convolutional neural network for modeling sentence pairs.* Transactions of the Association for Computational Linguistics, vol. 4, pages 259–272, 2016.

[Yin *et al.* 2020] Pengcheng Yin, Graham Neubig, Wen-tau Yih and Sebastian Riedel. *TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data.* In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8413–8426, Online, July 2020. Association for Computational Linguistics.

[Yu *et al.* 2020] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji and Meng Jiang. *A Survey of Knowledge-Enhanced Text Generation.* CoRR, vol. abs/2010.04389, 2020.

[Yue *et al.* 2021] Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin and Dayong Wu. *NeurJudge: A Circumstance-aware Neural Framework for Legal Judgment Prediction.* In SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, pages 973–982. ACM, 2021.

[Zamani *et al.* 2018] Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik G. Learned-Miller and Jaap Kamps. *From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing.* In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, pages 497–506. ACM, 2018.

[Zamani *et al.* 2020a] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett and Gord Lueck. Generating clarifying questions for information retrieval, page 418–428. Association for Computing Machinery, New York, NY, USA, 2020.

[Zamani *et al.* 2020b] Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N. Bennett, Nick Craswell and Susan T. Dumais. *Analyzing and Learning from User Interactions for Search Clarification.* In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pages 1181–1190. ACM, 2020.

[Zamani *et al.* 2022a] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler and Michael Bendersky. *Retrieval-Enhanced Machine Learning.* In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, page 2875–2886, New York, NY, USA, 2022. Association for Computing Machinery.

[Zamani *et al.* 2022b] Hamed Zamani, Johanne R. Trippas, Jeff Dalton and Filip Radlinski. *Conversational Information Seeking*, January 2022. arXiv:2201.08808 [cs].

[Zenke *et al.* 2017] Friedemann Zenke, Ben Poole and Surya Ganguli. *Continual Learning Through Synaptic Intelligence.* In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995. PMLR, 2017.

[Zhai & Lafferty 2004] ChengXiang Zhai and John D. Lafferty. *A study of smoothing methods for language models applied to information retrieval.* ACM Trans. Inf. Syst., vol. 22, no. 2, pages 179–214, 2004.

[Zhang *et al.* 2019a] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger and Yoav Artzi. *Bertscore: Evaluating text generation with bert.* arXiv preprint, 2019.

[Zhang *et al.* 2019b] Xingxing Zhang, Furu Wei and Ming Zhou. *HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization.* In Anna Korhonen, David R. Traum and Lluís Màrquez, editors, Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 5059–5069. Association for Computational Linguistics, 2019.

[Zhang *et al.* 2022] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou and Dawei Song. *A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models.* CoRR, vol. abs/2201.05337, 2022.

[Zukerman & Raskutti 2002] Ingrid Zukerman and Bhavani Raskutti. *Lexical Query Paraphrasing for Document Retrieval*. In COLING 2002: The 19th International Conference on Computational Linguistics, 2002.