Génération de questions et paradigme question/réponse pour l'exploration des collections de sciences humaines numériques

Frederic Bechet, Elie Antoine, Jeremy Auguste, Géraldine Damnati

Génération de questions et paradigme question/réponse pour l'exploration des collections de sciences humaines numériques

Frédéric Béchet[1] Elie Antoine[1] Jeremy Auguste[1,3] Géraldine Damnati[2]
(1) Aix-Marseille Université, CNRS, LIS {first.last}@lis-lab.fr
(2) Orange Innovation, DATA&AI, Lannion {first.last}@orange.com
(3) IrAsia - Institut de recherches Asiatiques

RÉSUMÉ

Cet article présente notre méthode de génération de questions, proposant d'exploiter l'analyse sémantique de textes pour sélectionner des réponses plausibles et enrichir le processus de génération par des traits sémantiques génériques. Ces questions générées sont ensuite utilisées à plusieurs fins, d'une part comme une méthode d'adaptation de modèles de compréhension de documents, et de l'autre comme liens explicables entre documents dans le cadre d'une collection d'archives numérisées pour les études en sciences sociales. Un autre intérêt de cette étude est l'évaluation des méthodes de génération de questions et de compréhension de documents sur un nouveau type de documents, pour aller au-delà des évaluations de référence traditionnelles.

ABSTRACT

Question Generation and Answering for exploring Digital Humanities collections

This paper presents our method for generating questions, proposing to exploit semantic analysis of texts to select plausible answers and to enrich the generation process with generic semantic features. These generated questions are then used for several purposes, on the one hand as a method for adapting models of document understanding, and on the other hand as explainable links between documents in the context of a collection of digitized archives for social science studies. Another interest of this study is the evaluation of question generation and document comprehension methods on a new type of documents, to go beyond traditional reference evaluations.

MOTS-CLÉS : Génération de questions, Compréhension de documents, Question/Réponse, Humanités numériques.

KEYWORDS: Question Generation, Machine reading Question Answering, Digital Humanities.

/!\ Cet article est un résumé/condensé de deux publications faites à ce sujets : /!

— Question Generation and Answering for exploring Digital Humanities collections (LREC 2022) (Bechet *et al.*, 2022)

— Génération de questions à partir d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents (TALN 2022) (Antoine *et al.*, 2022)

1 Introduction

Machine Reading Comprehension and Question Generation are two *mirror* tasks of Natural Language Processing (NLP). Traditionally handled by complex *pipelines* based on very different models, information retrieval for question answering models and parsing for question generation, they have recently been unified since the advent of *end-to-end* methods based on pre-trained language models.

Thus, as presented in (Du *et al.*, 2017), question generation can be modeled as a text generation task where a sequence-to-sequence model is trained to *translate* a sequence of words representing a sentence or a passage into another sequence of words representing a question, without going through an explicit linguistic analysis as it used to be the case. On the other hand, automatic document understanding can be seen as a labeling task consisting in learning, from a pair *(text,question)*, which words should be labeled with the labels *start of answer* and *end of answer* in the text.

The development of pre-trained language models used in conjunction with large corpora of triplets *question/answer/context* such as *SQuAD* (Rajpurkar *et al.*, 2016) can be used to directly train both translation models *answer-to-question* and labeling models *question-to-answer*. While the performance of these models is impressive on these reference corpora, generally containing text from Wikipedia and simple questions obtained by crowdsourcing, the generalization of these models to corpora containing more complex texts and less literal questions remains a challenge. It is in this context that we propose in this study a method for unsupervised adaptation of text comprehension models based on automatic question generation.

The contributions of this study are at two levels : on the one hand, the comparison of different methods of encoding semantic information for the generation of questions evaluated with respect to their capacity to train a question/answer model on a new corpus of texts ; on the other hand, the study of the generalization capacity of comprehension and question generation models on a new corpus containing texts and questions more complex than those that can be found in reference corpora such as *SQuAD* (Rajpurkar *et al.*, 2016) for English or *FQuAD* (d'Hoffschmidt *et al.*, 2020) for French.

In order to study how the current boost in performance in NLU models on benchmark data translates to real-life settings, the applicative framework considered here is the exploration of digitized collections by professional users that are used to analyze archives in order to perform Social Science research. We chose to focus on the question/answering paradigm, as asking questions and looking for answers is at the same time a natural way for researchers to explore archives and also the task that received the most attention in recent language understanding studies, especially since the release of large training data such as SQuAD (Rajpurkar *et al.*, 2016) [1].

In this paper we will present first the *self-management* corpus, a collection of a French journal ranging over 20 years from 1966 to 1986, which has been chosen as our archival material in the project, then we will highlight the differences between benchmark corpora usually based on *Wikipedia* and digitized archive collections. We will then present the question/answering paradigm, the annotation scheme developed in Archival and point out the differences between the kinds of questions that can be made by professional users and those used in Machine Reading datasets such as *SQuAD*. We will describe the question generation and question answering models that have been developed to adapt a Machine Reading model trained on Wikipedia to the *self-management* corpus of the project without any supervision. Finally we will present the first results obtained on the *self-management* dataset with this adapted Machine Reading model.

1. ttps://rajpurkar.github.io/SQuAD-explorer/

2 The self-management corpus

2.1 Origin of the collection

The "self-management" notion falls within the large spectrum of social sciences. It concerns daily social environment, economic life, as well as political life, education, ecology, culture, architecture, …. It addresses populations structure, the relationship of populations with resources, the political, legal and administrative framework of society and the authority relations between individuals and groups.

Since the 1960's, the FMSH [2] foundation's library has gathered a pluridisciplinary multilingual mixed collection (archives and documents) about self-management (*autogestion* in French). It gathers around 25000 pieces : books, journals, reports, leaflets, correspondences.

2.2 Corpus description

For this study, we are particularly interested in the *Autogestion* journal [3] which is distributed in its digitized form by the French Persée organization. We are using a version of the corpus that has been OCRized with Tesseract without manual corrections. Hence data are not free of OCR errors but the structure of the journal (mono-column, few figures) implies that the OCR quality is good (further studies could imply precise evaluation of OCR quality and impact of OCR errors on downstream NLP tasks but for this study, OCR output are taken *as is*).

The resulting corpus is composed of 46 issues ranging over 20 years, for an overall amount of 6298 pages and 1.98M tokens.

2.3 Specificities of texts from an NLP point of view

Most studies in Information Extraction or Question Answering are carried out on Wikipedia pages. Wikipedia documents are particularly well suited for theses tasks as they intrinsically dedicated to convey factual information. Another characteristic of Wikipedia is that articles are supposed to follow a Neutral Point of View policy [4]. Recent work (Bertsch & Bethard, 2021) aims at detecting so-called puffery (*i.e* sentences that do not respect that policy, which are tagged by editors as "peacock phrases") but this phenomenon remains very rare. On the contrary, texts that are relevant for Digital Humanities and studies related to Social Science are not only factual and neutral documents but also essays or articles that reflect the writer's point of view. Description of events are not only depicted by facts but with deeper analysis of the previous notions or influences that yielded this event as well as their consequences and how they influenced the thinking of other actors.

Références

ANTOINE E., AUGUSTE J., BECHET F. & DAMNATI G. (2022). Génération de question à partir

2. Fondation Maison des Sciences de l'Homme, https://www.fmsh.fr/
3. https://www.persee.fr/collection/autog
4. https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents. In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Traitement Automatique des Langues Naturelles*, p. 104–115, Avignon, France : ATALA. HAL : hal-03701494.

BECHET F., ANTOINE E., AUGUSTE J. & DAMNATI G. (2022). Question Generation and Answering for exploring Digital Humanities collections. In *13th Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France. HAL : hal-03719368.

BERTSCH A. & BETHARD S. (2021). Detection of puffery on the english wikipedia. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 329–333.

D'HOFFSCHMIDT M., BELBLIDIA W., HEINRICH Q., BRENDLÉ T. & VIDAL M. (2020). FQuAD : French question answering dataset. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1193–1208, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.107.

DU X., SHAO J. & CARDIE C. (2017). Learning to ask : Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1342–1352.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392 : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.