Un corpus annoté pour la génération de questions et l'extraction de réponses pour l'enseignement

Thomas Gerald, Sofiane Ettayeb, Ha Quang Le, Gabriel Illouz, Patrick Paroubek, Anne Vilnat

Un corpus annoté pour la génération de questions et l'extraction de réponses pour l'enseignement

Thomas Gerald[1, 2] Sofiane Ettayeb [1, 2] Ha Quang Le [3] Gabriel Illouz [1]
Patrick Paroubek [1] Anne Vilnat [1]
(1) CNRS/LISN, Paris Saclay, France
(2) SATT Paris Saclay, France
(3) ProfessorBob, Paris Saclay, France
thomas.gerald@universite-paris-saclay.fr,
sofiane.ettayeb@universite-paris-saclay.fr, ha-quang.le@polytechnique.edu,
vilnat@limsi.fr

RÉSUMÉ

Dans cette démonstration, nous présenterons les travaux en cours pour l'annotation d'un nouveau corpus de questions-réponses en langue Française. Contrairement aux corpus existant comme "FQuad" ou "Piaf", nous nous intéressons à l'annotation de questions-réponses "non factuelles". En effet, si dans la littérature, de nombreux corpus et modèles de questions-réponses pré-entraînés sont disponibles, ceux-ci ne privilégient que rarement les annotations s'appuyant sur un schéma de raisonnement issue de l'agrégation de différentes sources ou contextes. L'objectif du projet associé est de parvenir à la création d'un assistant virtuel pour l'éducation, ainsi des réponses explicatives, de raisonnement et/ou d'agrégation de l'information sont à privilégier. Notons enfin, que la volumétrie des données doit être conséquente, en particulier par la considération d'approches neuronales génératives ou extractives. Actuellement, nous disposons de 262 questions et réponses obtenues durant l'étape de validation de la campagne d'annotation. Une deuxième phase d'annotation avec une volumétrie plus importante débutera fin mai 2022 (environ 8000 questions).

ABSTRACT

An annotated corpus for abstractive question generation and extractive answer for education

In this demonstration, we present the work in progress for the annotation of a new corpus of questions/answers (**QA**) in French language. Contrary to existing corpora such as "FQuad" or "Piaf", we are interested in non-factual questions/answers. If many corpora and pre-trained **QA** models are available, annotations are seldom based on several-steps reasoning. The final objective of the associated project is to create a virtual assistant for education, thus, non-factual answers must be prefered. Finally, the number of annotations must be consequent, particularly with the consideration of neural approaches for the **QA** generation process. Today, we reached 262 questions and answers obtained with the validation step of the annotation campaign. A second annotation phase will start in the end of May 2022.

MOTS-CLÉS : question/réponse, extraction d'informations, système d'annotation.

KEYWORDS: question/answering, extraction d'informations, annotation system.

1 Description

Avec les nouveaux outils du TAL et la démocratisation de l'apprentissage profond, les applications en langage naturel se sont complexifiées. Parmi ces applications, nous nous intéressons à la génération et la sélection de questions et réponses pour la mise en place d'un assistant de cours en langue Française. Néanmoins, si des corpus de questions/réponses (**QR**) existent, aucun ne correspond aujourd'hui au niveau de complexité requis pour assister efficacement enseignants et élèves. Des corpus comme *Piaf* (Keraron *et al.*, 2020) et *Fquad*(d'Hoffschmidt *et al.*, 2020) ont mis à disposition de la communauté des collections de **QR**. Cependant, des réponses factuelles sont majoritairement attendues, celles-ci ne permettant pas d'aider efficacement l'apprenant à parfaire sa méthodologie ou ses connaissances. Pour parvenir à réaliser notre objectif, nous avons mis en place un système d'annotation de **QR** conçu pour ce type de données. Nous avons aujourd'hui terminé la première étape et validé le protocole d'annotation pour des documents portant sur les programmes d'histoire. Cette pré-campagne a été réalisée avec l'aide de trois annotateurs spécialisés dans l'enseignement secondaire. Les retours et remarques nous ont permis de définir et de préciser le type des données souhaité, en particulier sur les sources à choisir et la typologie des questions.

Pour l'étape de validation du protocole, nous nous somme appuyés sur deux ressources, l'une d'elle reprenant des articles de l'encyclopédie libre "Wikipedia", la seconde axée sur les programmes scolaires officiels en utilisant des manuels d'enseignement du secondaire [1]. Pour les ressources extraites de Wikipedia, nous avons filtré les articles relativement à leur adéquation avec les programmes de seconde et première en histoire. Pour cela, nous avons récupéré les articles s'appuyant sur des requêtes formées par les titres des chapitres du "bulletin officiel du ministère de l'éducation nationale" [2]. D'autres domaines et sources suivront dans les prochaines étapes de la campagne d'annotation.

Notre collaboration avec ProfessorBob nous a permis d'analyser les données issues d'interactions réelles entre étudiants et un assistant virtuel. Avec l'aide d'enseignants et leur retour sur les annotations, nous avons pu déterminer une typologie des questions souhaitées. Nous précisons dans le guide d'annotation quelles sont les formulations attendues, privilégiant les annotations où la réponse sélectionnée est explicative ou construite sur un schéma de raisonnement en plusieurs étapes. Pour reconnaître les différentes questions nous avons donc choisi de les classer en quatre types : 1) les questions factuelles ; 2) les questions de vocabulaire ; 3) les questions de cours où la réponse est explicative et la suite des passage sélectionnées répond explicitement à la question ; 4) les questions de compréhension où un raisonnement est nécessaire, la réponse est alors une liste de passages permettant de construire la réponse. Ces différents types sont renseignés par les annotateurs.

2 Démonstration

Dans cette démonstration, nous présenterons :

- La plate-forme d'annotation mise en place pour la campagne ;
- Les ressources utilisées, ainsi que les textes obtenus après traitement ;
- Le guide d'annotation [3] et les différents aspects des données souhaitées ;
- Les premiers résultats issus du traitement des données récoltées ;

1. https://www.lelivrescolaire.fr/
2. https://www.education.gouv.fr/programmes-scolaires-41483
3. https://docs.google.com/document/d/1KW-UMKHG-t9C9bD4jgRP0MduUhDq_fMM84fRK06iSyM

Références

d'Hoffschmidt M., Belblidia W., Heinrich Q., Brendlé T. & Vidal M. (2020). FQuAD : French Question Answering Dataset. In T. Cohn, Y. He & Y. Liu, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 de *Findings of ACL*, p. 1193–1208 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.107.

Keraron R., Lancrenon G., Bras M., Allary F., Moyse G., Scialom T., Soriano-Morales E.-P. & Staiano J. (2020). Project PIAF : Building a Native French Question-Answering Dataset. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis, Éds., *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, p. 5481–5490 : European Language Resources Association.