Classification de questions en langage naturel par le type sémantique des réponses attendues

Théo Oriol*, Mathieu Dodard*, Kévin Cousot*, Melissa Mekaoui*, Hani Guenoune*,**, Jean Bort*, Antoine Nguyen*, Thibaud Sanchez*, Philippe Garnier*, Cédric Lopez*

*Emvista, Cap Oméga, Rond-point Benjamin Franklin, Montpellier
prenom.nom@emvista.com,
https://www.emvista.com/
**LIRMM, 161 rue Ada, Montpellier
hani.guenoune@lirmm.fr
http://www.lirmm.fr/

Résumé. Les systèmes de question-réponse (QA, *Question Answering*) sont traditionnellement constitués des trois tâches suivantes : 1) analyse de la question, 2) analyse de l'ensemble documentaire contenant les réponses, 3) recherche et extraction des réponses. Dans cette dernière décennie, les systèmes de QA à base d'apprentissage prennent la forme d'un modèle *end-to-end*. Par conséquent, les trois étapes ne sont plus explicitement représentées. Il en résulte que les systèmes de QA à base d'apprentissage les plus récents commettent de nombreuses erreurs dès lors que la réponse n'est pas dans le texte ou qu'un raisonnement est nécessaire. En particulier, le type sémantique de la réponse attendue (TSA) peut être incohérent avec le type sémantique de la réponse retournée. Dans cet article, nous nous focalisons sur la tâche d'identification du TSA. Dans un premier temps, nous proposons une taxonomie pour représenter les TSA. Dans un second temps, nous expérimentons des modèles avec CamemBERT développés à partir du corpus de questions-réponses français FQUAD. L'évaluation est réalisée sur le corpus de questions-réponses français PIAF.

1 Introduction

Les systèmes de question-réponse (QA) ont pour objectif de retourner automatiquement les réponses aux questions posées par les humains. Les systèmes de QA sont indispensables pour faciliter l'accès à une information précise noyée dans un grand volume documentaire. Les systèmes actuels sont fondés sur l'hypothèse que la réponse se trouve dans le texte. Il en découle que, dans le cas contraire, les modèles les plus avancés demeurent confrontés à un problème d'incohérence sémantique de la réponse retournée vis-à-vis de la question posée. Par exemple, la question "Quel animal mange des céréales ?" posée sur le texte "La société Gradle fournit des céréales." retourne "Gradle" avec le modèle le plus récent traitant le français, CamembertQA [1] (Martin et al., 2020). L'incohérence sémantique entre la réponse attendue de

1. https ://fquad-demo.illuin.tech/