Système de question-réponse multilingue appliqué aux agents conversationnels

Wissam Siblini, Charlotte Pasqual
Axel Lavielle, Cyril Cauchois

Worldline, France
{wissam.siblini,charlotte.pasqual,cyril.cauchois}@worldline.com, lavielleaxel@gmail.com

Résumé. Les modèles de langages (e.g. BERT) permettent de résoudre avec brio des tâches de TALN complexes comme le question-réponse. Cependant, les jeux de données spécifiques à ces tâches sont principalement en anglais, ce qui rend difficilement compte des progrès dans les autres langues. Heureusement, les modèles commencent à être pré-entraînés dans des centaines de langues et ont une bonne capacité de transfert zero-shot d'une langue à l'autre. Dans cet article, nous montrons notamment que BERT multilingue, entraîné pour la tâche de question-réponse en anglais, est capable de généraliser au français et au japonais. Nous présentons alors une application pratique Kate, agent conversationnel dédié au support ressources humaines, qui répond aux questions posées par des utilisateurs dans plusieurs langues à partir de contenus de pages d'intranet.

1 Introduction

Depuis quelques années, on assiste à une révolution du domaine du Traitement Automatique du Langage Naturel (TALN) (Vaswani et al., 2017; Devlin et al., 2018), probablement motivée par des compétitions comme SQuAD (Rajpurkar et al., 2016) ou GLUE (Wang et al., 2018). Sur SQuAD, les modèles de langage tels que BERT (Devlin et al., 2018), XLNet (Yang et al., 2019) ; ont montré qu'ils pouvaient identifier, si elle existe, la réponse à une question dans une source donnée. Cette fonctionnalité est intéressante pour augmenter les agents conversationnels. En effet, ceux-ci sont souvent limités : ils identifient des intentions prédéfinies et fournissent des réponses scriptées, mais ne peuvent pas répondre à une requête inattendue. Pourtant les sociétés disposent d'un grand nombre de sources d'informations dans lesquels un modèle automatique de compréhension de texte serait susceptible d'identifier l'information recherchée. Malheureusement, la majorité des jeux de données (e.g. SQuAD) pour entraîner un tel modèle sont exclusivement en anglais. Comment résoudre cette tâche dans d'autres langages ? Recréer des ensembles de données étiquetées dans toutes les langues ciblées serait une solution peu flexible qui demanderait beaucoup de ressources. Une autre direction possible est le transfert zero-shot d'un modèle entraîné pour la tâche en anglais vers la langue cible (Loginova et al., 2018). Par exemple, les modèles de langage entraînés en multilingue semblent intégrer naturellement un alignement linguistique dans leur représentation des phrases, et obtiennent alors des performances étonnantes en transfert zero-shot. Dans cet article, nous démontrons cette