Université Claude Bernard Lyon 1

DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **25 juin 2024**

Nom de famille et prénom de l'auteur : **Monsieur BERRO Auday**

Titre de la thèse : *« Génération de paraphrases pour l'apprentissage automatique des services conversationnels »*

Résumé

Les Services de Dialogue (DS), comme les assistants virtuels et les bots orientés tâches, connaissent une adoption croissante grâce aux avancées des technologies open source, de l'IA et de la puissance de calcul. Ils améliorent les interactions homme-machine en facilitant les échanges en langage naturel. Ils ont transformé notre façon d'interagir avec les appareils, les sites web et les applications. Par exemple, un enfant de 2 ans peut écouter sa chanson préférée en disant simplement "Alexa, joue la chanson Baby Shark" avant même d'apprendre à utiliser un ordinateur.

Cependant, développer un bot reste un défi, notamment dans la traduction des énoncés des utilisateurs en intentions, en raison de la diversité des expressions linguistiques. Par exemple, pour l'énoncé "Quel temps fait-il à Lyon" le bot doit reconnaître l'intention (prévision météorologique) et les entités associées (lieu = Lyon). La même intention peut être exprimée différemment. Par exemple, un autre utilisateur pourrait demander "Quelles sont les prévisions météo pour Lyon".

Développer un bot implique la capacité à transformer une expression utilisateur en une ou plusieurs intentions, correspondant à l'identification des tâches que l'utilisateur souhaite accomplir (e.g. prévisions météorologiques). Le bot extrait ensuite les entités pertinentes (e.g. le lieu et la date de prévision). Enfin, il mappe l'intention et les paramètres vers des services back-end (e.g. appels API) pour obtenir les résultats. Cela se fait généralement en deux étapes: (i) entraîner un modèle de compréhension du langage naturel (NLU) pour mapper les énoncés des utilisateurs à des intentions

prédéfinies et extraire les entités associées et (ii) développer des fonctions pour mapper les intentions vers des formes exécutables (par exemple, API) et satisfaire les demandes des utilisateurs en effectuant des tâches. Ainsi, l'entraînement d'un NLU nécessite un grand ensemble d'énoncés pour chaque intention avec toutes les compositions possibles d'entités. Les énoncés qui se réfèrent à la même intention sont appelés paraphrases. La richesse et l'ambiguïté du langage humain soulignent l'importance de la paraphrase dans la construction de jeux de données diversifiés. La paraphrase est une tâche NLP cruciale pour créer des jeux de données diversifiés car elle reformule un énoncé tout en préservant son sens. Les approches traditionnelles comme l'embauche d'experts sont coûteuses, d'où l'intérêt croissant pour la génération automatisée de paraphrases.

Dans cette thèse, nous proposons d'exploiter les techniques de PG existantes afin de générer des ensembles de données de haute qualité pour la formation des bots. L'accent est mis sur la collecte d'un grand nombre d'énoncés tout en garantissant des critères de qualité linguistique spécifiques tels que la pertinence sémantique et la diversité. Les principales contributions comprennent la mise en œuvre et l'évaluation d'un pipeline PG de base et la résolution de problèmes tels que la pertinence sémantique et la diversité. Une taxonomie de 15 types d'erreurs identifiées dans les modèles de PG basés sur l'architecture transformer. Le développement d'un nouvel ensemble de données annotées où les paraphrases ont été annotées à l'aide de la taxonomie proposée. En outre, nous avons exploité cet ensemble de données annotées pour développer un modèle d'annotation de paraphrases multilabel. Inspirés par des études antérieures sur le crowdsourcing, nous étudions le potentiel des LLM, tels que GPT-3.5, pour des tâches de PG syntaxiquement diverses. Nous proposons de reproduire une étude existante sur le crowdsourcing qui propose un pipeline de paraphrases en plusieurs étapes qui guide le crowdsourcing pour produire des paraphrases syntaxiquement diverses. Nous proposons de substituer les crowdworkers humains par des LLM et nous effectuons une analyse comparative pour démontrer leur efficacité dans des tâches de PG contrôlées. Dans l'ensemble, cette thèse présente une exploration complète des techniques automatisées de PG pour relever les défis de l'acquisition d'ensembles de données de haute qualité pour construire des services de dialogue robustes et réactifs.

Mots-clés: Paraphraser, Génération Automatisée de Paraphrases, Génération de Paraphrases Contrôlées, Diversité Syntaxique, Services de Dialogue, Chatbots, Traitement Automatique du Langage Naturel (NLP), Taxonomie d'erreurs, Modèles de Langues, LLMs.