Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Sanjay Kamath RAMACHANDRA RAO : me@sanjaykamath.eu

Titre : Question-réponse utilisant des données et modèles hybrides

Mots-clés : question et réponses, traitement du langage naturel, apprentissage automatique, réseaux neuronaux, intelligence artificielle.

Title: *Question Answering with Hybrid Data and Models*

Keywords: *question-answering, natural language processing, machine learning, neural networks, artificial intelligence.*

Thèse de doctorat en informatique, LIMSI, CNRS, Université Paris-Saclay, sous la direction de Brigitte Grau (Pr, ENSIIE) et Yue Ma (MC, Université Paris-Saclay). Thèse soutenue le 06/02/2020.

Jury : Mme Brigitte Grau (Pr, ENSIIE, codirectrice), Mme Yue Ma (MC, Université Paris-Saclay, codirectrice), M. Nicolas Sabouret (Pr, Université Paris-Saclay, président), M. Patrice Bellot (Pr, Université Aix-Marseille, rapporteur), M. Mohand Boughanem (Pr, Université Paul Sabatier, Toulouse, rapporteur), Mme Catherine Berrut (Pr, Université Grenoble Alpes, examinatrice), M. Patrick Gallinari (Pr, Sorbonne Université, examinateur), Mme Anne Vilnat (Pr, Université Paris-Saclay, examinatrice).

Résumé : *La recherche de réponses à des questions relève de deux disciplines : le traitement du langage naturel et la recherche d'information. L'émergence de l'apprentissage profond dans plusieurs domaines de recherche tels que la vision par ordinateur, le traitement du langage naturel, la reconnaissance vocale, etc. a conduit à l'émergence de modèles de bout en bout et les travaux actuels de l'état de l'art en question-réponse (QR) visent à mettre en œuvre de tels modèles.*

Dans le cadre du projet GoASQ[1]*, l'objectif est d'étudier, comparer et combiner différentes approches pour répondre à des questions formulées en langage naturel sur des données textuelles, en domaine ouvert et en domaine biomédical. Le travail de thèse se concentre principalement sur : 1) la construction de modèles permettant de traiter des ensembles de données à petite et à grande échelle ; 2) l'exploitation de connaissances sémantiques pour répondre aux questions par leur intégration dans les différents modèles. Nous visons à fusionner des connaissances issues de textes libres, d'ontologies, de représentations d'entités, etc.*

Afin de faciliter l'utilisation des modèles neuronaux sur des données de domaine de spécialité, généralement de petite taille, nous nous plaçons dans le cadre de l'adaptation de domaine. Nous avons proposé deux modèles de tâches de QR différents, évalués sur la tâche BIOASQ de réponse à des questions biomédicales, et nous montrons par nos résultats expérimentaux que le modèle de questions-réponses ouvert *(extraction de la réponse étant donné un ensemble de paragraphes pertinents et non pertinents) convient mieux qu'une modélisation de type* compréhension machine *(extraction de la réponse étant donné un paragraphe pertinent) qui est la plus courament utilisée. Nous pré-entraînons le modèle de* compréhension machine*, qui sert de base à notre modèle, sur différents ensembles de données pour montrer la variabilité des performances lorsque ces modèles sont adaptés au domaine biomédical. Nous constatons que l'utilisation d'un ensemble de données particulier (ensemble de données SQUAD v2.0) pour le pré-entraînement donne les meilleurs résultats lors du test et qu'une combinaison de quatre jeux de données donne les meilleurs résultats lors de l'adaptation au domaine biomédical. Nous avons effectué des expériences à l'aide de modèles de langage à grande échelle, comme BERT*[2]*, qui sont adaptés à la tâche de réponse aux questions. Les performances varient en fonction du type des données utilisées pour pré-entraîner BERT. Nous en avons conclu que le modèle de langue appris sur des données biomédicales, BIOBERT, constitue le meilleur choix pour le QR biomédical.*

Étant donné que les modèles d'apprentissage profond visent à fonctionner de bout en bout, les informations sémantiques provenant de sources de connaissances construites par des experts n'y sont généralement pas introduites. Nous avons annoté manuellement et automatiquement un jeu de données par les variantes des réponses de BIOASQ et montré l'importance d'apprendre un modèle de QR avec ces variantes. Nous montrons l'utilité d'exploiter le type de réponse attendu *et* le type lexical de la réponse *en domaine ouvert et en domaine biomédical par différentes études. Ces types sont ensuite utilisés pour mettre en évidence les entités dans les jeux de données, ce qui montre des améliorations sur l'état de l'art. Par ailleurs l'exploitation de représentations vectorielles d'entités dans les modèles se montre positif pour le domaine ouvert.*

Une de nos hypothèses est que les résultats obtenus à partir de modèles d'apprentissage profond peuvent être encore améliorés en utilisant des traits sémantiques et des traits collectifs calculés à partir des différents paragraphes sélectionnés pour ré-

1. https://goasq.lri.fr/
2. https://github.com/google-research/bert

pondre à une question. Nous proposons d'utiliser des modèles de classification binaires pour améliorer la prédiction de la réponse parmi les K candidats à l'aide de ces caractéristiques, conduisant à un modèle hybride qui surpasse les résultats de l'état de l'art sur la plupart des ensembles de données.

Enfin, nous avons évalué des modèles de QR ouverte sur des ensembles de données construits pour les tâches de compréhension machine *et* sélection de phrases. *Nous montrons la différence de performance lorsque la tâche à résoudre est une tâche de QR ouverte et soulignons le fossé important qu'il reste à franchir dans la construction de modèles de bout en bout pour la tâche complète de réponse aux questions.*

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02890467

Yuming ZHAI : zhaiyuming9@hotmail.com

Titre : Reconnaissance des procédés de traduction sous-phrastiques : des ressources aux validations

Mots-clés : création de corpus, reconnaissance automatique, application en traitement automatique des langues.

Title: *Recognition of Sub-Sentential Translation Techniques: from Resources to Validation*

Keywords: *corpus creation, automatic recognition, application in natural language processing.*

Thèse de doctorat en informatique, LIMSI, CNRS, Université Paris-Saclay, sous la direction de Anne Vilnat (Pr, Université Paris-Saclay). Thèse soutenue le 19/12/2019.

Jury : Mme Anne Vilnat (Pr, Université Paris-Saclay, directrice), M. Alexandre Allauzen (Pr, École supérieure de physique et de chimie industrielles de la ville de Paris, président), Mme Amalia Todirascu (Pr, Université de Strasbourg, rapporteuse), M. Mathieu Lafourcade (MC, Université de Montpellier, rapporteur), Mme Emmanuelle Esperança-Rodier (MC, Université Grenoble Alpes, examinatrice), M. Philippe Langlais (Pr, Université de Montréal, Canada, examinateur), M. Gabriel Illouz (MC, Université Paris-Saclay, examinateur).

Résumé : *Les procédés de traduction constituent un sujet important pour les traductologues et les linguistes. Face à un certain mot ou segment difficile à traduire, les traducteurs humains doivent appliquer des solutions particulières au lieu de la traduction littérale, telles que l'équivalence idiomatique, la généralisation, la particularisation, la modulation syntaxique ou sémantique, etc.*

Ce sujet a reçu peu d'attention dans le domaine du traitement automatique des langues (TAL). Notre problématique de recherche se décline en deux questions : est-il possible

de reconnaître automatiquement les procédés de traduction ? Certaines tâches en TAL peuvent-elles bénéficier de la reconnaissance des procédés de traduction ?

Notre hypothèse de travail est qu'il est possible de reconnaître automatiquement les différents procédés de traduction (par exemple littéral versus non littéral). Pour vérifier notre hypothèse, nous avons annoté un corpus parallèle anglais-français en procédés de traduction, tout en établissant un guide d'annotation. Notre typologie de procédés est proposée en nous appuyant sur des typologies précédentes, et est adaptée à notre corpus. L'accord inter-annotateur (0,67) est significatif mais dépasse peu le seuil d'un accord fort (0,61), ce qui reflète la difficulté de la tâche d'annotation. En nous fondant sur des exemples annotés, nous avons ensuite travaillé sur la classification automatique des procédés de traduction. Même si le jeu de données est limité, les résultats expérimentaux valident notre hypothèse de travail concernant la possibilité de reconnaître les différents procédés de traduction. Nous avons aussi montré que l'ajout des traits sensibles au contexte est pertinent pour améliorer la classification automatique.

En vue de tester la généricité de notre typologie de procédés de traduction et du guide d'annotation, nos études sur l'annotation manuelle ont été étendues au couple de langues anglais-chinois. Ces langues partagent beaucoup moins de points communs que le couple anglais-français au niveau linguistique et culturel. Le guide d'annotation a été adapté et enrichi. La typologie de procédés de traduction reste identique à celle utilisée pour le couple anglais-français, ce qui justifie d'étudier le transfert des expériences menées pour le couple anglais-français au couple anglais-chinois.

Dans le but de valider l'intérêt de ces études, nous avons conçu un outil d'aide à la compréhension écrite pour les apprenants de français langue étrangère. Une expérience sur la compréhension écrite avec des étudiants chinois confirme notre hypothèse de travail et permet de modéliser l'outil. D'autres perspectives de recherche incluent l'aide à la construction de ressources de paraphrases, l'évaluation de l'alignement automatique de mots et l'évaluation de la qualité de la traduction automatique.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02460548