Journée Traitement Automatique des Langues & I.A.

Nancy, 6 juillet 2018

L'Association française pour l'intelligence artificielle (AFIA) et l'Association pour le traitement automatique des langues (ATALA) ont organisé conjointement la journée traitement automatique des langues et intelligence artificielle, TALIA 2018, le 6 juillet 2018 à Nancy.

Le traitement automatique des langues est un thème important de l'intelligence artificielle : la langue est au cœur de la communication humaine et est un véhicule privilégié d'enregistrement et de transmission d'information, de connaissance et de culture. Compréhension et production de langue, dialogue en langue naturelle, traduction, extraction d'information et réponse à des questions sont des exemples de fonctions et d'applications auxquelles s'attaque le traitement automatique des langues. Il mobilise lui-même divers champs de l'intelligence artificielle, comme l'apprentissage automatique et la représentation des connaissances, et joue un rôle clé dans l'acquisition de connaissances à partir de textes.

Ces dernières années ont vu l'émergence des réseaux de neurones profonds qui sont aujourd'hui intensivement utilisés dans le traitement automatique des langues. Après un saut qualitatif notable pour des tâches centrales comme par exemple la traduction automatique, ces réseaux ont montré certaines de leurs limites. On sait par exemple qu'il n'est pas facile de comprendre comment un résultat a été obtenu, que la qualité des résultats est souvent moins bonne qu'avec les méthodes classiques lorsque moins de données sont disponibles.

Cette journée visait à faire le point sur les méthodes actuellement employées en IA & TAL, notamment les travaux en cours sur les réseaux profonds et représentations continues de mots, leurs limites et les recherches entreprises pour les pallier.

Elle a été organisée autour de deux séminaires invités (1 heure chacun) et de cinq présentations d'articles (seules quatre ont pu être présentées, 30 minutes chacune) sélectionnés par un comité de programme de 23 membres. 40 à 50 personnes ont assisté à la journée.

Les deux invités étaient Chloé Braud (CNRS, Loria, Nancy) et Alexandre Allauzen (Université Paris Sud, Limsi, Orsay).

Chloé Braud nous a parlé de *plongements lexicaux pour l'analyse discursive automatique.* L'analyse discursive correspond à l'identification de liens sémantiques entre des groupes de mots, phrases ou propositions. C'est une tâche complexe, car cette identification repose sur de nombreuses informations : sémantique lexicale, syntaxe, temporalité, connaissances du monde, etc. Par ailleurs, il faut prendre en compte l'interaction entre les segments textuels à lier. Chloé Braud a présenté une étude montrant que l'utilisation de plongements lexicaux – des représentations denses pré-entraînées sur de larges jeux de données – permet d'atteindre des performances similaires aux études précédentes fondées sur l'utilisation de multiples ressources [Braud & Denis 2015]. Le défaut de ces représentations est qu'elles n'ont pas été construites spécifiquement pour la tâche : des expériences ultérieures montrent qu'il est probablement crucial d'effectuer une telle adaptation [Braud & Denis 2016]. En conclusion, Chloé Braud a fait le point sur les difficultés actuelles en présentant un

système d'analyse discursive translingue pour lequel les plongements lexicaux utilisés ne conduisent pas aux meilleures performances.

Alexandre Allauzen a abordé le problème des *modèles de langue neuronaux à grand vocabulaire.* Ces dernières décennies, les réseaux de neurones artificiels et plus généralement l'apprentissage profond ont renouvelé les perspectives de recherche en traitement automatique des langues. Certaines applications, comme la traduction automatique et la reconnaissance automatique de la parole, nécessitent la conception de modèles capables d'engendrer des phrases. Du point de vue de l'apprentissage automatique, l'enjeu est alors de modéliser des séquences de mots ou de symboles qui se caractérisent par des distributions particulières, parcimonieuses et impliquant un espace de réalisation, le vocabulaire, de grande dimension. Or, malgré les avancées récentes dans ce domaine, si les modèles neuronaux sont considérés comme « universels » dans leur conception, la diversité des langues implique une réalité bien différente. Selon les langues et leurs processus morphologiques, la dimension des vocabulaires et la notion de mot diffèrent grandement et altèrent la pertinence des modèles d'apprentissage considérés pourtant comme état de l'art. Ainsi, la manipulation efficace de vocabulaire de grande taille reste un défi. Cet exposé a abordé ce défi en s'intéressant aux architectures et aux critères d'apprentissage dédiés qui permettent d'appréhender et de mieux modéliser ce phénomène typique des langues naturelles.

Les cinq articles sélectionnés par le comité de programme représentaient des thèmes variés.

Frédéric Landragin et Bruno Oberle ont présenté un article intitulé *Identification automatique de chaînes de coréférences : vers une analyse des erreurs pour mieux cibler l'apprentissage*. Ils ont rapporté une étude qualitative préliminaire concernant l'analyse linguistique des erreurs commises par des systèmes de détection automatique de chaînes de coréférence. Ils ont souligné plusieurs cas de bruit et de silence, caractérisés par des gravités différentes, ainsi que des types d'erreurs spécifiques, notamment la construction de chaînes « fourre-tout » regroupant des expressions référentielles inexploitées par ailleurs. Dans le but de définir une méthodologie généralisable, ils proposent une première typologie d'erreurs et quelques pistes de réflexion pour leur prise en compte à terme dans les processus d'apprentissage, ce qui passe par des considérations sur les types d'hybridation à envisager pour ces processus.

Emmanuelle Esperança-Rodier a présenté un travail qu'elle a réalisé avec Nicolas Becker autour de la *Comparaison de systèmes de traduction automatique, probabiliste et neuronal, par analyse d'erreurs*. Cet article présente les travaux d'analyse d'erreurs de deux systèmes de traduction automatique maison (Laboratoire d'Informatique de Grenoble), l'un probabiliste et l'autre neuronal. Après une description du corpus et des systèmes, les auteurs analysent les deux systèmes en fonction d'une typologie d'erreurs en s'arrêtant sur quelques exemples de phrases pour lesquelles les deux systèmes ont effectué le même type d'erreurs.

Ahmed Mabrouk, Rim Hantach et Philippe Calvez ont présenté une approche efficace basée sur des graphes pour la représentation textuelle (*An Efficient Semantic Graph-Based Approach for Text Representation*). La représentation des documents est l'un des principaux problèmes dans le domaine de l'analyse des textes, notamment pour l'extraction de thèmes et le calcul de similarité entre des textes. L'approche standard utilisant une représentation par sac de mots ne permet pas de représenter les liens sémantiques entre les termes. Afin de surmonter cette limitation, les auteurs introduisent une nouvelle approche basée sur l'utilisation conjointe du graphe de co-occurrence obtenu à partir d'un corpus et d'un réseau sémantique de la langue anglaise (Wordnet). Pour ce faire, un algorithme de

désambiguïsation du sens des mots a été utilisé dans le but d'établir les liens sémantiques entre les termes étant donné le contexte sous-jacent. Les expérimentations réalisées sur des bases de données standard montrent une bonne performance de l'approche proposée.

Enfin, Loïc Grobol a présenté le travail effectué avec Marco Dinarelli sur la *Modélisation d'un contexte global d'étiquettes pour l'étiquetage de séquences dans les réseaux neuronaux récurrents*. Depuis quelques années, les réseaux neuronaux récurrents ont atteint des performances à l'état-de-l'art sur la plupart des problèmes de traitement de séquences. Notamment les modèles *sequence to sequence* et les CRF neuronaux se sont montrés particulièrement efficaces pour ce genre de problèmes. Dans cet article, les auteurs proposent un réseau neuronal alternatif pour le même type de problèmes, fondé sur l'utilisation de plongements d'étiquettes et sur des réseaux à mémoire, qui permettent la prise en compte de contextes arbitrairement longs. Les auteurs comparent leurs modèles avec la littérature, et remarquent que leurs résultats dépassent souvent l'état-de-l'art et en sont proches dans tous les cas. Leurs solutions restent toutefois plus simples que les meilleurs modèles de la littérature.

Malheureusement, dans un contexte de grèves de train, Mathieu Lafourcade et Alain Joubert on dû annuler leur déplacement à Nancy et n'ont pas pu présenter leur article *Production endogène de règles déductives dans le réseau JeuxDeMots.* À partir d'un réseau lexico-sémantique, il est possible de générer des règles de façon inductive à partir des faits présents. Ces règles permettent de densifier le réseau et d'en réduire les silences. Afin de minimiser l'émergence de relations qui pourraient être erronées, la question de la polysémie est abordée et un filtrage sur les règles présentant des exceptions est réalisé.

La journée a reçu un public fourni et varié issu de laboratoires d'informatique et de linguistique : le pari consistant à organiser une journée commune TAL et IA sur la plateforme AFIA a été un succès. Les organisateurs de la journée remercient à ce titre l'ATALA, l'AFIA, les organisateurs de la plateforme AFIA, le comité de programme, les orateurs et le public. Ils ont conclu la journée en annonçant l'organisation de la conférence TALN 2019 à Toulouse sur la plateforme AFIA (1-5 juillet 2019).

Didier Schwab
Pierre Zweigenbaum