Pré-GDR TAL : Axe Apprentissage et modélisation statistique pour le TAL

Coordinateur : Frédéric Béchet

19 juin 2018

1 Participants

Une première liste de participants potentiels dans le monde académique français a été contacté, dans les laboratoires suivants :

CEA-LIST	Paris	INRIA-MAGNET	Lille	INRIA-ALMANACH	Paris
IRISA	Rennes	IRIT	Toulouse	LIA	Avignon
LIG	Grenoble	LIMSI	Orsay	LIPN	Villetaneuse
LIS	Marseille	LIUM	LeMans	LLF	Paris
LORIA	Nancy	LS2N	Nantes		

Les contributeurs à ce document qui ont répondu positivement à la déclaration d'intérêt pour le GDR TAL et qui ont contribué à sa rédaction par leurs commentaires et proposition de textes sont listés dans le tableau suivant :

Nom	**Prénom**	**Laboratoire**	**ville**	**email**
Allauzen	Alexandre	LIMSI	Paris	allauzen@limsi.fr
Bechet	Frederic	LIS	Marseille	frederic.bechet@univ-amu.fr
Besacier	Laurent	LIG	Grenoble	laurent.besacier@imag.fr
Candito	Marie	LLF	Paris	marie.candito@linguist.univ-paris-diderot.fr
Cerisara	Christophe	LORIA	Nancy	christophe.cerisara@loria.fr
Claveau	Vincent	IRISA	Rennes	vincent.claveau@irisa.fr
Crabbé	Benoit	LLF	Paris	benoit.crabbe@linguist.univ-paris-diderot.fr
de la Clergerie	Eric	INRIA	Paris	Eric.De_La_Clergerie@inria.fr
Denis	Pascal	INRIA	Lille	pascal.denis@inria.fr
Esteve	Yannick	LIUM	LeMans	yannick.esteve@univ-lemans.fr
Favre	Benoit	LIS	Marseille	benoit.favre@univ-amu.fr
Ferret	Olivier	CEA-LIST	Paris	olivier.ferret@cea.fr
Gaussier	Eric	LIG	Grenoble	Eric.Gaussier@imag.fr
Lefevre	Fabrice	LIA	Avignon	fabrice.lefevre@univ-avignon.fr
Le Roux	Joseph	LIPN	Paris	leroux@lipn.fr
Linares	Georges	LIA	Avignon	georges.linares@univ-avignon.fr
Morin	Emmanuel	LS2N	Nantes	emmanuel.morin@univ-nantes.fr
Muller	Philippe	IRIT	Toulouse	Philippe.Muller@irit.fr
Nasr	Alexis	LIS	Marseille	alexis.nasr@univ-amu.fr
Ramisch	Carlos	LIS	Marseille	carlos.ramisch@univ-amu.fr
Sagot	Benoît	INRIA	Paris	benoit.sagot@inria.fr
Tomeh	Nadi	LIPN	Paris	nadi.tomeh@lipn.univ-paris13.fr
Wisniewski	Guillaume	LIMSI	Paris	Guillaume.Wisniewski@limsi.fr
Yvon	Francois	LIMSI	Paris	francois.yvon@limsi.fr

2 Thèmes et périmètre d'étude

Le Traitement Automatique des Langues et l'Apprentissage Automatique, en particulier l'apprentissage statistique, ont des relations anciennes qui ne se limitent pas à la déferlante actuelle de l'*apprentissage profond*.

Du côté de l'apprentissage, la langue, qu'elle soit écrite ou parlée, a toujours constitué un champ d'expérimentation fécond pour de nombreux modèles statistiques numériques. Ainsi les premiers modèles de calcul stochastique mis au point par Andreï Markov ont été appliqués dès 1913 à une problématique de linguistique statistique en modélisant les séquences de lettres du roman *Eugène Onéguine* d'Alexandre Pouchkine. De nos jours, comme c'est le cas pour le traitement d'images, tous les modèles développés dans la communauté de l'apprentissage profond sont systématiquement appliqués à des tâches de Traitement Automatique de la Langue. Ces modèles sont évalués sur des données annotées (*benchmarks*) faisant référence dans la communauté.

Du côté du Traitement Automatique des Langues, l'apprentissage s'est peu à peu imposé, à partir des années 1980, pour répondre à deux problèmes que rencontraient les modèles symboliques basés sur des représentations explicites de modèles linguistiques (règles, grammaires) telles que les grammaires génératives :

— le premier est l'ambiguïté : comment choisir une solution lorsque plusieurs sont jugées acceptables par un modèle ayant forcément une vision partielle de la langue et du monde ?

— le second est la robustesse : les modèles décrits dans un cadre de modélisation de la compétence sont difficiles à mettre au point (*all grammars leak*, comme disait Edward Sapir) et ne permettent pas (ou mal) de décrire des énoncés soumis aux aléas de la performance linguistique.

Ainsi, à la description formelle de langue, sous la forme de règles, s'est substituée peu à peu une description informelle dans laquelle le modèle linguistique à implémenter est décrit sous la forme d'un *guide d'annotation* permettant à des humains d'annoter des corpus à partir desquels un processus d'apprentissage automatique permet d'inférer des modèles prédictifs.

Ce principe d'*apprentissage supervisé* à partir de corpus annotés a permis de répondre de manière partiellement satisfaisante aux deux problèmes que sont l'ambiguïté et la robustesse, au prix d'un changement de paradigme : les modèles statistiques ne produisent pas des solutions *exactes* mais des solutions *probables*. Contrairement aux modélisations symboliques qui avaient pour but de séparer les hypothèses acceptables des hypothèses erronées, en fournissant des preuves d'acceptabilité ou de rejet, les modélisations statistiques *classent* entre elles les différentes solutions produites, selon leur probabilité d'être correctes, mais sans fournir de justification précise expliquant ce classement (début des modèles *boîtes noires*).

Un deuxième changement de paradigme est apparu vers la fin des années 1990 lorsque des applications commerciales concrètes basées sur le Traitement Automatique des Langues ont commencé à se généraliser (transcription de la parole, traduction automatique, classification de documents électroniques, correcteurs orthographiques, etc.). La disponibilité de corpus contenant à la fois les entrées et les sorties d'applications ont permis de développer des méthodes d'apprentissage prédisant directement le résultat final du traitement envisagé, sans respecter les niveaux linguistiques postulés par les théories.

L'application *phare* de ce type de méthodes est la traduction automatique qui fait correspondre à des textes en langue *source* leur traduction en langue *cible* en laissant le modèle d'apprentissage inférer directement la fonction de transfert à partir des suites de mots, sans représentations abstraites d'ordre syntaxique ou sémantique.

Cette évolution découle notamment du manque chronique de données linguistiquement annotées, qui sont très coûteuses à produire, dont la couverture est toujours insuffisante, en termes de phénomènes linguistiques et de langues couvertes, et dont l'impact est amoindri par le manque d'interopérabilité (manque de consensus sur la modélisation à utiliser). Le recours au crowdsourcing pour produire des données pousse également vers des données pouvant être traitées sur la base de la seule compétence langagière, mais sans connaissances en linguistique.

Ce type de modèles trouve son apogée avec les modèles *end-to-end* actuels dans lesquels seule compte la performance par rapport à l'application finale, quelles que soient les représentations intermédiaires produites par les modèles.

Alors que le TAL statistique a tendance à s'éloigner des modèles linguistiques en privilégiant les approches *end-to-end* applicatives, un nouveau mouvement est en train de voir le jour, inspiré par l'étude des mécanismes d'acquisition et de traitement du langage chez l'humain dans le cadre des sciences cognitives [1]. Dans cette nouvelle vision, les modèles développés en apprentissage automatique peuvent être confrontés à des données provenant de la psycholinguistique, de la psychologie ou de la neurologie. Ces comparaisons pourront aider à développer des modèles de TAL plus en conformité avec des données humaines et pourront symétriquement fournir des modèles de prédiction de comportements humains (dans le cadre de tâches linguistiques), avec l'espoir de rendre moins opaque la fameuse *boîte noire* des systèmes purement statistiques.

Pour résumer, 3 thèmes principaux nous semblent pouvoir être dégagés autour de l'axe TAL et apprentissage :

1. Apprentissage pour l'analyse linguistique automatique : des tâches telles que l'étiquetage en parties du discours, le chunking, l'analyse syntaxique, la résolution de coréférences, l'analyse discursive ou l'analyse sémantique peuvent être réalisées par des modèles à base d'apprentissage. Il s'agit du TAL statistique traditionnel, visant à produire des structures linguistiques formelles, intelligibles par l'homme. La modélisation linguistique reste nécessaire pour définir les tâches et les schémas d'annotation. Ces tâches constituent des *benchmarks* auxquels différentes architectures et paradigmes d'apprentissage peuvent être comparés.
2. Apprentissage pour la réalisation d'applications : ce thème concerne les applications ***end-to-end*** où les modèles d'apprentissage prédisent directement les sorties des applications concernées. Les applications de génération de texte telles que la traduction automatique, la transcription de parole, le résumé automatique ainsi que d'autres tâches telles que l'analyse de sentiments entrent dans ce thème dans la mesure où l'annotation nécessaire à l'apprentissage ne nécessite plus nécessairement de modèles linguistiques.
3. Apprentissage pour l'étude de l'acquisition du langage : dans ce thème ce sont les mécanismes d'apprentissage à partir de données du thème 2 qui sont étudiés pour pouvoir répondre à des questions telles que *Qu'apprennent ces modèles ?*, *Quelles sont les représentations obtenues ?*, *Comment se comparent-elles aux représentations du langage chez l'humain ?*.

3 État des lieux

Nous l'avons vu dans la section précédente, les liens entre TAL et apprentissage ne sont pas nouveau. Ce qui en revanche est nouveau est le coup de projecteur que reçoit le domaine à l'heure de la déferlante *deep learning* et de l'engouement du grand public pour l'*IA* représentée de nos jours par les méthodes d'apprentissage automatique. Une question importante qui se pose face à ce constat est : qu'est-ce qui a fondamentalement changé depuis 10 ans ?

Il y a 10 ans, avant l'avènement du deep learning, les chercheurs en TAL engagés dans une approche par modélisation statistique s'appuyaient déjà sur des algorithmes d'apprentissage automatique, appliqués sur les données à leur disposition. Les approches étaient diverses (SVM, arbres de classification, CRF, GMM, FSM, LSA, TF-IDF, LDA, ...) et une grande partie du travail des chercheurs consistait à exploiter les meilleurs approches possibles pour répondre aux problèmes de TAL qu'ils souhaitaient résoudre à partir d'un jeu de données à leur disposition. Cela nécessitait de leur part une bonne maîtrise théorique et pratique des différentes approches, d'une bonne connaissance de leurs données et une bonne compréhension du problème afin de le représenter de manière optimale pour le projeter dans le contexte de l'approche retenue.

Le développement récent de méthodes basées sur des réseaux de neurones profonds, aussi appelé *Deep Learning*, semblent offrir une alternative à ce modèle en proposant à la fois d'utiliser de très grandes quantités de corpus non-annoté via des méthodes de *plongement* ou *embedding* mais aussi d'apprendre de manière jointe les représentations intermédiaires (ou traits linguistiques) et les modèles d'analyse permettant de réaliser la tâche linguistique visée [2, 3].

L'avantage de telles méthodes est d'offrir une représentation continue, sous forme de vecteur, pour à la fois les données d'entrées (par exemple les mots) mais aussi chaque niveau intermédiaire. Ces représentations, une fois intégrées dans des réseaux de neurons peuvent être raffinées pour une tâche donnée, donnant beaucoup plus de flexibilité que les approches séquentielles classiques où chaque niveau d'analyse fourni une annotation symbolique servant d'entrée au niveau suivant.

Ces stratégies ont été employées avec succès dans des tâches telles que les modèles de langage [4], la traduction automatique [5], la reconnaissance d'entités nommées [6], ou encore l'analyse syntaxique et sémantique [7, 8, 9].

Si des gains notables ont pu être observés par rapport à l'état de l'art précédent basé sur des méthodes de classification et d'étiquetage supervisés telles que les Support Vector Machine (SVM) ou les Conditional Random Field (CRF), ces gains restent limités par rapport à ceux constatés dans la communauté du traitement automatique des images lorsque les réseaux profonds ont été introduits. Une limitation des méthodes à base de plongement et de réseaux profonds appliquées au traitement du langage réside sans doute dans la nécéssité de disposer de très grandes quantités de corpus annoté, ce qui est rarement le cas pour la plupart des tâches de TAL.

Il faut noter que cet engouement du *"tout deep learning"* n'a pas pour autant effacé toutes les méthodes précédentes. Ainsi l'intégration du deep learning avec d'autres approches plus anciennes prend de plus en plus d'importance en TAL, comme par exemple dans la combinaison de modèles neuronaux récurrents et de modèles CRF [**?**] ou encore dans l'intégration avec des modèles probabilistes que l'on retrouve dans toutes les méthodes *variationnelles*, comme les auto-encodeurs variationnels très utilisés en TAL [10, 11]. Ces dernières méthodes permettent de faire facilement de la régularisation de distribution a posteriori, ce qui était très difficile avec les méthodes bayésiennes. De plus, ces approches variationnelles intègrent aujourd'hui l'apprentissage adversarial, qui permet de

modéliser des distributions de plus en plus complexes [12]. On assiste ainsi à une converge entre deep learning et d'autres modèles probabilistes très prometteuse en TAL.

Pour le thème 1 évoqué dans le paragraphe précédent, tous les résultats *état de l'art* sur des benchmarks de tâches linguistiques sont obtenus avec des modèles à base d'apprentissage, éventuellement couplés à des méthodes symboliques. Les modèles dominants en apprentissage sont les *Réseaux de neurones profonds* sous la forme de perceptrons multicouches, réseaux récurrents ou convolutionnels. Différents mécanismes permettant de modéliser des dépendances à longue distance ont été proposés (modèles à mémoire, basés sur l'attention). Toutes les équipes de recherche du monde académique sur la linguistique computationnelle développent ce genre d'approche car de réels gains de performance ont pu être obtenus grâce à ces modèles, évaluées lors de campagnes d'évaluation (ou *shared tasks*) à l'échelle internationale (*CoNLL*, *Semeval*, ...). Les contributions de ces équipes portent le plus souvent sur les aspects *modélisation* plutôt que sur les techniques d'apprentissage. En France, les principaux laboratoires développant de telles méthodes sont : INRIA-Magnet, Inria-Almanach, LIMSI, LIPN, LIS, LLF, LORIA. **à compléter**

À ces équipes *traditionnelles* travaillant sur le TAL, se sont rajoutées des équipes spécialisées dans l'apprentissage automatique, que ce soit dans le monde académique comme l'équipe MILA à Montréal, ou celui des laboratoires privés, notamment des GAFAM (Google, Apple, Facebook, Amazon, Microsoft) ou Baidu en Chine pour la reconnaissance automatique de la parole. En effet de nombreux problèmes encore ouverts dans la communauté de l'apprentissage automatique se trouvent de manière particulièrement importante dans des tâches d'annotation linguistique : déséquilibres entre classes, manque de données, problème de généralisation, bruits dans les données, prédiction structurée, etc.

Si ces problèmes ne sont pas spécifiques au TAL, ils trouvent néanmoins un écho particulier dans des tâches linguistiques à cause de caractéristiques intrinsèques à la langue (*mur de brique* de la loi de Zipf, difficultés d'annotation de corpus, dépendance des modèles au domaine, structure séquentielle et hiérarchique du langage, etc.). Ainsi, de nombreuses applications en TAL se retrouvent dans des articles publiés dans des conférences d'apprentissage telles que NIPS ou ICML à l'international ou encore CAP en France.

La structure interne de certains modèles peut n'avoir aucune modélisation linguistique *a priori*, car ils sont basés sur des approches *end-to-end* prenant en entrée des phrases et produisant directement les analyses requises. Cependant le besoin en modélisation linguistique est ici indispensable de part la nature même des tâches d'analyse linguistique automatique.

Pour le thème 2 portant sur l'apprentissage pour la réalisation d'applications, ce sont les laboratoires privés qui dominent le plus souvent le domaine. Cela s'explique par la nécessité à disposer de très grands corpus d'apprentissage ainsi que de capacités de calcul suffisantes pour pouvoir mettre au point les modèles tout en trouvant l'architecture et le paramétrage optimaux pour une tâche donnée.

Une activité académique persiste néanmoins pour les tâches de transcription de parole et de traduction automatique car de grands corpus accessibles à la communauté académique existent. On peut tout de même penser que les systèmes actuels ont fait basculer ces domaines dans l'ère industrielle.

Pour des tâches applicatives telles que le dialogue homme-machine ou la compréhension du langage pour des assistants personnels, il est très difficile pour des laboratoires académiques d'accéder à des corpus car cela nécessite d'avoir accès à des systèmes existants et à de très nombreux utilisateurs pour collecter suffisamment de données.

Face à cette industrialisation du domaine, la communauté académique s'organise en collectant ses propres corpus et en les mettant à disposition de la communauté, notamment lors des campagnes d'évaluation (par exemple les corpus de dialogue collectés pour les campagnes DSTC), ou en récupérant des données *écologiques* avec une supervision gratuite telles que des corpus d'articles avec résumés, des avis de consommateurs avec des notes, etc. C'est cette piste qu'explore par exemple en France la communauté animant la campagne d'évaluation DEFT sur la fouille de textes.

En France, les principaux laboratoires menant des activités dans ce thème sont : CEA, IRISA, IRIT, LIA, LIG, LIMSI, LIS, LIUM, LORIA, LS2N **à compléter**

Pour ces approches dirigées par les applications, principalement pour celles adoptant des modèles end-to-end, la tendance actuelle est de supprimer toute modélisation linguistique explicite en laissant l'étape d'apprentissage de représentation choisir elle-même les niveaux d'analyse pour réaliser la tâche. Cette tendance est illustrée par les modèles actuels de *séquence-à-séquence* employés pour la transcription de parole où à une séquence de vecteurs acoustiques en entrée correspond une séquence de lettres en sortie représentant la transcription de la parole recherchée ; ou encore les modèles de traduction évoqués au parahraphe précédent. Dans les deux cas il n'y a aucune modélisation linguistique explicite, ni sur le choix des unités (phonèmes, mots, n-grammes), ni sur les niveaux

d'analyse (modèles acoustiques, modèles de langage, contraintes syntaxiques ou sémantiques).

Cette tendance à l'abandon des représentations linguistiques questionne légitimement la place des chercheurs en TAL par rapport aux chercheurs en apprentissage *pur* pour ce thème. On peut néanmoins remarquer qu'à l'exception de la transcription de parole pour laquelle la forme écrite peut être considérée comme une représentation objective de la forme orale (ce qui peut être discutable dans le cas de l'oral très spontané), un problème de ces approches réside dans la difficulté à définir une fonction objective à optimiser alors que l'appréciation d'une solution est faite de manière subjective par les humains (*qualité* d'une traduction, d'un résumé, d'un dialogue). Dans ce cas, des évaluations mettant en jeu des modélisations linguistiques pour, par exemple, apprécier la *qualité* d'un énoncé, peuvent avoir leur pertinence.

Le thème 3 porte sur l'étude de l'apprentissage du langage par les machines dans le cadre général de l'émergence du langage et de son acquisition par les humains. Ces travaux visent de manière générale à tisser des liens entre les traitements du langage réalisés par l'ordinateur et ceux réalisés par le cerveau humain. Il s'agit d'un type de travaux relativement nouveaux prenant en compte des données issues de l'activité électrique du cerveau ou de données comportementales, par exemple sur la plausibilité des représentations issues du TAL en termes de comportements psycho-cognitifs, et les modèles de la lecture. On trouve aussi dans ce thème les travaux autour de l'apprentissage du langage chez l'enfant, et le parallèle avec l'apprentissage des langues par les machines sans supervision explicite.

En France, les principaux laboratoires menant des activités dans ce thème sont : Institut ILCB, LLF, LIG, **à compléter**

4 Grands enjeux

Le Traitement Automatique des Langues est à un moment charnière de son histoire. Il se retrouve au centre de la nouvelle vague de l'*Intelligence Artificielle* telle qu'elle est définie et popularisée par les grands groupes technologiques tels que les GAFAM pour lesquels le traitement du langage est au coeur des vitrines technologiques telles que les assistants personnels.

Ce coup de projecteur s'accompagne d'une *industrialisation* du domaine qui n'est pas sans risque pour la recherche académique. La quantité de données ainsi que la puissance de calcul dont bénéficient les laboratoires industriels, couplées au développement d'équipes de recherche brillantes composées de chercheurs provenant du monde académique (chercheurs confirmés ou jeunes docteurs), ont accéléré le rythme de publication et réduit considérablement le délai entre l'apparition d'une nouvelle idée et son application systématique à de très grandes quantités de tâches et de cadres applicatifs liés au TAL. Dans ce contexte, il est difficile pour les laboratoires académiques de trouver une place à cause de leurs ressources limitées en termes de données, calcul et ressources humaines.

Paradoxalement, la mise à disposition de très nombreuses librairies logicielles à la communauté scientifique, souvent par ces mêmes laboratoires industriels, rend accessibles la plupart de ces innovations, en évitant un développement logiciel important.

Bien sûr, cette situation n'est pas sans danger : le risque de transformer la recherche académique en TAL à la simple utilisation de briques logicielles toujours plus performantes, sans maîtrise des mécanismes sous-jacents, est réel.

Les grands enjeux de l'axe TAL et apprentissage concernent donc à la fois la place du TAL par rapport à d'une part l'industrialisation du domaine, et d'autre part la communauté de l'apprentissage automatique. Plusieurs enjeux ou points de vue sont listés dans ce paragraphe comme autant de réflexions et de pistes de recherche pour cet axe.

4.1 L'apprentissage comme nouveau paradigme de programmation pour le TAL ?

Les méthodes actuelles, notamment celles basées sur l'apprentissage profond, nous donnent une nouvelle *boite à outils* très riche pour envisager la modélisation de la langue. De la même manière que la programmation logique a constitué un paradigme utile pour les méthodes symboliques à base de connaissances explicites, les méthodes et outils d'apprentissage actuels peuvent être vues un nouveau langage dans lequel nous pouvons modéliser le langage, soit d'un point de vue théorique, soit d'un point de vue applicatif.

Ainsi, réduire l'apprentissage profond à un ensemble de boîtes noires transformateurs de données serait une erreur, car l'évolution de ce domaine tend au contraire vers une convergence du domaine de la programmation informatique traditionnelle, et de l'apprentissage machine. Comme l'a indiqué Yann Lecun en Janvier 2018 : "deep learning is dead. Long live differentiable programming !", les chercheurs en TAL pratiquant l'apprentissage profond écrivent en fait des programmes informatiques qui modélisent une fonction complexe et paramétrique, l'apprentissage

des paramètres de cette fonction étant automatique et transparent. Il y a tout à parier que cette convergence se poursuive, et que ces nouveaux outils soient communément utilisés par les chercheurs en TAL au même titre que le language python l'est aujourd'hui, c'est-à-dire sans que l'outil ne soit le sujet principal de la recherche.

4.2 A-t-on besoin de nouvelles théories scientifiques pour représenter le langage ?

La linguistique a constitué tout naturellement la théorie scientifique qui a structuré toutes les recherches en TAL depuis 50 ans. Même si les modèles issus de la linguistique formelle ont perdu de leur importance au fur et à mesure de l'avancée des méthodes empiriques basées sur l'apprentissage, les niveaux de représentation du langage (phonétique, morphologie, syntaxe, sémantique) ont continué à être largement utilisés lors de la mise au point de systèmes de TAL. Les travaux récents sur les modèles *end-to-end* de type ***séquence-à-séquence*** remettent en jeu cette dépendance à la représentation linguistique traditionnelle de la langue. De nouveaux modèles issus de travaux en Sciences Cognitives pourront peut-être remplacer ou compléter les modèles existants ?

On peut s'attendre à ce que les modèles d'analyse syntaxiques soient utilisés dans un contexte expérimental d'études cognitives, psycholinguistiques et neurolinguistiques pour traiter des questions liées aux hypothèses de structuration du langage. Est-ce que les modèles structurés en arbre permettent de mieux prédire le comportement humain que des modèles peu structurés comme des modèle de séquence ?. Quel apport des modèles d'apprentissage profond et comment les interpréter [13] ? Comment mettre ces modèles en relation avec les modèles cognitifs de traitement de la mémoire [14] ?

4.3 Eclaircir la *boîte noire*, rendre les modèles moins opaque ?

La voie du *whatever works* a certainement permis d'accomplir des progrès remarquables pour les applications du TAL. Cependant, cette voie a ses limites, d'un point de vue théorique en ne permettant pas de sortir d'un empirisme devenu trop coûteux nécessitant un paramétrage constant des architectures et des paramètres pour chaque nouvelle application ou domaine à traiter ; et d'un point de vue pratique en ne contrôlant pas suffisamment la qualité des sorties des systèmes, générant ainsi des erreurs inacceptables pour un humain (un système bon *en moyenne* mais faisant des erreurs incompréhensibles pour un humain ne sera pas accepté). Par exemple les approches *end-to-end* pour le résumé par abstraction actuelles [15] savent au mieux produire des formes de titres pour un document, avec parfois des problèmes de cohérence par rapport au texte originel.

Les notions d'*explicabilité* des décisions prises par des modèles d'apprentissage sont au centre d'un grand nombre de recherches actuelles afin de mieux contrôler ce qu'apprennent ces modèles tout en évitant les décisions inacceptables pour des utilisateurs.

Pour les analyses linguistiques, étudier la nature des structures apprises par les réseaux profonds est une voie de recherche qui peut permettre de mieux comprendre leur fonctionnement et leurs limites, notamment à travers le type de structure syntaxique inférée par ces modèles [13, 16, 17].

Pour les applications il est également important d'interpréter les représentations intermédiaires apprises par les réseaux profonds afin de comprendre quelles informations ont été capturées et quelles représentations ont été apprises. Des travaux récents sur la tâche de reconnaissance automatique de la parole ont proposé d'analyser les représentations capturées par les SRAP profonds. (Mohamed et al. , 2012) [18] et (Belinkov et Glass, 2017) [19] ont analysé les représentations intermédiaires apprises (d'un SRAP profond) en utilisant la visualisation t-SNE. Ils essaient aussi de comprendre quelles couches capturent mieux les informations phonétiques en entraînant un classifieur de phonèmes peu profond. Par ailleurs, (Wu et King, 2016) [20] ont évalué les représentations de plusieurs variantes de LSTM pour une tâche de synthèse vocale.

Dans le cadre de la traduction automatique neuronale, les travaux de (Shi et al. , 2016) [21] et (Belinkov et al. , 2017) [19] ont essayé de comprendre les représentations apprises par un système de traduction neuronal. Ces représentations sont fournies à un classifieur peu profond afin de prédire des étiquettes syntaxiques [21], grammaticales ou sémantiques [19]. L'analyse montre que les couches inférieures (basses) sont meilleures pour l'étiquetage grammatical, tandis que les couches supérieures sont meilleures pour l'étiquetage sémantique.

4.4 Est-ce la fin des benchmarks sur des tâches linguistiques ?

Les méthodes end-to-end basées sur des applications vont-elles complètement remplacer les évaluations sur des tâches purement linguistique, telles que l'analyse syntaxique ?

Il y a une réelle interrogation sur l'utilité de tâches, artificielles, de production automatique de représentations linguistiques. Elles ont longtemps été considérées comme une étape nécessaire à toute tâche de TAL élaborée, mais

aujourd'hui même des tâches à forte composante sémantique (comme le question/réponse, l'inférence textuelle) sont abordables via une approche end-to-end.

Cependant la difficulté à évaluer les systèmes applicatifs de manière objective et le besoin d'évaluation des capacités de *généralisation* des modèles, notamment en évaluant comment extraire des connaissances générales à partir d'un modèle spécifique, donnent toujours une pertinence à ces tâches linguistiques génériques.

En outre il faudra sans doute sortir du paradigme de la seule évaluation quantitative dont l'effet pervers consiste à ne plus regarder que la *performance* des modèles sans se poser la question de la pertinence et de la réalité des gains obtenus.

4.5 Les méthodes supervisées sont-elles une impasse ?

La nécessité de disposer d'importants corpus annotés pour chaque tâche de TAL dans les approches supervisées, et le manque de généralité des modèles appris lors d'un changement de domaine ou de tâche, même léger, posent la question de la pertinence de ce type d'approche sur le long terme.

D'un point de vue industriel, le coût et la difficulté d'obtention de tels corpus d'apprentissage rendent ces approches souvent inefficaces ; d'un point de vue scientifique, le manque de généralité des modèles peut laisser à penser que l'approche supervisé n'est gagnante qu'à court terme et que d'autres méthodes sont nécessaires pour capturer des phénomènes plus généraux permettant une meilleure généralisation des modèles.

Les méthodes non ou peu supervisées sont une des réponses à ce problème, utilisant de la supervision indirecte ou bien en complétant les annotations existantes avec des règles explicites ou induites [22]. Cependant on peut penser qu'un changement de paradigme, lié à la vitesse grandissante de production des données et à leur quantité ébranle le domaine de l'apprentissage cette fois, au dépend de la notion traditionnelle d'apprentissage supervisé utilisant des annotations et en faveur d'un apprentissage multi-formes exploitant d'autres modèles *professeurs*, des annotations indirectes, le transfert entre tâches, la compétition ou la collaboration entre modèles, la prédiction du contexte, etc. Ceci impactera largement les méthodes actuelles déployées en TAL aujourd'hui, et notamment la création de corpus.

La plupart des modèles de TAL considèrent que l'ensemble de l'objet à analyser (mot, phrase paragraphe, document) est accessible lors de son analyse. De plus, les méthodes d'apprentissage nécessitent généralement d'avoir en totalité le corpus d'apprentissage à disposition avant d'entraîner les modèles. Ces deux prérequis ne sont pas valides dans le cadre du traitement du langage par l'humain où l'analyse se fait en flux et où l'apprentissage est continu. Implémenter des modèles de TAL en flux qui continuent à apprendre au fur et à mesure de leur utilisation, notamment via la collaboration humain/machine, est une piste de recherche prometteuse.

4.6 Compréhension de la langue : où en sommes nous ?

Les problèmes posés par les tâches de compréhension de texte, même en se limitant à des approches envisageant la compréhension comme une forme de question-réponse sont encore très loin d'être résolus. Les capacités inférentielles des systèmes de question-réponse actuels, même en se concentrant sur la partie finale de recherche d'une réponse dans un court passage, ne sont au mieux que de bons systèmes d'appariement plus ou moins sémantique entre question et phrase [23] qu'ils soient ou non neuronaux. Des travaux récents commencent à se pencher sur les problèmes liés aux inférences [24] mais c'est très embryonnaire et la route est encore longue.

5 Positionnement

5.1 Ethique

Problème de l'accès aux données. Dans le cadre de la nouvelle loi relative à la protection des données personnelles. Problèmes éthiques d'utilisation de ces données : nos modèles sont-ils bien anonymes ?

Problème de l'inégalité dans l'accès aux corpus. Les performances potentiellement atteignables sur les différentes tâches du TAL dépendent principalement de la qualité et de la quantité de données disponibles (ou pas !). C'est à ce niveau que se situe la plus grande fracture entre les grandes multinationales (Google, Apple, Facebook, Amazon, Microsoft, Samsung, Baidu) et le reste de la communauté. Cette fracture existe non seulement entre ces multinationales et les laboratoires académiques, mais aussi entre ces multinationales et la grande majorité des entreprises privées : grands groupes, PME, ou autres. Il est crucial que les laboratoires publics puissent accéder à ce type de données afin de pouvoir réaliser des travaux de recherche sur une échelle qu'il leur est impossible d'atteindre

actuellement, afin que les connaissances, compétences, savoir-faire que cela implique puissent être partagés avec le reste de la population : étudiants, entreprises, citoyens, politiques... Avec la hausse significative de la quantité de données, se pose aussi la question des infrastructures de calcul du milieu académique qui se doivent d'être également redimensionnées.

Problème des biais dans les données d'apprentissage. Des disparités importantes dans les performances des méthodes basées sur l'apprentissage automatique peuvent apparaître pour certaines catégories de personnes en fonction des biais dans les corpus d'apprentissage. Ainsi, la fameuse étude sur le *Gender Shades* [25] a montré des écarts importants entre les hommes et les femmes et la couleur de peau en termes de reconnaissance faciale pour plusieurs systèmes automatiques.

Le cas de la reconnaissance faciale et l'étude de Buolamwini (2018) ont permis de questionner l'idée selon laquelle les algorithmes étaient *objectifs*. Encore peu d'études aborde cet aspect pour le TAL. Le chat-bot Tay a été un parfait exemple de ce qui peut se passer lorsque l'on fournit des données biaisées à un système d'apprentissage automatique : le 23 mars 2016, Microsoft met en ligne sur Twitter un chatbot, Tay, censé incarner une adolescente. Le robot conversationnel de Microsoft visait comme public les Américains de 18 à 24 ans. L'intelligence artificielle était basée sur un fonctionnement simple : le vocabulaire et le raisonnement du chat-bot s'étoffent au fur et à mesure que les utilisateurs interagissaient avec lui. Mais la communauté 4Chan a décidé de pousser le système dans ses retranchements et par force de discours biaisés a réussi à faire tenir des propos misanthropes, sexistes, racistes et xénophobes au chatbot. Un exemple un peu moins médiatisé est celui des word embeddings ou plongements de mots. La représentation vectorielle des mots permet d'utiliser des opérations algébriques pour inférer sur leur sens. Mais des études ont récemment mis en avant le fait que ces plongements de mots capturaient également des stéréotypes sexistes ou racistes. Les travaux de Bolukbasi et al. (2016) [26] ont mis en avant l'existence de ces biais dans les plongements de mots entraînés sur le corpus Google News et préviennent du biais que cela peut induire dans les systèmes qui utiliseront de telles données. Dans leur article ils démontrent que les plongements de mots encodent des analogies du type `(man:computer programmer :: woman:homemaker)` qui pourrait se traduire comme *le programmeur est à l'homme ce que la ménagère est à la femme*. Caliskan et al. (2017) [27] ont trouvé des résultats similaires et montré l'existence de biais en terme de genre et de race. Ce genre d'études permet de nous rappeler le fait que les données ne sont pas plus brutes qu'elles ne sont neutres.

6 Interface

Le langage est partout ! Cet axe est au coeur d'activités de recherche avec d'autres GDR (le nouveau GDR apprentissage évidemment) et d'autres instituts (SHS, Santé, ...).

7 Programmatique

7.1 Communauté

des idées ?

7.2 Jeunes chercheurs

Montrer les limites des applications industrielles existantes du TAL, notamment tout ce qui concerne les capacités de compréhension encore très limités des machines.

7.3 Mediation scientifique

Intervenir dans le débat public sur l'IA.

8 Bibliographie

Références

[1] Morten H Christiansen, Nick Chater, and Peter W Culicover, *Creating language : Integrating evolution, acquisition, and processing*, MIT Press, 2016.

[2] T. Mikolov, K. Chen, G. Corrado, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," 2013, vol. abs/1301.3781.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, Eds., pp. 2787–2795. Curran Associates, Inc., 2013.

[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin, "A neural probabilistic language model," *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[5] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv :1406.1078*, 2014.

[6] Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu, "Evaluating word representation features in biomedical named entity recognition tasks," *BioMed research international*, vol. 2014, 2014.

[7] Mohit Bansal, Kevin Gimpel, and Karen Livescu, "Tailoring continuous word representations for dependency parsing," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

[8] Jacob Andreas and Dan Klein, "How much do word embeddings encode about syntax," in *Proceedings of ACL*, 2014.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," in *the Journal of Machine Learning Research 12*, 2011, pp. 2461–2505.

[10] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *stat*, vol. 1050, pp. 1, 2014.

[11] Yishu Miao, Lei Yu, and Phil Blunsom, "Neural variational inference for text processing," in *International Conference on Machine Learning*, 2016, pp. 1727–1736.

[12] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing, "Toward controlled generation of text," in *International Conference on Machine Learning*, 2017, pp. 1587–1596.

[13] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg, "Assessing the ability of lstms to learn syntax-sensitive dependencies," *Transactions of the Association of Computational Linguistics*, vol. 4, no. 1, pp. 521–535, 2016.

[14] Richard L Lewis and Shravan Vasishth, "An activation-based model of sentence processing as skilled memory retrieval," *Cognitive science*, vol. 29, no. 3, pp. 375–419, 2005.

[15] Abigail See, Peter J. Liu, and Christopher D. Manning, "Get to the point : Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2017, pp. 1073–1083, Association for Computational Linguistics.

[16] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni, "Colorless green recurrent networks dream hierarchically," in *Proceedings of NAACL-HLT*, 2018, pp. 1195–1205.

[17] Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A Smith, "What do recurrent neural network grammars learn about syntax ?," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, 2017, vol. 1, pp. 1249–1258.

[18] Abdel-rahman Mohamed, Geoffrey Hinton, and Gerald Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4273–4276.

[19] Yonatan Belinkov and James Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *Advances in Neural Information Processing Systems*, 2017, pp. 2438–2448.

[20] Zhizheng Wu and Simon King, "Investigating gated recurrent networks for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5140–5144.

[21] Xing Shi, Inkit Padhi, and Kevin Knight, "Does string-based neural mt learn source syntax ?," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1526–1534.

[22] Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré, "Training classifiers with natural language explanations," *arXiv preprint arXiv :1805.03818*, 2018.

[23] Danqi Chen, Jason Bolton, and Christopher D. Manning, "A thorough examination of the cnn/daily mail reading comprehension task," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2016, pp. 2358–2367, Association for Computational Linguistics.

[24] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel, "Constructing datasets for multi-hop reading comprehension across documents," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 287–302, 2018.

[25] Joy Buolamwini and Timnit Gebru, "Gender shades : Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[26] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai, "Man is to computer programmer as woman is to homemaker ? debiasing word embeddings," in *Advances in Neural Information Processing Systems*, 2016, pp. 4349–4357.

[27] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.