Pré GDR TAL: Axe de réflexion Ressources

June 13, 2018

1 Participants

Coordination : Philippe Muller, IRIT, Université de Toulouse.
Contributeurs :

Marianna Apidianaki, LIMSI-CNRS; Marie Candito, LLF, Univ. Paris Diderot; Iris Eskohl-Taravela, Univ. Paris Nanterre; Michael Filhol, LIMSI-CNRS, Univ. Paris Sud; Karën Fort, Paris Sorbonne; Nicolas Hernandez, LINA, Univ Nantes; Mathieu Lafourcade, LIRMM, Univ Montpellier; José Moreno, IRIT, Univ. Toulouse.

2 Définition et Périmètre

La problématique des ressources pour le TAL peut se décliner sur trois volets: les données, les connaissances, et les outils.

En tant que science expérimentale le TAL est fortement dépendant des données, qui nourrissent les analyses, permettent de régler les modèles, et d'évaluer et comparer les approches. Les données langagières prennent la forme de corpus, des sélections de documents écrits, oraux,sous des formes variées, et sont enrichies de connaissances sous formes d'annotations d'analyse à des niveaux divers (tokenization, morphologique, lexical, syntaxique, sémantique, pragmatique).

La constitution de corpus, annotés ou non, est la première pierre sur laquelle se fondent de nombreux efforts expérimentaux, et les questions majeures sont la méthodologie de constitution, l'évaluation de la validité ou de la qualité du recueil (qualification),et les moyens de rendre accessible de manière fiable et pérenne les données produites.

Une autre partie importante de l'activité du domaine est la constitution de connaissances linguistiques qui alimentent les recherches sur tel ou tel problème, ou sont issues d'étapes précédentes d'analyse: lexiques, grammaires, ontologies, représentations sémantiques ... Les problématiques importantes ici sont les questions de couverture, d'évaluation et de réutilisabilité.

Enfin les travaux reposent généralement sur l'implémentation d'analyseurs des différent niveaux de traitement, dont la réutilisabilité et l'interopérabilité est un enjeu important pour la progression de la recherche et des applications induites.

La particuliarité du thème des ressources en TAL est qu'elles sont coûteuses à construire et ont vocation à être utilisées par plusieurs chercheurs, équipes ou institutions. Il y a donc un très fort enjeu d'organisation de la production et de l'accès aux ressources, et un risque de "gaspillage" d'efforts si les ressources sont peu utilisées, ou si l'absence de concertation met des projets de constitution en concurrence.

3 Etat des lieux

La caractéristique la plus importante est l'accès toujours croissant à des données textuelles, sous des formes très variées: documents écrits, oraux, vidéo et des nouveaux media "intermédiaires": conversation par chat, forum, emails, microblogging. Les données collectées par des acteurs académiques ou non sont de plus en plus facilement rendues disponibles, et Twitter est un bon exemple de plateforme qui a généré beaucoup de travaux autour des données qu'elle suscite. Ces données sont de plus facilement croisées avec des informations extra-linguistiques par les réseaux d'échange qu'elles définissent. Elles mettent en jeu des aspects socio-linguistiques porteurs à la fois de promesses applicatives et d'inquiétudes sur la confidentialité et les droits d'utilisation.

Les ressources qui correspondent à des connaissances moins brutes offrent un paysage plus contrasté: si, là encore, le partage devient la norme en TAL, les freins à la production de connaissances réutilisables sont plus nombreux. Annoter des données avec des informations linguistiques est un processus couteux en temps et en main d'œuvre, difficile à mettre en œuvre et encore trop peu valorisé académiquement.

Du fait de ce coût, la production et l'utilisation de ressources s'organise autour de standards, qui émergent en général d'initiatives de production de ressources pour l'anglais par des acteurs majeurs du TAL américains, qui s'imposent ensuite au reste de la communauté. Pour les ressources francophones , la production ou le portage de données en français se fait à partir de ces standards, souvent en taille plus réduite par rapport aux équivalents en anglais. L'alternative d'un format d'annotation original condamne à un manque de visibilité internationale. L'initiative sur le schéma de dépendance syntaxique "Universal dependencies" est particulière à cet égard : sous l'impulsion de Joakim Nivre, la communauté s'est fédérée à partir de propositions de schémas universels d'étiquettage par Google, de dépendances par Stanford.

Le besoin de supervision pour la plupart des approches en apprentissage automatique motive l'invention de nouvelles pratiques de collectes (production participative notamment), qui posent d'autres problèmes. D'un autre côté, la production de connaissances linguistiques utilisables pour des traitements divers, comme des lexiques, dictionnaires, grammaires, pose des défis encore plus importants, en ajoutant également des questions de maintenance, d'accès, et d'utilisabilité par des publics divers.

Un problème important pour la valorisation de telles ressources est notamment le stockage pérenne et accessible. Dans le contexte français, il existe une structure pour cela (Ortolang[1]), et un effort au niveau européen (Clarin[2]) auquel la France n'est pas pleinement associé. Pour la constitution de corpus textuels, on peut souligner les efforts de l'Inist avec la plate-forme Istex, qui donne accès à des textes, scientifiques pour l'instant, avec une extension en cours à des données plus générales. Cependant, ces textes restent non redistribuables, ce qui rend peu réutilisable leur annotation éventuelle.

Des problèmes similaires se posent pour les outils logiciels: maintenance et interopérabilité sont cruciales pour que de telles ressources soient utiles, et nécessitent en amont un effort de normalisation important. Cela a fait l'objet d'un sous-groupe ISO (TC37/4) sur les ressources linguistiques, et de définition de cadres dans lesquels insérer des briques logicielles partageables (par exemple UIMA, et plus récemment dans le cadre du projet OpenMinted[3]).

En parallèle, il faut noter la pratique croissante de partage de code (par exemple, *via* la plate-forme ouverte github), qui favorise la reproductibilité des résultats expérimentaux et améliore grandement les comparaisons entre approches. La pérennité de ces efforts est sans doute faible si les questions ci-dessus ne sont pas réglées.

[1] `https://www.ortolang.fr`
[2] `https://www.clarin.eu`
[3] `http://openminted.eu`

On peut noter bien sûr une grande disparité des ressources selon les langues, l'anglais mobilisant bien plus d'efforts que les autres langues, avec plusieurs chaines de traitement bien répandues et assez complètes (par exemple Core-NLP de Stanford, OpenNLP, plus récemment Spacy). Ces chaines disposent parfois de modèles pour le français, avec des performances plus ou moins dégradées selon les étapes. Pour les données annotées, il existe des corpus de références pour la syntaxe (le corpus Sequoia [Candito and Seddah, 2012], le corpus arboré du français ou French Treebank [Abeillé et al., 2003] - qui n'est pas librement redistribuable),, la syntaxe profonde [Candito et al., 2014], en rôle sémantiques (Asfalda []), pour la substitution lexicale (Semdis []), en relations discursives (FDTB [], Annodis [Afantenos et al., 2012]). On peut citer pour les connaissances de nombreux lexiques généraux: Lefff [Sagot, 2010], thesaurus [], ou bases lexico-ontologiques (Wolf [**?**]), lexiques spécialisés: pour l'opinion/le sentiment (), XXX, ; on peut mentionner aussi des grammaires (...).

Parmi les ressources pour le français créées par production participative, on peut citer le corpus ZombiLingo [Guillaume et al., 2016] annoté en syntaxe de dépendances, le réseau lexical produit par JeuxDeMots [Lafourcade and Joubert, 2008] et tous les lexiques spécialisés en dépendant, dont un lexique polarisé, LikeIt [Lafourcade et al., 2015], et un lexique de sentiments, Emot [Lafourcade et al., 2016].

Ouvert: - question ressources multilingues ? - particularités de l'oral, des LS

4 Grands enjeux

Dans le contexte actuel, les thématiques prégnantes sur les ressources sont liées au besoin quantitatif de données pour les modèles supervisés, et au problème de maintenir une qualité suffisante pour que leur exploitation soit utile, en permettant de développer des modèles généraux et généralisables à de nouvelles données. Cela se traduit par des préoccupations diverses :

- collecte des données : organisation des besoins en ressources humaines, définition des objectifs, des schémas éventuels d'annotation.
- pérennité : stockage, maintenance, partage, standards de repésentation des informations
- contrôle de la qualité : éviter les biais de constitution, définir les usages "normaux", assurer la diversité et la représentativité.

Les enjeux pour le TAL en France sont donc liées à ces problématiques, avec un objectif de fournir une couverture des phénomènes et des applications possibles du TAL.

La communauté constate un manque de ressources sur certains aspects, notamment sémantiques, à grande échelle pour le français, nécessaires pour entraîner et évaluer des modèles statistiques. Ce manque limite le développement de systèmes de traitement sémantique et de compréhension pour cette langue. Les ressources nécessaires comprennent des corpus annotés à grande échelle avec des informations sémantiques, ainsi que des jeux de données encodant des connaissances comme les sens des mots et leurs relations, la paraphrase, l'inférence textuelle et les rôles sémantiques. Les ressources sémantiques (lexiques, dictionnaires, réseaux sémantiques) de bonne qualité actuellement disponibles pour le français sont soit payantes, ce qui limite leur utilisation, soit faibles en couverture, ce qui pose des contraintes pour le traitement de texte libre. Les ressources automatiquement créées sont souvent de qualité moyenne, ce qui rend leur utilisation dans des applications difficile. En outre, la disponibilité de corpus annotés en français est très faible. On retrouve ces problématiques dans le Groupe de Travail sur la compréhension.

Il y a également un besoin de partage des outils et modèles développés au sein des différentes équipes, actuellement pas disponibles pour être réutilisés par d'autres groupes. Le développement de

ressources sémantiques à grande échelle aidera le développement de systèmes de traitement sémantique (en France et à l'étranger). Le partage des outils facilitera la comparaison des résultats obtenus, ce qui fera avancer la recherche en sémantique du français, et favorisera la collaboration entre équipes.

Sur la couverture et la représentativité des données, il faut noter la disparité entre les langues, mais aussi à l'intérieur d'une même langue les questions de diversité régionale ou socio-linguistique sont rarement explicitement posées, alors même que les sources de données textuelles se sont élargies et seraient en théorie propices à prendre en compte plus de variétés de communication.

Une problématique majeure dans un contexte de production de données et de connaissances motivées par les nécessités techniques des modèles d'apprentissage est celle de l'évaluation des ressources produite, de l'estimation de leur "qualité" selon divers critères.

Cela commence par la constitution dans de bonnes conditions, notamment en évitant les biais, et par des questions méthodologiques définies *a priori*. Un aspect souvent oublié est celui des conditions d'utilisation des données, qui devrait imposer certaines limitations (portée des résultats, généralisation) souvent ignorées par les travaux qui les utilisent. Un effort dans l'explicitation des bonnes pratiques sur les données est sans doute nécessaire pour garantir des résultats, et on peut voir quelques frémissement sur ces questions, comme le manifeste "Datasheet for datasets" []. En ce qui concerne la traçabilité, les licences et les conditions de création, la Charte "éthique et big data" fournit un cadre qui permet de mieux documenter les ressources [Couillault et al., 2014].

A contrario, on peut constater certains dérives d'approches venant essentiellement du Machine learning, considérant toute donnée comme bonne et à prendre telle quelle, dont un exemple récent frappant est celui des données d'inférence textuelle, avec des données constituées clairement à la vavite (SNLI), et qui souffrent de biais importants. Ces biais, notés par au moins 3-4 articles récents [], minent la crédibilité de nombreux travaux sur l'inférence, utilisés plus largement comme évaluation d'approches sémantiques. Le problème des biais dans les données dépend bien sûr en premier lieu de la méthode de collecte, qui se décline sous des formes de plus en plus variées: depuis l'annotation experte jusqu'à la production participative, bénévole (à travers des cadres ludiques) ou non (comme Amazon Mechanical Turk), mais aussi parfois par de "l'augmentation de données", essentiellement avec de la supervision indirecte (par exemple, entrainer des représentations sémantiques en entrainant un modèle de langue) dont les biais sont difficiles à estimer.

Il y a là des lacunes méthodologiques qui demandent à être comblées, et qui vont sans doute de pair avec les besoins de traçabilité et d'explicabilité des résultats algorithmiques qui traversent tous les champs de l'intelligence artificielle. Une question reliée est celle de la place des connaissances et des ressources linguistiques élaborées quand la tendance des approches automatiques populaires actuelles est de se fonder sur des ressources assez pauvres (corpus bruts ou avec des prétraitements légers).

Au-delà de la qualité des travaux résultants de ces collectes, les biais posent aussi des questions éthiques, liées au traitement équitable ou non de populations possédant des traits source potentielles de discrimination. Nous revenons sur les aspects éthiques plus loin dans le document.

5 Positionnement

5.1 Éthique

La collecte de données est naturellement source de questions éthiques sur au moins les plans suivants :

- la possibilité d'usages détournés de données fournies volontairement, comme la surveillance de population, la commercialisation sans accord explicite notamment. Les pratiques de surveillance du gouvernement chinois ou le récent scandale impliquant Facebook et Analytica en sont des exemples frappants.

- les biais dans le recueil des données liées à des informations socialement sensibles se retrouvent facilement dans les modèles inductifs et peuvent propager des préjugés ou des stéréotypes, et on en a vu là aussi des exemples spectaculaires dans les problèmes de relations publiques de certains chatbots mal entrainés. Le traitement inéquitable de certaines catégories est plus pernicieux dans les modèles statistiques, et commence à être traité par des approches de "Fair learning", même si le domaine est assez nouveau. On a vu par exemple des articles montrant que les erreurs de reconnaissance faciale sont bien plus importantes sur certains populations que sur d'autres.

- le recueil ou l'annotation de données étant très onéreux, il devient courant d'employer des méthodes de production participative (*crowdsourcing*) en particulier de travail parcellisé (*microworking*), avec tous les problèmes de la juste rémunération du travail que cela représente [Fort et al., 2011].

5.2 Interface

Le TAL est au carrefour de plusieurs disciplines avec lesquelles il a partagé différemment au cours de son histoire: linguistique, traitement de la parole, recherche d'information, apprentissage automatique, ingénierie des connaissances.

Les ressources produites ou utilisées sont donc souvent mutualisables avec ces disciplines, même si en pratique elles sont plutôt dues à un domaine et exportées vers d'autres.

parole: certaines données proches / style oral de certains medium écrits

IC: lien extraction d'information à partir de textes / ontologies

ML: de plus en plus de tâches liées aux données textuelles servent de benchmarks pour approches (défis deft/cap en français par exemple)

Linguistique: controversé sur le lien linguistique / TAL dans les modèles automatiques, mais incontestable sur constitution corpus/annotations cf aussi le récent dépôt du GDR LIFT : "Linguistique informatique, linguistique de terrain"

coordination sur les données ?

6 Programmatique

6.1 Communauté

Un point crucial pour progresser sur les fronts mentionnés ci-dessus est d'arriver à concerter les efforts nécessaires au développement de ressources bénéfiques à l'ensemble de la communauté TAL. Cela peut se décliner sur les trois aspects: corpus, connaissances, outils, avec des modalités propres.

Pour les corpus, si on distingue la création et la diffusion, il est possible de s'appuyer sur l'existant (Istex pour la création, Ortolang pour la diffusion), mais ces cadres sont sans doute sous-exploités, et il faudrait diagnostiquer pourquoi.

Pour les outils, un cadre tel que OpenMinted peut être un point de départ pragmatique pour développer l'interopérabilité, mais on peut s'interroger sur l'étude préalable des besoins et pratiques de la communauté recherche qui a été faite.

Le plus gros travail à faire est sans doute sur le plan des connaissances: comme le montre le GT compréhension certains domaines sont loin de disposer des ressources existant sur l'anglais ou d'autres langues, et c'est là qu'une nécessité de pilotage et de concertation des efforts est la plus évidente pour optimiser la couverture.

point transverse :

valorisation création de données/ressources: comment ?

continuer à encourager partage

6.2 Jeunes chercheurs

Vis à vis des jeunes chercheurs en TAL, quelles sont les spécificités des ressources ? Les JC sont comme les autres susceptibles de contribuer à créer des ressources, que ce soit des corpus avec annotations au moins pour valider certaines expériences. Il faudrait évaluer dans quelle mesure la formation que recoivent les JC prend en compte cette dimension, et si une place est faite dans le contexte d'écoles d'étés par exemple. Ce serait le lieu pour des approches pluri-disciplinaires, notamment en lien avec la communauté plus linguistique (cf section précédente).

6.3 Médiation

Au vu des problèmes que peuvent poser l'exploitation de données personnelles et de l'utilisation croissante des productions textuelles disponibles sur internet (forums, twitter, etc), il y aurait un travail utile de pédagogie à faire vers le public, sur les possibilités d'usages détournés des données fournies volontairement (surveillance, commercialisation), qu'il est sans doute difficile d'envisager concrètement sans connaissances des possibilités techniques.

Par ailleurs, les efforts vers l'explicabilité des prédictions des modèles statistiques sont importants dans un contexte social, pour surveiller les reproductions de biais sociaux des modèles, voire la création de nouveaux biais qui passeraient inaperçus.

References

[Abeillé et al., 2003] Abeillé, A., Clément, L., and Toussenel, F. (2003). *Building a Treebank for French*, pages 165–187. Springer Netherlands, Dordrecht.

[Afantenos et al., 2012] Afantenos, S. D., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, M., Draoulec, A. L., Muller, P., Péry-Woodley, M., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M., and Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2727–2734.

[Candito et al., 2014] Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., and de la Clergerie, E. (2014). Deep syntax annotation of the sequoia french treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

[Candito and Seddah, 2012] Candito, M. and Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France.

[Couillault et al., 2014] Couillault, A., Fort, K., Adda, G., and De Mazancourt, H. (2014). Evaluating Corpora Documentation with regards to the Ethics and Big Data Charter. In *International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Islande.

[Fort et al., 2011] Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2):413–420.

[Guillaume et al., 2016] Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japon.

[Lafourcade and Joubert, 2008] Lafourcade, M. and Joubert, A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France.

[Lafourcade et al., 2015] Lafourcade, M., Le Brun, N., and Joubert, A. (2015). Collecting and evaluating lexical polarity with a game with a purpose. In *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Hissar, Bulgaria.

[Lafourcade et al., 2016] Lafourcade, M., Le Brun, N., and Joubert, A. (2016). Construire un lexique de sentiments par crowdsourcing et propagation. In *Proc. of Traitement Automatique des Langues Naturelles*, Paris, France.

[Sagot, 2010] Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.