PréGDR TAL
Axe de réflexion : Modèles computationnels de la langue

Nicholas Asher

Juin 2018

1 Participants

- Gerhard Jäger, Department of Linguistics , Universität Tübingen,
- James Pustejovsky, Department of Computer Science, Brandeis University
- Gemma Boleda, Dep, Department of Translation and Language Sciences, Pompeu Fabra University, Barcelona.
- Nabil Hathout, CNRS, laboratoire CLLE,
- Olivier Bonami, Université Paris 7, LLF.
- Pascal Denis, INRIA, Magnet.
- Jean Luc Schwarz, CNRS, Gipsalab Grenoble
- Sylvain Pogodalla, INRIA, Nancy.
- Philippe de Groote, INRIA, Nancy.
- Jean-Philippe Prost, Université de Montpellier, LIRMM
- Benoit Crabbé, Université Paris 7, LLF

2 Thèmes et périmètre d'étude

1/2 page
Cette axe de réflexion concerne les théories computationnelles pour la langue. Donc son périmètre est large; il comprend: la linguistique historique, la perception et la production de la parole, la phonologie, la morphologie, la syntaxe, la sémantique, la pragmatique et le discours/dialogue. Je n'ai pas eu de retours sur la phonologie faute de l'organisateur de ce thème. Par contre, sur les autres sous-thèmes j'ai eu des retours très interessants.

Un grand débat se situe autour de la question et le rôle des représentations symboliques voir logiques (lambda calcul) vis à vis des représentations statistiques voir auto-construites à partir de réseaux neuronaux ou autre méthode non supervisé (par exemple la réduction de dimensions en utilisant la technique Nonnegative matrix factorization (NMF) ou Singular Value Decomposition (SVD). Il y a des experts des deux côtés du débat comme ce rapport montrera. Il y a aussi des questions techniques et conceptuelles relevées comme la distinction entre la recursivité et la compositionalité, question importante car presque tout modèle de la langue comprend un élément de recursivité ou de compositionalité.

3 État des lieux

1 à 2 pages
Pour chaque thème où en est d'un point de vue international
Les travaux de référence, logiciels

3.1 Le débat sur les modèles de la langue

D'abord le débat sur le rôle des représentations construites de manière non supervisée. Voici les mots sur ce sujet de James Pustejovsky, qui est mondialement reconnu pour ses travaux en sémantique lexicale et qui a une expertise très large sur la computation et linguistique :

> I have worked in a number of CL, linguistic, and AI areas as funded by NSF, DARPA, IARPA, NIH, NGA, Dept. of Energy, and Andrew Mellon Foundation. Given the overwhelming concentration of attention, money, and media on deep learning from big data, it is hard to escape the feeling that everything will eventually "fall" to neural net learning. My feeling is very much the opposite. Neural nets as currently designed are not embedded within the "model-bias" which allows us (and other species) to do one-shot and two-shot learning, along with what I call "low-resource learning" problems. The model bias is the background radiation of communication. It doesn't show up in the data until you have a theoretical model of what you're looking for. This is what characterizes the problems of any discourse, dialogue, or multi-modal communication.

A la voix de Pustejovsky s'ajoute la voix de Denis Pascal, expert en apprentissage supervisé et non supervisé pour la sémantique et le discours : comme beaucoup de chercheurs, il trouve que la manque d'explicabilité dans les modèles non supervisés neuronaux nuit à la compréhension des phenomènes langagiers. La communauté TAL a priviligié la "prédictibilité" du modèle plutôt que sa force explicative. C'est un problème et un enjeu pour le futur.

Leur scepticisme n'est pas universellement reconnu. Voici une opinion de l'autre coté par Gemma Boleda, jeune lauréate d'un ERC et experte en sémantique lexicale:

> As is well known, in recent years there has been a huge revolution in AI, the deep learning revolution, that has swept the field. There is now a kind of mismatch between "old", rule-based and even statistical based approaches, and "new", continuous or deep learning models. Many researchers in the field feel that deep learning "works" but are suspicious because they see these models as black boxes. However, they have brought about improvements: 1) advances in computational models of language, with empirical improvements across the board (Machine Translation, language modeling, word representations, etc.); 2) easier cross-fertilization with other fields, in particular Computer Vision and Speech Processing, since in these fields deep learning has also proven successful and it is easy to build a single model that works with different types of information, either simultaneously (e.g. integrating visual information into semantic representations for words like "apple") or providing a mapping between the two (for instance, identifying the referent of a noun phrase in a picture). ... This is in my opinion not by chance or because of some quirk, but because they provide a better model of natural language than previous models, both symbolic / rule-based and previous-generation Machine Learning / statistical systems.

Ce débat a des échoes sur toutes les modélisations de la langue en TAL. Même dans la "historical linguistics", la méthode comparative bien confirmée de (Meillet 1954), qui itère sept étapes de recherche en séquence, est en train d'être remplacé au moins en partie par des méthodes statistiques (Jäger & List, 2016), bien que pour le moment il ne s'agit pas de "deep learning" pour cette sous-discipline par manque de données. Cependant en morphologie (Bonami & Sagot 2016), la compétition entre systèmes symboliques et systèmes de deep learning est d'actualité Les modèles "finite state transducers" pour la morphologie de Karttunen (2003) font face aux implémentations avec une logique nonmonotone simple comme celle de DATR. Comme la morphologie suppose des structures formelles relativement simples, les méthodes symboliques ont eu du succès dans ce domaine et continuent à être populaires. Néanmoins la construction des systèmes morphologiques de données non annotées commence à prendre de l'élan. En particulier Malouf (2016) a réçamment montré comment utiliser le cadre neuronal d'RNN (recursive neural network) pour construire des morceaux d'un système inflectionnel. Le réseau a comme entrée un pair d'un lexème identifiant et un "paradigm cell"; la sortie est une forme phonologique. Malouf donne un modèle pour des ensembles de données dans sept langues différentes avec un taux d'exactitude entre 86% et 99.9 %.

Passons maintenant à la syntaxe. L'évolution récente des travaux en analyse syntaxique automatique en TAL est surtout liée aux évolutions en modélisation statistique. Dans ce contexte, on a vu progressivement disparaître le recours à des grammaires formelles explicites. L'enjeu des modèles statistiques récents n'est plus tellement de prédire la grammaticalité de phrases comme cela était proposé initialement par Chomsky. En TAL, les modèles d'analyse syntaxique sont plutôt formulés dans le but de prédire une structure à partir d'une donnée qui est une séquence de mots. Le but recherché est la structuration de gros volumes de texte et la robustesse de traitement. Les modèles emblématiques qui illustrent ce changement de perspective sont les modèles d'analyse en dépendances. Les progrès réalisés et les jeux de données récemment réalisés comme le corpus multilingue Universal Dependencies ont considérablement facilité le traitement multilingue.

L'arrivée du Deep Learning n'a cependant pour l'instant pas changé fondamentalement la nature des algorithmes d'analyse syntaxique statistiques. Ils ont par contre contribué a améliorer les performances de manière très substantielle. Par contre ce nouveau paradigme pose une question sur la nature des structures et des représentations utilisées. Peut-on se passer de structures symboliques comme des structures en arbres lorsqu'on sait que des modèles de réseaux de neurones récurrents (LSTM) sont capables d'approximer des motifs de parenthésages typiques des langages libres de contexte. Mais l'interprétation et l'analyse de ces observations reste à être éclaircie.

Pour Jean Philippe Prost, "soit le deep learning nous met tous au chômage d'ici quelques années ... et le TAL devient rien de plus qu'une application industrielle supplémentaire ; soit, comme l'évoque notamment Chris Manning (2015), le langage pourrait s'avérer plus coriace que prévu (notamment parce que pas typiquement un "pur" problème de traitement du signal) et les années à venir pourraient voir un retour des investigations scientifiques et cognitives du langage (ce que Manning appelle de ses voeux)." Prost trouve que les modèles pour la syntaxe évoluent selon deux axes intéressants : - l'interface syntaxe-sémantique semble prendre un role plus important via son interaction avec le parsing sémantique, ou l'interconnexion de bases de connaissances. Et plus généralement il voit un avance vers une meilleure intégration de la syntaxe, de la sémantique, et des différentes dimensions linguistiques en général. Comme exemples pertinents, il mentionne les Sar-graphs (Uszkoreit et Xu, 2013 ; Krause et al., 2015), et le projet ERC de Laura Kallmeyer TREEGRASP, "Tree rewriting grammars and the syntax-semantics interface: From grammar development to semantic parsing", qui s'appuie sur des théories formelles (RRG et TAG) et le développement d'une méta-grammaire pour contribuer à des procédés de TAL statistique pour le parsing sémantique..

En sémantique en TAL vu les publications dans les conférences de réference (ACL, EMNLP, EACL, NAACL) on dirait que les modèles symboliques ne sont plus du tout d'actualité. Mais est-ce que les modèles symboliques n'ont plus rien à offrir la communauté TAL en sémantique? Il est intéressant de comparer l'état de la discipline avec par exemple la physique ou les modèles abstraites et leur approximations statistiques même neuronaux co-existent sans compétition. Pour donner un exemple un peu plus concret, la mécanique quantique donne un modèle abstrait qui doit prédire les proprietés d'une nano-structure particulière, mais les calculs pour faire cette prédiction sont intractables analytiquement. Donc les physiciens font recours à des méthodes d'apprentissage, voir des réseaux neuronaux, pour approximer le modèle abstrait. Il pourrait être ainsi pour les réseaux neuronaux en sémantique vis a vis le model abstrait de la sémantique proposé par la sémantique de la grammaire de Montague ou de la sémantique dynamique (voir Asher et al. 2016), sémantiques qui se basent sur des idées de sens et de référence qui ont été dévelopées pendant des millénaires.

Si nous passons à l'interprétation du discours ou a la pragmatique, nous voyons que pour le moment les données manquent pour construire des représentations discursives avec un contenu sémantique et pragmatique de façon complètement automatique. C'est-à-dire que l'apprentissage automatique des structures discursives et pragmatiques se fait de manière supervisée dans la grande majorité des cas. Cependant il faut noter le papier de Liu et Lapata dans *TACL* 2018 qui montre comment des architectures neuronales peuvent encoder des informations structurales pour des documents qui servent à des classifications de documents. Ceci dit, il semble que les architectures neuronales sont loins d'être capables de retrouver une structure discursive pour un texte en comparaison avec des methodes plus classique d'apprentissage supervisée avec des données annotés du point de vue d'un modèle abstrait de la structure discursive et son interprétation. Pour d'autres domaines de la pragmatique, par exemple la détection des implicatures ou des sens qui dépendent sur des informations extra-linguistiques voir sociolinguistiques, les modèles qui traitent ces tâches sont des modèles issus de la théorie des jeux (Burnett, à paraître; Asher et al., 2017; Asher & Paul, 2018) qui résistent à un traitement informatique pour le moment.

3.2 Nouveaux principes pour l'élaboration de modèles et de systèmes de perception et de production de la parole

Pour regarder de plus près ce débat, il est utile de regarder de plus près les modèles des phénomènes langagiers qui sont plus proches de l'architecture neuronale humaine, résumés ci-dessous par Jean Luc Schwartz.

Oscillations neuronales et codage multiplexe Le calcul neuronal apparaît organisé autour de mécanismes de synchronisations collectives dans des bandes de fréquence opérant jusqu'à un certain point en parallèle – bandes delta, theta, alpha, beta, gamma entre 1 et 120 Hz – qui semblent en charge de types de traitement spécifiques : détection syllabique dans la bande theta entre 4 et 8 Hz, caractérisation phonétique dans la bande gamma autour de 40 Hz, suivi des fluctuations prosodiques lentes dans la bande delta entre 1 et 3 Hz, processus attentionnels associés à la bande alpha vers 10 Hz, mécanismes prédictifs dans la bande beta vers 20 Hz. Ce principe de "codage multiplexe", avec bien évidemment des processus de couplage entre bandes, ouvre des perspectives nouvelles dans l'étude des traitements neuronaux, tant pour la perception que pour le contrôle.

Hiérarchies de traitement, interactions multisensorielles et couplage perception-action Ces principes de synchronisation sont associés à des mécanismes de propagation d'information à large distance dans le cerveau, en lien

avec de nombreuses connaissances nouvelles sur les architectures corticales sous-jacentes. On y retrouve des principes classiques de structuration hiérarchique et de combinaison de processus feedforward et feedback, bien intégrés dans le cadre des théories cérébrales du codage prédictif. Il apparaît également des principes nouveaux et potentiellement importants pour le développement de modèles computationnels : l'existence à tous niveaux de traitement de mécanismes d'interaction multisensorielle, et l'existence de liens structurants entre connaissances perceptives et motrices qui montrent que systèmes de perception et de production de la parole fonctionnent de manière étroitement complémentaire.

Repère développementaux Dans ce contexte, la connaissance des étapes permettant aux systèmes de perception et de production de la parole de se structurer – et très probablement de se co-structurer – dans le développement est cruciale pour la modélisation. Ces étapes peuvent servir de guide aux modèles computationnels en leur fournissant des repères temporels majeurs : connaissances sur l'état initial, perceptual narrowing permettant de faire converger sur la langue cible, les unes après les autres, les représentations prosodiques, les voyelles, les consonnes, puis l'acquisition des mots et des principes de combinaison syntaxique.

3.2.1 Nouveaux cadres de modélisation

Adossés à ces nouveaux principes de neurocalcul, ont émergé de nouveaux cadres de modélisation cognitive et neurocognitive.

Modélisation bayésienne La modélisation bayésienne permet de formuler précisément des hypothèses, de séparer état initial, connaissances acquises et principes d'inférence. Les modèles bayésiens, sous des formes variées, sont applicables à la perception de parole, à la lecture, à la compréhension du langage, mais aussi plus récemment à la production de la parole, aux interactions multisensorielles et aux relations percpetion-production.

Neurocalcul Les mécanismes de neurocalcul adossés aux principes d'oscillations et de synchronisations débouchent naturellement sur la mise en œuvre de systèmes dynamiques capables de générer des processus oscillatoires au niveau du neurone individuel ou du systèmes de neurones (de type colonne corticale). Ces principes se retrouvent d'ailleurs assimilés au sein des réseaux de neurones artificiels avec les réseaux récurrents et Long Short-Term Memory networks.

Robotique cognitive et systèmes multi-agents L'inscription des modèles dans un cadre développemental et perceptuo-moteur conduit tout naturellement vers la robotique – outil naturel pour traiter de liens sensori-moteurs, de relations entre variables proximales et distales, de coordination entre contrôle et perception, de méthodes d'exploration et de réduction de dimensionnalité – avec les déclinaisons que sont robotique développementale, cognitive et sociale, et systèmes multi-agents pour traiter des processus d'interaction.

Il faudrait remarquer que ces nouveaux cadres de modélisation pour la perception et production de la parole vaudrait aussi pour d'autres sous disciplines de la linguistique, voir la sémantique ou l'interprétation et l'étude de l'interaction.

4 Grands enjeux

1 à 2 pages
Le projet de recherche
Quels sont le ou les grands enjeux où on doit aller en fonction des forces existantes en France, où il faudrait aller.

4.1 Challenges généraux

Voici quelques perspectives pour la syntaxe. Pour Benoit Crabbé et Jean Philippe Prost, il prévoit que les travaux en analyse syntaxique "se reformulent plus directement comme une composante de modèles d'analyse sémantique pour lesquels l'hypothèse d'une structuration en arbres (ou en DAGs) apporte quelque chose" en donnant un structure pour construire une représentation sémantique de façon compositionnelle (avec un homomorphisme de la structure syntaxique dans l'interprétation sémantique qui peut etre soit representée par une formule logique comme traditionnellement (Liang et al 2015) soit via fonctions vectorielles (Socher et al. 2013).

Crabbé pointe aussi sur la perspective où les modèles d'analyse syntaxiques seront utilisés dans un contexte expérimental d'études cognitives, psycholinguistiques et neurolinguistiques pour traiter des questions liées aux hypothèses de structuration du langage. Et ceci me semble une question importante dans le débat qu'anime cette trame: 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 ? (illustration donnée par la discussion de Frank and Christiansen 2018).

Quel apport des modèles d'apprentissage profond et comment les interpréter (Linzen et al 2016) ? Et comment mettre ces modèles en relation avec les modèles cognitifs de traitement de la mémoire (Lewis and Vasishth 2005) ?

Pour James Pustejovsky, et d'autres un challenge général est de développer des modèles computationnels qui intègrent action, perception, gestuel, langage pour passer des réprésentations linguistiques à des modèles de communication. Donc il faudrait pousser vers une théorie intégrée de la communication et interprétation multimodale en dialogue qui prend en compte le contexte physique, sociale, et interationnel d'une conversation. Boleda mentionne aussi l'idée qu'il faudrait peut etre regarder les n-grams de caractères, c'est a dire des parties des mots ou sequences de mots pour des entrées au deep learning comme a proposé Hinrich Schütze.

En même temps il faudrait regarder de plus près la relation entre modèle abstrait de la langue et son approximation par un réseau neuronal ou se demander est-ce que le TAL n'a pas besoin de modèle abstrait. Une sous-défi de ce dernier challenge est de se demander quelles sont les limites du deep learning vis à les phénomènes langagiers (voir l'article de Linzen et al (2016); voir aussi Paperno et al. (2016) qui donnent l'établissement de référents dans un dialogue et le phénomène d'attachement longue distance en discours comme des défis non-résolus pour le deep learning). Abrusan et al (2017) donnent aussi l'exemple des inférences logiques validés par la sémantique abstraite des connecteurs et déterminants qui resisteraient aux méthodes d'aquisition de sens par généralisation sur des corpus en utilisant soit des méthodes statistiques classiques soit du deep learning.

4.2 Challenges dans la production et reconnaissance de la parole

Sur cette base on voit apparaître de nouveaux challenges pour le développement de modèles et de systèmes pour la perception et la production de la parole et la gestion des processus d'interaction langagière

Robustesse et plasticité Les technologies d'inspiration cognitive peuvent se différencier des systèmes provenant de principes d'apprentissage massif par leurs capacités de robustesse aux bruits et d'adaptation aux perturbations, tant en perception qu'en production, en lien avec les enjeux cliniques, les outils de réalité augmentée, les systèmes virtuels pour la compensation des handicaps et pour la rééducation.

Apprentissage Comment développer des systèmes capables d'apprendre à produire et percevoir de la parole en suivant les étapes décrites par les spécialistes du développement ? Mais aussi comment résoudre le problème du développement conjoint de toutes les composantes du langage ? Et du coup comment résoudre des problèmes similaires d'apprentissage complet de langues à partir d'exemples ? Et notamment comment résoudre ces problèmes avec peu de données, en tout cas moins que les systèmes "couteux" de l'apprentissage massif – et sans doute comme les résout le bébé, qui apprend sa langue avec beaucoup moins de données que les systèmes du commerce ?

Origine du langage Peut-on enfin utiliser ces développements pour progresser sur la question de l'émergence du langage et de la morphogenèse des systèmes sonores des langues du monde ? La modélisation est en effet un outil essentiel du test d'hypothèses et donc du progrès des connaissances, dans un domaine par nature non susceptible d'expérimentations contrôlées.

5 Positionnement

1/2 page

5.1 Éthique

questionnement sur les problèmes éthiques, sociétaux et légaux

5.2 Interface

Autre GDR, autre institut

6 Programmatique

1 page
Actions à entreprendre

6.1 Communauté

6.2 Jeunes chercheurs

Comment intéresser les jeunes chercheurs à ce projet

6.3 Médiation scientifique

Comment communiquer sur ce projet au grand public

7 Références

N. Asher & S. Paul, 'Strategic conversations under imperfect information: Epistemic Message Exchange Games', *Journal of Logic, Language and Information*, (43 pages), doi: 10.1007/s10849-018-9271-9.

N. Asher, S. Paul & A. Venant, 'Message Exchange Games', *Journal of Philosophical Logic*, 2017, vol 46.4, pp.355-404, doi:10.1007/s10992-016-9402-1

N. Asher, T. van de Cruys, A. Bride & M. Abrusan (2016). 'Integrating type theory and distributional semantics: a case study on adjective-noun compositions', *Computational Linguistics*, 42.4, 2016, pp.703-725

H. Burnett (forthcoming) 'Signalling Games, Sociolinguistic Variation and the Construction of Style', *Linguistics Philosophy*

Frank and Christiansen (2018). 'Hierarchical and sequential processing of language A response to: Ding, Melloni, Tian, and Poeppel (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience'. *Language, Cognition and Neuroscience*. .

Richard L. Lewis and Shravan Vasishth (2005). ' An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval', *Cognitive Science* 29 (2005).

Percy Liang and Christopher Potts (2014). 'Bringing machine learning and compositional semantics together'. *Annual Review of Linguistics*.

Tal Linzen, Emmanuel Dupoux Yoav Goldberg (2016). 'Assessing the ability of LSTMs to learn syntax-sensitive dependencies'. *Transactions of the Association for Computational Linguistics* 4, 521–535.

Mikolov, T., Yih, W., & Zweig, G. (2013). 'Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Atlanta, Georgia: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/N13-1090

Paperno, D., G. Kruszewski, A. Lazaridou, Q. Ngoc, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, R. Fernandez. 2016. 'The LAMBADA dataset: Word prediction requiring a broad discourse context'. *Proceedings of ACL 2016 (54th Annual Meeting of the Association for Computational Linguistics)*, 1525-1534, Berlin, Germany, August. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts, Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)