Préface

1. Introduction

L'apprentissage automatique (ou apprentissage artificiel) est, suivant la définition de Tom Mitchell dans (Mitchell, 1997), l'étude des algorithmes qui permettent aux programmes de s'améliorer automatiquement par expérience. Le domaine a connu ces dernières années un développement considérable, et ses interactions avec le TAL sont de plus en plus étroites et fréquentes, comme l'illustre par exemple (Manning et Schütze, 1999).

Du côté des linguistes, les intérêts pratiques de ce rapprochement sont nombreux. En effet, la constitution manuelle de ressources spécifiques à une langue donnée est une tâche longue et fastidieuse, qui doit être recommencée pour chaque langue différente, et pour chaque sous-domaine spécifique d'une langue. À condition de disposer de données initiales suffisantes et adaptées, l'apprentissage automatique offre une alternative séduisante. Il permet d'obtenir ou d'améliorer à moindres frais des ressources, et de s'assurer qu'elles sont robustes et à large couverture. La démarche inductive, employée depuis longtemps en linguistique de corpus, peut ainsi être opérationnalisée à grande échelle, et son efficacité évaluée de façon systématique. Dans sa composante plus théorique, l'apprentissage automatique contribue aussi, *via* certains résultats d'apprenabilité de classes de grammaires formelles, aux débats sur l'acquisition des langues récurrents depuis les années 50.

De leur côté, les spécialistes de l'apprentissage automatique voient dans le TAL un domaine d'application privilégié, pourvoyeur potentiel de problèmes difficiles et de grandes quantités de données. La fouille de textes a ainsi été à l'origine d'innovations conceptuelles importantes ces dernières années. Mais la prudence est souvent de mise quant à l'interprétabilité des résultats obtenus. Les méthodes employées sont de plus en plus fondées sur des mathématiques raffinées, apparemment réservées aux spécialistes. Dans ce contexte, la linguistique a-t-elle encore son mot à dire ? Comment combiner des connaissances linguistiques déjà acquises avec des programmes d'apprentissage automatique ? Quand bien même il peut les interpréter, quelle confiance un linguiste peut-il accorder aux résultats de ces programmes ?

C'est pour affronter ce questionnement contemporain que la revue TAL a décidé de consacrer un numéro aux *relations entre apprentissage automatique et traitement automatique des langues (TAL)*, particulièrement quand ils s'intéressent tous les deux aux textes. Six articles (parmi quatorze soumis), très représentatifs des différentes tendances actuelles, ont été sélectionnés. Mais, avant de présenter leur

contenu, il nous a semblé utile de faire un petit détour historique et réflexif pour comprendre les liens qu'entretiennent depuis leurs origines l'apprentissage automatique et le TAL. La première partie de cette introduction est donc consacrée à un survol historique comparatif des deux disciplines. Elle se focalise toutefois prioritairement sur l'apprentissage automatique, supposé moins familier aux lecteurs de la revue. Dans un deuxième temps, nous aborderons les problématiques des travaux contemporains, qui renouvellent complètement les relations entre les deux domaines. Il ne restera plus qu'à conclure en laissant la place aux contributions sélectionnées.

2. Une brève histoire de l'apprentissage automatique et du TAL

L'apprentissage automatique et le TAL partagent le projet de doter les machines de certaines capacités humaines évoluées. En ce sens, ils sont tous les deux les héritiers plus ou moins directs de l'intelligence artificielle. Cela fait ainsi près de 60 ans que les deux domaines cohabitent au sein d'une même communauté de recherche. Nous verrons pourtant que, malgré cette filiation commune, ils ont longtemps suivi des chemins parallèles avant de commencer à collaborer de manière fructueuse, depuis une vingtaine d'années. Pour ce rapide et légèrement acrobatique survol historique, nous nous appuierons notamment sur (Crevier, 1999) et (Cornuéjols et Miclet, 2002).

2.1. *Les intuitions fondatrices*

Le langage et l'apprentissage sont des capacités fondamentales des êtres humains, et ont d'ailleurs été identifiés comme telles dès les tout premiers projets de construction d'une « machine intelligente ». L'article programmatique fondateur de l'intelligence artificielle, le fameux *Computing Machinery and Intelligence* d'Alan Turing (Turing, 1950), les évoque tous les deux de manière plus ou moins indirecte. Il commence par introduire le « jeu de l'imitation » qui sera plus tard reformulé en « test de Turing », et que l'on pourrait résumer ainsi : un agent artificiel pourra être considéré comme intelligent s'il est indiscernable d'un être humain lors d'une interaction langagière à distance. Bien que Turing ne le dise pas explicitement, et ne semble pas y accorder beaucoup d'importance, ce test donne au langage naturel un statut particulier : c'est un peu plus que le véhicule de la pensée, c'est en quelque sorte son *symptôme*, sa manifestation la plus incontestable. Plus intéressant encore, et rarement rappelé : dans le même article, après avoir passé en revue les mauvaises raisons de ne pas croire à l'existence possible d'une machine intelligente, Turing s'attaque à la difficulté probable de la programmer. Il se lance alors dans un plaidoyer en faveur d'une « machine-enfant » dont la compétence principale serait d'être capable d'*apprendre*, et qu'il suffirait donc d'éduquer correctement pour qu'elle atteigne, voire dépasse, les performances des adultes humains. La toute fin du texte (qui mérite décidément d'être régulièrement relu) évoque même brièvement

le langage comme un des premiers domaines qui pourraient lui être enseignés. Les chercheurs en intelligence artificielle n'ont cessé de réinventer ce rêve, sous différentes formes.

La caractérisation du *principe d'induction*, en vertu duquel on peut inférer des règles générales à partir d'exemples particuliers, est un problème qui mobilise les philosophes depuis au moins Hume, *via* notamment Popper. La formalisation des probabilités par Kolmogorov, dans les années 1930, est née aussi de cet effort, tandis que les travaux de Shannon permettent, lors de la décennie suivante, de mesurer la quantité d'information d'un message. Mais l'histoire de l'apprentissage automatique proprement dit commence sans doute avec McCulloch et Pitts qui introduisent, en 1943, un modèle formel élémentaire du fonctionnement des neurones à base de « rétroaction » (McCulloch et Pitts, 1943). Cette intuition inspire aussi les pionniers d'une « science cybernétique » qui, à l'instar de Norbert Wiener, tentent dans les années 50 de construire des animaux artificiels doués de capacités d'apprentissage par essais/erreurs. La psychologie de cette période, dominée par le behaviorisme et l'apprentissage par renforcement, va dans le même sens. C'est aussi à la même époque qu'Arthur Samuel, à IBM, développe un programme de jeu de dames américain dont la fonction d'évaluation s'améliore par la pratique.

2.2. *Les débuts incertains*

Les années 60 sont une période de rationalisation et de remise en question. Elles voient émerger à la fois les premiers modèles théoriques de l'apprentissage automatique et les premiers résultats qui montrent leurs limites. Ainsi, après avoir été promu par Rosenblatt, le modèle des perceptrons, ancêtre des réseaux de neurones artificiels, est sévèrement critiqué par Minsky et Papert dans un livre (Minsky et Papert, 1969) qui entraîne un arrêt de 15 ans des recherches sur le sujet.

De son côté, en posant les bases des « probabilités algorithmiques » et de l'« inférence inductive », Solomonoff contribue à formaliser les conditions de l'apprentissage (Solomonoff, 1964). Cet objectif est aussi celui de Gold, qui propose de modéliser l'acquisition de leur langue maternelle par les enfants *via* la notion d'apprenabilité « à la limite » de classes de grammaires (Gold, 1967). L'intérêt de cette formalisation est apparemment atténué par les résultats négatifs qui l'accompagnent : dans le modèle de Gold, aucune des classes de grammaires de la hiérarchie de Chomsky-Schützenberger n'est apprenable par exemples positifs seuls, c'est-à-dire à partir d'exemples de phrases syntaxiquement correctes d'une langue quelconque.

Ce résultat corrobore, en quelque sorte, les prises de position de Chomsky lui-même qui, à la même époque, s'attaque de front au behaviorisme. S'il n'a jamais travaillé sur l'apprentissage automatique proprement dit, on lui doit l'argument de la « pauvreté du stimulus », selon lequel les enfants seraient exposés à de bien faibles données, en regard des remarquables capacités langagières qu'ils acquièrent en un

temps record. Cela justifie, à ses yeux, l'existence d'une « capacité de langage » innée et spécifique à l'espèce humaine (Chomsky, 1980 ; Piatelli-Palmarini, 1979). Cet argument a le mérite de mettre l'accent sur la complexité de la tâche d'acquisition d'une langue naturelle, qui avait été largement sous-estimée par les tenants de l'apprentissage par renforcement. Même s'il est aujourd'hui contesté (Pullum, 2002), il a souvent été repris par des praticiens de l'apprentissage automatique, pour justifier des *biais* ou *connaissances a priori* qu'ils intégraient à leurs programmes.

La statistique textuelle se développe dès les années 60-70 (Benzecri, 1982). Mais la communauté de recherche qui se constitue alors (encore représentée de nos jours par les conférences JADT) n'interagit pas vraiment avec les théoriciens de l'apprentissage automatique ni avec les linguistes de la tradition chomskyenne. Le traitement de la parole (dans la lignée de laquelle se développeront les conférences JEP) commence aussi très tôt à faire appel à des « modèles de langues » promis à un certain avenir.

Mais, de manière générale, les années 70 sont marquées en intelligence artificielle par la prédominance des modèles symboliques de représentation des connaissances. C'est aussi le cas en linguistique formelle, que ce soit pour l'expression de la syntaxe (formalismes LFG, HPSG, grammaires catégorielles, TAG, etc.), de la sémantique (réseaux sémantiques, formalismes de Schank, *frames* de Minsky, graphes conceptuels de Sowa, etc.), ou de leurs relations (Winograd, Montague, etc.). Et c'est vrai également en apprentissage automatique symbolique où les travaux pionniers ne manquent pas. Les plus connus sont ARCH, de Wilson, qui apprend à reconnaître les empilements de blocs qui constituent une « arche », les programmes de découvertes mathématiques AM puis EURISKO de Lenat, ou encore META-DENDRAL de Mitchell, dédié à l'acquisition de règles pour un système expert. Mais ces avancées, à base d'heuristiques, sont plus empiriques que conceptuelles. Et les programmes conçus sont toujours très spécifiques des domaines sur lesquels ils visent à acquérir des connaissances.

Malgré certains partis pris communs, le TAL et l'apprentissage automatique interfèrent donc encore assez peu entre eux ou alors, un peu plus tard, dans le cadre de modèles généraux de la cognition (ACT d'Anderson ou SOAR de Newell) qui, malgré leur ambition, n'ont pas vraiment donné lieu à des applications pratiques.

2.3. *Le retour de l'apprentissage automatique*

Dans les années 80, c'est presque simultanément que les premiers résultats négatifs des années 60 sont contrebalancés par de nouveaux plus favorables : les réseaux de neurones réémergent alors, accompagnés de nouveaux algorithmes d'inférence par descente de gradient plus puissants et efficaces que les précédents, tandis qu'Angluin montre que certaines classes non triviales de grammaires sont tout de même apprenables par exemples positifs seuls dans le modèle de Gold (Angluin,

1980 ; Angluin, 1982). Par ailleurs, Valiant propose un nouveau modèle de l'apprenabilité au sens PAC (« probablement approximativement correct ») (Valiant, 1984), plus réaliste que celui de Gold.

Ces avancées sont assez représentatives des travaux en apprentissage automatique, et plus généralement en intelligence artificielle, dans ces années-là. D'un côté, avec les réseaux de neurones artificiels, on dispose de techniques d'apprentissage « numériques » opérationnelles et efficaces sur les données réelles, mais dont les résultats sont difficiles à interpréter. De l'autre, avec les modèles symboliques dont est issue, entre autres, l'inférence grammaticale ou, un peu plus tard, la PLI (programmation logique inductive), on a accès à des résultats théoriques bien fondés, accompagnés de théorèmes garantissant une certaine convergence et donnant lieu à des objets compréhensibles, mais dont les algorithmes sont difficiles à mettre en œuvre en pratique, parce qu'ils sont d'une complexité élevée et requièrent des données non bruitées.

Cette dichotomie reflète le débat, très prégnant dans les années 90, entre approches « connexionniste » et « cognitiviste ». L'idée qui prédomine alors est que les modèles de type connexionniste, de par leur inspiration dans le substrat « matériel » du fonctionnement du cerveau humain, sont plus aptes à modéliser des facultés « de bas niveau » comme les perceptions sensorielles. Mais, pour la représentation des connaissances ou le raisonnement, ce sont plutôt les modèles symboliques qui sont encore privilégiés. Les deux approches ne sont pourtant pas incompatibles. Comme le formule alors explicitement Smolensky dans une tentative de synthèse (Smolensky, 1992), un « symbole » n'est peut-être rien d'autre qu'une étiquette associée à une configuration globale, stabilisée par apprentissage, d'un réseau de neurones. La connaissance symbolique est dans ce cas envisagée comme *le passage à la limite, l'horizon de l'apprentissage numérique ou statistique* qui n'en est qu'une approximation imparfaite et provisoire.

2.4. *Le triomphe de l'apprentissage automatique*

L'intelligence artificielle a connu depuis lors une mutation profonde. L'objectif initial de reproduire, voire d'imiter, les capacités de l'esprit humain (parfois désigné aussi comme le projet de l'« IA forte »), a laissé progressivement la place à l'objectif plus pragmatique de tirer le meilleur profit possible des capacités spécifiques des ordinateurs (« IA faible »). On est en quelque sorte passé de l'« intelligence artificielle » à l'« intelligence des machines », tandis que les sciences cognitives ont pris le relais dans le champ de l'étude et de la modélisation de l'esprit humain (Gardner, 1993).

Or, ces capacités spécifiques des ordinateurs sont plutôt à chercher du côté des possibilités de stockage, de traitement et d'échange de données. Autant de paramètres qui, justement, atteignent des seuils critiques dans les années 90, au moment où Internet et les ordinateurs individuels se banalisent. Cette évolution est

sensible dans tous les domaines de l'intelligence artificielle. Que ce soit pour la reconnaissance des formes, le raisonnement, la programmation de stratégies pour les jeux, etc. : la démarche empirique, « *bottom-up* », fondée sur la force brute du calcul et l'accumulation d'exemples prend alors partout le pas sur la modélisation de connaissances symboliques.

L'apprentissage automatique suit le même chemin : il cesse de se situer systématiquement en référence aux capacités des humains pour se concentrer sur les moyens d'exploiter au mieux les données stockées dans la mémoire des ordinateurs. Il rejoint aussi la démarche des statistiques, dont il s'était longtemps tenu éloigné. La théorie de l'apprentissage automatique progresse aussi à cette époque : le *no-free-lunch theorem* de (Wolpert, 1992), en montrant qu'aucun algorithme n'est meilleur que tous les autres sur l'ensemble de tous les problèmes possibles, sème, un temps, le trouble. Il formalise en quelque sorte l'intuition suivant laquelle sans biais, c'est-à-dire sans restriction sur l'espace des hypothèses possibles, l'induction est impossible. En ce sens, il ouvre aussi la porte à l'usage de stratégies d'apprentissage variées pour répondre à différents besoins.

Les années 1990-2000 voient ainsi l'émergence de multiples algorithmes qui se révèlent efficaces sur différents problèmes : arbres de décision, classification bayésienne, SVM, modèles graphiques, etc. Ces algorithmes dits « supervisés » nécessitent de disposer d'exemples étiquetés en quantité suffisante, mais reposent surtout sur des hypothèses numériques ou statistiques de mieux en mieux comprises (Quilan, 1993 ; Kearns et Vazirani, 1994 ; Vapnik, 1995 ; Mitchell, 1997 ; Vapnik, 1998). Le clustering et la découverte de règles d'association, qui relèvent de l'apprentissage non supervisé, connaissent aussi un grand développement.

Parallèlement, des corpus réels de grande dimension commencent à être disponibles : dans le sillage de la fouille de données, la fouille de textes devient un domaine en pleine expansion. La revue TAL s'est fait l'écho de cette évolution dès 1995, en consacrant un numéro double aux « *Traitements probabilistes et corpus* » (TAL, 1995). Celui-ci donne un panorama assez varié de travaux à base de corpus. Les questions de normes d'étiquetage y sont très prégnantes.

3. Les visages contemporains de l'apprentissage automatique appliqué au TAL

À l'heure actuelle l'apprentissage automatique, principalement représenté dans la communauté francophone par les conférences CAP (anciennement JFA) et EGC, est devenu une composante fondamentale de l'intelligence artificielle. Il a atteint un degré de maturité tel qu'il est impossible de l'ignorer dès qu'il s'agit de manipuler de grandes quantités de données de quelque nature que ce soit. C'est aussi vrai pour les textes, et le domaine du TAL s'en est trouvé bouleversé. Tous les niveaux d'analyse et tous les domaines applicatifs sont concernés. Mais la manière de

concevoir les liens entre apprentissage automatique et connaissances a aussi beaucoup évolué. C'est ce que nous explorons dans les sections qui suivent.

3.1. *État des lieux de l'apprentissage automatique*

L'apprentissage automatique est actuellement un domaine vaste et complexe qui ne se limite pas, comme on le croit trop souvent, aux traitements numériques ou statistiques. L'appel à communication de ce numéro voulait évoquer un plus vaste paysage, en citant plusieurs critères de classification possibles. Il y était ainsi question d'approches théoriques – liées à l'apprenabilité et la non-apprenabilité suivant des critères formels – ou empiriques – liées à l'utilisation d'algorithmes exploitant des données, annotées ou non, et s'appuyant sur un protocole expérimental. Il y était aussi évoqué que les méthodes d'apprentissage mises en œuvre pouvaient être symboliques (inférence grammaticale, PLI, etc.), à base de modèles probabilistes, statistiques ou numériques (modèles bayésiens, SVM, etc.), ou de similarités (voisinages, analogies, *memory-based learning*, etc.). Et encore, cet inventaire ne mentionnait ni l'apprentissage par renforcement ni les algorithmes génétiques, il est vrai plus rarement utilisés en TAL. (Cornuéjols et Miclet, 2002) donne un panorama beaucoup plus complet de l'apprentissage automatique dans son ensemble et illustre à sa façon la difficulté d'être exhaustif en la matière.

Il aurait été aussi possible de structurer cet appel d'une autre façon, en se focalisant plus sur la dimension applicative de l'apprentissage automatique et en s'appuyant sur les différentes tâches génériques auxquelles s'attaquent les algorithmes actuels les plus courants. Certaines de ces tâches, comme le *clustering*, la classification[1] (Sebastiani, 2002), l'annotation... sont étudiées depuis longtemps ; d'autres, comme l'ordonnancement de données, ont émergé plus récemment. Leur identification a permis une rationalisation du domaine : les progrès en apprentissage automatique sont maintenant systématiquement quantifiés, plusieurs algorithmes étant mis en concurrence pour résoudre la même tâche avec les mêmes données. Cette rationalisation a entraîné en retour un affinement croissant des programmes employés, devenus de plus en plus efficaces au fur et à mesure que leurs fondements mathématiques devenaient plus complexes. Les SVM (« *Suport Vector Machines* » ou « machines à vecteurs supports » en français) ont ainsi supplanté les réseaux de neurones pour les tâches de classification, de même que les CRF (« *Conditional Random Fields* », (Lafferty *et al.*, 2001) ou « champs markoviens conditionnels » en français) sont en train de prendre le relais des HMM pour celles d'annotation.

Pour le non-spécialiste qui souhaite mettre en œuvre des techniques d'apprentissage automatique, l'essentiel du travail consiste désormais souvent à

1. Le vocabulaire employé par les statisticiens et par les informaticiens diffère parfois : ici, nous utilisons la terminologie des informaticiens qui définissent la classification comme une catégorisation supervisée, alors que le *clustering* est non supervisé.

ramener le problème qu'il veut traiter à une de ces tâches génériques. C'est un travail de modélisation, qui peut aller d'une simple mise au format de ses données à une profonde reformulation de son problème. Il n'a par exemple pas été évident tout de suite que le problème de l'extraction et du typage des entités nommées dans un texte serait bien traité en le reformulant comme une tâche d'annotation de ce texte (Sarawagi, 2008). Quant au choix de l'algorithme lui-même, l'efficacité n'est pas toujours le seul critère à prendre en compte. D'autres paramètres peuvent justifier l'utilisation d'un programme d'apprentissage plutôt qu'un autre, comme le nombre et le type d'exemples qu'il requiert, sa capacité à intégrer des connaissances externes, ou encore l'interprétabilité de ses résultats.

Cette nouvelle structuration du domaine montre qu'un renversement profond a eu lieu. Les tâches d'apprentissage automatique sont devenues de plus en plus génériques, mais les algorithmes qui les traitent sont, de leur côté, de plus en plus capables de prendre en compte, dans leurs modèles, des connaissances externes. C'est un point fondamental sur lequel nous reviendrons plus loin. De fait, au lieu d'apparaître comme un acquis définitif ou comme un horizon, les connaissances relatives au domaine traité sont désormais intégrées dans la formulation du problème. Cette évolution est particulièrement sensible en TAL où la modélisation des connaissances a une longue histoire. Depuis plusieurs années déjà, une des préoccupations majeures des recherches en TAL est ainsi la combinaison entre connaissances linguistiques et apprentissage automatique. Les programmes de recherche actuels mettent presque systématiquement en avant des allers-retours féconds entre connaissances symboliques externes et connaissances acquises à partir de données, et tentent de faire collaborer les traitements manuels avec des traitements numériques ou statistiques. Cette hybridation nouvelle ne va pas sans heurts, mais elle peut aussi prendre plusieurs formes. C'est ce que nous allons voir dans les sections qui suivent.

3.2. *Apprentissage automatique et connaissances linguistiques : affrontements*

Il semble à première vue que les ressources obtenues par apprentissage automatique et celles construites « à la main » relèvent d'approches irréconciliables. Il existe, par exemple, divers étiqueteurs en « parties du discours » (*part of speech*) pour le français produits manuellement : ce sont en général des produits commerciaux payants. Les ressources libres (étiqueteur de Brill (Brill, 1992), TreeTagger (Schmid, 1994)[2]) ont, elles, été apprises automatiquement à partir de corpus. Dans le domaine de l'analyse syntaxique, les grammaires du français écrites à la main dominent encore (*cf.* les campagnes d'évaluation Easy[3] puis Passage[4]), mais des travaux sont en cours pour acquérir automatiquement une grammaire à partir du *FrenchTreebank* (Abeillé *et al.*, 2003), en s'inspirant de ce qui a déjà été

2. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
3. http ://www.technolangue.net/article198.html
4. http ://atoll.inria.fr/passage/eval2.fr.html

fait pour l'anglais avec le *Penn Treebank* (Hockenmaier et Steedman, 2002 ; Collins, 2003 ; Collins, 2004).

Quand de nouvelles problématiques applicatives émergent en TAL, comme la reconnaissance et le typage des entités nommées ou la classification des textes d'opinion, l'évolution des travaux suit souvent un cours comparable : d'abord, le problème est abordé en construisant et en exploitant des ressources spécialisées (dictionnaires, patrons écrits à la main). Puis, des corpus de référence commencent à être disponibles et les méthodes d'apprentissage automatique deviennent applicables. Elles permettent d'obtenir à moindres frais des programmes de bonne qualité qui, tôt ou tard, concurrencent les ressources patiemment construites à la main.

Les deux types de ressources ont chacun leurs intérêts et leurs limites. Les modèles symboliques produits manuellement sont en général interprétables par les humains. Mais leurs principaux défauts sont leur sensibilité aux erreurs et leur faible évolutivité. Quand on produit à la main une ressource, il est impossible de prévoir à l'avance tous les cas possibles. Les situations non anticipées (mots inconnus, constructions non répertoriées, etc.) mettent en échec le programme qui, au mieux, ne peut fournir qu'une réponse par défaut. Les modèles statistiques ou numériques appris automatiquement peuvent, au contraire, fournir une réponse en toutes circonstances qui sera la « moins mauvaise possible », en s'appuyant sur une combinaison de facteurs observables disponibles. Cela signifierait-il que le patient travail des linguistes est en passe de devenir inutile ?

3.3. *Apprentissage automatique et connaissances linguistiques : hybridations*

Heureusement pour les linguistes, la situation n'est pas si sombre. L'opposition classique et caricaturale entre approches symboliques et statistiques a déjà, en effet, largement perdu de sa pertinence. Tout d'abord, le domaine du symbolique s'est depuis longtemps ouvert à diverses hybridations. De nombreux modèles de représentation des connaissances intègrent, dans leur définition même, des règles symboliques et des valeurs numériques. Ainsi, les grammaires ou les automates probabilistes (Manning et Schütze, 1999), les réseaux bayésiens ou les modèles graphiques sont des objets typiquement hybrides. Même les méthodes d'apprentissage automatique historiquement fondées sur des modèles symboliques ont pris un tournant pragmatique et sont devenues capables de se « frotter » aux données réelles : la PLI a évolué vers l'apprentissage relationnel, voire même statistique (Getoor et Taskar, 2006), et l'inférence grammaticale s'est beaucoup diversifiée (de la Higuera, 2010). Plusieurs campagnes de compétitions consacrées à l'identification de grammaires à partir d'exemples (Abadingo, Gowachin, Omphalos[5], ou les actuels Stamina[6] et Zulu[7]) font progresser les algorithmes dans le

5. Voir le site consacré à l'inférence grammaticale (malheureusement peu à jour) : http://labh-curien.univ-st-etienne.fr/informatique/gi/

sens d'un passage à l'échelle et d'une moindre sensibilité aux données erronées. Même si encore peu de travaux combinent apprentissage symbolique et apprentissage numérique ou statistique, on peut parier que l'acquisition d'analyseurs syntaxiques à partir de corpus arborés va de plus en plus faire appel à des techniques venues à la fois de l'inférence grammaticale et de l'apprentissage statistique.

Ensuite, comme nous l'avons déjà suggéré dans la section 3.1, les ressources construites à la main peuvent souvent être réinvesties par les algorithmes d'apprentissage eux-mêmes pour enrichir les exemples et améliorer la qualité de ce qui est appris. La plupart des meilleurs algorithmes d'apprentissage actuels sont capables d'intégrer de telles connaissances. Ainsi, les SVM requièrent la définition d'un « noyau » qui caractérise les distances entre données. Il est possible, dans la définition de ce noyau, de prendre en compte une multitude d'informations, qui codent des connaissances externes. De même, les CRF sont fondés sur l'affectation de « poids » à des « caractéristiques » (*features*). Les caractéristiques sont des fonctions booléennes fournies au système. Toutes les connaissances linguistiques supposées utiles à la résolution de la tâche peuvent être intégrées au CRF par le biais de ces caractéristiques. Dans les deux cas, les connaissances symboliques externes peuvent donc être vues comme des « atomes de connaissances » que les algorithmes se chargent de combiner entre eux *via* des paramètres numériques fixés par apprentissage. Le programme ainsi défini est donc à la fois capable de tenir compte de toutes les informations symboliques qu'on lui a fournies, mais aussi de paramétrer leur importance relative et de donner une réponse adaptée en toutes circonstances. C'est sans doute là le mode de combinaison entre approches symboliques et numériques le plus élégant et efficace actuellement disponible. Au lieu d'espérer atteindre des connaissances symboliques par passage à la limite de modèles numériques, il amène à considérer qu'elles sont en fait *premières mais locales* et ont besoin d'être intégrées à plus grande échelle par le biais de valeurs numériques. Plusieurs articles de ce numéro en sont une parfaite illustration.

Notons enfin que les compétences des linguistes sont également précieuses pour comprendre et interpréter les résultats d'un programme d'apprentissage automatique. Le reproche qui a été longtemps fait aux modèles statistiques de ne pas être « interprétables » a ainsi de moins en moins lieu d'être. Là encore, plusieurs des articles qui suivent le montrent de façon convaincante.

3.4. *Les articles de ce numéro*

Il est donc temps d'en venir brièvement au contenu de ce numéro. Les articles qui y figurent sont très représentatifs de l'évolution que nous venons de tracer à grands traits. Ils traitent de niveaux d'analyse variés, allant de la syntaxe à la sémantique en passant par la classification de phases de dialogues ou la traduction

6. http ://stamina.chefbe.net/
7. http ://labh-curien.univ-st-etienne.fr/zulu/

automatique. Mais, plutôt que de les agencer en fonction de leur domaine applicatif, nous avons choisi de les présenter en fonction de la manière dont ils articulent apprentissage automatique et traitement des langues.

Les deux premiers articles seraient ainsi à ranger dans la lignée historique qui va des données aux connaissances, en mettant en œuvre des stratégies d'apprentissage non supervisées, mais paramétrées, contrôlées et évaluées par une expertise humaine. Le premier d'entre eux, dû à Salma Jamoussi, montre qu'en appliquant certains algorithmes de *clustering* à des textes, il est possible, dans le même temps, d'extraire des listes de mots représentatifs de leur contenu sémantique. Les diverses techniques testées sont paramétrées par différents choix possibles de distances et de représentations des textes, qui peuvent s'interpréter comme différentes hypothèses distributionnelles. Dans celui signé par Thierry Charnois, Marc Plantevit, Christophe Rigotti et Bruno Crémilleux, des méthodes d'identification de motifs séquentiels fréquents sont employées pour constituer des patrons d'extraction d'entités nommées et de relations qui les relient. Cette dernière tâche est typiquement de celles qu'il est encore difficile d'aborder par apprentissage automatique, parce qu'elle se ramène difficilement à une tâche générique plus simple (comme la classification ou l'annotation). Les patrons obtenus sont interprétables et ont été soumis avec succès à des experts humains.

Les quatre articles suivants, quant à eux, illustrent parfaitement l'*intégration de connaissances linguistiques dans un mécanisme d'apprentissage supervisé numérique ou statistique sophistiqué*. Les deux premiers, en anglais, mettent en effet en œuvre des SVM, outils actuellement les plus performants pour les tâches de *classification*. Celui de Pierre Andrews et Suresh Manandhar traite justement d'un problème de classification, relativement original, qui consiste à évaluer l'accord entre interlocuteurs dans un dialogue. L'article se concentre sur les caractéristiques (« *features* ») linguistiques à intégrer au calcul du noyau du SVM pour atteindre les meilleurs taux de reconnaissance. L'article suivant, de Lilja Øvrelid, Jonas Kuhn et Kathrin Spreyer, utilise exactement la même méthodologie, mais pour une toute autre tâche : acquérir un analyseur syntaxique efficace dans différentes langues. Pour cela, le problème de l'analyse syntaxique est tout d'abord ramené à une série de classifications élémentaires. Pour apprendre à réaliser ces classifications, les données issues de grammaires symboliques existantes sont transformées, là encore, en « caractéristiques » prises en compte dans le noyau du SVM. C'est une excellente illustration de la méthodologie évoquée en section précédente.

Les deux derniers articles, enfin, portent sur les CRF séquentiels, le meilleur modèle graphique actuel pour apprendre à annoter des textes. Les CRF aussi requièrent des « caractéristiques » et la phase d'apprentissage consiste à trouver les poids relatifs de chacune d'entre elles pour étiqueter correctement une séquence. L'article de Nataliya Sokolovska, Olivier Cappé et François Yvon décrit une stratégie qui permet de sélectionner les caractéristiques les plus utiles lors de la phase d'apprentissage, garantissant ainsi d'énormes gains en temps de calculs. Mais elle montre aussi que les résultats d'un tel modèle restent interprétables

linguistiquement. Les expériences qui valident ce nouvel algorithme d'inférence portent sur le *chunking* et la reconnaissance des entités nommées, deux tâches très bien traitées par annotation. Enfin le dernier article d'Alexandre Allauzen et Guillaume Wisniewski décrit aussi l'utilisation de CRF, cette fois pour réaliser des alignements mots à mots multilingues pour la traduction automatique. Encore une fois, une tâche complexe est ramenée à une autre plus générique, mais capable d'intégrer dans ses caractéristiques des informations fournies par une autre ressource (dans ce cas, des catégories morphosyntaxiques).

4. Conclusion

Ce court panorama, malgré d'inévitables simplifications, met en avant certaines lignes de force indiscutables. En quelques décennies, l'apprentissage automatique, comme le TAL, a connu des fluctuations majeures. Il est passé de la résolution de problèmes spécifiques aboutissant à des connaissances spécialisées à l'étude de tâches génériques capables d'intégrer dans ses modèles des connaissances du domaine. Cette évolution est un signe de maturité qui lui permet d'être applicable à des domaines très variés. On peut aussi voir cet aboutissement (provisoire) comme un retour à l'intuition initiale, suivant laquelle l'apprentissage est une capacité générale et universelle. C'est vrai pour les humains, c'est en train de le devenir pour les machines.

Le TAL, qui n'a jamais non plus renoncé à l'idéal d'universalité des sciences du langage, a tout à gagner à cette nouvelle maturité. Mieux, l'apprentissage automatique pourrait lui permettre en quelque sorte de se réconcilier avec lui-même : la rupture historique, en son sein, entre « grammaires formelles » et « théorie de l'information », qui remonte aux controverses entre Chomsky et Harris, a de moins en moins lieu d'être, quand on regarde de près les travaux actuels qui combinent les deux approches (Pereira, 2000). Il n'y a plus vraiment de contradictions à construire manuellement des ressources ou des modèles formels et à les exploiter dans un programme d'apprentissage automatique à partir de données.

Les acquis de l'apprentissage automatique doivent désormais faire partie du bagage de base de tout bon praticien du TAL.

5. Bibliographie

Abeillé A., Clément L., Toussenel F., *Treebanks : Building and Using Parsed Corpora*, Kluwer, chapter Building a Treebank for French, p. 165-188, 2003.

Angluin D., « Inductive Inference of Formal Languages from Positive Data », *Information and Control*, vol. 45, n° 2, p. 117-135, May, 1980.

Angluin D., « Inference of Reversible Languages », *Journal of the ACM*, vol. 29, n° 3, p. 741765, July, 1982. Benzecri F., *Histoire et préhistoire de l'analyse des données*, Dunod, 1982.

Brill E., « A simple rule-based part of speech tagger », *Proceedings of the third conference on Applied natural language processing*, p. 152-155, 1992.

Chomsky N., *Rules and representations*, Basil Blackwell, 1980.

Collins M., « Head-Driven Statistical Models for Natural Language Parsing », *Computational Linguistics*, vol. 29, n° 4, p. 589-637, 2003.

Collins M., *New Developments in Parsing Technology*, Kluwer, chapter Parameter Estimation for Statistical Parsing Models : Theory and Practice of Distribution-Free Methods, 2004.

Cornuéjols A., Miclet L., *Apprentissage artificiel ; concepts et algorithmes*, Eyrolles, 2002.

Crevier D., *A la recherche de l'intelligence artificielle*, champs, Flammarion, 1999.

Gardner H., *Histoire de la révolution cognitive*, Payot, 1993.

Getoor L., Taskar B., *Introduction to Statistical Relational Learning*, MIT Press, 2006.

Gold E., « Language Identification in the Limit », *Information and Control*, vol. 10, p. 447-474, 1967.

Higuera (de la) C., *Grammatical Inference, Learning Automata and Grammars*, Cambridge University Press, 2010.

Hockenmaier J., Steedman M., « Generative Models for Statistical Parsing with Combinatory Categorial Grammars », *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 335-342, 2002.

Kearns M. J., Vazirani U. V., *An Introduction to Computational Learning Theory*, MIT Press, 1994.

Lafferty J. D., McCallum A., Pereira F. C. N., « Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. », *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, p. 282-289, 2001.

Manning C., Schütze H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, 1999.

McCulloch W., Pitts W., « A Logical Calculus of the Ideas Immanent in Nervous Activity », *Bulletin of Mathematical Biophysics*, 1943.

Minsky M., Papert S., *Perceptrons*, MIT Press, 1969.

Mitchell T., *Machine Learning*, McGraw-Hill, 1997.

Pereira F., « Formal grammar and information theory : Together again ? », *Philosophical Transactions of the Royal Society*, vol. 358, p. 1239-1253, 2000.

Piatelli-Palmarini M., *Théories du langage, théories de l'apprentissage, le débat entre Jean Piaget et Noam Chomski*, Le Seuil, 1979.

Pullum G., « Empirical assessment of stimulus poverty arguments », *The Linguistic Review*, vol. 19, p. 9-50, 2002.

Quilan J. R., *C4.5 : Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

Sarawagi S., « Information Extraction », *Foundations and Trends in Databases*, vol. 1, n° 3, p. 261-377, 2008.

Schmid H., « Probabilistic part-of-speech tagging using decision trees », *Proceedings of International Conference on New Methods in Language Processing*, p. 44-49, 1994.

Sebastiani F., « Machine Learning in Automated Text Categorization », *ACM Computing Surveys*, vol. 34, n° 1, p. 1-47, 2002.

Smolensky P., *IA connexioniste, IA symbolique et cerveau*, Folio essai, Gallimard, p. 77-107, 1992.

Solomonoff R. J., « A Formal Theory of Inductive Inference », *Information and Control*, vol. 7, p. 1-22,224-254, 1964. TAL, *Traitements probabilistes et corpus*, vol. 36, revue de l'ATALA, num 1-2, 1995.

Turing A., « Computing Machinery and Intelligence », *Mind*, vol. 49, p. 433-460, 1950.

Valiant L. G., « A Theory of the Learnable », *Communications of the ACM*, vol. 27, n° 11, p. 1134-1142, 1984.

Vapnik V. N., *The nature of statistical learning theory*, Springer Verlag, 1995.

Vapnik V. N., *Statistical Learning Theory*., John Wiley, 1998.

Wolpert D., « No free lunch theorem for optimization », *IEEE Transactions on Evolutionary Computation*, vol. 1, n° 1, p. 467-482, 1992.

Remerciements

Ce numéro a été coordonné avec Mark Steedman, de l'université d'Edimbourg.

Nous remercions chaleureusement tous les relecteurs qui y ont contribué : Pieter Adriaans (HSC Lab, Université d'Amsterdam, Pays-Bas), Massih Amini (LIP6, Paris et ITI-CNRC, Canada), Walter Daelemans (CNTS, Université d'Anvers, Belgique), Pierre Dupont (Université Catholique de Louvain, Belgique), Alexander Clark (Royal Holloway, Université de Londres, Grande-Bretagne), Hervé Dejean (Xerox Center, Grenoble), George Foster (ITI-CNRC, Canada), Colin de la Higuera (Laboratoire Hubert Curien, Université de St Etienne), François Denis (LIF, Université de Marseille), Patrick Gallinari (LIP6, Université de Paris 6), Cyril Goutte (ITI-CNRC, Canada), Laurent Miclet (Enssat, Lannion), Richard Moot (CNRS, Bordeaux), Emmanuel Morin (LINA, Université de Nantes), Jose Oncina (PRAI Group, Université d'Alicante, Espagne), Pascale Sébillot (IRISA, INSA Rennes), Marc Tommasi (LIFL-Inria, Université de Lille), Menno van Zaanen (ILK, University of Tilburg, Pays-Bas).

Isabelle Tellier
LIFO
Université d'Orléans 6, rue Léonard-de-Vinci BP 6759 45 067 Orléans Cedex
France
isabelle.tellier@univ-orleans.fr
http://www.univ-orleans.fr/lifo/Members/Isabelle.Tellier/