Notes de lecture

Rubrique préparée par Denis Maurel

Université François Rabelais Tours, LI (Laboratoire d'informatique)

Jean-Gabriel GANASCIA, Communication et connaissance. Supports et médiations à l'âge de l'information, *CNRS éditions*, 2006, 178 pages, ISBN 2-271-06415-5.

Lu par **Jean-Baptiste BERTHELIN**

LIMSI-CNRS

Cet ouvrage collectif comporte un bilan et une étude prospective de l'impact des technologies de l'information et de la communication dans plusieurs domaines. Il aborde en particulier des questions sur le matériel, les algorithmes, l'ergonomie des interfaces, la gestion des corpus, ainsi que les enjeux juridiques, éthiques et administratifs de cette évolution des techniques et des usages.

Réunis autour de Jean-Gabriel Ganascia, une quarantaine de spécialistes de tous les aspects du traitement de l'information nous font profiter de leur vision des changements induits par les rapides évolutions des techniques et des usages dans ce secteur. Leurs domaines de compétence vont du plus concret (les composants matériels des nouveaux équipements) au plus abstrait (les aspects politiques et juridiques des nouveaux comportements d'utilisateurs). Leurs contributions sont écrites avec un grand souci de lisibilité et donnent par conséquent une image claire des acquis et des défis qui caractérisent chaque sous-domaine.

De la sorte, le lecteur apprendra pourquoi dix nanomètres est une taille critique pour des composants, comment se servir du spin des électrons et des phénomènes d'intrication quantique ou comment concilier recherche de preuve de programme et gestion de l'approximation dans les données. En matière de génie logiciel, une intéressante discussion porte sur le statut de la programmation par objets et la possibilité de lui associer des théories scientifiques pour lui donner du sens.

D'autres contributions examinent les effets de la croissance prodigieuse des réseaux et de l'accès d'un grand nombre d'utilisateurs à une immense quantité de données, qui ne sont ni stables, ni homogènes, ni clairement hiérarchisées. En particulier, les linguistes disposent désormais de grands corpus de textes, pour lesquels ces questions se posent avec acuité. La diversité des comportements d'utilisateurs interdit en effet toute normalisation excessive. Cependant, il est possible et souhaitable d'appliquer à ces corpus une norme descriptive minimale, afin de tenir compte des éclairages dont la linguistique peut les faire bénéficier. Une

telle démarche aurait aussi l'avantage de contribuer à la préservation de la langue française, enjeu majeur de notre temps.

Plus fondamentalement, une réflexion est engagée, et doit se poursuivre, sur les mutations subies par le concept de document. Pendant plusieurs siècles, ce terme a désigné une surface de papier arrangée plus ou moins en forme de cahier. Désormais, c'est un morceau de mémoire informatique dont l'archivage, la consultation et la mise à jour mobilisent des instruments autres que ceux des scribes du passé. Par exemple, les liens entre documents sont souvent matérialisés de telle sorte que le lecteur les suivra sans effort et se lancera dans l'exploration, forcément partielle, d'une véritable galaxie de textes. Un tel parcours ne peut que gagner à être intelligemment éclairé et balisé, d'où le besoin de nouveaux outils de marquage et de commentaire.

Les dernières pages de ce recueil sont consacrées à des points de vue sociologiques, historiques et anthropologiques sur les nouvelles approches de l'information et des documents. Elles posent, entre autres, la question d'adapter les outils logiciels aux différentes modalités du travail d'érudition. Certains chercheurs peuvent vouloir traiter, successivement, des résultats relativement simples, alors que d'autres procéderont par approfondissements et affinages progressifs d'une même notion subtile.

Par-delà ces questionnements individuels, l'incidence des nouvelles technologies sur les activités collectives est considérable. Qu'il s'agisse des mécanismes de prise de décision à plusieurs ou de l'encadrement des nouveaux usages par les anciennes normes juridiques (appelées, de ce fait, à évoluer), la réflexion des spécialistes est sollicitée. C'est particulièrement vrai lorsqu'il s'agit de fonder une éthique de la nouvelle communication, autrement dit, d'élaborer un système de valeurs et un code de comportement dans cet environnement en mutation. Cependant, les fondements traditionnels d'un système de valeurs, qu'ils soient religieux, culturels ou méthodologiques, n'ont pas forcément cours auprès des participants aux nouvelles formes d'interaction. Il convient donc de réfléchir à ce qui pourrait être la source d'un consensus dans ce domaine.

Au total, cet ouvrage est à la fois une somme et un manifeste. Une somme, car il s'efforce d'apporter exhaustivement des réponses à une foule de questions sur les récentes et futures mutations des systèmes de traitement de l'information. Un manifeste, car chacun des auteurs adopte une attitude résolument progressiste, montrant que la communauté des chercheurs, si on lui procure des moyens convenables, peut inlassablement faire évoluer les environnements complexes de notre univers informationnel. Sa lecture est donc vivement recommandée aux étudiants, aux chercheurs et à quiconque éprouve, à l'égard des questions qu'il traite, une légitime curiosité.

Walter DAELEMANS, Antal VAN DEN BOSCH, Memory-Based Language Processing, *Cambridge University Press*, 2005, 198 pages, ISBN 0521808901.

Lu par **François YVON**

Ecole Nationale Supérieure des Télécommunications (GET/ENST et CNRS/LTCI)

Depuis plus de dix ans, Walter Daelemans et Antal van den Bosch s'intéressent à l'utilisation de méthodes d'apprentissage automatique en traitement automatique des langues (TAL) et plus spécifiquement à l'utilisation de méthodes à base de cas[1], c'est-à-dire pour lesquelles le processus d'apprentissage se ramène à une simple mémorisation des exemples proposés et ne construit aucune abstraction des données disponibles. Ils proposent ici une synthèse illustrée par de nombreux exemples de l'ensemble de leurs travaux. Cette synthèse s'accompagne d'une présentation des algorithmes et logiciels libres[2] développés dans le cadre de ces recherches, ce qui permettra au lecteur de répéter les expériences décrites dans le livre. Cette synthèse est enfin l'occasion de discuter les avantages et difficultés d'une approche à base de cas, à la fois d'un point de vue linguistique, algorithmique et expérimental.

Memory-Based Language Processing est organisé en sept chapitres. Le premier chapitre est une brève introduction à l'ouvrage, qui s'attache à défendre l'hypothèse que de nombreuses tâches du TAL peuvent être reformulées comme des problèmes de catégorisation supervisée, impliquant des représentations « plates » (structures attributs-valeurs) des entités linguistiques. Cette hypothèse ne va pas de soi : les données manipulées en TAL sont souvent structurées et présentent des dépendances complexes qui les prédisposent mal à l'utilisation de méthodes de catégorisation.

Le second chapitre vise à situer les méthodes à base de cas dans le cadre plus général de la linguistique à base de corpus, d'une part, en soulignant, de manière un peu convenue, leur proximité avec les méthodes issues de la linguistique structurale (Halliday, Firth, Harris, etc) ; et, d'autre part, dans celui des études conduites en psycho-linguistique. La seconde filiation revendiquée des auteurs est une tradition issue de la reconnaissance des formes et de l'intelligence artificielle, qui a conduit au développement de l'algorithme des *k*-plus-proches voisins et des méthodes de raisonnement à base de cas. Les développements plus formels auxquels ces méthodes ont donné lieu dans le domaine des statistiques sont, en revanche, passés sous silence, manifestant le parti-pris de limiter les développements mathématiques et de rendre ainsi le texte accessible au plus grand nombre.

[1] De préférence à *à base d'exemples* ; on parle également d'*apprentissage paresseux*.

[2] TIMBL et MBT, voir `http://ilk.uvt.nl/mblp`.

Le troisième chapitre détaille pas-à-pas un exemple de mise en œuvre de cette méthodologie sur la tâche de génération du pluriel des noms en allemand. Cette tâche est rendue difficile par la diversité des procédés morphologiques en compétition, parmi lesquels le procédé régulier (suffixation de -s) est loin d'être le plus fréquent. Daelemans et van den Bosch montrent comment se ramener à un problème de catégorisation consistant à associer, à une représentation phonologique de la base, une étiquette de classe correspondant au procédé de formation du pluriel. La catégorisation d'une nouvelle forme s'effectue en la comparant aux exemples connus et en lui assignant la catégorie de l'exemple *le plus proche*. Se trouve ainsi établie l'importance de la *métrique* utilisée pour comparer les exemples. Les auteurs consacrent une large place à présenter les différentes métriques disponibles dans Timbl et à les comparer empiriquement. Cette discussion donne lieu à une présentation très pédagogique des techniques d'optimisation des paramètres, des protocoles d'évaluation, ainsi que les méthodes de comparaison des performances. Le chapitre se conclut par une discussion des problèmes algorithmiques posés par la méthode (qui implique un calcul de distance avec tous les exemples disponibles) et une présentation des solutions retenues dans Timbl.

Le chapitre 4 est consacré à des applications plus ambitieuses : la transcription orthographique-phonétique et l'analyse morphologique. Ces applications sont plus difficiles, car le caractère séquentiel des entrées et des sorties joue un rôle essentiel, ce qui rend moins aisé l'utilisation d'un cadre de catégorisation. Pour la première application, Daelemans et van den Bosch utilisent une reformulation classique: un mot est transcrit phonème par phonème de la gauche vers la droite ; chaque phonème est l'étiquette de classe assignée à une fenêtre orthographique de taille fixe. Cette approche demande que les entrées orthographiques et les sorties phonémiques soient préalablement alignées[3]. La tâche d'analyse morphologique pose des problèmes plus ardus encore, qui sont résolus comme suit: chaque fenêtre orthographique centrée sur la $i^{ème}$ lettre est catégorisée en deux classes principales selon qu'il s'agit ou non d'une frontière de morphème ; dans l'affirmative, l'étiquette encode également le type du morphème. Si cette représentation du problème permet d'atteindre des performances remarquables, on notera qu'elle repose sur l'hypothèse linguistiquement contestable que les mots se décomposent intégralement en morphèmes et qu'elle ne se généralise pas à toutes les langues.

---

[3] Signalons une petite approximation méthodologique dans la présentation de l'algorithme d'alignement décrit p. 62 : l'algorithme décrit n'est pas exactement une instance de l'algorithme *Expectation Maximization* (EM), mais d'une version heuristique dans laquelle l'étape E est rendue déterministe. Il est également dommage que le modèle probabiliste sous-jacent (une forme dégénérée de modèle de Markov à états cachés (HMM)) ne soit pas explicitement mentionné. Ceci aurait donné à l'occasion aux auteurs de justifier l'étrange détour consistant à utiliser un HMM en prélude à l'utilisation de méthodes à base de cas.

Daelemans et van den Bosch considèrent au chapitre 5 des tâches d'annotation syntaxique. Si l'analyse syntaxique classique (la construction d'une représentation de la structure interne d'un énoncé, sous la forme d'un arbre ou un graphe de dépendance) n'est pas directement un problème de catégorisation, il est possible de s'en approcher en enchaînant (par *pipe-line*) trois tâches plus simples, qui sont réalisables par des méthodes à base de cas : étiquetage morpho-syntaxique, analyse de surface (*chunking*) et identification de (certaines) relations de dépendances. Pour chaque tâche, les auteurs présentent une nouvelle bordée de résultats expérimentaux, qui montrent que cette approche est aussi performante que les modèles probabilistes tels que les modèles de Markov cachés ou les champs aléatoires conditionnels. Ce chapitre introduit, en parallèle, une évolution de Timbl, nommée Mbt, spécialement optimisée pour le traitement de données syntaxiques.

Le sixième chapitre analyse plus finement les mérites des approches à base de cas en TAL. En s'appuyant sur une multitude de résultats expérimentaux obtenus sur six tâches différentes, couvrant un large spectre d'applications[4], les auteurs s'attachent à évaluer le bénéfice que représente la mémorisation de *l'ensemble des exemples disponibles*. Ce trait est caractéristiques des méthodes paresseuses et les distingue des méthodes *à base de modèles*, qui, construisant une abstraction des données, s'avèrent souvent en peine de capturer les multiples micro-régularités présentes dans les données linguistiques. En comparant les résultats obtenus par l'approche paresseuse à ceux obtenus par un apprenti à base de règles; en mesurant les pertes de performance enregistrées lorsqu'on *édite* la base d'exemples (en supprimant les instances les plus exceptionnelles, ou, au contraire, les plus prédictibles), Daelemans et van den Bosch font une démonstration très convaincante du fait qu'en TAL, toutes les données disponibles contribuent aux performances en généralisation. Cette discussion est assortie d'une présentation des principales méthodes d'édition des bases de données.

Le septième chapitre revient sur la question de la prédiction de séquences de catégories (transcription orthographique-phonétique, étiquetage morpho-syntaxique…). Dans ce cas, l'approche de Timbl consiste à effectuer une série de catégorisations à chaque position de la séquence d'entrée, chaque décision étant indépendante de ses voisines. Cette approche ne permet de modéliser qu'indirectement les dépendances entre les catégories qui constituent la séquence de sortie, conduisant par exemple à produire des suites d'étiquettes très improbables, voire mal-formées au regard des contraintes de la tâche. Pour pallier ce problème, Daelemans et van den Bosch introduisent deux « astuces » dont la portée est assez générale. La première, dénommée *stacking*, consiste à enchaîner (à mettre en série) des classifieurs : le premier prédit la séquence de sortie position par position, sur la

[4] Dont trois nouvelles : la génération de diminutifs, la prédiction des rattachements prépositionnels et l'identification d'entités nommées viennent compléter leur jeu de test.

base de fenêtres de contextes décrivant l'entrée. Le second fonctionne à l'identique, à ceci près qu'il dispose également des étiquettes prédites par le premier classifieur pour affiner ses prédictions : si celles-ci sont correctes, alors l'étiquette à la position *j* dépend directement des étiquettes des positions voisines. Ce procédé peut être généralisé à un nombre arbitraire de classifieurs. La seconde astuce consiste à prédire des *catégories composites*, par exemple des trigrammes de catégories : pour une tâche de phonétisation, on assigne à chaque fenêtre orthographique centrée sur la position *i* une étiquette correspondant aux phonèmes *i*-1, *i* et *i*+1. Procéder de cette manière conduit à prédire chaque étiquette trois fois (pour trois positions différentes), un mécanisme de vote permettant d'attribuer la catégorie finale. Le bénéfice, très net, qu'apporte ces deux raffinements (qui peuvent être utilisés ensemble) est évalué empiriquement sur diverses tâches.

Le livre se termine par une bibliographie très substantielle, bien que marquée par une forte sur-représentation des travaux des auteurs, qui rend pas complètement compte de la diversité des travaux conduits dans le domaine ; puis sur un index.

Mon principal regret est que les auteurs n'aient pas davantage cherché à élargir la discussion à d'autres méthodes statistiques communément utilisées en TAL. En ce sens *Memory-Based Language Processing* prend l'exact contre-pied de *Foundations of Statistical Natural Language Processing* (Manning et Schütze, 1999), qui est très discret sur les méthodes à base de cas et en est donc très complémentaire. Je reste également un peu frustré que la discussion se cantonne le plus souvent à discuter des aspects pratiques et des performances que permet d'obtenir la méthode. Utiliser des techniques de catégorisation impose parfois d'effectuer des reformulations contre-intuitives et/ou complexes, voire de se plier à des hypothèses linguistiques, qui restreignent le champ d'application de ces techniques. C'est un des sujets sur lequel on aurait pu s'attarder, pour cerner plus clairement les limites de l'approche et pour évoquer les manières de la rendre mieux à même de traiter des données fortement structurées.

On ne peut toutefois que sortir de cette lecture impressionné par la quantité et la qualité du travail expérimental effectué par les auteurs et leur co-auteurs, qui leur a permis d'accumuler une expertise sur une vaste gamme de tâches du TAL. S'appuyant sur cette expertise unique, Daelemans et van den Bosch développent un argumentaire très convaincant sur les avantages des méthodes à base de cas. L'autre point fort de ce livre est qu'il constitue un guide très pédagogique, illustrant avec force détails la mise en œuvre d'une démarche d'apprentissage dans un cadre de TAL. À ce titre, cet ouvrage sera le plus utile à des lecteurs peu familiers avec le domaine de l'apprentissage automatique et désireux de tester pratiquement l'efficacité de méthodes de l'état de l'art, en particulier celles développées par les auteurs (Timbl et Mbt). Il me semble également que la disponibilité de ces outils et leur facilité d'utilisation les désigne comme des systèmes de référence avec lequel comparer les performances de techniques d'apprentissage dans un cadre de TAL.