Notes de lecture

Rubrique préparée par Denis Maurel

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

Afra ALISHAHI, Computational Modeling of Human Language Acquisition, *Morgan & Claypool publishers*, 2011, 93 pages, ISBN 978-1-6084-5339-9.

Lu par **Pascal AMSILI**

Université Paris-Diderot, LLF

L'acquisition des langues intéresse les chercheurs depuis des siècles et cette question a été explorée avec toutes sortes de méthodes, mais c'est seulement très récemment que la modélisation informatique est venue enrichir l'inventaire des méthodes disponibles. C'est un mouvement assez naturel, si l'on s'avise que la recherche en TAL fait appel de façon de plus en plus massive à des méthodes d'apprentissage automatique. Il faut aussi ajouter que des données expérimentales et des corpus pertinents ont été rendus accessibles récemment. Cette situation rend la parution du livre d'Afra Alishahi très utile, aussi bien pour les chercheurs venus du TAL qui s'intéressent à l'application de leurs méthodes à la question de l'acquisition des langues, qu'aux linguistes et psycholinguistes qui veulent savoir ce que l'informatique peut apporter à leur domaine de recherche.

Ce livre est le onzième d'une série récemment publiée chez Morgan & Claypool, sous la direction de Graeme Hirst, titrée « *Synthesis Lectures on Human Language Technologies* ». Le concept est de fournir des textes relativement courts qui font un survol récent et informé de différentes questions relevant des technologies linguistiques, au sens le plus large possible. Il réalise parfaitement les objectifs de la série et propose un survol synthétique, et relativement complet, de la question.

L'ouvrage est organisé en six chapitres qui peuvent être regroupés en trois parties distinctes. La première partie – premier et deuxième chapitres – présente la problématique et les principales méthodologies du domaine. Dans le premier chapitre, on montre quelques-uns des grands débats qui ont animé le domaine de l'acquisition des langues, avec la question de la modularité du langage et celle de l'opposition entre la vue innéiste du langage, célèbre pour l'argument de la pauvreté du stimulus proposé par Chomsky en 1965, et la vue « *usage-based* », qui explore sans *a priori* les possibilités d'induire une compétence linguistique à partir des données typiquement accessibles aux enfants pendant leur apprentissage. Quelques pointeurs sur les résultats mathématiques pertinents pour ce débat sont redonnés plus loin dans l'ouvrage (p. 45-46). Cette présentation rapide du contexte ne sera sans doute pas très utile aux chercheurs ayant déjà une formation en linguistique, mais

elle a le mérite de replacer ces recherches dans leur contexte, qui reste un contexte très débattu aujourd'hui. Le chapitre 1 présente aussi des arguments en faveur de l'ajout de méthodes de modélisation informatique à l'inventaire des méthodes de la psycholinguistique. Le chapitre 2 présente la problématique de la modélisation computationnelle et discute en particulier deux aspects importants qui sont d'une part la question de la plausibilité cognitive des modèles que l'on produit (question rarement pertinente en TAL, ce qui rend beaucoup de modèles venus du TAL inappropriés pour les sciences cognitives), et, d'autre part, la question de l'évaluation, sachant que là encore les habitudes venues du TAL doivent être adaptées, puisque l'on ne dispose en général pas de « *gold standard* » pour les tâches modélisées.

La deuxième partie regroupe les trois chapitres suivants, qui sont chacun consacrés à un domaine particulier : l'apprentissage du vocabulaire (chapitre 3), l'apprentissage de la morphologie et de la syntaxe (chapitre 4) et, enfin, l'apprentissage de la relation syntaxe-sémantique, tout du moins en ce qui concerne les structures argumentales et les rôles thématiques (chapitre 5). Enfin, la troisième partie, représentée par le chapitre 6, conclut l'ouvrage avec quelques remarques conclusives.

Chacun des chapitres de cette deuxième partie est organisé de manière similaire, avec une division en sous-domaines (par exemple la morphologie, les parties du discours et la grammaire pour le chapitre 4), et chaque fois une présentation de la problématique, avec une revue assez exhaustive des modèles récents qui ont pu être proposés et une « étude de cas », censée présenter de façon plus approfondie un modèle particulier.

La structure de l'ouvrage, et son contenu, en font un excellent outil pour tout chercheur qui s'intéresse à la modélisation de l'acquisition des langues, qu'il vienne du champ du TAL ou de la psychologie cognitive, et qui va trouver ici un texte bref, dans lequel se trouvent listées la plupart des références pertinentes dans le domaine. De plus, l'autonomie des différentes parties contribue à faire de ce livre un texte de référence facile à consulter et qui pourra même être utile aux chercheurs déjà engagés dans le champ du TAL ou autres, qui trouveront un inventaire de références utiles.

Cependant, le projet de l'ouvrage conduit à une limite qui peut engendrer une certaine frustration : le niveau de détail des discussions et des explications est souvent insuffisant en soi et rend nécessaire la consultation des sources directes, mais ce serait malvenu d'en faire le reproche à l'auteure compte tenu du projet et des contraintes de la collection.

On peut cependant exprimer quelques regrets, qui n'enlèvent rien à l'intérêt de l'ouvrage. Tout d'abord, les études de cas ne sont pas la partie la plus réussie de l'ouvrage. On comprend le but, qui était de donner un peu de chair à un texte qui est souvent assez abstrait et comprend surtout des inventaires. Mais le niveau de détail des études de cas est souvent insuffisant pour qu'elles puissent vraiment être

exploitées (on suppose dans ces études une compétence accrue du lecteur). Par ailleurs, leur choix semble assez arbitraire, car il ne s'agit pas toujours d'un modèle particulièrement influent, ou représentatif. Enfin, il semble que l'importance donnée aux approches neuromimétiques et, en corrélation, la faible représentation des approches bayésiennes, ne reflètent pas la réalité du domaine aujourd'hui.

Pour des raisons évidentes de place, l'auteure est conduite à une certaine simplification des positions quand il s'agit de rendre compte des quelques grands débats qui traversent (quelquefois très violemment) la communauté. Ainsi l'opposition entre ce qu'on appelle le « *semantic bootstrapping* » et les approches dites « *usage-based* » est résumée en quelques lignes d'une façon forcément un peu réductrice. Mais il faut créditer l'auteure de sa neutralité assumée, qui la conduit à présenter dans tous les cas les réussites et les manques des approches en présence.

Sur un plan différent, il faut noter que tous les domaines de l'acquisition ne sont pas couverts. En particulier, les domaines de la phonologie, de la phonétique, et de la pragmatique, ne sont pas du tout évoqués, alors que dans ces domaines aussi, la convergence entre les psycholinguistes et les informaticiens s'est réalisée (en particulier dans le domaine de la phonologie).

D'autres domaines pertinents ont été négligés, là encore sans doute pour des raisons de place. D'abord, on ne mentionne que l'acquisition de la première langue, alors que de nombreux travaux existent sur l'apprentissage d'une langue seconde (en linguistique, en psycholinguistique et en modélisation). Il faudrait ajouter le domaine de l'acquisition pathologique, qui est certes un domaine plus délicat, mais où, là encore, une convergence est ébauchée entre les psycholinguistes et les informaticiens. Enfin, il serait intéressant d'ajouter à ce panorama les domaines où la psycholinguistique et la modélisation trouvent des points de rencontre, mais qui ne concernent pas directement l'acquisition. Par exemple, les études psycholinguistiques sur le temps de traitement (par exemple en syntaxe) font intervenir une notion de complexité qui peut être approchée par des modèles informatiques venant de la communauté du « *parsing* » (probabiliste).

Mais on s'éloigne clairement du domaine couvert par cet ouvrage de synthèse, dont la lecture doit être absolument recommandée, malgré les quelques remarques qui précèdent, à tout chercheur ou étudiant intéressé par l'application des techniques du TAL à des investigations psycholinguistiques.

Ismaïl BISKRI, Adel JEBALI, Traitement automatique des langues naturelles : de l'analyse à l'application, *Hermès-Lavoisier*, 2011, 256 pages, ISBN 978-2-7462-3138-7.

Lu par **Elisa Lavagnino**

Centre de recherche CeRTeM – Université de Gênes

L'ouvrage met en relief la multidisciplinarité du traitement automatique des langues : la réussite du traitement des langues repose sur une étude correcte des principes théoriques qui inspirent les sciences humaines pour arriver ensuite aux applications informatiques. Dans cette perspective, l'ouvrage introduit les bases théoriques sur lesquelles s'appuient les investigations présentées dans les chapitres suivants.

Le livre commence par une description de la théorie des actes de discours et présente les lois qui règlent la logique de la communication à travers la théorie des universaux linguistiques qui permettent à toutes les langues naturelles de remplir les fonctions d'expression et de communication. L'auteur définit les investigations sur les universaux linguistiques comme interdisciplinaires, utilisant les ressources des différentes sciences qui s'occupent de la communication et de l'action (la logique, la philosophie du langage, la linguistique, l'anthropologie, les sciences cognitives, la psychologie, les mathématiques, l'informatique, etc.).

Le deuxième chapitre présente le rôle de la causalité intentionnelle dans la compréhension du comportement du locuteur en relation avec son état mental. L'auteur établit une liaison entre raison et comportement, cette relation peut être de nature logique et causale. À la base de ses travaux, on retrouve le besoin fondamental de communication des agents humains, ce processus englobe le codage et le décodage de la signification et la maîtrise d'un système communicatif commun. La causalité intentionnelle montre la relation de cause à effet entre les états mentaux et les comportements langagiers et non verbaux des agents. Les raisons qui sont à la base du comportement humain rationnel sont logiques et causales et justifient le besoin d'étudier la causalité intentionnelle dans les sciences humaines.

Le troisième chapitre introduit l'ébauche d'un projet plus grand fondé sur la conviction suivante : les travaux plus récents en philosophie analytique permettent de modéliser les attitudes, les actes illocutoires et les actions, à travers une définition de leur structure interne qui est suffisamment élaborée pour permettre le développement d'une logique complète. Ces structures peuvent constituer la base, pour l'implémentation, d'un modèle multi-agent qui dévoile les relations combinatoires entre les éléments du système (agents, attitudes et propositions). Le constat à la base des travaux de Bergier est que les sources d'informations numériques sont considérables et riches dans le domaine des renseignements, notamment les données sous forme discursive : presse, blogs, e-mails, textos et conversations. L'hypothèse est que ces données cachent des renseignements sur les agents et sur leurs implications futures. Une théorie adéquate sur l'action et les attitudes pourrait contribuer à leur dévoilement et modélisation. Les avantages de cette étude résident dans le fait qu'elle ne s'arrête pas qu'aux modèles concernant un seul agent, mais plutôt à ce qu'un groupe d'agents ou une collectivité veut faire.

Le chapitre suivant décrit une étude de corpus visant à l'exploration contextuelle .dont le but est la fouille sémantique de thèses en ligne. Cette étude répond à la demande de plus en plus croissante en linguistique de corpus en outils d'annotation

de textes qui visent au filtrage et à la recherche de l'information, à l'aide à la lecture et à l'analyse des textes. À travers l'exploration contextuelle, on peut retrouver des indicateurs et des indices linguistiques qui peuvent répondre à ces besoins. L'étude utilise la plate-forme EXCOM qui organise, selon des formatages comparables, des classes de marqueurs d'opérations linguistiques et des classes d'indices linguistiques contextuels. Ces classes contribuent ainsi à faire prendre automatiquement des décisions pour lever l'indétermination sémantique de certaines unités et attribuer automatiquement une annotation (grammaticale, sémantique ou discursive) à un segment textuel, dans le but de pouvoir l'extraire ou le stocker dans une base exploitable pour des recherches ultérieures (résumés automatiques, création d'ontologies, etc.). Les deux points de vue considérés sont celui de la définition et celui de la bibliosémantique. Le but est de faciliter la recherche des définitions ou des références bibliographique dans des thèses ou des travaux fondamentaux pour l'avancement dans un domaine, qui souvent ne sont pas pris en compte.

L'exploration contextuelle est aussi reprise dans le chapitre 5 où les travaux présentés visent à l'analyse de textes philosophiques assistés par ordinateur afin d'identifier des relations causales qui sont décrites dans l'ouvrage de Bergson. L'étude vise à approfondir la dimension psychologique du concept de langage. Le but des auteurs est de présenter une méthode de catégorisation d'un concept philosophique qui soit assistée informatiquement afin de produire à l'intérieur des parcours de lecture de la concordance. La plate-forme sémantique d'annotation utilisée est EXCOM, la même que celle employée dans le chapitre précédent. Cette étude s'inscrit dans une méthodologie d'analyse conceptuelle plus large qui vise à l'analyse de la concordance du terme « langage » dans l'ouvrage de Bergson.

Le chapitre 6, « *Les marqueurs de sujet de l'arabe standard : analyse et formation* », introduit les grammaires catégorielles et l'intérêt de la linguistique formelle envers la dichotomie entre les unités de la langue qui fonctionnent comme opérateurs et celles qui fonctionnent comme opérandes. Le chapitre montre que les grammaires catégorielles peuvent rendre compte de certaines spécificités propres à la langue arabe, en particulier les aspects qui concernent l'asymétrie sur le plan de l'accord. Après la description des particularités du système pronominal de l'arabe standard, notamment des formes indépendantes et des formes conjointes qui sont marqueurs d'arguments (sujet ou objet), les auteurs cherchent à différencier les marqueurs de sujet des marqueurs d'objet, en soulignant l'ambiguïté des marqueurs de sujet. Les résultats de cette analyse ont été intégrés dans un modèle formel appelé « grammaire catégorielle combinatoire applicative ».

Le chapitre 7 présente une étude sur les types catégoriels et leur contexte d'utilisation à travers l'apprentissage automatique fondé sur des principes bayésiens. De fait, la difficulté principale dans des recherches de type catégoriel est d'établir le type qui doit être attribué à chaque unité lexicale.

Le dernier chapitre de cet ouvrage cherche à nommer des classes d'unités lexicales et à représenter les relations qui les relient. Le domaine pris en

considération est la paléoanthropologie. Les auteurs s'appuient sur l'application des représentations évolutives fondées sur la topologie et des logiques non monotones déjà utilisées en linguistique. Ils introduisent une échelle de typicalité qui permet de définir des degrés, où des entités individuelles qui se réfèrent à une classe sont plus ou moins typiques selon les traits qui les caractérisent, ainsi qu'un concept de frontière à épaisseur pour modéliser l'échelle de typicalité : plus il y a d'éléments avec un degré faible de typicalité qui sont rattachés à une classe, et plus la frontière est épaisse. Cette étude peut s'étendre vers une formalisation de la notion de classes de typicalité, vers l'application à d'autres domaines apparentés et vers la mise au point d'une interface graphique qui permet de visualiser les relations de parenté des éléments.

Les investigations présentées dans cet ouvrage démontrent qu'une approche multidisciplinaire qui s'appuie sur les applications informatiques et le traitement automatique des langues est primordiale pour tester la robustesse des modèles linguistiques fondés sur la théorie des actes du discours et plus généralement sur les théories linguistiques formelles et avancer ainsi dans les analyses.

Hang LI, Learning to Rank for Information Retrieval and Natural Language Processing, *Morgan & Claypool publishers*, 2011, 101 pages, ISBN 978-1-6084-5707-6.

Lu par **Marie CANDITO**

Université Paris Diderot – Alpage, équipe projet INRIA

L'ouvrage traite de la tâche d'apprentissage automatique de modèles d'ordonnancement d'objets, principalement dans le cadre de l'apprentissage supervisé. L'ouvrage fournit des exemples de tâches de TAL ou de recherche d'information qui relèvent de l'apprentissage d'ordonnancement. Il distingue les modèles créant un ordre sur des objets et les modèles qui agrègent plusieurs ordres préexistants. Il fournit une typologie des approches utilisées (point à point, par paire, par liste), et fournit une revue détaillée de quatorze méthodes, avec la partie théorique et l'évaluation sur des jeux de données standard.

Cet ouvrage présente un panorama des connaissances théoriques et des techniques utilisées pour apprendre des modèles d'ordonnancement[1] d'objets, essentiellement dans le cadre de l'apprentissage automatique supervisé. L'auteur cite en introduction quelques exemples où cette tâche est centrale pour différentes applications en TAL et en recherche d'information :

1. Pour traduire *ranking* j'utiliserai *ordonnancement* ou *ordre* selon que l'on parle de la tâche ou de son résultat.

– la réponse d'un système de recherche documentaire met en œuvre l'ordonnancement des documents réponses censé refléter un ordre de pertinence à la requête ;

– la réponse d'un système de filtrage collaboratif met en œuvre l'ordonnancement de produits censé refléter un ordre de préférence de ces produits par l'utilisateur ;

– la tâche d'un métamoteur de recherche implique d'ordonner la fusion des résultats de recherche de moteurs sous-jacents ;

– en traduction automatique, une approche peut impliquer un premier ordonnancement de phrases candidates à la traduction, permettant d'isoler les x premières phrases candidates et d'appliquer une étape de réordonnancement des x premières phrases candidates en utilisant un modèle plus sophistiqué.

L'ouvrage comporte sept chapitres, de taille inégale. Le chapitre 1 fournit une définition formelle de l'apprentissage d'ordre, et introduit la distinction entre deux tâches d'apprentissage : apprendre à créer un ordre (*ranking creation*) et apprendre à faire une agrégation d'ordres (*ranking aggregation*). La première tâche consiste à apprendre un modèle permettant d'ordonner des objets relativement à une « demande ». Par exemple, les demandes sont respectivement les requêtes ou les phrases sources dans le cas de la recherche documentaire et de la traduction automatique. La deuxième tâche consiste à apprendre à ordonner l'union de plusieurs listes ordonnées d'objets, étant donné une demande. C'est typiquement le cas pour un métamoteur de recherche.

Le chapitre 2 décrit formellement l'apprentissage de modèles de création d'ordre. L'auteur prend le cas de la recherche documentaire pour exemple conducteur pour tout le chapitre, et fait valoir que les modèles historiques de recherche documentaire (modèles par similarité vectorielle ou modèle probabiliste), sont fixés (avec seulement quelques paramètres à régler). Au contraire, l'apprentissage de modèles discriminants permet d'introduire de multiples « signaux » de pertinence.

Dans ce cas les données d'entraînement sont un ensemble d'exemples, où chaque exemple est constitué d'une requête et d'une liste ordonnée de documents. Chaque couple requête et document est représenté par un vecteur de traits. L'information supervisée est contenue dans l'ordre associé à ces couples, ordre pouvant être partiel, sous la forme de labels de pertinence associés aux documents.

Le chapitre détaille les techniques de création de données d'entraînement (manuelles ou automatiques, ce qui dans le cas de la recherche documentaire peut être fait en mettant à profit les clics opérés par les utilisateurs dans la liste de résultats) ainsi que les métriques d'évaluation de modèles d'ordonnancement.

L'auteur fait ensuite le lien entre l'apprentissage d'ordonnancement et trois autres tâches d'apprentissage que sont la classification, la régression et la classification ordinale, toutes trois pouvant servir d'approximation pour apprendre

un modèle d'ordonnancement. Enfin l'auteur introduit la distinction centrale entre trois catégories d'approche pour l'apprentissage de modèles d'ordonnancement :

– dans l'approche point à point, la structure des données d'entraînement n'est pas prise en compte : chaque couple vecteur (requête + document) et son label ou score de pertinence constitue un exemple isolé. Les techniques classiques de classification ou de régression peuvent être utilisées ;

– dans l'approche par paires, des exemples binaires sont construits à partir des données d'entraînement : pour chaque paire de documents, xi xj, on dérive un exemple positif si xi est ordonné plus haut que xj, et sinon il est négatif. Un classifieur classique peut être entraîné pour classifier des paires de documents. Dans le cas linéaire (ou linéaire après utilisation d'un noyau) ce même classifieur peut être utilisé directement pour ordonner les documents ;

– enfin, dans l'approche plus récente par liste, on utilise pour chaque exemple d'entraînement la liste ordonnée d'objets (et pas les points individuels ou des paires uniquement).

L'auteur liste les caractéristiques et les techniques relevant de chacune de ces approches, qui seront détaillées au chapitre 4, et fournit des résultats d'évaluation sur différents jeux de données librement disponibles.

Le chapitre 3 décrit formellement l'apprentissage de modèles pour l'agrégation d'ordre. Le chapitre 4, largement le plus long, est consacré à une revue détaillée de onze méthodes d'apprentissage d'ordonnancement (variantes pour ordonnancement de perceptron, de SVM, variantes appliquées à la recherche d'information, variantes fondées sur le *boosting*...) et trois méthodes d'apprentissage d'agrégation d'ordres (*Borda count*, chaînes de Markov, *CRanking*). Pour chacune des méthodes sont présentés le modèle mathématique et l'algorithme d'apprentissage.

Le chapitre 5 présente très succinctement des applications qui peuvent se modéliser comme une tâche d'apprentissage de création ou d'agrégation d'ordre : la recherche documentaire sur le Web, le filtrage collaboratif, l'extraction de phrases clés, le résumé d'après une requête (typiquement choisir les phrases ou extraits à présenter, pour un document réponse, à un utilisateur de moteur de recherche), le réordonnancement dans le cadre de la traduction automatique.

Le chapitre 6 décrit assez rapidement (5 pages) les aspects théoriques de l'apprentissage d'ordonnancement, et le chapitre 7 est consacré aux pistes de recherche actuelles dans ce domaine : l'auteur classe en huit catégories les travaux actuels du domaine (comme la création automatisée de données d'entraînement, l'apprentissage semi-supervisé, l'apprentissage de traits, l'adaptation de domaine...), il en décrit brièvement les enjeux et pointe vers les références de travaux correspondants.

L'ouvrage s'adresse à un public connaissant les principes de base d'apprentissage automatique et offre un bon aperçu des connaissances dans le

domaine ciblé, avec une classification des approches, et une description des techniques relevant de chacune d'elles. Cependant, certaines parties cruciales pourraient être plus développées, en particulier la distinction point à- point, par paire et par liste.

Chu-ren HUANG, Nicoletta CALZOLARI, Aldo GANGEMI, Alessandro LENCI, Alessandro OLTRAMARI, Laurent PREVOT (éditeurs). Ontology and the lexicon. A Natural Language Processing Perspective. *Cambridge University Press.* 2010. 339 pages. ISBN 978-0-5218-8659-8.

Lu par **Yannis HARALAMBOUS**

Télécom Bretagne

Cet ouvrage fait, de manière spectaculaire, le point sur les interactions entre ontologies et ressources lexicales, et plus généralement, entre les disciplines de TAL et d'extraction de la connaissance. Chaque chapitre est écrit par une équipe différente, mais, néanmoins, on s'écarte du modèle « actes d'un colloque » puisqu'il y a une cohérence interne, des renvois fréquents entre chapitres et une bibliographie commune. À l'image de cette vaste question qui est la synergie entre les points de vue des deux disciplines, les contributions sont tantôt théoriques, tantôt applicatives, tantôt des discussions générales et tantôt des focalisations sur des sujets bien précis. Cette diversité, qui fait partie de la richesse de l'ouvrage, m'a conduit à présenter ici plus en détail la partie I qui traite des aspects fondamentaux.

Partie I : aspects fondamentaux

Le premier (et sans doute le plus important) chapitre pose les bases des interactions entre ontologies et lexiques. Il commence par définir les ontologies comme étant des « spécifications de conceptualisations partagées ». Or, le flou artistique qui règne est tel que cette définition pourrait très bien être appliquée aussi aux lexiques. C'est l'interprétation de ces deux notions, la « conceptualisation » et la « spécification », qui fait la différence entre les deux types de ressources. En effet, le lexique classe les mots, et va, par exemple, se servir amplement de relations de synonymie. L'ontologie, en revanche, fait abstraction des mots et s'intéresse aux concepts ; la relation de synonymie y est inconnue puisqu'on s'intéresse au sens et non aux mots qui le représentent. Mais, pauvres humains que nous sommes, comment représenter des concepts, qui font abstraction de mots, autrement que par des mots ? Différentes solutions ont été données, comme l'utilisation d'un mot « tout en précisant que le choix du mot est sans importance » (ce qui est légèrement irréaliste puisqu'il n'y a jamais de véritable synonymie) ou alors la donnée de plusieurs mots représentant le même sens ou la donnée d'une description textuelle du concept (glose). L'illustre ressource WordNet combine les deux dernières solutions.

Autre terme important dans la définition des ontologies : la « spécification ». Les ontologies dites « formelles » sont décrites dans un langage formel (par exemple la logique du premier ordre, ou une logique de description). Lorsqu'on fait le compromis de se servir de textes, l'ontologie devient semi-formelle (c'est le cas de WordNet), voire même informelle. Il est beaucoup plus difficile d'imaginer des « lexiques formels », puisque, contrairement aux concepts, les mots échappent à la rigueur du langage formel.

Dans la suite, il est question de « relations ». Les auteurs mettent en garde le lecteur du fait que souvent on utilise les mêmes noms pour des relations au niveau des ontologies et des lexiques. Ainsi, les auteurs précisent que l'hyperonymie est une relation lexicale qui ne doit pas être confondue avec la relation « *is_a* » des ontologies, mais, hélas, ne donnent pas d'exemple de cas problématiques. Quoi qu'il en soit, les relations, de part et d'autre, font partie intégrante de l'interface entre ontologies et lexiques, qui est le principal sujet de l'ouvrage. Après un rappel des bases théoriques de cette interface dans les cadres cognitif, philosophique, lexicographique et linguistique, les auteurs terminent le chapitre en insistant sur le fait qu'ils ne voient aucune primauté de l'ontologique vis-à-vis du lexical ou *vice versa*, et que chacune des deux approches peut servir à améliorer les ressources de l'autre.

Le chapitre 2, très court et factuel, a été coécrit par Christiane Fellbaum, l'une des instigatrices du projet WordNet. Il s'agit de comparer WordNet et l'ontologie de haut niveau SUMO. WordNet est qualifié d'« ontologie lexicale » puisque les concepts (appelés « synsets », c'est-à-dire ensembles de synonymes) sont décrits par des termes d'une langue donnée. Un exemple intéressant qui illustre la différence entre ontologies formelle et lexicale est la dépendance de la deuxième de la langue utilisée. Ainsi, les auteurs donnent l'exemple des « véhicules sur rails » : il n'existe en anglais aucun terme pour ceux-ci, et de ce fait aucun synset dans WordNet. Mais le concept de « véhicule sur rail » existe bel et bien dans une ontologie formelle « par la géométrie des relations » puisqu'il s'agit du parent commun des concepts de « train » et de « tramway » dans le graphe. Autre différence importante : un lexique doit refléter l'état d'une langue et les lexicographes n'ont pas le droit d'omettre tel ou tel mot en le considérant comme superflu ou inintéressant. Une ontologie formelle est une construction de l'esprit, où les concepts et les relations ne sont déterminés que par des axiomes formels, au point où on a l'impression que la représentation linguistique des concepts devient inutile...

Après ces considérations qui clarifient en grande partie le premier chapitre, les auteurs décrivent leur projet : la mise en correspondance entre SUMO et WordNet. Une description des différentes étapes de ce projet est suivie de celle des résultats obtenus. Les auteurs soulignent d'ailleurs le fait que si les WordNets des autres langues en faisaient autant, SUMO permettrait la mise en correspondance des différents WordNets, dans l'esprit de ce qui a déjà été réalisé pour le projet EuroWordNet. Les auteurs affirment que la rigueur du langage formel utilisé par

SUMO justifie la pertinence de ce projet de « WordNet global » pour le TAL multilingue.

Le chapitre 3 peut, à première vue, paraître très similaire au chapitre 2. En effet, ici aussi il s'agit de mise en correspondance entre WordNet et une ontologie formelle – dans ce cas, il s'agit de DOLCE (ontologie descriptive pour l'ingénierie linguistique et cognitive). Mais les auteurs ne s'arrêtent pas à la mise en place de cette correspondance. Ce qu'ils cherchent, c'est l'« ontologisation » de WordNet, l'interprétation des relations lexicales de WordNet en tant que relations ontologiques, et sa validation à l'aide du langage formel de l'ontologie. Ils appellent le résultat de ce projet ambitieux, OntoWordNet. Le principal ingrédient de l'ontologisation de WordNet (ou de son « nettoyage », comme ils disent), est l'application d'une méthodologie de création d'ontologies, appelée « OntoClean » et développée par les mêmes auteurs. Grâce aux contraintes de cette méthodologie, les auteurs proposent une nouvelle organisation de la partie supérieure de WordNet et une fusion de celle-ci avec l'ontologie DOLCE, qui est longuement décrite dans le chapitre. En guise de conclusion, les auteurs se demandent si le WordNet réorganisé qu'ils proposent apportera un gain de performance à la multitude d'applications de cette ressource en TAL.

Le chapitre 4 est coécrit par Collin F. Baker, un des collaborateurs de Fillmore et artisans de son projet FrameNet. Il s'agit dans ce chapitre d'utiliser FrameNet comme outil de jonction du texte avec des ontologies formelles, et permettre ainsi l'utilisation de moteurs d'inférence sur la connaissance extraite du texte. Après une introduction à FrameNet, à la sémantique des cadres et à divers types de relations entre les cadres, les auteurs comparent les types sémantiques de FrameNet aux concepts des ontologies. Ils concluent que FrameNet étant un projet de nature lexicographique, il est normal qu'il s'attache plus à la modélisation de la langue qu'à celle du monde. La complémentarité entre ces deux approches les conduit tout naturellement à l'utilité d'une mise en correspondance entre FrameNet et des ontologies de haut niveau (comme SUMO) ou, plus généralement, avec des ontologies décrites en KIF, CycL ou OWL. Mais pour cela il faut réécrire FrameNet dans l'un de ces langages. Ils choisissent donc de le réécrire en OWL-DL et décrivent leurs choix de conception de cette nouvelle ontologie, appelée ontologie FrameNet. Ainsi, tout texte annoté en FrameNet donne lieu à un sous-graphe de cette ontologie et là-dessus on peut utiliser des moteurs d'inférence fondés sur les logiques de description. Les auteurs terminent le chapitre en insistant sur le fait qu'ils ne voient aucune primauté de l'ontologique vis-à-vis du lexical ou *vice versa*, et que chacune des deux approches peut servir à améliorer les ressources de l'autre.

Le chapitre 5 est une synthèse des différents projets d'interaction entre ontologies et ressources lexicales.

Les parties II à IV

La partie II traite de la découverte et de la représentation de systèmes conceptuels. Le chapitre 6 se pose la question de la génération semi-automatique

d'ontologies. Dans le chapitre 7, il est question de la ressource OntoSem. Dans le chapitre qui suit, deux linguistes de Hong Kong reprennent les 540 radicaux de Xu Shen et les font correspondre à des concepts de l'ontologie de haut niveau SUMO. Le chapitre 9, très théorique, présente un cadre formel pour représenter des primitives de la théorie de grammaire de construction incarnée et l'utilise pour définir une ontologie.

La partie III traite de l'interface entre ontologies et ressources lexicales. Dans le premier chapitre, on trouve une synthèse et une classification de ressources existantes qui combinent des ontologies de haut niveau et WordNet. Le chapitre 11 présente BOW, un WordNet ontologique bilingue (chinois/anglais) et le chapitre 12, LingInfo, un modèle d'inclusion d'information linguistique dans les ontologies. Enfin, dans le chapitre 13, les auteurs se posent la question de la fusion d'une ontologie générale avec une ontologie spécialisée.

La partie IV porte sur l'apprentissage et l'utilisation de la connaissance ontologique. Le chapitre 14 est une longue discussion sur l'utilité de l'interface Ontolex pour le TAL, le processus de création de connaissances « Ontolexicales » par le biais du TAL, ainsi que le choix de corpus et les méthodes d'extraction de connaissances. Le chapitre 15 est dédié à une ressource spécifique : l'ontologie Oméga et, dans le chapitre 16, il est question de l'acquisition de connaissances lexico-sémantiques pour un système de questions/réponses. Enfin, le dernier chapitre présente la construction semi-automatique d'une ontologie thaïe sur l'agriculture.

Conclusion

Situé à la frontière des communautés du TAL et d'extraction de la connaissance, ce livre ravira les membres de l'une de ces communautés qui souhaitent savoir comment interagir avec l'autre et profiter d'avancées communes. Il fait le point sur une multitude de ressources et d'approches théoriques et, en le parcourant, on ne cesse de retrouver des pistes et des idées intéressantes. Le lecteur non averti peut être découragé par la densité du premier chapitre (qui ne traite, après tout, que de « fondamentaux ») – dans ce cas, il serait plus judicieux de commencer la lecture par le deuxième, voire même les chapitres 5 et 14 qui sont tout à fait abordables.

Je recommande ce livre à toute personne souhaitant acquérir une vision étendue et approfondie des interactions et synergies du traitement automatique des langues et de l'extraction de connaissances.

Violeta SERETAN, Syntax-Based Collocation Extraction, *Springer*, 2011, 217 pages, ISBN 978-9-4007-0133-5.

Lu par **Agnès TUTIN**

LIDILEM, Université Grenoble 3 – Stendhal

L'ouvrage de Violeta Seretan, réalisé à partir de sa thèse de doctorat à l'Université de Genève au LATL sous la direction d'Eric Wehrli, porte sur l'extraction des collocations à partir de techniques syntaxiques et statistiques. L'approche proposée est délibérément multilingue – l'application visée à long terme est la traduction automatique – et exploite une analyse syntaxique profonde, réalisée à partir de l'analyseur FIPS (Wehrli, 2007). L'expérimentation est menée sur une large échelle et sur plusieurs langues.

Après l'introduction qui traite la problématique de l'extraction des collocations fondée sur la syntaxe, le deuxième chapitre aborde la notion controversée de collocations à travers des exemples comme *widely available, to meet a requirement*. L'intérêt de telles expressions est souligné par de nombreux linguistes, mais comme le montre Violeta Seretan, la définition de cette notion reste bien malaisée et les contours flous, même si de nombreux linguistes s'accordent à dire que les collocations se situent à mi-chemin entre des expressions libres et des expressions complètement figées. Violeta Seretan présente les types de critères exploités pour ces définitions (critères statistiques et linguistiques) et les principaux modèles linguistiques qui recourent à cette notion (avec une acception souvent variable selon les modèles) comme le contextualisme anglais ou la théorie sens-texte. La définition qu'elle propose pour son propre travail recourt aux critères suivants : les collocations sont des éléments reliés par des relations syntaxiques, préfabriqués, imprédictibles, récurrents et de longueurs indéterminées, les collocations au-delà de deux éléments étant des collocations enchâssées. En outre, Violeta Seretan adopte la distinction introduite par Hausmann entre la base, l'élément autosémantique de la collocation et le collocatif, l'élément synsémantique, distinction à l'œuvre dans le modèle des fonctions lexicales de Mel'čuk.

Le chapitre 3 expose les techniques d'extraction des collocations, essentiellement à base statistique. Les collocations étant généralement considérées comme des associations binaires, différentes mesures d'association sont employées, couplées avec des filtrages linguistiques au besoin. L'exploitation des mesures d'association repose sur l'idée que les candidats à la collocation seront des éléments qui tendent à apparaître en cooccurrence de façon significative, c'est-à-dire davantage que ce que la distribution normale de chacun des éléments laisserait attendre, à partir d'un calcul des fréquences des éléments dans les textes, consignées dans une table de contingence. La méthode générale consiste à identifier des candidats selon des critères spécifiques (mots apparaissant dans une fenêtre de mots ou entretenant une relation syntaxique spécifique) puis dans un second temps à calculer les mesures d'association reliant les candidats. Violeta Seretan présente alors de façon très didactique les différents types de tests : t-score, z-score, chi-carré, log-likelihood ratio, l'information mutuelle, dont aucun ne semble faire l'unanimité même si le log-likelihood ratio semble plus robuste que les autres mesures par rapport aux effets de fréquence. Elle montre ensuite l'intérêt des prétraitements linguistiques, de la lemmatisation, pour les langues à morphologie riche comme le français, et surtout l'analyse syntaxique, pour les langues à ordre des mots libres, et

présente les expérimentations d'extraction dans leur diversité, dont certaines sont déjà assez anciennes, effectuées dans différentes langues, principalement l'anglais, l'allemand et le français.

Le chapitre 4 est le cœur de l'ouvrage puisqu'il présente l'expérimentation propre menée par Violeta Seretan, qui consiste en une extraction des collocations dans un contexte multilingue fondée sur une analyse syntaxique profonde. L'analyse syntaxique réalisée à l'aide de FIPS, dont la modélisation est inspirée de la grammaire générative, traite plusieurs types de phénomènes complexes (passivation, relativation, interrogation, dislocation…) et permet ainsi de réaliser une analyse syntaxique « profonde » où des associations lexicales rencontrées dans différentes alternances syntaxiques seront rapprochées. Une fois les candidats identifiés à partir des patrons syntaxiques divers (du type nom-adjectif ou sujet-verbe, par exemple, y compris quelques associations incluant des mots grammaticaux comme verbe-préposition) dans le corpus, une mesure d'association, le log-likelihood ratio, est appliquée. L'extraction à base syntaxique est comparée à une extraction morphosyntaxique dans une fenêtre de plusieurs mots, à la fois sur un corpus monolingue de la partie française du corpus Hansard, et sur quatre langues du corpus Europarl (anglais, français, italien, espagnol). Les résultats, évalués par des linguistes sur plusieurs types d'expressions, montrent que la méthode à base syntaxique est nettement meilleure au niveau de la précision (de 9 % dans l'extraction monolingue, de presque 20 % en moyenne dans l'extraction des expressions polylexicales et des collocations pour les quatre langues) en particulier pour les paires les moins fréquentes, de nombreuses erreurs étant liées aux erreurs d'identification des relations syntaxiques avec la méthode de l'étiquetage. Des études ponctuelles sur quelques expressions montrent également que le rappel apparaît meilleur, sauf dans certains cas mal traités par l'analyse syntaxique comme les anaphores ou les noms « transparents » (exemple : *espèce,* dans *jouer une espèce de rôle*).

Le dernier chapitre est consacré aux extensions de la méthode dans plusieurs directions. La première concerne l'extraction de collocations complexes, au-delà de deux éléments. La méthode sélectionnée ici est de réaliser un post-traitement à partir des collocations extraites, essentiellement des collocations partageant une base commune comme dans *prendre une décision difficile* (*prendre une décision + décision difficile*), en utilisant les mesures d'association entre n-grammes extraits, méthode qui apparaît particulièrement encourageante pour les expressions polylexicales à trois termes. La deuxième extension concerne un thème tout à fait intéressant, celui de la mise au jour par induction de patrons syntaxiques récurrents d'associations, utilisant également le log-likelihood ratio, qui confirme toutefois la productivité des patrons classiques pour les collocations lexicales et met en évidence quelques patrons traditionnellement associés aux collocations grammaticales. Enfin, une expérimentation menée sur une extraction de collocations multilingues, utilisant des corpus alignés et l'analyse syntaxique profonde, donne de bons résultats pour la

précision et le rappel et montre que l'utilisation d'un dictionnaire bilingue n'accroît que marginalement la qualité des résultats.

Commentaire

L'ouvrage de Violeta Seretan constitue un très bon état de l'art concernant l'extraction des collocations pour qui s'intéresse à cette problématique. La présentation synthétique, claire, comporte de nombreux tableaux et des annexes détaillées permettant d'appréhender concrètement la problématique. Les différentes mesures d'association utilisées et les méthodes d'évaluation y sont expliquées de façon très didactique, même pour les non-spécialistes en méthodes statistiques comme l'auteure de ces lignes. L'analyse quantitative et qualitative des résultats y est soigneusement effectuée, et elle montre clairement, mais de façon non caricaturale l'intérêt (et les difficultés de mise en œuvre) des méthodes à base syntaxique et leur complémentarité avec des méthodes plus rustiques, ce que d'autres expérimentations avaient déjà montré mais sur des études de moindre envergure et surtout dans un contexte monolingue. La réflexion autour de l'évaluation humaine des résultats obtenus, non triviale pour ce type d'expressions, n'est pas éludée et la proposition dans la deuxième expérimentation d'évaluer finement les résultats en plusieurs types d'expressions apparaît tout à fait intéressante. L'analyse qualitative occupe une part non négligeable et permet de comprendre la complémentarité des techniques employées.

La synthèse des recherches en linguistique sur la question des collocations va au-delà d'un simple état de l'art en TAL et révèle une très bonne connaissance de la problématique non seulement en TAL, mais également en linguistique théorique où la difficulté de définir la notion de collocation est clairement soulignée. L'approche adoptée par l'auteure s'inscrit dans la lignée de Hausmann et Mel'čuk même si sa conception des collocations n'est pas toujours conforme à ce modèle en intégrant des éléments grammaticaux généralement plutôt associés à des colligations, par exemple des structures du type ADJ-PREP qui ne correspondent tout à fait au schéma de la dissymétrie base et collocatif.

Comme le montre l'auteure, la méthode à base syntaxique et statistique apparaît clairement plus satisfaisante que la méthode fondée sur de simples fenêtres. Toutefois, l'utilisation de patrons syntaxiques prédéfinis présente également quelques limites dans la mesure où elle ne permet pas de repérer des expressions ne correspondant pas à des schémas moins classiques, par exemple des associations de type ADJ-ADJ comme *ivre mort* ou *amoureux fou* ou des collocations plus complexes comme des *similes* du type *bête à manger du foin* ou *laid à faire peur*, qui répondent à des schémas moins productifs mais néanmoins réguliers. La disponibilité des corpus analysés syntaxiquement devrait cependant pouvoir permettre le repérage d'expressions lexicalisées au-delà des expressions binaires même si les calculs statistiques s'en trouvent complexifiés.

Enfin, si Violeta Seretan montre clairement l'intérêt des techniques utilisées pour l'extraction de lexiques de collocations multilingues, on peut regretter, dans

l'ouvrage, l'absence d'expérimentation dans l'étude en traduction automatique qui aurait démontré l'intérêt de ce type de ressources pour cette application.

Pour conclure, l'ouvrage de Violeta Seretan présente une très bonne synthèse des techniques d'extraction à base syntaxique et statistique qui intéressera non seulement les spécialistes du TAL mais aussi les linguistes et les chercheurs en traduction. On peut désormais le considérer comme un ouvrage de référence sur la question des expressions polylexicales et des collocations en TAL.

Jörg TIEDEMANN, Bitext Alignment, *Morgan & Claypool publishers*, 2011, 153 pages, ISBN 978-1-6084-5510-2.

Lu par François YVON

Université Paris Sud, Limsi/CNRS

Le traitement automatique de collections documentaires multilingues est un sujet d'actualité, tant ces collections sont des ressources critiques pour développer des systèmes de traduction automatique, en particulier des systèmes statistiques. S'appuyant sur une expertise peu commune sur les problématiques liées à l'alignement de corpus bilingues, Jörg Tiedemann offre, avec « Bitext Alignment *», un tour d'horizon complet et fort bien documenté de l'état de l'art, qui permet de mieux comprendre les recherches actuelles sur ces questions.*

Bitext Alignment s'ouvre sur un très court chapitre, qui introduit brièvement la notion de *bitexte* (associant un texte en langue source et sa traduction en langue cible) ainsi que les différents types d'alignements qui peuvent être envisagés pour expliciter les correspondances entre fragments sources et fragments cibles, et les applications possibles de ces alignements.

Cette mise en bouche se prolonge au chapitre 2, dans lequel l'auteur s'attache à clarifier un certain nombre de notions et de termes relatifs à ces alignements de *bitextes*, à définir les différentes formes (symétriques, asymétriques, fonctionnels, hiérarchiques, etc.) qu'ils peuvent prendre, enfin à dresser une cartographie des indices d'association, des méthodes et des algorithmes que ces points de vue variés impliquent. Le chapitre se clôt par une rapide discussion de la question de l'évaluation des alignements. Dans l'ensemble, ces clarifications terminologiques sont d'autant plus appréciables que la littérature du domaine est parfois un peu flottante. Le second mérite de ce chapitre est de mettre en évidence la similitude entre les problématiques d'alignements de phrases et celles qui concernent les alignements de mots et de les présenter dans un cadre unifié. Cette présentation met donc ainsi en évidence l'importance de la segmentation en unités élémentaires qui servent de support à ces liens d'alignements. Le lecteur est ainsi mis en garde : les alignements ne se conçoivent, ne se calculent et ne s'évaluent qu'en relation avec des segmentations source et cible, qui déterminent la granularité de ces alignements et en limitent la précision. Le reste du livre est alors organisé, de manière classique,

en considérant successivement plusieurs niveaux de granularité des unités considérées : le document, la phrase, le mot et, enfin, le syntagme.

Le chapitre 3 est donc consacré à l'alignement de documents, c'est-à-dire à la construction de bitextes. Après avoir précisé cette notion et mis en évidence la gradualité, l'existence de degrés variables de parallélismes entre textes, J. Tiedemann souligne que, quelle que soit la forme sous laquelle on les envisage, ces ressources restent relativement difficiles d'accès, à l'exception de quelques sources bien identifiées : institutions internationales, communauté du logiciel libre, brevets, etc. Ces sources ne couvrent que peu de paires de langues, de domaines et de genres, et les producteurs professionnels de traductions sont en général peu enclins à ouvrir leurs archives. Pour contourner ces limites, il faut s'en remettre à des techniques automatiques par exploitation de sites Internet multilingues, ou encore par fouille de *corpus comparables,* deux techniques qui sont rapidement évoquées dans la seconde partie de ce chapitre.

Supposant construits ces bitextes, l'auteur présente au chapitre 4 les méthodes utilisées pour construire des alignements de phrases. Pour ce niveau de granularité, les textes cibles tendent à reproduire le séquencement des textes sources dont ils sont issus, ce qui contraint très fortement les alignements possibles. S'appuyant sur cette observation, plusieurs méthodes heuristiques très simples ont été proposées dès le début des années 1990, qui s'appuient soit sur la forte corrélation entre les longueurs des phrases parallèles, soit sur la présence dans des phrases parallèles de « cognats », soit encore qui combinent ces deux types d'informations. J. Tiedemann fait de ces travaux bien connus et déjà anciens une présentation particulièrement claire, qui s'appuie sur une fine connaissance de l'efficacité de ces divers indices. En comparaison, les rares travaux plus récents, qui s'appuient souvent sur des stratégies de raffinement itératif, sont plus brièvement évoqués : sans doute leur présentation présupposerait de connaître les techniques de base pour aligner les mots, qui ne sont présentées que plus tardivement. Le chapitre se clôt par une étude portant sur l'alignement de sous-titres, qui illustre la nécessité de s'appuyer sur des heuristiques et des ressources qui soient bien adaptées aux spécificités des documents traités.

Par comparaison avec les alignements de phrases, les alignements de mots sont incomparablement plus difficiles à construire, puisqu'à ce niveau de granularité, l'ordre respectif des unités n'est en général pas préservé au cours de la traduction. Ces alignements, qui ont concentré l'essentiel de l'effort de recherche de ces dernières années, font l'objet du copieux chapitre 5. L'auteur y fait une présentation très complète des méthodes visant à aligner des phrases dans leur intégralité, qui débute par les modèles génératifs « classiques » des années 1990 (IBM (1-5) et HMM), se poursuit par l'étude de diverses méthodes heuristiques alternatives, visant en particulier à produire des alignements symétriques, et s'achève par l'exposé de quelques tentatives plus récentes et moins bien connues, qui s'appuient sur des méthodes d'apprentissage discriminants. Au regard de l'exhaustivité de ces premières sections, les quelques pages consacrées à l'extraction de dictionnaires et

de terminologies bilingues données en clôture du chapitre, et qui font la part belle aux alignements différentiels, laissent le lecteur un peu sur sa faim.

Le chapitre 6 aborde, pour finir, un sujet moins bien documenté, celui de la construction d'alignements hiérarchiques. Après avoir exposé les méthodes permettant de construire des alignements d'arbres syntaxiques préexistants, et décrit certaines de leurs limitations, en particulier la difficulté d'appareiller des représentations élaborées indépendamment en langues source et cible, J. Tiedemann revient dans le détail sur les propositions déjà anciennes de D. Wu portant sur l'utilisation de « grammaires synchrones ». Une brève conclusion, des pointeurs vers des ressources publiques, et une très riche bibliographie (plus de 250 références) clôt le livre.

Au final, l'impression qui subsiste est mitigée : s'il faut saluer la volonté de couvrir l'intégralité du spectre des méthodes d'alignement, ce qui, en 125 pages, relevait de la gageure, on regrettera que les développements les plus détaillés et les plus techniques ne portent, à une ou deux exceptions près, que sur des méthodes bien connues et abondamment documentées par ailleurs (méthode de Church & Gale pour l'alignement de phrases ; modèles IBM pour l'alignement de mots, grammaires ITG pour les alignements hiérarchiques, etc.). Quoique l'auteur s'en défende, on pourra déplorer également un certain déséquilibre dans la présentation, qui fait la part belle aux outils utilisés en traduction statistique : il était toutefois difficile de présenter l'alignement hors de toute application. Il reste que ce livre, qui condense une expertise rare, aussi bien pratique que théorique, constitue une mise à jour bienvenue de l'état de l'art dans le domaine, qui renouvelle les ouvrages déjà anciens de Véronis et de Melamed. Il fournit, aux étudiants et aux chercheurs intéressés par le traitement des corpus multilingues, et novices sur ces questions, un point d'entrée fort utile dans cette foisonnante littérature.

Éric GAUSSIER, François YVON, Modèles statistiques pour l'accès à l'information textuelle, *Hermès-Lavoisier*, 2011, 482 pages, ISBN 978-2-7462-2497-1.

Lu par **Benoît Sagot**

Alpage (INRIA – Université Paris 7)

L'ouvrage coordonné par Éric Gaussier et François Yvon et consacré aux modèles statistiques pour l'accès à l'information textuelle rassemble des contributions de plus d'une vingtaine d'auteurs. Il a pour double objectif de présenter les thématiques de recherches qui visent à améliorer l'accès à l'information contenue dans les données textuelles tout en proposant un panorama des techniques statistiques mises en œuvre, lesquelles ont une portée plus générale, et notamment dans tous les domaines du traitement automatique des langues et de la linguistique de corpus.

L'ouvrage est organisé en quatre parties, consacrées successivement à la recherche et à l'extraction d'informations, à la classification et au *clustering* de textes, à la traduction automatique et à des applications émergentes telles que l'exploration de données textuelles. Les différents chapitres sont donc organisés autour du domaine de la fouille de textes, mais couvrent un champ plus vaste que ce seul domaine. Un des points forts de cet ouvrage réside dans ce que les auteurs font des allers-retours permanents entre modèles statistiques employés et problématiques concrètes. Enfin, il se termine par une annexe qui constitue une introduction aux modèles probabilistes mis en œuvre dans les autres chapitres. Cette annexe propose un tour d'horizon clair, concis et précis de différents types de modèles statistiques, ainsi qu'un bref rappel de la théorie des probabilités. Cet ouvrage peut donc servir tout à la fois d'introduction aux approches statistiques pour le traitement de données textuelles et de panorama de l'état de l'art en fouille de textes et dans certains domaines connexes.

Le lecteur moins familier avec les modèles statistiques et la théorie des probabilités tirera le meilleur parti de cet ouvrage en en commençant l'étude par l'annexe A, rédigée par François Yvon. Cette annexe rassemble en effet en une cinquantaine de pages l'ensemble des bases nécessaires à la compréhension du reste de l'ouvrage : la catégorisation supervisée, et notamment le modèle de Bernoulli et le modèle multinomial, l'apprentissage non supervisé au moyen de modèles de mélange, les modèles de Markov pour la modélisation de séquences, et enfin les modèles de Markov cachés. Les dernières pages sont consacrées à quelques rappels de théorie des probabilités.

La première partie de l'ouvrage traite des problématiques de recherche d'information. Le premier chapitre se concentre sur la tâche d'appariement entre documents et requêtes dans un système de recherche d'information. Les mots et leur fréquence dans un ou plusieurs documents étant considérés comme des variables aléatoires, diverses approches probabilistes peuvent être employées. Les auteurs les classent en trois familles : les modèles d'ordonnancement probabiliste, et notamment le modèle Okapi, les modèles de langage et les approches informationnelles. Le chapitre se termine par un rapide comparatif expérimental dans deux configurations différentes. Le deuxième chapitre couvre deux grandes familles de modèles d'ordonnancement : l'ordonnancement d'instances, qui est mis en rapport avec une tâche de classification, et l'ordonnancement d'alternatives. Les auteurs appliquent respectivement ces deux paradigmes au résumé automatique de textes et à la recherche d'information.

Les quatre chapitres suivants, qui forment la deuxième partie de l'ouvrage, couvrent un large éventail de problèmes de classification et de *clustering* (ou partitionnement). Le chapitre 3 étudie les modèles de régressions logistiques, y compris diverses régularisations utilisées pour sélectionner les modèles, et les applique à la catégorisation de textes. Les méthodes à noyaux font, quant à elles, l'objet du quatrième chapitre. Après une introduction générale à ce type d'approches, l'auteur ne cache pas l'importance du choix des noyaux. Il revisite trois

algorithmes classiques dans leur version à noyaux, puis décrit quelques noyaux fréquemment utilisés pour l'analyse de textes, y compris des noyaux adaptés aux arbres et aux graphes. Le chapitre 5 présente un tout autre type de modèles, à savoir les modèles génératifs, et les applique pour l'accès à l'information textuelle. Les modèles génératifs ont pour but de fournir un modèle global (joint) des données observées et des données cachées, contrairement aux modèles discriminants qui cherchent à produire un modèle des données cachées fonction des valeurs observées. L'auteur présente les modèles à base de thèmes (*topic models*). Il rentre alors en détail dans les modèles sur lesquels ils reposent : modèles de thèmes (et notamment la Latent Dirichlet Allocation, LDA et son estimation par échantillonnage de Gibbs), modèles de termes (ainsi la multinomiale à composante de Dirichlet) et modèles de similarité entre documents (y compris les noyaux de Fisher). Le chapitre se termine par une annexe qui recense quelques logiciels qui implémentent des modèles à base de thèmes. Le dernier chapitre de la seconde partie est consacré aux champs markoviens conditionnels (CRF) et à leur application à l'extraction d'informations. C'est à la description de ce domaine d'application du traitement automatique des langues que s'attellent tout d'abord les auteurs. De façon très intéressante, les auteurs mentionnent les approches non statistiques qui peuvent être mises en œuvre dans ce domaine, et expliquent pourquoi les modèles statistiques les plus courants ne sont pas véritablement adaptés à la tâche. Ils vont jusqu'à indiquer que les CRF ont parmi leurs avantages celui d'ouvrir des pistes pour l'exploitation de connaissances linguistiques non statistiques dans les modèles statistiques. Les auteurs rentrent alors dans les détails des CRF, modèle d'étiquetage qui ont des points communs avec divers types de modèles. Ils closent ce chapitre par un exposé de l'intérêt applicatif des CRF et un inventaire des outils disponibles.

La troisième partie de l'ouvrage ne comporte qu'un seul chapitre, consacré à la traduction automatique. Les auteurs, Alexandre Allauzen et François Yvon, rappellent brièvement l'histoire de la traduction automatique, et son importance pour l'accès à l'information. Le chapitre, comme l'ouvrage dans son ensemble, étant consacré aux approches statistiques, il commence par une présentation des plus anciens modèles statistiques pour la traduction automatique, les modèles à base de mots. Les principes généraux étant alors posés, ils rentrent alors dans la description des modèles « état-de-l'art », qui sont à base de segments (*phrase-based*) : techniques d'alignement au niveau des mots (*word-alignment*), modélisation des divergences entre langue source et langue cible quant à l'ordre des mots, et finalement techniques de « décodage », c'est-à-dire pour la traduction effective d'un nouvel énoncé. Enfin, la problématique difficile de l'évaluation de systèmes de traduction automatique est abordée. Le chapitre se termine par deux sections complémentaires : la première traite de récentes avancées dans le domaine (utilisation du contexte en langue source, modèles à base de segments hiérarchiques, exploitation de ressources linguistiques externes) ; la seconde donne là aussi quelques liens vers des outils et ressources disponibles.

La quatrième et dernière partie, intitulée « *Applications émergentes* », comprend deux chapitres. Le premier traite des méthodes et interfaces pour l'accès à des informations élaborées. L'auteur présente différentes applications des méthodes d'analyse de données et des visualisations qui leur sont associées dans le cadre de l'accès à l'information. L'idée générale derrière ces travaux est de ne pas se contenter d'une simple extraction de connaissances pertinentes, par exemple en réponse à une requête, mais également d'appliquer des traitements en aval afin d'en présenter une synthèse à l'utilisateur qui réponde au mieux à ses besoins. Enfin, dans le neuvième chapitre, se pose la question suivante : peut-on voir la détection d'opinions comme un problème de classification thématique ? Les auteurs expliquent dès le début de ce chapitre quelles sont en effet les difficultés auxquelles on est confronté lorsque l'on s'attelle à la tâche de fouille d'opinions (ou analyse de sentiments, en anglais *sentiment analysis* ou *opinion mining*). Ils indiquent notamment les phénomènes complexes qu'il faut savoir prendre en compte (difficultés lexicales, syntaxiques, sémantiques voire pragmatiques). Ils montrent toutefois comment l'on peut apporter une réponse positive à la question posée. Après un bref aperçu des campagnes d'évaluation organisées dans ce domaine (TREC, TAC, DEFT), les auteurs proposent un certain nombre de solutions pratiques, et notamment l'utilisation conjointe de plusieurs systèmes dans une approche par fusion. Ils étudient alors plus en profondeur comment améliorer un indice de similarité par la prise en compte de critères discriminants et comment choisir au mieux les dimensions utilisées pour représenter des documents sous forme de vecteurs. Diverses expériences sont décrites, notamment sur des données clients issus de France Télécom.

Cet ouvrage propose donc un tour d'horizon assez complet des différents domaines qui traitent de l'accès à l'information textuelle, et notamment la fouille de textes, la classification de documents, l'analyse de sentiments et la traduction automatique. Étant consacré aux approches statistiques, il n'est pas surprenant que les solutions discutées dans les différents chapitres relèvent toutes de ce paradigme. Elles en couvrent les principaux aspects, ce qui fait de l'ouvrage un moyen de se familiariser avec les techniques statistiques les plus utilisées pour le traitement de la langue, tout en couvrant de façon exhaustive leurs applications pour l'accès à l'information textuelle. Bien qu'il serait un peu en marge de la thématique centrale de l'ouvrage, on pourrait regretter, à côté du chapitre sur la traduction automatique, l'absence d'un chapitre sur les méthodes statistiques pour l'analyse syntaxique et sémantique de textes (analyse syntaxique, désambiguïsation lexicale, construction de représentations sémantiques). Le lecteur aurait peut-être bénéficié de discussions sur l'intérêt respectif des approches statistiques et des autres types d'approches, ainsi que des limitations des approches purement statistiques, à l'image de ce qu'ont proposé les auteurs du chapitre 6, et, dans une moindre mesure, du chapitre 7. Ceci étant dit, l'ouvrage coordonné par Éric Gaussier et François Yvon est certainement destiné à devenir un incontournable : il couvre de façon claire et exhaustive un domaine et une famille d'approches résolument actuels, pour lesquels il manquait un état de l'art de référence.

Juan-Manuel TORRES-MORENO, Résumé automatique de documents, une approche statistique, *Hermès-Lavoisier*, 2011, 260 pages, ISBN 978-2-7462-3212-9.

Lu par **Aurélien BOSSARD**

Limsi-CNRS

Le livre de Juan-Manuel Torres-Moreno présente un panorama des approches statistiques du résumé automatique. D'après l'auteur, cet ouvrage s'adresse à un public large : étudiants, linguistes, informaticiens ou mathématiciens...

Ce livre est organisé en trois parties. La première partie introduit le résumé automatique et présente des généralités sur les méthodes utilisées ainsi que les approches d'évaluation. La deuxième partie rentre dans les détails des méthodes du résumé automatique mono et multidocument tandis que la troisième partie est consacrée aux systèmes émergents : résumé multilingue et résumé de documents spécialisés. Des annexes concluent l'ouvrage, dont une destinée aux néophyte du TAL, leur sera utile, car elle explique des techniques largement utilisées dans le domaine et nécessaires à la compréhension de la majorité des systèmes de résumés exposés tout au long de l'ouvrage

Un premier chapitre introductif pose la problématique du résumé. En particulier, sont abordés les besoins en résumé et les définitions du résumé. Un bref historique des recherches sur le résumé automatique est également dressé.

Le chapitre 2 présente les trois grandes familles de résumé automatique : par génération, par extraction et par compression. Les principaux algorithmes de chacune de ces méthodes et les principales références y sont cités.

Le chapitre 3 est consacré à l'évaluation des résumés : les différentes familles d'évaluations y sont évoquées ; évaluation manuelle avec et sans référence, évaluation semi-automatique avec référence, et évaluation automatique sans référence. Sont également présentées les principales campagnes d'évaluation et leurs tâches : TSC, DUC et TAC.

Le chapitre 4 est consacré au résumé automatique monodocument. Les principales approches de l'état de l'art, depuis Luhn (1958) jusqu'à LexRank (Erkan et Radev, 2004) et PageRank (Mihalcea, 2005) y sont abordées. On peut d'ailleurs se questionner sur la catégorisation de l'algorithme LexRank dans les méthodes monodocuments, alors même que les auteurs l'ont développé spécifiquement pour le résumé multidocument. Ce chapitre présente également des approches développées par l'auteur :

– CORTEX, un système de résumé automatique fondé sur une combinaison de mesures vectorielles et l'utilisation d'un algorithme de décision afin de pondérer les phrases et sélectionner les meilleures phrases à extraire dans le résumé ;

– ENERTEX, un système de résumé automatique fondé sur l'énergie textuelle des phrases.

Le chapitre se poursuit par la présentation de méthodes de résumé fondées sur l'analyse rhétorique, l'analyse de chaînes lexicales et de méthodes de compression de phrases. Le chapitre conclut sur la nécessité de coupler les paradigmes statistiques de résumé automatique par extraction avec des paradigmes de compression et de reformulation afin de franchir un palier dans les performances des systèmes de résumé automatique.

Le chapitre 5 s'intéresse au résumé automatique multidocument guidé par une thématique. Ce type de résumé est essentiel pour la veille d'information et a fait l'objet de plusieurs campagnes d'évaluation DUC et TAC. C'est sous l'optique de ces campagnes que sont présentées les problématiques et les protocoles d'étude de ce chapitre. Les méthodes de résumé automatique de l'état de l'art, MEAD et MMR, y sont présentées, ainsi que les méthodes CATS et SUMUM développées à l'Université de Montréal et NEO-CORTEX, une version multidocument de l'algorithme CORTEX. Le chapitre se concentre également sur le résumé de mise à jour, qui mêle détection de la nouveauté et résumé automatique. Le chapitre conclut par la complexité relative du résumé multidocument par rapport au résumé monodocument, et la difficulté de la tâche de résumé de mise à jour.

Les deux derniers chapitres sont consacrés au résumé *multi* et *cross-lingue*, qui mêle résumé et traduction automatiques, et au résumé de documents spécialisés. Le résumé automatique s'est longtemps focalisé sur le discours général, et l'auteur propose d'appliquer le résumé automatique aux documents de chimie organique. Les deux chapitres présentent les enjeux linguistiques de telles tâches et concluent sur l'importance de l'hybridation de systèmes.

Globalement, le panorama dressé dans cet ouvrage est intéressant. Les personnes qui souhaitent se renseigner sur le résumé automatique y trouveront les principales méthodes de résumé, expliquées clairement. La didactique de l'ouvrage est intéressante, puisque l'auteur a pensé à une annexe qui présente les notions essentielles à la compréhension du contenu du livre par des néophytes du TAL.

Cependant, plusieurs reproches peuvent être faits à l'ouvrage. Premièrement, la qualité de rédaction n'est souvent pas à la hauteur du contenu. Nombre de formulations sont très approximatives, et mieux vaut ne pas s'attarder sur l'avant-propos, clairement en deçà du reste du livre. On peut également regretter que l'auteur ne prenne pas plus position sur les différentes méthodes, et n'en délimite pas plus les champs d'applications ou les limitations. Ainsi, les questions que les lecteurs se poseront à la lecture des résultats de l'évaluation entièrement

automatique FRESA, qui donne de meilleurs scores à des résumés automatiques qu'à des résumés de référence écrits par des humains, resteront sans réponse.

En résumé, cet ouvrage est à recommander pour des étudiants ou des chercheurs qui souhaitent s'initier au résumé automatique, sans chercher un regard critique sur la recherche dans ce domaine.