Notes de lecture

Rubrique préparée par Denis Maurel

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

Geoffrey BARBIER, Zhuo FENG, Pritam GUNDECHA, Huan LIU. Provenance Data in Social Media. ***Morgan & Claypool publishers*****. 2013. 72 pages. ISBN 978-1-6084-5783-0.**

Lu par Atefeh FARZINDAR

NLP Technologies inc., Montréal, Canada

Les médias sociaux ont éliminé les barrières de la communication, nous permettant désormais d'échanger sans égard aux axes spatio-temporels. L'information issue des médias sociaux, mise à la disposition du public pratiquement sans frais, pose un nouveau défi aux utilisateurs qui se sentent préoccupés par la fiabilité de cette information. Par exemple, il peut être difficile de comprendre la motivation derrière la déclaration d'un utilisateur à l'endroit d'un autre sans d'abord connaître la personne à l'origine du message. En outre, de fausses informations peuvent être propagées à travers les réseaux sociaux, causant inévitablement des dommages. Les données de provenance relatives à une déclaration publiée dans les médias sociaux peuvent contribuer à mettre fin aux rumeurs, à clarifier les opinions et à confirmer des faits. Toutefois, les données de provenance associées à ces types de déclarations ne sont pas facilement accessibles pour les utilisateurs actuels. En fait, ces données ne peuvent être fournies à un utilisateur que si un changement est apporté à l'infrastructure des médias sociaux ou si ce dernier a recours à un service d'abonnement. La recherche de données de provenance révèle un espace de problèmes intéressant nécessitant l'élaboration et la mise en application de nouvelles mesures afin d'offrir aux utilisateurs de médias sociaux des données de provenances pertinentes.

Les données de provenance dans les médias sociaux

Au cours des dernières années, tant l'intérêt du public envers les médias sociaux que le rôle qu'ils jouent dans la société moderne se sont accrus. La possibilité pour les utilisateurs de créer et de partager du contenu au moyen d'un ensemble varié de plates-formes comme les blogues, les microblogues, les wikis, les sites de partage et de collaboration multimédia, ainsi que les réseaux sociaux se trouve au cœur de cet intérêt. Les médias sociaux ont été utilisés pour recueillir de l'information concernant des événements de grande envergure tels que des incendies, tremblements de terre et autres catastrophes ayant un impact sur les instances gouvernementales ou autres organisations. Aussi, les particuliers utilisent les médias sociaux pour trouver des renseignements fiables sur ce qui se passe autour d'eux, leur permettant ainsi d'en tirer parti dans un court délai.

Habituellement, la provenance d'un objet offre des renseignements quant à son propriétaire, à sa source ou à son origine. Dans les médias sociaux, les renseignements relatifs à la provenance permettent aux utilisateurs d'accorder une certaine valeur à l'information reçue. Les données des médias sociaux (les attributs, les liens, le contenu) peuvent être utilisées pour déterminer la provenance de l'information. Le fait de connaître la provenance d'une information publiée dans les médias sociaux – soit de savoir dans quelle mesure elle a été modifiée, comment elle a été propagée dans ces médias et le rôle qu'a joué son propriétaire dans sa transmission – offre un contexte supplémentaire à cette information. Les utilisateurs peuvent alors se servir de ce contexte pour déterminer la validité ainsi que la valeur et la confiance à accorder à l'information reçue.

Ce livre se concentre sur le problème de la fiabilité de l'information reçue par un utilisateur à travers les réseaux sociaux et présente des pistes de solutions, notamment à l'aide de la provenance de l'information.

Le chapitre 2 présente une analyse des attributs de provenance. Les attributs de provenance d'un utilisateur peuvent comprendre son nom, son emplacement, son sexe, sa profession, ses affiliations politiques et religieuses, le contenu de l'information qu'il génère ainsi que la liste des destinataires potentiels qui pourraient avoir participé à sa retransmission (par exemple, des renseignements ajoutés dans des *retweets*, sur Twitter). Ces attributs peuvent s'avérer essentiels à la tâche d'identification de la provenance de l'information. Entre autres, ils contribuent à limiter les sources potentielles et à donner plus de crédibilité à une information.

Le chapitre 3 décrit la recherche de la provenance à l'aide des renseignements fournis par le réseau comme tels. Deux approches ont été présentées afin de conduire cette recherche. La première consiste en l'utilisation des renseignements disponibles afin d'effectuer une recherche directe sur la provenance. On aura recours à cette approche lorsque tous les destinataires sont connus pour un élément d'information. Avec la seconde approche, on cherche à découvrir le flux de propagation de l'information à partir des sources des destinataires connus, en se rapprochant le plus possible de la source d'origine ; ensuite, en se fondant sur le flux de propagation, il suffit d'identifier cette source. Cette approche peut être utilisée même si seuls quelques destinataires sont connus.

Le chapitre 4 présente la recherche de données de provenance. Afin de déterminer la source, l'approche qui s'appuie sur les attributs de provenance utilise seulement les renseignements tirés du contenu, alors que l'approche fondée sur le parcours de l'information n'utilise que les renseignements offerts par la structure même du réseau. Cependant, le défi est ouvert pour savoir comment il serait possible d'appliquer des attributs de provenance afin de conduire une recherche de la provenance plus précise. Le chapitre 4 décrit un *framework* (cadre de référence) afin d'utiliser les renseignements portant sur le réseau, déjà existants en plus des attributs et de l'historique de propagation dans le but d'obtenir la provenance de l'information. Selon lui, ce *framework*, fondé sur une méthode itérative, vise à résoudre le problème relatif aux données de provenance en utilisant alternativement le parcours de l'information, les attributs de provenance et l'historique de la propagation.

Commentaire

Au final, l'ouvrage de G. Barbier *et al.* étudie une question importante, celle de comprendre les données de provenance dans les médias sociaux et les graphiques. Ce livre est publié par Morgan & Claypool Publishers, dans la série « *Synthesis lectures on data mining and knowledge doscovery* » (Synthèse des conférences sur le forage de données et la découverte de connaissances). À l'origine, le terme de provenance est utilisé dans le domaine de l'art afin de découvrir l'origine ou la source d'une œuvre d'art, ce qui peut alors augmenter sa valeur. Ceci permet de tirer des conclusions quant à la sincérité et l'exactitude du témoignage. Dans les médias sociaux, le fait de connaître la provenance d'un élément d'information (comme un gazouillis qui a été modifié pour ensuite être propagé) fournit un contexte supplémentaire.

En annexe, on nous présente un outil en ligne[1] pour la collecte de données de provenance appelé le « *Provenance Data Collector* ». Il mise sur la récupération des valeurs des attributs de provenance d'un utilisateur donné de Twitter. Cet outil dispose d'une interface utilisateur qui permet la récupération rapide du nombre d'attributs de provenance souhaité. Cependant, il présente certaines limites ; entre autres, dans le cas des valeurs des attributs du président Barack Obama (@barackobama), l'affiliation politique n'est pas détectée, pas plus que ne l'est le sexe dans le cas du chanteur Justin Bieber (@justinbieber) !

Gil FRANCOPOULO. LMF Lexical Markup Framework. ISTE Ltd/John Wiley & sons, Inc. 2013. 267 pages. ISBN 978-1-8482-1430-9.

Lu par **Antonio BALVET**

Université Lille 3, UMR STL

La spécification LMF, issue d'une initiative conjointe entre le groupe ISO TC37/SC34 et notamment le projet européen e-content LIRICS (www.lirics.loria.fr), fait l'objet d'une norme ISO 24613 : 2008 depuis novembre 2008. Cette norme, qui a mobilisé une communauté d'une soixantaine d'experts, s'est nourrie des échanges scientifiques et des réalisations issues de plusieurs grands projets de recherche : GENELEX, EDR, EAGLES, MULTEXT, PAROLE, SIMPLE et ISLE, pour n'en citer que quelques-uns. Le méta-modèle LMF, qui s'appuie sur d'autres standards tels que XML, Unicode et UML, notamment, a été proposé afin d'unifier les différents schémas de structuration existants pour l'édition de ressources lexicales électroniques, tout en garantissant leur interopérabilité, leur pérennité, la dimension multilingue, l'expressivité et la rigueur nécessaires à la réalisation de ressources linguistiques pour le TAL. LMF a également été pensé pour permettre l'édition collaborative, asynchrone, modulaire de ressources linguistiques électroniques impliquant des coûts de

[1] http ://blogtrackers.fulton.asu.edu/Prov_Attr/

développement élevés. Les ressources au format LMF sont ainsi conçues comme autant de modules interopérables et intégrables, permettant d'aller au-delà des lexiques « statiques » : LMF a été conçu de manière à s'interfacer de façon directe avec des analyseurs morphologiques, syntaxiques ou sémantiques automatiques ainsi que des ressources du Web sémantique. Pour toutes ces raisons, le méta-modèle de ressources lexicales LMF pourrait à terme être adopté par l'ensemble de la communauté des éditeurs de ressources linguistiques, avec toutes les retombées positives qui découleraient de l'adoption d'un tel standard (services Web, API lexicales...).

L'ouvrage expose à la fois le contexte technique et scientifique qui a donné naissance au standard LMF, ainsi que des exemples d'applications concrètes de cette norme. Les trois premiers chapitres présentent les aspects les plus fondamentaux du standard LMF et de ses interactions avec d'autres standards, tel que ISOCat Data Registry[2]. Les chapitres suivants sont consacrés à l'application de la norme LMF pour l'édition, tant de ressources linguistiques monolingues, dans plusieurs langues différentes, dont des langues encore peu dotées, que pour des ressources multilingues. Ces présentations sont l'occasion d'exposer les choix linguistiques et techniques nécessaires à l'implémentation du modèle et à la réalisation des différentes ressources, qui vont de lexiques sémantiques (WordNet multilingue, UBY) à des lexiques d'entités nommées (ProLMF, Global Atlas) en passant par des lexiques génériques pour des langues non européennes (arabes, africaines et asiatiques) ou encore des lexiques ciblant des unités lexicales spécifiques (unités polylexicales en néerlandais). Les aspects techniques liés à l'élaboration de ressources conformes au modèle LMF, dans leurs aspects les plus concrets (adaptation de la spécification, édition collaborative, conversion de formats, fusion de données linguistiques existantes vers le format LMF et sérialisation de ressources lexicales LMF) sont surtout abordés dans les derniers chapitres, qui présentent des retours sur expériences pratiques de l'implémentation des spécifications LMF dans différents cadres applicatifs.

À nos yeux, l'apport principal de l'ouvrage réside dans les chapitres dédiés aux différentes applications du modèle LMF, ainsi qu'aux retours d'expériences concrets. En effet, la spécification elle-même, soit les trois premiers chapitres, est déjà exposée de façon détaillée sur www.lexicalmarkupframework.org, où le lecteur trouvera également des exemples d'entrées en plusieurs langues, ainsi que des publications, et des projets de recherche associés à la norme LMF[3]. Il ressort

[2] Qui vise à normaliser la désignation des propriétés linguistiques des entrées lexicales en fixant une terminologie lexicographique associée à des identifieurs uniques et pérennes sémantiquement « neutres ».

[3] Voir notamment les ressources mises à disposition par le projet UBY, présenté dans l'ouvrage (chapitre 10), pour l'édition de ressources lexicales sémantiques multilingues (DTD, schéma de base de donnée SQL). La plate-forme COLDIC (N. Bel, S. Espeja, M. Marimon, M. Villegas, Univ. Pompeu Fabra), de son côté, est conçue comme un service Web pour l'édition de lexiques LMF. Elle n'est malheureusement disponible qu'en version alpha à l'heure actuelle (www.coldic.sourceforge.net).

clairement des différentes contributions que, malgré les investissements consentis[4], la spécification LMF est loin d'être totalement fixée. Certains aspects de la norme restent encore à définir (ex. : sérialisation XML *vs.* RDF de lexiques LMF). D'autre part, bien qu'au niveau conceptuel, l'interopérabilité entre LMF et ISOcat Data Category Registry soit garantie, sur le plan pratique, les quelques lexiques existants conformes à LMF ne peuvent pas, à l'heure actuelle, se fonder pleinement sur le répertoire ISOcat, ce qui laisse présager des opérations de révision, de refonte et de reformatage de ces ressources dans un avenir proche.

La lecture de l'ouvrage donne le sentiment d'un travail de conception très abouti, d'une recherche permanente de l'adhésion de la communauté, passant par une certaine neutralité théorique : les lexiques LMF sont réputés compatibles avec les principales approches formelles telles que HPSG, LFG, CCG, etc. Toutefois, malgré son potentiel et sa popularité grandissante dans la communauté du TAL, il apparaît également que l'avenir de la norme LMF dépende entièrement de sa diffusion et de l'adhésion des développeurs de ressources lexicales. Au-delà de la communauté TAL « dure », on peut également s'interroger sur l'adoption de LMF par les acteurs majeurs de l'édition lexicographique « traditionnelle », y compris pour ceux ayant adopté XML comme format maître de structuration et de représentation des informations lexicographiques.

Souhaitons à LMF le même succès que d'autres spécifications contestées à leurs débuts, parfois encore qualifiées de « lourdes » et de « verbeuses » par certains développeurs, et aujourd'hui totalement incontournables : XML, le langage Java ou encore OWL et RDF.

Vivi NASTASE, Preslav NAKOV, Diarmuid Ó SÉAGHDHA, Stan SZPAKOWICZ. Semantic Relations Between Nominals. *Morgan & Claypool publishers*. 2013. 107 pages. ISBN 978-1-6084-5979-7.

Lu par **Thierry HAMON**

Université Paris-Nord – Laboratoire d'informatique médicale et bio-informatique

---

Ce livre a pour objectif de faire une synthèse des travaux dédiés à l'acquisition de relations sémantiques entre groupes nominaux au sein de la phrase, et marquées linguistiquement dans les textes. Des approches s'intéressant à l'identification de relations entre composants de groupes nominaux sont également mentionnées. L'ouvrage, relativement court, est composé de cinq chapitres. Il propose de nombreux exemples et une bibliographie riche (utile pour approfondir le sujet). Les auteurs ne se contentent pas de passer en revue les approches d'extraction de relations et les ressources recensant ces relations, mais guident également le lecteur en apportant des conseils sur le contexte dans lequel l'usage de ces approches est pertinent. On peut cependant regretter que les travaux cités portent exclusivement sur des

[4] Notamment par l'Union européenne, ainsi qu'en termes d'expertise scientifique de la part des membres des comités ISO responsables de la norme.

données en anglais et que la portabilité des approches sur d'autres langues ne soit jamais discutée.

Le premier chapitre délimite les objectifs de l'étude et définit les différents concepts qui seront évoqués par la suite. Ainsi, les auteurs se concentrent sur les relations sémantiques entre groupes nominaux à l'intérieur d'une phrase. Les travaux de reconnaissance ou de désambiguïsation de ces entités en relation sont écartés de cette synthèse. Ce chapitre introductif se termine par la description de plusieurs ressources et corpus issus de campagnes d'évaluation.

Le chapitre 2 commence par un rappel historique de l'intérêt que présentent les relations sémantiques pour représenter des connaissances ou pour accéder aux informations contenues dans les textes (depuis les philosophes de l'Antiquité jusqu'aux derniers travaux en intelligence artificielle). La présentation se poursuit par la tentative d'un inventaire des différents types de relations sémantiques étudiés dans la littérature, mais aussi tels qu'ils sont définis dans les ressources ontologiques et les bases de connaissances. Ce panorama montre en particulier qu'il existe une grande variété dans les relations sémantiques et qu'il est illusoire de vouloir les répertorier de manière exhaustive. L'ensemble de ces relations peuvent cependant partager certaines caractéristiques et être ainsi organisées selon qu'elles sont ontologiques ou idiosyncratiques, selon leur niveau de généralité, leur spécificité ou non à un domaine, etc.

Le chapitre 3 est consacré à l'extraction de relations à l'aide d'approches par apprentissage supervisé. Ces approches sont typiquement dédiées à l'analyse de collections de textes relativement peu volumineuses étant donné la nécessité de disposer de données annotées et de tester différentes configurations en fonction des relations visées. Une première section de ce chapitre est consacrée au passage en revue des jeux de données disponibles (issus de campagnes d'évaluation comme MUC/ACE ou SEMEVAL, ou proposés par des ressources constituées manuellement). Dans une deuxième section, les auteurs s'intéressent à la définition des caractéristiques (issues du texte ou de ressources externes) décrivant les relations ou les entités en relation afin de spécifier les modèles d'apprentissage automatique ou d'extraction. La troisième section de ce chapitre est consacrée à la présentation des méthodes d'apprentissage exploitées. Dans cette section, les auteurs ne font aucun présupposé sur les connaissances du lecteur concernant la maîtrise ou la familiarité avec les méthodes d'apprentissage automatique, mais rappellent tout d'abord les principes de l'apprentissage supervisé avant de décrire les algorithmes d'apprentissage généralement mis en œuvre pour identifier des relations dans les textes (classification des relations avec des méthodes à noyau, étiquetage de séquences à l'aide de modèles graphiques). Le chapitre se conclut par un certain nombre de conseils en vue d'aider le lecteur à choisir la meilleure approche et à modéliser au mieux les exemples.

Le chapitre 4 s'intéresse à l'acquisition de relations à l'aide de méthodes peu ou pas supervisées. Ces méthodes visent plutôt à traiter de gros ou très gros volumes de données, comme des collections de pages Web ou d'articles scientifiques. Il s'agit alors d'extraire des connaissances relationnelles taxinomiques ou ontologiques, ou des connaissances associées à des événements. Cet aspect est tout d'abord illustré à l'aide de travaux visant à fouiller des dictionnaires pour en extraire des relations organisées hiérarchiquement. Les auteurs s'intéressent ensuite principalement à l'acquisition de relations à l'aide de patrons lexico-syntaxiques et à l'amorçage de ce type d'approche grâce à des ressources existantes. La présentation de deux projets (*Machine Reading* et *Never-Ending Language Learner*) explorant le Web pour acquérir « en continu » des relations conclut ce chapitre.

Le dernier chapitre est une très courte conclusion ouvrant des perspectives sur l'hybridation des méthodes présentées dans les deux chapitres précédents.

Sylviane CARDEY. Modelling Language. ***John Benjamins publishing company*****. 2013. 194 pages. ISBN 978-9-0272-4996-8.**

Lu par **Nadège LECHEVREL**

Laboratoire d'histoire des théories linguistiques (CNRS, université Paris 7)

Modelling Language *présente les aspects méthodologiques et les applications d'un modèle linguistique et logico-mathématique pour le traitement et la production automatique des langues. Il s'agit d'un ouvrage fidèle à la ligne éditoriale de la collection « Natural Language Processing » des éditions John Benjamins dont l'objectif est de publier des travaux concernant l'usage de technologies informatiques innovantes dans le domaine du traitement automatique des langues.*

Structure et contenu de l'ouvrage

L'ouvrage se divise en trois parties, chacune introduite par un texte qui en présente les grandes lignes. Les trois parties sont encadrées par une préface et un prologue, une introduction et une conclusion, un épilogue, une bibliographie concise et un index.

La première partie reprend et résume les grands débats de la linguistique théorique sur le langage et son fonctionnement afin d'expliquer les choix qui ont conduit à l'élaboration d'une approche microsystémique dans le domaine du traitement automatique des langues. L'auteur souligne tout au long de l'ouvrage que le but recherché n'est pas de travailler en vain à un modèle permettant d'aborder le système complet que forme une langue naturelle, mais de travailler à partir d'un découpage microsystémique des langues, élaboré en réponse à un besoin spécifique.

La partie centrale de l'ouvrage présente le modèle linguistique et son corollaire logico-mathématique. Le modèle de linguistique microsystémique a été développé au Centre Tesnière vers la fin des années 1990, puis affiné dans les années 2000. Pour être traitées automatiquement de façon efficace, les langues doivent être découpées en microsystèmes définis par le linguiste. Les microsystèmes peuvent être mis en relation entre eux et permettent de travailler en intension, c'est-à-dire de décrire et de nommer tous les éléments jugés pertinents pour une situation donnée nécessitant une opération automatique. Cela n'exclut pas que certains ensembles soient définis en extension, c'est-à-dire sous forme d'énumération (par exemple autour du système Labelgram). Une présentation synthétique des notions clés de la théorie des ensembles, que l'on trouve à la base du modèle mathématique, ainsi que de nombreux exemples et schémas (auxquels on pourrait peut-être ajouter quelques schémas de Venn) facilitent grandement la compréhension de la modélisation proposée.

La partie suivante présente en détail les différentes applications du modèle dans les domaines de la santé, de la sécurité et de l'industrie, et la façon dont il a servi à construire des correcteurs automatiques (orthographe, grammaire), des étiqueteurs morphosyntaxiques (avec approche de désambiguïsation, notamment intralinguistique), des systèmes de *sense mining*, des dictionnaires multilingues et des langues contrôlées.

On regrette du point de vue de la structure de l'ouvrage une trop grande hiérarchisation des sections ou des chapitres trop « courts » qui en rendent la lecture parfois difficile.

Commentaire

De façon convaincante, l'auteur montre qu'un modèle microsystémique permet une mise en algorithmes efficace pour traiter des langues de façon automatique. Ce modèle, conçu exclusivement sur objectifs, vise à répondre à des problèmes spécifiques dans tous les domaines nécessitant une qualité de l'analyse linguistique et une performance des outils. De nombreux exemples issus de l'expérience de l'auteur ou de la littérature mettent à la portée du lecteur certains des problèmes rencontrés dans le traitement automatique des langues – tel le thaï, dont l'écrit ne contient pas d'espace entre les mots et dont la longueur des phrases augmente la difficulté de la reconnaissance des graphèmes et des tonèmes – et les solutions apportées par l'analyse en microsystèmes. C'est le dernier chapitre de l'ouvrage, consacré à la langue orale, qui illustre probablement le mieux les difficultés à traiter des langues naturelles et les aléas d'une approche fondée sur la norme.

Ce sont sur les enseignements de la première partie, de nature plutôt épistémologique, que nous terminerons cette note de lecture. Il ressort de la présentation des grands débats théoriques linguistiques (sur une large période historique allant de l'Antiquité aux théories les plus marquantes du XX^e^ siècle), une problématique qui n'est pas étrangère à l'historien des théories linguistiques, celle de leur « faible cumulativité ». On proposerait simplement une approche de la problématique différente qui soulignerait les évolutions conjointes des paradigmes structuralistes et générativistes en lien avec le domaine du traitement automatique des langues, en ne traitant par exemple que des aspects historiques concernant la norme, le mot et la tension entre modèle formel et modèle linguistique, aspects qui entreraient davantage en résonance avec le reste de l'ouvrage.

Anders SØGAARD. Semi-Supervised Learning and Domain Adaptation in Natural Language Processing. ***Morgan & Claypool publishers*****. 2013. 93 pages.**
ISBN 978-1-6084-5985-8.

Lu par **Fabrice LEFÈVRE**

Université d'Avignon, LIA-CERI

Ce livre introduit les concepts de base permettant l'application des techniques de l'apprentissage automatique au traitement de la langue écrite. L'objectif est de montrer comment les techniques d'apprentissage vont permettre de passer du cas simple de l'apprentissage supervisé (où des données sont disponibles pour représenter le problème traité) au cas plus complexe des apprentissages semi-supervisés et non supervisés afin de permettre d'utiliser des données incomplètes (par exemple non étiquetées). L'objectif est ainsi d'apporter une réponse aux problèmes de manque de données linguistiques dans certains cadres d'application, mais aussi et surtout de permettre l'adaptation des systèmes développés à de nouveaux domaines, ou même à des domaines inconnus.

L'ouvrage s'inscrit dans une collection originale débutée en 2008 et qui compte maintenant vingt et une monographies. La série *Synthesis Lecture on Human Language Technologies* consiste en de courts livres de cinquante à cent cinquante pages traitant de tous les thèmes relatifs au traitement de la langue naturelle, à l'informatique linguistique, à la recherche d'information ou encore au dialogue homme-machine et à la compréhension de la parole. Ces textes ont la caractéristique de faire une présentation assez courte et pédagogique de leur matière. Il ne s'agit donc pas d'être exhaustif sur les sujets abordés mais plutôt d'accompagner un lecteur novice vers un domaine nouveau en étant le plus concret et synthétique possible.

Ainsi, Anders Søgaard, de l'université de Copenhague, s'attache à nous transmettre les notions de base de l'apprentissage semi-supervisé et de l'adaptation entre domaines pour le traitement automatique de la langue. Comme le signale l'auteur dès le résumé, il ne s'inquiète pas tant des garanties théoriques (« *cet algorithme n'est jamais complètement mauvais* ») que de fournir d'utiles règles intuitives (« *dans cette situation l'algorithme peut fonctionner très bien* »). C'est donc appuyé sur quelques exemples pratiques d'application du TAL (classification thématique de documents, étiquetage en parties du discours ou encore analyse en dépendances) que l'auteur va balayer l'application des techniques de l'apprentissage automatique au TAL et plus particulièrement nous guider progressivement vers la problématique de l'apprentissage sous condition de biais. En effet, Søgaard y revient avec constance, les tâches de TAL réelles sont soumises à la difficulté d'appliquer des modèles (tests) sur des données différentes de celles disponibles pour leur mise au point (apprentissage). Il est donc nécessaire de développer des stratégies permettant de prendre en compte le mieux possible cette caractéristique, plutôt que de se limiter à assurer les meilleures performances possibles sur un domaine connu et maîtrisé uniquement. Dans cette optique, l'ouvrage parcourt les techniques de prédiction supervisées et non supervisées, avant d'aborder l'apprentissage semi-supervisé permettant d'aboutir à des techniques d'apprentissage sous condition de biais.

L'ouvrage est divisé en six chapitres. Le premier chapitre constitue une brève introduction (sept pages) qui situe le contexte général du TAL, et notamment la représentation vectorielle, et quelques applications qui serviront d'illustrations tout au long du document, avant de justifier la problématique de l'adaptation au domaine. La présentation générale du TAL se poursuit dans le chapitre 2 qui consacre une trentaine de pages à passer en revue les techniques de base des apprentissages supervisés pour le TAL. Ainsi quelques méthodes de classification emblématiques sont décrites pour la partie supervisée (k plus proches voisins, bayésien naïf, perceptron), avec leur version utilisant des pondérations d'observation. Puis, le principe des algorithmes de regroupement automatique (*clustering*) sert de support à la description des approches non supervisées, illustrées par l'algorithme des k moyennes. Cette présentation se conclut par la mise en pratique des approches sur les tâches d'étiquetage en parties du discours (POS) et d'analyse en dépendances. Clairement non exhaustif et parcellé de légères imprécisions, ce chapitre permet malgré tout une bonne entrée en matière (que les experts peuvent toutefois éviter).

Le chapitre 3 (quatorze pages) permet d'entrer dans le vif du sujet. L'apprentissage supervisé, dont Søgaard a justifié en introduction qu'il serait sa réponse au problème de l'adaptation au domaine, est structuré selon trois groupes principaux : les méthodes de type *wrapper* d'approches supervisées (*self-training, co-training, tri-training*...), les combinaisons d'approches supervisées et non supervisées (*clusters-as-features*) et enfin les méthodes fondées sur la technique du plus proche voisin (*label propagation*...). Globalement, l'effort de formalisation est louable, et permet une lecture plus structurée des méthodes et de leurs rapports. La classification proposée est toutefois hasardeuse sur de nombreux points. Ainsi plutôt qu'une technique de *wrapping* parmi les autres il semble que l'algorithme *Expectation-Maximisation* (algorithme EM) fournit un cadre général à tout algorithme de ce type (*i.e.* possédant des variables « cachées » ou non connues, *unlabeled*, qu'ils convient d'évaluer avant de mettre à jour le modèle considéré). D'ailleurs, l'auteur semble pris dans cette contradiction, comme le montre le code Python proposé où le *self-training* est décrit sous la forme d'un algorithme EM. De même le rôle particulier qu'il fait jouer à l'algorithme du plus proche voisin ne semble pas non plus se justifier pleinement. Ce sont les usages que l'ont fait des informations du classifieur qui sont pertinents dans ce cas pour introduire de la supervision de manière automatique.

Le chapitre suivant met en pratique les approches du précédent pour résoudre en pratique le problème de l'apprentissage sous condition de biais. Après une discussion sur les multiples justifications possibles de ce biais (échantillonnage imparfait, distributions non accordées ou encore problèmes disjoints) et leurs effets sur les distributions statistiques des données, une approche par transfert est envisagée et étudiée selon trois angles : le transfert des données, le transfert des caractéristiques ou le transfert des paramètres. Quelques expériences sur la catégorisation de documents viennent éclairer les techniques présentées. Mais il n'est toutefois pas possible de conclure réellement sur l'intérêt comparé des unes et des autres, le lecteur étant renvoyé à la lecture de quelques articles supplémentaires pour aller plus loin dans cette analyse.

Enfin le chapitre 5 s'intéresse à l'adaptation à un domaine inconnu. Malheureusement ce qui devrait être la partie la plus intéressante du livre n'occupe que sept pages (dont une et demie de code !). Aussi, les notions d'apprentissage adverse (*adversarial*) ou de méta-apprentissage sont clairement survolées alors que par l'intérêt qu'elles revêtent quant à leur potentiel pour certaines applications du TAL elles auraient du être discutées très largement. Cette déception ne sera pas amoindrie par la suite. L'ajout d'un chapitre 6 portant sur l'évaluation nous laisse sur notre faim, tant on peine à comprendre le niveau où il se place. Au final, il constitue surtout un plaidoyer pour une évaluation globale des méthodes de TAL au sein des applications où elles sont utilisées plutôt que de manière séparée. On regrettera vraiment l'absence d'une conclusion générale qui aurait pu utilement remettre l'ensemble des techniques présentées en perspective.

Tout au long des chapitres, la présentation des techniques est ponctuée par la description des algorithmes en Python (en lien avec la librairie Scikit-Learn). L'idée est tout à fait pertinente, au lieu d'une description formelle de l'algorithme, on accède ainsi directement à une version utilisable (par exemple dans le cadre de travaux pratiques). Des exemples pratiques viennent aussi illustrer parfois les explications. Toutefois, alors que les chapitres initiaux font une large part à la description de l'impact des méthodes sur les applications envisagées, la problématique de l'apprentissage avec biais dans son ensemble, pourtant centrale à l'ouvrage, est elle plus pauvre en résultats pratiques. Un recours plus systématique à ces exemples aurait certainement constitué une réponse acceptable au problème déjà soulevé de la difficulté de conclure sur les bénéfices réels des techniques exposées.

En conclusion, cet ouvrage tient bien les engagements de la série dans laquelle il s'inscrit, version textuelle d'une *synthesis lecture*, il n'a pas la prétention d'être un ouvrage de référence sur le sujet. À ce titre, il se révèle facile et agréable à lire. Sa lecture sera sans aucun doute très profitable aux personnes découvrant le sujet (étudiants, doctorants...) et au minimum rafraîchissante et non dénuée d'intérêt pour les spécialistes du TAL.