Notes de lecture

Rubrique préparée par Denis Maurel

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

Michael PIOTROWSKI. Natural Language Processing for Historical Texts. *Morgan & Claypool publishers*. 2012. 145 pages. ISBN 9-781-608459469.

Lu par **Serge ROSMORDUC**

Conservatoire national des arts et métiers

L'ouvrage est un panorama de l'utilisation du TAL pour les corpus de textes historiques, d'un point de vue pratique. La principale caractéristique de ces textes est la variabilité, tant en synchronie qu'en diachronie, ce qui pose des problèmes, tant pour l'acquisition des textes que pour leur exploitation. La saisie des corpus et les procédés d'analyse simple, comme l'étiquetage, forment l'essentiel de l'ouvrage. Les méthodes décrites sont essentiellement utilisables pour des textes imprimés, dans des états anciens de langues encore vivantes, ce qui constitue néanmoins une masse documentaire très conséquente.

Ce livre se veut un court état de l'art sur les problèmes liés à l'utilisation du TAL pour les textes historiques. L'auteur a abordé le domaine en travaillant à la numérisation des archives juridiques de la Suisse, et s'est donc intéressé essentiellement à des corpus modernes et contemporains.

Les deux premiers chapitres du livre expliquent en quoi le traitement des textes anciens est particulier, en insistant beaucoup sur les problèmes de standardisation du langage, et en particulier de l'orthographe. L'auteur définit ensuite son public : chercheurs en TAL d'un côté, philologues, linguistes, historiens… de l'autre. Le contexte de la plupart des travaux décrits par l'auteur est la numérisation de fonds anciens, dont les originaux sont imprimés ou manuscrits ; la plupart des traitements automatiques effectués sur ces fonds sont généralement des statistiques simples. Un certain nombre d'études plus ambitieuses, visant par exemple à extraire automatiquement des constructions verbales, ou à typer des paragraphes, ont vu le jour. Il s'agit dans ce cas de travaux de chercheurs dotés d'une double compétence. On assiste néanmoins à un changement de paradigme, avec l'institutionnalisation d'un certain nombre de projets, l'utilisation de plus en plus fréquente d'outils de TAL déjà existants, des publications et des conférences sur le sujet. L'apport principal de l'étude des corpus anciens est la modélisation des variations synchroniques et diachroniques du langage, qui peuvent s'appliquer à d'autres domaines (par exemple celui des SMS).

Le chapitre 3 est consacré aux variations de graphies, qui posent des problèmes pour nombre d'opérations : recherches, statistiques et reconnaissance optique de caractères. L'auteur décrit un certain nombre d'expériences, en particulier l'utilisation de correcteurs orthographiques pour normaliser les formes, ainsi que l'utilisation d'étiqueteurs entraînés sur les langues modernes correspondant aux textes anciens.

Le chapitre 4 traite de l'acquisition des textes. Les problèmes de reconnaissance de l'écriture manuscrite sont évoqués brièvement ; suivent des conseils pratiques pour l'acquisition, avec une discussion des formats d'image, puis des expériences avec divers logiciels d'OCR. Les résultats peuvent être améliorés en utilisant plusieurs logiciels, et en alignant les analyses. Pour améliorer la qualité d'un corpus, il est possible d'utiliser des méthodes de « *crowd sourcing* », qui permettent de le faire corriger par ses utilisateurs eux-mêmes, voire par l'internaute lambda. Le coût de la saisie manuelle des textes est ensuite étudié, ainsi que des logiciels d'aide à la transcription de manuscrits : GIDOC, T-PEN, OTTO.

Le chapitre 5 porte sur le codage du texte et les annotations. Après une présentation d'Unicode et de quelques codages *ad hoc*, l'auteur conclut que « *Unicode is therefore the only reasonable character encoding for any new project involving historical texts* ». Il présente ensuite la TEI (*Text Encoding Initiative*), en portant une attention particulière aux annotations ecdotiques (ratures, ajouts, notes marginales, etc.).

Le chapitre 6 revient en détail sur la gestion des variations orthographiques, en examinant essentiellement les états anciens de langues modernes. Le système proposé consiste à canoniser le texte dans l'orthographe actuelle, ce qui simplifie en outre les recherches pour les non-spécialistes. Quelques algorithmes de calcul de distance d'édition et de détection d'erreurs sont présentés.

Les outils de TAL, à proprement parler, comme les étiqueteurs, lemmatiseurs et analyseurs syntaxiques forment le chapitre 7. Pour l'étiquetage au sens large, l'auteur évoque longuement l'adaptation d'étiqueteurs entraînés sur une langue contemporaine vers un état ancien de cette même langue. L'analyse syntaxique, assez rarement mise en œuvre pour les langues anciennes, est évoquée rapidement.

Pour finir, le chapitre 8 donne une liste non exhaustive de corpus historiques pour l'arabe, le chinois, le néerlandais, l'anglais, le français, l'allemand, les langues nordiques, le latin, le grec et le portugais.

En cent quarante-cinq pages, l'auteur ne présente évidemment pas un état de l'art complet. Il s'agit plutôt d'une introduction, enrichie de nombreuses références. L'ouvrage aborde aussi bien des points très pratiques, comme le coût d'une saisie manuelle, qu'un certain nombre de points plus théoriques, comme la représentation des variations linguistiques.

La vision proposée du traitement des textes historiques est très largement tirée de la pratique de l'auteur lui-même, qui travaille sur des corpus imprimés relativement

récents et volumineux ; d'où l'importance accordée aux questions d'OCR. Il évoque assez peu, en revanche, le traitement des langues non indo-européennes, les écritures non alphabétiques ou les logiciels de travail collaboratif. Il n'y a rien, par exemple, sur *epidoc*, dont le développement a une importance certaine pour les études classiques. Le cadre choisi représente cependant des collections très nombreuses et volumineuses, pour lesquelles la numérisation sera quasi indispensable.

On reste un peu sur sa faim d'un point de vue théorique ; l'approche étant éminemment pratique et *ad hoc*. L'éclairage donné sur les textes historiques, s'il est partiel, n'en est pas moins intéressant et l'ouvrage propose une riche bibliographie et webographie qui seront utiles pour aborder le développement d'un corpus historique.

Inderjeet MANI. Computational Modeling of Narrative. *Morgan & Claypool publishers*. 2013. 124 pages. ISBN 9781608459810.

Lu par **Laurent PRÉVOT**

Aix-Marseille Université & CNRS, Laboratoire Parole et Langage, UMR 7309

L'objectif principal (et déclaré) du livre est principalement d'offrir aux informaticiens un « primer *» de narratologie. Mani espère cependant aussi que l'expression dans un cadre computationnel des concepts principaux de narratologie permette de raffiner et de clarifier certains de ces concepts.*

La partie « nouvelle » de cet ouvrage pour un spécialiste en intelligence artificielle ou de la représentation des connaissances est relativement restreinte (et se situe principalement dans le chapitre 4). L'ouvrage pourra être conseillé au praticien du TAL, non spécialiste en IA, ou, comme une lecture préalable, pour un étudiant (informaticien ou linguiste) qui souhaiterait aborder le sujet. Ce livre lui donnera efficacement accès à ce domaine qu'est la modélisation computationnelle des narratives. Par ailleurs, grâce à l'éclectisme des exemples et l'enthousiasme de l'auteur, l'ouvrage crée un effet de motivation pour travailler ce sujet. Il faudra cependant aller plus loin que la bibliographie proposée pour obtenir un réel état de l'art.

La présentation (en XML) du langage de description développé par l'auteur (NarrativeML) offre une sorte de fil rouge lors des conclusions de chaque chapitre. Sans réellement contribuer à une meilleure présentation des concepts, il montre concrètement l'intégration des différents aspects de la narratologie dans un format d'annotations précis (des exemples d'annotations auraient cependant été sans doute plus parlants que la seule DTD).

En résumé cet ouvrage constitue un point de départ intéressant pour quiconque souhaiterait se lancer dans l'exploration de ce sujet. Les concepts, les exemples et les discussions forment un tout assez homogène et stimulant.

Le premier chapitre peut-être décomposé en deux parties. Tout d'abord, un tour d'horizon des grands concepts utilisés en narratologie (narrateur, niveaux de narration, temps, audience) illustré par des exemples littéraires. La présentation est du niveau de la vulgarisation, mais n'est pas ennuyeuse grâce aux exemples. Le tout est homogène et permet pour le novice intéressé de se familiariser avec un domaine. La deuxième partie présente un peu plus le modèle de Mani. Le niveau de prérequis est moins clair. Par exemple, les propriétés des relations binaires (réflexivité, etc.) sont réintroduites, alors que pour un public d'informaticiens cela semble superflu. Inversement, la connaissance de PropBank, une ressource certes connue mais assez spécifique, est pratiquement présupposée alors qu'elle mériterait sans doute une introduction. Par ailleurs l'introduction de l'ontologie de base est réellement elliptique. La plupart des lecteurs venant de la représentation des connaissances auraient sans doute aimé que ce passage (une page !) ait un petit plus d'épaisseur.

Le deuxième chapitre s'attaque à la question des personnages en tant qu'agents rationnels. Il débute par une présentation de problématiques et de concepts classiques en IA : la reconnaissance et la génération de plans, les scripts. Puis le chapitre se focalise sur l'utilisation de ces notions dans un cadre narratif. Une section intéressante, mais un peu courte concerne l'intégration de l'interaction dans la narration (par exemple pour la conception de jeux). L'empan de ce chapitre est vaste et le traitement en reste donc superficiel. Il a cependant le mérite de rassembler des notions et une bibliographie de base (qui mériterait bien d'être étoffée notamment sur les aspects TAL et discours) sous un angle novateur.

Dans le troisième chapitre Mani traite de sa spécialité : le temps. L'état de l'art et le niveau des discussions, tout en restant accessibles, deviennent un peu plus pointus. Une bonne place est faite à l'algèbre de Allen qui est intensivement utilisé dans les modèles et systèmes présentés dans le chapitre. La discussion sur les difficultés liées à la richesse des représentations des structures temporelles est intéressante mais un peu rapide. Par exemple, le chapitre effleure la question des multiples manières d'exprimer une structure temporelle à partir d'un jeu de relations donné, ou encore la question de l'équivalence entre plusieurs descriptions formelles (y compris à partir d'un même jeu de relations). La couverture bibliographique n'inclut cependant pas certains travaux récents dans ce domaine. Par ailleurs, pour un ouvrage où la question de l'évaluation est importante, on peut regretter que l'auteur se contente de livrer des scores variés (F-score, taux d'accord, Kappa) sans proposer une réflexion sur la valeur relative de ces évaluations, et ce en fonction des tâches évaluées. De plus la question du passage de ces expériences d'annotations à la valeur des ressources produites pour l'apprentissage automatique (un des objectifs évoqués dans ce chapitre) mériterait d'être un peu plus discutée. C'est un chapitre qui sera plus proche des intérêts du lecteur moyen de la revue TAL. Il le rapprochera davantage de l'état de l'art du domaine (comparé aux chapitres 1 et 2), mais il laissera sur leur faim les spécialistes du traitement automatique du temps ou du discours.

C'est dans le chapitre 4, traitant de l'intrigue (*plot*), que le rapprochement entre narratologie et TAL est le plus fécond. C'est aussi dans ce chapitre que le lecteur se sent le plus proche de l'état de l'art du domaine. De nombreux systèmes automatiques, plus ou moins directement inspirés de travaux de narratologie sont décrits assez finement et comparés. Là encore, l'auteur fait un travail de défrichage d'un domaine riche pour aller à l'essentiel des notions utilisées dans les systèmes automatiques. Les linguistes seront toutefois étonnés de ne pas voir du tout apparaître dans ce chapitre le travail du sociolinguiste William Labov dont le travail sur la narration a généré tout un paradigme d'études des narratives.

Le chapitre conclusion, bien que très succinct apporte des considérations finales pertinentes et offre des ouvertures stimulantes tant en terme d'applications du domaine qu'en terme de développements scientifiques.

Patrice BELLOT. Recherche d'information contextuelle, assistée et personnalisée. *Hermès-Lavoisier*. 2011. 302 pages. ISBN 978-2-7462-2583-1.

Lu par **Marilyne LATOUR**

ReportLinker.com - Département R&D

« Montre-moi qui tu es et je te dirai comment rechercher », c'est en quelques mots le thème abordé dans cet ouvrage dirigé par Patrice Bellot. Le livre aborde en effet les principales solutions mises en œuvre dans les systèmes de recherche d'informations (SRI) actuels pour prendre en compte l'utilisateur (son profil, ses compétences) ainsi que son contexte de recherche en recherche d'informations (RI). Les objectifs étant une meilleure assistance, une adaptation personnalisée ainsi qu'une modélisation de la langue. Ces solutions s'articulent autour de trois axes : [1] la notion de contexte de RI ainsi que la robustesse des modèles informatiques et linguistiques employés sur des textes et requêtes réels, [2] la personnalisation de la RI et la place de l'utilisateur dans le processus de recherche, [3] l'assistance aux utilisateurs, l'aide à la navigation et les interfaces spécifiques pour la visualisation interactive d'informations.

Dans le chapitre 1, l'auteur s'intéresse aux contextes et aux spécificités des requêtes formulées dans un SRI pour inférer des caractéristiques sur les besoins informationnels, comme le type de requête, le type de réponse ou encore le niveau de réponse attendu. L'auteur dresse une typologie des requêtes : requêtes récurrentes – populaires ou répétées –, requêtes reformulées (avec leurs historiques) et requêtes difficiles. L'auteur propose ensuite de les classifier en fonction de leur appartenance au sein d'un même groupe (les mesures de similarité traditionnelles n'étant pas adaptées pour déterminer la ressemblance de requêtes). Diverses classifications sont présentées : classification thématique (qui s'appuie sur le contenu des requêtes), classification fondée sur les caractéristiques des requêtes (comme la taille, la clarté, la différence d'informations entre les thèmes…), classification fondée sur les performances auxquelles elles conduisent.

Le chapitre 2 aborde le problème de la robustesse en RI. Cette dernière constitue en effet une question essentielle tout en restant une notion peu précise, très dépendante du type de traitement à effectuer ainsi que du type de texte. Pour le traitement automatique des langues (TAL), la robustesse d'une application se mesure à sa capacité à résister aux erreurs. Les auteurs s'attachent à décrire les situations qui conduisent les systèmes à des erreurs (défauts de couverture, erreurs de traitement, entrées corrompues par un prétraitement, entrées contenant des phénomènes spécifiques difficiles ou impossibles à traiter) et présentent quelques techniques robustes (notamment des techniques développées au LPL) ainsi que des techniques d'étiquetage efficaces. Ils abordent également le problème de l'analyse syntaxique (superficielle symbolique et superficielle stochastique) et présentent une approche syntaxique intrinsèquement robuste : les Grammaires de Propriétés (GP).

Dans le chapitre 3, l'auteur s'intéresse à la fois aux approches d'évaluation de SRI traitant de corpus ou de requêtes bruités et aux techniques proposées dans la littérature pour intégrer le bruit au sein des modèles d'accès à l'information. L'auteur présente une étude réalisée sur le système SQuaLIA qui montre successivement des modules d'analyse de la question en type de réponse attendue, de recherche de documents et de passages susceptibles de contenir une réponse appropriée, puis d'appariement de la question avec les entités candidates correspondant au type de réponse attendue dans les passages sélectionnés. Enfin, elle présente une méthode d'évaluation applicable à des systèmes complexes et modulables.

Dans le chapitre 4 qui concerne plus spécifiquement les documents audio, les auteurs indiquent comment il est possible de retrouver la réponse précise à une question dans un enregistrement audio en utilisant des méthodes de QAst (Question Answering on Speech Transcripts). Pour les auteurs, plusieurs points sont à retenir : tout d'abord [1] l'importance de préparer les données pour que tous les documents aient une forme commune (qu'ils soient issus de transcriptions manuelles ou d'un quelconque système de reconnaissance automatique de la parole) et [2] prendre en compte le niveau phonétique pour une amélioration réelle dans la précision des réponses trouvées par le système.

Le chapitre 5 fait part d'une réflexion sur la place de l'usager dans le développement des SRI. Les auteurs font une synthèse de la réflexion sur la modélisation de l'usager dans le cadre d'une démarche de RI. Ils traitent des approches classiques de modélisation de l'usager développées en informatique mais également des applications de ces modèles et présentent également les résultats des études menées en ergonomie cognitive sur l'influence des caractéristiques de la tâche, de l'outil et de l'usager sur l'utilisation d'un SRI. Ils discutent, en outre, de la complémentarité des deux approches et des différences de points de vue. Enfin, ils dressent un bilan des limites et des enjeux de la prise en compte dans les processus de RI.

Le chapitre 6 traite de la RI collaborative ; celle-ci étant vue comme l'ensemble des approches techniques permettant ou facilitant la collaboration dans le processus de RI. D'un point de vue opérationnel, la RI collaborative peut également être définie comme l'ensemble des approches et des techniques qui exploitent les liens et les interactions entre les utilisateurs d'un SRI, que ces liens soient directs ou indirects, explicites ou implicites. L'auteur détaille notamment la RI collaborative asynchrone, où les utilisateurs ne travaillent pas forcément de

façon simultanée (les approches s'appuient alors le plus souvent sur les traces des actions passées d'utilisateurs pour mieux servir les demandes ultérieures d'autres utilisateurs) ; il traite aussi de la RI collaborative synchrone, c'est-à-dire celle dont les collaborateurs sont soit en présence soit à distance mais de manière simultanée. Pour la RI collaborative synchrone, les collaborateurs partagent en général le même besoin informationnel et cherchent explicitement à la résoudre de façon collaborative. Pour la RI collaborative asynchrone, le groupe demeure souvent implicite et le besoin informationnel ne se trouve partagé qu'*a posteriori*, lors du rapprochement qui est fait pour l'exploitation des traces passées. Pour l'auteur, on peut mettre en place un outil facilitant la mise en commun et la comparaison des requêtes, des résultats et même des différentes stratégies de recherche. Enfin, l'auteur propose un cadre d'analyse pour décrire les travaux et les avancées du domaine de la RI collaborative.

Dans le chapitre 7, l'auteur met en avant l'existence de grandes lacunes concernant l'adaptation des SRI aux utilisateurs ayant des capacités de lecture limitées (pathologies langagières, personnes ne maîtrisant pas suffisamment la langue d'un document ou encore personnes se retrouvant face à un contenu pour lequel il est nécessaire d'avoir une certaine expertise). Répondre à un tel besoin de personnalisation nécessite de définir de nouvelles mesures en prenant en compte la lisibilité d'un document et offrir, par exemple, la possibilité de présenter d'abord les documents les plus simples (*i.e.* les plus lisibles). L'auteur identifie les modèles cognitifs de la lecture et présente les principaux travaux qui ont abordé le problème de l'estimation automatique de la lisibilité d'un texte. Il propose, en outre, une manière d'exploiter concrètement la lisibilité au sein d'un système de RI et définit la (les) dyslexie(s) comme sujet d'études. Dans l'idéal, le profil serait appris automatiquement et représenterait les capacités de lecture de l'utilisateur dans un cadre générique, mais aussi en fonction des thématiques rencontrées. Selon un processus interactif, il serait possible d'associer à des requêtes, et par la suite à des thématiques, des listes de documents que l'utilisateur aurait trouvés non seulement pertinents mais aussi utilisables.

Le chapitre 8 dresse un état de l'art de la navigation et du résumé dans les documents audio. L'auteur détaille une expérimentation prouvant l'utilité du résumé de la parole. Il est en effet possible de retrouver des éléments factuels et définitoires mentionnés explicitement dans des enregistrements de paroles préparées ou spontanées. L'auteur propose un ensemble de méthodes pour faciliter l'accès aux collections de documents audio. La plupart d'entre elles reposent sur une transcription automatique du discours parlé alors que d'autres tentent de localiser l'information importante en analysant directement l'acoustique.

Le chapitre 9 concerne la visualisation interactive d'informations. L'auteur commence par analyser les besoins d'interactivité en RI en fonction des tâches envisageables. Il détaille ensuite les différents styles d'interactions classiques : interaction à facettes, filtrage dynamique, brossage (ou *brushing*), interfaces zoomables ou déformantes, et enfin interaction coopérative et dispositifs d'affichage distribués.

Enfin, le chapitre 10 présente la RI à travers les nouveaux dispositifs mobiles (notamment *via* les téléphones intelligents comme les *smartphones*) dont la particularité est l'interface limitée de la zone de saisie du texte (la taille du clavier étant réduit du fait des dimensions du dispositif utilisé). Pour l'auteur, l'ingénierie des langues peut offrir des outils à

même de compenser ces insuffisances : c'est le cas de la prédiction linguistique qui fait l'objet de ce chapitre et que l'auteur propose d'étudier sous plusieurs modèles plus ou moins évolués afin de comparer leurs performances respectives.

Cet ouvrage permet d'appréhender en quelques chapitres les enjeux actuels de la personnalisation de la prise en compte de l'utilisateur et du contexte de recherche en RI. En outre, il en ressort que les travaux futurs devront forcément être pluridisciplinaires pour que les SRI soient plus efficaces : ces systèmes devront être orientés modèles théoriques de la RI, bien sûr, mais aussi ouverts aux apports de la linguistique, de l'informatique, des sciences cognitives, de la sociologie et de la sémiologie...

Cyril GROUIN, Dominic FOREST. Expérimentations et évaluations en fouille de textes : un panorama des campagnes DEFT. *Hermès-Lavoisier*. 2012. 248 pages. ISBN 978-2-7462-3836-7.

Lu par **Laurence KISTER**

Université de Lorraine, Atilf UMR 7118

Cet ouvrage collectif propose un panorama des campagnes DEFT et constitue en quelque sorte, un état de l'art des activités en fouille de textes sur des documents francophones. Il situe rapidement les campagnes DEFT dans le contexte international des campagnes d'évaluation avant de consacrer un ou deux chapitres à chacune des sept campagnes menées entre 2006 et 2011 (la première campagne, celle de 2005, ayant fait l'objet d'un ouvrage spécifique).

Le premier chapitre rappelle les objectifs généraux de la fouille de textes (accès au contenu par extraction et organisation de l'information disponible dans les textes) et les objectifs des différents défis (thématiques des campagnes, corpus utilisés, tâches demandés). Il présente également les différentes mesures (micro-moyenne, macro-moyenne, précision, rappel, f-mesure) nécessaires à l'évaluation.

Le défi 2006 s'intéresse aux ruptures thématiques. Il porte sur différents types de textes : discours politiques (Giscard d'Estain, Mitterrand, Chirac), textes juridiques (lois de l'Union européenne) et un ouvrage scientifique *Apprentissage artificiel* (Curnuéjols et Miclet, 2002). Il fait l'objet de deux chapitres. Les auteurs de l'un des chapitres postulent un apprentissage en fonction de plusieurs points de vue (débuts/attaques de phrase d'un point de vue formel, présence de traits rhétoriques dans la phrase, thématique/sémantique développées dans chaque paragraphe) pour aboutir à une coopération des approches dans la phase finale d'apprentissage. Les techniques non supervisées utilisées permettent de valider le repérage des « thèmes rhétoriques » et de l'intensité alors que la détection des ruptures thématiques demande à être approfondie. La méthode paraît reproductible si l'on dispose de différents objets à identifier, descriptibles selon divers points de vue. Les auteurs du deuxième chapitre proposent de mettre en lumière les liens entre

la finalité de la segmentation et l'approche à utiliser. Pour cela, ils se fondent sur deux approches : la segmentation lexicale et l'utilisation de plusieurs sources d'implémentation inspirée de TextTiling [Hearst 1997] complétées par une phase d'apprentissage automatique. La segmentation automatique utilise différentes informations relatives à la distribution des mots dans le texte (cohésion lexicale, chaînes lexicales, taux d'introduction de termes nouveaux) pour proposer des frontières possibles. Il reste à inclure des informations provenant de thesaurus et du traitement des anaphores.

Le défi 2007 porte sur l'attribution automatique de valeurs d'opinion à partir de critiques de livres, de films et de spectacles (corpus « *À voir, à lire* »), de quatre mille critiques de jeux vidéo, de relectures d'articles scientifiques (*JADT*, *RFIA* et *TALN*) et d'interventions de députés à l'Assemblée nationale. Il consiste à identifier des parties pertinentes pour la classification automatique de textes. Les auteurs utilisent une méthode symbolique (analyse syntaxique avec un analyseur fonctionnel et relationnel), une méthode statistique (n-grammes) et une méthode hybride qui combine les deux méthodes précédentes. Les résultats de la méthode hybride sont plus précis que ceux des autres méthodes. De fait, la méthode hybride permet de superposer différentes couches d'annotations, donc d'informations.

Le défi 2008 consiste en une classification en genres de textes en lien avec la télévision (économiques, artistiques, sportifs) et en catégories de documents (sociétaux, actualité/informations françaises, actualité/informations internationales, scientifiques, littéraires) à partir de documents extraits du *Monde* et de Wikipédia en français. Après une évaluation des méthodes de classifications classiques les auteurs évaluent différentes méthodes numériques. Les méthodes retenues et mises en œuvre ont permis d'élaborer un système de catégorisation valide sur plusieurs langues.

Le défi 2009 s'intéresse à la fouille d'opinions à partir de documents rédigés en trois langues (français, anglais et italien). Le corpus se compose d'articles de presse (*Le Monde*, *The Financial Times*, *Il Sole 24 Ore*) et de débats parlementaires au Parlement européen. Les tâches consistent à repérer le caractère subjectif ou objectif des documents dans leur intégralité, à déterminer plus finement les passages subjectifs dans les textes, et à identifier le parti politique d'appartenance des parlementaires. Il fait l'objet de deux chapitres. L'auteur de l'un des deux chapitres propose d'optimiser la technique classique des machines à supports vectoriels. Les gains restent limités mais l'intérêt de la prise en compte d'une variable continue (longueur des articles) est mis en évidence. Il propose aussi de distinguer par une procédure d'apprentissage les catégories éditoriales plutôt que de les construire à partir des termes d'indexation. Les auteurs du deuxième chapitre adaptent des méthodes d'analyse de textes dans leur globalité puis proposent une analyse des différents passages des textes. La deuxième tâche est au centre des préoccupations de l'équipe qui travaillait, préalablement à Deft 2009, à la détection des segments phrastiques ou intraphrastiques qui expriment une évaluation pour les catégoriser en fonction de leur modalité, de leur configuration d'énonciation et de leur valeur

axiologique. L'équipe est satisfaite des résultats obtenus quant aux évaluations et à la subjectivité dans des textes sans contrainte de domaine thématique.

Le défi 2010 s'intéresse à la datation et à l'origine géographique d'articles de journaux francophones. Les documents, tous issus de Gallica, correspondent à des sujets variés et à des résumés de débats de la période 1800-1944. Ils proviennent de *La Croix, Le Figaro, le Journal des Débats* et des *Décrets* (devenu *Journal de l'Empire* puis *Journal des débats politiques et littéraires*). Ce défi fait l'objet de deux chapitres. La première équipe travaille à partir des entités nommées pour effectuer un traitement logique et des marqueurs caractéristiques d'une époque. Elle ne s'intéresse qu'à la datation des articles. La combinaison de ces traitements est associée à un apprentissage statistique pour éviter les biais et l'absence d'information. Pour cela la fréquence d'apparition des mots et leur enchaînement sont pris en compte pour estimer les décennies possibles des extraits proposés. Le système, après différents filtrages, retient la date d'apparition la plus ancienne des mots. Deux modules complémentaires sont envisagés : un module de « conversion » des préfixes désignant la fonction des personnes pour améliorer le repérage des entités nommées et un module fondé sur la datation des inventions nombreuses sur la période considérée. La seconde équipe s'intéresse aux deux tâches proposées et remarque leurs difficultés inégales de réalisation : la classification géographique s'est révélée relativement aisée (notamment en raison des variations lexicales géographiques) tandis que celle par décennies est relativement complexe. Les auteurs proposent de fusionner les différentes approches testées et de réduire de manière automatique le poids des différents paramètres en utilisant des approches probabilistes.

Le défi 2011 comporte deux tâches : identifier l'année de publication d'articles de presse et apparier des articles et leur résumé. Le corpus se compose des articles proposés en 2010 auxquels s'ajoutent des articles extraits de *La presse* et *Le Temps.* Les deux équipes qui présentent leurs travaux ont des approches différentes. La première postule la prise en compte des caractéristiques des tâches d'apprentissage et le choix judicieux d'une technique classique d'apprentissage adaptée aux tâches et à l'espace de représentation des données sans aucun recours à des connaissances externes. Elle conclut sur la difficulté de la tâche de datation (données bruitées, classification en classes continues, diversité des exemples, etc.) et l'aisance à réaliser l'appariement article/résumé. La seconde équipe met en évidence la nécessité de mettre en œuvre différentes méthodes utilisant plusieurs ressources : utilisation et combinaison d'indices chronologiques, prise en compte de l'évolution de la langue et utilisation de techniques de catégorisation de textes par apprentissage. Les résultats qu'elle obtient sont mitigés en raison de la petite quantité de données de références disponibles et de la qualité médiocre des documents numérisés.

Les défis 2010 et 2011 étaient soumis à la même contrainte : Gallica ne pouvait en aucun cas constituer une ressource puisque tous les articles de presse en étaient extraits.

Cet ouvrage dresse un panorama intéressant des différents défis (sur les genres et les thèmes, sur la fouille d'opinions et sur des aspects diachroniques) et des différentes manières de les aborder (base de connaissances d'experts et apprentissage attribué). Les équipes qui ont relevé les défis présentent les points positifs de leurs démarches avant d'exposer les éventuelles limites de celles-ci, de proposer des suggestions d'amélioration ou d'évoquer des perspectives de poursuite de travaux après les défis.

Alexander CLARK, Chris FOX, Shalom LAPPIN, The Handbook of Computational Linguistics and Natural Language Processing, *Wiley-Blackwell*, 2010, 775 pages, ISBN 978-1405155816.

Lu par **Nathalie FRIBURGER**

Université François Rabelais Tours, laboratoire d'informatique

---

Ces trente dernières années, les techniques de TAL sont passées de cas d'études, prototypes et théories à de réelles applications, notamment avec Internet. Considérant l'utilisation de plus en plus grande des technologies de traitement automatique des langues, les éditeurs de ce livre ont voulu fournir une référence dans les domaines de la linguistique computationnelle et du traitement automatique des langues.

Ce livre s'annonce comme étant un panorama complet des méthodologies, concepts et applications en linguistique computationnelle et traitement automatique des langues. Les éditeurs présentent leur livre comme *« designed as a reference and source text for graduate students and researchers from computer science, linguistics, psychology, philosophy and mathematics »*.

Ce « Handbook » de 802 pages est divisée en vingt-deux chapitres, organisés en quatre thèmes : les fondements formels, les méthodes courantes, les domaines d'application, les applications. L'idée générale étant de proposer des articles, traitant principalement des « états de l'art », dans tous les aspects de la linguistique computationnelle. Ce livre contient aussi les biographies de tous ses contributeurs, majoritairement des chercheurs du monde anglo-saxon (Londres, Cambridge, Edinburgh, Belgique, Amsterdam, Chicago, NYU, San Diego, etc.) très connus dans leurs domaines respectifs. Dans l'introduction, on trouve un résumé détaillé pour chacun des chapitres présentés dans le livre.

Cette note de lecture évoque rapidement chacun des vingt-deux chapitres et leur cohésion à l'intérieur des différentes parties de cet ouvrage.

Partie 1 – Les fondements formels

À travers quatre chapitres, cette première partie définit les langages formels, la complexité des tâches du TAL, la modélisation statistique et l'analyse syntaxique.

Le chapitre 1 présente la théorie des langages formels de manière complète, en commençant par les définitions de base (alphabets, chaînes, mots, lettres, concaténation, etc.) puis les langages réguliers, automates à états finis et les opérations de minimisation, déterminisation, etc. Ce chapitre aborde aussi les langages hors contextes, grammaires et méthodes à états finis, dérivations, hiérarchie de Chomsky, concepts très importants pour le domaine de la linguistique computationnelle. Les explications sont très claires et peuvent être accessibles mêmes pour des non-initiés (aux mathématiques). Le deuxième chapitre, plus ardu, présente la complexité en temps et espace pour plusieurs classes de langage et tâches de TAL. Il commence par faire un panorama de la théorie de la complexité et des machines de Turing, puis présente les classes de complexité les plus courantes. Ce chapitre rappelle les problèmes liés à la complexité des algorithmes dont on ne parle pas si souvent dans les articles de TAL. Le chapitre 3 concerne la modélisation statistique des langages : l'état de l'art est assez court. Ce chapitre présente principalement et longuement le modèle structuré que l'auteur du chapitre utilise lui-même dans ces travaux : SLM (*Structured Language Model*). De ce fait, ce chapitre n'est pas aussi intéressant que les autres chapitres de ce livre. Le chapitre 4 clôt la partie sur les fondements et présente l'analyse syntaxique (*parsing*) et les algorithmes de reconnaissance de la structure syntaxique d'une phrase pour des grammaires hors contexte (CYK, Earley), mais aussi des algorithmes probabilistes. Ensuite ce chapitre montre comment un algorithme de reconnaissance peut être étendu pour devenir un algorithme de *parsing*.

Cette partie du livre est très réussie et peut être très utile pour qui souhaite avoir rapidement un aperçu des fondements du traitement automatique des langues.

Partie 2 – Les méthodes courantes

Cette partie présente les méthodes d'apprentissage supervisées ou non (maximum d'entropie, fondée sur la mémoire), les arbres de décision, les réseaux de neurones puis expose l'annotation linguistique et les problèmes liés à l'évaluation.

Le chapitre 5 introduit le concept de maximum d'entropie et les différents modèles existants, puis leurs usages (délimiter les phrases, étiquetage en parties du discours, résolution d'ambiguïtés, reconnaissance d'entités nommées, etc.). Le chapitre 6 présente l'apprentissage fondé sur la mémoire (*memory-based learning*) ainsi que son application dans de nombreuses tâches du langage naturel. Ce type d'apprentissage consiste en un stockage d'exemples dans la mémoire, le traitement s'appuyant sur la similarité avec les exemples stockés. Ce chapitre décrit les applications possibles en traitement des langues naturelles : la morphophonologie, la syntaxe et la sémantique, l'analyse de textes, le dialogue et le discours, la génération et la traduction. Le chapitre 7 présente les arbres de décision, technique permettant de résoudre les problèmes de classification (tels que *POS tagging*, désambiguïsation du sens des mots, etc.). L'auteur présente la manière de créer des arbres de décision à partir de données d'entraînement, les applications en TAL, puis leurs avantages et inconvénients (limitations de cette technique et solutions permettant de dépasser ces

limites). Le chapitre 8 présente l'apprentissage non supervisé et l'induction de grammaire. Il compare l'apprentissage supervisé et non supervisé en termes de précision et de coût. Le chapitre 9 présente les réseaux de neurones et leur architecture afin de faire de l'apprentissage non supervisé en TAL (notamment pour le *parsing* et la modélisation). Les chapitres 5 à 9 présentent des méthodes d'apprentissage variées mais classiques. Le chapitre 8 aurait pu prendre la première place de cette partie car il expose et compare les méthodes supervisées et non supervisées d'apprentissage (coûts et bénéfices pour les tâches du TAL).

Le chapitre 10 traite du processus d'annotation linguistique (corpus, consistance, outils, évaluation). Il présente les éléments de base pour obtenir une annotation consistante citant, pour la structure syntaxique, le *treebanking* (différents *tree banks* mais pas le français), pour la structure sémantique, l'étiquetage du sens et du rôle sémantique, la coréférence, les annotations d'opinions. L'évaluation des systèmes de TAL est un problème évoqué par le chapitre 11. Les concepts fondamentaux de l'évaluation sont expliqués à travers les défauts et qualités de différents types d'évaluations (manuelles, automatiques, intrinsèques, par les composants, accords interannotateurs, etc.). Le chapitre présente le partitionnement des données pour l'évaluation (entraînement, développement, test), la validation croisée, la comparaison des performances des systèmes et des études de cas. Les chapitres 10 et 11 se complètent. Le chapitre 11 est très intéressant et permet d'avoir une bonne idée des problèmes liés aux campagnes d'évaluation si on n'y a pas encore participé.

Partie 3 – Domaines d'application

Cette partie, composée de six chapitres, s'intéresse aux domaines de la reconnaissance de la parole, à l'analyse syntaxique, la morphologie, la sémantique, le dialogue et la psycholinguistique, soit les différents niveaux d'analyse du langage.

Le chapitre 12 présente le domaine de la reconnaissance de la parole (modélisation acoustique, utilisation et rôle des HMMs) et les problèmes de robustesse et de passage à grande échelle. Il présente une étude de cas : AMI système. L'analyse syntaxique statistique est évoquée largement dans le chapitre 13, principalement par l'étude des grammaires probabilistes lexicalisées (PCFG, Collins). Ces modèles ayant été principalement développés pour l'anglais, l'auteur s'interroge sur l'opportunité de les appliquer sur d'autres langues. Le chapitre 14 détaille les problèmes liés à la segmentation et à la morphologie, dont celui de trouver les limites correctes de mots ainsi que la construction du lexique d'un langage en présentant de nombreuses approches (notamment celles liées aux travaux de Zellig Harris). Le chapitre 15 s'intéresse aux questions de sémantique computationnelle et à leur usage dans les phrases et les discours. Puis le chapitre 16 présente les phénomènes liés au dialogue qu'il faut modéliser. Il propose le modèle de dialogue Kos et le compare à d'autres approches. La psycholinguistique computationnelle est décrite dans le chapitre 17 à travers la manière dont l'être humain reconnaît la structure et le sens des phrases dans un discours et s'intéresse particulièrement à la résolution des ambiguïtés de structures.

Partie 4 – Applications

Les applications évoquées dans cette partie sont l'extraction d'information, la traduction, la génération, l'analyse du discours et les systèmes de questions-réponses.

Le chapitre 18 explore le sujet de l'extraction d'information telle que les noms, entités, relations et événements présents dans un texte. Les systèmes d'extraction utilisent des règles ou des méthodes d'apprentissage supervisé, semi ou non supervisé. Ce chapitre se conclut sur les perspectives de ces applications sur le Web. Le chapitre 19 présente la traduction automatique statistique selon des modèles de langage fondés sur les mots ou sur les phrases. L'auteur présente ensuite son propre outil hybride de traduction. La génération de langage naturel est décrite au chapitre 20. Ce chapitre présente de nombreux systèmes de génération et leur rôle dans le TAL. Le chapitre 21 montre ensuite l'analyse de la structure du discours notamment à travers la résolution d'anaphores, nécessaire dans bien des tâches (résumé, génération, extraction d'information). Le dernier chapitre de ce livre aborde les systèmes de questions-réponses d'abord limités à de petits ensembles de données puis créés pour répondre à des questions à l'aide de contenu Web en utilisant des méthodes d'apprentissage.

Conclusion

Donner un avis sur un tel ouvrage n'est pas chose aisée. Les informations apportées par les différents chapitres se croisent et se complètent. Ce livre confirme que les approches symboliques et modèles statistiques se mêlent à travers les approches hybrides dans tous les types de tâches présentées. Comme le disent les éditeurs, ce livre n'est pas exhaustif puisque les pistes de recherche qui ne sont pas encore assez couvertes ne sont pas évoquées. Toutefois, on peut mettre un léger bémol sur certains chapitres qui ne sont pas aussi « états de l'art » que d'autres.

L'ensemble du livre est réellement un *handbook*, une somme de savoirs passés et récents permettant d'avoir un premier aperçu de tout ce qui fait la linguistique computationnelle et le TAL ainsi que des pistes solides pour débuter une bibliographie dans un domaine. Ce livre est donc conseillé si l'on veut avoir toujours à portée de main un premier état de l'art des différents domaines du TAL.

Nicolas TURENNE. Besoins informationnels et extraction d'information. Vers une conscience artificielle. *Hermès-Lavoisier*. 2013. 288 pages. ISBN 978-2-7462-4507-5.

Lu par **Jorge Garcia FLORES**

LIPN – CNRS UMR 7030

L'ouvrage de M. Turenne propose la théorie suivante : les capacités langagières de l'être humain ont généré au fil de l'évolution un nouveau besoin cognitif qui est le besoin d'information. Ce besoin existerait chez l'humain au même titre que les autres besoins physiologiques de respirer, de se nourrir et de se reproduire. L'ouvrage définit ce besoin informationnel non comme une consommation constante de données, mais plutôt comme une activité dialogique de la conscience dont l'échange d'informations avec l'environnement produirait un sens existentiel, que l'auteur relie à la notion d'automotivation. À partir de cette perspective, le besoin informationnel humain produirait une série de traces observables et analysables par des techniques de fouilles de textes. Le corolaire à cette théorie serait que « si l'automotivation produit de l'information [...] qui est analysable et détectable par des agents intelligents de traitement de l'information et de la connaissance [...] de manière réciproque l'information qui est capturée peut être produite, ce qui est une démarche [...] de création potentielle de conscience » (page 249).

D'où une clé pour relire le titre de l'ouvrage, « besoin informationnel », surtout à ne pas confondre avec la notion de besoin informatique (en anglais : *requirement*), mais plutôt à considérer comme un besoin cognitif, philosophique, psychologique et social de l'humain pour motiver (linguistique et « informationnellement ») son élan vital. Les sept premiers chapitres de l'ouvrage (soixante-dix pages) sont consacrés à un bref état de l'art de trois notions : la conscience, le besoin et l'automotivation en philosophie, sociologie, psychologie et neurosciences. L'ouvrage fait d'abord un point terminologique, puis philosophique sur la notion de conscience (de Saint-Augustin à Piaget en passant par Nietzsche). Ensuite il survole quelques considérations théologiques (du bouddhisme au marketing) pour atterrir dans la pyramide des besoins de Maslow et continuer son tour de force avec les théories de la motivation en psychologie (on fait référence, entre autres, aux théories de l'auto-efficacité, de l'autodétermination et du contrôle). Quant à la motivation en neurosciences, l'auteur passe en revue des travaux récents consacrés aux mécanismes neurophysiologiques qui relient le système motivationnel aux émotions, au langage et à la gestion de besoins. Vu l'étendu du sujet, le traitement est forcément superficiel.

Si, comme cela a été le cas pour nous, le lecteur se voit attiré par le deuxième syntagme du titre, « extraction d'information », il doit alors commencer par le chapitre 8, « La modélisation du langage », pour trouver une remarquable synthèse historique des sciences du langage. La linguistique de corpus y est introduite d'une

façon extrêmement didactique, avec une prose qui maîtrise parfaitement son sujet. Des sections sont également consacrées au structuralisme, aux grammaires génératives, aux grammaires fonctionnelles, à la linguistique distributionnelle, à la théorie Sens-Texte ainsi qu'aux fondements du traitement du langage naturel. Pour le lecteur qui vient de l'informatique ou des mathématiques, cette section est une très éclairante porte d'entrée aux problématiques propres aux sciences du langage.

Le sous-titre de l'ouvrage, « vers une conscience artificielle », est le centre des trois chapitres suivants (9, 10 et 11). Ces chapitres mettent en relation l'hypothèse d'une automotivation cognitive traduite par un besoin informationnel avec son corolaire qu'est la modélisation possible d'une conscience artificielle fondée sur l'automotivation. L'auteur propose alors un modèle de la motivation s'appuyant sur les systèmes multi-agents auto-organisés, ainsi que sur quelques formalismes probabilistes et logiques, comme un modèle probabiliste de l'utilité définissant une fonction d'utilité du moi à deux arguments : les actions du moi effectif et les croyances dans tous les moi possibles. De même, l'auteur a recours à un modèle fondé sur la logique auto-épistémique pour définir un système d'autocroyance. Ainsi, M. Turenne propose une architecture cognitive d'un sous-système d'automotivation qui s'appuie sur quatre composants : l'organisation du langage, le système motivationnel, les facteurs de contrôle et le système neurophysiologique de régulation générale.

Au douzième chapitre, l'ouvrage se focalise enfin sur l'extraction d'information. Les cent dernières pages offrent au lecteur intéressé dans le traitement automatique des langues, un état de l'art sérieux, méthodique et riche en références à des travaux scientifiques récents en extraction de connaissances et fouille de textes.

Le chapitre 12, « Impact de l'automotivation sur l'information écrite », est une introduction à l'écosystème de la fouille de textes. L'auteur passe en revue des méthodes de détection d'entités nommées, de recherche documentaire, de suivi de thème, de classification des textes, d'extraction de concepts, d'analyse de sentiments et de modélisation de relations entre entités. L'auteur offre une perspective historique du domaine sans perdre de vue la portée épistémologique des méthodes mises en jeu. Le chapitre revisite aussi les divers niveaux de représentation des données textuelles : le caractère, le mot, le terme, l'entité nommée, l'étiquette grammaticale, l'étiquette sémantique et la relation ontologique, ainsi que les divers modèles (formels, relationnels, arborescents, vectoriels, probabilistes) qui aident à mettre en relation ces niveaux de représentation. Enfin, ce chapitre consacre deux sections à la notion de bruit lexical et à la fouille du Web (*Web mining*).

Le chapitre 13, « Techniques non transversales de la fouille de données », a trait aux spécificités de la représentation des données textuelles. Les concepts de base comme les n-grammes, la mesure tf-idf, la lemmatisation et les parties du discours sont introduits de manière claire et didactique. Une section très intéressante est consacrée aux dix-neuf lois de statistique textuelle, de la très populaire loi de Zipf à

la loi moins connue de Piotrowski. Une section est également consacrée à l'analyse de surface (*shallow parsing*).

Le chapitre 14 (« Les techniques transversales de la fouille de textes ») offre un état de l'art presque exhaustif en fouille de textes. L'auteur passe en revue diverses techniques d'extraction d'entités nommées. Des sections spécifiques sont consacrées aux approches d'accès au contenu de textes basées sur la fréquence et la cooccurrence lexicale, comme la LSA (analyse sémantique latente) et les méthodes de *ranking*. Les modèles de classifications textuelles classiques sont expliqués en détail : k-moyens (*k-means*), naïve bayésiennes, k-plus proches voisins, *clustering* hiérarchique, *clustering* basé sur une densité, champs conditionnels (CRF), réseaux de neurones, arbres de décision, ainsi que les approches de classification s'appuyant sur des séparateurs à vaste marge (SVM). Enfin, l'auteur analyse les méthodes de fouille basées sur les graphes et fait un point sur les mesures classiques d'évaluation (précision, rappel, F-mesure) d'une tâche de fouille de textes, pour clore le chapitre avec une réflexion sur le choix d'un modèle approprié aux diverses tâches de fouille.

Le quinzième et dernier chapitre, « Domaines d'intérêt de la fouille de textes », passe en revue les diverses domaines d'application de la fouille de textes : gestion des archives, veille, aide à la décision, sciences humaines et sociales, et sciences biomédicales.

Par la qualité de ses références et par sa capacité de synthèse, les chapitres 8, 12, 13, 14 et 15 proposent un remarquable état de l'art en extraction d'information et fouille de textes. M. Turenne offre un panorama très stimulant des défis scientifiques du domaine. Son propos a aussi la vertu d'essayer de donner une transcendance philosophique et épistémologique à une discipline qui, ces derniers temps, pourrait être prise pour une sous-branche de la statistique appliquée, plus que pour une héritière des sciences du langage ou de l'intelligence artificielle.

Quant à l'ouvrage pris dans son ensemble, l'auteur se voit obligé d'insister à plusieurs reprises sur le fait que son propos se veut « *moins mystique qu'universitaire* ». Sa théorie cognitive sur le besoin informationnel et l'automotivation est certainement intéressante, mais sa prose peine parfois à se focaliser sur son propos, prise comme elle est dans une propension encyclopédique qui gagne en extension interdisciplinaire ce qu'elle perd en intention et en profondeur d'analyse. Un travail d'édition plus soigné aurait peut être réservé certaines de ces pages pour un autre ouvrage de nature plus spéculatif sur la philosophie des sciences autour de la fouille de textes.