Notes de lecture

Rubrique préparée par Denis Maurel

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

Annelies BRAFFORT. La Langue des Signes Française (LSF). Modélisations, ressources et applications. *ISTE Editions*. 2016. 224 pages. ISBN 978-1-78405-050-4.

Lu par **Stéphanie GOBET**

Université de Poitiers – Laboratoire Forell

L'ouvrage dirigé par Annelies Braffort s'adresse en particulier à des étudiants aguerris, ou des ingénieurs et chercheurs spécialisés en TAL. Il est constitué de cinq chapitres dont le fil conducteur est la langue des signes abordée sous différents aspects : l'histoire, la pédagogie, la modélisation informatique et ses applications. Nous attirons l'attention sur les chapitres 2 et 3 qui exigent une lecture spécialisée en TAL.

Le premier chapitre introduit l'ouvrage en proposant une partie qui a pour vocation de vulgariser les recherches en linguistique des langues des signes afin que le lecteur, non initié à ces langues dont la modalité est visuo-gestuelle, découvre les travaux et leurs applications. Les auteurs, après un rappel sociolinguistique concernant le statut de la LSF et de ses locuteurs, abordent différents domaines, comme la phonologie, la poésie… Comme le rappellent les auteurs tout au long de leur chapitre, la langue des signes est devenue un objet pertinemment scientifique bousculant les modèles classiques, tout en en proposant de nouveaux, comme en TAL par exemple.

Le deuxième chapitre se focalise, de façon très pédagogique, sur la question des corpus en langue des signes. Après avoir rappelé les notions de transcription et d'annotations, tout en évaluant les apports et les difficultés liés à la modalité visuo-gestuelle de la langue des signes, les auteurs exposent la typologie des corpus à partir d'exemples et posent la problématique de la transcription des LS qui peut se réaliser sous forme graphique ou sous forme de gloses. Par conséquent, ce chapitre a pour intérêt de présenter les types de corpus existants, comment ils sont recueillis en langue des signes, avec quel protocole, mais surtout quelles en sont les exploitations possibles. La présentation comparative de différents projets (CREAGEST, TALS, DICTA-SIGN, par exemple) est l'occasion de discuter du traitement des données selon les objectifs du protocole et de l'archivage de ces données visuelles.

Le troisième chapitre traite plus précisément et de manière très technique de la modélisation linguistique appliquée à la langue des signes. Les auteurs définissent dans un premier temps les différents types de données disponibles (brutes, degré

intermédiaire, etc.) pour les chercheurs en TALS (traitement automatique des langues signées). Des exemples illustrent les propos des auteurs, mais surtout permettent de visualiser le type de données et leur traitement. Il existe aujourd'hui différentes technologies et différents logiciels pour les représenter (image, rotoscopie, modèle Zebedee, signeur virtuel, etc.). Comme l'exposent les auteurs, les modèles linéaires ne permettent pas de représenter la multilinéarité spécifique aux langues des signes. Par conséquent, les modèles doivent être repensés et ont donné lieu à de nouveaux travaux concernant la traduction automatique appliquée aux langues des signes.

Les deux derniers chapitres sont des mises en application de la modélisation informatique en langue des signes.

L'avant-dernier chapitre a une visée plus spécifiquement pédagogique. Il s'adresse à tous ceux qui réfléchissent aux supports pédagogiques. Après être revenus sur le statut de la langue des signes au sein du système éducatif et sur le combat mené pour sa reconnaissance par la loi (en France et dans le monde), les auteurs abordent la question de la trace écrite pour les langues à modalité visuo-gestuelle. Comme il est explicitement exposé, le support vidéo est devenu, au-delà d'une simple trace, un outil pédagogique et didactique. Bien que non reconnue par l'Éducation nationale, la vidéo répond aux besoins des enseignants lors de la création de cours, mais est aussi devenue le support de glossaires. Des sites tels que Wikisign ont été conçus comme des ressources lexicales. Les auteurs, à travers cet article, attirent l'attention du lecteur sur la problématique de la langue des signes qui n'a pas d'équivalent écrit, alors que l'écrit est nécessaire en pédagogie et en didactique lors de la transmission du savoir. Face à ce besoin, des ressources pédagogiques ont vu le jour, telles que LogiSignes dont les logiciels sont décrits à la fin du chapitre. Pour toute personne réfléchissant à l'amélioration de ses cours, les auteurs apportent ici des exemples concrets – tout en les analysant –, offrant ainsi aux pédagogues des outils pour améliorer leurs pratiques.

Le cinquième et dernier chapitre traite de l'application de la modélisation informatique des langues des signes au grand public. L'innovation, en termes de logiciels et de ressources, a créé de nouveaux besoins pour la communauté sourde. L'un d'entre eux est la transmission de l'information au grand public. Cette transmission, générée par la vidéo, doit répondre à différentes contraintes exposées par les auteurs et peut se réaliser différemment. Les auteurs expliquent, en particulier, comment la LS peut s'articuler dans un document, soit par incrustation soit par juxtaposition, avec des éléments tels que l'image, le pictogramme, la vidéo, voire la conjonction de plusieurs éléments comme il apparaît sur le site Internet Macif Sourds. Toutefois, la question de cette articulation a soulevé le questionnement dû à la multimodalité des langues des signes. Ces dernières n'étant pas linéaires, la problématique de l'articulation des LS dans le document est confrontée à des problématiques autres telles que l'interaction et la navigation dans l'ensemble des documents. La typologie des LS implique donc de développer de nouveaux systèmes afin que l'accès à l'information pour la communauté sourde soit optimal. Les auteurs signalent tout de même l'existence d'applications à destination du grand public tout en rappelant la difficulté d'évaluer ces outils. Ce chapitre

conclut sur un aspect abordé tout au long de l'ouvrage : le signeur virtuel. L'avatar, dont le processus de création a été présenté dans le chapitre précédent, constitue une avancée dans le domaine public par son confort de visualisation, à condition de répondre à des critères tels que la visibilité ou encore l'esthétique.

Bien que l'ouvrage concerne davantage les spécialistes en TAL, il a été conçu de manière très pédagogique, les illustrations et exemples permettant de visualiser les concepts traités pour tout lecteur souhaitant avoir une première approche des travaux en TAL. Toutefois, les chapitres deux et trois requièrent des prérequis en traitement de corpus et modélisation informatiques.

Jannik STRÖTGEN, Michael GERTZ. Domain-Sensitive Temporal Tagging. *Morgan & Claypool Publishers*. 2016. 133 pages. ISBN 978-1-62705-459-1.

Lu par **Yannis HARALAMBOUS**

IMT – UMR CNRS 6285 Lab-STICC

Le titre de cet ouvrage nous renseigne déjà sur deux de ses particularités : primo, Temporal Tagging *indique qu'il s'agit non pas d'extraction d'informations temporelles en général, mais seulement de balisage temporel. L'ouvrage se concentre donc sur l'extraction, la classification et la normalisation d'expressions temporelles, et ne s'occupe pas de leur interprétation, de l'extraction d'évènements et de relations entre ces évènements. Secundo,* Domain-Sensitive, *c'est-à-dire « spécifique à un domaine donné », représente la véritable innovation de l'ouvrage, puisqu'il s'agit de comparer le balisage temporel de quatre types de documents, de natures assez différentes.*

Le premier auteur, Jannik Strötgen, post-doctorant à l'Institut Max-Planck de Sarrebruck, est un spécialiste de l'annotation temporelle aussi bien du point de vue du Web sémantique, que de celui de l'étude de corpus historiques et littéraires. Le deuxième auteur, Michael Gertz, professeur à l'université Heidelberg et directeur de l'Institut d'informatique de celle-ci, participe à un grand nombre de projets, dans différents domaines du traitement automatique de langue et de la modélisation. D'ailleurs, cet ouvrage est fondé sur la thèse de doctorat du premier auteur, dont le deuxième auteur était directeur de thèse.

Structure de l'ouvrage

L'ouvrage est divisé en six chapitres, dont les trois premiers sont introductifs, le quatrième traite du cœur de la question (spécificités du balisage temporel selon le domaine d'application), le cinquième énumère un certain nombre d'outils, méthodes et projets, et le sixième conclut avec des directions de recherches futures. L'ouvrage peut paraître court (133 pages en tout, dont seulement 38 pour le quatrième chapitre, qui est le plus important), mais il est dense et contient énormément d'informations, notamment sur les ressources externes, les projets, et les défis.

Chapitres 1 à 3

L'introduction réussit sans trop de mal à nous convaincre de l'importance et de la complexité de l'extraction d'informations temporelles, notamment en illustrant

cette tâche par une page Wikipédia munie d'un ensemble de flèches enchevêtrées, reliant les expressions temporelles contenues dans le texte et les points ou intervalles de la frise chronologique correspondante.

Le deuxième chapitre présente de manière très simple la mesure standard du temps et les quatre types d'expressions temporelles : explicite (« je suis né le 16 février 1962 »), sous-spécifiée (« j'ai terminé mes lectures vendredi »), relative (« ils se sont vus hier »), et implicite (« les vacances commencent le lundi de Pâques de cette année »). Ce chapitre est tellement clair et concis, que l'on regrette presque l'abondance encyclopédique à laquelle un autre auteur (comme Umberto Eco[1], pour ne pas le nommer) se serait certainement prêté : il n'est question ni de calendrier hégirien (avec ces mois de vingt-neuf à trente jours), ni de calendrier républicain (avec des décades à la place des semaines), ni des turpitudes du passage du calendrier julien au calendrier grégorien, passage différent dans chaque pays, qui s'est étalé sur cinq siècles, ni de la célèbre interprétation du mot « maintenant » par Michel Serres comme « Main tenant, tenant en main le monde »... Les auteurs n'étaient certainement pas obligés de parler de tout cela, mais cela aurait pu être une agréable surprise pour le lecteur.

Le troisième chapitre est très hétérogène. Il commence par une description technique et détaillée des normes XML afférentes (TIDES et TimeML), se poursuit par une section sur l'évaluation des outils d'analyse, enchaîne naturellement en énumérant les différentes compétitions, et enfin, donne une liste de corpus de type journalistique existant dans différentes langues.

Chapitre 4

Le quatrième chapitre part du constat que l'écrasante majorité des textes analysés sont des textes journalistiques (en anglais : *news style*), textes pour lesquels l'annotation temporelle est cruciale, et qui sont rédigés de manière assez similaire. Les auteurs définissent un domaine comme un « groupe de documents ayant les mêmes caractéristiques vis-à-vis de la tâche de balisage temporel » et énumèrent des corpus de documents temporellement balisés de domaines autres que le domaine journalistique : on y trouve des textes d'histoire, des collections de pages Wikipédia, des SMS, ainsi que des récits cliniques et des rapports de pathologie.

Pour mettre de l'ordre dans cette nébuleuse de méthodes de représentation et d'utilisation de la temporalité, ils décrivent quatre grands domaines (au sens du balisage temporel) :

1) *les documents journalistiques* : on y trouve un grand nombre d'expressions temporelles sous-spécifiées et relatives ; le temps de référence est souvent le temps de création du document ; les temps des verbes sont importants pour normaliser les expressions sous-spécifiées ; néanmoins, le temps du verbe peut induire en erreur,

1 En effet, dans le *Pendule de Foucault*, chap. 71, Eco décrit un rendez-vous manqué entre Rose-Croix anglais et français, le même jour étant pour les premiers le 13 juin 1584 et pour les deuxièmes le 23 juin, puisque la réforme du calendrier a aboli dix jours en 1583 en France, mais n'a été adoptée en Angleterre qu'en 1752.

ainsi dans la phrase « la réunion a été reportée à lundi de la semaine prochaine », le verbe est au passé, mais la date donnée est dans le futur (sans doute parce que l'information implicite est « la réunion a été reportée et aura lieu le lundi de la semaine prochaine », où le verbe est bien au futur) ;

2) *les documents narratifs*, l'exemple paradigmatique étant les pages Wikipédia : on y trouve un grand nombre d'expressions explicites, mais aussi des expressions relatives à ces dernières ; le temps de référence change assez souvent et on est amené à déterminer les expressions sous-spécifiées par rapport à celui-ci ; de par cette relativité, des erreurs peuvent se propager, et on trouve parfois des phénomènes qui peuvent être modélisés par des automates de type fini, par exemple les ères « av. J.-C. » et « apr. J.-C. » : l'auteur ayant donné une date av. J.-C. aura tendance à ne plus spécifier explicitement l'ère jusqu'à l'occurrence d'une date dans l'autre ère, et *vice-versa* ;

3) *les documents en langage informel ou familier*, l'exemple paradigmatique donné étant des corpus de SMS : il faut interpréter les parophonies (« 2m1 » pour « demain ») et détecter les erreurs ; le temps de référence est le temps de publication ; le contexte est souvent manquant ; les temps des verbes ne sont pas toujours cohérents (« j'arrive demain », le verbe est au présent alors que l'action se situe au futur) ;

4) *les documents autonomiques* : le terme *autonomique* provient de la médecine et se réfère à des processus biologiques *involontaires* par opposition à des processus *indépendants* qui sont qualifiés d'*autonomes* (par exemple, le système nerveux autonome est appelé *autonomic* en anglais au lieu d'*autonomous*). Les auteurs désignent par ce terme les documents contenant une majorité d'expressions temporelles qui ne peuvent pas être normalisées dans un cadre temporel absolu, mais doivent rester dans un cadre temporel local (exemple : les fameux « T = 0, T + 3, T + 6, T + 12 » que l'on trouve dans les plannings). En guise d'exemple, ils mentionnent les textes scientifiques et les récits cliniques. Dans ce type de documents, il s'agit de détecter et de déterminer le « temps zéro » et d'interpréter la sémantique des expressions relatives.

Les auteurs comparent ces quatre domaines, et donnent des stratégies d'analyse circonstanciées.

Chapitres 5 et 6

Le cinquième chapitre nous offre un panorama des outils existants et de différentes compétitions du domaine, y compris celles comportant des corpus multilingues. Le sixième chapitre comprend une conclusion et une liste de directions de recherches futures proposées : la comparaison des annotations temporelles dans les différentes langues, l'identification du temps de référence, la compréhension des cadres temporels locaux, l'adaptation à des disciplines spécifiques (par exemple, la paléontologie ou le sport), l'identification automatique du domaine, l'évaluation de la confiance des expressions temporelles, le traitement des zones horaires, les références temporelles implicites (« pendant la campagne présidentielle 2017 », « à la saint-glinglin »), etc.

Conclusion

Les points positifs de cet ouvrage sont nombreux : il est concis et extrêmement clair, et il fourmille de liens vers des outils, des ressources, des projets et des compétitions.

En guise de point négatif, on peut noter que le choix des « domaines » opéré par les auteurs peut sembler un peu arbitraire et que les auteurs ne font appel à aucune théorie linguistique pour fonder cette classification. Quoi qu'il en soit, les exemples donnés sont intéressants et le lecteur ayant parcouru le quatrième chapitre est tout à fait en mesure d'appliquer les mêmes méthodes à d'autres types de documents auxquels il peut être confronté[2]. Pour conclure, nous considérons que cet ouvrage est bien indiqué à ceux qui ont des expressions temporelles à baliser, et même à ceux qui souhaitent s'y investir et participer aux diverses compétitions.

Karën FORT. Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects. *Wiley-Iste*. 2016. 164 pages. ISBN 978-1-84821-904-5.

Lu par **Aurélien BOSSARD**

Université Paris 8 – Laboratoire d'informatique avancée de Saint-Denis

---

L'autrice part du constat que les ressources textuelles annotées sont un élément décisif pour tout système de traitement automatique du langage. Elles sont en effet nécessaires à la fois pour l'entraînement des systèmes et pour leur évaluation. L'autrice définit ce qu'est une annotation et les enjeux généraux de l'annotation puis décrit le sujet, les enjeux spécifiques, et dresse un état des lieux de l'annotation participative et/ou collaborative.

L'ouvrage d'environ cent vingt pages plus l'appendice a pour objectif de décrire les méthodologies et les outils existants d'annotation collaborative et participative pour le traitement automatique du langage ainsi que d'en montrer les forces et les faiblesses. L'annotation collaborative et sa petite sœur, l'annotation participative, sont des enjeux majeurs du TAL.

Dans une introduction très documentée et illustrée, l'autrice définit ce qu'est une annotation, en replaçant le processus d'annotation dans une chronologie qui court des débuts de l'écriture à nos jours.

---

2 Sans oublier, en tant qu'exemple extrême, les textes à frise chronologique circulaire, comme les récits de voyages temporels, qui sont un véritable sous-genre de la science-fiction, et dont H. G. Wells (*La machine à explorer le temps*), J. Finney (*Le voyage de Simon Morley*), R. A. Heinlein (*Vous les zombies*), R. Silverberg (*Les temps parallèles*), C. Webb (*Les quinze premières vies d'Harry August*) et notre cher R. Barjavel national (*Le voyageur imprudent*) sont des illustres représentants.

Le chapitre 1 est dédié à l'annotation collaborative. Y sont évoqués les enjeux ainsi que les méthodologies de la constitution collaborative de données annotées. Y sont également décrites clairement différentes méthodologies ainsi que leurs points communs et leurs différences, depuis l'identification des acteurs de l'annotation, en passant par l'écriture de guides d'annotation, jusqu'aux différents outils d'évaluation de la qualité d'annotation.

Ce chapitre met également en évidence, après avoir présenté les formats d'annotation et les outils existants, les défis posés à ces derniers : la généricité, la prise en compte de l'aspect collaboratif et les biais associés, ainsi que la place du processus d'annotation dans une campagne d'annotation.

Le deuxième et dernier chapitre est consacré à l'annotation participative dont l'autrice avait déjà montré qu'elle est une évolution naturelle de l'annotation collaborative et qui remet la participation citoyenne au centre d'un processus scientifique. Après avoir défini l'annotation collaborative, l'autrice montre les spécificités de ce type d'annotation : nécessité d'évaluer et de prendre en compte la compétence des participants, problématiques liées aux quantités des participants et de leur contribution, motivation de ces derniers (financière comme pour *Amazon Mechanical Turk*, ludique au travers de jeux sérieux...).

Ce chapitre passe en revue les différentes utilisations que l'on peut faire d'un participant dans le cadre des jeux sérieux, de la simple utilisation des capacités innées jusqu'à celle des capacités d'apprentissage des joueurs. Le chapitre clôt avec les problèmes éthiques liés aux jeux sérieux et aux plateformes participatives rémunérées et démontre la nécessité de l'utilisation d'une charte éthique pour les scientifiques et les financeurs.

L'ouvrage comporte également un appendice qui référence, de manière très exhaustive, les outils existants dans les domaines de l'annotation collaborative et de l'annotation participative, les classe en trois catégories (génériques, orientés tâches et spécifiques au TAL) et les décrit ainsi que leurs points forts et leurs faiblesses.

L'ouvrage se termine par un glossaire réduit au strict minimum et une bibliographie très complète.

Commentaires

Cet ouvrage, extrêmement bien documenté et complet, est, à ma connaissance, la seule référence à l'heure actuelle pour l'annotation collaborative et participative dans le domaine du traitement automatique du langage. Il repose sur des années de recherche et de collaborations de la part de l'autrice. L'ouvrage, par l'effort pédagogique auquel a consenti l'autrice, est accessible à toute personne qui s'intéresse au traitement automatique du langage, mais également utile, par la qualité et la complétude de l'état de l'art et le recul de l'autrice sur le sujet, à un chercheur expérimenté.

Le traitement automatique du langage est fortement dépendant de la quantité et de la qualité des ressources annotées, à une heure à laquelle les systèmes d'apprentissage supervisés prennent une place de plus en plus prépondérante. Ainsi,

constituer et disposer de ressources annotées fiables est essentiel. C'est pourquoi nous ne pouvons que recommander la lecture de cet ouvrage à tout chercheur qui se prépare soit à mettre en place une campagne d'annotation collaborative, soit à utiliser des données issues de l'annotation collaborative et/ou participative. En effet, tous les problèmes, qu'ils soient d'ordres méthodologique, quantitatif, qualitatif ou éthique ne sont évidents ni à cerner, ni à résoudre.

Xavier-Laurent SALVADOR. XML pour les linguistes. *L'Harmattan*. 2016. 189 pages. ISBN 978-2-34309-956-9.

Lu par **Lydia-Mai Ho-Dac**

Université de Toulouse – CLLE-ERSS

« XML pour les linguistes » *se donne pour objectif de présenter le langage XML et ses applications à des chercheurs en lettres et sciences humaines et sociales à même de produire des ressources langagières dans un format moins explicite et partagé que l'XML. L'objectif est clairement de rassurer un public parfois frileux devant la technologie informatique et de le convaincre que le langage XML est adapté aux recherches en linguistique. La présentation commence par les éléments historiques dont hérite le langage XML pour introduire de façon quelque peu littéraire le fonctionnement du langage XML. Viennent ensuite des parties plus didactiques qui expliquent la syntaxe XML, l'utilité et la création de DTD et schémas XSD, les normes XML (TEI, RDF, OWL) et les langages de requête et de transformation disponibles (Xpath, XSL, XQuery). La dernière partie présente trois projets menés par l'auteur mettant en jeu des ressources XML variées (corpus littéraires, dictionnaires, transcriptions) et l'outil Isilex développé par l'auteur.*

Cet ouvrage se donne pour objectif de présenter le langage XML et son potentiel à des étudiants et chercheurs en lettres et sciences humaines et sociales. Tout en définissant les règles de base du langage XML, l'auteur justifie, documente et illustre l'utilisation de cette norme pour des études en linguistique et plus largement des études manipulant du matériau langagier.

Les notions informatiques présentées sont régulièrement reliées à des notions historiques (le chevron, caractère utilisé par les copistes bibliques pour distinguer des éléments textuels différents, le codex comme éternelle base de représentation d'un texte, etc.) dans un souci de montrer que finalement rien n'est vraiment nouveau. Cette attention semble directement adressée à un public frileux à l'égard des technologies informatiques et que l'auteur souhaite convaincre de la simplicité et du bien-fondé du langage XML.

Les avantages du langage XML pour la recherche en lettres et sciences humaines et sociales sont documentés et illustrés à travers une variété d'exemples de ressources en XML (dictionnaires, lexiques, corpus de textes bruts et annotés, etc.) et de projets collaboratifs.

Organisation et contenu

L'ouvrage commence par une introduction à certaines notions fondamentales pour manipuler des formats numériques (le « plasma numérique » selon l'auteur) : octets, caractères et encodage, documents numérisés *vs* électroniques, formats de fichiers, etc.

Le chapitre 2 décrit le langage XML et un certain nombre de normes associées. Il commence par un aperçu des origines du langage XML en tant que convention typographique, puis de la syntaxe XML présentée comme permettant de décrire et déclarer des objets langagiers, à la manière des didascalies (« ceci est un paragraphe »). Après une définition des espaces de nommage, décrits comme des dialectes de XML, une large partie de chapitre 2 est dédiée à la nécessité de normaliser la description et l'encodage des éléments d'une ressource, ainsi qu'aux moyens utilisés pour assurer cette normalisation et, à terme, une pérennisation de la ressource. Sont parcourus les formalismes XML utilisés par les outils de traitement de texte (Open Office et MSWord), la gestion et la création de DTD et de schémas XML, la norme TEI pour l'encodage de ressources textuelles, le modèle RDF pour la structuration des données du Web sémantique, et le format OWL pour l'encodage de ressources terminologiques et les ontologies du Web.

Le troisième chapitre est consacré aux langages de requête et de transformation XPath, XSL et XQuery. La présentation est orientée de façon à montrer toute la valeur ajoutée d'une ressource structurée en langage XML. Le chapitre commence par les bases de la syntaxe d'une requête XPath et d'une transformation XSLT, avant de décrire le langage de requête et de transformation XQuery, langage qui sera utilisé dans les applications et projets présentés dans la suite de l'ouvrage. Le choix pour le XQuery est justifié par son potentiel (c'est un langage de programmation en tant que tel) et la possibilité de l'utiliser *via* le logiciel BaseX, une application multiplateforme *open source* développée à l'université de Konstanz. La fin de ce chapitre illustre le potentiel de la combinaison XML, Xpath et XQuery par des « exemples de manipulations de corpus en XQuery » avec le logiciel BaseX : exploration de corpus bruts ou annotés, annotation de corpus et mise en place d'une interface en ligne pour la consultation et l'interrogation d'une ressource XML. Le chapitre fini sur une liste éclectique d'outils permettant la création, la manipulation et/ou l'exploitation de ressources XML : Oxygen. ToolBox, TXM, Transcriber, PRAAT, ELAN, et Isilex, développé par l'auteur et utilisé dans les projets décrits dans le chapitre 4.

Le quatrième et dernier chapitre présente trois projets menés par X.-L. Salavador avec le logiciel Isilex. Ces projets illustrent des cas de (1) détection automatique de thématiques dans des textes littéraires ; (2) de visualisation d'annotations sous forme de graphes ; (3) de construction et gestion collaborative du dictionnaire Crealscience, dictionnaire de français scientifique médiéval ; (4) de transformation du format XML vers le format LaTeX ; (5) de transcription collaborative pour une version numérique et consultable de l'Exode de la *Bible historiale*.

Commentaire

« *XML pour les linguistes* » est conçu comme un livre qui se lit plus qu'un manuel qui s'utilise. L'auteur met l'accent sur le fait que le langage XML hérite de concepts issus de la tradition de l'édition et de l'écriture depuis ses origines, parfois au détriment d'indications simples pour la bonne prise en main du langage XML par un public néophyte.

Le caractère didactique est toutefois présent dans les trois premiers chapitres, notamment grâce à un lexique qui fournit des définitions complètes des termes utilisés dans l'ouvrage et également à un certain nombre d'encadrés offrant des résumés didactiques (principales règles de syntaxe XML, procédures pour construire un projet collaboratif de base de données XML, etc.) et des mini-tutoriels (gérer des fichiers XML en ligne de commande, etc.). On regrettera cependant le peu d'outils pour faciliter la navigation à l'intérieur du document : pas de renvoi depuis le lexique aux pages pertinentes, absence de table des encadrés, des figures et exemples utilisés, faible nombre de titres de section pour pointer sur des aspects spécifiques du langage XML.

Cet ouvrage s'adresse donc à des étudiants et chercheurs en lettres et sciences humaines et sociales qui ne cherchent pas un manuel, mais plutôt une sorte de cours à lire pour comprendre pourquoi et comment utiliser le langage XML. La première partie peut tout à fait être utilisée en support de cours par des néophytes. La partie consacrée à la manipulation de ressources XML par le langage XQuery et le logiciel BaseX est davantage construite comme un tutoriel adressé à un public plus averti ayant certaines compétences en programmation et gestion de systèmes informatiques, des « ingénieurs du texte » selon les propos de l'auteur. Le dernier chapitre sert, quant à lui, à présenter des projets menés par l'auteur, sans spécialement adopter une orientation didactique.

Yoav GOLDBERG. Neural Network Methods for Natural Language Processing. *Morgan & Claypool Publishers*. 2017. 287 pages. ISBN 978-1-62705-298-6.

Lu par **Franck Burlot**

Limsi – CNRS

L'ouvrage propose une initiation à la fois aux méthodes neuronales et aux différentes architectures qui constituent l'état de l'art en traitement automatique des langues (TAL). Il est destiné aussi bien aux étudiants qu'aux ingénieurs et chercheurs professionnels en TAL et ne suppose que quelques connaissances mathématiques basiques. Les qualités didactiques déployées par l'auteur permettent d'aborder progressivement des architectures complexes et d'assimiler les connaissances nécessaires pour être indépendant dans l'exploration de la littérature spécialisée. Les modèles principaux sont exposés (perceptron, réseaux convolutifs et récurrents) et appliqués de manière pratique à différentes tâches de classification et de

génération (analyse des sentiments, étiquetage, traduction automatique). L'importance des modèles neuronaux dans le TAL ne cesse de grandir et cet ouvrage est une ressource unique pour se former efficacement à l'apprentissage profond.

L'ouvrage consiste en une introduction aux réseaux de neurones appliquée au traitement automatique des langues (TAL). Il évoque les tâches caractéristiques du domaine, pour lesquelles on observe, depuis quelques années seulement, le succès grandissant des méthodes dites « d'apprentissage profond ». L'auteur a publié en 2016 dans la revue « *Journal of Artificial Intelligence Research* » un rapport de soixante-seize pages sur le même sujet, intitulé *A Primer on Neural Network Models for Natural Language Processing*, destiné à la communauté du TAL. L'ouvrage présenté ici reprend ce rapport, l'étend, et s'ouvre ainsi à un public plus large en manifestant un effort didactique considérable et en ajoutant une présentation assez détaillée du TAL. S'il peut donc être intéressant pour un spécialiste d'apprentissage automatique, son propos principal reste le fonctionnement des modèles neuronaux auxquels il s'agit d'initier les spécialistes, ou simplement amateurs, du TAL. En effet, outre quelques connaissances très basiques en algèbre linéaire (opérations de matrices) et en statistiques (chaînes de Markov), la connaissance d'aucun domaine particulier n'est nécessaire. L'auteur précise en préface qu'il ne s'agit nullement d'une introduction à l'apprentissage automatique, néanmoins il ne fait référence à aucune connaissance particulière à avoir acquise avant la lecture de l'ouvrage, qui s'ouvre sur un chapitre simple et clair décrivant les modèles linéaires.

L'ouvrage est publié dans la collection « *Cours de synthèse sur les technologies du langage humain* » et il s'agit bien là d'un cours, dont la valeur pédagogique est indéniable. Les différents modèles neuronaux présentés sont inclus dans le cadre d'une progression systématique, où chaque nouveau détail fait référence à ce qui a été présenté précédemment. Ainsi, après la description de la régression linéaire, le perceptron multicouche est décrit comme une simple « superposition de modèles linéaires séparés par une fonction de non-linéarité ». Cette simplicité dans la formulation est constante et conduit naturellement à des architectures très complexes en fin d'ouvrage.

Le cours présente la théorie nécessaire à la compréhension des modèles neuronaux, sans toutefois noyer le lecteur dans des descriptions abstraites, et introduit régulièrement des notions très pratiques. Ainsi, une section est consacrée à la notion de graphe computationnel, qu'il est essentiel d'acquérir puisqu'elle est employée dans la plupart des modules d'implémentation de réseaux de neurones (Theano, TensorFlow, DyNet). Par ailleurs, l'auteur donne régulièrement des conseils pratiques concernant des détails d'implémentation ou d'utilisation. Il met par exemple le lecteur en garde contre l'utilisation telle quelle de représentations continues de mots préentraînées et propose plusieurs méthodes efficaces pour les intégrer dans un nouveau modèle. Ces détails sont précieux, puisqu'ils sont parfois omis dans la littérature et traités comme allant de soi. C'est là une importante ambition de l'auteur : rendre le lecteur capable d'aborder la littérature technique et scientifique sur le sujet. En effet, certains articles sont parfois elliptiques et omettent des détails importants, comme la présence d'une fonction d'activation ou d'un *softmax*. C'est dans ce même souci qu'il s'arrête rigoureusement sur les flous

terminologiques courants dans un domaine qui connaît actuellement un développement spectaculaire. Ainsi, la « log-vraisemblance négative » est aussi désignée comme « l'entropie croisée », une « fonction non linéaire » comme une « fonction d'activation », une « convolution » comme un « filtre », etc.

La bibliographie est abondante. L'auteur y inclut des ouvrages et articles fondateurs, ainsi que des descriptions de modèles très récents au moment de la publication, au risque qu'ils soient dépréciés rapidement.

La rédaction de l'ouvrage est assez claire, mais comprend un très grand nombre de coquilles qui ne gênent toutefois la compréhension que rarement. Ces négligences dans le travail d'édition semblent courantes chez Morgan & Claypool Publishers. La structure en quatre parties est judicieuse et permet une progression efficace.

La première partie introduit les notions de base (perceptron, rétropropagation, optimisation). L'auteur se désolidarise à juste titre de l'analogie entre les réseaux de neurones artificiels et l'activité cérébrale humaine, qui connaît un heureux succès dans la presse, et présente la notation mathématique qui sera utilisée dans tout l'ouvrage. Cette partie est probablement celle qui nécessitera le plus d'effort de la part du lecteur peu familiarisé à l'apprentissage automatique.

La deuxième partie propose un tour d'horizon des tâches les plus célèbres du TAL, qui concentreront cette fois l'attention du lecteur spécialiste d'apprentissage automatique non initié au domaine. Elle aborde également la question des enjeux liés aux différentes représentations symboliques et continues de la langue. La tâche de modélisation de la langue est présentée en détail et conduit aux algorithmes les plus employés pour apprendre des représentations continues de mots, ainsi que les différentes manières d'utiliser ces représentations. La partie se clôt sur la présentation complète d'un modèle d'inférence du sens des phrases, qui met en application de nombreuses notions abordées plus tôt.

La troisième partie évoque les deux types d'architectures couramment utilisées en TAL : les réseaux convolutifs et récurrents, avec une interprétation très intuitive des différentes composantes d'un LSTM. La performance des réseaux récurrents dans la génération de texte est mise en valeur et le modèle d'encodeur-décodeur avec mécanisme d'attention est présenté, là encore, de manière très progressive et intuitive, ce qui permet de dresser une description assez détaillée du nouvel état de l'art en traduction automatique, dont certaines composantes sont apparues moins d'un an avant la publication de l'ouvrage.

La quatrième partie est plus disparate et évoque des sujets plus avancés. Après une présentation des réseaux récursifs pour l'encodage des arborescences, l'auteur reprend les questions liées à l'étape du décodage et à l'exploration de l'espace de recherche dans une sortie structurée, qui relèvent plutôt de l'apprentissage automatique général. Cette section est brève et apparaîtra probablement comme complexe au lecteur non familiarisé au domaine. L'ouvrage se clôt sur la combinaison de différents modèles et introduit notamment l'apprentissage multitâche.

À l'issue de cette lecture, le spécialiste de TAL qui découvre les modèles neuronaux est en mesure d'aborder une grande partie de la littérature de l'état de l'art et n'a plus qu'à choisir un module pour se lancer dans ses premières implémentations. Cette ambition rendue réaliste fait de l'ouvrage l'entrée la plus efficace dans le domaine de l'apprentissage profond.

Jeffrey HEINZ, Colin de la HIGUERA, Menno van ZAANEN. Grammatical Inference for Computational Linguistics. *Morgan & Claypool Publishers*. 2015. 136 pages. ISBN 978-1-60845-977-3.

Lu par **Isabelle TELLIER**

Université Paris 3 – Sorbonne Nouvelle – Laboratoire LaTTiCe, UMR 8094

Ce livre propose un panorama synthétique et pédagogique de l'inférence grammaticale, domaine qui se consacre à l'apprentissage automatique de modèles de langages (comme les automates finis ou les grammaires formelles) à partir d'exemples. Ses composantes « théoriques » aussi bien qu'« empiriques » sont abordées, et autant que possible illustrées par des exemples empruntés au traitement automatique des langues.

L'inférence grammaticale est une branche de l'apprentissage automatique qui étudie comment il est possible d'acquérir par programme une « grammaire » ou un autre modèle formel de langage[3] (un automate fini, par exemple) à partir de données (en général des séquences de symboles, mais aussi éventuellement des arbres d'analyse) qu'il génère (ou pas). Ses principes et ses résultats sont souvent mal connus, parce qu'elle vise l'identification d'*objets symboliques structurés* et ne rentre pas, de ce fait, dans le cadre classique de l'apprentissage statistique. Ce sont d'ailleurs souvent des chercheurs venant de l'informatique théorique, de la « théorie des langages » et de la combinatoire qui ont le plus contribué à son développement. Quant aux spécialistes du traitement automatique des langues, à qui s'adresse prioritairement le livre, s'ils connaissent en général les grammaires formelles, peu sont familiers avec l'inférence grammaticale. Les problèmes auxquels elle s'attaque pourraient pourtant intéresser nombre d'entre eux. L'ouvrage arrive donc à point nommé pour combler un fossé entre différentes communautés. La tâche est délicate, car leurs cultures et leurs références ont largement divergé. Mais le pari du rapprochement vaut d'être tenté.

Le premier chapitre de l'ouvrage introduit les spécificités de l'inférence grammaticale, pour laquelle la lisibilité et l'interprétabilité de la structure cible sont fondamentales. Le champ est extrêmement vaste ; les auteurs reconnaissent avoir dû faire des choix. Ils tiennent toutefois à préserver une certaine variété représentative et s'attacheront donc à présenter à la fois les aspects « formels » (modèles théoriques

3 J'utilise dans ce texte le terme « langage » (qui est une traduction littérale de l'anglais « language ») dans son sens informatique et non linguistique, comme une combinatoire de symboles régulés par des règles explicites.

d'« apprenabilité » et résultats associés) et « empiriques » (comportement de certains algorithmes sur certaines données) du domaine. Mais c'est sur la branche « formelle », historiquement la première et scientifiquement la plus féconde, que l'accent est d'abord mis. Les préliminaires indispensables de la théorie des langages (langages, grammaires, automates finis...) figurent ainsi également dans ce chapitre.

Le suivant est consacré aux modèles d'apprentissage utilisés en inférence grammaticale « formelle » et à quelques résultats associés. Dans ce domaine, en effet, on ne juge pas du succès d'un programme de la même façon qu'en apprentissage statistique. Un modèle d'apprentissage est caractérisé par un « scénario » qui précise les conditions dans lesquelles un algorithme accède à ses données d'entrée (une par une, toutes d'un coup ou *via* des requêtes, par exemple) et par l'ensemble des critères qu'il doit satisfaire. Ces critères peuvent combiner la qualité du résultat produit avec d'autres propriétés (par exemple de convergence, de complexité...). Plusieurs de ces « modèles formels » (modèle de Gold, « PAC learning »...) sont ainsi détaillés. Chacun peut être vu comme une manière de délimiter précisément les frontières de ce qui est « apprenable » par programme. Quelques résultats saillants d'apprenabilité au sens de ces modèles sont ensuite évoqués. Ils concernent les automates finis (déterministes ou non), les expressions régulières, les HMM (*Hidden Markov Models* ou « chaînes de Markov cachées »), les transducteurs et les grammaires algébriques (ou « *context free* ») ainsi que leur version probabiliste, quand cela a un sens. Pour les linguistes ou les spécialistes du TAL qui les découvriront, ces résultats seront à la fois impressionnants et frustrants : impressionnants parce que ce sont des théorèmes mathématiques qui formalisent des notions habituellement laissées à l'intuition, frustrants parce que, comme tous les « modèles », ils fixent des conditions simplificatrices dans lesquelles les situations humaines ont du mal à rentrer.

Le troisième chapitre se concentre sur la famille de langages ayant donné lieu au plus grand nombre de travaux en inférence grammaticale : les langages réguliers (ou rationnels). C'est l'occasion de se pencher de plus près sur le fonctionnement de quelques algorithmes de référence, capables d'identifier un automate fini à partir de séquences qu'il reconnaît ou produit (ce que l'on appelle des « exemples positifs ») et, éventuellement, de séquences qu'il ne reconnaît pas (« exemples négatifs »). C'est l'occasion aussi de s'attarder sur une opération qui joue un rôle fondamental dans la plupart de ces algorithmes : la « fusion d'états ». Fusionner deux états d'un automate fini a pour effet de *généraliser* le langage reconnu. À condition d'être appliqué à bon escient (c'est-à-dire en évitant de *surgénéraliser*), c'est une clé possible du processus *d'induction* que doivent opérer ces algorithmes. Les intérêts et limites de cette opération sont longuement discutés dans ce chapitre. Ils sont notamment illustrés sur la tâche originale de reconnaissance automatique des schémas accentuels (« *stress patterns* » en anglais) mis en œuvre dans deux langues exotiques (le pintupi et le kwak'wala), à partir d'exemples de suites de syllabes accentuées. Cette application, qui fera plaisir aux linguistes, est une des rares, inspirée par les langues naturelles, qui figure dans les premiers chapitres du livre. Les schémas accentuels sont, en effet, caractérisables par un nombre restreint de

symboles et suivent des règles simples (supposées régulières au sens de la théorie des langages), ce qui les rend bien adaptés à ce type d'approche.

Le quatrième et dernier chapitre aborde les langages non réguliers (principalement algébriques ou « *context free* »), dans une perspective d'inférence grammaticale « empirique ». C'est celui qui devrait intéresser le plus les spécialistes du TAL, car il évoque des travaux en lien avec l'analyse syntaxique des langues naturelles. Avant de présenter quelques systèmes, les principes sur lesquels ils sont fondés sont étudiés. Intervient notamment alors le critère fondamental de *substituabilité*, également familier des linguistes, qui permet ici d'identifier les séquences ou les sous-structures pouvant être produites par le même symbole non terminal. Les systèmes eux-mêmes (Emile, ABL, Adios, CCM, DMV, U-DOP) sont souvent peu connus en dehors du domaine, et ont des capacités très variables. Le problème de leur évaluation donne d'ailleurs lieu à des développements détaillés. On peut regretter que des approches plus récentes, comme Mate ou MaltParser, qui visent à apprendre non pas une grammaire formelle, mais directement un analyseur syntaxique (un « *parser* ») à partir d'un corpus arboré, ne soient pas mises en regard de ces systèmes.

Une brève conclusion, enfin, récapitule l'ensemble et signale quelques éléments qui n'ont pu être développés. Il est dommage, par exemple, que certains travaux cherchant à intégrer au sein de l'inférence grammaticale des notions linguistiques (la sémantique, par exemple) n'aient pas pu être évoqués faute de place. Au final, le livre réalise un certain tour de force : il balaie un spectre très large de travaux souvent assez difficiles d'accès, en esquivant les preuves formelles pour se focaliser sur les principes sous-jacents. Ce parti pris lui permet de rester accessible, compact et néanmoins rigoureux. Comme la plupart des ouvrages de sa collection, il se veut plus une porte d'entrée d'un domaine qu'un exposé exhaustif. Avec ses schémas, ses exemples, ses « encadrés », ses résumés en fin de chapitre et sa bibliographie, il jouera parfaitement son rôle pédagogique d'initiateur, aussi bien que d'« aiguilleur » pour les lecteurs curieux d'en savoir plus. Il n'est pas certain que cet effort éditorial suffise pour autant à convertir les praticiens du TAL à l'inférence grammaticale. Mais le paysage scientifique ainsi dévoilé vaut néanmoins le détour.