Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Thierry POIBEAU. Machine Translation. *The MIT Press*. 2017. 285 pages. ISBN 978-0-262-53421-5.

Lu par **Claire LEMAIRE**

Université de Grenoble Alpes – LIDILEM/GETALP

Machine Translation *se propose, en quinze chapitres de six à vingt-six pages, de donner une vue d'ensemble sur les progrès réalisés en traduction automatique depuis la Seconde Guerre mondiale. L'auteur aborde le sujet sous quatre angles : par l'histoire de la traduction automatique, les méthodes utilisées pour développer des outils, les types d'évaluation de ces outils, et enfin par l'industrie de la traduction automatique.*

L'ouvrage s'ouvre sur la difficulté de la conception d'outils de traduction automatique et sur le rappel que ces outils sont prévus pour des textes techniques et informatifs et non pour des textes littéraires. L'auteur s'interroge ensuite sur l'acte de traduire, et sur ce que représentent de bonnes traductions. Il propose les cinq critères que sont le respect de la structure du texte source, de son ton, de son style et de ses idées, et la fluidité de la lecture. Par ailleurs, la qualité dépend également du destinataire et de l'usage du texte, or deux tendances coexistent en traductologie : soit le traducteur cherche à rester proche de la forme de la langue source, soit au contraire proche de la forme de la langue cible. Enfin, la difficulté d'une modélisation informatique réside dans la subjectivité des critères de qualité, dans l'ignorance des processus cognitifs humains à reproduire, ainsi que dans le manque de précision des langages naturels.

1. Histoire de la traduction automatique (TA)

L'historique présente tout d'abord une figure du triangle de Vauquois, qui décrit les différents niveaux de transfert d'une langue à une autre lors du processus de TA, et se poursuit par une présentation des différents types de systèmes. Reprenant les publications, entre autres de John Hutchins, l'auteur s'intéresse aux réflexions qui ont précédé la question de la TA, retrace l'évolution en Occident des concepts de langue universelle et de code numérique, ainsi que des premières machines qui ont suivi. Sont ensuite décrits les premiers essais de TA, par transfert direct, puis par transfert syntaxique avec les systèmes à base de règles, aux États-Unis et en Europe, sur une période qui part de 1949 avec le premier document précisant le problème de la TA, le fameux mémorandum de Warren Weaver, et qui se termine en 1959 avec

le rapport de Bar-Hillel et son questionnement sur le bien-fondé de travaux sur le sujet.

La dernière partie de cet historique relate tout d'abord la suppression d'une grande partie des budgets, dans le monde anglophone, due au rapport *Automatic Language Processing Advisory Committee* (ALPAC), en 1966, annonçant la quasi-impossibilité de traduire automatiquement, puis la période plus calme qui a suivi, au cours de laquelle se sont tout de même poursuivis des travaux dans différents centres de recherche et dans différentes sociétés commerciales. Cette période difficile pour le domaine se termine au début des années 1980, avec une recrudescence de l'intérêt pour la TA et une diversification des travaux et des méthodes.

2. Méthodes utilisées aujourd'hui pour développer des outils

Cette partie commence par aborder les méthodes *empiriques*, par opposition à la méthode *experte*. Alors que la méthode experte est fondée sur la modélisation des langues concernées, une méthode empirique repose sur un gigantesque corpus de phrases bilingues. Dans la première méthode empirique, la TA *basée sur l'exemple*, on entre un segment (un ou plusieurs mots) à traduire et, si un segment semblable est stocké dans le corpus bilingue, le système le retrouve et le propose. Si le système ne retrouve pas de segment semblable, il produit une traduction par analogie.

Dans une deuxième méthode empirique appelée TA *probabiliste* (ou *statistique*), on commence par une longue phase d'entraînement sur le corpus bilingue. On entre une phrase à traduire, le programme calcule, grâce à des modèles d'alignement, la phrase cible la plus probable. L'auteur décrit également les différentes approches de création et de traitement des corpus bilingues, puis les modèles d'alignement d'IBM, utilisés pour entraîner les systèmes, qui commencent par des probabilités de traduction au niveau lexical et qui se terminent par un réordonnancement des mots dans la langue cible.

La méthode experte, également appelée *par règles* (pour des raisons historiques), est aujourd'hui intéressante pour traduire des langues peu dotées, puisqu'il n'existe pas assez de données pour entraîner un système probabiliste. L'auteur propose Apertium qui est un logiciel libre adapté pour développer rapidement son propre système de TA, mais qui ne fonctionne bien que pour des langues proches syntaxiquement.

La troisième méthode empirique est le *deep learning*, également appelée *traduction neuronale*, car elle fait appel aux réseaux neuronaux profonds déjà connus en intelligence artificielle. À partir du corpus bilingue, on indique une correspondance exacte entre un segment source et un segment cible à obtenir en sortie. Ensuite, on prépare le segment source en définissant des caractéristiques (genre, lemme, etc.) pondérées et en attribuant un vecteur. Dans le moteur neuronal, on applique une fonction mathématique, puis on récupère le résultat que l'on décode. L'auteur précise que la prochaine étape en traduction neuronale sera donc de maîtriser ces caractéristiques et leur pondération, pour l'instant très obscures.

3. Types d'évaluation de ces outils

Les premières méthodes d'évaluation faisaient appel à des humains, soit des monolingues répondant par QCM à des questions de compréhension d'un texte automatiquement traduit, soit des traducteurs post-éditant ou jugeant directement les traductions. Puis ont suivi les méthodes d'évaluation automatique, qui comparent une traduction du système de TA à une traduction de référence produite par un humain. Si les deux textes sont strictement identiques, le score maximal est atteint. BLEU est la mesure la plus utilisée, NIST reprend l'idée précédente et accorde plus de poids aux groupes de mots plus rares. METEOR ajoute une recherche de mots porteurs de sens (et de leurs lemmes) ; meilleure, mais moins facile à utiliser, cette dernière méthode a moins de succès.

L'auteur reconnaît qu'aucune des méthodes automatiques n'évalue réellement la qualité des traductions, mais constate que les scores fournis donnent tout de même des indications sur le système de TA testé, soit par rapport à ses résultats précédents, soit par rapport aux autres systèmes de TA entraînés avec les mêmes données.

Depuis 2005, la conférence *Workshop on Machine Translation* (WMT) propose des campagnes d'évaluation. Les résultats des tests sont certes fonction des systèmes, mais également des langues : les plus riches morphologiquement, comme l'allemand et ses mots composés, sont pénalisées et les couples de langues proches sont favorisés.

4. Industrie de la TA

Les acteurs principaux sont Systran, Google, PROMT, IBM et Microsoft. Depuis sa création en 1968 (sous le nom de Latsec Inc.), Systran a été en partenariat avec les services du département de la Défense des États Unis, puis avec la Commission européenne. La firme, qui propose une solution gratuite en ligne *Systranet*, a été rachetée en 2014 par le Coréen CSLi, développeur des systèmes d'analyse vocale et de TA utilisés par les téléphones et tablettes Samsung. ProMT, le concurrent russe de Systran, vend des services de traduction dans des sites Web, payables au nombre de mots par mois, tout comme IBM avec sa plate-forme *IBM WebSphere*. Google a développé *Google Translate* initialement basé sur le système *Systran*, l'auteur met cependant en garde les entreprises sur l'absence totale de confidentialité de ce système qui conserve tous les textes sources des utilisateurs. Enfin, Microsoft propose *Bing Translator* dans ses logiciels. Pour conclure, l'auteur note que Google, mais aussi Apple et Facebook, rachètent des start-up du domaine et montent des centres de recherche en traduction neuronale recrutant actuellement des chercheurs au niveau mondial, puis cite les applications consommatrices de TA.

L'ouvrage se referme sur une TA pleine d'avenir et sur les tout derniers systèmes neuronaux capables de prendre en compte, au cours du processus, des phrases entières.

Un glossaire d'une quarantaine de mots du TAL et un index complètent le livre, ainsi qu'une bibliographie des sources d'inspiration de l'auteur, particulièrement détaillée et bienvenue pour quiconque souhaite se forger une petite culture dans le domaine.

Caroline BARRIÈRE. Natural Language Understanding in a Semantic Web Context. *Springer*. 2016. 318 pages. ISBN 978-3-319-41335-8.

Lu par **Thierry POIBEAU**

Lattice-CNRS

Le livre de Caroline Barrière se présente comme un point d'entrée pour les spécialistes du Web sémantique qui souhaiteraient s'initier au traitement automatique des langues (TAL), même si le contenu peut bien évidemment aussi intéresser un public plus large, ayant tout simplement un intérêt pour le TAL. L'autrice détaille avant tout les techniques d'extraction d'informations (c'est-à-dire l'extraction d'informations ciblées à partir de textes non structurés, pour remplir des bases de données structurées). Ce domaine est particulièrement intéressant pour le Web sémantique, qui a précisément pour but de produire des bases de connaissances et/ou d'utiliser des bases de connaissances dans le cadre d'applications « intelligentes », s'adaptant au contexte ou pouvant, par exemple, engager une forme de dialogue avec l'utilisateur. Du coup, le livre laisse volontairement de côté certains domaines importants du TAL, comme la recherche d'informations ou la traduction automatique, qui sont a priori *jugés moins pertinents dans ce contexte. L'ouvrage est ainsi homogène, cohérent et très progressif. Le titre est toutefois un peu large : ce qui est visé, c'est une compréhension locale et limitée du contenu textuel ; il existe d'autres approches de la compréhension, plus ambitieuses, mais évidemment beaucoup moins opérationnelles, qui ne sont pas prises en considération dans cet ouvrage.*

Le livre est divisé en quatre parties. La première partie est consacrée à la présentation des techniques de reconnaissance d'entités nommées dans les données textuelles. La deuxième porte sur la notion de corpus et la recherche de séquences (n-grammes) dans des textes monolingues ou bilingues. La troisième explore des questions d'annotation de corpus (annotation morphosyntaxique et syntaxique) ; sur cette base, la section présente ensuite différentes méthodes de calcul de similarité entre séquences textuelles, et enfin le processus qui consiste à relier les formes de surface aux formes normalisées que l'on peut trouver dans des ressources comme des thésaurus ou des référentiels (techniques dites de « liage d'entité » ou « *entity linking* »). Enfin, la quatrième partie aborde les liens entre syntaxe et sémantique, et la question des relations entre entités au sein du texte. Le livre comprend ensuite trois annexes : un aperçu du Web sémantique pour le lecteur qui ne maîtriserait pas ce domaine ; des informations sur les plates-formes et les outils de TAL disponibles ; des listes de relations typiques, souvent utilisées dans les applications de Web sémantique. Enfin, le livre comprend un glossaire de deux cents termes couramment utilisés en TAL.

Le contenu de l'ouvrage est très pédagogique. Chaque notion nouvelle est introduite, expliquée et replacée dans le contexte plus large de l'ouvrage. Plusieurs

algorithmes classiques du domaine sont présentés (sous forme de pseudo-code) et expliqués en détail. Le lecteur est constamment invité à vérifier les résultats et mener des expériences par lui-même (en utilisant par exemple les ressources détaillées dans l'annexe 3). Chaque chapitre est accompagné d'exercices permettant au lecteur de vérifier qu'il a bien compris le contenu, les enjeux et les problèmes susceptibles de se poser. Chaque chapitre inclut enfin des lectures complémentaires permettant d'aller plus loin que ce qui est présenté dans le livre. De ce point de vue, on peut dire que le livre est clair, bien écrit, et remplit parfaitement sa fonction ; il est progressif et amène le lecteur pas à pas à aborder des questions de recherche plus ambitieuses (de la recherche d'entités à la recherche de relations entre entités, en passant par la question du « liage » avec une base de données). La focalisation de l'ouvrage sur le domaine de l'extraction d'informations (recherche et analyse des entités du texte, recherche de relations entre entités en vue de compléter des bases de données structurées) semble aussi pertinente, même si cela amène l'autrice à laisser de côté de nombreux autres domaines qui auraient aussi pu être pertinents (comme la recherche d'informations sémantiques, qui entretient pourtant aussi des liens étroits avec le monde du Web sémantique).

Pour ce qui est des limites de l'ouvrage, on regrettera juste que l'accent soit essentiellement mis sur les techniques symboliques ou statistiques simples, et que les approches à base d'apprentissage artificiel soient le plus souvent à peine évoquées dans les lectures complémentaires, sans que ce choix ne soit véritablement expliqué. Les approches par apprentissage sont pourtant aujourd'hui la norme pour ce qui concerne la reconnaissance des entités nommées, et on regrettera ainsi que ni le format d'annotation BIO (*begin*, *inside*, *outside*), ni les algorithmes de reconnaissances de séquences (par exemple les CRF, *Conditional random fields*, pourtant omniprésents pour la reconnaissance des entités nommées) ne soient présentés. Quoi que l'on pense de ces techniques, elles ont acquis un poids tel qu'il semble difficile de les passer sous silence dans un ouvrage sur le TAL focalisé sur les techniques d'extraction d'informations. Dans le chapitre sur la similarité sémantique, la notion de plongements de mots (*word embeddings*) est rapidement évoquée, mais elle aurait aussi pu être davantage détaillée. Il est vrai que celle-ci a peut-être moins d'importance dans une perspective d'application dans le monde du Web sémantique que pour le TAL au sens large, encore que cela serait à prouver.

En conclusion, on appréciera un ouvrage très pédagogique et pertinent pour le lecteur souhaitant se former aux domaines abordés par l'ouvrage. On peut regretter le peu de détails concernant les techniques récentes d'apprentissage artificiel, mais le lecteur pourra s'appuyer sur les lectures complémentaires indiquées dans l'ouvrage (très complet de ce point de vue !), pour aller plus loin et s'initier par lui-même aux techniques les plus récentes. Notons enfin que l'ouvrage aborde un large choix de notions, dans la mesure où la reconnaissance des liens entre entités implique une analyse morphosyntaxique, syntaxique et même sémantique. Ce livre permet donc d'établir un pont solide entre traitement automatique des langues et

ingénierie des connaissances, et répond, à ce titre, à un besoin certain, que ce soit dans le monde universitaire ou dans le monde industriel.

Shay COHEN. Bayesian Analysis in Natural Language Processing. *Morgan & Claypool publishers*. 2016. 246 pages. ISBN 978-1-62705-873-5.

Lu par **Jose G. MORENO**

Université Paul Sabatier Toulouse III – IRIT

L'ouvrage est organisé en huit chapitres et une annexe. Il est d'un très haut niveau du point de vue formel et très riche en commentaires scientifiques. C'est aussi une source complète de références qui constitue un état de l'art de chaque sujet traité. Dans certains cas, quand l'auteur ne développe pas complètement un sujet, il indique des références précises pour trouver des informations complémentaires. Cependant, même si l'ouvrage est très complet, il requiert une connaissance élevée de l'inférence bayésienne dans son ensemble. Il est très adapté à des étudiants en doctorat sur le sujet ou des sujets plus proches de l'apprentissage automatique que du traitement automatique des langues (TAL). Le résumé en fin de chaque section ne résume pas forcément la section, et donne plus des informations pertinentes que manquantes. Le revers de ce livre est qu'il ne constitue pas un matériel pédagogique adapté pour un cours introductif de l'inférence bayésienne, alors qu'il sera clairement utile dans un cours avancé.

Le chapitre 1 est un rappel sur la probabilité et les statistiques relatives au TAL bayésien. Il aborde des concepts de base tels que les variables aléatoires, l'indépendance entre les variables aléatoires, l'indépendance conditionnelle et les valeurs attendues des variables aléatoires ; les statistiques bayésiennes sont présentées brièvement, en montrant la façon dont elles diffèrent des statistiques fréquentistes (ou statistiques classiques). La majeure partie de ce chapitre peut être omise si le lecteur possède une connaissance de l'inférence bayésienne. Cependant, il est recommandé de jeter un coup d'œil pour identifier les notations, revoir des concepts aussi bien en informatique qu'en statistique et identifier les exemples de TAL pour les chapitres suivants dans lesquels les concepts sont utilisés dans des contextes réels sans donner trop d'exemples.

Le chapitre 2 présente l'analyse bayésienne en TAL à l'aide de deux exemples : (1) le modèle d'allocation latente de Dirichlet (LDA) et (2) la régression bayésienne pour le texte. Ce chapitre donne également un aperçu général du sujet. L'histoire générative est présentée et sera utilisée dans différents chapitres jusqu'au dernier. Le modèle graphique (*plate notation*) est présenté pour le modèle LDA, mais n'est pas trop utilisé dans les chapitres suivants.

Le chapitre 3 traite d'une composante importante de la modélisation bayésienne : l'*a priori*. Les *a priori* les plus couramment utilisés dans le TAL bayésien sont présentés, tels que la distribution de Dirichlet, les antécédents non informatifs

(inappropriés et uniformes ou basés dans l'information de Fischer comme l'*a priori* de Jeffreys), la distribution normale et d'autres. Ces concepts sont primordiaux pour la suite des explications, en particulier la distribution de Dirichlet qui sera rappelée jusqu'à la fin du livre. Les positions données aux *a priori* sont d'une grande importance pour les modèles bayésiens et leur combinaison est possible dans des modèles hiérarchiques d'*a priori*.

Le chapitre 4 examine des idées qui regroupent les statistiques fréquentistes et bayésiennes à travers l'estimation de la distribution postérieure. Il détaille les approches permettant de calculer une estimation ponctuelle d'un ensemble de paramètres tout en maintenant une mentalité bayésienne.

Le chapitre 5 aborde l'une des principales approches d'inférence dans les statistiques bayésiennes : la chaîne de Markov Monte-Carlo (MCMC). Il détaille les algorithmes d'échantillonnage les plus couramment utilisés dans le TAL bayésien, tels que l'échantillonnage de Gibbs et l'échantillonnage de Metropolis-Hastings. Dans ce chapitre, le problème de convergence MCMC est abordé. La recommandation finale des auteurs est d'être prudent lors de la présentation des résultats avec la MCMC, car la convergence de la chaîne n'est pas une garantie et, dans la plupart des cas, il faut être très attentif pour ne pas manquer la bonne configuration du modèle.

Le chapitre 6 traite d'une autre approche d'inférence importante dans le TAL bayésien, à savoir l'inférence variationnelle. Il décrit l'inférence variationnelle du champ moyen, l'algorithme de maximisation des vraisemblances variationnelles et les méthodes d'échantillonnage. Ce sont des approches pour obtenir une inférence approximative. La différence principale entre ces méthodes est que la dernière est présentée comme un problème d'optimisation. Dans le chapitre 5, l'auteur rappelle que les modèles, type MCMC, ne font pas de l'optimisation sur une fonction unique, comme c'est le cas pour l'échantillonnage de Gipps, leur objectif consiste à trouver la chaîne qui correspond dans l'espace des possibilités. Dans ce cas, les itérations sont fixes sans garantie de convergence. Les méthodes d'inférence variationnelle sont une option viable pour atténuer les problèmes de convergence.

Le chapitre 7 traite d'une technique importante de modélisation en TAL bayésien, à savoir la modélisation non paramétrique. Les modèles non paramétriques comme le processus de Dirichlet et le processus de Pitman-Yor sont présentés. Ce chapitre est particulièrement intéressant pour les chercheurs des méthodes non supervisées. L'explication du processus de Dirichlet et ses relations avec l'algorithme k-moyennes sont claires et intéressantes.

Le chapitre 8 traite des modèles de grammaires formelles pour le TAL, comme les grammaires probabilistes non contextuelles et les grammaires synchrones. La façon de les cadrer dans un contexte bayésien est aussi abordée en détail, en particulier à l'aide de modèles tels que les grammaires d'adaptation, l'utilisation de hiérarchies dans le processus de Dirichlet et d'autres. Les lectures complémentaires recommandées par l'auteur sont une riche collection de références pour les chercheurs intéressés par les grammaires bayésiennes et leurs applications.

Commentaires

Pour résumer, il est important de préciser que ce livre est très technique et précis. Malgré sa qualité, il n'est pas recommandé aux débutants. Le lecteur doit avoir une bonne connaissance des modèles probabilistes pour pouvoir suivre l'auteur. Il ne faut pas s'attendre à des exemples explicatifs et il faut l'utiliser comme une référence technique, plutôt en apprentissage automatique qu'en TAL. Côté exercices, ceux-ci sont aussi très focalisés sur l'apprentissage automatique et ont peu ou rien à voir avec des exercices réels en TAL. En revanche, le livre comprend deux annexes qui fournissent des informations supplémentaires afin de faciliter sa lecture. Pour les novices, les annexes doivent être abordées avant de démarrer ce livre. Enfin, signalons que ce livre contient des conseils sur le plan théorique et pratique de l'inférence bayésienne. Par exemple, après une explication détaillée de l'inférence variationnelle, il énonce des méthodes récentes qui diminuent la charge théorique du développeur d'une nouvelle méthode sans propager des erreurs de différentiation (très fréquents parmi les débutants selon le chapitre 5).

Ce livre est recommandé aux chercheurs (confirmés ou pas) qui veulent vérifier leur état de l'art, trouver une explication à certains résultats, trouver des sources d'inspiration ou une notation formelle pour présenter un nouveau modèle. Comme l'auteur le souligne dans ces derniers mots, il existe aujourd'hui une énorme possibilité, pour les chercheurs intéressés, à profiter des avantages du TAL bayésien et des approches neuronales en TAL. Le formalisme existant dans le TAL bayésien est clairement déficient dans le TAL neuronal, mais la performance de ce dernier est plus que remarquable. Ce livre est une belle présentation du point de vue bayésien pour le TAL, et le plus probable est qu'il restera une référence pendant un certain temps.