Notes de lecture

Rubrique préparée par Denis Maurel

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

Satoshi SEKINE, Elisabete RANCHHOD, Named Entities : Recognition, classification and use, *John Benjamins publishing company*, 2009, 168 pages, ISBN 978 90 272 2249 7.

Lu par **Éric Charton**

Associé de recherche – Projet GITAN – École Polytechnique de Montréal

Cet ouvrage collectif se donne pour objectif de dresser un panorama des aspects variés de la recherche sur la reconnaissance d'entités nommées. À travers sept chapitres, le livre examine sur une période d'une quinzaine d'années un ensemble d'aspects liés à la reconnaissance d'entités nommées. Il examine également sous plusieurs angles l'application des techniques d'étiquetage et de repérage à des applications concrètes.

Ce livre présente une synthèse de la recherche sur les entités nommées (EN) qui couvre une quinzaine d'années. La période concernée est située par les éditeurs de l'ouvrage entre 1991 et 2006. Cette période est d'importance car son terme correspond à une période charnière, au cours de laquelle on a vu émerger de nouvelles méthodes d'étiquetage d'entités nommées utilisant notamment des ressources conceptuelles issues du Web, et que l'ouvrage passe complètement sous silence. Après un inventaire assez exhaustif de la problématique particulière de la reconnaissance d'entités nommées et de l'évaluation des systèmes (principalement en référence aux campagnes d'évaluation MUC, CONLL et ACE), viennent les sept chapitres du livre. Ils abordent l'étude des EN sous l'angle des systèmes à base de règles et des systèmes à base d'apprentissage automatique, mais à travers des angles thématiques très spécifiques. Ces spécificités font que l'ouvrage se veut plus un assemblage de descriptifs de recherches originales qu'un état de l'art et des méthodes d'étiquetage d'entités nommées.

Le premier chapitre aborde le problème de la normalisation des modèles utilisés dans les étiqueteurs à champs conditionnels aléatoires (CRF). Un modèle de détection, intitulé le LOP, est combiné d'après plusieurs modèles CRF. Des expériences sont menées avec ce modèle sur les corpus de CONLL 2003 à quatre classes.

Le deuxième chapitre envisage la question des entités nommées par leur rapport avec les conjonctions. Les auteurs étudient notamment la problématique d'une entité intégrant une connexion logique (comme par exemple dans la séquence « *Seshayaee*

*Paper **and** Boards Limited* ». Bien que le problème traité reste strictement cantonné à l'étude des entités à conjonctions, ce chapitre dresse également un panorama très exhaustif des études précédentes menées sur les systèmes à base de règles. Les auteurs proposent, pour résoudre la difficulté qu'ils étudient, d'entraîner plusieurs détecteurs d'EN avec des motifs de détection (Perceptron, arbres logistiques, K + proche voisins, Naives Bayes) et en comparent les résultats en utilisant un corpus financier qu'ils ont eux-mêmes mis au point.

Les trois chapitres suivants étudient la question de l'étiquetage des EN dans une perspective plurilinguistique. Le troisième chapitre soulève la question de la localisation des EN complexes contenues dans des textes rédigés en espagnol, en utilisant des règles (succession de noms, de prénoms, utilisation des prépositions exploitées). Le quatrième chapitre propose une méthode de détection des EN et de translittération pour le bengali. Trois modèles d'extraction à base de règles et un modèle à base de HMM sont appliqués à un corpus bengali mis au point par les auteurs. Le cinquième chapitre aborde la question des propriétés morphologiques et sémantiques des noms propres. L'idée est de définir une méthode de localisation d'une entité nommée de type nom propre dans un contexte multilingue. Une mise en perspective de la problématique est réalisée avec des exemples en serbe, français, polonais et anglais. Après une étude des mécanismes de dérivation, de morphologie et de structure des noms propres, une proposition de définition ontologique de noms propres est décrite en fin d'article. Cette formalisation vise à fournir une aide à la localisation de noms propres dans un contexte multilingue.

Le sixième et dernier chapitre étudie la question de la reconnaissance interlingue des EN. Après une intéressante étude du rôle joué par la reconnaissance d'entités nommées dans le système de gestion de flux d'actualité *NewExplorer*[1], les auteurs décrivent un algorithme d'extraction d'EN multilingue principalement fondé sur des motifs de détection. Suit la description des heuristiques complémentaires mises en place pour renforcer le système (variantes morphologiques, translittérations, détection de coréférences, méthodes de normalisation). On notera la description succincte d'une méthode de détection reposant sur les données fournies par des réseaux sociaux (LinkedIn et MySpace) très innovante et prometteuse, bien que peu détaillée.

Globalement, cet ouvrage présente des méthodes déjà assez anciennes pour un domaine dont les progrès ont été constants depuis 2006. La fenêtre temporelle concernée n'a pas permis de considérer les nombreuses propositions récentes exploitant des contenus encyclopédiques ou des ressources lexicales dans des étiqueteurs. La question de l'importance des EN dans de nouveaux domaines tels que la bio-informatique est également absente. Ce livre offre néanmoins un contenu varié et un ensemble de propositions bien présentées, qui permettront à un lecteur

1. http : //emm.newsexplorer.eu/

non spécialiste d'enrichir ses connaissances sur la question de l'étiquetage d'entités nommées.

Jian-Yun NIE, *Cross Language Information Retrieval*, Morgan & Claypol Publishers (Synthesis Lectures on Human Language Technologies), 2010, 125 pages, ISBN 9781598298635.

Lu par **Thierry POIBEAU**

Laboratoire LaTTiCe, CNRS

La collection « Synthesis lectures in human language technologies *» vise à fournir une introduction à un domaine de recherche (en environ 125 pages). Elle concerne donc des ouvrages faisant le point sur une question, destinés à des étudiants avancés ou des chercheurs qui souhaitent une bonne synthèse sur un sujet donné.* Cross-Language Information Retrieval *est issu de cette collection : l'ouvrage porte sur la recherche d'information «* cross-lingue *», c'est-à-dire mettant en jeu des documents de différentes langues, voire des requêtes et des documents de langues différentes. L'auteur, Jian-Yun Nie, a déjà publié de nombreux articles sur la question et ce livre constitue une bonne introduction à ce domaine de recherche.*

Contenu de l'ouvrage

Cross-Language Information Retrieval est composé de cinq chapitres. L'introduction est relativement longue (30 pages) et contient des rappels nécessaires sur la recherche d'information. L'auteur passe notamment en revue les principaux modèles de recherche d'information (modèles booléens, vectoriels, probabilistes, etc.). L'auteur survole ensuite certains problèmes posés suivant les langues considérées (racinisation ou décomposition des mots composés pour les langues indo-européennes, segmentation du chinois ou du japonais, etc.) et peut enfin aborder les questions posés par la recherche *cross-lingue* : faut-il plutôt traduire les requêtes ou les documents ? Faut-il passer par un langage pivot ? L'introduction se termine par un rapide survol des approches et des besoins. Si ceux-ci ne sont pas encore flagrants pour les applications grand public (le Web regorge de documents en langue étrangère, mais il n'est pas dit qu'un moteur de recherche *cross-lingue* serait tellement utilisé dans les faits), il est certain que des domaines comme la veille technologique ou l'analyse de brevets ont des besoins urgents et importants en la matière.

Le deuxième chapitre concerne les systèmes de traduction automatique : il s'agit d'un passage obligé vu leur importance pour traiter les aspects multilingues du problème. Le chapitre est censé porter sur les approches manuelles (*using manually constructed translation systems and resources*) : cet intitulé est quelque peu réducteur car peu de systèmes sont aujourd'hui mis au point manuellement. De fait, ce chapitre présente aussi le système de traduction mis au point par Google, qui n'a rien de manuel. Le contraste serait sans doute plus convaincant si l'auteur parlait de

systèmes à base de règles (souvent mises au point manuellement) *versus* systèmes à base de traitements statistiques calculés automatiquement. Mais, là encore, les oppositions ne sont pas si nettes et tendent à s'estomper[2]. La deuxième partie du chapitre porte sur le couplage d'un système de traduction avec un système de recherche d'information. De multiples problèmes peuvent se poser, de la présence de mots inconnus à la question de la traduction de la requête (souvent difficile vu que les mots sont la plupart du temps ambigus et apparaissent quasiment sans contexte du fait de la longueur très limitée de la plupart des requêtes).

Le troisième chapitre (*translation based on parallel and comparable corpora*) porte sur l'utilisation de corpus parallèles et/ou comparables pour la traduction. L'auteur présente les modèles statistiques utilisés en traduction automatique (modèles IBM notamment) et montre comment ceux-ci peuvent également être utilisés pour la recherche d'information *cross-lingue*. Les deux domaines sont en effet très liés dans la mesure où la recherche d'information peut être vue comme un problème de traduction. Néanmoins, les besoins sont parfois différents : le modèle IBM-1 est généralement considéré comme insuffisant pour la traduction, du fait de la non-prise en compte de l'ordre des mots. J.-Y. Nie montre, à l'inverse, que ce modèle peut être suffisant pour la traduction de requêtes dans la mesure où, dans ce contexte, la traduction ne vise qu'à chercher des équivalences entre langues sans prise en compte de la syntaxe. Le chapitre considère également des problèmes plus « pointus », comme la recherche d'équivalences entre des langues très différentes (recherche de translittération et/ou d'équivalents sémantiques, par exemple entre l'anglais et le chinois dont les systèmes d'écriture sont très éloignés).

Le chapitre 4 est plus court et décrit des méthodes moins standard en recherche d'information *cross-lingue*. Il est ainsi possible de procéder de manière interactive et de vérifier auprès de l'utilisateur que les documents retournés sont pertinents ou non, pour pouvoir ainsi relancer la recherche pour obtenir des résultats plus précis. Différentes stratégies sont aussi possibles pour essayer d'améliorer la traduction (multiplier les sources de connaissances, essayer différentes stratégies de traduction, etc.), chacune permettant une légère augmentation de la qualité des résultats obtenus.

Enfin, l'ouvrage se termine par un chapitre sur les perspectives de recherche de ce domaine très riche (*a look into the future : toward a unified view of monolingual IR and CLIR ?*). Comme le montre le titre, l'auteur défend l'idée que les problèmes posés par la recherche d'information monolingue sont largement identiques à ceux de la recherche d'information *cross-lingue*. En effet, même dans un cadre

2. L'auteur rappelle le triangle de Vauquois (1968) pour la traduction automatique, mais celui-ci n'est probablement plus d'actualité pour la plupart des systèmes. Dans l'ouvrage, le système de traduction censé symboliser une approche manuelle est celui de Systran. Or, si celui-ci a longtemps été exclusivement symbolique, il intègre maintenant une composante statistique importante.

monolingue, la question de la désambiguïsation et des équivalents sémantiques de la requête se pose. Il faut dans tous les cas déterminer le sens de la requête afin de fournir une réponse satisfaisante à l'utilisateur ; pour l'auteur, le processus d'analyse de la requête afin d'obtenir une représentation efficace peut, dans tous les cas, être assimilé à une traduction.

L'ouvrage se termine par une bibliographie très riche de onze pages, soit plus de cent soixante-dix références.

Commentaire

L'ouvrage est une très bonne introduction au domaine de la recherche d'information *cross-lingue*. Il fournit un bon aperçu de ce domaine très riche, en présentant les différentes facettes du problème, les solutions développées et les problèmes ouverts. La présentation est claire et les perspectives bien mises en avant. Il s'agit donc d'un bon ouvrage de synthèse qui peut intéresser un large public à condition d'avoir un minimum de connaissances en mathématiques afin de suivre l'exposé sur certains points, notamment pour les parties consacrées à la traduction automatique à partir de modèles statistiques.

C'est sans doute là la principale limite de tels ouvrages en 125 pages : le format ne permet pas d'entrer dans les détails et les développements sont forcément rapides, surtout quand ils concernent comme ici des aspects linguistiques (problèmes de l'ambiguïté des requêtes, de la correspondance entre différents types d'écritures, etc.) et techniques relativement vastes (modèles de traduction automatique, modèles de recherche d'information, couplage des deux).

Comme nous l'avons souligné, il s'agit malgré tout d'une très bonne synthèse, notamment pour le chercheur qui veut se faire une idée de l'état des recherches dans ce vaste domaine. La bibliographie fournit en outre toutes les pistes pour aller plus loin. On ne peut donc que recommander cet ouvrage qui constitue un bon point de départ pour découvrir la recherche d'information *cross-lingue*.

Jimmy LIN, Chris DYER, Data-Intensive Text Processing with MapReduce, *Morgan & Claypool Publishers*, 2010, 165 pages, ISBN 9781608453429.

Lu par **Clément DE GROC**

Syllabs – LIMSI-CNRS – Université Paris Sud

Le livre de Jimmy Lin et Chris Dyer est dédié à l'application du modèle de programmation MapReduce pour le traitement automatique des langues fondé sur de grandes masses de données. MapReduce est un modèle de programmation inventé chez Google permettant de traiter des ensembles de données volumineux sur un grand nombre de machines en parallèle. Les auteurs motivent et détaillent le fonctionnement du modèle puis introduisent différents patrons de conception autour de ce dernier. Ces patrons représentent des solutions

réutilisables pour des problèmes récurrents et permettent la réalisation d'algorithmes efficaces et passant à l'échelle. Ils sont mis en application dans le cadre de plusieurs tâches concrètes de recherche d'information et d'apprentissage statistique. L'ouvrage permet une compréhension du modèle en profondeur et la réalisation d'algorithmes efficaces en pratique.

Le phénomène « *big data* »

Le chapitre introductif motive l'utilisation de méthodes et algorithmes pour le traitement de grandes masses de données, notamment au travers de récents travaux menés par les grands noms de l'industrie de la recherche d'information. Mercer avait su, dès 1985, résumer les conclusions de ces derniers en une phrase : « *There is no data like more data* » (Il n'y a pas de meilleures données que plus de données). Partant du postulat que la prochaine décennie sera donc centrée sur le défi « *big data* », le livre poursuit sur la présentation détaillée du modèle de programmation MapReduce inventé chez Google.

Le modèle MapReduce

MapReduce est un modèle de programmation dérivé des langages de programmation fonctionnels et notamment des fonctions *Map* et *Reduce*. Partant d'un ensemble de données partitionné, un traitement indépendant est appliqué en parallèle sur chacune des partitions (fonction *Map*). Les résultats de ces traitements sont ensuite agrégés (fonction *Reduce*), puis renvoyés à l'utilisateur. L'application de ces fonctions sur une grappe de machines (*cluster)* de taille variable est alors abstraite, tout comme les détails (et potentielles difficultés) techniques dus à la parallélisation d'un programme informatique.

Les auteurs s'attachent dans le deuxième chapitre aux détails du modèle, que ce soit concernant l'algorithmique ou la gestion matérielle de clusters. Leur propos est clair notamment grâce à l'utilisation de schémas et d'exemples concrets à chaque introduction de nouveau concept. Lin et Dyer ont choisi Apache Hadoop[3] comme implémentation de référence mais restent généralement à un niveau d'abstraction élevé permettant de transposer les notions étudiées à d'autres implémentations. Les algorithmes sont écrits en pseudo-code, ce qui augmente leur lisibilité tout en permettant une mise en application rapide. Ce chapitre permet une compréhension en profondeur du modèle : où se trouvent les données, quand sont-elles transférées, quand sont-elles disponibles globalement ou localement.

Patrons de conception et cas d'utilisations

Le chapitre suivant présente différents patrons de conception (*design patterns*) pour le modèle MapReduce. Certains sont dédiés à une amélioration des performances (*In-mapper combining*) ou à ajouter de nouvelles fonctionnalités absentes de l'implémentation open source Apache Hadoop (*value-to-key conversion*). D'autres (*pairs, stripes)* permettent de modéliser l'apparition

3. http : //hadoop.apache.org/

d'événements joints (des cooccurrences par exemple) et offrent une alternative entre passage à l'échelle et performances. Enfin, le patron *order inversion* permet aux fonctions Reduce d'accéder à des valeurs intermédiaires avant de traiter les données qui ont généré ces valeurs. Cela permet, par exemple, de calculer la fréquence normalisée de termes dans un corpus en une seule tâche MapReduce, fournissant à la fonction Reduce le nombre de termes total et la fréquence d'un terme simultanément.

Les patrons sont introduits à chaque fois à partir d'exemples d'applications concrètes, pour la plupart issues du domaine de la recherche d'information et de la fouille de textes : calcul du nombre d'occurrences de termes, matrices de cooccurrences, temps moyen de session utilisateur à partir de traces (*logs*). Le dernier cas d'application présenté est la jointure – au sens des bases de données relationnelles – de deux jeux de données. Il fournit au lecteur une première idée du fonctionnement interne des langages développés comme surcouches décisionnelles de MapReduce tels que Apache Hive[4] et Pig[5].

Le chapitre 4 est dédié à la recherche d'information et centré sur la création d'un index inversé compressé. Nous sont donc rappelés quelques rudiments de la recherche d'information, à savoir : le parcours (*crawling*) du Web, la compression d'index (codage différentiel, codage de gamma, codage de Golomb) et le rapatriement des résultats en réponse à une requête utilisateur. Le chapitre, didactique, permet de voir comment, itérativement, nous pouvons améliorer un algorithme d'inversion d'index naïf pour assurer son passage à l'échelle et son efficacité. Enfin, les auteurs mettent en avant le fait que MapReduce ne répond pas aux besoins des requêtes utilisateurs, notamment en raison de son temps de réponse trop long.

Le chapitre 5 s'oriente alors vers les graphes et leurs algorithmes (parcours en largeur et PageRank) et nous présente comment modéliser un algorithme itératif à l'aide de plusieurs appels à MapReduce. Nous constatons, dans ce chapitre, une des limites majeures du modèle, à savoir l'impossibilité de partager une structure globale entre les fonctions Map distribuées. Dans le cadre d'un algorithme de plus court chemin, il faut donc, par exemple, recourir à un parcours en largeur coûteux en remplacement de l'algorithme de Dijkstra.

Le dernier chapitre avant la conclusion est dédié à la famille d'algorithmes espérance-maximisation (EM) et leur application aux modèles de Markov cachés (HMM) avec l'algorithme Forward-Backward. Après un rappel complet de ces notions, les auteurs montrent que l'algorithme s'intègre naturellement au modèle en réalisant une itération de l'algorithme par tâche MapReduce. La fin du chapitre décrit brièvement plusieurs tâches connexes faisant appel aux HMM ou aux

4. http : //hive.apache.org/

5. http : //pig.apache.org/

algorithmes EM comme la traduction statistique, l'alignement mot à mot et l'entraînement de modèles logistiques (maximum d'entropie, champs aléatoires conditionnels).

Le livre se conclut sur une courte synthèse des limites du modèle MapReduce et notamment sur le fait que MapReduce s'avère limité au traitement *batch* (le jeu de données doit être disponible dès le départ dans son intégralité) et ne permet donc pas d'apprentissage actif, en ligne ou de traitements en temps réel. Une ouverture vers de récentes propositions de modélisations alternatives palliant certaines limites du modèle clôt finalement ce dernier chapitre.

Jimmy Lin et Chris Dyer nous proposent une introduction très agréable à MapReduce au travers d'une démarche progressive et d'applications concrètes. Leur ouvrage s'adresse à un public d'informaticiens, éventuellement novices dans les domaines de la recherche d'information, du traitement automatique des langues ou de l'apprentissage statistique. Ils proposent une vision pratique et détaillée de l'application de différents algorithmes sur le modèle MapReduce. Cette dernière complètera assurément une vision plus technique du modèle qu'il faudra acquérir dans un livre dédié à l'implémentation choisie.

Nizar Y. HABASH, Introduction to Arabic Natural Language Processing, *Morgan & Claypool publishers*, 2010, 167 pages, ISBN 9781598297959.

Lu par **Abdelmajid BEN HAMADOU**

Laboratoire MIRACL, Institut Supérieur d'informatique et du Multimédia, Sfax, Tunisie

---

L'arabe, par sa forme d'écriture (de gauche à droite, signes diacritiques optionnels, pas de majuscule, jusqu'à quatre variantes pour la même lettre...), sa phonologie (vingt-huit consonnes, trois voyelles courtes et trois voyelles longues...), sa morphologie (dérivation à partir de racines, existence d'infixes en plus des préfixes et suffixes, agglutination...) et sa syntaxe (assez libre, absence possible des désinences...) est considéré parmi les langues difficiles à appréhender sur le plan du traitement automatique. Les problèmes de recherche que ces caractéristiques posent, nécessitent une formalisation assez poussée et des outils robustes d'analyse. Cet ouvrage, avec ses huit chapitres, ses cinq annexes et sa bibliographie apporte un éclairage intéressant de ces problèmes surtout pour les étudiants-chercheurs et ingénieurs débutants dans le domaine du TALN arabe. Il s'intéresse à l'arabe moderne, dit aussi standard, sans pour autant négliger ses dialectes. Chaque chapitre aborde aussi bien la facette linguistique que celle technique au travers de la présentation des tâches TALN associées les plus importantes. Aussi, pour aider les lecteurs, notamment ceux non arabophones, à mieux assimiler les concepts étudiés, un parallèle avec d'autres langues, notamment l'anglais, est effectué chaque fois qu'il est nécessaire et tous les exemples en arabe sont translittérés.

Après un bref aperçu de l'arabe standard et dialectal dans le chapitre 1, le chapitre 2 détaille les spécificités de l'écriture arabe au niveau de son alphabet, des signes diacritiques qui accompagnent cet alphabet et des problèmes de codage des caractères. Les tâches de voyellation automatique et de lecture optique des caractères arabes ont été abordées avec les réels défis qu'elles posent.

Le chapitre 3 traite la phonologie et l'orthographe arabes en exposant les traits distinctifs des phonèmes et la manière de les transcrire graphiquement. Les tâches de translittération des noms propres, de correction orthographique des textes arabes et de reconnaissance et synthèse de la parole ont été présentées d'une manière assez synthétique.

La morphologie de l'arabe a occupé la part la plus importante dans cet ouvrage (deux chapitres) eu égard à sa richesse, son caractère agglutinatif qui est en étroite liaison avec la syntaxe et la multitude des tâches TAL associées. Le chapitre 4 commence par une présentation détaillée de la terminologie utilisée pour décrire la morphologie arabe. Le lecteur appréciera, à ce propos, la clarification des confusions que l'on a tendance à effectuer dans l'usage de certains termes comme affixes, proclitiques, racines, radicaux, modèles et patrons morphologiques. Ensuite, il détaille la structure du mot arabe et les différentes formes dans lesquelles il peut apparaître dans un texte (agglutinée, dérivée, fléchie,..). Enfin il aborde les transformations morphophonologiques qui peuvent affecter le mot (assimilation, gémination...) et leur impact sur l'orthographe. Le chapitre 5 discute des tâches TAL associées à la morphologie arabe : analyse et génération morphologiques, *tokenization* et étiquetage grammatical. Une bibliographie est donnée pour chacune de ces tâches. Bien qu'elles soient quasiment limitées à la communauté anglophone, celles-ci donnent une idée assez complète des travaux accomplis pour l'arabe.

Le chapitre 6 aborde la syntaxe en montrant comment la langue arabe partage avec les autres langues plusieurs notions (verbe, nom, sujet, complément d'objet...), mais possède des notions spécifiques comme *Al-Idhafa*[6] et *Al-Tamyiz*[7]. Ce chapitre commence par tracer un croquis des structures syntaxiques, puis présente les efforts de création de corpus arborés arabes (*Arabic Treebanks*).

Le chapitre 7 donne un bref aperçu de la sémantique arabe et introduit le projet « *Arabic Wordnet* » et les efforts de création de ressources pour les tâches d'extraction des informations.

---

6. *Idhafa* se traduit par « annexion ». Quand un nom *mudâf* est suivi de son complément *mudâf ilayhi,* on dit qu'ils sont en annexion. Exemple : *moudirou Al-madrasati,* (le directeur de l'école). Dans ce cas, le nom ne peut pas être défini par l'article *Al,* car il l'est par son complément.

7. *Tamyiz* se traduit par « spécification » ou « discernement ». C'est un complément invariable qui suit un nom ou une phrase pour apporter plus de sens et lever l'ambiguïté. Exemple : *Indi ritlou* ***Asalin*** (j'ai une livre de **miel).**

Le denier chapitre aborde d'une manière assez synthétique la question de la traduction automatique pour introduire les concepts de base associés et donner une comparaison sommaire, dans le contexte de la traduction, entre l'arabe et trois autres langues, le chinois, l'anglais et l'espagnol, sur les plans de l'orthographe, la morphologie et la syntaxe.

Cet ouvrage comporte cinq annexes qui listent les principaux consortiums, réseaux, conférences, livres, dictionnaires, ressources disponibles et outils s'intéressant au traitement automatique de l'arabe.