Notes de lecture

Rubrique préparée par Denis Maurel

Université François-Rabelais de Tours, LI (Laboratoire d'informatique)
denis.maurel@univ-tours.fr

Denis MAUREL, Franz GUENTHNER, Automata and Dictionaries, *King's College Publications*, 2006, 240 pages, ISBN 1-904-987-32-X.

par Aurélie NEVEOL

Equipe CISMeF (Rouen) et National Library of Medicine (Bethesda, Etats-Unis)
aurelie.neveol@insa-rouen.fr

Automata and Dictionaries présente l'utilisation d'automates et de transducteurs comme formalismes de représentation pour des dictionnaires électroniques de langue naturelle. Après un tour d 'horizon sur les enjeux de la construction de tels dictionnaires et sur les applications du TAL utilisatrices de ce type de ressources, les auteurs rappellent les définitions formelles et les algorithmes liés à la manipulations des automates et transducteurs. Chaque étape est illustrée par des exemples concrets constitués par des dictionnaires jouets. Cet ouvrage constitue une bonne introduction technique à la manipulation d'automates dictionnaires et renvoie le lecteur souhaitant approfondir le sujet vers les publications adéquates. On peut cependant regretter que le premier chapitre, plus général, n'ait pas la clarté des suivants.

Résumé de l'ouvrage

Le premier chapitre se fixe pour triple objectif d'exposer les difficultés liées à l'élaboration de dictionnaires pour une langue donnée, les raisons pratiques motivant cette entreprise et les spécificités engendrées par la réalisation de cette tâche au format électronique. Les auteurs partent du constat qu'un flou théorique demeure sur le contenu exact des dictionnaires à élaborer. Ils doivent encoder à la fois les propriétés lexicales des unités de la langues, les liens qui les unissent ainsi que les règles qui régissent leurs associations avec d'autres unités – ce, de manière exhaustive. En pratique, les modalités de recueil et de représentation de ces informations ne sont pas unanimement établies : notamment, une étape aussi fondamentale que le découpage des *unités lexicales* pose problème. Cependant, au delà de l'aspect descriptif de la langue, les dictionnaires sont nécessaires dans de nombreuses applications du Traitement Automatique de la Langue telles que l'analyse syntaxique, la recherche d'information, la traduction automatique ou encore

la correction orthographique. Comme pour les autres applications, les dictionnaires disponibles sont néanmoins insuffisants – en terme de couverture et d'informations encodées, pour obtenir les performances souhaitées. Ainsi, il est nécessaire de poursuivre le travail amorcé pour élaborer des dictionnaires électroniques aussi exhaustifs que possible, utilisables par l'ensemble de ces applications. Pour ce faire, il convient d'apprécier l'ampleur de la tâche du fait du nombre de langues à décrire, du nombre de domaines de spécialité à couvrir et de la diversité des applications auxquelles les dictionnaires sont finalement destinés. Les auteurs estiment que ce chantier doit donner lieu à un effort collaboratif rassemblant les chercheurs des différentes communautés scientifiques concernées.

Le deuxième chapitre illustre le fonctionnement et l'intérêt des automates à états finis comme structure de stockage des dictionnaires électroniques. Pour ce faire, un dictionnaire simple composé des jours de la semaine est utilisé.

Le troisième chapitre présente des définitions formelles autour des automates et transducteurs. Les notions introduites sont illustrées à l'aide de l'exemple donné au chapitre précédent.

Les chapitres quatre et cinq définissent et explicitent les notions d'automate déterministe et minimal.

Les chapitres cinq et six synthétisent les algorithmes de construction des automates et transducteurs introduits par différents chercheurs. Un algorithme particulier est détaillé et implémenté sur un dictionnaire comportant quelques entrées choisies pour illustrer les étapes clés.

Finalement, le chapitre huit aborde des questions liées à la complexité des algorithmes, l'espace de stockage ou le temps d'exécution nécessaire à leur mise en oeuvre.

Les chapitres 4 à 7 se terminent par quelques exercices d'application.

Commentaire

La clarté des chapitres 2 à 8 fait défaut au début de l'ouvrage et l'objectif pédagogique s'en trouve atteint. La longueur des phrases, le style empesé ainsi que l'absence de définitions et de références pour certaines des notions introduites nuisent à la lisibilité de cette partie. Le public novice visé (étudiants) pourra éprouver quelques difficultés à appréhender le sujet traité avec la vision globale souhaitée par les auteurs.

Dans un ouvrage traitant à la fois d'automates et de dictionnaires électroniques, certaines ouvertures sur d'autres travaux connexes dans ces deux domaines auraient pu avoir leur place. Par exemple, l'affirmation en préface qu'aucune méthode

statistique n'a permis de construire de ressources linguistiques conséquentes[1] porte à controverse à la lumière du travail récemment entrepris à l'INRIA sur l'élaboration du Lexique des Formes Fléchies du Français[2]. Une discussion contrastant les deux approches (essentiellement manuelle *vs.* statistique) et l'utilisabilité des ressources produites (DELA *vs.* L*efff*) aurait eu un intérêt à la fois pédagogique et pratique. Dans une moindre mesure, une rapide comparaison entre l'utilisation des automates appliqués à des textes en langue naturelle par opposition à d'autres formes de textes telles que les séquences génomiques aurait permis une vue plus complète de la portée des outils décrits.

On peut également espérer que le nombre non négligeable de coquilles figurant dans le « tirage préliminaire » que j'ai consulté ont pu être corrigées dans la version finale.

Dans l'ensemble, malgré quelques écueils, l'ouvrage fournit une présentation simple, claire et illustrée de l'utilisation d'automates et de transducteurs pour la construction de dictionnaires électroniques. Les principes de base de la théorie des automates sont exposés, et nombre de références plus détaillées sont proposées aux lecteurs souhaitant approfondir le sujet. Dans les chapitres 6 et 7, les algorithmes les plus complexes de l'ouvrage sont exposés de manière tout à fait abordable. Automata and dictionaries constitue une bonne introduction à la problématique des dictionnaires électroniques sous forme d'automates.

[1] p.8 : « [T]here has not been any substantial progress in creating (…) sophisticated linguistic databases (e.g. monolingual (…) lexica (…)) on the basis of statistical approaches (…) »

[2] le L*efff*, disponible sous licence libre avec les publications associées sur http://www.lefff.net.