Notes de lecture

Rubrique préparée par Denis Maurel

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

Sébastien HARISPE, Sylvie RANWEZ, Stefan JANAQI, Jacky MONTMAIN. Semantic Similarity from Natural Language and Ontology Analysis. *Morgan & Claypool publishers*. 2015. 236 pages. ISBN 978-1-627-05446-1.

Lu par **Laurie ACENSIO**

Lexiane Formation

Parmi la diversité des mesures sémantiques issues de la littérature scientifique, cet ouvrage distingue deux catégories de méthodes d'analyse : une approche à base de corpus de textes en langage naturel et une approche à partir d'une base de connaissances. Ces techniques d'apprentissage dites topologiques fondées sur les ontologies sont liées à l'émergence des applications sémantiques. Le calcul de similarité distributionnelle à base de corpus, plus traditionnel dans son approche, connaît également un regain d'intérêt inhérent à la disponibilité de vastes ressources textuelles. Par conséquent, le recours à des mesures sémantiques devient une préoccupation essentielle dans le processus de raisonnement et d'interprétation.

Contenu et structure

À travers quatre chapitres, les auteurs ont principalement développé les mesures de similarité fondées sur la structure de la connaissance du domaine et sur une analyse statistique de corpus. Cet ouvrage s'inscrit dans la continuité de la thèse soutenue par l'auteur principal Sébastien Harispe de l'équipe KID[1] du laboratoire LGI2P[2] à l'École nationale supérieure des mines d'Alès, en 2014.

Le premier chapitre présente les notions théoriques et applicatives des mesures sémantiques selon les approches disciplinaires (intelligence artificielle, psychologie, sciences cognitives). De nombreuses références sont citées dont nous pouvons relever les plus matures : les travaux de Tversky et de Collins dans le domaine de la psychologie cognitive et ceux de Lesk et de Rada dans le domaine informatique. Ces travaux démontrent une nécessité et une volonté d'évaluer la proximité conceptuelle sur des réseaux sémantiques et la mémoire sémantique humaine bien avant l'émergence des applications du Web sémantique de ces dernières années. Les modèles de calcul de la similarité sémantique sont utilisés dans divers contextes

1. Knowledge representation & Image analysis for Decision.
2. Laboratoire de génie informatique et d'ingénierie de production.

applicatifs avec pour objectif d'utiliser des connaissances supplémentaires pour raisonner sur leurs données.

Les auteurs terminent par une classification schématique des méthodes existantes à savoir : les mesures fondées exclusivement sur les corpus textuels (unités de langage), les bases de connaissances (concepts) et les méthodes hybrides qui tiennent compte à la fois des relations dans le graphe (chemins et directions du concept) et des propriétés des concepts (spécificités et profondeur du concept).

Les auteurs ont choisi de focaliser les deux chapitres suivants principalement en distinguant deux approches de calcul : des unités lexicales *via* l'analyse de corpus de textes en langage naturel et les concepts et instances extraits d'une base de connaissances hiérarchisée (par exemple, une ontologie).

Le deuxième chapitre se focalise particulièrement sur les méthodes de sémantique distributionnelle fondées sur une analyse des unités de langage de corpus authentiques et contextualisés. Les méthodes de calcul de proximité sémantique les plus répandues comme la LSA (*Latent Semantic Analysis*), la ESA *(Explicit Semantic Analysis)* et le HAL (*Hyperspace Analogue to Lanuage*) sont détaillées à travers plusieurs exemples de matrices. Ces méthodes sont essentiellement statistiques (fréquences, cooccurrences des mots, représentations vectorielles) et par conséquent dépendantes de la nature et de la taille du corpus qui est considéré comme un espace sémantique représentatif. Un approfondissement en annexe traite des cooccurrences de termes, capturées dans un corpus au moyen d'une technique de décomposition de matrices standard (décomposition en valeurs singulières). *A contrario*, les méthodes à base de connaissances sont fondées sur la connaissance lexicale préalablement définie par une intervention humaine.

Le troisième chapitre constitue la partie le plus argumentée de l'ouvrage en se concentrant sur les mesures fondées sur une structuration de la connaissance sous forme de graphes sémantiques, plus couramment appelés ontologie. Contrairement aux techniques de raisonnement déductif fondées sur des faits, les capacités de l'inférence des ontologies relative à un domaine dédié permettraient de résoudre les incertitudes et les ambiguïtés relatives à certaines applications dont notamment les systèmes de recherche d'information, de recommandation ou de traduction automatique.

Le positionnement des auteurs est de considérer l'ontologie comme un graphe sémantique, dont les nœuds sont des concepts, et qui est restreint à une relation hiérarchique (is-a) commune à la majorité des ontologies. Bien que la démarche la plus intuitive soit d'utiliser la longueur des chemins pour mesurer la distance entre des concepts, des travaux récents démontrent que tous les chemins ne sont pas sémantiquement cohérents. Cela implique, qu'il n'y a pas d'unicité de chemins et que plusieurs chemins candidats sont envisageables, ce qui peut résulter d'un graphe cyclique mais également acyclique. Par la suite, deux niveaux sont principalement distingués dont les mesures qui évaluent le degré de similarité entre deux concepts et par groupe de concepts. Pour chacune des parties, des tableaux récapitulatifs simplifiés sont présentés après chaque section.

Le chapitre conclut sur les approches hybrides en proposant l'exemple d'une ressource textuelle et ouverte comme l'encyclopédie libre de Wikipédia. Ce corpus a l'avantage d'être continuellement mis à jour contrairement à d'autres ressources lexicales de grande taille (par exemple WordNet).

Le quatrième chapitre aborde l'évaluation des mesures sémantiques. Parmi les critères les plus fréquemment cités dans la littérature on trouve l'exactitude, la précision, la robustesse des résultats, la complexité algorithmique, et les propriétés mathématiques et sémantiques. Cependant, les auteurs soulignent un manque de protocole d'évaluation malgré le lancement récent de la campagne STS[3]. Ils notent également que les approches à base de corpus ou à base de connaissances posent des problèmes de stockage souvent spécifiques à un domaine donné.

En fin d'ouvrage, des annexes pratiques proposent un aperçu global des techniques de modélisation statistiques avec un focus particulier sur l'algorithme de décomposition de la valeur singulière, et sur les modèles de langues qui s'appuient sur les n-grammes et les réseaux de neurones.

La dernière annexe présente les outils principalement issus de la communauté scientifique classés selon les définitions propres à l'ouvrage.

Commentaires

L'ouvrage est davantage approprié pour être abordé comme un inventaire parmi la variété des approches des mesures sémantiques. Bien que disparate dans ses principes théoriques et ses résultats d'expérimentations, il démontre l'intérêt de cette thématique au sein de la communauté scientifique à travers de nombreuses références bibliographiques. Ce domaine de recherche relativement récent implique, par nature, un manque de consensus scientifique mais révèle un potentiel grandissant dans l'amélioration des modèles sémantiques.

Il est à noter que les deux premiers chapitres constituent une introduction pédagogique indispensable pour aborder cette problématique. Le schéma proposé en fin du deuxième chapitre propose un aperçu global intéressant pour différencier les méthodes. Cependant, les chapitres suivants ne suivent pas cette classification et sont plus confus dans la structuration des enchaînements.

Il semble que les auteurs aient voulu clairement différencier les méthodes à base de corpus de celles à base de connaissances. Cependant, les résultats d'expérimentations semblent divergents et ne révèlent pas d'une manière significative la pertinence d'une méthode par rapport à une autre. Par exemple, la famille des modèles statistiques en annexe aurait été plus compréhensible si elle avait été présentée en introduction de l'ouvrage pour faciliter la comparaison et démontrer ainsi les performances de chacune des familles des modèles (temps de calculs, qualité des résultats, facilité d'employabilité).

3. Semantic Textual Similarity Campaign.

L'évaluation applicative est restreinte à des expérimentations dans des contextes bien identifiés (bioinformatique), elle est donc peu significative. De ce fait, le choix de la méthode à utiliser se fera selon l'interprétation du lecteur. De même, les problématiques récurrentes du TAL liées à la négation, à l'antinomie, et aux contraintes des ressources lexicales (multilingues, par exemple) sont peu évoquées.

Jacqueline LÉON, *Histoire de l'automatisation des sciences du langage.* ENS Éditions. 2015. 218 pages. ISBN : 978-2-84788-653-5.

Lu par **Thierry POIBEAU**

Laboratoire LATTICE, CNRS

L'ouvrage de Jacqueline Léon aborde en neuf chapitres et environ 220 pages les débuts de l'analyse automatique des langues. La période couverte s'étend pour l'essentiel de l'immédiat après-guerre jusqu'aux années 1960, même si certains chapitres évoquent des périodes plus récentes. L'ouvrage couvre assez largement le domaine, puisque les chapitres traitent de sujets aussi variés que la linguistique appliquée, la linguistique de corpus, la traduction automatique ou plus généralement le traitement automatique de la langue.

Dans le premier chapitre, Jacqueline Léon défend l'idée que la traduction automatique a ouvert un nouveau champ de recherche, en rupture avec la linguistique de l'époque, mais dans la continuité des « sciences de la guerre » définies comme « l'interaction entre sciences de l'ingénieur et sciences fondamentales, comprennent notamment les mathématiques, la logique, la physique, les neurosciences, l'acoustique, et les sciences nouvellement apparues que sont la cybernétique et la théorie de l'information ».

Le deuxième chapitre montre comment les problèmes dans le domaine de la traduction automatique ont entraîné une réorientation des recherches vers l'analyse syntaxique. Ce champ de recherche, rapidement autonome, renoue en partie avec les liens rompus de l'analyse linguistique traditionnelle. Jacqueline Léon montre enfin l'évolution des recherches dans les années 1960, partant de l'analyse syntaxique proprement dite pour se tourner progressivement vers la compréhension de texte et plus globalement l'intelligence artificielle.

Le chapitre 3 est l'occasion d'un détour vers la linguistique appliquée. Jacqueline Léon évoque des pistes permettant de cerner les liens entre avancées technologiques et enseignement des langues dans la période de l'après-guerre.

Le chapitre 4 examine le statut du terme « théorie de l'information » dans l'après-guerre. Jacqueline Léon montre que le terme est souvent cité dès les années 1950, mais avec de multiples acceptions, dans des cadres très différents. À travers l'étude de ce terme particulier, il s'agit de voir comment une notion est reprise, modifiée et réinterprétée par différents champs scientifiques. Ce chapitre est aussi l'occasion d'examiner en détail qui sont les chercheurs qui jouent le rôle de « passeur » entre domaines scientifiques, et comment ceux-ci s'influencent les uns les autres.

Le chapitre 5 examine le cas des néo-bloomfieldiens et leur rapport à la notion de formalisation. Ce chapitre est notamment l'occasion d'examiner les recherches de Harris aux États-Unis et de voir les liens entre l'approche chomskyenne et les travaux de Harris de l'époque.

Le chapitre 6 examine d'autres traditions linguistiques. Il reprend des études antérieures de l'autrice consacrées, par exemple, aux travaux menés à partir des années 1950 à Cambridge au sein du CLRU (Cambridge Language Research Unit) autour de Margaret Masterman. La traduction russe est aussi abordée à travers les travaux de Mel'čuk. La deuxième partie de ce chapitre montre comment la traduction automatique fait apparaître de nouveaux objets linguistiques comme les mots composés et leur formalisation dans différents types de travaux menés en France à l'époque.

Le chapitre 7 porte justement sur la tradition française et décrit donc une histoire peut-être plus connue des lecteurs de la revue TAL. Il rappelle en tout cas les débuts de l'ATALA, ceux de la revue qui s'est appelée successivement « *La Traduction Automatique* » puis « *TA Informations* », et la création des centres CETA (Centre d'étude pour la traduction automatique) avec deux centres, à Paris et Grenoble, mais le premier sera rapidement fermé. Il rappelle aussi le parcours de Maurice Gross, qui tient une place incontournable dans le paysage français de l'époque.

Le chapitre 8 revient aux travaux de Harris qui ont effectivement connu une grande prospérité en France. Deux types de travaux sont particulièrement intéressants à ce titre : ceux de Gardin dans le domaine de la documentation automatique, poursuivis sous d'autres formes par Maurice Gross au LADL, d'une part, et, d'autre part, les travaux de Michel Pêcheux en analyse du discours.

Le chapitre 9 porte sur les sources anglo-saxonnes de la linguistique de corpus que l'autrice classe en différents courants, notamment « *corpus-driven* » (Halliday, Sinclair) *versus* « *corpus-based* » (Quirk, Leech). Le tournant statistique des années 1990 en traitement des langues est juste esquissé, et mis en rapport avec le courant lié aux corpus dans le monde anglo-saxon dans les années 1960.

Le livre se termine sur une conclusion, une bibliographie très complète et deux index (des noms et des notions).

Commentaires

L'ouvrage de Jacqueline Léon est une étude fouillée, précise et extrêmement bien documentée de l'automatisation des sciences du langage dans les années 1950 et 1960. La couverture de l'ensemble du domaine, ainsi que la mise en perspective des recherches menées dans différentes régions du monde rendent la lecture très attrayante.

La description est particulièrement intéressante quand elle met en relation les liens entre automatisation et analyse linguistique traditionnelle. L'automatisation a en partie tourné le dos à l'analyse linguistique, avant d'y revenir régulièrement au fur et à mesure des difficultés rencontrées. Par exemple, Jacqueline Léon montre bien comment l'échec relatif des premiers systèmes de traduction automatique a

poussé à aller vers une automatisation de la syntaxe et ainsi à renouer progressivement avec un domaine classique de l'analyse linguistique. À l'inverse, la linguistique de corpus permet de faire évoluer la linguistique générale en prenant mieux en compte les aspects quantitatifs.

L'étude est convaincante car très bien documentée et repose sur des années de recherche de la part de l'autrice. Le lecteur pourra parfois être étourdi par les noms d'organismes, de systèmes, d'acteurs du domaine, mais ceci est inévitable si l'on veut appuyer le propos sur des données précises. Globalement, l'ouvrage est l'occasion d'une réflexion sur le statut des recherches qui sont menées dans le cadre de l'automatisation des langues, sur le rapport entre linguistique et automatisation, et surtout l'occasion de renouer avec les origines du domaine, trop souvent ignorées. Nous recommanderons donc chaudement la lecture de ce livre à tout chercheur en traitement automatique des langues, quel que soit son domaine de spécialité. Au-delà, l'ouvrage touchera tous les lecteurs intéressés par l'automatisation des sciences.