Notes de lecture

Rubrique préparée par Denis Maurel

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

Martha PALMER, Daniel GILDEA, Nianwen XUE, Semantic Role Labeling, *Morgan & Claypool Publishers*, 2010, 91 pages, ISBN 9781598298314.

Lu par **Isabelle TELLIER**

Lifo, université d'Orléans

Ce livre porte sur l'étiquetage des rôles sémantiques, une tâche à l'articulation entre syntaxe et sémantique consistant à annoter les arguments d'un prédicat en fonction de leur relation avec lui. Il aborde aussi bien les théories linguistiques fondatrices que les ressources existantes, ainsi que les techniques utilisées pour apprendre à annoter automatiquement un corpus. Il est illustré de nombreux exemples et s'appuie sur l'expérience de ses auteurs.

Introduction

L'étiquetage des rôles sémantiques (*Semantic Role Labeling* ou SRL en anglais) est une tâche identifiée depuis longtemps mais elle a connu un grand développement ces dernières années, notamment pour avoir fait l'objet de plusieurs compétitions lors de conférences CoNLL récentes. La tâche consiste à associer à des constituants en position d'arguments d'un prédicat la nature de leur relation avec ce prédicat. Les relations possibles sont du type : « Agent », « Patient », « Thème », « Bénéficiaire », « Localisation », etc. Le livre est une exploration synthétique et pédagogique des problèmes que pose cette tâche et des solutions qui ont été adoptées pour y répondre.

Contenu de l'ouvrage

Le premier chapitre porte sur la notion de rôle sémantique elle-même et montre qu'elle constitue une phase clé de l'articulation entre syntaxe et sémantique. Plusieurs linguistes, parmi lesquels Fillmore, Jackendoff, Dowty et Levin, se sont attachés à préciser cette notion. Les difficultés sont nombreuses : la distinction entre arguments et modifieurs d'un verbe est parfois problématique et la liste des rôles sémantiques possibles qui y sont associés ne fait pas consensus. Si l'identification des « Agents » ne pose en général pas trop problème, d'autres rôles sont plus sujets à interprétations. Distinguer un « Patient » d'un « Thème » peut ainsi s'avérer assez délicat. Le chapitre illustre, *via* de nombreux exemples, les différentes conceptions possibles de la notion de rôle sémantique et les différents critères et niveaux de

granularité envisagés pour caractériser et classifier ces rôles. Il constitue une introduction claire et stimulante au domaine.

Le deuxième chapitre, plus bref et plus pragmatique, s'intéresse aux ressources lexicales effectivement disponibles qui mettent en œuvre l'une ou l'autre de ces théories. Pour la langue anglaise, il en recense trois : FrameNet, VerbNet et PropBank. FrameNet s'appuie explicitement sur les « frames sémantiques » de Fillmore. Comme leur nom l'indique, ces « frames » sont fondées sur des critères sémantiques bien plus que syntaxiques, et contiennent des informations riches et détaillées. VerbNet, lui, s'inspire des travaux de Levin. Il se présente comme une hiérarchie structurée de classes de verbes aux comportements similaires, à laquelle s'ajoutent une correspondance explicite entre rôles sémantiques et constructions syntaxiques ainsi que des liens avec d'autres ressources (Wordnet, OntoNotes, FrameNet et PropBank). PropBank, enfin, est de nature un peu différente. C'est une version enrichie du Penn TreeBank avec des annotations sur les constituants en fonction de leurs relations avec les différents verbes de chaque phrase. Ces annotations sont beaucoup plus neutres linguistiquement que les précédentes : elles se contentent de numéroter les arguments sans leur associer d'étiquette sémantique précise et d'identifier éventuellement les modifieurs. PropBank était en effet essentiellement destiné, à l'origine, à servir de corpus de référence pour des systèmes d'apprentissage automatique. Le chapitre se clôt sur un tableau récapitulatif qui résume les propriétés de chacune de ces ressources et sur l'évocation du projet en cours, Semlink, qui vise à jeter des passerelles entre elles.

Le chapitre suivant est consacré aux travaux exploitant des méthodes d'apprentissage automatique pour l'étiquetage de rôles sémantiques. La tâche est alors généralement reformulée comme une séquence de deux tâches de classification : d'abord celle consistant à distinguer les constituants qui sont des arguments des autres, ensuite celle consistant à leur associer la bonne étiquette. Intelligemment, le chapitre ne détaille pas les techniques d'apprentissage automatique elles-mêmes, mais se concentre plutôt sur les « *features* » (traits, caractéristiques) qu'elles intègrent pour parvenir à leur fin. Sans surprise, ce sont les catégories « *part of speech* » des constituants, celles de leur tête ou encore les chemins qui les relient au verbe qui sont les plus utiles. Les mesures d'évaluation sont évoquées et l'impact des structures disponibles (suivant qu'un arbre d'analyse syntaxique complet est disponible et correct ou non et suivant la théorie qu'il implémente) sur les résultats est également discuté.

Enfin, le quatrième et dernier chapitre aborde le problème de l'annotation en rôles sémantiques dans un contexte multilingue. Il cherche ainsi à répondre à des questions du type : peut-on facilement transférer un étiquetage d'une langue à une autre ? Peut-on aligner les annotations de deux langues différentes ? Dans quelle mesure les rôles sémantiques sont-ils spécifiques d'une langue donnée ? Ces différentes problématiques sont illustrées par l'exemple du couple anglais-chinois, qui a fait l'objet de nombreuses expérimentations. Une section, courte mais

originale, est également consacrée à l'étiquetage des rôles des prédicats non verbaux.

Conclusion

Le livre est compact, agréable à lire et bien documenté. Il ne se présente pas comme exhaustif mais doit plutôt être vu comme un point d'entrée vers les travaux les plus récents du domaine. Les auteurs en sont eux-mêmes des acteurs reconnus, leurs références sont visiblement très à jour. Les chercheurs français intéressés par cette tâche sont encore peu nombreux, ce livre constituera donc pour tous ceux qui souhaitent s'y initier une excellente introduction.

Maciej PIASECKI, Stanisław SZPAKOWICZ, Bartosz BRODA, A Wordnet from the Ground Up, *Oficyna Wydawnicza Politechniki Wrocławskiej*, 2009, 222 pages, ISBN 978-83-7493-476-3.

Lu par **Izabella THOMAS**

Centre de recherche en linguistique et TAL L. Tesnière, Université de Franche-Comté, Besançon

---

L'ouvrage de M. Piasecki, S. Szpakowicz et B. Broda est consacré à la description du processus de construction d'une ressource équivalente à Princeton WordNet pour le polonais. Deux postulats sous-tendent ce processus. Premièrement, le réseau lexical polonais doit être construit à partir de zéro, en évitant d'être inadéquatement calqué sur une langue différente (en l'occurrence l'anglais). Deuxièmement, il n'est pas possible de construire une ressource linguistique de haute qualité de façon entièrement automatique. En revanche, le travail de linguiste peut être largement facilité par le développement de procédés et d'outils automatiques lui suggérant, de la façon la plus exacte possible (similarité sémantique, relation lexico-sémantique particulière, point d'attachement au réseau existant) d'éventuels élargissements du réseau par de nouvelles unités lexicales. Ainsi, le processus décrit dans l'ouvrage est centré sur trois thématiques parallèles : construction manuelle du réseau lexical « de base », conception des outils d'aide aux linguistes, développement des procédés pour l'acquisition semi-automatique de nouveaux candidats à l'élargissement du réseau. Bien que le travail entrepris s'inspire constamment d'expériences déjà connues, force est de constater qu'il engage nécessairement une adaptation des procédés existants. Adaptation, d'une part, à la spécificité d'une langue hautement flexionnelle à l'ordre syntaxique libre telle que le polonais, et, d'autre part, au manque de ressources secondaires dédiées au polonais.

Le projet décrit dans le livre est né en 2005 de la nécessité de créer une ressource lexicale pour le polonais, équivalente aux « wordnets » développés depuis les années 90 pour la plupart des langues européennes (et au-delà) sur le modèle de Princeton WordNet (PWN). Ces réseaux constituent souvent la principale ressource lexico-sémantique dans de nombreuses recherches et applications en TAL

(désambiguïsation lexicale automatique, recherche d'informations mono et multilingues, traduction automatique, extraction d'informations, etc.). Ils sont, par conséquent, devenus indispensables dans le paysage du TAL. Les motivations, les objectifs et les postulats de base concernant la construction du plWordNet sont décrits dans le chapitre 1.

Bien que le réseau polonais s'appuie sur le modèle de PWN et son concept de synset, sa construction, contrairement à une majorité d'autres projets (y compris EuroWordNet), prend un chemin particulier : pour rendre compte de la spécificité de la langue polonaise, les auteurs défendent la nécessité de construire le réseau lexical polonais à partir de zéro, c'est-à-dire non pas en traduisant puis en ajustant le PWN au polonais, mais en s'appuyant uniquement sur des ressources monolingues, afin d'obtenir un réseau lexical totalement fiable sur le plan linguistique et de très haute qualité. Cette décision est justifiée par les problèmes rencontrés lors d'un essai de traduction de PWN en polonais, lequel a démontré l'inadéquation entre des structures hyperonymiques de haut niveau, un manque d'équivalents lexicaux (soit en polonais, soit en anglais), l'introduction de concepts artificiels et non lexicalisés dans une langue.

La construction du réseau est subdivisée en deux étapes : la première, entièrement manuelle, mais fondée sur un grand corpus (IPI PAN, 254 millions de mots), consiste à développer un réseau de base *(core plWordNet)* composé d'environ 16 500 unités lexicales et 9 000 synsets. Le réseau de base est ensuite utilisé comme une des ressources exploitées pour l'expansion semi-automatique, expansion qui constitue la seconde étape du développement. L'ensemble du réseau construit à l'aide de ces deux méthodes est composé de 27 000 unités lexicales et de 17 700 synsets.

Le chapitre 2, le plus « linguistique » de l'ensemble du livre, décrit les difficultés liées à la création du réseau de base et introduit les notions centrales pour son développement : celle d'unité lexicale (LU), le « synset » et les relations lexico-sémantiques (RLS) qui relient les unités (et les « synsets ») entre elles. Les RLS (synonymie, antonymie, hyperonymie/hyponymie, etc.) sont redéfinies à l'aide de tests (présentés en annexe 1). Les synsets sont construits pour trois catégories grammaticales : les noms, les verbes et les adjectifs. Le réseau obtenu est richement caractérisé par plusieurs paramètres : le nombre de LU par synset, le nombre de synsets composés de *n* LU, le taux de polysémie par LU, le nombre des instances pour différents RLS, etc.

Puisque le développement manuel d'un réseau lexical s'avère très coûteux, les auteurs présentent le développement de méthodes semi-automatiques (chaque proposition doit être validée par un linguiste) d'enrichissement du réseau lexical. Celles-ci concernent d'une part le calcul de similarité sémantique entre les LU (chapitre 3), et d'autre part, l'extraction des occurrences des RLS (chapitre 4). Aucune méthode n'est exclue *a priori* : sont considérées aussi bien les méthodes

fondées sur les règles (décrites manuellement ou extraites automatiquement), que les méthodes purement statistiques et probabilistes. Chaque procédé analysé est introduit par un état de l'art, puis adapté à une tâche spécifique et évalué, soit de façon automatique, soit manuellement, lorsque les ressources existantes ne permettent pas l'évaluation automatique. Aucun des procédés appliqué séparément ne donne de résultats considérés comme valables (le seuil « psychologique » étant fixé à 50 % de réponses directement utilisables par un linguiste). Par conséquent, une solution hybride est proposée. Elle se fonde sur la combinaison des résultats de plusieurs algorithmes et se voit complétée par une méthode appelée activation de la zone d'attachement (*activation–area attachement*), qui a pour objectif d'indiquer l'emplacement possible d'une nouvelle LU dans un réseau hyperonymique déjà existant. Les résultats ainsi obtenus sont finalement évalués à plus de 60 % de LU pour lesquelles au moins une des propositions est directement utilisable par un linguiste. Un outil « convivial » d'aide (*Word-Net Weaver*) permettant, entre autres, de visualiser les suggestions générées par l'ensemble des procédures automatiques est conçu pour faciliter le travail du linguiste.

Une évaluation générale, proposant une comparaison avec les réseaux développés pour d'autres langues, est proposée dans le chapitre 5. Les résultats sont prometteurs, surtout en ce qui concerne le procédé d'expansion semi-automatique. En plus d'accélérer de cinq à six fois le processus de construction, elle permet d'améliorer la qualité du réseau lui-même, en suggérant des erreurs dans l'organisation des relations d'hyperonymie, ainsi qu'en complétant les relations hyponymiques. Les points faibles identifiés proviennent du manque de corpus suffisants et d'outils d'analyse syntaxique efficaces pour le prétraitement du corpus. Certaines limitations des outils d'extraction sont également mises en cause. Les procédés d'enrichissement automatique sont développés uniquement pour les unités nominales et concernent exclusivement la relation d'hyperonymie, laissant de côté tout autre type de RLS.

L'intérêt de ce livre, riche en informations pour les chercheurs intéressés par le développement de réseaux lexicaux, réside en plusieurs points. Non seulement il s'agit d'une première publication concernant une langue slave, et donc typologiquement différente de l'anglais, mais elle permet d'appréhender, de façon complète et didactique, toute la problématique liée à la construction d'un réseau *ex nihilo*. Le parti pris entièrement assumé de la qualité au détriment, d'une part, de la vitesse de construction, et, d'autre part, de sa compatibilité avec les réseaux existants pour d'autres langues (le problème de l'alignement de synsets avec d'autres « wordnets » existants est identifié, mais n'est pas approfondi), permet de mieux entrevoir l'intérêt de la collaboration entre linguistes « purs » et ceux développeurs d'outils informatiques Cette collaboration résulte dans la création d'un outil permettant d'associer le meilleur de chacune de ces approches, au service de la qualité.