Coupler syntaxe et sémantique dans une même base de connaissances linguistiques

Jean-Philippe Prost

Coupler syntaxe et sémantique dans une même base de connaissances linguistiques

Jean-Philippe Prost[1]
(1) Laboratoire Parole et Langage (LPL), Aix-Marseille Université, France
Rapport de recherche, mars 2021
Jean-Philippe.Prost@univ-amu.fr

RÉSUMÉ

Cet article soutient qu'une approche holistique de la modélisation computationnelle de la connaissance linguistique faciliterait le traitement automatique du langage naturel. Nous nous concentrons sur la question de la combinaision homogène de sources hétérogènes de connaissances linguistiques telles que syntaxe et sémantique lexicale au sein d'une unique structure de graphe. Nous présentons une architecture de réseau linguistique multi-couches qui aborde le problème. Nous proposons en particulier de connecter une couche de grammaire par dessus un réseau lexico-sémantique existant pour le français. La solution que nous présentons s'appuie sur les Grammaires de Propriétés pour modéliser la couche de grammaire comme une sorte de réseau de contraintes. Le réseau résultant doit permettre le traitement intégré et simultané des connaissances syntaxiques et sémantiques, contrairement à leur traitement séquentiel dans une architecture habituelle en cascade.

ABSTRACT

Merging Syntax and Semantics in a Natural Language Knowledge Graph

This paper argues that a holistic perspective on the computational modelling of linguistic knowledge would ease natural language processing. We focus on the problem of homogeneously combining heterogeneous sources of linguistic knowledge such as syntax and lexical semantics within a unique graph structure. We introduce a Multi-Level Language Network (MLLN) architecture to address the problem. Such a structure contributes to bridging the gap between syntax and semantics, and we propose to plug a syntax layer on top of an existing lexico-semantic network for French to support our claim. The solution we introduce models the syntax layer as a kind of constraint network with Property Grammar. The resulting network is expected to enable the integrated and simultaneous processing of syntactic and semantic knowledge, as opposed to their sequential processing through the usual pipeline computing architecture.

MOTS-CLÉS : Base de connaissances, Réseau linguistique, Grammaires de Propriétés, JeuxDeMots, Syntaxe, Sémantique.

KEYWORDS: Knowledge graph, language network, Property Grammar, JeuxDeMots, Syntax, Semantics.

1 Introduction

Depuis des décennies la pratique courante en TAL consiste à modéliser et traiter le langage naturel le long de dimensions distinctes (lexique, syntaxe, sémantique, etc.), typiquement à travers des

architectures en cascade, conformément à des stratégies de résolution de problèmes dites de *diviser pour mieux régner*.

Cette approche en cascade présente l'inconvénient majeur de ne pas permettre, par essence même, de traitement intégré sur plusieurs dimensions, telles la syntaxe et la sémantique. cette intégration offrirait clairement de pouvoir raisonner sur la base de l'ensemble des connaissances disponibles, y compris sur les interactions entre connaissances relevant de dimensions différentes.

Cependant, les représentations de connaissances syntaxiques ne sont généralement que peu compatibles, sur la forme, avec les connaissances d'ordre sémantique, ou pragmatique, pour n'évoquer que ces deux exemples. Le traitement automatique des langues, que ce soit en compréhension ou en production, bénéficierait donc grandement d'un modèle computationnel de représentation des connaissances qui permette cette intégration des dimensions.

Nous pensons qu'une approche holisitique du langage naturel et de sa modélisation computationnelle est possible, qui repose sur une représentation homogène, au sein d'une unique structure, les connaissances présentes sur toutes les dimension linguistiques.

Nous pensons également que la modélisation computationnelle des connaissances linguistiques relatives à la grammaire doivent, et peuvent, au même titre que d'autres dimensions linguistiques et connaissances du monde, faire l'objet d'une mémorisation à long terme, dans une base de connaissances abstraites qui profitent à toutes sortes de traitements et d'utilisations. Contrairement à bon nombre d'approches qui se concentrent sur l'acquisition une fois pour toute d'une grammaire, généralement à partir de données annotées, cette base doit être struturée de telle sorte qu'elle puisse être évolutive et puisse bénéficier d'améliorations continues dans le temps.

Dans cet article nous présentons une architecture préliminaire de Réseau Linguistique Multi-Couches (RLMC) qui s'attaque au problème de la représentation homogène d'informations hétérogènes pour la constitution d'une base de connaissances holistique sur le langage naturel. Nous décrivons en particulier la structure d'une couche de grammaire, que nous proposons de brancher par dessus un réseau lexico-sémantique. Pour la couche lexico-sémantique nous nous appuyons sur JDMRezo (**?**), le réseau lexico-sémantique pour le français.

2 État de l'art

Données linguistiques et Bases de Connaissances Les connaissances linguistiques représentées dans les grandes Bases de Connaissances et autres ressources sont essentiellement organisées et structurées le long des différentes étapes de l'architecture en cascade traditionnelle. Les ressources référencées dans LLOD [1], par exemple, concernent séparemment différentes dimension linguistiques : lexiques et dictionnaires (par ex. WordNet, WordNet-RDF, BabelNet, Wiktionary) pour les mots et la sémantique lexicale ; terminologies de domaines spécifiques, ontologies et bases de connaissances sur la connaissance du monde (par ex. YAGO, DBPedia) ; corpus annotés pour la syntaxe ; et quelques autres pour les méta-données et autres sortes de ressources.

LLOD met en évidence l'absence totale de ressource grammaticale abstraite. Seuls des corpus annotés syntaxiquement sont répertoriés, mais aucune base de connaissances qui rende compte de connaissances grammaticales abstraites pour le language. Aucune ressource non plus qui interface

1. http ://linguistic-lod.org/

réellement syntaxe et sémantique lexicale à un même niveau d'abstraction. Si l'intention de LLOD va clairement dans le sens d'une meilleure interopérabilité des ressources disponibles sur les différentes dimensions linguistiques, l'alignement des connaissances qui permettrait leur réelle intégration, et donc un modélisation holistique du langage, fait encore défaut.

Les SAR-Graphs introduisent un type de ressource différent, qui inclue des données syntaxiques, et est liée à différentes ressources sémantiques telles que WordNet, BabelNet, et YAGO. Les SAR-Graphs sont décrits comme *des graphes de relations associées sémantiquement* (*graphs of Semantically-Associated Relations*) (**?**). Ils abordent la question de l'intégration de différents aspects des connaissances linguistiques dans une structure homogène appelée un *réseau de langage* (*language network*) (**??**). Les SAR-Graphs intègrent sémantique lexicale et relations sémantiques sur des faits et des événements à travers l'utilisation de patrons syntaxiques appris à partir de textes parsés automatiquement. Les relations quant aux faits et événements sont extraites de base de connaissances telles Freebase (**?**)[2] . Pour chaque relation sémantique de haut niveau (c'est-à-dire qui modélise de la connaissance du monde comme des faits et des événements) trouvée dans une base donnée, un SAR-Graph fusionne toutes les structures en dépendance parsées automatiquement qui correspondent au patrons syntaxiques extraient (séparément) pour cette relation. Par exemple, dans (**?**) le SAR-Graph pour la relation *Prize_Arwarding_Event*⟨*Winner, Prize_Name, Prize_Area, Year*⟩ couvre la phrase "Mohamed ElBaradei won the 2005 Nobel Prize for Peace on Friday" avec la structure (partielle) illustrée en Figure 1. Les patrons d'extraction syntaxique sont appris avec le système DARE d'extraction

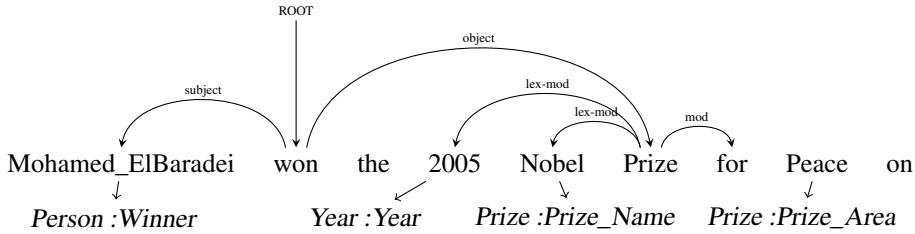

FIGURE 1 – Exemple de SAR-Graph, sans les relations lexico-semantiques)

de relation (**??**), distinct de la ressource principale.

Néanmoins, les SAR-Graphs n'incluent toujours pas de niveau de connaissances grammaticales abstraites. En l'occurrence la grammaire est contenue dans le système DARE à travers l'utilisation de MINIPAR, un parseur en dépendances à base de principes (**?**).

Les réseaux syntaxiques complexes L'étude de la syntaxe des langues naturelles à travers des *réseaux syntaxiques complexes* (**?**) s'appuie sur des structures de graphe de relations de dépendances entre mots ou formes de mots (**?**). Les mots (lemmes) sont des noeuds et les relations de dépendances des arcs. Le terme de *dépendance* doit être pris ici dans le sens du formalisme des Grammaires de dépendances (**??**). La construction d'un tel réseau syntaxique nécessite l'analyse préalable en dépendances de différents corpus, manuellement et/ou automatiquement. La littérature ne semble rapporter aucune tentative de représenter dans ces réseaux d'autres formalismes grammaticaux, comme les grammaires syntagmatiques (**?**). De même, ces réseaux ne semblent pas non plus intégrer de relations lexicales ou sémantiques.

2. Maintenant incluse dans le Google Knowledge Graph.

Les réseaux sémantiques Parmi la faune de réseaux sémantiques WordNet, probablement l'un des plus utilisés d'entre eux, est un réseau lexical pour l'anglais qui s'appuie sur la notion de *synset*, qui peut être vue comme une sorte de concept. WOLF (**?**) en est une version française, et EuroWordnet (**?**) une version multilingue. BabelNet (**?**) est un autre réseau lexical multilingue, construit à partir de co-occurrences de termes dans Wikipedia. Le Réseau Lexical du Français (RLF) (Polguère 2014) (**?**) est une ressource lexicale du français fondée sur la notion de fonction lexicales telle que définie par Igor Mel'čuk. La ressource, construite essentiellement manuellement, comprend environ 10000 vocables.

JDMRezo est un réseau lexico-sémantique libre construit et validé semi-automatiquement, d'une part de façon collaborative (crowdsourcing) à travers une couche de différents jeux sérieux [3], et d'autre part grâce à différents mécanismes d'inférence automatique et d'acquisition à partir de ressources externes. Le réseau, de taille continuellement croissante, est constitué actuellement de plus de 14 millions de noeuds, plus de 300 millions de relations et plus de 150 types de relations [4], ce qui en fait de très loin le réseau lexico-sémantique libre le plus étendu au monde. Les noeuds sont des termes, des concepts et des informations symboliques. Les relations sont notamment d'ordre lexical, morphologique, pragmatique, logique, ontologique, etc.

3 La couche de grammaire

Parmi les cadres formels pour MTS tels que les Grammaires de Propriétés (PG) (**?**) la grammaire d'un langage naturel peut être représentée par un graphe (V, E), où les sommets de V modélisent des catégories morpho-syntaxiques, et les arcs dans E modélisent des relations syntaxiques unaires ou binaires entres sommets. PG pré-définit la sémantique de 8 relations (Constituance, Dépendance, Accord, Linéarité, Obligation, Unicité, Exigence, Exclusion) (**??**) appelées *Propriétés*.

Ainsi, par exemple, l'interprétation dans la théorie des graphes de la propriété de précédence linéaire (ou Linéarité) NP :D $\prec$ N, qui établit que dans un syntagme nominal (NP) le Déterminant (D) précède le Nom (N) (dans le sens de lecture), correspond à la relation typée $\prec$ dans la Figure 2. On notera que dans cette représentation les structures classiques, syntagmatiques et en dépendances, correspondent simplement à un type de relation particulier (*Constituency* et *Dependency*). Dans le cas de la structure syntagmatique une partie droite de règle de réécriture est représentée par un noeud de type *aggregat*, lui-même relié par une relation de type *rewrites* (*réécrit*) à un noeud représentant la partie gauche.

Le graphe représenté en Figure 2 illustre la décomposition de la règle de réécriture NP $\rightarrow$ D N$(w1)$ en relations typées, dont les types correspondant aux propriétés de Grammaire de Propriétés.

3.1 Construction d'une première couche de grammaire

? a montré comment extraire une grammaire GP partielle à partir d'un corpus annoté en arbres syntagmatiques. Les propriétés de Dépendance et d'Obligation qui manquent dans (**?**) peuvent dorénavant être aisément extraites d'un corpus tel que le French Treebank actuel (**?**), annoté à la fois en dépendances et en structures syntagmatiques. Précisons que dans ce cas la sémantique que nous

3. http://jeuxdemots.org
4. D'après les valeurs affichées sur http://jeuxdemots.org en date du 16 février 2020.

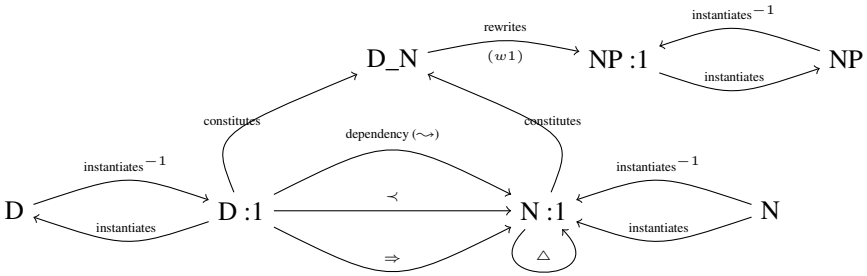

FIGURE 2 – Graphe de grammaire (NP → D N)($w1$)

adoptons pour la propriété d'Obligation correspond à la tête de syntagme, ou à la catégorie destination d'une dépendance.

Une fois la grammaire PG de corpus extraite, **?** a montré comment la compiler. Cette compilation inclut notamment de charactériser chacune des règles, c'est-à-dire de vérifier la satisfaction ou la violation de chaque propriété (i.e. relation) pour chaque règle de la grammaire. Il est évident que dans le cas présent l'opération serait circulaire, puisque l'ensemble des propriétés a été établit précisément à partir des règles de réécriture observées sur corpus. Nous pouvons donc partir du principe que toutes ces propriétés sont trivialement satisfaites pour le corpus qui a servi à l'extraction de la grammaire.

De même que **?**, nous augmentons ces propriétés observées avec des propriétés issues de règles de réécritures théoriques mais non observées sur corpus. L'intérêt principal de ces compléments théoriques est d'augmenter artificiellement la couverture de la grammaire, notamment avec des règles qui s'avèrent seulement partiellement grammaticales en ceci qu'elles ne satisfont qu'un sous ensemble des propriétés établies par observation sur corpus. Par exemple, si la grammaire observée se réduisait au seul graphe $G_{\}}$de la Figure 2, une règle théorique complémentaire pourrait être la règle factice (R2*)(NP → D N N). Lorsqu'on soumet (R2*) à la vérification de satisfaction des propriétés dans $G_{\}}$il en résulte un ensemble P_{R2*} de propriétés satisfaites (notamment les Linéarités entre D et chacun des deux N), et un ensemble P_{R2*} de propriétés violées (notamment l'Unicité de la tête de syntagme N). Ces règles partiellement grammaticales offrent comme intérêt majeur, lorsque la grammaire correspondante est utilisée à des fins de parsage, de permettre des analyses robustes et approchées d'énoncés partiellement grammaticaux (**?**).

Le caractère "violée" d'une propriété, et donc d'une relation dans notre graphe de grammaire, est représenté par une pondération de la relation concernée de valeur arbitrairement négative, tandis que le caractère "satisfaite" est, lui, représenté par une valeur positive. La relation `rewrites` de réécriture, entre le noeud aggrégat représentant la partie droite de règle et le noeud de partie gauche, est elle pondérée d'un score significatif de grammaticalité, conformément à **?**. La formulation du calcul de ce score a déjà fait l'objet d'études approfondies (**??**). Elle pourra l'être encore, de façon à inclure notamment la prise en compte non seulement de relations syntaxiques mais également lexico-sémantiques. Néanmoins, cet aspect de la discussion dépasse le cadre de cet article.

3.2 Ancrage syntaxe–sémantique

Un ancrage relativement naturel de la couche de grammaire peut se faire avec JDMRezo, grâce aux parties du discours associées aux termes (relation `r_pos`). Les relations syntaxiques sont alors de types spécifiques qui correspondent pour l'essentiel aux propriétés de GP, mais sont représentées de façon strictement identique aux relations sémantiques. Idem pour les noeuds, les catégories morpho-syntaxiques utiles à la couche de grammaire étant une extension du domaine de valeurs du type `n_pos` qui, dans JDMRezo, permet de définir les parties du discours. L'ancrage se ramène alors principalement à un travail de mise en correspondance des valeurs de catégories morpho-syntaxique, entre d'un côté le ou les schémas d'annotation utilisés dans les corpus qui servent de base pour l'extraction de la grammaire, et les valeurs utilisées dans la couche sémantique pour le type `n_pos`.

4 Conclusion

Nous avons présenté une architecture de réseau linguistique multi-couches qui permet une représentation homogène, au sein d'une même structure, de connaissances linguistiques en provenance de différentes dimensions telles la syntaxe et la sémantique. Cette architecture s'appuie sur une structure de graphe, compatible avec un réseau lexico-sémantique tel que JDMRezo. L'intérêt d'une telle architecture tient principalement dans l'homogénéisation de connaissances linguistiques hétérogènes. Il tient également dans ce qu'elle permet de constituer une base de connaissances qui porte à la fois sur les relations lexico-sémantiques entre termes et concepts, et sur les relations syntaxiques abstraites qui constituent la grammaire de la langue concernée.

Les perspectives de travail sont nombreuses, à commencer par l'implantation du modèle proposé. Le modèle peut également être amélioré de diverses façon. Un axe d'amélioration à explorer concerne l'ajout d'une mémoire d'analyse à travers une forme de lexicalisation de la grammaire. L'idée serait d'inclure dans le réseau des analyses, éventuellement partielles, pour des énoncés ou des morceaux d'énoncés.

Un autre axe concerne l'extension du réseau à d'autres niveaux de connaissances, tels les faits et événements, à l'image des SAR-Graphs.

Enfin, un axe primordial concerne l'exploitation de cette base de connaissances et des interactions entre différentes couches de réseau, notamment pour l'analyse sémantique de texte. On pourra notamment se demander dans quelle mesure les algorithmes existants de parcours de graphe peuvent contribuer à résoudre des problèmes de TAL tels que l'analyse sémantique.