Apprentissage de patrons lexico-syntaxiques à partir de textes

Valentina Dragos*, Marie-Christine Jaulent*

* INSERM, UMRS 872, eq. 20
15, Rue de l'école de médecine
75006 Paris
valentina.dragos@upmc.fr
marie-christine.jaulent@upmc.fr

Résumé. Ce papier présente une approche d'apprentissage de patrons lexico-syntaxiques à partir de textes annotés. Les patrons lexico-syntaxiques sont utilisés pour identifier des relations lexicales dans les corpus textuels. Leur construction manuelle est une tâche fastidieuse et des solutions permettant l'apprentissage sont souhaitables. Nous proposons une approche d'apprentissage qui repose sur l'utilisation des chemins de dépendance pour représenter les patrons et l'implémentation d'un algorithme de classification. L'approche a été appliquée dans le domaine biomédical pour identifier des patrons lexico-syntaxiques exprimant des relations fonctionnelles.

1 Introduction

Les patrons lexico-syntaxiques sont des structures représentant des schémas récurrents du langage. Ils sont utilisés dans le domaine du traitement automatique de la langue pour repérer des schémas langagiers dont ils sont l'abstraction. Les informations extraites sont des relations lexicales, voir par exemple Hearst (1992) ou Chagnoux et al. (2008) ou des relations spécifiques à un domaine particulier,cf. Koike et al. (2005).

La construction des patrons lexico-syntaxiques suppose un travail d'abstraction réalisé en amont, qui vise à définir une relation lexicale particulière et à identifier les différents contextes d'apparition de cette relation. Les éléments véhiculant la relation sont ainsi mis en évidence et synthétisés sous la forme d'un patron lexico-syntaxique. La construction manuelle des patrons présente deux inconvénients. Le premier est lié à leur portabilité. En changeant le domaine ou le type de langage d'un corpus de nouveaux patrons doivent être définis. Le deuxième inconvénient est lié à la mise en oeuvre d'applications à grande échelle. Dans ce cas, identifier un ensemble de patrons est loin d'être une tâche triviale. Ces deux inconvénients ont ouvert la voie à des travaux visant l'apprentissage de patrons lexico-syntaxiques. Nous proposons une approche d'apprentissage de patrons lexico-syntaxiques à partir de textes annotés. L'approche est générique et indépendante du domaine d'application. Nous l'avons utilisée dans le domaine biomédical pour identifier des patrons exprimant des relations fonctionnelles.

2 Apprentissage automatique de patrons

L'approche proposée apprend des patrons lexico-syntaxiques à partir de textes annotés par les experts du domaine. Il s'agit de phrases recueillies dans la littérature scientifique, dans lesquelles se retrouve l'information recherchée. L'annotation des phrases rend explicite l'information recherchée, comme on peut voir dans l'exemple suivant : *Endofin | early endosome| **Endofin** is associated with **early endosomes**.* Les phrases annotées sont étiquetées à l'aide d'un étiqueteur lexical, ce qui permet d'associer à chaque mot sa catégorie lexicale : *Endofin/NN early/JJ endosome/NN Endofin/NN is/VBZ associated/VBN with/IN early/JJ endosomes/NNS./*. Deux étapes sont nécessaires pour apprendre des patrons lexico-syntaxiques. La première utilise les chemins de dépendance pour construire un espace de représentation de patrons. La deuxième utilise un algorithme de classification qui permet d'identifier, dans l'espace de représentation précédemment construit, des chemins de dépendance similaires et de les regrouper dans des catégories homogènes. Ci-dessous sont détaillées ces deux étapes.

Représentation de patrons par des chemins de dépendance : Trois modèles sont utilisés pour représenter les patrons lexico-syntaxiques : les structures prédicat-argument, le modèle de sous-arbres et les chemins de dépendance. Les structures prédicat-arguments sont des triplets constitués d'un verbe et de ses arguments. Elles sont utilisées par les approches d'extraction d'information centrées sur les verbes, voir Wattarujeekrit et al. (2004) et présentent l'inconvénient de ne pas pouvoir mettre en évidence les informations qui ne s'expriment pas à l'aide de verbes. Le modèle des sous-arbres, Greenwood et al. (2005) représente un patron lexico-syntaxique sous la forme d'un arbre dont la racine est un verbe. Les noeuds de l'arbre sont des mots et les branches de l'arbre correspondent aux liens syntaxiques existants entre ces mots. Ce modèle enrichit le précédent, en mettant en évidence plusieurs liens syntaxiques. Cependant, il reste centré autour des verbes, ce qui constitue sa limitation.

Le modèle des chemins de dépendance représente un patron lexico-syntaxique sous la forme d'un chemin de dépendance. Les chemins de dépendance sont construits à partir de résultats fournis par un parseur. Un parseur construit un arbre de dépendance en identifiant les liens syntaxiques existants entre les différents mots d'une phrase. Un chemin de dépendance représente la succession des arcs reliant deux mots dans un arbre. Le modèle des chemins de dépendances a l'avantage de ne pas imposer de contrainte de modélisation. Il devient ainsi possible de mettre en évidence différents types de patrons véhiculant l'information recherchée, sans imposer des contraintes concernant les catégories constituantes.

Nous avons choisi le modèle des chemins de dépendance pour représenter les patrons lexico-syntaxiques. Ce formalisme a été adopté car il n'impose pas la construction de schémas langagiers centrés autour des verbes. Nous utilisons des arbres de dépendance construits à partir des résultats fournis pas un étiqueteur lexical, dont les branches correspondent à l'ordre de succession des mots dans la phrase. Cela nous permet de mettre en évidence le contexte lexical de chaque mot. Pour identifier les entités de l'annotation dans la phrase nous avons utilisé des heuristiques d'appariement de chaînes de caractères, afin de prendre en compte différents phénomènes langagiers, tels que la transposition ex. *signal transduction* et *transduction of signal*. Un algorithme de classification regroupe les chemins de dépendance similaires dans des catégories homogènes. Il est décrit *infra.*

Classification des chemins de dépendance : L'algorithme de classification repose sur la définition d'une mesure de similarité qui identifie les chemins de dépendance identiques, similaires ou distincts. Deux chemins de dépendance sont considérés identiques si : 1) les mêmes

catégories lexicales apparaissent dans leur structure et 2) l'ordre d'enchaînement des catégories lexicales est le même. Deux chemins de dépendance sont similaires si on peut obtenir des chemins identiques et appliquant, sur chacun d'entre eux plusieurs transformations lexicales. Les transformations lexicales consistent à remplacer une association de catégories lexicales par une autre, selon six schémas identifiés empiriquement. Le premier schéma consiste à remplacer le groupe *adverbe, verbe* par un *verbe*. La motivation de ce choix réside dans le fait que la présence d'un adverbe modifie uniquement certains aspects de l'information véhiculée par le verbe. Ainsi, les structures *pap 2* ***plays a role*** et *pap 2* ***also plays a role*** contiennent des informations similaires.

Le deuxième schéma substitue le groupe *verbe, adverbe, adjectif* par le groupe *verbe, adjectif*. Elle s'applique dans le cas de l'utilisation des adjectifs déverbaux. Ainsi, les structures : *stamp1* ***is localized*** et *GEF1* ***is predominantly localized*** fournissent des informations similaires. Le troisième schéma de transformation remplace le groupe *adjectif, nom* par le *nom*. Linguistiquement, la substitution est motivée par le fait que le déterminant d'un nom modifie des aspects de l'information véhiculée par le nom. Ainsi, des informations similaires sont exprimées par les constructions : *grasp 55 may play* ***a role*** et *pypaf7 play* ***a important role***. Le quatrième schéma de transformation consiste à remplacer le groupe *verbe modal, adverbe, verbe* par le groupe *verbe modal, verbe*. Elle est illustrée par l'exemple suivant : *irak1b* ***can activate*** et *IRF4* ***can functionally interact***.

Les deux dernières transformations reposent sur la même motivation que la transformation no. 3. Ainsi, le groupe *adjectif, nom* peut être remplacé par le *nom* (5ème schéma) et le groupe *adjectif, adjectif, nom* peut être remplacé par le groupe *adjectif, nom* (6ème schéma). Deux chemins de dépendance sont distincts si'ls ne sont pas identiques ou similaires.

3 Apprentissage de patrons lexico-syntaxiques pour l'annotation fonctionnelle

Ce paragraphe décrit une expérimentation réalisée dans le domaine biomédical pour apprendre des patrons exprimant des relations fonctionnelles. Une relation fonctionnelle représente une association entre un gène (ou un produit protéinique du gène) et une fonction.

Contexte de l'expérimentation : la tâche d'annotation fonctionnelle : L'annotation fonctionnelle concerne l'attribution d'un rôle fonctionnel (ou d'une fonction) aux produits protéiques des gènes. Dans le domaine de la biologie, une fonction peut être représentée par : une fonction moléculaire, un processus biologique ou un composant cellulaire d'un produit de gène. Un couple *protéine, fonction* ou *gène, fonction* représente une annotation fonctionnelle.

Données utilisées : Nous avons utilisé un ensemble des données utilisées dans le cadre de la compétition BiocreAtIvE 2005, voir Hirschmann et al. (2005). Les phrases ont été étiquetées en utilisant l'outil GeniaTagger, Tsuruoka et al. (2005). Nous avons à disposition 3468 annotations fonctionnelles, représentant un volume de 5600 Ko.

Résultats obtenus : Nous avons construit deux types de chemins de dépendance. Les chemins de dépendance lexicaux sont constitués uniquement de catégories lexicales. Les chemins de dépendance mixtes sont constitués de catégories lexicales et des mots clés *Gene* et *Function*. Le premier remplace la séquence de catégories lexicales correspondant au gène, alors que le deuxième mot remplace la séquence de catégories lexicales correspondant à la fonction. Ces

chemins de dépendance correspondent aux patrons lexico-syntaxiques spécifiques au domaine. Ainsi, *nn md vb dt jj nn* est un chemin lexical ayant comme instance *Mical may be a cytoskeletal regulator*. *Gene vbz to Func* est un chemin mixte instancié par *Stamp 1 colocalizes to early endosome*.

Nous avons identifié 397 chemins de dépendance lexicaux, auxquels correspondent 429 instances linguistiques et 394 chemins de dépendance mixtes, ayant 416 instances linguistiques. En appliquant l'algorithme de classification, nous avons obtenu 41 catégories de chemins de dépendance mixtes et 68 catégories de chemins de dépendance lexicaux. Chaque catégorie met en évidence des patrons lexicaux similaires. Ainsi, on retrouve dans une même catégorie les patrons *Gene vbz jj jj Func* et *Gene vbz Func*, dont les instances sont *light induces several distinct signal* et respectivement *vhl inhibits apoptosis*.

Evaluation des résultats : Pour évaluer les résultats nous utilisons un ensemble de patrons lexico-syntaxiques acquis manuellement à partir d'un corpus constitué de résumés d'articles scientifiques, voir Jilani (2009) et nous les considérons comme un gold standard. Il s'agit de patrons mixtes, structures en 6 types : *(1)Gène Verbe Function*, *(2)Fonction [mot +] verbe modal verbe préposition Gène*, *(3)Gène [mot +] verbe modal verbe préposition Fonction*, *(4)Gène [mot+] auxiliaire verbe préposition Fonction*, *(5)[mot+] Gène verbe préposition Fonction*, et *(6)Gène conjonction [mot+] préposition Fonction*. La présence du symbole + indique des mots pouvant avoir une ou plusieurs occurrences. Les éléments figurant entre [] peuvent ne pas apparaître.

Le protocole d'évaluation proposé est fondé sur la comparaison des patrons issus de l'apprentissage avec le gold standard. L'évaluation a été réalisée manuellement, en comparant chacune des catégories obtenues avec les types de patrons existants. Une catégorie est apparentée à un type de patron si, parmi les patrons qu'elle contient, on peut identifier un patron correspondant au type de patron en question. Ainsi, la catégorie constituée de *Gene VBZ JJ Func* et *Gene VBZ Func* sera associée au patron de type 1 *Gene VBZ Func*.

Parmi les 41 catégories de patrons identifiées par l'algorithme, 5 ont été assimilées aux patrons de type 1, 7 aux patrons de type 3, 2 aux patrons de type 2, 10 aux patrons de type 4 et 17 catégories n'ont pas été associées à un type de patrons. Si on considère comme non valides les patrons qui n'ont pas été assignés à un type de patrons, on peut calculer les valeurs de la précision et du rappel : Précision = 58.3 et Rappel = 66.66. Parmi les 6 types de patrons du gold standard, 2 n'ont pas été identifiés par notre approche d'apprentissage. Il s'agit du type 5, qui reconnaît des schémas langagiers dont le gène est précédé par un ou plusieurs mots. La manière dont nous construisons les chemins de dépendance rend impossible l'indentification d'un tel type de patron. Le deuxième type de patron non reconnu est le type 6. Le résultat pourrait être expliqué par la taille critique du corpus utilisé dans le cadre de notre travail. La plupart des patrons appris relèvent du type 4 et 3, qui correspondent aux structures employées souvent dans les articles scientifiques du domaine bio médical.

L'évaluation des patrons lexicaux n'a pas été réalisée car nous ne disposons pas actuellement d'un gold standard de patrons lexicaux. Ainsi, nous envisageons un protocole d'évaluation différent, qui consisterait à utiliser les patrons identifiés pour extraire des informations à partir de textes et à faire appel à un expert du domaine afin de valider les informations extraites, ce qui permettra d'évaluer , indirectement, les patrons utilisés.

4 Travaux connexes

Les patrons lexico-syntaxiques ont été introduits par les travaux de Hearst (1992) dans le but d'extraire des relations lexicales. Les limitations de la construction manuelle des patrons ont ouvert la voie aux travaux visant à automatiser leur acquisition à partir de textes. Ce paragraphe présente une sélection de ces travaux.

Snow et al. (2005) proposent une méthode pour automatiser l'acquisition de patrons lexico-syntaxiques permettant d'identifier des relations d'hyperonymie. Les travaux font partie d'un projet plus ample, visant l'enrichissement automatique de ressources lexicales. Les patrons sont représentés sous forme de chemins de dépendance et la construction de ces chemins est guidée par l'utilisation des mots clés identifiés empiriquement (ex. *such as*, *and*). Ces mots clés sont remplacés dans l'arbre de dépendance par de nouveaux liens, ce qui permet d'augmenter le nombre de connexions de l'arbre et de construire ainsi de nouveaux chemins de dépendance. En exploitant les arbres ainsi enrichis, les auteurs définissent un patron lexico-syntaxique comme un chemin de dépendance contenant au plus quatre noeuds. Les auteurs proposent également une procédure d'évaluation semi automatique. Ainsi, les phrases contenant des paires de noms se trouvant ou pas dans une relation d'hyperonymie sont catégorisées par l'algorithme proposé, et les résultats fournis sont analysés manuellement.

Greenwood et al. (2005) proposent une approche d'apprentissage de patrons lexico-syntaxiques capable d'identifier des interactions géniques. Les patrons sont représentés par le modèle des sous arbres. Un patron est représenté par un arbre ayant comme racine un verbe, et plusieurs sous-arbres dont les noeuds contiennent : des mots en forme lemmatisée, des catégories lexicales de mots (verbe, nom, etc.) ou des catégories fonctionnelles des mots (gène, protéine, etc.). Les patrons sont générés itérativement, à partir de documents textuels, et sont comparés entre eux afin d'identifier des classes similaires. L'approche est évaluée en utilisant un corpus annoté manuellement.

Les travaux présentés *supra* permettant l'apprentissage de patrons lexico syntaxiques à partir de textes. Le processus d'apprentissage est guidé par des connaissances linguistiques ou par des connaissances du domaine. Des mesures statistiques sont parfois utilisées pour caractériser la pertinence d'un patron. La validation des résultats demeure un processus manuel. La solution que nous avons développée se rapproche de la solution proposée par Snow et al. (2005). Cependant, nous avons utilisé des connaissances linguistiques acquises empiriquement pour la construction des chemins de dépendance, ce qui offre à notre solution un degré plus élevé de généralité.

5 Conclusion

Ce papier présente une approche d'apprentissage de patrons lexico-syntaxiques à partir de textes annotés. L'intérêt d'une telle approche est de pallier les inconvénients sous-jacents à la construction manuelle de patrons lexico-syntaxiques. L'approche repose sur la représentation des patrons sous forme de chemins de dépendance et l'implémentation d'un algorithme de classification. Elle a été utilisée dans le domaine biomedical pour apprendre des patrons exprimant des annotations fonctionnelles. Les résultats obtenus ont été évalués en comparant les patrons appris avec un gold standard. Plusieurs directions de recherche ont été identifiées pour améliorer ces résultats. Ainsi, nous pouvons prendre en compte des ressources lexicales ou

des ontologies pour améliorer la qualité des annotations. Nous pouvons également utiliser des exemples négatifs de la relation recherchée, afin de renforcer le processus d'apprentissage.

Références

Chagnoux, M., N. Hernandez, et N. Aussenac-Gilles (2008). An interactive pattern based approach for extracting non-taxonomic relations from texts. In P. Buitelaar, P. Cimiano, G. Paliouras, et M. Spilliopoulou (Eds.), *Workshop on Ontology Learning and Population*, pp. 1–6. University of Patras.

Greenwood, M. M. S., Y. Guo, H. Harkema, et A. Roberts (2005). Automatically acquiring a linguistically motivated genic interaction extraction system. In *In Proceedings of the 4th Learning Language in Logic Workshop (LLL05*.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING*, pp. 539–545.

Hirschmann, L., A. Yeh, C. Blasche, et A. Valencia (2005). Overview of biocreative : critical assessment of information extraction for biology. *BMC Bioinformatics 6*, 27–48.

Jilani, I. (2009). *Extraction automatique des connaissances à partir de textes bio-médicaux*. Thèse de doctorat, Université Pierre et Marie Curie.

Koike, A., Y. Niwa, et T. Takagi (2005). Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics 21*, 1227–1236.

Snow, R., D. Jurafsky, et A. Y. Ng (2005). Learning syntactic patterns for automatic hypernym discovery. In L. K. Saul, Y. Weiss, et L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, pp. 1297–1304. Cambridge, MA: MIT Press.

Tsuruoka, Y., Y. Tateishi, J. Kim, T. Ohta, J. Mcnaught, S. Ananiadou, et J. Tsujii (2005). Developing a robust part-of-speech tagger for biomedical text. pp. 382–392.

Wattarujeekrit, T., P. K. Shah, et N. Collier (2004). Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics 5*.

Summary

This paper presents an automatic approach to learn lexico-syntactic patterns from texts. Such patterns are used to identify lexical relations within textual corpora. Their manual construction is an expensive and time consuming task. Therefore, new approaches are needed allowing us to automatically construct lexico-syntactic patterns. We propose a generic approach to learn patterns from texts. Our solution uses dependency paths as representation model and implements a classification algorithm. The approach was applied in the biomedical field in order to identify lexico-syntactic patterns expressing relationships between genes or proteins and functions.