Attribution de rôles sémantiques aux actants des lexies verbales

Fadila Hadouche[1] Guy Lapalme[1] Marie-Claude L'Homme[2]

(1) RALI, (2) OLST
Université de Montréal, C.P 6128 Succursale Centre-ville, Montréal, Québec, Canada H3C 3J7
hadouchf@iro.umontreal.ca, lapalme@iro.umontreal.ca, mc.lhomme@umontreal.ca

Résumé

Dans cet article, nous traitons de l'attribution des rôles sémantiques aux actants de lexies verbales en corpus spécialisé en français. Nous proposons une classification de rôles sémantiques par apprentissage machine basée sur un corpus de lexies verbales annotées manuellement du domaine de l'informatique et d'Internet. Nous proposons également une méthode de partitionnement semi-supervisé pour prendre en compte l'annotation de nouvelles lexies ou de nouveaux rôles sémantiques et de les intégrés dans le système. Cette méthode de partitionnement permet de regrouper les instances d'actants selon les valeurs communes correspondantes aux traits de description des actants dans des groupes d'instances d'actants similaires. La classification de rôles sémantique a obtenu une F-mesure de 93% pour Patient, de 90% pour Agent, de 85% pour Destination et de 76% pour les autres rôles pris ensemble. Quand au partitionnement en regroupant les instances selon leur similarité donne une F-mesure de 88% pour Patient, de 81% pour Agent, de 58% pour Destination et de 46% pour les autres rôles.

Abstract

In this paper, we discuss assigning semantic roles to actants of verbal lexical units in French specialized corpus. We propose a machine learning classification of semantic roles based on a corpus of verbal lexical units, which are annotated manually in the Informatics and Internet domain. We also propose a semi supervised clustering method to consider the annotation of new verbal lexical units or new semantic roles and integrated them in the system. Clustering is used to group instances of actants according to their common values corresponding to the features describing these actants into groups of similar instances of actants. The classification model give an F-measure of 93% for Patient, 90% for Agent, 85% for Destination and 76% for other roles. When partitioning by grouping instances according to their similarity gives an F-measure of 88% for Patient, 81% for Agent, 58% for Destination and 46% for other roles.

Mots-clés : Rôles sémantiques, traits syntaxiques, classification, partitionnement semi-supervisé
Keywords: Semantic roles, syntactic features, classification, semi supervised partitioning

1 Introduction

L'annotation automatique des structures actantielles joue un rôle important dans les applications du traitement automatique des langues naturelles telles que l'extraction d'informations [2, 9], la traduction automatique [1] les questions/réponses [7] et le résumé automatique [6]. Ainsi Gerhard Fliedner a montré que l'utilisation d'annotations automatiques améliore les performances de ces applications et propose des outils de construction d'un lexique d'annotation sémantique [4]. Plusieurs techniques ont été explorées pour induire automatiquement des rôles sémantiques : méthodes d'apprentissage non supervisé employant des informations lexicales pour développer le classificateur ou méthodes d'apprentissage supervisé se basant sur les données d'entraînement extraites des ressources lexicales PropBank[1] ou FrameNet[2], particulièrement pour l'anglais, qui ont aussi servi comme corpus pour entraîner le classificateur.

Dans notre travail, nous traitons les unités lexicales ou lexies[3] verbales du français apparaissant dans des contextes, c.-à-d. des phrases tirées d'un corpus de textes français des domaines de l'informatique et de l'Internet. L'annotation assigne aux actants de ces lexies verbales des étiquettes de rôles sémantiques par exemple Agent, Patient, Destination, Instrument, Source, Lieu, Moyen, etc. Par exemple

[Agent Vous] CLIQUEZ sur [Patient le bouton]

Dans cette phrase, Vous et le bouton réalisent les deux actants de la lexie verbale CLIQUER. Le premier actant, le pronom personnel Vous, le sujet de la lexie verbale joue le rôle sémantique d'Agent et le deuxième actant, le bouton, qui occupe la place du complément joue le rôle de Patient. L'identification de rôles sémantiques peut aussi concerner les circonstants[4] mais, nous nous limitons ici aux actants. Les actants se distinguent des circonstants en ce sens qu'ils participent au sens d'une lexie prédicative. Certains rôles dans notre corpus tels que Assaillant, Lieu, Instrument, Source, Récipient, Cause, Environnement et Matériau ne sont pas suffisamment représentés par des exemples. Ceci nous a conduit à rassembler ces rôles dans une seule classe appelée Autre. Nous avons ainsi restreint la liste des rôles à prendre en compte dans l'apprentissage machine. La liste finale des rôles étudiés ici est donc Agent, Patient, Destination et Autre.

Nous voulons automatiser la tâche d'attribution automatique de rôles sémantiques aux actants des termes spécialisés. Nous nous sommes basés sur des descriptions actancielles (arguments) existantes réalisées à la main et enrichies de rôles sémantiques inspirés du modèle de *Frame Semantics* [3]. Ce corpus est constitué de phrases du français tirées d'un dictionnaire spécialisé dans le domaine de l'informatique (DicoInfo[5]) construit sur la base de la Lexicologie Explicative et combinatoire (LEC). Les unités lexicales verbales apparaissant dans ces phrases sont annotées manuellement. L'annotation manuelle de phrases reste une tâche fastidieuse pour les annotateurs et très exigeante en temps, en moyenne 2 heures par lexie. Nous proposons un modèle d'annotation automatique pour accélérer le temps d'annotation et faciliter la tâche de l'annotateur. Cette tâche d'annotation devient une validation des annotations faites par notre système automatique. En évitant une grande partie du travail routinier pour la majorité des cas simples, le linguiste pourra se concentrer sur les cas plus difficiles et plus intéressants.

Dans cet article, nous présentons les traits syntaxiques qui décrivent les actants extraits à partir de l'analyseur syntaxique du français Syntex[6] sur lesquels sont basées la classification et une méthode de partitionnement semi-supervisé.

2 Définition des traits d'attribution de rôles sémantiques

L'annotation automatique des rôles sémantiques est basée essentiellement sur des traits ou caractéristiques syntaxiques qui décrivent les actants d'une lexie verbale dans une phrase. Pour définir ces traits, nous nous sommes inspirés des traits utilisés pour l'annotation de rôles sémantiques des unités lexicales de l'anglais et des descriptions des données dans notre corpus annoté manuellement.

Nous avons constaté que les fonctions et groupes syntaxiques ne peuvent pas à eux seuls de définir un rôle sémantique car des rôles sémantiques différents peuvent avoir des fonctions et des groupes syntaxiques

[1] PropBank http://verbs.colorado.edu/~mpalmer/projects/ace.html
[2] FrameNet http://framenet.icsi.berkeley.edu
[3] Dans le présent article, **lexie** et **unité lexicale** sont utilisées dans le même sens.
[4] Certains auteurs désignent les actants par arguments et les circonstants par adjoints.
[5] DicoInfo : Dictionnaire fondamental de l'informatique (http://olst.ling.umontreal.ca/dicoinfo)
[6] Analyseur en dépendances, Bourigault et al., (http://w3.erss.univ-tlse2.fr/textes/pagespersos/bourigault/syntex.html)

similaires [5]. Ces derniers ne déterminent pas un rôle d'une manière unique. Donc ils ne peuvent pas seuls construire les traits sur lesquels est basé notre modèle. Nous avons défini d'autres caractéristiques complémentaires décrites au Tableau 1

Caractéristique	Signification	Ag	Pa	De	Au
Lexie	Une lexie peut affecter un rôle sémantique à ses actants : ex. une lexie L1 peut affecter un rôle R1 pour son actant sujet alors qu'une autre lexie L2 lui affectera un rôle R2.	14	77	40	37
Tête de la lexie	La tête est sous forme d'une préposition quand la lexie est précédée par une préposition telle que de ou à. Un actant sujet d'une lexie précédée par la préposition à est patient.	0	75	0	0
Catégorie lexie	La catégorie grammaticale de la lexie (conjugué, participe passé, participe présent, etc.)	0	72	0	0
Mot actant	La réalisation de l'actant de la lexie.	75	81	42	42
Tête de l'actant	Les actants prépositionnels jouent des rôles sémantiques selon la tête (préposition utilisée) : à, avec, dans, sur, pour, etc.	0	72	0	0
Fonct-synt-act	Ce sont les fonctions syntaxiques (sujet, objet, lien-indirect, modificateur, tête, complément) rencontrées dans notre corpus	35	81	55	0
Group-synt-act	Sont telles que SN (syntagme nominal), SP (syntagme prépositionnel), etc.	61	76	0	0
Position	Position où apparaît l'actant par rapport à la lexie. « avant ou après »	0	72	0	0
Distance	Le nombre de mots plein qui séparent un actant de sa lexie	5	72	0	2
Ordre	Ordre de l'actant par rapport aux autres actants de la lexie dans une même phrase. C.-à-d., s'il est 1er actant, 2ème actant, etc.	50	75	60	40
Nombre_actant	Nombre d'actants que possède la lexie dans le contexte en question	0	72	2	0
Verbe	De préciser le verbe qui apparait entre la lexie et le mot actant. Ce verbe peut être l'auxiliaire être ou avoir. Il peut être aussi un verbe modal tel que permettre, demander, pouvoir, vouloir etc. qui affectent des rôles sémantiques spécifiques.	44	74	5	0
Nombre_verbe	Nombre de verbes qui apparaissent entre la lexie et son actant	37	87	0	0

Tableau 1 Caractéristiques ou traits décrivant les actants d'une lexie verbale. Les 4 dernières colonnes présentent la contribution de ces traits à la classification des rôles sémantiques en termes de F_mesure en pourcentage (%) pour les rôles *Agent*, *Patient*, *Destination* et *Autres*.

Ces traits décrivent les actants de lexies verbales dans des contextes. Par exemple, dans la phrase

[Agent Vous] CLIQUEZ sur [Patient le bouton]

les actants Vous et le bouton de la lexie CLIQUEZ sont décrits par les traits du Tableau 2 comme suit

lexie	tête	Cat.	mot	tête	FS	GS	Pos	dist	ord	Verb	NbrA	NbrV
CLIQUER	?	Vconj	Vous	?	sujet	Pro	avant	0	1	?	2	0
			bouton	?	Objet	SP	après	1	2	0	2	0

Tableau 2 Description des réalisations d'actants Vous et bouton de la lexie cliquer

Chaque ligne de ce type décrit un actant d'une lexie. Nous disposons de 3301 lignes ou instances correspondant à chacun des actants de différentes lexies dans des contextes différents rencontrés dans notre corpus. La valeur « ? » ou valeur absente ou indique que ce trait ne concerne pas l'actant en question dans ce contexte.

3 Classification de rôles sémantiques par Weka[7]

Après avoir testé plusieurs classificateurs, il s'est avéré que RandomForest est celui qui a donné les meilleurs résultats sur nos données décrites par les traits du Tableau 1. Nous avons d'abord testé ce classificateur en tenant compte des traits individuellement dont les résultats sont donnés dans le Tableau 1. Nous constatons que s'il est assez facile d'identifier le rôle Patient, les autres rôles sont beaucoup plus difficiles à départager. La combinaison de ces traits permet de distinguer ou de démarquer ces rôles sémantiques plus efficacement que de les prendre individuellement. Dans le Tableau 3, nous donnons les résultats en précision, rappel et F-mesure, du classificateur RandomForest en considérant tous les traits. Ces résultats sont calculés sur la validation croisée (10 folds) du corpus d'entraînement composé de 3301 actants de lexies différentes.

	RandomForest		
	Précision	Rappel	F_mesure
Agent	0,94	0,86	0,90
Patient	0,90	0,97	0,93
Destination	0,90	0,82	0,85
Autre	0,85	0,69	0,76

Tableau 3 Combinaison de tous les traits dans la classification des rôles

La classification des rôles sémantiques donne donc de très bons résultats sur les rôles proposés pour les 3301 actants rencontrés dans le corpus. Or nous savons que les rôles sémantiques ne constituent pas une liste finie. Ces derniers sont découverts au fur et à mesure que l'annotation de nouvelles lexies ou données est réalisée. De nouveaux rôles peuvent apparaître. L'objectif premier de notre travail était d'accompagner le linguiste annotateur pendant l'annotation des rôles sémantiques, notre système devrait donc être capable de donner à l'annotateur la possibilité non seulement de choisir le rôle proposé mais aussi d'en définir un nouveau. Pour intégrer cette façon de faire dans le système d'annotation manuel, nous proposons donc de classifier les instances d'actants décrites à l'aide des traits définis ci-dessus, au lieu de simplement classifier les rôles sémantiques. Nous regroupons toutes les instances proches ou qui partagent les mêmes valeurs des traits suggérés dans les mêmes groupes. Nous proposons de faire un partitionnement semi-supervisé sur l'ensemble de nos données. C'est-à-dire nous regroupons les instances dans des partitions selon leur similarité et ensuite classifions les rôles sémantiques dans ces groupes. Pendant le regroupement, les rôles sémantiques sont ignorés.

4 Partitionnement semi-supervisé

Le partitionnement est un processus de regroupement des données dans des groupes afin que des objets dans un même groupe aient une similarité plus élevée entre eux qu'avec les instances d'un autre groupe. Dans notre cas, nous procédons au regroupement semi supervisé dont les instances à regrouper possèdent des étiquettes qui sont les rôles sémantiques. Dans les regroupements et la mesure de similarité, les étiquettes de rôles sémantiques sont ignorées. Une fois les groupes formés, nous affectons aux instances leurs étiquettes ainsi nous classifions les étiquettes dans chaque groupe. Nous regroupons nos données selon leur mesure de similarité. Cette notion de similarité est basée sur les caractéristiques ou traits communs entre les instances. Elle est égale au nombre de valeurs communes pour ces caractéristiques divisé par le nombre de valeurs de ces caractéristiques pour les deux actants en comparaison.

Nous avons utilisé l'algorithme CHAMÉLÉON (Hierarchical Clustering Using Dynamic Modeling) [Karipys 1999] qui fusionne les différents groupements trouvés dans sa phase de partitionnement de manière dynamique. Il assure la réalisation de deux conditions: maximiser la similarité intra-groupe et minimiser la similarité inter-groupes. Il opère en deux phases partitionnement ensuite fusion.

5 Expérimentation

Dans notre expérimentation, nous avons divisé notre ensemble de données (3301 instances d'actant) en deux parties. Nous en avons utilisé les 2/3 (2200 instances) pour la formation des groupes, et 1/3 (1101 instances) pour tester si les groupements sont bien formés.

[7] Weka est développé à l'Université de Waikato en Nouvelle-Zélande (www.cs.waikato.ac.nz/ml/weka).

Nous avons appliqué les deux phases de partitionnement et de fusion de l'algorithme CHAMÉLÉON pour former les groupes. Nous avons calculé les similarités de toutes les instances d'actants, deux à deux en utilisant la formule de similarité (1) et nous avons formé une matrice de similarité qui correspond au graphe initial. Ce dernier est partitionné en sous-graphes en utilisant l'algorithme de partitionnement du package Metis (Karypis 1998) basé sur *k-way partitioning*. Ces différentes partitions ou groupes sont fusionnés en mesurant la similarité entre eux. Ceci est interprété par le calcul de deux mesures entre ces groupes : interconnectivité relative (RI) qui est obtenue par le rapport entre la somme des poids minimaux des arcs qui ont permis le partitionnement de ce groupe par rapport à la somme des poids des arcs dans chaque groupe; et proximité relative (RC) qui est obtenu plutôt par la moyenne des poids au lieu de leur somme. La fusion se base sur ces deux mesures qui consiste à maximiser $RI(Ci,Cj)* RC(Ci,Cj)^{\alpha}$. Le choix de ces paramètres k nombre de partitions et α est fait en testant plusieurs valeurs. Avec chacune de ces valeurs de k (5, 10, 15, 20) et de α (0.5, 1, 1.5, 2, 2.5, 3), nous avons formé des groupements à base des 2200 instances en appliquant l'algorithme CHAMÉLÉON. Une fois les groupes formés avec ces valeurs, nous avons essayé de classifier les 1101 instances de test dans ces groupes. Nous avons évalué le résultat en termes de précision, définie par le rapport entre le nombre d'instances de test correctement classifiées et la somme du nombre d'instances correctement classifiées et du nombre d'instances incorrectement classifiées; et de rappel défini par le rapport entre le nombre d'instances de test correctement classifiées et le nombre total d'instances de chaque classe. Nous avons pris les valeurs des paramètres qui maximisent la F-mesure. Nous avons choisi un nombre de partitions k=20, ce qui permet à CHAMÉLÉON de fusionner les partitions qui se ressemblent. Si ce nombre est faible, CHAMÉLÉON risque de ne pas trouver les groupes naturels. Nous avons choisi α=1 qui donne une même importance aux deux mesures RI et RC.

Pour chaque groupe formé, nous cherchons des représentants qui peuvent mieux décrire l'ensemble des éléments pouvant appartenir à ce groupe. Ces représentants sont les instances portant un rôle sémantique à associer à chaque groupe. Nous avons proposé de calculer un seuil qui permet de sélectionner les instances représentantes du groupe. Nous avons calculé pour chaque instance d'actant son importance ou son poids dans le groupe. Ce poids est calculé en utilisant la mesure de similarité (1). Le poids P_i d'une instance i dans un groupe C est donné par la somme des similarités de cette instance i avec les autres instances j du groupe C divisé par la somme des similarités entre toutes les paires d'instances k, j du groupe C. Le poids P_i est donné par :

$$P_i = \frac{\sum_{j \in C} Sim(i,j)}{\sum_{k,j \in C} Sim(k,j)} \qquad (2)$$

Le seuil est défini par la moyenne des poids de toutes les instances du groupe C. Nous sélectionnons toutes les instances ayant un poids supérieur ou égal à ce seuil comme instances représentantes du groupe C. les groupes obtenus sont formés de ces instances représentantes.

Pour tester si les groupes obtenus sont bien formés, nous avons pris les instances de notre ensemble de test et nous avons cherché à identifier le groupe dans lequel on devrait la classifier. Le Tableau 4, montre les résultats obtenus par CHAMÉLÉON vs classification en utilisant les mesures de précision, de rappel et de F-mesure.

Rôles sémantiques	**RandomForest**			**CHAMÉLÉON**		
	Précision	**Rappel**	**F-mes**	**Précision**	**Rappel**	**F-mes**
Patient	0,90	0,97	0,93	0,88	0,89	0,88
Agent	0,94	0,86	0,90	0,75	0,88	0,81
Destination	0,90	0,82	0,85	0,66	0,53	0,58
Autre	0,85	0,69	0,76	0,62	0,37	0,46

Tableau 4 résultats de la classification vs partitionnement semi-supervisé

La classification donne de meilleurs résultats que le partitionnement semi-supervisé. La classification est la meilleure pour classifier les rôles sémantiques et prédire des rôles, déjà existants pour de nouvelles instances des lexies déjà vues. Par contre, elle ne permet pas de prendre en compte de nouveaux rôles et de nouvelles lexies.

Afin d'annoter de nouvelles lexies ou de nouveaux rôles, nous avons proposé une approche de partitionnement semi supervisé, qui permet de regrouper les instances semblables dans des groupes, afin d'intégrer les nouveaux rôles suggérés par l'annotateur dans des groupes correspondants.. Nous estimons avoir obtenu des taux de précision et rappel acceptables. Ce modèle a donné une F-mesure de 88% pour

Patient, de 81% pour Agent dont le nombre d'exemples est respectivement de 1992 et de 792. Les résultats obtenus sont comparables à certains travaux similaires effectués pour d'autres langues. Padò et Lapata se sont basés sur le FrameNet anglais pour annoter les rôles sémantiques en français. Ils ont testé en 2007 une projection anglais-français [8] et ils ont obtenu une F-mesure de 69%. Quand aux rôles dont le nombre d'exemples ne dépasse pas les 300, ce modèle de partitionnement donne une F-mesure de 58% pour Destination et de 46% pour Autre. Nous constatons dans ce cas que le nombre d'exemples est important pour atteindre de bons résultats tels que le cas de Patient et Agent.

Conclusion

Nous avons testé le classificateur RandomForest sur nos données qui a obtenu de bons résultats à savoir une F-mesure de 93% pour Patient, de 90% pour Agent, de 85% pour Destination et de 76% pour la classe Autre. Le corpus annoté et vérifié par des linguistes a été un grand avantage par rapport à des études similaires effectuées par d'autres équipes pour d'autres langues. Le modèle de classification permet d'annoter de nouvelles instances de lexies déjà traitées, mais nous ne pouvons donner que le rôle le plus proche dans le cas d'une nouvelle lexie avec de nouveaux rôles. Afin d'intégrer de nouveaux rôles, nous avons proposé un partitionnement semi-supervisé pour lesquels nous recalculons les représentants de ces groupes en tenant compte des rôles intégrés. Ce modèle de partitionnement a donné une F-mesure de 88% pour Patient, de 81% pour Agent. Quand aux rôles dont le nombre d'exemples ne dépasse pas les 300 la F-mesure est de 58% pour Destination et de 46% pour Autre. Nous avons utilisé un corpus basé sur DicoInfo, une ressource construite sur la base de la LEC, une théorie différente de celle de FrameNet démontrant ainsi qu'il est possible d'affecter des rôles sémantiques dans d'autres langues sans pour autant structurer les données sous la forme ou la hiérarchie de Framenet ou en passant par la notion de frames.

Remerciements

Nous tenons à remercier les annotateurs de l'OLST, particulièrement Annaïck Le Serrec, Janine Pimentel, Marie-Ève Lanville et Susanne Desgroseilliers, d'avoir mis à notre disposition le corpus d'annotation sur lequel nous avons travaillé et qui ont validé les annotations de notre système. Nous remercions également le CRSH pour le support financier.

Références

1. Boas H.C., Bilingual Framenet dictionaries for machine translation. In Proceedings of the Third International Conference on Language Ressources and Evaluation (LREC), 2002, Las Palmas de Gran Canaria, Spain.
2. Bunescu R.C. et Mooney R.J. *A shortest path dependency kernel for relation extraction*. in *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processind (HLT-EMNLP-2005)*. 2005. Vancouver, B. C., Canada.
3. Fillmore C.J., *Frame semantics*, in *Linguistic Society of Korea*, ed. Linguistics in the Morning Calm. Seoul:Hanshin. 1982. p. 111-137.
4. Fliedner G., Tools for building a lexical semantic annotation. Proceedings of Lorraine-Saarland Workshop on Prospects and Advances in the Syntax/Semantic Interface, 2003, Loria, Nancy, France, p. 5-9.
5. Hadouche F., Annotation syntaxico-sémantique des actants en corpus spécialisé. Thèse de Ph.D, Université de Montréal, 2011.
6. Melli G., et al. *Description of squash, the SFU question answering summary handler for the DUC-2005 summarization task*. in *in Proceedings of the HLT/EMNLP Document Understanding Workshop (DUC)*. 2005. Vancouver, Canada.
7. Narayanan S. et Harabagiu S. *Question answering based on semantic structures*. in *In Proc. of the 20th International Conference on Computational Linguistics (COLING)*. 2004. Genève, Suisse.
8. Padò S. et Pitel G., Annotation précise du français en sémantique de rôles par projection cross-linguistique. Actes de la conférence Traitement Automatique des Langues Naturelles (TALN), 2007, Toulouse, France.
9. Surdeanu M., et al., Using predicate-argument structures for information extraction. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL '03), 2003, Sapporo, Japan, p. 8-15.