Identification des participants de lexies prédicatives : évaluation en performance et en temps d'un système d'annotation automatique

Fadila Hadouche[1], Suzanne DesGroseilliers[2], Janine Pimentel[2], Marie-Claude L'Homme[2], Guy Lapalme[1]
1) RALI
2) OLST
Université de Montréal, C.P. 6128 Succ.
Centre-ville, Montréal, H3C 3J7, Québec, Canada
hadouchf@iro.umontreal.ca,{suzanne.desgroseilliers,janine.pimentel, mc.lhomme}@umontreal.ca,lapalme@iro.umontreal.ca

Résumé

Afin d'alléger le travail d'annotation de contextes illustrant le comportement syntaxico-sémantique des termes du domaine de spécialité de l'informatique et de l'Internet en français, une méthode d'annotation automatique a été conçue. Dans cet article, nous proposons d'évaluer une partie du système d'annotation automatique de lexies verbales. Nous évaluons la méthode automatique de la tâche d'identification d'actants (également appelés *arguments*) et circonstants (en anglais, *adjuncts*) en performance et en temps. Nous avons procédé en deux étapes. Dans la première, deux terminologues repèrent les erreurs d'identification les plus fréquentes. Dans la seconde, ils identifient manuellement les actants et circonstants dans des contextes d'un échantillon de termes afin de mesurer le temps nécessaire pour réaliser cette tâche. Ce dernier est comparé au temps requis par un réviseur qui corrige les sorties automatiques de ce même échantillon. Nos évaluations montrent que 80 % des sorties automatiques sont justes et que les terminologues gagnent 60 % de temps en corrigeant les sorties automatiques par rapport à un travail réalisé manuellement.

1 Introduction

L'annotation des contextes illustrant des termes permet aux terminologues de décrire leurs propriétés syntaxico-sémantiques dans les dictionnaires spécialisés et les bases de données. Toutefois, cette tâche est souvent longue et ardue. Afin de l'alléger, nous avons conçu une méthode d'annotation automatique.

Le système d'annotation automatique que nous avons proposé (Hadouche, 2010) réalise trois tâches principales : identification des participants de la lexie à annoter, distinction entre actants et circonstants[1], et attribution de rôles sémantiques aux actants. Dans cet article, nous proposons d'évaluer ce système automatique pour les deux premières tâches.

La méthode automatique utilisée pour réaliser ces tâches est basée sur un modèle d'apprentissage machine. Nous avons entraîné un classificateur RandomForest[2] du package Weka (Witten *et al.*, 2005) en validation croisée de 10. Ce classificateur est basé sur des traits de description que nous avons définis. Dans un premier cas, Nous avons pris en compte les dépendances syntaxiques comme un trait en nous basant sur l'analyseur syntaxique Syntex (Bourigault, 2007). Dans le deuxième cas, nous avons proposé des traits sans inclure les dépendances syntaxiques.

Nous avons testé ce système sur une quarantaine de nouvelles lexies verbales, non annotées manuellement, ayant approximativement une

[1] La distinction entre « actant » et « circonstant » s'appuie sur Mel'čuk (2004). Les actants sont définis comme des participants obligatoires contribuant au sens d'unités lexicales de sens prédicatif. Ce sont ceux qui apparaissent dans la définition de l'unité lexicale. Par contre, les circonstants ne contribuent pas au sens de ces unités, bien qu'ils puissent entretenir un lien syntaxique avec elles. Par exemple, dans le contexte *L'usager a configuré son nouveau micro-ordinateur hier*, l'unité prédicative *configurer* possède deux participants obligatoires ou actants (*l'usager* et *son nouveau micro-ordinateur*) et un participant optionnel ou circonstant (*hier*).

[2] RandomForestTree qui consiste en un ensemble d'arbres de décision. Dans notre cas, nous avons utilisé un nombre de 10 arbres.

quinzaine de contextes chacune. Des terminologues ont révisé les sorties de ce système sur 500 contextes au total. En plus de la performance, nous mesurons le temps dont un terminologue a besoin pour réviser ces sorties. Sur un échantillon de 5 lexies verbales de 20 contextes, nous avons observé un gain de temps important permis par l'identification automatique, c'est-à-dire en faisant uniquement la révision des sorties.

2 Pourquoi annoter les termes de nature prédicative ?

La description des propriétés syntaxico-sémantiques des termes dans les dictionnaires spécialisés et bases de données constitue une source de renseignements extrêmement utiles sur plusieurs plans. Par exemple, elle permet de mettre en évidence les constructions syntaxiques que les termes admettent ainsi que le comportement de leurs participants (nombre, rôles sémantiques, modalités de combinatoire avec le terme, etc.). Bien que ces propriétés soient rarement expliquées dans les dictionnaires généraux, voire dans les dictionnaires spécialisés, un certain nombre de ressources lexicales générales dont FrameNet (Ruppenhofer *et al.*, 2010), ou spécialisées dont le DiCoInfo[3] (la ressource utilisée dans le cadre de notre travail, L'Homme, 2008), commencent à fournir ce genre de renseignements.

Les terminologues du DiCoInfo décrivent les structures actancielles des termes prédicatifs de la façon suivante:

Naviguer : Agent{internaute 1} ~ dans Lieu{Internet 1} avec Instrument{navigateur 1}

Ensuite, ils utilisent une méthode d'annotation formelle pour faire émerger les propriétés syntaxico-sémantiques des termes à partir de contextes prélevés dans un corpus. Prenons les contextes suivants du terme **naviguer** :

[Act,Instrument,Lien indirect, SN Un logiciel] permet de **NAVIGUER** [Act, Lieu, Complément, SP sur le web]

[Act,Agent,Sujet, SN Vous] **NAVIGUEZ** [Act, Lieu, Complément, SP sur internet] [Circ, Mode, Complément, Prop sans bien le comprendre].

Les éléments qui sont explicités dans ces annotations sont : le terme prédicatif faisant l'objet de l'annotation (*naviguer*), les participants et leur nature d'actant ou de circonstant[4] (ici les éléments entre crochets), le rôle sémantique des participants (Agent, Instrument, Lieu, Mode), leur fonction syntaxique (Sujet, Lien indirect, Complément) et leur groupe syntaxique (SN, SP, Proposition).

Pour chaque unité prédicative, les annotateurs annotent manuellement une quinzaine de contextes en utilisant le format .XML et des balises correspondant à chaque élément décrit.

Les éléments explicités dans les annotations sont présentés sous forme de tableaux récapitulatifs (Tableau 1).

NAVIGUER 1		
Actants		
Agent	Sujet (SN)	Vous
Lieu	Complément (SP -sur)	internet web
Instrument	Lien indirect (SN)	logiciel
Autres		
Mode	Complément (Prop)	Sans bien le comprendre

Tableau 1. Tableau récapitulatif des informations syntaxico-sémantiques du terme **naviguer** dans les exemples de contextes ci-dessus.

Cette méthode d'annotation des termes prédicatifs s'inspire largement de celle mise au point dans le cadre du projet de FrameNet. Bien que FrameNet et le DiCoInfo s'intéressent aux patrons syntaxico-sémantiques des lexies, le choix de rôles sémantiques est différent. Les terminologues du DicoInfo utilisent des rôles s'inspirant plutôt d'une liste plus restreinte proposée par Fillmore (1968) pouvant être appliquée à l'ensemble des termes. Les lexicographes de FrameNet choisissent quant à eux des étiquettes qui, la plupart du temps, sont applicables à l'intérieur d'un frame seulement.

L'annotation manuelle de phrases dans lesquelles apparaissent les lexies prédicatives reste une tâche fastidieuse pour les annotateurs et très exigeante en temps (en moyenne 1 heure par lexie), d'où l'intérêt de faire appel à une méthode automatique d'annotation.

3 Méthode informatique

Nous avons proposé un modèle de classification binaire pour les tâches d'identification et de distinction des participants actants et circonstants. Ce travail a porté sur les lexies de nature verbale et uniquement sur le français. Il existe également

[3] http://olst.ling.umontreal.ca/dicoinfo

[4] Les terminologues tiennent compte des circonstants uniquement dans les annotations de contextes. Les circonstants permettent de saisir des nuances sémantiques pouvant aider à la production de textes spécialisés et servant à mieux identifier les équivalences inter linguistiques (Pimentel et al., 2011).

des travaux qui portent sur l'anglais et d'autres langues, lesquels sont basés sur la ressource FrameNet de l'anglais. SemEval-2007[5] a été consacré aux rôles sémantiques et consistait à annoter les rôles sémantiques des unités lexicales en anglais. Plusieurs auteurs ont proposé des modèles statistiques ou d'apprentissage machine afin d'automatiser la tâche d'annotation (Palmer *et al*. 2010, Màrquez *et al*. 2008). Ces modèles reposent sur des traits de description qui sont basés, dans la majorité des cas, sur des arbres syntaxiques (Gildea *et al.* 2002, Johanson *et al*. 2005, Surdeanu *et al*. 2003).

Pour les langues dont les annotations sont inexistantes, on exploite les ressources disponibles pour d'autres langues, par exemple l'anglais, afin de construire un corpus d'annotations sur lequel peuvent se baser les modèles d'apprentissage machine. Dans le cas de la langue française, on a proposé une approche de projection d'annotations de l'anglais (FrameNet) vers le français (Padò *et al*., 2007). Dans notre cas, nous n'avons pas recours à des ressources parallèles de ce genre. Nous avons l'avantage de disposer d'un corpus de données annotées manuellement par des spécialistes terminologues. Une centaine de lexies prédicatives verbales du domaine de spécialité de l'informatique et de l'Internet en français, ayant une quinzaine de contextes chacune, sont annotées avec le plus grand soin.

Nous avons expérimenté un modèle d'apprentissage machine sur ces lexies annotées manuellement. Nous nous sommes inspirés des travaux faits sur l'anglais et nous avons proposé un modèle basé sur des traits syntaxiques que nous avons définis à partir des données de notre corpus. Nous avons utilisé Weka qui fournit des implémentations des algorithmes d'apprentissage les plus connus que nous pouvons appliquer sur les contextes des lexies verbales de notre corpus, en utilisant des traits de classification construits à partir des informations syntaxiques pour retrouver les participants actants et circonstants. Nous avons réalisé une classification binaire des participants (Oui/Non). Ensuite, nous avons testé plusieurs classificateurs de Weka et nous avons opté pour le classificateur RandomForest celui-ci ayant donné de meilleurs résultats. Ce classificateur a été testé selon deux approches :

1) en utilisant les dépendances syntaxiques entre les unités du contexte avec la lexie verbale en étude. Dans ce cas, nous avons utilisé l'analyseur syntaxique Syntex;

2) sans utiliser les liens syntaxiques de Syntex.

Le classificateur utilisé est basé sur des traits de classification que nous avons proposés pour les deux tâches et selon les deux approches décrites ci-dessus (Hadouche *et al*. 2010).

3.1 Traits de classification des participants

Pour classifier les participants, nous avons utilisé les traits suivants :

Lexie : Unité lexicale verbale à l'étude (c.-à-d. : accéder, imprimer, etc.)
CatLexie : Catégorie grammaticale de **Lexie** (c.-à-d. : VInf, VConj, VPpa, etc.)
Mot : Unité lexicale candidate comme participant (dans la plupart des cas, il s'agit de la tête d'un syntagme) dans le contexte où la lexie apparaît.
CatMot : Catégorie grammaticale du **Mot** (c.-à-d. Nom, Adverbe, Pronom, etc.)
Position : Position du **Mot** par rapport à **Lexie**. Sa valeur est « avant » si **Mot** apparaît avant **Lexie** et elle est « après » si **Mot** est après **Lexie**.
Distance : Le nombre de mots qui séparent **Lexie** du **Mot.**

Dans le cas où nous utilisons l'analyseur syntaxique Syntex, nous avons proposé deux autres traits dépendant des liens syntaxiques. Ces traits sont :

Lien-syntaxique-cg: Chemin de **Lexie** à **Mot**. Ce chemin est un ensemble de liens syntaxiques trouvés par Syntex allant de **Lexie** jusqu'à **Mot**. Il s'agit d'une combinaison de toutes les catégories grammaticales des mots qui se retrouvent sur ce chemin (ie : Nom, Prep, Adv, etc.).
Lien-syntaxique-fn : Dans ce cas le chemin de liens syntaxiques entre **Lexie** et **Mot** est une combinaison de toutes les fonctions syntaxiques de ces liens (SUJ, OBJ, etc.)

Dans le cas où nous n'utilisons pas l'analyseur syntaxique Syntex, nous ajoutons plutôt les traits suivants : 1) les **catégories grammaticales** des mots séparant le mot candidat de la lexie, 2) le **nombre de verbes** entre la lexie et le mot candidat s'ils existent, 3) la **valeur du verbe ou sa catégorie**, 4) la **valeur de la préposition,** du **pronom relatif** et de la

[5] 4th International Workshop on Semantic Evaluations (http://nlp.cs.swarthmore.edu/semeval/tasks/)

coordination s'ils existent entre la lexie et le mot candidat.

3.2 Traits de classification des actants et des circonstants

Les participants identifiés à la section 3.1 sont soit des noms, des pronoms ou des adverbes. Les participants régis par une préposition sont difficiles à distinguer, étant donné qu'ils peuvent être actants pour une lexie et circonstants pour une autre. Dans ce cas, nous avons proposé de calculer la fréquence relative entre la lexie et son participant prépositionnel. Cette notion de fréquence relative est utilisée par Messiant dans la réalisation du lexique « LexSchem » présentant la sous-catégorisation des verbes français (Messiant *et al.*, 2008).

On obtient la fréquence relative en calculant le rapport entre le nombre de contextes où la lexie verbale apparaît avec une même préposition P et le nombre total de contextes où la lexie apparaît. Cette fréquence est donnée par :

$$fréquence\_relative = \frac{\#(Lexie, P)}{\#(Lexie)}$$

Si cette fréquence est élevée, elle peut indiquer que la relation sémantique entre la lexie verbale et ce participant est très étroite, et donc que ce participant est vraisemblablement un actant. En revanche, si elle est faible, cela indique que la relation sémantique n'est pas étroite et que ce participant est probablement un circonstant. Nous avons donc suggéré de prendre la fréquence relative comme un trait dans la classification binaire (actant/circonstant) en plus des traits proposés en 3.1

Le classificateur RandomForest a été entrainé et testé sur le corpus des annotations manuelles. En considérant une validation croisée de 10 *folds*, il a donné pour la tâche d'identification des participants une F-mesure de 86 % en utilisant l'analyseur syntaxique Syntex et une F-mesure de 76 % sans utiliser l'analyseur syntaxique. Et pour la tâche de distinction entre actants et circonstant, il a donné une F-mesure de 96 % en considérant les dépendances de l'analyseur syntaxique Syntex et une F-mesure de 94 % sans utiliser ces dépendances syntaxiques.

4 Évaluation

L'évaluation de l'identification automatique des participants actants et circonstants des lexies verbales dans des contextes a été divisée en deux étapes : la première a porté sur le repérage des erreurs de la méthode automatique (section 4.1), tandis que la deuxième a visé à calculer la différence de temps entre une annotation manuelle et la correction des sorties automatiques (section 4.2).

4.1 Évaluation en performance

Cette évaluation a consisté à repérer les erreurs de l'identification automatique des actants et circonstants de lexies verbales dans le but d'apporter des rectificatifs au système dans les cas où c'est possible. L'évaluation de la tâche d'identification automatique a été confiée à deux terminologues. On leur a demandé de vérifier s'ils annoteraient les contextes de la même façon que le système automatique.

4.1.1 Méthode d'évaluation

Les terminologues ont d'abord reçu un fichier .XML contenant 500 contextes de 40 unités prédicatives annotées automatiquement en faisant appel à Syntex. Afin de réviser les sorties automatiques, les terminologues se sont basés sur les renseignements déjà décrits dans la base de données DiCoInfo, notamment la structure actancielle des unités prédicatives (cf. section 2). Les terminologues ont vérifié : 1) si les participants des unités prédicatives avaient été tous bien identifiés (les actants et les circonstants des termes) ; 2) si la réalisation du participant était complète; 3) si, le cas échéant, les antécédents de ces mêmes participants avaient été repérés (par exemple, les antécédents des pronoms) ; 4) et, si les groupes et fonctions syntaxiques de chaque participant avaient été correctement attribués (par exemple, Syntagme Nominal et Sujet du verbe).

Avec cette méthode les terminologues étaient en mesure d'identifier les erreurs commises par le système automatique. Les terminologues ont consigné toutes les informations relatives à l'annotation automatique dans un fichier Excel.

Les erreurs les plus courantes commises par le système sont les suivantes :

1. participants absents ;
2. antécédents absents ;
3. réalisations des participants incomplètes.

Participants absents

Dans le contexte ci-dessous, le système a repéré le premier participant de l'unité prédicative *journaliser* (*on*) correspondant à un actant, mais il n'a pas repéré les trois autres participants du

verbe, soit un second actant (*les actions*) et deux circonstants (*en utilisant le système de fichier* et *en utilisant le système de messages du middleware*).

[1] *Par exemple, on peut* ***journaliser*** ~~*les actions*~~ ~~*en utilisant le système de fichier*~~ *ou* ~~*en utilisant le système de messages du middleware*~~.

Antécédents absents

Dans certains contextes, les participants des unités prédicatives apparaissent sous la forme de pronoms. Le système d'annotation automatique est censé identifier correctement les unités lexicales, dans le contexte, auxquelles les pronoms font référence, voire les antécédents.

[2] *Quelques soucis ont été cependant rapportés avec* ~~*Pine*~~ *qui semble gérer difficilement les certificats venant d'une autorité qu'il ne* ***reconnait*** *pas.*

Dans ce cas, le système a repéré le premier participant de l'unité prédicative (*qu'*) et il a également correctement identifié son antécédent (*une autorité*). Pourtant, bien que le deuxième participant ait bien été repéré (*il*), le système n'a pas relevé son antécédent (*Pine*).

Réalisations incomplètes

Dans l'exemple [3], le système a bien identifié le premier participant, mais le syntagme nominal qui compose ce participant est incomplet dans la mesure où le déterminant *certains* n'a pas été repéré.

[3] *Dans la mesure où le CMOS est une mémoire lente,* ~~*certains*~~ *systèmes* ***recopient*** *parfois le contenu* ~~*du CMOS*~~ *dans la RAM (mémoire rapide).*

4.1.2 Résultats

Les terminologues évaluateurs ont fourni un fichier Excel composé de plusieurs colonnes correspondant aux éléments qui ont été évalués et les corrections proposées. Afin de calculer le taux de performance du système, nous nous sommes basés sur les valeurs de certaines colonnes. La colonne *antécédent* contenait « oui » si l'antécédent avait été identifié correctement et par « non » si ce n'était pas le cas. La colonne *participant* contenait « oui » si le participant avait été bien identifié, par « non » si le participant identifié n'était pas correct et par « absent » s'il n'était pas identifié du tout. Les colonnes *actant* et circonstant contenaient « oui » si l'actant ou le circonstant était correctement identifié et par « non » dans le cas contraire. On a procédé de la même manière en ce qui a trait à la colonne *fonction syntaxique*.

Nous avons calculé la précision en utilisant le nombre de « oui » par rapport au total du nombre de « oui » et de « non ». Le rappel correspond au rapport entre le nombre de « oui » et le nombre total de « oui » et d'« absent ». La F-mesure a donné un taux de 80 % de participants correctement identifiés. Bien que ce taux soit très bon, il n'est pas dissocié du temps attribué pour vérifier et valider les annotations. Donc, il est nécessaire de mesurer ce temps pour pouvoir évaluer l'apport du système automatique au travail du terminologue.

4.2 Évaluation en temps

La deuxième partie de l'évaluation a porté sur la comparaison du temps nécessaire à un terminologue pour annoter les contextes manuellement et celui requis pour corriger les sorties automatiques.

4.2.1 Méthode

Nous avons fait appel à deux annotateurs et un réviseur. Deux annotateurs-terminologues (A et B), les mêmes ayant participé à la première étape de l'évaluation, ont procédé à l'annotation manuelle des contextes liés à cinq unités prédicatives verbales (*acheminer*, *allouer*, *connecter*, *formater* et *relancer*) comprenant vingt contextes chacune. L'annotateur A est un locuteur du français langue seconde avec assez d'expérience en annotation de contextes. L'annotateur B est un locuteur francophone avec moins d'expérience d'annotation. Le réviseur est un locuteur francophone avec une grande expérience d'annotation et de révision d'annotation. Les annotations ont été faites dans un fichier .XML soit de la même façon dont les terminologues associés au DiCoInfo travaillent habituellement. Le réviseur, quant à lui, a procédé à la correction d'un fichier .XML dans lequel les contextes des mêmes unités prédicatives avaient été annotés automatiquement. Il a vérifié si le système avait identifié tous les participants relatifs à la structure actancielle d'une unité prédicative donnée et si ces participants avaient été annotés correctement. Le cas échant, le réviseur a ajouté les participants oubliés et a corrigé ceux comportant des erreurs. Afin que les données ne soient pas biaisées, les annotateurs n'ont pas reçu le fichier .xml déjà

annoté automatiquement. Le réviseur a procédé à la correction de l'annotation dans ce fichier sans communiquer avec les annotateurs. Les annotateurs ont annoté manuellement le fichier côte à côte sans toutefois se consulter.

4.2.2 Résultats

Le temps d'identification manuelle et automatique est mesuré en minutes. Nous avons mesuré le temps requis par les annotateurs A et B afin d'effectuer l'annotation ainsi que celui du réviseur afin de faire la correction des sorties automatiques.

Nous avons constaté que le temps requis par les annotateurs et le réviseur est variable. On peut expliquer ces variations selon trois facteurs. D'abord, les temps varient en fonction des verbes, c'est-à-dire que certains verbes demandent plus de temps d'annotation et de révision que d'autres. Par exemple, un terme comme *acheminer* est long à annoter ou à réviser, car il a plusieurs actants souvent exprimés sous la forme de pronoms avec des antécédents en contexte. Ensuite, les temps varient d'un annotateur à l'autre. Du fait de son expérience, le temps d'annotation de l'annotateur A est toujours inférieur à celui de l'annotateur B. Enfin, les temps varient selon que l'on utilise l'identification manuelle des actants et des circonstants ou l'identification automatique. En effet, le temps nécessaire à la correction des sorties automatiques effectuée par le réviseur est toujours inférieur à celui qui est nécessaire à l'annotation manuelle des contextes.

Le Tableau 2 présente le gain de temps fournit par le système automatique correspondant à chaque lexie verbale par rapport à chaque annotateur.

Gain/annotateur / Lexies	Gain/A (%)	Gain/B (%)
Relancer	57	90
Acheminer	42	60
Allouer	50	64
Connecter	43	60
Formater	69	70
Moyenne	52	69

Tableau 2. Gain de temps en utilisant le système automatique par rapport aux temps des annotateurs A et B et la moyenne des gains sur toutes les lexies.

Selon la complexité de la lexie à l'étude et l'expérience de l'annotateur, le gain de temps est proportionnellement significatif.

Ainsi, si l'on se fie au gain de temps par rapport au verbe *formater* (verbe considéré comme étant plus facile à annoter que les autres pour des raisons mentionnées précédemment), on peut effectivement avancer que le temps requis par le terminologue pour réviser les contextes de trois lexies correspond au temps requis pour en annoter une seule manuellement. De façon similaire, si l'on se fie au temps requis en minutes, le terminologue pourrait réviser tous les contextes annotés automatiquement des cinq lexies à l'étude, pour un temps total de 94 minutes, au lieu d'annoter manuellement les deux lexies les plus difficiles de l'échantillon (*relancer* et *acheminer*), ce qui prendrait 87 minutes pour l'Annotateur A et 125 minutes pour l'Annotateur B.

Selon le Tableau 2, le gain de temps par rapport à l'annotateur A, ayant une expérience dans l'annotation, est en moyenne de 52 %. Quant au gain de temps par rapport à l'annotateur B ayant moins d'expérience est en moyenne de 69 %. Ce qui implique en moyenne, par rapport aux deux annotateurs, un gain de 60 %. On peut donc dire qu'il est possible de réviser l'équivalent de trois lexies verbales dans un temps correspondant à l'annotation manuelle des contextes d'une seule lexie. En comparaison avec l'annotation manuelle des contextes, la méthode d'annotation automatique permet donc de gagner, en moyenne, 60 % du temps de travail.

5 Conclusion

Dans cet article nous avons décrit un système d'annotation automatique de contextes et illustre son application aux termes de nature prédicative d'une ressource lexicale spécialisée, le DiCoInfo. Nous avons également proposé une méthode afin d'évaluer la performance de ce système d'annotation et de quantifier le gain de temps si le terminologue utilise ce système dans sa tâche d'identification d'actants et circonstants des lexies prédicatives. L'évaluation a montré que le terminologue peut travailler au moins deux fois plus rapidement que lorsque cette tâche n'est pas automatisée. Selon les résultats de l'évaluation, le terminologue gagne un temps considérable en révisant les sorties automatiques.

Ce travail nous a également permis de montrer que, peu importe si la méthode est manuelle ou automatique, le temps d'annotation des contextes illustrant le comportement linguistique des termes dépend de la difficulté du terme lui-même.

Les annotateurs font souvent face au même genre de problèmes : ambiguïtés syntaxiques, verbes ayant plusieurs actants exprimés en contexte, antécédents. En fait, c'est la combinaison plus ou moins marquante de ces facteurs qui détermine le temps d'annotation nécessaire. Ainsi, un terme comme *acheminer*, dont la structure actancielle se compose de quatre actants, demande presque deux fois plus de temps d'annotation qu'un verbe tel que *relancer* qui n'en a que deux.

L'annotation manuelle de contextes reste une tâche fastidieuse pour les annotateurs et très exigeante en temps. Le système automatique proposé facilite la tâche de l'annotateur. Cette tâche d'annotation, qui est devenue une validation des annotations faites par notre système automatique, évite donc une grande partie du travail routinier pour la majorité des cas simples et permet aux terminologues de se concentrer sur les cas plus difficiles.

Dans le futur, nous souhaitons évaluer notre système sur un nombre plus important de verbes. En outre, nous pensons qu'il serait intéressant d'étendre ce travail pour d'autres domaines de spécialité et d'autres langues.

Remerciements

Nous remercions le Conseil de recherches en sciences humaines (CRSH) du Canada qui a financé ce projet.

Références

Bourigault, D., Un analyseur syntaxique opérationnel: Syntex, Mémoire d'Habilitation à Diriger les Recherches, Laboratoire CLLE-ERSS, CNRS & Université de Toulouse-le Mirail, 2007, 158 p.

DiCoInfo. Dictionnaire fondamental de l'informatique et de l'Internet. (http://olst.ling.umontreal.ca/dicoinfo)

Fillmore, C., The Case for Case. In Bach and Harms (Ed.): Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1968, 1-88.

Gildea, D. et Jurafsky, D., Automatic labeling of semantic roles. Computational Linguistics, 2002, 28(3): 245-288.

Hadouche, F. Annotation syntaxico-sémantique des actants en corpus spécialisé. Thèse PhD, Université de Montréal, 2010, pp 155.

Hadouche, F., Lapalme, G. et L'Homme, M.C., Identification des actants et circonstants par apprentissage machine, TALN 2010, Université de Montréal, Québec, 2010.

Johanson, R. et Nugues, P., Sparse bayesian classification of predicate arguments. Proceedings of CoNLL'05, Ann Arbor, Michigan, 2005, p. 117-180

L'Homme, M.C., Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. Traduire, 2008, 217: 78-103.

Màrquez, L., Carreras, X., Litkowski, K.C. et Stevenson, S., Special issue on semantic role labeling. Computational Linguistics, 2008, 34(3) :317

Messiant, C., Korhonen, A. et Poibeau, T., LexSchem. A large subcategorization lexicon for french verbs. In LREC 2008 Proceedings. Marrakech, Maroc, 2008.

Padò, S. et Pitel, G., Annotation précise du français en sémantique de rôles par projection cross- linguistique. Actes de la conférence TALN, Toulouse, France, 2007.

Palmer, M., Gildea, D. et Xue, N., Semantic role labeling. Synthesis Lectures on Human Language Technologies, 2010, p. 103.

Pimentel, J. et L'Homme, M.C., Annotation syntaxico-sémantique de contextes spécialisés: application à la terminographie bilingue. In Van Campenhoudt, et al. (Ed.): Passeurs de mots, passeurs d'espoir. Lexicologie, terminologie et traduction face au défi de la diversité, 2011, p. 651-670.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. et Scheffczyk, J., FrameNet II: Extended Theory and Practice. (http://framenet.icsi.berkeley.edu/).

Surdeanu, M., Harabagiu, S., Williams, J. et Aarseth, P., Using predicate-argument structures for information extraction. Proceedings of the ACL '03, Sapporo, Japan, 2003, p. 8-15.

Witten, I. et Frank, E., Data mining: Practical machine learning tools and techniques. 2nd Edition. San Francisco: Morgan Kaufmann, 2005.