Typage sémantique de verbes avec LVF, pour la résolution d'anaphores

Elisabeth Godbert
Aix-Marseille Université, LIF-CNRS UMR 7279, 163 avenue de Luminy, 13288 Marseille Cedex 9
Elisabeth.Godbert@lif.univ-mrs.fr

Résumé. Le travail présenté ici s'intéresse à la détection automatique de relations de coréférence dans un corpus de dialogues oraux enregistrés dans le centre d'appel de la RATP ; chaque dialogue du corpus met en interaction un opérateur et un client, et la majorité des relations de coréférence sont des anaphores pronominales sur les entités dont parlent ces deux personnes. Par exemple : *J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ? - Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.*
On observe que la sémantique des entités dont on parle est un trait essentiel à prendre en compte, et que dans de nombreux cas il n'y a pas conservation des traits de genre et nombre entre l'antécédent et sa reprise anaphorique.
Nous décrivons les choix que nous avons faits pour typer sémantiquement les verbes en utilisant des données du dictionnaire électronique LVF, et pour effectuer manuellement la classification des noms. Puis nous présentons les méthodes que nous avons mises en oeuvre pour la résolution des anaphores pronominales.

Abstract. This paper focuses on automatic detection of coreference and anaphoric relations in a corpus of dialogues recorded in the call-center of the RATP. The majority of the coreference relations are third person anaphoric personal pronouns, for example : *I lost my portable phone in the bus 45, where may I hope to get it back ? - Phone to the Lost property service, they will tell you if it was brought back.*
It can be seen that it is essential to take into account the semantic types of entities mentioned in the dialogue, and that in numerous cases there is no preservation of gender and number features between the antecedent and the anaphoric element. We describe how we have defined semantic types for verbs, by using data of the electronic dictionary LVF, and manually classified nouns. Then we describe the methods used for the resolution of pronominal anaphora.

Mots-clés : coréférence, anaphore, typage sémantique de verbes, LVF (Les Verbes Français).

Keywords: coreference, anaphora, semantic types of verbs, LVF (Les Verbes Français).

1 Introduction

La détection des relations de coréférence, ou anaphores, permet le suivi des entités mentionnées dans les documents. Elle est nécessaire, par exemple, pour l'extraction d'information et le résumé automatique de textes, qui sont des domaines très actifs en TAL. Par ailleurs, la constitution de corpus annotés permet le développement et l'entraînement de modèles statistiques de TAL.
Le travail présenté ici s'intéresse à l'enrichissement d'un corpus annoté, par la détection automatique de relations de coréférence. Nous travaillons sur le corpus RATP-DECODA (Bechet *et al.*, 2012), élaboré dans le cadre du projet DECODA ("DEpouillement automatique de COnversations provenant de centres D'Appels"), dont le cadre applicatif est le centre d'appel de la RATP. Créé à partir de l'enregistrement de bases de données de messages oraux de très grande taille contenant des interactions entre un opérateur et un client, ce corpus est composé de 2100 dialogues et correspond à environ 74 heures d'enregistrement, avec un total d'environ 600 000 mots. Ces dialogues ont été transcrits manuellement, puis plusieurs phases de traitement ont permis d'obtenir leur annotation linguistique à plusieurs niveaux : détection des disfluences, repérage des entités nommées, découpage en parties de discours et chunks, et production d'une analyse syntaxique en dépendances par le système MACAON (Bechet *et al.*, 2012; Bazillon *et al.*, 2012; Nasr *et al.*, 2011).

Chaque dialogue du corpus mettant en interaction un opérateur et un client de la RATP, la majorité des relations de coréférence sont des anaphores pronominales à la troisième personne, qui portent sur les entités dont parlent les deux interlocuteurs.

Par exemple :

1. *Mon fils a eu un problème avec son bus ce matin ; il a perdu plus de 30 minutes à attendre.*
2. *Mon fils a eu un problème avec son bus ce matin ; il est passé avec 30 minutes de retard.*
3. *J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ?*
Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.

On voit sur ces exemples que la sémantique des entités dont on parle est un trait essentiel à prendre en compte, et qu'il n'y a pas toujours conservation des traits de genre et nombre entre l'antécédent et sa reprise anaphorique. La recherche de l'antécédent du pronom (*il, le, ils*) ne pourra aboutir que si l'on a préalablement fait un typage sémantique des verbes et de leurs actants, ainsi que des noms qui apparaissent dans le dialogue et qui seront des antécédents potentiels.

Le corpus RATP-DECODA relevant d'un domaine applicatif très particulier, le contenu des dialogues à traiter ne contient qu'un nombre restreint de noms et de verbes : on y compte (en lemmes distincts) 2232 noms communs et 1024 verbes.

Nous montrons dans la section 2 comment nous avons établi un typage sémantique de ces 1024 verbes en utilisant les données du dictionnaire électronique LVF "Les Verbes Français", dans lequel on trouve pour chaque verbe ses différents sens et constructions, et (parmi d'autres) des informations d'ordre sémantique.

Nous décrivons ensuite dans la section 3 les méthodes que nous avons mises en oeuvre pour traiter les anaphores pronominales des types suivants :
1. Reprise anaphorique par un pronom personnel à la troisième personne (à l'exclusion des pronoms réflexifs) : *le bus 105 ... il va en direction de ...*
2. Reprise anaphorique associative pronominale : *le Service clientèle du RER ... ils sont en pause déjeuner*.
3. Relation cataphorique entre un pronom personnel et son référent : *il a été sympa, le chauffeur*.

2 Typage des noms et des verbes

2.1 LVF : Les Verbes Français

LVF est un dictionnaire électronique développé par J. Dubois et et F. Dubois-Charlier (Dubois & Dubois-Charlier, 1997) et disponible librement [1]. Il contient 25 610 entrées verbales représentant 12 310 verbes différents, dont 4 188 à plusieurs entrées. La classification des verbes repose sur l'hypothèse qu"'il y a adéquation entre les schèmes syntaxiques de la langue et l'interprétation sémantique qu'en font les locuteurs de cette langue". Ce dictionnaire contient, pour chaque verbe, les informations suivantes :

1. la classe selon certains principes de classification,
2. le sens donné par un synonyme, un parasynonyme, une définition ou une explication,
3. le domaine d'emploi principal (géologie, psychologie, etc.) et le niveau de langue,
4. la conjugaison et l'auxiliaire,
5. la syntaxe du verbe : intransitif, transitif direct ou indirect, pronominal ; nature des sujets et des compléments,
6. les dérivations (noms d'action, d'instrument, d'agent, de résultat, adjectifs verbaux),
7. les termes (nom ou adjectif) dont le verbe est éventuellement lui-même dérivé,
8. le type de dictionnaire où l'entrée est répertoriée,
9. et à chaque entrée sont associées une ou plusieurs phrases simples, illustrant le sens et la construction syntaxique.

Ces informations nous fournissent des éléments très pertinents pour effectuer un typage sémantique des verbes. En particulier, pour chaque entrée verbale :
- L'information du point 1 ci-dessus, dite "classe", code la classe sémantique à laquelle appartient cette entrée verbale, par exemple "verbe de communication", "verbe de mouvement", etc. Il existe 54 classes, elles-mêmes découpées en sous-classes et sous-types.
- L'information du point 5, dite "syntaxe du verbe", donne le nombre d'actants ou compléments du verbe et la nature de chaque actant : *humain, animal, chose, complétive*, etc.

1. http ://talep.lif.univ-mrs.fr/FondamenTAL/

2.2 Classification des noms

La classification des 2232 lemmes de noms communs du corpus RATP-DECODA a été faite manuellement, dans une hiérarchie très simple, dans laquelle apparaissent en premier lieu : la classe "Tout", divisée en "Humain" et "Non-Humain", puis les sous-classes "Véhicule", "Objet-Concret", "Objet-Abstrait".
Ensuite d'autres sous-classes sont définies, associées aux types d'entités nommées qui ont été répertoriées dans le corpus : pour ce qui concerne les noms propres et les entités nommées, le corpus RATP-DECODA nous fournit l'annotation de ces entités avec leur nature (*date, adresse, organisation*, etc.). Ceci nous permet de classer ces entités automatiquement dans la hiérarchie. En particulier, tout ce qui est répertorié comme une organisation (SNCF, RATP, le Service des objets trouvés, etc.) est placé dans une sous-classe de la branche "Humain" : en effet, le corpus contient de nombreuses reprises anaphoriques associatives du type de *Téléphonez au Service des objets trouvés, ils vous diront...*

2.3 Typage des verbes

Comme dit précédemment, nous avons choisi de faire un typage sémantique des verbes à partir des données de LVF.
L'une des difficultés est que dans LVF un verbe peut avoir plusieurs (et même de nombreuses) entrées, correspondant chacune à une construction ou un sens particulier. Comme notre domaine applicatif est très particulier, nous avons choisi d'élaborer un typage relativement simple dans lequel chaque verbe n'a qu'une entrée, qui correspond à l'usage de ce verbe dans le corpus.

Pour faciliter cette opération de typage, nous avons procédé de la façon suivante, en effectuant d'abord un traitement automatique (points 1 et 2 ci-dessous) à partir des données LVF, puis une post-interprétation manuelle (points 3 et 4) :
1. Pour chacun des 1024 verbes, nous avons rassemblé tous les types sémantiques éventuels de ses sujets et compléments d'objet direct en utilisant les champs "OPERATEUR" et "CONSTRUCTION" de LVF, et en gardant l'information "Humain", "Non-Humain", ou "Tout" ;
2. Nous en avons fait une synthèse, pour en tirer ce que nous appelons le type sémantique de base du verbe : nous avons gardé pour le sujet le type qui est donné dans la première entrée du verbe, et pour le complément d'objet direct la classe minimale qui couvre les types de tous les compléments ;
3. Puis, pour chaque verbe, nous avons vérifié manuellement que dans le contexte du corpus DECODA le typage automatique obtenu par 1 et 2 était pertinent, et, si besoin, nous l'avons modifié ;
4. Pour les verbes qui admettent un complément d'attribution, nous en avons ajouté manuellement le type.

A l'issue de ce traitement, nous avons obtenu un typage des verbes et de leurs actants où apparaissaient trois types : "Humain", "Non-Humain", "Tout".
Le tableau 1 montre les trois premières étapes du typage des verbes *boire* et *contrôler* . "Humain" y est noté "1", "Non-Humain" y est noté "3" et "Tout" y est noté "9".
Pour le verbe *boire*, le résultat obtenu dans la synthèse a été modifié, mais uniquement pour le complément d'objet, car nous avons jugé que dans notre application il n'y avait pas lieu de tenir compte des deux derniers sens du verbe illustrés par les exemples *Le buvard boit l'encre. On est bu après cette réunion.* (ce dernier sens est d'usage populaire).

Extrait de LVF	boire 01 ;(#) ;T1300 - A10 ; boire 02 ;(#) ;T1300 - A10 ; boire 03 ;(qc) ;T3306 ; boire 04 (être) ;(#) ;A10 - T3100 ;	contrôler 01 ;(#) ;T1400 ; contrôler 02 ;(#) ;T1900 ; contrôler 03 ;(#) ;T1900 - P1000 ; contrôler 04 ;(#) ;T1300 ; contrôler 05 ;(#) ;T1900 ; contrôler 06 ;(#) ;T1308 ;
Nature des actants sujets et objets	boire ;-1-1-1-1-qc-3-1-3- ;-3-3-3-1- ;	contrôler ;-1-1-1-P1-1-1-1- ;-4-9-9-3-9-3- ;
Synthèse	boire ;1 ;9 ;	contrôler ;1 ;9 ;
Modification éventuelle	boire ;1 ;3 ;	

TABLE 1 – Les trois premières étapes du typage sémantique des verbes

Les verbes de mouvement (*monter, descendre, avancer, passer,...*) ont de très nombreuses occurrences dans le corpus. Pour traiter correctement les anaphores pronominales qui sont des actants de ces verbes, nous en avons raffiné le typage : en utilisant l'information "classe" de LVF, nous avons repéré dans notre corpus tous les verbes de mouvement et verbes locatifs, puis, pour ceux d'entre eux qui peuvent avoir pour sujet une personne ou un véhicule, nous avons affecté à leur

actant sujet le type mixte "Humain-ou-Véhicule".
Le tableau 2 donne un extrait du tableau final de typage des verbes.

verbe	sujet	compl-objet-direct	compl-attribution
atteindre	Humain-ou-Véhicule	Tout	
attendre	Humain-ou-Véhicule	Tout	
atterrir	Tout		
attester	Humain		Humain
attraper	Humain	Tout	
attribuer	Humain	Non-Humain	Humain
augmenter	Tout	Non-Humain	Humain
autoriser	Humain	Tout	Humain

TABLE 2 – Un extrait du tableau de typage sémantique des verbes

Notons qu'apparaissent dans le corpus une dizaine de verbes très familiers et spécifiques du domaine, qui ne sont pas répertoriés dans LVF, dont : *bugger, checker, gourrer, recréditer, redispatcher, remagnétiser, repoinçonner,....*

3 Résolution d'anaphores

3.1 Résolution d'anaphores pronominales

Notre objectif est ici de traiter les anaphores pronominales à la troisième personne, c'est-à-dire d'identifier les antécédents, ou éventuellement des pronoms coréférents, des pronoms personnels *il, elle, ils, elles, le, la, l', les, lui, leur*.

Comme nous l'avons dit dans la section 1, nous avons choisi de donner un poids essentiel à la sémantique pour la recherche des antécédents de pronoms, entre autres parce que l'on constate qu'il n'y a pas toujours conservation des traits de genre et de nombre entre un nom et le pronom qui le reprend. Un antécédent ne sera donc retenu que s'il est de type sémantique compatible avec le pronom sur lequel on travaille.
Considérons l'exemple suivant :
- j'ai un passe, je voudrais le faire refaire ;
- il faut passer dans une Agence Intégrale ;
- oui mais je ne sais pas où elles sont en fait ;
- [...]
- allez sur place, ils vont vous refaire un passe.
Notre système est conçu pour donner ici les résultats suivants : l'antécédent de *elles* est *Agence* et l'antécédent de *ils* est également *Agence*.

Le traitement de ces pronoms *il, elle, ils, elles, le, la, l', les, lui, leur* est effectué en plusieurs étapes :

1. On identifie les pronoms à traiter, ce qui revient, à l'inverse, à identifier les pronoms *il* et *le* que nous ne traiterons pas car ils sont utilisés dans des formes impersonnelles : *il y a*, *il faut*, *je le sais*, etc. L'identification de ces formes impersonnelles est faite via les relations syntaxiques données dans le corpus, et les pronoms associés se voient affecter le type "imp".

2. Pour chaque pronom non typé "imp", les liens de dépendance nous permettent de trouver le verbe dont il est actant, et son rôle (sujet, objet direct, complément d'attribution). En utilisant le typage des verbes (voir 2.3) nous attribuons au pronom un type sémantique.

3. La recherche de l'antécédent de chaque pronom se fait en remontant dans le dialogue, et en y cherchant une entité de type sémantique compatible. Cette recherche se fait en plusieurs passes, chacune d'elles remontant plus ou moins loin dans le dialogue. En particulier :
a) dans un premier temps on recherche un nom de même type, de même genre et de même nombre que le pronom, en ne remontant que 40 mots ou 10 tours de parole ;
b) puis, si cette première recherche n'a pas abouti, on assouplit peu à peu les contraintes sur le nom, pour finalement, au bout de plusieurs passes, ne garder que la contrainte sur le type sémantique si les passes précédentes ont été infructueuses ; pour cette passe ne portant que sur le type sémantique, on remonte encore à 40 mots ou 10 tours de parole ;

c) si c'est toujours infructueux, on recherche de nouveau un nom de mêmes type, genre et nombre que le pronom, mais en remontant beaucoup plus haut (100 mots ou 30 tours de parole).
Par ailleurs, une autre passe est intercalée dans le b), dans laquelle on recherche une entité coréférente sous la forme d'un pronom et non d'un nom, en ne remontant que très peu dans le dialogue (20 mots ou 4 tours de paroles) ; si la recherche est fructueuse, les deux pronoms sont notés "coréférents".

Reprenons les trois exemples donnés dans l'introduction :

1. *Mon fils a eu un problème avec son bus ce matin ; il a perdu plus de 30 minutes à attendre.*
Le sujet du verbe *perdre* est de type "Humain", l'antécédent ne peut être que *fils*.

2. *Mon fils a eu un problème avec son bus ce matin ; il est passé avec 30 minutes de retard.*
Le sujet du verbe *passer* est de type "Humain-ou-Véhicule", l'antécédent est le mot le plus proche : *bus*.

3. *J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ?*
Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.
Le complément d'objet direct du verbe *récupérer* est de type "Non-Humain-Concret" (disjoint de "Véhicule"), l'antécédent de *le* est le mot *portable*.
Le sujet du verbe *dire* est de type "Humain", ce typage permet de trouver que l'antécédent de *ils* est le Service des objets trouvés ; l'antécédent de *il* est *portable* car le complément du verbe *rapporter* est de type "Non-Humain-Concret".

Si l'on n'a trouvé aucune entité acceptable comme antécédent ou coréférent du pronom, le système indique l'échec de sa recherche pour ce pronom. En particulier, la recherche échoue :
a) lorsque l'antécédent existe mais est trop éloigné dans le dialogue ;
b) lorsqu'il n'y a ni antécédent ni pronom coréférent, comme pour le premier *ils* dans :
- il y a un préavis de grève aujourd'hui ;
- oh, ils nous cassent les pieds ; je sais que vous n'y êtes pour rien mais ils nous cassent les pieds.
Ici, les prédictions du système sont :
- le premier *ils* n'a ni antécédent ni coréférent ;
- le deuxième *ils* n'a pas d'antécédent non plus, mais il est noté coréférent du premier.

3.2 Résolution de relations de type cataphore

Le cas des cataphores est très particulier. Une cataphore est une anaphore où la reprise sémantique est située avant son antécédent (que l'on peut appeler conséquent), comme *Vous l'avez acheté où le ticket ?*. Or, dans le corpus sur lequel nous travaillons, l'annotation syntaxique en dépendances permet d'identifier immédiatement les cataphores. Par exemple, dans l'analyse de *Vous l'avez acheté où le ticket ?*, les mots *l'* et *ticket* sont tous les deux identifiés comme complément d'objet de *avez acheté*. De même, dans l'analyse de *Il passe devant l'hôpital le bus*, *Il* et *bus* sont tous les deux sujets de *passe*.
Notre système se contente donc d'utiliser ces dépendances pour afficher les relations de type cataphore.
En fait, cette recherche est la première qui est effectuée par le système. Car si un pronom a été identifié comme cataphore d'un nom qui le suit dans le dialogue, il ne faut pas en chercher un antécédent en remontant dans le dialogue. La résolution d'anaphores pronominales décrite en section 3.1 ne se fait donc que sur les pronoms non identifiés comme cataphore.

4 Remarques finales

Nous avons montré comment les données de LVF nous ont fourni des informations très intéressantes pour définir un typage sémantique des verbes du corpus DECODA. Les mêmes méthodes pourraient certainement être utilisées pour d'autres applications, dès lors que leur domaine sémantique n'est pas trop étendu.

Ce travail est en cours, il a débuté assez récemment, et l'on pourra y apporter de nombreuses améliorations.
En particulier, nous envisageons d'élargir les possibilités pour le typage des verbes : nous avons choisi pour le moment de définir pour les verbes un typage simple, dans lequel chaque verbe n'a qu'une entrée, mais on peut penser que c'est insuffisant pour les verbes qui sont couramment utilisés et qui ont de nombreux sens et/ou constructions. Il sera intéressant de voir comment raffiner le typage en attribuant à chaque verbe plusieurs "cadres sémantiques" tirés des données de LVF et pertinents pour notre domaine d'application, et ensuite adapter les processus de traitement des anaphores.

Nous envisageons par ailleurs de traiter d'autres types d'anaphores à plus ou moins long terme, dont :
- En premier lieu les reprises directes et indirectes qui permettent de traiter par exemple *Je voulais prendre le 107 ; mais j'attends ce/le bus depuis une heure.*
- Les anaphores associatives qui demandent de prendre en compte des relations de méronymie, mais c'est un travail complexe. L'entité anaphorique peut y être introduite par un adjectif possessif (*la cliente a perdu sa carte*), ou par un article défini (*le bus a été accidenté ; l'arrière est enfoncé*).

Une évaluation de la version actuelle du système va être faite très prochainement. On utilisera pour cela un sous-ensemble privilégié du corpus RATP-DECODA, composé de 102 dialogues, dit le *GOLD corpus*, qui a déjà été utilisé comme étalon au cours de la phase d'annotation syntaxique (Bechet *et al.*, 2012). En premier lieu, le *GOLD* va être annoté automatiquement en coréférences par notre système ; puis une validation manuelle sera effectuée, qui permettra d'évaluer la qualité des prédictions faites par le système dans son état actuel. Le *GOLD* servira alors de référence, et l'on pourra entrer dans un processus itératif pour tenter d'obtenir un taux acceptable de réussite.

La constitution de corpus annotés, dits corpus de référence, est un enjeu important pour le TAL, car il permet l'entraînement et le test de systèmes statistiques de TAL.
Pour ce qui concerne l'élaboration de corpus annotés en relations de coréférence, on peut en premier lieu citer les travaux de (Tutin *et al.*, 2000), qui ont permis d'annoter manuellement en coréférences un corpus de texte contenant environ un million de mots ; puis les travaux décrits dans (Muzerelle *et al.*, 2014), qui ces denières années ont abouti à la constitution du corpus ANCOR-Centre, composé de trois corpus de parole conversationnelle annotés manuellement en relations de coréférences, sur un total d'environ 450 000 mots.
Ces corpus sont suffisamment larges pour répondre aux besoins des méthodes de TAL basées sur l'apprentissage.

L'annotation manuelle de grands volumes de données étant très longue et coûteuse, le projet ANR ORFEO (Outils et Ressources pour le Français Ecrit et Oral) est parti d'une démarche alternative, avec pour objectif la constitution d'un Corpus d'Etude pour le Français Contemporain (CEFC) qui rassemblera des données à partir de différents corpus. Ces données, obtenues automatiquement ou semi-automatiquement, seront de nature diverse, dont l'alignement texte et son, et des annotations morphologiques, syntaxiques, sémantiques, conversationnelles et prosodiques. Cela correspond à un traitement massif de données écrites ou orales, qui n'est pas parfait mais qui donne accès à un grand volume de données.
Le travail qui a été présenté dans cet article participe au projet ORFEO, par l'ajout de relations de coréférence dans le corpus RATP-DECODA.

Remerciements

Je suis très reconnaissante à Frédéric Béchet et Alexis Nasr pour les conseils et pour l'aide qu'ils m'ont apportés pour ce travail.

Références

BAZILLON T., DELPLANO M., BECHET F., NASR A. & FAVRE B. (2012). Syntactic annotation of spontaneous speech : application to call-center conversation data. In *Proceedings of the 8th international conference on Language Resources and Evaluation, LREC 2012, Istambul, Turkey*.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BÈZE M., MORI R. D. & ARBILLOT E. (2012). DECODA : a call-center human-human spoken conversation corpus. In *Proceedings of the 8th international conference on Language Resources and Evaluation, LREC 2012, Istambul, Turkey*.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). *Les verbes français*. Larousse-Bordas.

MUZERELLE J., LEFEUVRE A., SCHANG E., ANTOINE J.-Y., PELLETIER A., MAUREL D., ESHKOL I. & VILLANEAU J. (2014). ANCOR-Centre, a large free spoken french coreference corpus : Description of the resource and reliability measures. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference, LREC 2014, Reykjavik, Iceland*.

NASR A., BECHET F., REY J. & ROUX J. L. (2011). MACAON : a linguistic tool suite for processing word lattices. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : demonstration session*.

TUTIN A., TROUILLEUX F., CLOUZOT C., GAUSSIER E., ZAENEN A. & ANDG. ANTONIADIS S. R. (2000). Annotating a large corpus with anaphoric links. In *Proceedings of Discourse, Anaphora and Reference Resolution Conference, DAARC-2000, Lancaster, UK*.