Apprentissage de relations prédicat-argument pour l'extraction d'information à partir de textes conversationnels

Narjès Boufaden (1), Guy Lapalme (1)
(1) RALI - Université de Montréal
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal, Québec, H3C 3J7 Canada
{boufaden,lapalme}@iro.umontreal.ca

Mots-clefs : Apprentissage de relations prédicat-argument, extraction d'information

Keywords: Learning predicat-argument relations, information extraction

Résumé Nous présentons les résultats de notre approche d'apprentissage de relations prédicat-argument dans le but de générer des patrons d'extraction pour des textes conversationnels. Notre approche s'effectue en trois étapes incluant la segmentation linguistique des textes pour définir des unités linguistiques à l'instar de la phrase pour les textes bien formés tels que les dépêches journalistiques. Cette étape prend en considération la dimension discursive importante dans ces types de textes. La deuxième étape effectue la résolution des anaphores pronominales en position de sujet. Cela tient compte d'une particularité importante des textes conversationnels : la pronominalisation du thème. Nous montrons que la résolution d'un sous ensemble d'anaphores pronominales améliore l'apprentissage des patrons d'extraction. La troisième utilise des modèles de Markov pour modéliser les séquences de classes de mots et leurs rôles pour un ensemble de relations données. Notre approche expérimentée sur des transcriptions de conversations téléphoniques dans le domaine de la recherche et sauvetage identifie les patrons d'extraction avec un F-score moyen de 73,75 %.

Abstract We present the results of our approach for the learning of patterns for information extraction from conversational texts. Our three step approach is based on a linguistic segmentation stage that defines units suitable for the pattern learning process. Anaphora resolution helps to identify more relevant relations hidden by the pronominalization of the topic. This stage precedes the pattern learning stage, which is based on Markov models that include *wild card* states designed to handle edited words and null transitions to handle omissions. We tested our approach on manually transcribed telephone conversations in the domain of maritime search and rescue, and succeeded in identifying extraction patterns with an F-score of 73.75 %.

1 Introduction

Nous présentons notre approche d'apprentissage de patrons dans le contexte de l'extraction d'information à partir de textes conversationnels spécialisés. Cette étape est la dernière de notre approche proposée pour l'extraction d'information à partir de ces textes que nous avons présentée dans nos travaux précédents (Boufaden *et al.*, 2002; Boufaden *et al.*, 2005). Nous proposons une modélisation utilisant des modèles de Markov pour apprendre des relations prédicat-argument (séquences de classes sémantiques étiquetant le verbe et ses arguments) et les rôles[1] des arguments à partir de textes étiquetés sémantiquement. Ces textes sont des transcriptions[2] manuelles de conversations téléphoniques portant sur des incidents survenus en mer. Ce sont des compte rendus où les locuteurs se communiquent des informations sur un incident, par exemple un bateau en difficulté, sur les conditions météorologiques lors d'une mission de recherche ou sur le lieu de l'incident. Un exemple de conversation est donné au tableau 1.

Le système repose sur trois étapes et prend en entrée des séquences de classes sémantiques étiquetant les mots clé des énoncés où les étiquettes sont définies dans une ontologie du domaine. La première étape segmente les conversations en unités linguistiques à l'instar de la phrase pour les textes bien formés tels que les dépêches journalistiques (section 2.1). Cette étape prend en considération la dimension discursive très importante dans ce types de textes (Levelt, 1989). La deuxième effectue la résolution des anaphores pronominales en position de sujet (section 2.2). Cette étape tient compte d'une particularité des textes conversationnels : la pronominalisation du thème. Nous montrons que la résolution d'un sous ensemble des anaphores pronominales améliore l'apprentissage des patrons d'extraction. La troisième utilise les modèles de Markov pour modéliser les séquences de classes de mots et leurs rôles pour un ensemble de relations données (section 3). La comparaison de notre approche avec celles développées pour les textes bien formés montrent la pertinence de notre approche (section 4).

2 Problématique de l'apprentissage des patrons d'extraction

Un patron d'extraction est une structure qui permet le repérage des informations que nous voulons extraire et établit une relation entre ces éléments d'information. Il se caractérise par des contraintes syntaxiques (position des arguments dans une relation **sujet-verbe-objet**) et sémantiques (type de classes sémantiques) permettant le filtrage d'un sous-ensemble d'énoncés qui contiennent des informations pertinentes au domaine d'application. Parmi les principales difficultés de l'apprentissage des patrons d'extraction à partir de textes bien formés mentionnés dans la littérature (Grishman, 1998; Surdeanu *et al.*, 2003), nous retenons: (1) la diversité des constructions phrastiques contenant l'information pertinente et (2) l'association de nouveaux éléments d'information à des objets référencés par une anaphore.

Dans le contexte des textes conversationnels, ces difficultés sont amplifiées. D'une part, les irrégularités langagières telles que les répétitions et les reprises modifient la structure syntaxique des énoncés, tandis que l'aspect conversationnel a pour effet de répartir l'information sur plus d'un énoncé, par exemple lors d'échanges de type question-réponse. D'autre part, la présence importante de pronoms notamment à l'intérieur des unités thématiques augmente le nombre de

[1] Un rôle est un nom de champ défi ni dans un formulaire.

[2] Ces textes ont été fournis par le Centre de Recherche de la défense Canadienne. Ils ne sont pas annotés prosodiquement et nous n'avions pas les enregistrements originaux pour reconstituer la prosodie.

No	Loc	Énoncé
4	b:	ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat VESSEL on the South Coast of Newfoundland, just in LOCATION the area quite between Fortune Bay and Trepassey. LOCATION .....*Incident* .....
5	b:	it's on the south east coast of Newfoundland. LOCATION .....*Incident* .....
6	b:	this is been going on for, for 24 hours that the case has, TIME or almost anyway, and we had an DFO King Air up flying AIRCRAFT STATUS this morning. TIME .....*Search-unit* .....
7	b:	they did a radar search for us in that area. STATUS MEANSOFDETECTION LOCATION
8	a:	yes. .....*Search-unit* .....

Table 1: Exemple de conversation dans le domaine de Recherche et sauvetage. Les mots soulignés sont les informations que nous voulons extraire. Les étiquettes sous les barres en soulignés sont des classes de mots importants. Les pointillés sont les frontières des unités linguistiques que nous détectons dans la section (2.1). *Incident* et *Search-unit* sont des exemples de relations que nous voulons modéliser par des modèles de Markov.

relations partielles (par opposition à une relation complète où tous les arguments sont définis). L'approche que nous proposons tient compte de ses difficultés. Tout d'abord, nous effectuons une segmentation en paires d'adjacence[3] qui détecte, par exemple, les paires de type question-réponse pour repgrouper dans une seule unité linguistique les éléments d'information présents dans une question et sa réponse. Ensuite, nous procédons à la résolution des anaphores pronominales en position de sujet pour diminuer le nombre des relations partielles. Enfin, nous relaxons la contrainte de contiguité des arguments de la relation "sujet-verbe-objet", en apprenant les patrons à partir de séquences d'étiquettes sémantiques de longueur variable.

2.1 Segmentation en unités linguistiques

À l'instar des travaux en segmentation linguistique de conversations (Stolcke, 1997), nous avons utilisé un modèle de Markov d'ordre 1 pour modéliser des séquences de traits composés de marques lexicales telles que `ok`, `well` et `?` caractéristiques des paires d'adjacence, mais aussi la longueur d'un énoncé ainsi que l'identité de locuteur. Contrairement aux approches proposées, nous n'avons pas utilisé la prosodie car celle-ci est absente de nos textes. Le modèle contient deux états représentant la classe des énoncés indépendants (E) et la classe des énoncés complétant une paire d'adjacence (PA). Nous avons validé notre modèle en effectuant 10 validations croisées sur notre corpus contenant 64 conversations (3481 énoncés) avec 80 % réservé

[3]Les paires d'adjacence sont deux tours de parole, chacun venant d'un locuteur distinct où le premier tour nécessite un second tour de parole d'un certain type (source `http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/`).

à l'entraînement. La moyenne des erreurs de classification obtenue à partir des 10 validations croisées est de 15,9 %. L'analyse des erreurs de classification a montré que la source principale des erreurs est due à l'absence de marques lexicales pour certains énoncés de la classe PA. Dans ces cas, l'information prosodique absente dans nos transcriptions permettrait de combler le manque d'information lexicale.

2.2 Résolution des anaphores pronominales

Nous nous intéressons aux anaphores pronominales `they`, `we`, `she`, `he` et `it` en position de sujet[4]. Notre approche se base sur la structure thématique des conversations et sur une liste des étiquettes sémantiques[5] extraites à partir de chaque énoncé d'une unité thématique. L'importance de la structure thématique a déjà été soulignée pour la résolution des coréférences dans les conversations (Grosz *et al.*, 1995).

Le choix d'un antécédent est dirigé par deux contraintes de compatibilité: **sémantique** et **thématique**. La première fixe des associations possibles entre les étiquettes sémantiques et les pronoms. Tandis que la seconde fournie un antécédent par défaut, lorsqu'aucun antécédent compatible avec l'anaphore n'a été détecté dans les énoncés précédents de l'unité thématique courante ou de la précédente portant sur le même thème. Les valeurs par défaut sont les étiquettes les plus fréquentes calculées sur 31 conversations du corpus.

L'évaluation de notre approche a été effectuée sur 31 conversations de notre corpus, soit 161 anaphores pronominales en position de sujet. Le taux moyen d'erreurs de résolution obtenu est de 79,5 %. Bien que le résultat soit encourageant, certains choix de notre approche ont contribué à augmenter le taux d'erreurs, en particulier, le choix d'une approche linéaire (non hiérarchique) de segmentation en unités thématiques (Boufaden *et al.*, 2002) dans la segmentation automatique et la simplicité de notre approche dans le calcul des antécédents par défaut qui se base sur les fréquences obtenues sur le corpus.

3 Apprentissage des patrons d'extraction

Le but de cette étape est d'exploiter les associations entre les étiquettes sémantiques afin d'apprendre des patrons d'extraction qui expriment une relation prédicat-argument où les arguments ont un rôle spécifique pour une relation donnée. Des exemples d'étiquettes sémantiques utilisées sont présentées dans l'extrait de conversation du tableau 1.

3.1 Approche

Nous avons considéré cinq relations dans nos expériences:

1. *Missing-object* qui décrit Le bateau en difficulté, c'est-à-dire sa description, le nom de son propriétaire.
2. *Incident* qui décrit le type d'incident, la cause, le type d'appel de détresse.

[4]Levelt (Levelt, 1989), montre que les pronoms position de sujet sont souvent le résultat de la pronominalisation du thème d'une unité thématique.

[5]La structure thématique et les étiquettes sémantiques sont générées de manière automatique par des systèmes développés dans nos travaux précédents (Boufaden *et al.*, 2005).

Schémas d'extraction	Modèle de Markov	Rappel	Précision	F-score
Incident (62 énoncés)	Ordre 1	59,0 %	79,6 %	
	Ordre 2	63,8 %	85,0 %	72,9 %
Search-mission (27 énoncés)	Ordre 1	79,0 %	89,5 %	83,9 %
	Ordre 2	70,7 %	81,4 %	
Search-unit (93 énoncés)	Ordre 1	53,3 %	75,2 %	
	Ordre 2	52,9 %	76,9 %	62.7 %
Missing-object (38 énoncés)	Ordre 1	54,4 %	71,7 %	
	Ordre 2	70,8 %	80,8 %	75,5 %

Table 2: Rappel, précision et F-score de l'apprentissage des patrons d'extraction pour les formulaires *Incident*, *Mission*, *Search-unit* et *Missing-object*. Le rappel et la précision sont obtenus par la méthode de validation croisée "Leaving one out" pour les deux modèles de Markov. Le F-score est la moyenne des F-scores du meilleur modèle.

3. *Search-unit* qui parle de la ressource utilisée dans une mission de recherche.
4. *Mission* qui décrit le lieu de la mission, les conditions météorologiques, la date.

Pour chaque type de relation, nous avons modélisé les séquences des étiquettes avec un modèle de Markov. Nous avons entraîné chaque modèle sur un sous-ensemble du corpus qui contient des exemples positifs du type de relation ciblée.

3.2 Expériences et résultats

Nous avons effectué deux expériences afin de déterminer l'ordre du modèle de Markov qui donne les meilleures performances pour chaque patron d'extraction. Nous avons testé un modèle de Markov d'ordre 1 et un modèle d'ordre 2. Étant donné la taille modeste des corpus d'entraînement ($<$100) pour les différents patrons d'extraction, nous avons opté pour une validation croisée avec l'approche "Leaving one out". Les rappels[6], précisions et F-scores des meilleures performances sont indiquées au tableau 2.

Nous constatons que le patron d'extraction associé à la relation *Search-mission* présente une meilleure performance avec le modèle de Markov d'ordre 1, tandis que les autres patrons d'extraction *Missing-object*, *Incident* et *Search-unit* montrent de meilleurs résultats avec les modèles d'ordre 2.

Le choix de l'ordre du modèle dépend du taux des étiquettes sémantiques ayant plusieurs rôles possibles. Par exemple, dans l'unité thématique *Mission*, l'étiquette la plus fréquente est WEATHER-CONDITIONS avec une fréquence relative de 37,7 %. Cette dernière a un seul rôle dans la relation *Mission*, contrairement à l'étiquette NUMBER qui peut avoir le rôle d'une date ou d'une position géographique (en degré par exemple). Le choix de l'ordre dépend également du bruit introduit par les irrégularités langagières, notamment les reprises, agrandit la taille du contexte nécessaire pour désambiguïser un rôle.

[6]Le rappel correspond au nombre de rôles corrects générés par le système sur le nombre de rôles dans le corpus de test, tandis que la précision est le nombre de rôles corrects générés par le système sur le nombre de rôles qu'il fournit.

4 Conclusion

Nous avons analysé la problématique de l'apprentissage des patrons d'extraction pour des textes complexes peu étudiés en EI: les transcriptions de conversations. Nous avons modélisé les patrons d'extraction par des modèles de Markov qui associent des rôles aux arguments des prédicats avec un F-score de 73,75 %. Bien que les modèles de Markov aient été utilisés pour l'apprentissage de patrons (Seymore *et al.*, 1999), peu de travaux les ont utilisés pour apprendre les rôles sémantiques. De ces travaux, nous retenons ceux de Gildea (Gildea & Palmer, 2002) effectués sur des textes journalistiques avec un F-score de 82 %. D'autres approches ont été utilisées, notamment les arbres de décisions sur des textes bien formés avec un F-score de 83,7 % (Surdeanu *et al.*, 2003). Cependant, cette approche ne permet pas de tenir compte des séquences de longueurs variables que l'on retrouve avec les textes conversationnels.

Nous avons ajouté une étape de résolution des anaphores pronominales en amont de l'étape d'apprentissage de patrons. Notre approche a permis un taux de résolution des anaphores de 79,5 % améliorant ainsi le F-score moyen pour l'apprentissage de patrons de 68,6 %. Quelques travaux Surdeanu (Surdeanu & Harabagiu, 2002) ont utilisé une approche similaire pour améliorer l'extraction des informations en résolvant les coréférences aux entités nommées.

Références

BOUFADEN N., LAPALME G. & BENGIO Y. (2002). Découpage thématique des conversations: un outil d'aide à l'extraction. In *Actes de la 9 e conférence annuelle sur le traitement automatique des langues naturelles (TALN 2002)*, volume I, p. 377–382, Nancy, France.

BOUFADEN N., LAPALME G. & BENGIO Y. (2005). Repérage de mots informatifs à partir de textes conversationnels. *Traitement Automatique de la Langue*, **45**(3).

GILDEA D. & PALMER M. (2002). The necessity of syntactic parsing for predicate argument recognition. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL 2002)*, p. 239–246, Philadelphie, Pennsylvanie.

GRISHMAN R. (1998). Information extraction and speech recognition. In *Proceedings of the DARPA Broadcast Transcription and Understanding Workshop*, Lansdowne, Virginie: Morgan Kaufmann Publishers.

GROSZ B., JOSHI A. & WEINSTEIN S. (1995). Centering: A Framework for Modeling the local Coherence of Discourse. *Computational Linguistics*, **21**(2), 203–225.

LEVELT W. J. M. (1989). *Speaking: From Intention to Articulation*. ACL-MIT Press Series in Natural Language Processing. MIT Press.

SEYMORE K., MCCALLUM A. & ROSENFELD R. (1999). Learning hidden Markov structure for information extraction. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction*, p. 37–42, Orlando, Floride.

STOLCKE A. (1997). Modeling linguistic segment and turn boundaries for n-best rescoring of spontaneous speech. In *Proceedings of EUROSPEECH 1997*, volume 5, p. 2779–2782, Rhodes, Grèce.

SURDEANU M., HARABAGIU S., WILLIAMS J. & AARSETH P. (2003). Using predicate-argument structures for information extraction. In E. HINRICHS & D. ROTH, Eds., *Proceedings of ACL 2003*, p. 8–15.

SURDEANU M. & HARABAGIU S. M. (2002). Infrastructure for Open-Domain Information Extraction. In M. MITCHELL, Ed., *Proceedings of HLT 2002*, p. 325–330, San Diego, Californie.