TALN 2002, Nancy, 24–27 juin 2002

Découpage thématique des conversations: un outil d'aide à l'extraction

Narjès Boufaden Guy Lapalme Yoshua Bengio
{boufaden, lapalme, bengioy}@iro.umontreal.ca
Département d'Informatique et Recherche Opérationnelle
Université de Montréal, Quebec Canada

Mots-clefs – Keywords

Découpage thématique, analyse de conversations, extraction d'information
topic segmentation, conversation analysis, information extraction

Résumé

Dans cet article, nous décrivons la complexité du traitement automatique des conversations. En particulier, nous étudions la problématique de l'extraction d'information à partir des conversations et nous présentons le découpage thématique comme un outil d'aide à l'extraction.

1 Introduction

Le découpage thématique consiste à diviser un texte en passages cohérents. Chaque passage est un ensemble de phrases ou d'énoncés qui partagent le même thème. Le découpage thématique est très utile pour certaines applications du Traitement Automatique de la Langue (TAL), telles que la recherche d'information, le résumé automatique ou la résolution d'anaphores. Dans cet article, nous montrons qu'il est aussi utile pour l'extraction d'information à partir de conversations.

Notre travail s'inscrit dans le cadre d'un projet mené par le département de la défense canadienne et a pour but l'exploitation des comptes rendus téléphoniques de missions de recherche et sauvetage maritime. Nous avons défini une approche d'extraction composée de trois étapes. La première, le découpage thématique sépare la conversation en segments cohérents possédant des propriétés qui facilitent le processus d'extraction. La deuxième étape, le processus d'extraction, prend en entrée un segment thématique, en extrait les informations pertinentes et les attribue aux champs des formulaires prédéfinis. La troisième étape permet de résoudre la surgénération des réponses pour un champ ; conséquence directe des répétitions et des négociations entre interlocuteurs.

Dans cet article, nous nous consacrons à l'étude de la première étape de notre approche. Nous étudions le découpage thématique comme un outil d'aide à l'extraction d'information à partir de conversations. Nous décrivons la problématique du découpage thématique et présentons l'approche utilisée.

Le corpus que nous avons étudié est composé de conversations téléphoniques transcrites manuellement (figure 1). Ces conversations sont des dialogues informatifs dans lesquels un interlocuteur rapporte l'état d'avancement d'une mission de recherche et sauvetage. Notre étude a été effectuée sur 95 conversations totalisant environ 39,000 mots.

2 Traitement automatique du langage et complexité des conversations

En dépit de la maturité des recherches en TAL pour les textes écrits, celles dédiées aux traitement des conversations sont encore au stade embryonnaire. Un des rares systèmes élaboré en traitement automatique des conversations est le système MIMI pour la génération automatique de résumés à partir de conversations dont le thème est la réservation de chambres d'hôtel (Kameyama et al., 94). Ce nombre restreint de travaux est en partie dû à la complexité de la structure des conversations comparativement à celle des textes écrits. Contrairement au texte écrit, une conversation est un ensemble **d'énoncés** résultant **d'interactions spontanées entre plusieurs locuteurs**. La définition du concept **énoncé** et l'étude du **processus interactionnel** entre les locuteurs sont les difficultés majeures le traitement automatique des conversations.

La première problématique reliée au traitement des conversations est l'absence d'un consensus sur la définition de ce que représente un énoncé. Pour les textes écrits, la phrase est l'unité linguistique utilisée pour le traitement du langage naturel. Elle est généralement bien définie au niveau de sa composition syntaxique et est délimitée par une ponctuation explicite. Pour les conversations, il est difficile d'identifier une unité linguistique analogue à la phrase. La définition d'une grammaire pour les conversations est complexe sinon impossible à cause de

1	C	*Maritime operation centre, (INAUDIBLE) hello.*	S_1
2	O	*Hi, Mr. Green, it's captain Mr. Red*	
3	C	*Yes.*	
		. . . . .	. .
4	O	*Ha, I don'nt know if I was handled over to you at all, but we've got an overdue boat on the south coast of Town2, just in the area quite between Town1 and Town3.*	S_2
5	O	*It's on the south east coast of Town2.*	
		. . . . .	. .
6	O	*This is been going on for, for 24 hours that the case has, or almost anyway, and we had an Aircraft1 up flying this morning*	S_3
		. . . . .	. .
7	O	*They did a radar search for us in that area.*	S_4
8	C	*Yes.*	
		. . . . .	. .
9	O	*And their search turned up nothing.*	S_5
10	C	*yeah.*	
		. . . . .	. .
	. . .		
18	C	*We're wondering about aircraft for tomorrow.*	S_6
19	C	*Is Aircraft2 gonna be available ?*	
20	O	*You're on another radar search or loose end ?*	
21	C	*Yeah.*	
22	O	*You feel that's gonna be of some use ?*	S_7
23	C	*Well it won't hurt.*	
		. . . . .	. .
	. . .		
31	C	*Thanks.*	S_8
32	O	*All right.*	
33	O	*Bye*	

FIG. 1 – Extrait d'un compte rendu entre deux locuteurs : `Caller` (C) et `Operator` (O). Pour des raisons de confidentialité certaines entités nommées ont été remplacées par des noms génériques.

la présence d'extra-grammaticalités (Lee et al., 95; Boufaden et al., 98). Plusieurs linguistes et psycholinguistes se sont attardés sur la définition d'un énoncé (Traum et al, 1997). Un énoncé peut être un tour de parole (Sacks et al., 74), un segment délimité par des informations de nature prosodique telles que l'intonation ou la durée d'une pause (Takagi et al., 96) ou une unité linguistique (avec des propriétés syntaxiques et/ou sémantiques) (Ford et al., 91). Cette problématique a été partiellement résolue par l'adoption de l'unité linguistique pour la transcription des textes oraux. Le découpage linguistique garantit des propriétés syntaxiques qui facilite les TAL.

La deuxième problématique est la nature fondamentalement contextuelle des énoncés. Un locuteur interagit en s'assurant que ces propos restent cohérents avec ceux des autres participants. Il s'en suit que l'interprétation d'un énoncé est fortement influencée par le contexte d'énonciation. Les paires question-réponse présentes dans les conversations sont des exemples triviaux qui illustrent la dépendance contextuelle des énoncés. Cette particularité met l'emphase

sur la nécessité de tenir compte de la dimension discursive dans l'élaboration de TAL pour les conversations.

Plusieurs applications dédiées aux textes écrits ont utilisé le segment thématique comme unité de traitement. En particulier, le découpage thématique est une étape fondamentale en recherche d'information. Nous parlons alors de segmentation de textes où il s'agit de délimiter des segments thématiques cohérents dans un ensemble de données non partitionné (Beeferman et al., 99).

Pour des applications telles que les systèmes de résumé automatique (Salton et al., 93; Boguraev et al., 2000), il s'agit plutôt d'un découpage en sous-thèmes. Un exemple de découpage en sous-thèmes est celui donné par Hearst. Il cite pour exemple un article dont le thème principal est la description de la vie sur la terre et sur les autres planètes. Il décrit son contenu comme une succession des sous-thèmes suivants (les numéros indiquent les paragraphes) (Hearst, 94) :

1-3	Introduction :The search of life in space
4-5	The moon's chemical composition
6-8	How early earth-moon proximity shaped the moon
9-12	How the moon helped life evolve on earth
13	Improbability of earth-moon system
14-16	Binary/trinary star systems make life unlikely
17-18	The low probability of nonbinary/trinary systems
19-20	Properties of earth's sun that facilitate life
21	Summary

Dans le cadre de notre application, nous nous intéressons au découpage en sous-thèmes comme un outil d'aide à l'extraction d'information. Toutefois nous gardons les termes "découpage thématique" pour décrire nos travaux.

3 Découpage thématique pour l'extraction d'information

Une étape cruciale pour l'extraction d'information est la localisation des énoncés contenant de l'information pertinente. Cette étape relativement maîtrisée pour les textes écrits est grandement facilitée par les propriétés syntaxiques et sémantiques de la phrase.

Requête : *What is the location of the overdue boat ?*
Réponse : `the south east coast of Town2,`
`just in the area quite between Town1 and Town3`

Passage

4 o *Ha, I don't know if I was handled over to you at all, but we've got an overdue boat on the south coast of Town2, just in the area quite between Town1 and Town3.*
5 o *It's on the south east coast of Town2.*

FIG. 2 – Recherche d'information dans le but de l'extraction d'information

Pour les conversations, ces propriétés ne sont pas toujours vérifiées. D'une part, la structure syntaxique d'un énoncé est souvent altérée par les extra-grammaticalités. D'autre part, la spon-

tanéité des échanges et les négociations ont un impact direct sur la structure de la conversation qui parfois paraît décousue. L'ensemble de ces caractéristiques font que l'information recherchée peut apparaître plus d'une fois d'un énoncé à un autre et elle peut être modifiée au fur et à mesure de la conversation. Il en résulte une sur-génération de réponses potentielles pour un champ donné. Nous utilisons le découpage thématique comme un pré-traitement pour l'extraction d'information dans le but de maximiser la probabilité d'extraire la bonne information parmi toutes celles résultant de répétitions ou tout simplement d'une négociation entre les interlocuteurs. Le résultat espéré de cette étape est une conversation séparée en une suite de passages cohérents dont certains contiennent de l'information destinées à remplir un champ de formulaire.

En partant de l'hypothèse qu'une information peut être explicitée plusieurs fois de manières différentes à cause des extra-grammaticalités, par exemple, nous pouvons formuler notre problème d'extraction d'information comme suit :

> *chercher le passage de la conversation qui permet de répondre à une requête portant sur le contenu d'un champ de formulaire donné, puis extraire les informations pertinentes à partir de ce passage.*

Le découpage thématique préconisé pour la tâche d'extraction génère des passages qui en plus d'être cohérents respectent deux contraintes :

1. Être assez petits pour ne répondre qu'à un nombre restreint de requêtes (champs de formulaire), généralement on espère répondre à un champ à la fois.
2. Être assez grands pour englober tous les énoncés adjacents qui contiennent de l'information répondant à la requête.

Considérons la requête *What is the location of the overdue boat ?* (figure 2). Le premier énoncé qui répond à la requête est l'énoncé `4`. La première contrainte identifiée pour le découpage thématique implique que cet énoncé appartient au passage recherché puisque la partie soulignée de l'énoncé `4` est une réponse partielle à la requête (figure 2). Par contre, celle-ci ne contient pas toute l'information correcte. En effet, l'énoncé `5` (`It's on the south east coast of Town2 `) est une reprise partielle de l'énoncé `4` qui corrige en partie l'information donnée dans l'énoncé `4` (`...on the south coast of Town2`). L'application de la deuxième contrainte permet d'inclure l'énoncé `5` dans le passage. Si l'énoncé `5` n'avait pas été inclus dans le passage, la réponse à la requête aurait été erronée.

En résumé, le découpage thématique ainsi défini fournit au système d'extraction d'information en une seule unité toutes les réponses potentielles à un champ. Une approche d'extraction d'information adaptée à la structure des conversations doit être centrée sur le segment thématique et non sur l'énoncé. Le découpage thématique devient une étape préliminaire de l'extraction d'information, comme nous le montrons dans la figure 3.

4 Notre Approche

Le but de notre étude est d'automatiser le découpage thématique afin d'intégrer ce processus comme étape préliminaire à l'extraction d'information. À cette fin nous avons identifié les informations discriminantes pour la localisation des frontières et avons utilisé une approche probabiliste basée sur un modèle de Markov caché d'ordre 1. La détermination des marques discriminantes pour le découpage repose sur deux aspects :

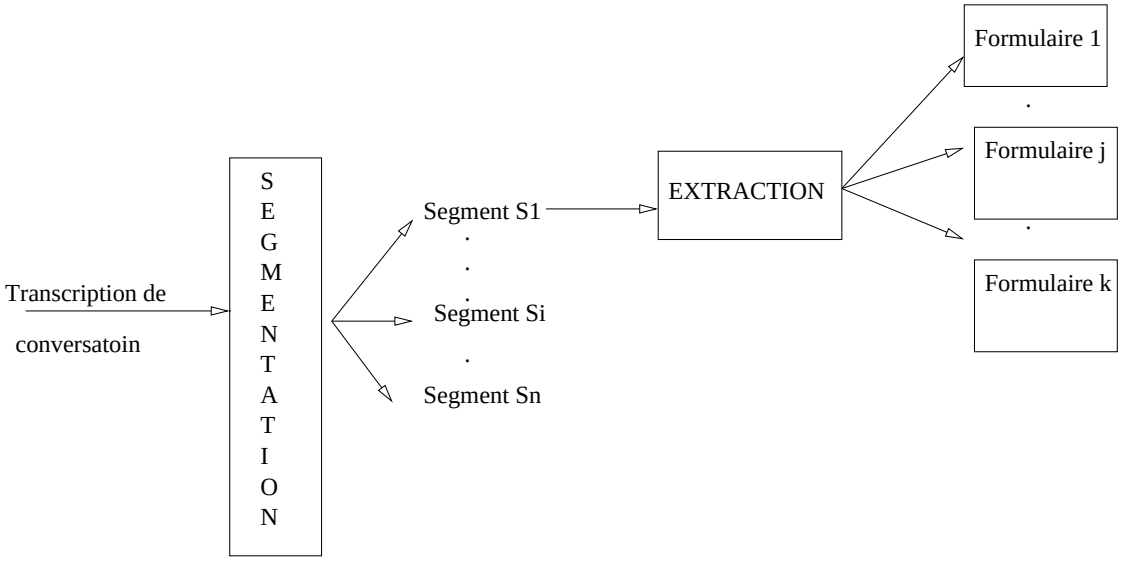

FIG. 3 – Une stratégie d'extraction d'information utilisant le segment thématique comme unité de traitement

– La cohésion entre les énoncés d'un même thème et les changements de thèmes : *quelles sont les marques lexicales qui indiquent une dépendance entre deux énoncés adjacents* ? Plusieurs travaux en analyse de discours (Halliday et al., 1976; Brown et al., 83) et en découpage thématique ont souligné l'importance des marques lexicales et syntaxiques comme information indiquant les changements de thèmes ou la cohésion entre les énoncés.
– L'aspect collaboratif des conversations. Plusieurs travaux ont décrit les règles qui régissent les changements de thème dans les dialogues (Sacks et al., 74; Maynard, 1980). Ils ont souligné l'importance du rôle du locuteur pendant le processus développemental. Ils se basent sur l'hypothèse que les changements de thème se réalisent selon un ensemble de schémas bien définis. Dans une conversation entre deux locuteurs, chaque locuteur montre son intérêt et sa compréhension de ce qui est communiqué grâce à des réponses typiques tels que `ok, yeah, right`. En fonction du rôle du locuteur dans le processus développemental du thème, ces réponses peuvent être perçues comme un incitateur à continuer le thème ou au contraire comme un inhibiteur dans le but d'interrompre le thème. En particulier (Maynard, 1980) parle de locuteur initiateur du thème **topical speaker**, comme étant le locuteur qui verbalise son intention communicative, par opposition au destinataire **recipient** qui va inciter à développer le thème ou au contraire changer de thème. Dans (Boufaden et al., 2001) nous avons montré que cette information améliore les résultats du découpage.

Ainsi, les marques retenues pour le découpage en thèmes sont les suivantes :

discursives marquant le rôle de chaque locuteur lors du processus développemental du thème. Un locuteur peut-être l'initiateur d'un thème (S) ou le destinataire (R).

syntaxiques comme les adverbes temporels (`therefore`, `then`, `before`), des conjonctions et disjonctions (`or`, `and`, `but`) qui apparaissent en début d'énoncé ainsi que le point d'interrogation sont aussi pris en compte.

lexicales par exemple (`ok`, `right`, `I mean`, `thank you`, `hum`) qui apparaissent en début d'énoncé, ou qui constituent l'énoncé et l'occurrence de même mots entre deux énoncés adjacents.

interruptions qui sont des coupures dans les énoncés dues au début d'un nouveau tour de parole sans qu'il y ait eu complétion du tour de parole précédent. Dans le corpus, ce

phénomène a été transcrit par des points de suspension (. . .). Cette marque est souvent liée à un changement de thème en particulier lorsqu'elle est provoquée par le destinataire.

Toutes ces marques sont extraites automatiquement à partir des transcriptions de conversation. En particulier, l'extraction du trait discursif a nécessité l'élaboration d'un modèle de langage qui permet la prédiction du rôle du locuteur à partir de marques lexicales et syntaxiques.

Les détails techniques de notre approche pour la prédiction du trait discursif ainsi que pour le découpage thématique ont été décrits et commentés dans (Boufaden et al., 2001).

L'approche basée sur les traits linguistiques et les interruptions nous a permit d'obtenir une moyenne pondérée de 81,3% pour la précision et de 81,9% de rappel. Les résultats représentent les moyennes des scores obtenus pour 10 validations croisées.

5 Résumé et travaux futurs

Dans cet article, nous avons étudié les problèmes liés à l'extraction d'information à partir des conversations et présenté le découpage thématique comme un outil facilitant la tâche d'extraction. Une des difficultés du découpage thématique est la définition du thème. Brown et Yule (Brown et al., 83) ont qualifié la notion de thème comme étant intuitive et difficile à expliciter. Pour contourner ce problème, ils ont proposé d'étudier ce qui permet d'identifier les changements de thèmes.

Dans notre expérimentation, nous étions confrontés au même problème. Cependant les contraintes que nous avions imposées pour le découpage ont grandement facilité le processus. Les frontières des segments thématiques sont disposées de manière à rassembler les énoncés adjacents qui contribuent à répondre à un nombre minimum de champs, en général un champ, tout en s'assurant d'avoir tous les énoncés qui contiennent des réponses potentielles à ce champ. Par exemple, les segments S_4 et S_5 (figure 1) pourraient être considérés comme un seul segment thématique où l'on indique une recherche par radar et le résultat de cette recherche. Toutefois la présence d'un champ spécifique pour le résultat de la recherche et d'un autre pour le type de recherche dans le formulaire prédéfinis pour la recherche fait que le système va générer deux segments.

Notre approche est une combinaison de techniques utilisées dans différentes approches de segmentation de textes. Nous utilisons certaines marques lexicales et syntaxiques qui permettent de détecter les changements de thème. Nous avons aussi utilisé la présence de répétitions des mots entre les énoncés qui est un facteur important dans la détection de la cohésion et ainsi que le trait discursif pour détecter les changements de thèmes. Le trait discursif a été choisi car il permet une interprétation plus rigoureuse de certaines des marques lexicales. Une évaluation croisée de notre système a donné une moyenne pondérée de 81,3% pour la précision et de 81,9% de rappel.

À l'instar des nombreux travaux effectués en découpage thématique, très peu sont consacrés aux conversations. De ce fait il est difficile de comparer notre travail à d'autres recherches du même domaine.Toutefois, ceux-ci sont assez concluants pour permettre le passage à l'étape d'extraction d'information. La deuxième étape de notre projet consiste à extraire les informations à partir des segments thématiques. Notre but est de définir une approche d'extraction qui est centrée sur l'utilisation du segment thématique comme unité d'extraction, en plus d'être robuste pour extraire l'information en dépit des altérations de la structure syntaxique des énoncés.

Références

J. Lafferty D. Beeferman, A. Berger. Statistical models for text segmentation. *Machine Learning*, 34(1-3), Fevrier 1999.

Branimir Boguraev and Neff Mary. Discourse segmentation in aid of documment summarization. In *Proceedings of the 33th Hawaii International Conference on Systems Sciences*, Maui, HI, 2000.

Narjès Boufaden, Sylvain Delisle, and Bernard Moulin. Analyse syntaxique robuste de dialogues retranscrits : peut-on vraiment traiter l'oral à partir de l'écrit ? In *Actes de traitement automatique des langues naturelles*, Paris, France, Juin 1998.

N. Boufaden, G. Lapalme, and Y. Bengio. Topic segmentation : A first stage to dialog-based information extraction. In *Natural Language Processing Rim Symposium, NLPRS'01*, pages 273–280, 2001.

Guilian Brown and Yule Beorge. *Discourse Analysis*. Cambridge Textbooks in Linguistics Series. Canbridge University Press, 1983.

C. Ford and S. Thompson. On projectability in conversation : Grammar, intonation, and semantics. In *The second Internatioanl Cognitive Linguistics Association Conference*, August 1991.

M.A.K Halliday and R. Hassan. *Cohesion in English*. Longman, London, 1976.

Marti Hearst. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9–16, New Mexico State University, Las Cruces, New Mexico, 1994.

M. Kameyama and I. Arima. Coping with aboutness complexity in information extraction from spoken dialogues. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-94*, Yokohama, Japan, September 1994.

K-J. Lee, C-J. Kweon, J. Seo, and G-C. Kim. A robust parser based on syntactic information. In *Proceedings of EACL*, pages 223–228, 1995.

D.W. Maynard. Placement of topic changes in conversation. In *Semiotica*, volume 30, pages 263–290. Mouton Publishers, 1980.

H. Sacks, E.A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking in conversation. *Language*, 50 :696–735, 1974.

Gerard Salton, Allan James, and Buckley Chris. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, pages 49–58, Pittsburgh, PA, 93.

K. Takagi and S. Itahashi. Segmentation of spoken dialogue by interjection, disfluent utterances and pauses. In *Proceedings of the 4rd International Conference on Spoken Language Processing*, pages 693–697, Philadelphia, October 1996.

Traum, D. et Heeman, P. (1997). Utterance units in spoken dialogue. Dans Maier, E., Mast, M. et LuperFoy, S., éditeurs, *Dialogue Processing in Spoken Language Systems*, Lecture Notes in Artificial Intelligence. Springer-Verlag.

G. Tür, D. Hakkani-Tur, A. Stolcke, and E. Shriberg. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational linguistics*, 1(27) :31–57, 2001.