Segmentation en thèmes de conversations téléphoniques :traitement en amont pour l'extraction d'information

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

Mots-clefs – Keywords

segmentation en thèmes, analyse des conversations, extraction d'information
topic segmentation, conversation analysis, information extraction

Résumé - Abstract

Nous présentons une approche de découpage thématique que nous utiliserons pour faciliter l'extraction d'information à partir de conversations téléphoniques transcrites. Nous expérimentons avec un modèle de Markov caché utilisant des informations de différents niveaux linguistiques, des marques d'extra-grammaticalités et les entités nommées comme source additionnelle d'information. Nous comparons le modèle obtenu avec notre modèle de base utilisant uniquement les marques linguistiques et les extra-grammaticalités. Les résultats montrent l'efficacité de l'approche utilisant les entités nommées.

We study the problem of topic segmentation as a means to facilitate information extraction from manually transcribed convesrations. We experiment with a first order HMM using a combination of linguistic-level cues and named entities. We compare the results of our linguistic-levels cues based model with the named entities based model. Results show the effectiveness of named entities as an additional source of information for topic segmentation.

1 Introduction

Une étape cruciale de l'extraction d'information est la localisation des énoncés contenant de l'information pertinente. Cette étape maîtrisée pour les textes écrits structurés ne l'est pas encore pour les textes oraux. Les conversations (Figure 1) présentent plusieurs particularités compliquant l'extraction d'information notamment l'aspect collaboratif des conversations et la présence d'extra-grammaticalités. Ces deux caractéristiques font que (1) les éléments d'une réponse ne se trouvent pas nécessairement dans le même énoncé et (2) il faut pouvoir reconstituer la réponse correcte à partir d'un segment dont la structure grammaticale est altérée par les extra-grammaticalités. Dans nos travaux antérieurs, nous soutenions que le découpage thématique peut faciliter l'extraction d'information à partir des conversations (Boufaden et al., 2001; Boufaden et al., 2002). Nous avons élaboré un système de découpage thématique qui détecte les changements de thèmes à partir de marques lexicales, syntaxiques, discursives et des interruptions. Dans notre premier système la marque discursive était ajouté manuellement ce qui ne permettait pas un découpage complètement automatisé. Dans cet article, nous présentons, tout d'abord, les résultats de l'automatisation du calcul de la marque discursive. Ensuite, nous proposons l'utilisation des entités nommées comme source d'information additionnelle pour améliorer le découpage thématique.

1	C	*Maritime operation centre, (INAUDIBLE) hello.*
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.*
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 Airplane1 up flying this morning*
7	O	*They did a radar search for us in that area.*
8	C	*Yes.*
		. . . . .
9	O	*And their search turned up nothing.*
10	C	*yeah.*
		. . . . .
11	C	*Thanks.*
12	O	*All right.*
13	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. Les lignes pointillées sont les frontières des segments thématiques.

2 Éxpériences et résultats

Notre approche pour le découpage thématique repose sur l'utilisation d'informations linguistiques (Halliday et al., 1976; Maynard, 1980) et extra-linguistiques pour détecter les change-

ments de thèmes. Les informations linguistiques sont essentiellement :
– des mots tels que *ok, right, well* que nous appelons marques lexicales,
– des adverbes temporaux, conjonctions qui sont des marques syntaxiques,
– le rôle du locuteur dans le développement du thème que nous appelons marque discursive. 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 thène. Nous avons montré que cette information améliore les résultats de la segmentation (Boufaden et al., 2001).

La marque extra-linguistique que nous utilisons est l'interruption transcrite dans les conversations par des points de suspension. Les statistiques faites sur notre corpus ont montré une corrélation entre les interruptions et les changements de thèmes.

2.1 Modèle de langue

Dans (Boufaden et al., 2001), nous avons montré que le problème de détection d'un changement de thème peut être transposé en un problème de classification des énoncés. Nous émettions l'hypothèse qu'entre chaque vecteur de marques se situe une frontière qui permet de délimiter deux classes d'énoncé. Nous avons construit un modèle de Markov caché d'ordre 1 composé de cinq états (Figure 2) où chacun des états représente une classe d'énoncé :
– Les énoncés qui indiquent un début de conversation. Généralement ils contiennent des salutations ainsi que l'identification des locuteurs. Ces énoncés sont représentés par la classe `BC` (Begin Conversation)
– Les énoncés qui clôturent une conversation sont représentés par la classe `EC` (End Conversation). Ces énoncés contiennent souvent des expressions typées tels que *talk to you later, bye, have a good day*.
– Les énoncés qui débutent un nouveau thème forment la classe `TC` (Topic Change).
– Les énoncés qui font partie du corps d'un thème sont représentés par la classe `NO-TC` (No Topic Change).
– Les énoncés qui clôturent un thème sont représentés par la classe `ET` (End of Topic). Ces énoncés sont souvent composés d'unités lexicales tels que *ok, right, well*.

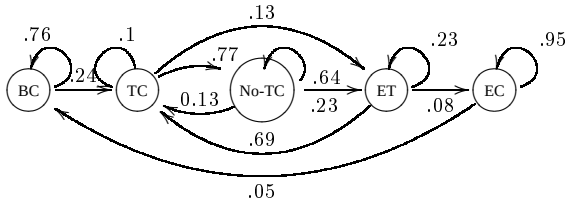

FIG. 2 – HMM d'ordre 1 pour la segmentation en topique

La Figure 2 illustre notre modèle de langue. Les valeurs représentées au dessus des arcs sont les probabilités $P(q_i|q_j)$ de générer l'état q_i sachant que l'on est dans l'état q_j.

Nous avons montré que, parmi les combinaisons de marque possibles (lexicale-discursive, lexicale-discursive-syntaxique, lexicale-discursive-syntaxique-interruption) pour la segmentation, la meilleure performance était obtenue avec la combinaison de toutes les marques c'est-à-dire lexicale, syntaxique, discursive et interruption (modèle SLDI) utilisée avec un modèle de Markov caché d'ordre 1. Les résultats rapportés dans (Boufaden et al., 2001) étaient basés sur un corpus d'entraînement où la marque discursive était ajoutée manuellement. Afin de rendre le découpage complètement automatisé, nous avons implémenté un modèle de Markov d'ordre 1 qui permet la génération automatique du trait discursif. Ensuite, dans le but d'améliorer les résultats du système de découpage thématique, nous avons extrait automatiquement les entités nommées pour les utiliser comme une source d'information additionnelle. Dans ce qui suit, nous décrivons les résultats de ces deux expériences.Tous les résultats présentés ici sont obtenus par validation croisée et avec des proportions de 85% pour l'apprentissage et 15% pour le test. Le corpus de base pour la segmentation est composé de 65 conversations, environ 3,700 énoncés.

2.2 Calcul de la marque discursive

Pour prédire les traits discursifs, la première idée était de considérer le locuteur (Operator `O` et Caller `C`) comme un trait discriminant en plus des marques syntaxiques et lexicales utilisées pour la segmentation. Afin de ne retenir que les traits les plus intéressants pour le modèle, nous avons testé différentes combinaisons entre les traits locuteur, syntaxique et lexical. Les combinaisons retenues sont celles où le locuteur est utilisé conjointement avec les traits lexical et syntaxique et une autre où le trait locuteur n'est pas considéré. Nous avons entraîné deux modèles de Markov caché sur 82 conversations qui ont été manuellement annotées avec le trait discursif et ils ont été testés sur 13 conversations. Les tableaux 1 et 2 représentent respectivement le taux d'erreur de classification, la précision et le rappel pour les classes destinataire (`R`) et initiateur de thème (`S`).

Trait discursif	R	S	Moyenne pondérée
(+) locuteur	24.1%	19.9%	21.7%
(-) locuteur	21.3%	20.6%	20.9%

TAB. 1 – Taux d'erreur de prédiction du trait discursif pour les modèles de Markov d'ordre 1 avec locuteur et sans locuteur

Trait discursif	Rappel	Précision
R	74.8%	78.5%
S	79.3%	79.7%
Moy.pondérée	77.3%	78.9%

TAB. 2 – Rappel et Précision par trait discursif pour le modèle sans locuteur

Il est intéressant d'observer que seules les marques lexicales et syntaxiques suffisent à déterminer le rôle du locuteur dans le processus développemental. L'utilisation de la marque discursive générée automatiquement a diminué légèrement les performances du système de découpage thématique. Le taux d'erreur moyen pondéré de découpage était de 16.5% avec la marque discursive ajoutée manuellement, tandis qu'avec celle calculée automatiquement il est de 18.5%.

2.3 Entités nommées source additionnelle d'information

Le but de cette expérimentation est d'améliorer les performances du système présenté dans (Boufaden et al., 2001). Un des problèmes soulignés dans (Boufaden et al., 2001) était le

manque de marques dans certains énoncés tels que ceux qui commence un nouveau thème ainsi que ceux débutant une conversation. Nous avons remarqué que 30.7% des énoncés classés BC et 46.8% des énoncés que nous avons classés TC dans le corpus d'entraînement contiennent uniquement la marque discursive. Par contre, nous avions aussi constaté la présence d'entités nommées dans ces mêmes énoncés. Lors d'un début de conversation les locuteurs se présentent et identifient l'organisme auquel ils appartiennent, ce qui implique la présence d'entités nommées de type ORGANISME et PERSONNE. À chaque changement de thème de nouveaux objets sont introduits et à cause de la nature informative de nos conversations ces objets, correspondent souvent à des entités nommées tels que les types d'avion, de bateaux, d'organisme ou les lieux. Suite à ces observations, nous avons procédé à l'extraction automatique des entités nommées PERSONNE, ORGANISME, AVION, BATEAU et LIEUX pour les intégrer à l'ensemble des marques utilisées pour le découpage thématique. D'emblée, cette procédure a permis de diminuer les pourcentages d'énoncés annotés uniquement avec le trait discursif à 16.5% pour la classe BC et 35.9% pour la classe TC (par rapport à 30.7% et 46.8%). Ensuite, nous avons utilisé les types d'entités nommées avec les anciennes marques pour réentraîner notre modèle de Markov. Les résultats de cette expérience sont représentés dans les tableaux 3 et 4. La colonne "(+) entités nommées" fait référence au système qui utilise les entités nommées comme source additionnelle d'information. La colonne "(-) entités nommées" représente le système de base qui utilise les marques linguistiques et extra-linguistiques. Pour les deux modèles les marques sont extraites de manière automatique.

Classe d'énoncé	(+) entités nommées	(-) entités nommées
BC	24.0%	39.4%
EC	12.1%	12.1%
TC	38.6%	39.2%
No-TC	9.9%	9.4%
Moy. pondérée	17.0%	18.1%

TAB. 3 – Taux d'erreurs par classe d'énoncé avec le modèle de Markov d'ordre 1 entraîné sur toutes les marques plus (+) les entités nommées, et sans (-) les entité nommées

Classe d'énoncé	(+) entités nommées		(-) entités nommées	
	Préc.	Rapp.	Préc.	Rapp.
BC	83.0%	76.0%	78.9%	60.6%
EC	82.6%	87.9%	80.4%	87.9%
TC	67.3%	61.4%	67.5%	86.0%
No-TC	87.0%	90.1%	85.8%	90.6%
Moy. pondérée	82.6%	83.0%	81.3%	81.9%

TAB. 4 – Rappel et Précision par classe d'énoncé

C'est au niveau de la classe BC que l'on observe la plus grande amélioration. Dans le système qui utilise les entités nommées le taux d'erreurs a diminué pour passer de 39.4% à 24%. Aussi, le rappel a significativement augmenté pour passer de 60% à 76%, ce qui indique que le système détecte plus d'énoncés de la classe BC, mais aussi se trompe moins dans sa classification puisque la précision a aussi augmenté pour passer de 78.9% a 83%.

Toutefois, nous ne pouvons attester des mêmes améliorations pour les autres classes. En particulier pour la classe TC, nous avons diminué le taux d'erreur de 1.5% ce qui est un maigre

résultat comparativement à celui de 39.9% pour la classe BC. Nous pensons que le manque de raffinement du module d'extraction d'entités nommées en est la cause principale.

3 Conclusion et travaux futurs

La majorité des méthodes de segmentation en thèmes, que ce soit dans le cadre d'applications telles que la recherche d'information ou dans des applications dédiées à la segmentation utilisées dans les conférence TDT (Topic Detection and Tracking)(Allan et al., 1998), utilisent des unités lexicales et la prosodie modélisés par des approches statistiques tels que les modèles de Markov et/ou des arbres de décisions (Litman et al., 1995; Laferty et al., 1999). Dans notre approche, nous avons utilisé des unités lexicales dans le processus de segmentation, toutefois, nous avons ajouté deux autres sources d'information : le trait discursif pour modéliser l'aspect collaboratif des conversations et les catégories d'entités nommées pour enrichir notre modèle. Les résultats montrent que les entités nommées accroissent les performances du système globalement puisque le score pondéré pour le rappel est passé de 81.3% à 82.6% et de 81.9% à 83% pour la précision. La plus grande amélioration a été enregistrée pour la classe BC avec 39.9% de diminution du taux d'erreur. Toutefois, plusieurs améliorations doivent être apportées au module d'extraction des entités nommées pour améliorer les résultats de la classe TC. Enfin, à notre connaissance peu ou pas de travaux en segmentation de dialogues ont été publiés, de ce fait il est difficile d'évaluer nos résultats comparativement à d'autres travaux. 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 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. Allan, J. Carbonnel, G. Doddington, J. Yamron, and Y. Yang. Topic Detection and Tracking pilot study final report. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 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.

N. Boufaden, G. Lapalme, and Y. Bengio. Découpage thématique : un outil d'aide à l'extraction d'information. In *TALN 2002*, Nancy, France, Juin 2002.

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

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

W.J.M. Levelt. *Speaking From Intention to Articulation*. MIT Press, 1989.

D.J. Litman and R.J. Passonneau. Combining multiple knowledge sources for discourse segmentation. In *Proc. of ACL'95*, pages 108–115, 95.

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 *Language*, volume 50, pages 696–735. 1974.