Multi-modal discursive analysis of online written conversations

Bibliography

Soufian Salim

October 9, 2014

Chapter 1

Discourse structure

La théorie des actes du langage [1] propose de décrire les énonciations en termes des fonctions communicatives portées par chacun d'eux (e.g. question, réponse, remerciement...). Ainsi, dans la plupart des travaux, c'est en termes d'actes du langage[1] que les interactions entre participants d'une conversation sont modélisées. Austin considère les énonciations comme des actions effectuées par le locuteur ; on trouve ici l'idée selon laquelle tout acte d'énonciation serait la réalisation d'un acte social. Les verbes qui spécifient ces actions sont appelés *verbes performatifs*, comme quand on dit "Je vous confère le titre de capitaine". Mais les actes du langage ne sont pas constitués uniquement de ces types de verbes. [2] propose cinq classes d'actes du dialogue : les assertifs (assertion, affirmation, etc.), les directifs (ordre, demande, conseil, etc.), les promissifs (promesse, offre, invitation, etc.), les expressifs (félicitation, remerciement, etc.) et les déclaratifs (déclaration de guerre, nomination, baptême, etc.).

Les travaux existants concernant les actes du langage s'intéressent dans leur large majorité à classifier les énoncés selon telle ou telle taxonomie, dont il existe un très grand nombre [3]. Le tableau 1.1 détaille deux taxonomies fondatrices : celle de Austin et celle de Searle. Le tableau 1.2 présente quelques taxonomies récemment employées dans le cadre de l'analyse de courriels. L'usage d'algorithmes de classification supervisée [4] représente l'approche dominante pour déterminer l'acte porté par une phrase ou un message.

[1] Aussi connu dans certains travaux sous le nom d'actes du dialogue (*dialog act*), ou d'actes du discours (*speech act*).

Acte	Description ou exemples	Référence
Verdictif	acquitter, condamner, décréter...	
Exercitif	dégrader, commander, ordonner, pardonner, léguer...	
Promissif	promettre, faire vœu de, garantir, parier, jurer de...	[1]
Comportatif	s'excuser, remercier, déplorer, critiquer...	
Expositif	affirmer, nier, postuler, remarquer...	
Assertif	affirmation d'un état de fait	
Directif	tentative de pousser un interlocuteur à faire quelque chose	
Promissif	engagement de la part du locuteur	[2]
Expressif	expression d'un état psychologique	
Déclaratif	déclaration ayant un impact direct	

Table 1.1: Taxonomies fondatrices pour la catégorisation des actes du langage.

Acte	Corpus ou type de corpus	Référence
Divulgation Édification Conseil Confirmation Question Reconnaissance Interprétation Réflexion	multi-domaines	[5]
Question-requête Question ouverte Engagement à la 1ère personne Expression à la 1ère personne Autres énoncés à la 1ère personne Autres	messagerie d'entreprise	[6]
Acceptation Reconnaissance/appréciation Motivateur d'action Mécanisme de politesse Question rhétorique Question ouverte Question à choix multiple Question en "wh*" Question binaire Rejet de réponse Affirmation Réponse incertaine	BC3	[7]

Table 1.2: Exemples de taxonomies des actes du langage spécifiques à l'analyse de courriels.

Chapter 2

Message processing

Jusqu'à présent, assez peu de travaux adressent le problème de la segmentation de courrier électronique. [8] propose de segmenter les courriels en zones prototypiques telles que la contribution de l'auteur, les citations de messages originaux, la signature, ou encore la formule d'ouverture ou de fermeture. Pour ce faire, il utilise un système basé sur les SVM (machines à vecteurs de support, ou *Support Vector Machines*[1]) et atteint un précision de 87% pour une segmentation en neuf zones. Notre travail contraste en ce que nous nous concentrons sur la segmentation de la contribution de l'auteur (ce que nous appelons le "nouveau contenu").

[9] identifie des groupes de phrases thématiquement proches au travers de multiple messages d'un fil de discussion, sans distinguer courriels et messages de forums. Notre problème diffère, d'une part parce que nous cherchons en premier lieu à effectuer une segmentation rhétorique et non thématique, et en second lieu en ce que nous ne nous intéressons qu'à la cohésion entre phrases consécutives, et non entre phrases distantes.

En ce qui concerne la segmentation de textes d'une manière générale, la plupart des travaux portant sur le sujet ne considèrent que l'aspect thématique des segments. Dans le domaine, il est important de mentionner notamment l'algorithme *TextTiling*, basé sur la notion de rupture lexicale [10]. Il s'agit de l'un des algorithmes les plus communément utilisés pour la segmentation automatique de texte. Si l'algorithme détecte une rupture dans la cohésion lexicale du texte (entre deux blocs consécutifs), il place une frontière pour indiquer un changement thématique. Bien que *TextTiling* soit capable de fonctionner correctement à l'échelle d'un courriel, il ne répond pas directement à notre problème puisque, comme nous l'avons dit, nous cherchons à effectuer une segmentation rhétorique et non thématique.

Nous sommes au courant des travaux récents portant sur la segmentation de texte linéaire, tels que [11], qui tente de résoudre le problème en modélisant le texte sous la forme d'un graphe de phrases et y appliquant des méthodes de regroupement ou de découpage. Cependant, en raison de la petite taille des messages (et par conséquent du modeste volume de matériau lexical que nous avons à disposition), il ne nous est généralement pas possible d'exploiter ce genre de méthode.

[1] `fr.wikipedia.org/wiki/Machine_à_vecteurs_de_support`

Chapter 3

Corpora

La plupart des travaux portant sur les actes du langage et les messages évitent d'annoter eux-mêmes leurs corpus et préfèrent faire appel à des corpus distribués dans la communauté scientifique. Cependant, et notamment en raison de problématiques dues au respect de la vie privée, peu de conversations sont disponibles publiquement.

Le corpus Enron[1] contient plus de 600 000 courriels envoyés par 158 employés de la compagnie Enron [12]. En 2010, EDRM a publié une version étendue de ce corpus, contenant plus de 1,7 millions de messages[2].

Le W3C[3] est le résultat de la récupération de 50 000 fils de conversation tirés du *World Wide Web Consortium*. Le corpus est constitué de courrier de type "entreprise". Les messages extraits de la liste de diffusion de w3c.org est constituée d'environ 200 000 documents. Il est utilisé par [13] pour la modélisation non supervisée d'actes du dialogue dans les courriels. [4] l'utilise en tant que jeu de données non annotées pour une tâche de classification semi-supervisée des actes du langage.

Le *British Columbia Conversation Corpus* (BC3) est utilisé par [4] comme jeu de données annotées pour une tâche de classification semi-supervisée des actes du langage. Il contient 40 fils de discussion tirés du corpus W3C. Chaque fil a été annoté par trois annotateurs différents. Les métadonnées produites comportent notamment des résumés (par extraction) et des actes de discours (*Propose*, *Request*, *Commit* et *Meeting*) [7].

Bien que ce dernier corpus présente certains avantages évidents, ils sont tous constitués de courriels d'entreprise et aucun n'est directement pertinent dans le contexte de l'amélioration de systèmes d'assistance aux utilisateurs. Nous obtiendrons donc nos données autrement, comme nous le montrerons plus tard.

[1] `http://www.cs.cmu.edu/~./enron/`
[2] EDMR Enron Email Data Set v2 Now Available: `http://www.edrm.net/archives/6462`
[3] `http://research.microsoft.com/enus/um/people/nickcr/w3csummary.html`

Bibliography

[1] John Langshaw Austin. *How to do things with words*, volume 1955. Oxford university press, 1975.

[2] John R Searle. *A taxonomy of illocutionary acts*. Linguistic Agency University of Trier, 1976.

[3] David R Traum. 20 questions on dialogue act taxonomies. *Journal of semantics*, 17(1):7–30, 2000.

[4] Maryam Tavafi, Yashar Mehdad, Shafiq Joty, Giuseppe Carenini, and Raymond Ng. Dialogue act recognition in synchronous and asynchronous conversations. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2013)*, SIGDIAL'13, 2013.

[5] Andrew Lampert, Robert Dale, and Cécile Paris. Classifying speech acts using verbal response modes, 2006.

[6] Rachele De Felice, Jeannique Darby, Anthony Fisher, and David Peplow. A classification scheme for annotating speech acts in a business email corpus. *ICAME Journal*, 37:71–105, 2013.

[7] J. Ulrich, G. Murray, and G. Carenini. A publicly available annotated corpus for supervised email summarization. In *AAAI08 EMAIL Workshop*, Chicago, USA, 2008. AAAI.

[8] Andrew Lampert, Robert Dale, and Cécile Paris. Segmenting email message text into zones. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 919–928, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[9] Shafiq Joty, Giuseppe Carenini, and Raymond Ng. Topic segmentation and labeling in asynchronous conversations. *Journal of AI Research (JAIR)*, 47:521–573, 2013.

[10] Marti A Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64, 1997.

[11] Anna Kazantseva and Stan Szpakowicz. Linear text segmentation using affinity propagation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 284–293, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[12] Bryan Klimt and Yiming Yang. Introducing the enron corpus. In *CEAS*, 2004.

[13] Shafiq Joty, Giuseppe Carenini, and Chin-Yew Lin. Unsupervised modeling of dialog acts in asynchronous conversations. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 1807–1813. AAAI Press, 2011.