Chaînes de référence et genre textuel pour la détection automatique de thèmes

Laurence Longo, Amalia Todirascu
LiLPa (Linguistique, Langues, Parole) – Fonctionnements Discursifs & Traduction
Université de Strasbourg, 67084 Strasbourg Cedex
longo@unistra.fr, todiras@unistra.fr

Nous présentons un module d'identification automatique des chaînes de référence (Longo et Todirascu, 2010a, 2010b) intégré dans un système de détection automatique des thèmes en français. La détection automatique de thèmes consiste à identifier les éléments d'un texte qui indiquent son sujet, ses acteurs. Cette tâche permet d'améliorer la navigation textuelle (Nomoto et Matsumoto, 1996) ou la recherche d'information (Callan *et al.*, 1992), (Salton *et al.,* 1993). Au niveau discursif, le thème constituerait un thème composite (Bilhaut, 2006), un agrégat des différents thèmes des phrases qui composent un paragraphe, un texte (Goutsos, 1997). La détection de thèmes est une tâche difficile, vu la multitude de procédés linguistiques de cohésion et cohérence lexicale qui participent à l'identification du thème discursif.

Différents travaux ont proposé d'exploiter ces procédés pour la segmentation thématique (Allan *et al*, 1998) et la détection de thèmes (Hernandez, 2004 ; Bilhaut, 2006) en prenant en compte des marqueurs linguistiques et structurels (Piérard et Bestgen, 2006 ; Ferret *et al*., 2001) et des critères statistiques (Litman et Passoneau, 1995). Nous nous situons dans cette lignée. Notre système de détection automatique de thèmes a comme finalité l'indexation des documents par thèmes. Les thèmes associés au document sont une agrégation des thèmes identifiés dans chaque segment. Dans notre approche, le texte est d'abord découpé en segments thématiquement homogènes à l'aide de réseaux lexicaux (Choi *et al.*, 2001)[1] puis nous identifions automatiquement plusieurs marqueurs linguistiques de continuité ou de rupture thématique : les introducteurs de cadre de discours (Charolles, 1997) listés par (Porhiel, 2004), les répétitions lexicales, les anaphores (Kleiber, 1994) et les chaînes de référence (Cornish, 1995 ; Corblin, 1995 ; Schnedecker, 1997). Les introducteurs de cadres thématiques (e.g. « concernant », « en matière de », etc.) indiquent explicitement le thème du paragraphe (Charolles, 1997 ; Porhiel, 2004). Aussi, les thèmes du discours (Givon, 1983 ; Chafe, 1994 ; Lambrecht, 1994) s'expriment au moyen de marqueurs référentiels tels que les chaînes de référence (CR). Les CR sont des suites d'expressions coréférentielles (Charolles, 1988).

Nous utilisons ainsi une combinaison entre des marqueurs de cohésion « descendants » (les introducteurs de cadre) et « remontants » (les CR) (Charolles et Péry-Woodley, 2005). Nous avons défini plusieurs classes de règles heuristiques, pour chaque catégorie de marqueurs, pour détecter les thèmes possibles de chaque segment. Par exemple, lorsque le premier élément d'une CR coïncide avec le thème introduit par l'introducteur de cadre, le thème est renforcé par les deux marqueurs donc le thème est un thème discursif. La sortie de notre système se présente sous la forme d'une liste de thèmes associés à chaque segment, comme par exemple « les attentes des usagers », « les associations consuméristes », « la réforme de l'Etat ».

[1] La délimitation en segments thématiques permet de se positionner directement dans une partie du document, rendant ainsi la navigation plus aisée.

Nous faisons l'hypothèse que les CR représentent des indices fiables pour participer à la détection des thèmes du discours. En effet, les liens référentiels permettent au lecteur de se focaliser sur un référent unique, qui peut constituer le thème du paragraphe. Le participant le plus mentionné au niveau du paragraphe thématique et ultérieurement au niveau du discours constitue le « thème continu » (Givon, 1983). Nous considérons qu'une CR inclut au moins trois expressions référentielles (ou maillons) (Schnedecker, 1997). De plus, la redénomination du nom propre, alors que le contexte ne le nécessite pas, ouvre une nouvelle CR et un nouveau cadre thématique. Ces deux critères distinguent les CR de la notion de coréférence (suite d'expressions référentielles référant à la même entité dans un texte) et d'anaphore (relation entre deux expressions référentielles). Par exemple (les maillons de la CR sont en gras) :

« **Le Centre de documentation d'études juridiques, économiques et sociales (CEDEJ),** installé au Caire, ne recoupe que partiellement les objectifs de son homologue libanais. Né dans le cadre de la coopération culturelle française de type classique, **il** a opéré, voilà trois ans environ, une mutation qui a suscité l'intérêt des partenaires égyptiens en multipliant **ses** activités avec l'aide de collègues dont quelques volontaires du service national actif. » (Le Monde Diplomatique).

Dans une étude antérieure (Longo et Todirascu, 2009), nous avons comparé les CR issues de divers genres textuels (journaux, textes de loi, textes littéraires) pour déterminer les contraintes qui conditionnent la composition des CR (matériau linguistique privilégié, distance entre les maillons, longueur des CR). Ces spécificités des CR dépendant du genre textuel ont été intégrées à *RefGen*.

RefGen procède en plusieurs étapes. Le texte brut est annoté morpho-syntaxiquement à l'aide de l'étiqueteur TTL (Ion, 2007), proposant des informations détaillées : genre, nombre, temps, mode. Puis, diverses expressions référentielles candidates au poste de maillons de CR sont annotées (entités nommées, groupes nominaux simples et complexes, pronoms *il* impersonnels) via le module *RefAnnot*. Le texte enrichi en annotations passe ensuite dans le module de calcul de la référence (*CalcRef*). Ce module prend en compte a) les paramètres dépendant du genre textuel pour sélectionner les premiers maillons potentiels des CR ainsi que b) un score d'accessibilité global calculé à partir de l'échelle d'accessibilité d'(Ariel, 1990) qui classe les expressions référentielles (noms propres, groupes nominaux, pronoms, déterminants possessifs) suivant leur degré d'activation en mémoire. Puis, la sélection des paires antécédent – anaphore possibles s'effectue par la validation d'une série de contraintes (lexicales, syntaxiques et sémantiques) fortes et faibles, identifiables à partir des annotations. Les paires ainsi formées sont regroupées suivant la propriété de transitivité afin de construire les CR.

RefGen a été évalué suivant les mesures utilisées dans les campagnes d'évaluation de la coréférence : MUC (Vilain *et al.*, 1995), CEAF (Luo, 2005), B^3 (Bagga et Baldwin, 1998) et BLANC (Recasens et al., 2010). Pour l'évaluation, un corpus de référence libre de droits de 15 192 tokens issu de genres divers (journalistique, littéraire, juridique) a été annoté manuellement (1061 expressions référentielles et 267 CR) à l'aide de la plateforme *Glozz* (Wildlöcher et Mathet, 2009). Les résultats de l'application des métriques montrent des variations significatives suivant le genre textuel : la f_mesure du corpus journalistique est comprise entre 44,7% pour CEAF et 59,2% pour BLANC, celle du corpus littéraire varie entre 36% pour MUC et 69% pour B^3 et la f_mesure du corpus juridique varie de 14,3% pour MUC à 81,3% pour B^3. Ces tendances seraient à confirmer sur un plus large corpus. Nous prévoyons aussi de comparer notre système à des systèmes symboliques tels qu'Anasem (Victorri, 2005) ou par apprentissage comme BART (Versley *et al.*, 2008).

Références

Allan, J., Carbonell, J., Dodington, G, Yamron, G, Yang, Y (1998) Topic detection and tracking pilot study final report, Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop

Bagga, A., Baldwin, B. (1998). Algorithms for scoring coreference chains. *Actes de LREC Workshop on Linguistic Coreference*, Granada, Spain, 563 - 566.

Bilhaut, F. (2006). Analyse automatique de structures thématiques discursives, application à la recherche d'information. Thèse de doctorat, université de Caen.

Callan, J. P., Croft, W. B. and Harding, S. M. (1992). The INQUERY retrieval system. *Actes de International Conference on Database and Expert Systems Application* (Berlin and New York: Springer-Verlag), 78-83.

Chafe, W. L. (1994). Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing. Chicago: University of Chicago Press.

Charolles, M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques*, 57, 3–15.

Charolles, M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. *Cahier de Recherche Linguistique 6*, 1–73.

Charolles, M., Péry-Woodley, M.-P. (2005). Introduction, dans Charolles, M., Péry-Woodley, M.-P., eds., Les adver¬biaux cadratifs, *Langue Française*, 148, 3-8.

Choi, F. Y. Y., Wiemer-Hastings P. and Moore J. (2001). Latent semantic analysis for text segmentation, *Actes de NAACL'01*, p. 109-117.

Corblin, F. (1985). Les chaînes de référence : analyse linguistique et traitement automatique. *Intellectica*, vol. 1, n° 1, 123–143.

Corblin, F. (1995). Les formes de reprise dans le discours : Anaphores et chaînes de référence. Rennes : Presses Universitaires de Rennes.

Cornish, F. (1995). Références anaphoriques, références déictiques, et contexte prédicatif et énonciatif. *Sémiotiques 8*, 31–57.

Ferret, O., Grau, B., Minel, J-L. et Porhiel, S. (2001). Repérage de structures thématiques dans des textes, *actes de TALN*, Tours.

Givon, T. (1983). Topic Continuity in Discourse: A Quantitative Cross Language Study. Typological Studies in Language 3. Amsterdam: John Benjamins.

Goutsos, D. (1997). Modeling Discourse Topic: sequential relations and strategies in expository text, *Advances in Discourse Processes*, vol. LIX, Norwood: Ablex Publishing Corporation.

Hernandez, N. (2004). Description et Détection Automatique de Structures de Texte, Thèse de doctorat, Université Paris-Sud XI.

Ion, R. (2007). TTL: A portable framework for tokenization, tagging and lemmatization of large corpora, Bucharest: Romanian Academy.

Kleiber, G. (1994). Anaphores et Pronoms. Louvain-la-Neuve : Editions Duculot.

Lambrecht, K. (1994). « Information structure and sentence form: Topic, focus, and the mental representation of discourse referents ». *Cambridge Studies in Linguistics,* vol. 71, Cambridge: Cambridge University Press.

Litman, D. J., Passoneau, R. J. (1995). Combining multiple knowledge sources for discourse segmentation ». *Actes d' ACL*, 108-115.

Longo, L. et Todirascu, A. (2010a). RefGen: a Tool for Reference Chains Identification. *Actes de CLA'10* (Computational Linguistics-Applications), IMCSIT (International Multiconference on computer Science and Information Technology), pages 447–454, 18-20 octobre 2010, Wisla, Pologne.

Longo, L. et Todirascu, A. (2010b). RefGen: Identifying Reference Chains to Detect Topics. *Actes des 4th International Workshop on Distributed Agent-Based Retrieval Tools (DART 10),* 18 juin 2010, Genève, Suisse.

Longo, L., Todiraşcu, A. Une étude de corpus pour la détection automatique des thèmes*. Actes des 6èmes journées de linguistique de corpus*, 10-12 septembre 2009, Lorient.

Luo, X. (2005). On coreference resolution performance metrics. *Actes de HLT-EMNLP*, Vancouver, Canada, 25-32.

Nomoto, T., Matsumoto, Y. (1996). Exploiting Text Structure for Topic Identification, *Actes des 4th Workshop on Very Large Corpora*, 101-112.

Piérard, S. & Bestgen, Y. (2006b). Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes, *TAL*, volume 47 – n° 2.

Porhiel, S. (2004). Les introducteurs de cadres thématiques, *Cahiers de Lexicologie*, 85, 2004/2, 9-45.

Recasens, M., Hovy, E. (2011). BLANC: Implementing the Rand Index for coreference evaluation. *Natural Language Engineering*, 17(4) :485-510. Cambridge University Press 2010.

Salton, G., Allan J., and Burckley, C. (1993). Approaches to Passage Retrieval in Full Text Information Systems. In Korfhage *et al.,* 49-58.

Schnedecker, C. (1997). Nom propre et chaînes de référence. *Recherches Linguistiques*, 21. Paris : Klincksieck.

Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A. (2008). BART: A Modular Toolkit for Coreference Resolution. Actes de LREC.

Vilain, M., Burger, J., Aberdeen, J. Connolly, D., Hirschman, L. (1995). A Model-Theoretic Coreference Scoring Scheme. *Actes de MUC-6*, 45-52.

Victorri, B. (2005) Le calcul de la référence, *Sémantique et traitement automatique du langage naturel*, Patrice Enjalbert (Ed.) 133-172

Wildlöcher, A. and Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus, *Actes de TALN 2009*, session poster, Senlis, France.