Analyse automatique de structures thématiques discursives

Application à la recherche d'information

Frédérik Bilhaut

Thèse de Doctorat de l'Université de Caen
Sous la direction de Patrice Enjalbert (GREYC)

Résumé

Le problème de l'analyse thématique des textes, qui vise l'étude de leur structure selon des critères relatifs à la répartition de leur contenu informationnel, est d'une importance capitale dans le contexte de l'accès assisté à l'information. Quel est le sujet d'un texte ou d'un passage ? À quel propos apporte-t-il de l'information ? Comment cette information est-elle répartie dans le discours ? Telles sont les questions auxquelles on doit pouvoir répondre pour sélectionner les documents pertinents relativement à une requête, pour aider à la navigation dans des documents longs, ou encore pour produire une résumé d'un texte.

Cette question n'est toutefois pas toujours abordée explicitement en ces termes, et nombre de travaux en recherche d'information (RI) ou en traitement automatique des langues (TAL) gardent certaines distances avec la notion de thème. En RI "classique", la phase d'indexation vise bien l'analyse et la description du contenu informationnel, mais met le plus souvent en oeuvre des méthodes numériques s'appliquant à la surface du texte. Dans ce cas, le document est analysé comme un tout indivisible, et son "thème" représenté par un ensemble de formes quantitativement significatives (typiquement des mots-clefs). En TAL, beaucoup de travaux portent sur la segmentation linéaire, qui revient essentiellement à détecter des ruptures thématiques au fil du texte. Dans ce cas, on dispose bien d'une structure intra-documentaire, mais les méthodes employées sont là aussi souvent numériques et surfaciques, et restent limitées quant à la représentation des thèmes des segments obtenus (quand celle-ci existe en tant que telle). En règle générale, les apports de disciplines telles que la linguistique ou les sciences de l'information quant à la notion de thème restent relativement peu considérés.

Notre travail consiste à envisager l'apport de théories de cet ordre dans le cadre de l'analyse automatique de la structure thématique du discours. Il s'agit plus précisément, tout en conservant une réelle visée applicative, de se concentrer sur la réalité linguistique et documentaire que recouvre par la notion de thème, plutôt que sur les contraintes opératoires qui prévalent habituellement, notamment en termes de temps de calcul et de ressources spécifiques à une langue ou à un domaine. Nous nous confrontons ainsi à d'autres problèmes tout aussi épineux, puisqu'il n'existe pas à ce jour de théorie consensuelle concernant la notion de thème au niveau documentaire ou même discursif. Il ne s'agit donc pas seulement d'opérationnaliser des modèles descriptifs existants, mais de chercher à recenser parmi les théories existantes des éléments susceptibles de servir au mieux les objectifs applicatifs de la recherche d'information. C'est l'objectif que nous nous sommes fixé, en tentant de conserver en parallèle une vue sur les approches théoriques de la notion de thème et une visée applicative marquée.

À cette fin, après avoir posé la problématique générale de l'accès assisté à l'information, nous passons en revue dans notre **première partie** diverses conceptions des notions de *thème*, de *topique*, de *sujet*

ou encore d'*à propos*. Il est bien sûr totalement impossible d'être exhaustif dans cette entreprise, et notre approche à consisté à rendre compte de façon assez précise, parmi les travaux relevant de différentes disciplines, d'un nombre limité de propositions susceptibles de trouver une application pertinente au problème qui nous occupe.

Il s'agit en premier lieu de recherches touchant aux sciences de l'information et à l'ingénierie documentaire (Hutchins, Hjørland, Bruza), qui nous amènent tout d'abord à prendre la mesure de l'ambivalence de la notion de thème définie en termes d'à propos, qui se confronte à la difficulté de situer *en pratique* une frontière franche entre "ce dont on parle" et "ce que l'on dit", et qui peut conduire à hésiter entre une vision du thème comme *ossature* synthétisant l'information véhiculée, et à une vision du thème comme *support* de cette information. En adoptant plutôt ce second point de vue, nous nous concentrons sur les propositions qui approchent la notion de thème comme *point de contact* entre un document (ou une base documentaire) et un (groupe d')utilisateur(s), entre l'information véhiculée et le "socle informationnel" sur lequel elle s'appuie, entre les présupposés du scripteur et les connaissances du lecteur. Nous considérons par suite d'autres approches qui nous invitent à considérer la notion de thème et la relation d'à propos comme intrinsèquement relatives, dépendant de considérations pragmatiques voire épistémologiques, mais dont on peut néanmoins définir et étudier les propriétés de façon rigoureuse et formelle, par exemple avec des outils logiques.

Nous considérons dans un second temps différents travaux en linguistique, où la notion de thème est bien sûr abondamment débattue. Dans ce domaine, le terme recouvre des concepts parfois très différents, et nous nous concentrons sur ceux qui font intervenir la notion d'à propos, et qui sont donc susceptibles de faire écho aux propositions précédentes. Nous tentons de les aborder par ordre de grain linguistique, de la proposition au discours. Il s'agit tout d'abord de théories liées à la *structure informationnelle* (Chafe, Lambrecht), qui consiste essentiellement à étudier le rôle thématique des constituants de la proposition ou de la phrase. Nous envisageons par la suite différentes théories portant sur des grains plus importants, notamment avec la notion de *progression thématique* au niveau inter-phrastique (Combettes), puis celle de *thème discursif* (Brown et Yule, Van Dijk). Enfin, nous considérons des travaux qui ne sont pas explicitement liés à la notion de thème, mais qui sont d'une grande importance dans l'analyse du discours, dont la *théorie du centrage* (Grosz et al.), celle de la *structure rhétorique* (Mann et Thompson), et surtout l'hypothèse de l'*encadrement du discours* (Charolles).

Finalement, nous envisageons des travaux portant explicitement sur l'analyse thématique en traitement automatique des langues. Il s'agit plus précisément de deux approches foncièrement opposées, dont la confrontation montre bien la largeur du spectre des méthodes applicables au problème. Il s'agit tout d'abord de la méthode dite *text-tiling* et de ses dérivés (Hearst, Ferret), qui s'appuient comme nous l'évoquions plus haut sur des méthodes numériques pour aboutir à une segmentation thématique linéaire. Nous évoquons d'autre part une approche visant à obtenir une structuration thématique hiérarchique en se fondant sur la dualité thème / rhème et en exploitant des indices d'ordre linguistique (Lucas).

Nous concluons l'ensemble de ce parcours bibliographique par un bilan visant à en faire émerger un certain nombre de lignes de forces qui sous-tendent notre approche de la notion de thème en tant qu'objet *discursif*, *sémantique*, *structuré*, et lié à la structure des *connaissances* liées à un domaine.

Notre **seconde partie** est consacrée à la description de systèmes et de modèles de traitement automatique des langues s'appuyant sur ces principes. Nous décrivons tout d'abord des travaux portant sur la *recherche d'information géographique*, réalisés au sein du projet GeoSem. Ce projet met en oeuvre des analyses profondes des textes et des cartes dans les documents géographiques, et notamment, pour ce qui est du texte, des analyses d'ordre sémantique et discursif aboutissant *in fine* à une indexation fine des documents exploitable en RI intra-documentaire. Nous présentons nos apports à ce projet, qui concernent tout d'abord l'analyse sémantique des expressions temporelles, tâche qui vise à reconnaître et marquer automatiquement ces expressions dans les textes, et surtout à calculer une représentation symbolique de leur valeur sémantique. Nous évoquons d'autre part le problème de la mise en relation en discours des références à des faits socio-géographiques avec leur localisation spatio-temporelle, tâche qui constitue une

certaine forme d'analyse thématique visant à établir une indexation par passages sous la forme de triplets "phénomène-espace-temps". Nous présentons enfin un prototype de moteur de recherche que nous avons conçu pour exploiter les résultats obtenus et permettre l'expérimentation *in situ* des différentes techniques développées dans la cadre de GeoSem.

Par suite, nous décrivons deux axes de recherche qui ont découlé de ce projet, et auquel nous nous sommes tout particulièrement intéressé. Le premier concerne l'*analyse automatique des cadres de discours temporels*. Nous nous basons ici sur l'hypothèse de l'encadrement de Michel Charolles, qui décrit des segments dits "cadres de discours", homogènes par rapport à un critère sémantique (en l'occurrence une localisation temporelle) spécifié par une expression détachée en initiale de phrase dite "introducteur de cadre". Ces derniers sont présentés comme des marqueurs d'indexation permettant de "répartir les contenus propositionnels dans des blocs homogènes relativement à un critère spécifié par le contenu de l'introducteur", fonction dont on voit immédiatement l'intérêt dans le contexte de la RI. Nous proposons une méthode développée en collaboration avec l'ERSS qui, en s'appuyant sur différents pré-traitements tels que l'étiquetage morpho-syntaxique ou l'analyse sémantique des expressions temporelles, permet de reconnaître automatiquement les bornes des cadres temporels, en proposant de premières solutions au délicat problème de l'analyse automatique de la portée des introducteurs. Nous décrivons le procédé d'évaluation que nous avons commencé à mettre en oeuvre pour tester les résultats obtenus, ainsi que les premières conclusions, encourageantes, que l'on peut en tirer.

Le second concerne la notion de *thème composite*, que nous avons développée dans le prolongement des travaux précédemment décrits, et qui constitue un modèle pour l'analyse thématique d'une certaine variété de structures discursives liées à la notion d'univers de discours. Nous présentons le modèle en lui-même avant de décrire la méthode d'analyse thématique automatique qui en découle. Nous envisageons les liens entre ce modèle et des concepts existants tels que l'encadrement du discours, la structure informationnelle ou encore la théorie de la structure rhétorique, avant d'introduire la notion d'*axe sémantique* que nous posons comme pivot entre l'organisation des connaissances d'un domaine et la structure thématique des textes qui s'y rapportent.

Notre **troisième et dernière partie** décrit la plate-forme *LinguaStream*[1], que nous avons développée parallèlement aux travaux précédemment évoqués pour faciliter leur élaboration, et qui est devenue une plate-forme générique pour le traitement automatique des langues. Elle a pour ambition de simplifier la réalisation d'expériences non triviales sur corpus, ainsi que le cycle d'évaluation / ajustements qui en découle. Sans outil adapté, le coût de développement induit par chaque nouvelle expérience devient en effet un frein considérable à l'approche expérimentale, et pour répondre à cette problématique, LinguaStream facilite la mise en oeuvre de procédés complexes tout en impliquant un investissement technique minimal. Elle a été utilisée pour implémenter tous les procédés d'analyse ici présentés, et est également utilisée à des fins de recherche ou d'enseignement dans plusieurs laboratoires. Son développement est aujourd'hui assuré au sein d'une équipe à laquelle prennent notamment part Antoine Widlöcher et des membres de l'ERSS.

Elle permet la conception et l'évaluation de *chaînes de traitement* par assemblage de modules d'analyse de types et de niveaux variés : morphologique, syntaxique, sémantique, discursif ou encore statistique. Chaque palier de la chaîne se traduit par la découverte et le marquage de nouvelles annotations, sur lesquelles pourront à leur tour s'appuyer les analyseurs subséquents. En fin de chaîne, différents outils permettent de visualiser les documents analysés avec leurs annotations. Plusieurs mécanismes d'élaboration des composants de traitement sont proposés : règles morphologiques, grammaires d'unification, expressions rationnelles, lexiques sémantiques, grammaires de contraintes, moteurs d'inférences, etc. La plupart d'entre eux s'appuient sur des formalismes déclaratifs, certains étant couramment utilisés en TAL et d'autres originaux. Chaque composant d'analyse est réutilisable immédiatement dans d'autres chaînes de traitement, et peut être remplacé par un autre composant fonctionnellement équivalent. Un environnement

[1] `http://www.linguastream.org`

graphique prend en charge les différents aspects de l'élaboration d'une chaîne de traitement complète.

À travers cette plate-forme qui se veut avant tout un "laboratoire virtuel" pour le TAL, nous proposons un certain nombre de principes méthodologiques applicables aux problématiques de l'annotation des documents électroniques et surtout de la constitution de procédés d'analyse complexes fondés sur la formalisation de modèles d'ordre linguistique. Nous développons notamment la notion de *perspective d'analyse*, qui vise à associer à chaque composant d'un traitement un point de vue spécifique sur le texte, ou encore l'opportunité d'exploiter la *complémentarité des modèles d'analyse*, qui permet de faire collaborer différents formalismes préférentiellement adaptés à un ou des niveaux linguistiques particuliers.

Sommaire

Sélection de publications

[1] F. Bilhaut, T. Charnois, P. Enjalbert, and Y. Mathet. Passage Extraction in Geographical Documents. In *Proceedings of New Trends in Intelligent Information Processing and Web Mining (IIPWM'03)*, pages 121–130, Zakopane, Pologne, 2003.

[2] F. Bilhaut and P. Enjalbert. Discourse Thematic Organisation Reveals Domain Knowledge Structure. In *Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI'05)*, pages 2815–2831, Pune, India, 2005.

[3] F. Bilhaut, L.-M. Ho Dac, A. Borillo, T. Charnois, P. Enjalbert, A. Le Draoulec, Y. Mathet, H. Miguet, M.-P. Péry-woodley, and L. Sarda. Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. In *Actes de la 10e Conférence Traitement Automatique du Langage Naturel (TALN'03)*, pages 315–320, Batz-sur-Mer, France, 2003.

[4] F. Bilhaut and A. Widlöcher. LinguaStream : An Integrated Environment for Computational Linguistics Experimentation. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL'06)*, Trento, Italy, 2006.