TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique

F. Bilhaut*, M. Ho-Dac**, A. Borillo**, T. Charnois*, P. Enjalbert*,
A. Le Draoulec**, Y. Mathet*, H. Miguet**, M.-P. Péry-Woodley**,
L. Sarda**

* GREYC, Université de Caen, Campus II
{fbilhaut, charnois, patrice, mathet}@info.unicaen.fr

** ERSS, Université Toulouse Le Mirail
{hodac, aborillo, draoulec, miguet, pery, lsarda}@univ-tlse2.fr

Mots-clefs – Keywords

Analyse automatique de discours, Cadres de discours, Recherche d'information, Document géographique
Automatic discourse analysis, Discourse framing, Information retrieval, Geographical documents

Résumé - Abstract

Cet article concerne la structuration automatique de documents par des méthodes linguistiques. De telles procédures sont rendues nécessaires par les nouvelles tâches de recherche d'information intradocumentaires (systèmes de questions-réponses, navigation sélective dans des documents...). Nous développons une méthode exploitant la théorie de l'encadrement du discours de Charolles, avec une application visée en recherche d'information dans les documents géographiques - d'où l'intérêt tout particulier porté aux cadres spatiaux et temporels. Nous décrivons une implémentation de la méthode de délimitation de ces cadres et son exploitation pour une tâche d'indexation intratextuelle croisant les critères spatiaux et temporels avec des critères thématiques.

This paper proposes linguistics-based methods for the automatic identification of text segments. Such procedures are required by new tasks appearing in intra-document information retrieval (question-answer systems, selective browsing). Our method is based on Charolles' theory of discourse framing and focuses on temporal and spatial frames. We describe an implementation of our method for determining frame boundaries and its exploitation for intradocument indexing combining spatial, temporal and thematic criteria.

1 Cadres de discours et recherche d'information

D'une manière générale, les travaux sur l'indexation et la recherche d'information font peu référence à la structuration discursive des documents. Les méthodes employées sont le plus souvent fondées sur des critères statistiques ou numériques, portant sur la distribution des mots et leur récurrence (Hearst, 1994). Par ailleurs, tant que l'unité pertinente est le document, la structuration interne de celui-ci n'a pas lieu d'intervenir. Mais les applications en traitement automatique des langues ont des frontières mouvantes ; les finalités comme les méthodes se redéfinissent, se redistribuent constamment. On peut identifier actuellement une évolution vers un rapprochement entre recherche d'information, synthèse de documents et systèmes de questions-réponses, et ce à travers le développement de systèmes de navigation sélective. C'est ce type d'application que nous visons ici dans l'univers de l'information géographique[1]. De tels systèmes ne se contentent pas de sélectionner les documents pertinents, ils guident l'utilisateur à l'intérieur de ces documents, vers les segments qui correspondent au plus près à une requête ou à un profil (Desclés & Minel, 2000; Voorhees, 2001). La question de la définition des unités d'information à indexer et/ou marquer en vue de ces parcours devient cruciale.

Les techniques mises en œuvre dans la synthèse de documents et dans les systèmes de questions-réponses sont utiles pour identifier des "zones" contenant des éléments lexicaux pertinents. Mais reste à délimiter des segments de texte formant des unités d'information. Certaines approches statistiques cherchent à identifier des ruptures isolant des segments "thématiquement homogènes" ("text-tiling", (Hearst, 1994)). D'autres, dont nous nous rapprocherons, reposent sur le filtrage de marques discursives (cf. les "cue phrases" de (Hovy & Lin, 1999), les "expressions prototypiques" de (Desclés & Minel, 2000)), mais là encore ces marques sont repérées pour contribuer au calcul du score de saillance des unités les contenant, plutôt que pour leur fonction de structuration.

Le présent article se situe dans une problématique générale différente, celle du repérage de structures proprement *discursives*, et de l'exploitation de cette structuration dans des tâches de recherche d'information (au sens défini plus haut, c'est-à-dire visant une indexation et une navigation *intradocumentaire*). Nous proposons une approche qui se fonde sur un mode de structuration spécifique, "l'encadrement du discours" mis au jour par (Charolles, 1997). Charolles identifie des segments – "cadres de discours" – homogènes par rapport à un critère sémantique (par exemple une localisation spatiale ou temporelle) spécifié par une expression détachée en initiale de phrase (donc aisément repérable) : un *introducteur de cadre* (dorénavant IC). Les IC 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" (op. cit. : 24). On voit immédiatement le bénéfice potentiel du repérage de telles structures pour la recherche d'information et la navigation intratextuelle.

Les documents de géographie humaine constituent un cadre particulièrement attractif à la fois pour l'observation de ces structures, et pour leur exploitation dans une tâche de recherche d'information. En effet, la propriété caractéristique de l'information géographique est d'ancrer ce que nous désignons comme phénomène (le "quoi") dans une localisation spatiale (le "où") et, fréquemment, temporelle (le "quand"). Cette propriété se reflète dans les textes, où l'on observe immédiatement une présence massive d'expressions spécifiant des critères spatiaux/temporels (telles que "Dans l'Ouest de la France" ou "Dans les années 1950, voir exemple [1])[2]. Cor-

[1]Projet GéoSem, soutenu par le programme "Société de l'Information", départements STIC et SHS du CNRS.

[2]Bien que notre corpus comporte plusieurs documents, l'étude présentée ici se fonde principalement sur *Atlas*

rélativement, une requête associera naturellement ces trois "dimensions" de l'information, par exemple : "Evolution des effectifs scolaires dans l'Est dans les années 1980". La réponse devra consister en passages reliant les deux ou trois critères de recherche. Parmi les divers indices textuels permettant de relier ces 3 critères dans le texte, les cadres de discours constituent un élément privélégié.

> [1] **De 1965 à 1985**, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible *dans le Sud-Ouest et le Massif central*, modérée *en Bretagne et à Paris*, l'augmentation a été considérable *dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne* où les effectifs ont souvent plus que doublé.

Nous nous intéressons particulièrement dans cet article au problème délicat de la "portée" des expressions introductrices de cadres, élément clé de leur fonction de structuration du discours et donc de la délimitation des unités d'information pertinentes pour la recherche d'information (section 2). La section 3 présente la mise en œuvre informatique du modèle et son application à l'indexation intradocumentaire de documents géographiques. Partant de ces résultats, la section 4 proposera quelques éléments de réflexion générale sur l'apport d'une approche discursive – telle que la théorie de l'encadrement – au problème de la segmentation thématique.

2 Critères d'évaluation de la portée

Il est indispensable, pour la délimitation de segments, de rechercher des indices susceptibles d'indiquer la fin d'un cadre, c'est-à-dire susceptibles d'indiquer au lecteur que le critère d'interprétation fourni par l'IC ne s'applique plus. Si les mêmes types d'indices semblent devoir s'appliquer aux cadres spatiaux et temporels, le domaine temporel a été pour l'instant le terrain privilégié de ce type d'étude. Trois grands types d'indices de fin de cadre temporel ont été mis au jour : (i) occurrence d'une expression temporelle référant à un temps qui n'est pas sémantiquement compatible avec la période dénotée par l'IC ; (ii) changement de temps verbal : passage (dans un sens ou l'autre) du non-présent (passé-composé, imparfait, etc.) au présent ; (iii) changement de paragraphe.

(Le Draoulec & M.-P. Péry-Woodley, 2001) met en évidence la complexité des interactions entre ces différents indices, qui apparaissent rarement seuls. Mais ce qui nous intéresse ici, dans la mesure où il s'agit de permettre une application informatique du repérage de ces indices à la fermeture des cadres, c'est d'explorer les limites de leur fiabilité. De fait, il apparaît que le seul indice absolument fiable est l'introduction d'un nouvel IC de type incompatible. Pour les autres, leur simple présence ne marque pas à coup sûr la fin d'un cadre : ils ne valent que dans certaines configurations. Ainsi, concernant la présence d'expressions temporelles non IC, il faudrait tenir compte du fait qu'elles peuvent apparaître par exemple dans des relatives, et donc en arrière-plan par rapport à la ligne générale du discours – dans ce type de situation, l'expression temporelle est impropre à déclencher la fermeture du cadre.

Le passage d'un non-présent à un présent ne marque pas nécessairement non plus la fermeture d'un cadre. D'autres indices doivent s'y associer tels que des marques d'un passage au mode

---

de la France Scolaire de la Maternelle au Lycée, R. Hérin et R. Rouault, 1994, 56346 occurrences.

commentaire (utilisation des pronoms "on" ou "nous") ; ou encore, une référence adjectivale au temps d'énonciation (cf. par exemple "actuelle" dans "la tendance actuelle").

Le changement de paragraphe est un indice de fin de portée plus fort. Cependant, quelques réserves sont là encore à faire. En effet, dans le cas des énumérations où chaque item peut correspondre à un paragraphe, un critère spatial/temporel introduit en amorce peut étendre sa portée sur l'ensemble des items.

Les indices mentionnés ici, et malgré les réserves émises à leur propos, seront utilisés comme indices de fin de portée dans la mise en œuvre informatique. A ce stade, cette mise en œuvre, autant qu'une première ébauche de l'application visée, constitue un banc d'essai, permettant ultérieurement d'affiner l'analyse linguistique.

3 Mise en oeuvre informatique

La mise en oeuvre automatique de l'approche linguistique vise à établir des liens entre les phénomènes et leur localisation dans l'espace et/ou le temps, se traduisant par une indexation multi-facettes. L'implémentation de la méthode d'analyse a été intégralement réalisée grâce à la plate-forme *LinguaStream*[3]. La chaîne de traitement sera présentée ici succintement, en se focalisant sur les aspects liés aux cadres du discours (le processus complet est décrit dans (Bilhaut & al., 2003)). Outre les manipulations permettant d'obtenir un document XML exploitable par la plate-forme, cette chaîne intègre en premier lieu des analyseurs morphologiques classiques, procédant au découpage en mots simples, puis à l'étiquetage et à la lemmatisation.

La première étape concerne l'extraction et l'analyse sémantique des expressions spatiales, temporelles, et dénotant les phénomènes. Dans les deux premiers cas, la recherche des syntagmes correspondants repose sur une analyse morpho-syntaxique locale. Réalisés en Prolog, les analyseurs sont basés sur des grammaires d'unification, la reconnaissance et l'analyse sémantique du syntagme étant effectuées simultanément par compositionnalité.

Les expressions dénotant les phénomènes sont quant à elles extraites à l'aide d'une grammaire reconnaissant les syntagmes nominaux composés, d'après les étiquettes morphologiques et sans utilisation de lexique. Cette étape générant un grand nombre de candidats, différents traitements de nature statistico-terminologique sont appliqués, reposant principalement sur des procédés classiquement utilisés en recherche d'information, basés sur l'étude statistique de la distribution des formes (de type $tf \cdot idf$ – Term Frequency by Inverse Document Frequency).

A l'issue de ces différentes analyses, l'annotation sémantique des document qui en résulte permet de procéder à leur analyse au niveau discursif. En s'appuyant sur des modèles linguistiques, et particulièrement celui de l'encadrement du discours, elle permet d'établir des corrélations entre expressions spatiales/temporelles et "phénomènes" de façon plus pertinente que par simple étude de leur co-présence dans le discours, qui conduirait trop fréquemment à l'établissement de relations sémantiquement invalides. Les cadres spatio-temporels jouant un rôle prépondérant dans cette analyse, la détection des IC correspondants constitue l'étape suivante du traitement. En s'appuyant sur le marquage des expressions spatiales et temporelles, l'opération consiste à détecter leur présence en position détachée en début de phrase. Lorsqu'un IC est rencontré, la valeur sémantique correspondante devient un critère d'interprétation qui sera associé aux phénomènes rencontrés à l'intérieur du cadre introduit. Quand l'IC est rencontré à l'intérieur

[3] http://www.info.unicaen.fr/~fbilhaut/linguastream.html

d'un cadre de type différent (ici temporel/spatial), le nouveau critère vient enrichir le contexte courant, et pourra alors aboutir à la localisation des phénomènes à la fois dans l'espace et le temps. Ce mécanisme nécessite bien-sûr la détection de la borne finale des cadres, problème complexe pour lequel les critères de portée décrits en section 2 sont appliqués.

A l'issue de ce processus, le système produit une indexation multi-facettes des documents, composée de triplets (phénomène/espace/temps). Cette indexation est exploitée par un moteur de recherche capable de répondre à des requêtes portant simultanément sur les 3 dimensions. Les résultats sont présentés à l'utilisateur en mettant en valeur des passages déterminés dynamiquement, délimités principalement à l'aide des bornes des cadres sélectionnés.

L'évaluation du système est difficile, en raison de la complexité du processus dans son ensemble, et surtout de la subjectivité de la notion de portée. De plus, il n'existe pas à notre connaissance d'analyseur équivalent permettant de procéder à une évaluation par comparaison. Nous privilégions donc l'évaluation basée sur l'annotation manuelle. Concernant l'analyseur de cadres temporels, une évaluation quantitative à partir de notre corpus d'étude annoté manuellement mesure un taux de précision de 88% et un taux de rappel de 79%. Afin d'obtenir une mesure plus significative, une phase d'évaluation systématique est en cours, reposant sur l'annotation manuelle des cadres dans un vaste corpus.

4 Les cadres pour la structuration thématique

A la lumière de cette expérience, nous souhaitons conclure en soulignant l'intérêt du modèle de l'encadrement dans le contexte plus général de l'analyse thématique automatique du discours. Ce type de méthode vise à délimiter automatiquement des segments textuels relatifs à un même thème, et éventuellement à en produire une représentation symbolique. Comme évoqué en section 1, on trouve majoritairement des approches s'appuyant sur un modèle linéaire du discours, qui se trouve alors segmenté en fragments adjacents et non superposés auxquels on attribue une certaine homogénéité thématique, dans la lignée de (Hearst, 1994). Mais cette approche trouve d'importantes limites lorsqu'elle est appliquée au document géographique, qui fait émerger une conception différente de la notion de thème. On peut en effet considérer que dans ce type de discours, certaines références spatiales ou temporelles font partie intégrante du thème que l'on peut attribuer à un segment textuel donné, par exemple "La France des années 1980" ou "Le retard scolaire dans l'Ouest depuis 1985".

Les méthodes s'appuyant sur la distribution des formes de surface se révèlent ici inefficaces puisque la compatibilité sémantique entre références spatiales ou temporelles ne se traduit quasiment jamais par une similitude lexicale, et le phénomène d'isotopie dans un cadre spatio-temporel est rarement engendré par une récurrence lexicale ou sémantique. L'analyse des introducteurs associée à un calcul de portée semble donc constituer une approche adaptée à ce problème, puisque l'analyse d'un seul syntagme aisément identifiable automatiquement permet d'obtenir un ancrage thématique pertinent pour l'ensemble d'un segment. L'utilisation de méthodes d'analyse sémantique telles que présentées dans la section 3 permet en outre d'en obtenir une représentation symbolique, et non purement quantitative ou numérique.

Dans (Ferret & al., 2001), les auteurs proposent l'exploitation des introducteurs de cadres thématiques comme indices venant confirmer ou infirmer les choix de segmentation thématique issus de critères statistiques. Cette approche semble cependant négliger la fonction structurante des cadres du discours, qui ne peut se réduire à la fonction d'"indice" de leurs introducteurs,

mais qui pose les problèmes difficiles de la portée et des relations inter-cadres. La nature hiérarchique de ce modèle autorise en effet la représentation de structures imbriquées, qui ne peuvent être représentées dans un modèle linéaire. Si des travaux tels que (Porhiel, 2001) s'attachent à la détection des cadres dits "thématiques" en intégrant la notion de portée, on se heurte alors au problème de la marginalité de ce type de cadres (du moins dans notre corpus, raison pour laquelle nous utilisons des approches statistiques pour la dimension "phénomènes"). Il faut donc souligner l'intérêt d'une approche intégrant le modèle de l'encadrement comme élément de structuration non linéaire *et* des méthodes de nature statistique, qui pourraient fournir des indices supplémentaires de fin de portée, ou apporter des éléments de structuration lorsque les introducteurs ne sont pas présents. Inversement, les critères statistiques peuvent tirer parti de la détection des introducteurs de cadres, puisque ces derniers renforcent la saillance thématique des formes qui les suivent. Cette idée est exploitée dans le système d'indexation présenté ici, pour renforcer la pondération attribuée aux phénomènes situés dans le champ direct d'un introducteur spatial ou temporel. Le mécanisme mis en place repose sur le fait que, dans le discours géographique, les phénomènes thématiquement pertinents ont de fortes chances d'être localisés spatialement et/ou temporellement : les introducteurs de cadres spatiaux et temporels fonctionneraient ainsi également comme introducteurs thématiques.

Références

F. Bilhaut, T. Charnois, P. Enjalbert, Y. Mathet, 2003, *Passage extraction in geographical documents*, New Trends in Intelligent Information Processing and Web Mining, Zakopane, Poland (à paraître).

T. Charnois, Y. Mathet, P. Enjalbert, F. Bilhaut, *Geographic reference analysis for geographic document querying*, Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL-HLT 2003), 31 Mai - 1er juin, Edmonton, Alberta, Canada.

M. Charolles, 1997, *L'encadrement du dicours - Univers, champs, domaines et espace*, Cahier de recherche linguistique, 6, pp. 1-60.

J.-P. Desclés, J.-L. Minel, 2000, *Résumé automatique et filtrage des textes*, In Ingénierie des langues, ed. J.-M. Pierrel, Hermès.

O. Ferret, B. Grau, J.-L. Minel, S. Porhiel, 2001, *Repérage de structures thématiques dans des textes*, Actes de TALN 2001, Tours, pp. 163-172.

M. Hearst, 1994, *Multi-paragraph segmentation of expository texts*, 32th Annual Meeting of the Association for Computational Linguistics, 9:16.

M. Ho-Dac, A. Le Draoulec, M.-P. Péry-Woodley, 2001, *Cohabitation des dimensions temps, espace et "phénomènes" dans un texte géographique*, Cahiers de Grammaire 26, "Sémantique et Discours", pp. 125-142.

E. Hovy, C. Lin, 1999, *Automated text summarization in SUMMARIST*, In I. Mani and M. Maybury (1999), (eds.) Advances in Automatic Text Summarization. Cambridge : The MIT Press, pp. 81-94.

A. Le Draoulec, M.-P. Péry-Woodley, 2001 *Corpus-based identification of temporal organisation in discourse*, Proceedings of the Corpus Linguistics 2001 Conference, Lancaster, pp. 159-166.

S. Porhiel, 2001, *Linguistic expressions as a tool to extract thematic information*, Corpus Linguistic, Lancaster, pp. 477-482.

E. Voorhees, 2001, *Overview of the TREC 2001 Question Answering Track*, http://trec.nist.gov/pubs/trec10/t10_proceedings.html.