La recherche d'information évolutive dans des documents de type encyclopédique : l'apport de techniques linguistiques

Marion Laignelet

ERSS - UMR 5610
Université Toulouse 2 - Le Mirail
5 allée A. Machado
31000 TOULOUSE

Société INITIALES
49 rue de Chio
34000 Montpellier
marion.laignelet@univ-tlse2.fr

RÉSUMÉ. *Dans cet article nous présentons la notion d'information évolutive : le développement de ce concept s'inscrit dans le cadre d'un projet de recherche industriel visant la recherche automatique de segments textuels nécessitant une mise à jour de l'information dans un but éditorial. Pour répondre à cet objectif nous faisons l'hypothèse de la nécessité d'associer des techniques issues de la recherche d'information à des techniques linguistiques.*

ABSTRACT. *In this article we present the concept of evolving information : the development of this concept lies within the scope of an industrial research project aiming at the automatic search for textual segments requiring an update of information with a leading aim. To meet this aim we make the assumption of the need for associating techniques resulting from the search for information linguistic techniques.*

MOTS-CLÉS : *information évolutive, RI, EI, segmentation, EN, cadres de discours*

KEYWORDS: *evolving information, IR, IE, segmentation, nammed entities, discourse framing*

CORIA 2007, pp. 449–454
Saint-Étienne, 28–30 mars 2007

1. Introduction

Dans cet article, nous souhaitons montrer en quoi la linguistique et les techniques développées en traitement automatique des langues (T.A.L.) sont en mesure de fournir des solutions concrètes à une problématique industrielle précise dans le cadre d'un projet unissant une société d'édition-packaging (INITIALES, Montpellier) et un laboratoire de recherche en linguistique (ERSS, Toulouse). L'intérêt, pour les premiers, est de trouver une réponse à une contrainte croissante dans le milieu de l'édition : le problème de la mise à jour de l'information et des contenus encyclopédiques. En tant que chercheur en linguistique, nous souhaitons montrer que pour y répondre, les méthodes linguistiques et de T.A.L. sont incontournables. Cette étude se situe donc à la limite entre plusieurs domaines et fait appel aux technologies développées au sein de plusieurs disciplines dont la recherche d'information (R.I.) mais également l'extraction d'information (E.I.) et le T.A.L. Dans une première partie, nous décrirons la notion de mise à jour de l'information ainsi que les segments que nous cherchons à repérer automatiquement. Puis un point méthodologique explicitera notre démarche. Enfin, nous montrerons l'intérêt de considérer conjointement des techniques propres à la R.I. et des approches linguistiques.

2. Information évolutive et mise à jour de l'information

Un segment d'information évolutive (ou SEDIS-ε[1]) est un segment textuel susceptible de contenir une ou plusieurs informations qui présente(nt) la particularité de pouvoir évoluer dans le temps et/ou qui relativement à des besoins éditoriaux nécessiterai(en)t d'être réactualisé(es).

1. Actualité

§ Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux. Toutefois, il convient de rappeler un certain nombre de découvertes très récentes. En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

1.1. Un vaccin contre le sida ?

§ Des recherches portant sur les prostituées [...]. La recherche se tourne justement aujourd'hui vers des vaccins qui [...]. Des expériences ont été faites pour [...]. En juin 2003, une équipe de biologistes américains a obtenu des résultats qui pourraient laisser envisager [...]. Les chercheurs sont parvenus [...]. Cette découverte pourrait aboutir à la mise au point d'un antigène [...]. [. . .]

Source : Corpus ATLAS (fiche Médecine - Le Sida)

Figure 1. *Un segment d'information évolutive*

Dans l'exemple 1, l'auteur exprime une issue possible et probable concernant les recherches sur le sida. La fiche de laquelle est extrait ce passage a été éditée et distribuée dans le courant de l'année 2003 et elle est destinée à être distribué dans le cadre

1. De SEGment de DIScours et ε pour faire référence à la notion d'*information évolutive*.

d'éditions dites « au long cours » : un client peut s'abonner à l'encyclopédie en 2007 et est susceptible de recevoir cette fiche écrite en 2003. Cependant, pendant ce laps de temps (de 2003 à 2007), soit certaines des prédictions formulées par l'auteur se seront réalisées, soit elles auront été repoussées par les scientifiques, ou encore de nouvelles données peuvent entrer en jeu. Il est donc tout à fait souhaitable que ce segment de texte ait été préalablement mis à jour. Nous distinguons deux types de SEDIS-ε. Tout d'abord, les segments textuels nécessitant une ***mise à jour*** : l'information n'est plus vraie ou ne s'est pas vérifiée (c'est souvent le cas lorsque l'auteur fait des prédictions sur un fait ou un événement). Les SEDIS-ε à **réactualiser** quant à eux sont des segments dans lesquels l'information restera vraie dans l'absolu mais, en vue d'une ré-édition, les événements et dates associés doivent être modifiés pour faire référence à un moment plus proche du moment de lecture/réédition.

> [...] On peut ainsi savoir qu'une personne est infectée longtemps avant que la maladie ne se déclare. Il n'existe pas à l'heure actuelle de vaccin contre le sida. [. . .]
>
> Source : corpus ATLAS (fiche Médecine - Le Sida

Figure 2. *Une mise à jour*

> L'organisation mondiale de la santé (OMS) estime, en effet, à 160 millions le nombre annuel de nouveaux cas dans le monde en 2002. [. . .]
>
> Source : ATLAS (fiche Médecine - Les maladies du travail)

Figure 3. *Une réactualisation*

Visant une application concrète, le contexte de la phrase est considéré comme la taille minimale requise pour que la personne chargée de mettre à jour l'information ait un contexte d'interprétation suffisant : c'est ce que nous appelons des **segments d'interprétation**. Il se définit comme un segment textuel de longueur indéterminée présentant une homogénéité sémantique (temporelle, spatiale, etc), pouvant contenir des éléments ne nécessitant pas de mise à jour et enfin, qui contient des SEDIS-ε **minimaux**[2] et/ou des indices. C'est précisément dans ce cas que l'association des techniques linguistiques à des techniques plus robustes nous semble pertinente. Nous en verrons un exemple dans la dernière partie (exemple 4, p. 5). Ce travail s'inscrit dans le cadre des études sur la cohérence discursive. Nous définissons un texte comme un ensemble structuré et hiérarchisé au sein duquel interagissent des sous-ensembles cohérents entre eux et en eux-mêmes que l'on appelle des segments discursifs (Pery-Woodley, 2005). De nombreuses recherches (menées en linguistique ou en psycholinguistique) cherchent à rendre compte de ces processus de (re-)construction de la cohérence que ce soit à travers la délimitation de segments de discours ou de la description des relations entre ces segments (*cf.* RST, SDRT, segmentation thématique,

2. Dans l'exemple 3, cela correspond à la valeur chiffrée et la date soit aux deux expressions soulignées.

etc.). De nombreux travaux visent cet objectif de segmentation automatique : certains dans un objectif de résumé automatique (Saggion *et al.,* 2002, Minel, 2002, Marcu, 2000), de recherche/sélection d'information importante (Rossi *et al.,* 1991) ou encore d'aide à la navigation. Dans tous les cas, la segmentation est envisagée pas rapport à une application précise, un point de vue sur les textes (Hernandez, 2004). En ce qui nous concerne, il s'agit de délimiter des segments discursifs particuliers, les SEDIS-ε.

3. Méthodologie

La première étape a pour objectif la description linguistique des SEDIS-ε. Cette étape est fondamentale pour envisager un travail de formalisation de ces objets particuliers, puis d'implémentation. Nous travaillons sur un corpus constitué de 92 textes rééls, non modifiés pour la tâche. Il s'agit de fiches encyclopédiques éditées et accessibles sur le marché de l'édition (propriété des Editions Atlas). Sur ces 92 fiches, nous avons procédé à l'annotation manuelle des SEDIS-ε de 38 d'entre elles. D'un côté, cette annotation manuelle permet de constituer un corpus de données textuelles nécessaires pour la description (linguistique) des SEDIS-ε. D'un autre côté, et avec les fiches non annotées, nous disposerons d'un corpus de réference utile pour valider nos résultats à l'issue du projet[3]. Nous travaillons sur la plateforme LinguaStream[4] (Widlöcher *et al.,* 2005) car elle permet d'effectuer des traitements et des analyses de types et de niveaux linguistiques variés (morphologique, syntaxique, sémantique, discursif ou encore statistique). Entièrement basée sur le langage XML, LinguaStream permet également de travailler directement sur notre corpus annoté manuellement des SEDIS-ε. Les divers indices[5], qui sont potentiellement des marqueurs de SEDIS-ε, sont ainsi projetés sur le corpus annoté manuellement. Nous observons ensuite quels indices apparaissent effectivement dans un SEDIS-ε et nous quantifions leur distribution, à l'intérieur d'un SEDIS-ε ou non. Des résultats ont été publiés dans (Laignelet, 2006). Les indices tels que les sigles sont repérés à l'aide d'un système d'annotation par expressions régulières permettant de spécifier des contraintes directement sur la forme de surface des unités marquées et d'associer des annotations en fonction des patrons reconnus. Nous repérons également les marqueurs de temps (syntagmes nominaux temporels, adverbiaux temporels, adverbes, etc) en les caractérisant selon qu'ils sont ou bien déictiques, ou bien porteurs d'une référence temporelle proche du "moment de lecture". Nous exploitons des indices tels que des valeurs chiffrées, des superlatifs et nous projetons de prendre en compte les noms de pays, de ville ou encore les noms propres. Au repérage de ces indices et à leur annotation sémantique, nous ajoutons une analyse linguistique discursive faisant appel notamment à la position de ces marqueurs dans le document et/ou dans la phrase. Ainsi le fait qu'un de ces marqueurs apparaisse dans un titre (le

3. Nous sommes également en train d'augmenter la taille de notre corpus avec de nouveaux textes encyclopédiques.
4. http ://www.linguastream.org
5. des adverbiaux temporels, des syntagmes nominaux temporels, des superlatifs, des lexiques spécifiques (temps, évolution, etc), des superlatifs, des marqueurs aspecto-verbaux, des marqueurs de la prise en charge de l'énonciateur, des sigles, des noms propres, etc.

premier exemple de la page 2 est caractéristique) est pris en compte pour déterminer s'il s'agit d'une borne initiale de SEDIS-ε ou plus précisément de segment d'interprétation.

4. Combiner les différents indices pour le repérage et la segmentation des SEDIS-ε

Notre hypothèse est que c'est par la combinaison d'indices de niveaux linguistiques différents (lexical, morpho-syntaxique, discursif) que la description puis la formalisation de ces segments prend consistance. A la suite de (Hernandez, 2004) ou encore (Enjalbert, 2005), nous mettons en place un système associant des techniques telles que la recherche d'entités nommées (ainsi que leur typage sémantique) et des techniques linguistiques présentant la caractéristique d'être moins robustes mais plus fines. Nous avons montré dans (Laignelet, 2006) que prendre en considération les indices de manière isolée est insuffisant pour déterminer si le segment (*a minima* la phrase) dans laquelle l'indice est présent peut être considéré comme un SEDIS-ε. Dans cette optique, associer un repérage d'entités nommées à la délimitation de cadres de discours ou la prise en considération des titres représente un gain pour le repérage (automatique) des SEDIS-ε dans la mesure où ils permettent l'ouverture de segments d'interprétation. L'exemple qui suit montre l'intérêt de travailler sur des indices ayant la capacité à fonctionner comme des cadres de discours (Charolles, 1997).

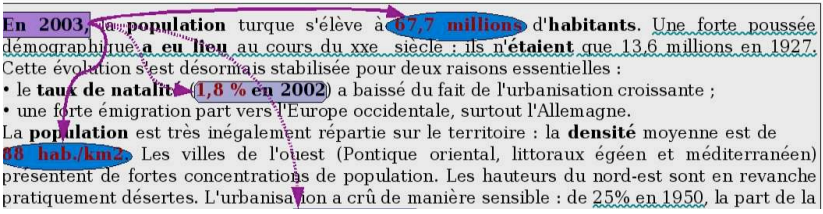

Figure 4. *Un segment d'interprétation*

Dans cet exemple, le segment s'ouvre sur un introducteur de cadre temporel. L'intérêt de considérer l'IC temporel *« En 2003 »* (dans l'encadré) est que le critère sémantique (*la référence temporelle « 2003 »*) qu'il véhicule est valable pour l'ensemble du paragraphe. Ainsi, les deux valeurs chiffrées dans les ovales ont une relation (temporelle) à travers l'expression « En 2003 ». Les deux éléments dans les encadrés arrondis sont également des informations à mettre à jour du fait de leur proximité temporelle. Dans ce cas, il est important de noter que toutes les informations contenues dans ce segment ne sont pas à mettre à jour et notamment les propositions soulignées (en ondulé), pour lesquelles une référence temporelle différente est explicitement signalée. Considérer le modèle de l'encadrement du discours permet de considérer une nouvelle technique de segmentation prenant en compte des critères sémantiques plus fins, la principale difficulté étant de repérer automatiquement la borne finale du segment

d'interprétation. Nous menons actuellement une expérimentation mettant en jeu une segmentation automatique du type *TextTiling* (Hearst, 1994). L'objectif est de mesurer l'intérêt de la prise en compte des ruptures thématiques pour le repérage des segments d'interprétation.

5. Conclusion

Si nous comparons les objectifs de ce travail à ceux de la R.I. "traditionnelle", ce qui nous différencie est que nous ne cherchons pas à vérifier si un document ou un segment de texte est pertinent (*relevant*) par rapport à une requète donnée ; nous recherchons les documents ou les segments textuels présentant une qualité particulière, celle de posséder des informations de nature évolutive. La démarche proposée pour répondre à cet objectif spécifique se situe dans la continuité de nombreux travaux en T.A.L. en cela qu'elle cherche à associer des traitements et des analyses de types et de niveaux différents. Dans le cadre de ce travail, le développement de techniques linguistiques telle que la prise en compte de l'encadrement du discours ou des titres associées au repérage d'indice plus surfastiques ou encore de traitements statistiques (TextTilling) permet d'améliorer les résultats de manière qualitative.

6. Bibliographie

Charolles M., « L'Encadrement du Discours, Univers, Champs, Domaine et Espaces », *Cahiers de Recherche linguistique*, 1997.

Enjalbert P., *Sémantique et TALN*, Hermes, 2005.

Hearst M., « Multi-paragraph segmentation of expository texts », *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, 1994.

Hernandez N., Description et Détection Automatique de Structures de Texte, PhD thesis, Université de Paris XI, Décembre, 2004.

Laignelet M., « Repérage de segments d'information évolutive dans des documents de type encyclopédique », *Actes de la 13ème conférence sur le traitement automatique des langues naturelles*, RECITAL, Presses Universitaires de Louvain, Louvain, Belgique, 2006.

Marcu D., *The Theory and Practice of Discourse Parsing and Summarization*, The MIT Press, 2000.

Minel J.-L., *Filtrage sémantique, du résumé automatique à la fouille de textes*, Hermès, 2002.

Pery-Woodley M.-P., *Discours, corpus, traitements automatiques*, Hermès, 2005.

Rossi J., Bert-Erboul A., « Sélection des informations importantes et compréhension de textes. », *Psychologie Française*, 1991.

Saggion H., Lapalme G., « Generating informative and indicative summaries with SumUM », *Computational Linguistics*, vol. 28, n° 4, p. 497-526, Décembre, 2002.

Widlöcher A., Bilhaut F., « La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus », *Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN)*, Dourdan, France, 2005.