RÉGAL (Résumé Guidé par les Attentes du Lecteur) : un modèle d'exploration sémantique de textes guidé par les points de vue du lecteur

Gérard Sabah

RÉGAL (Résumé Guidé par les Attentes du Lecteur) : un modèle d'exploration sémantique de textes guidé par les points de vue du lecteur

Responsable scientifique : Gérard SABAH

Gérard SABAH
Laboratoire : LIMSI
BP 133, 91403 Orsay Cedex
Tél. : 01 69 85 80 03
Fax : 01 69 85 80 88
E-mail. : gs@limsi.fr

Sous-thèmes dont relève ce projet :
Traitement automatique de la langue
Compréhension et production de textes
Lexique
Sémantique
Syntaxe

Équipes partenaires

- Équipe LaLICC (FRE2520 du CNRS, Université Paris-Sorbonne : 96 Boulevard Raspail – 75006 Paris)
- UMR LATTICE – Langues, Textes, Traitements Informatiques et Cognition (ENS-Ulm : 1 rue Maurice Arnoux – 92120 Montrouge)
- CEA/Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue (LIC2M) (18, rue du Panorama – BP 6 – 92265 Fontenay aux Roses Cedex)

Résumé signalitique

Afin de pallier les défauts des moteurs de recherche actuels (essentiellement trop grand nombre de documents proposés comme réponse à une requête), notre projet visait à offrir des outils de visualisation rapide des textes sélectionnés afin que l'utilisateur puisse évaluer leur pertinence par rapport à son besoin. Ce mécanisme se fonde sur un modèle d'analyse et de représentation de textes permettant de s'adapter aux différents points de vue d'un lecteur, ainsi que sur les techniques d'analyse nécessaires à la production de résumé.

La dimension originale de notre projet consiste à produire un résumé « dynamique », fonction des désirs et des besoins d'un utilisateur, donnant ainsi la possibilité d'une exploration personnalisée d'un texte. Il s'applique à tout type de texte, quel que soit le sujet traité. Le modèle est implémenté dans la plate-forme Filtext de l'équipe LALICC qui avait déjà développé, avec ContextO, un système de résumé par extraction.

Principaux résultats

- Les modèles, linguistiques et informatiques, sur lesquels s'appuient les réalisations ont été définis.
- Les règles permettant la reconnaissance des introducteurs de cadre ont été intégrées dans la plate-forme ContextO.
- L'intégration des deux approches pour la segmentation d'un texte a été réalisée.
- La construction d'une structure globale du texte existe. Il reste à l'étendre pour intégrer certaines marques linguistiques. Néanmoins, le système peut être aussi utilisé tel quel.
- La visualisation d'un texte est montrée par un prototype qui va permettre de préciser le modèle de navigation. Pour cela, il faudra une collaboration avec des ergonomes afin de mettre en place des protocoles d'utilisation du système et recueillir des données sur le type de navigation mise en œuvre par chaque utilisateur.
- Enfin, les divers modules développés en tant que systèmes autonomes (ce qui facilite leur test) ont été intégrés dans la plate-forme ContextO

Mots-clés : Analyse thématique de texte • analyse sémantique • exploration contextuelle • résumé dynamique • adaptation à l'utilisateur

Nombre de participants : 32 (Linguistique : 8 ; Informatique : 15)

Nombre total d'homme-mois : 131

Rappel des enjeux et objectifs fixés à l'origine

Le projet soumis vise au développement d'un modèle d'analyse et de représentation de texte permettant de s'adapter aux points de vue d'un lecteur lorsqu'il consulte un document afin de lui fournir l'information qu'il juge pertinente. Cette problématique repose sur les techniques d'analyse nécessaires à la production de résumé, la dimension originale de notre projet consistant à produire un résumé « dynamique », fonction des désirs et besoins d'un utilisateur, donnant la possibilité d'une exploration personnalisée d'un texte. Ce modèle est destiné à être implémenté dans la plate-forme Filtext du CAMS, qui a déjà développé, avec ContextO, un système de résumé par extraction.

Ce modèle met en jeu deux idées essentielles qui expliquent les développements prévus dans le projet soumis :

- le système de fouille et de filtrage doit être adapté aux besoins des utilisateurs et conçu de telle sorte que ceux-ci puissent l'enrichir en fonction des informations qu'ils recherchent ;
- ce système, pour répondre aux attentes du maximum d'utilisateurs, doit reposer sur des indicateurs linguistiques indépendants des sujets abordés dans les textes traités, l'intégration de connaissances du domaine demeurant possible, pour améliorer ses performances.

La conception et le développement d'un tel système supposent une collaboration étroite et de longue haleine entre linguistes et informaticiens. Les améliorations prévues dans le projet soumis vont dans le sens d'un approfondissement et d'un élargissement de cette collaboration. Ils portent sur deux points qui sont étroitement liés, à savoir : le repérage des unités thématiques et la segmentation des données textuelles.

En ce qui concerne le repérage des unités thématiques, il s'agira de faire collaborer des procédures de calcul prenant en compte des indicateurs lexicaux, à même de fournir très rapidement des indications sur le thème d'un segment de texte et les changements thématiques d'un segment à un autre, avec des marqueurs linguistiques porteurs d'informations quant au rôle des segments du point de vue argumentatif ou discursif. Ces procédures, déjà explorées dans les travaux de l'équipe L & C du LIMSI, seront affinées et intégrées dans ContextO.

En ce qui concerne la segmentation, la plupart des systèmes s'en tiennent au découpage en paragraphes qui est une unité assez grossière et parfois peu motivée. Pour étoffer cette dimension, il est prévu d'intégrer dans le système les expressions introductrices de cadres de discours (groupes adverbiaux détachés en tête de phrases) qui signalent la façon dont un rédacteur répartit les informations dont il fait état dans des rubriques homogènes. L'analyse linguistique de ces marqueurs et les principes gouvernant la mise en place des cadres de discours seront pris en charge par l'équipe LATTICE.

La fouille et le filtrage d'informations textuelles sont des activités auxquelles les lecteurs se livrent en permanence. On ne dispose pas de données psychologiques précises sur ces activités. Il est probable que les connaissances du domaine jouent un rôle important dans ces activités, toutefois, on peut penser que les lecteurs s'appuient aussi pour les mener à bien sur des indicateurs linguistiques du genre de ceux pris en compte dans le système ContextO.

Les contraintes liées à l'implémentation, en ce qu'elles obligent à expliciter le maximum de facteurs à même de peser sur une décision d'extraction, constituent une excellente base pour la formulation d'hypothèses neuro-psychologiques sur les démarches naturelles de fouille et de filtrage. Une des dimensions du projet soumis consistera à rechercher des partenaires en vue de l'opérationnalisation à court terme de ces hypothèses.

Résumé des résultats effectivement atteints

• Les modèles, linguistiques et informatiques, sur lesquels s'appuient les réalisations ont été définis (LATTICE).

• Les règles permettant la reconnaissance des introducteurs de cadre ont été intégrées dans la plate-forme ContextO (LATTICE + LALICC).

• Un nouveau module de segmentation a été réalisé par le CEA, afin de disposer d'un module plus paramétrable et plus standard vis-à-vis de l'état de l'art.

• Une intégration des deux approches pour la segmentation d'un texte a été réalisée au LIMSI, validée par des tests préalables (ce qui a donné lieu à un article présenté à TALN 2001). Ce résultat est directement issu de l'interaction entre les différents participants, et, même si il y a encore des éléments à valider et à ajouter dans le modèle, cela constitue un résultat qui n'aurait pu voir le jour sans ces interactions, aussi bien grâce aux aspects théoriques apportés par chacun en ce qui concerne la structuration d'un texte, qu'aux aspects informatiques qui ont permis de concrétiser les propositions faites par chacun.

• La construction d'une structure du texte (réalisée par le LIMSI et LALICC) existe. Il reste à l'étendre pour intégrer certaines marques linguistiques. Néanmoins, le système peut être aussi utilisé tel quel. Nous avons défini un format standard de représentation des textes, fondé sur la technologie XML. Ce choix a été motivé par la volonté de construire des modèles et outils réutilisables pour d'autres applications, et nous a permis de réaliser l'intégration des différents modules, tous fournissant un même format de représentation, et est issu d'un travail principalement entre le CEA et le LIMSI.

• De ce fait, les divers modules développés en tant que systèmes autonomes (ce qui facilite leur test) ont été intégrés dans la plate-forme ContextO (travail commun entre le LIMSI, le CEA et LALICC).

• La visualisation d'un texte se réalise par l'intermédiaire de deux prototypes, réalisés par LALICC et le LIMSI, qui vont permettre de préciser le modèle de navigation. Pour cela, il faudra une collaboration avec des ergonomes afin de mettre en place des protocoles d'utilisation du système et recueillir

des données sur le type de navigation mise en œuvre par chaque utilisateur.

• Dans le cadre d'une collaboration avec le projet LACO 57, dirigé par J. Pynte, une équipe composée de L. Sarda (Toulouse), J. Pynte & S. Colonna (Aix-en-Provence) M. Charolles (Paris) s'est mise en place afin d'élaborer une première expérience neuro-psycholinguistique. Les premières passations ont eu lieu en juin 2002, dans le laboratoire « Parole et Langage » d'Aix, sous la direction de J. Pynte. L'expérimentation porte sur l'incidence de la place des compléments de lieu sur le temps de lecture d'une phrase.

Ces résultats obtenus auprès de 18 sujets demandent à être confirmés sur une population plus importante mais ils sont encourageants et appellent une ou plusieurs expérimentations complémentaires. Ce travail va être conduit dans les semaines qui viennent et donnera lieu à une présentation le 22 novembre lors de la journée de synthèse du présent projet, journée ouverte à laquelle seront conviés des psycholinguistes susceptibles d'être intéressés par des expérimentations dans ce domaine.

Publications issues du projet

Michel Charolles et B. Lamiroy 2002, « Syntaxe phrastique et transphrastique : du but au résultat », in H. Nolke & H.L. Andersen eds. *Macrosyntaxe et macrosémantique*, Actes du colloque international dAarhus, 17-19 mai 2001, Bern, Peter Lang, 383-419.

Javier Couto, 2001, ContextO, *Los sistemas de exploracion contextual de cara al usuario*, Mémoire de Master, Université de la République, Uruguay.

Olivier Ferret, Brigitte Grau, Jean-Luc Minel, Sylvie Porhiel, 2001, Repérage de structures thématiques dans des textes, *TALN 2001*, p. 163-172, Tours.

Ghassan Mourad, 2001, *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des applications informatiques SegAtex et CitaRE.* Thèse de doctorat, Université Paris-Sorbonne, Paris.

Sylvie Porhiel (à paraître) « Organizing linguistic data : Thematic introducers as an example », *Coyote Papers* 12.

Géraldine Schrepfer, à par., « Sur la portée textuelle des expressions introductrices de cadres de discours en selon X : les indices de clôture des univers énonciatifs. », Actes du colloque *La médiation. Marquages en langue et en discours*, 6-8 décembre 2000, Université de Rouen.

Géraldine Schrepfer, à par., « Quelques remarques pour une analyse sémantique comparative de *selon*, *d'après* et *pour* énonciatifs », Actes du colloque *Ci-dit Le discours dans tous ces états*, Bruxelles, 8-11 2001.

Géraldine Schrepfer, *Les expressions introductrices de cadres de discours énonciatifs et leur portée textuelle : les expressions en « selon X »*, thèse de doctorat, Université de Paris III, 2002.