Un résumeur automatique de textes multilingues intégré dans une plate-forme de veille; application à la langue arabe

Abderrafih Lehmam (1), Philippe Bouvet (2)

(1) & (2) Société Pertinence Mining - Paris, France
Site Internet : http://www.pertinence.net

Résumé

Dans cet article nous présentons une solution de résumé automatique de texte multilingue. Cette solution se fonde uniquement sur des techniques linguistiques codées de façon sophistiquée en XML et en Java. Des fonctionnalités avancées augmentent la pertinence des résultats par des moyens pragmatico-linguistiques. En particulier, le moteur de résumé s'appuie à la fois sur une base de connaissances linguistiques et sur des thésaurus de divers domaines. L'association des deux bases permet d'intégrer des fonctionnalités intelligentes d'extraction de connaissances. Tous les termes descripteurs d'un domaine donné sont présentés sous forme d'une liste à partir de laquelle il est possible d'extraire les synonymes se trouvant dans le texte source ou son résumé automatique. Ces fonctionnalités nouvelles de fouille de texte permettent d'explorer des documents en naviguant à travers les termes spécifiques d'un domaine et leurs synonymes et produire automatiquement ou interactivement un résumé. Ces résultats sont exploités sur une plate-forme de veille qui analyse des sites Internet en plusieurs langues pour en extraire des informations liées à l'actualité.

Mots-clés

Pertinence, traitement linguistique de l'information textuelle, résumeur automatique, résumé automatique de texte, cohésion, multilinguisme, plate-forme de veille, fouille de texte.

Abstract

In this article, we present a solution for the automatic summarization of multilingual text. This solution is based solely on linguistic techniques with a sophisticated encoding in languages such as XML and Java. It includes advanced functionalities to make the results more relevant via pragmatic and linguistic means. In particular, to enhance relevance, the summarization engine draws upon a linguistic knowledge base and thesauri on various domains. This association of knowledge sources makes it possible to offer smart knowledge extraction functionalities; a list of all the descriptors of a specific domain is used to extract only the synonyms occurring in the source text or its automatically generated summary. Thanks to

these new text mining functionalities, the user can explore documents by navigating through domain-specific terms and their synonyms, and read an automatically or interactively generated summary. The results are exploited on a watch platform which analyzes Internet sites to extract information pertaining to current events in several languages.

Keywords

Relevance, automatic watch application, automatic text summarization, multilingual summarizer, coherence, text mining.

1 Introduction

Il est maintenant reconnu que la masse d'information véhiculée à travers les réseaux (Internet, intranets et extranets) est de plus en plus profuse et incontrôlable. La tendance est celle d'une surinformation tentaculaire et le problème de la maîtrise de l'information risque de devenir très délicat voire nocif à moins que des outils de collecte, de traitement, de diffusion ciblée et d'exploitation de l'information de façon multilingue ne viennent aider les utilisateurs à mieux gérer ces avalanches d'informations. Dans ce cadre, nous exposerons dans cet article deux applications en traitement efficient de l'information textuelle utilisant des techniques linguistiques et des algorithmes informatiques avancés. Dans un premier temps, nous allons décrire la solution professionnelle Pertinence Information Network[1] (PIN) ; une solution de veille collectant des contenus Internet/intranet en quatorze langues[2] et ayant comme particularité de résumer, dans la langue d'origine, le contenu du texte placé (par procédé de *push*) dans les boites de messagerie des utilisateurs de PIN. Dans un second temps nous présenterons l'apport de techniques linguistiques de résumé automatique ayant servi à la construction de Pertinence Summarizer. Ce dernier produit automatiquement et à la volée les résumés de chacun des documents textuels expédiés électroniquement aux utilisateurs de PIN. Les fondements théoriques linguistiques et informatiques de Pertinence Summarizer étant déjà décrits dans LEHMAM et BOUVET (2001), nous nous limiterons ici à l'aspect fonctionnel des deux applications.

2 Plate-forme multilingue d'alertes sur l'actualité : Pertinence Information Network

Pertinence information Network est une plate-forme documentaire qui peut être utilisée dans différents contextes pour différents usages (gestion de communautés virtuelles, gestionnaire de documents, gestionnaire de bookmarks, localisation de compétences, traitement des demandes client, constitution d'une base de documents pour les chercheurs et pouvant servir dans leurs domaines de recherche, plate-forme de veille en multilingue… etc.). PIN est alimentée de milliers de sources d'information multilingues liées à l'actualité mais aussi de

[1] http://www.pertinence.net/pin

[2] Allemand, anglais, arabe, chinois, coréen, espagnol, français, grec, italien, japonais, portugais, néerlandais, norvégien et russe.

nombreuses autres sources issues de différentes thématiques. Tous les documents textuels sont analysés en temps réel en vue d'expédier aux utilisateurs, ayant défini des centres d'intérêt à l'aide de mots, noms propres, expressions, citations, phrases..., une lettre d'information ou "Newsalerte" d'information (Figure1) recensant les documents pertinents. L'utilisateur sélectionne ses sources d'information en fonction de ses centres d'intérêt et de la langue parmi les 14 disponibles. Il choisit ensuite d'être alerté par email de la parution d'une information en fonction de ses mots clés. Cette alerte est entièrement paramétrable. Les documents de l'alerte peuvent être résumés automatiquement avec Pertinence Summarizer[3]. Afin de pouvoir recevoir une alerte, l'utilisateur, au préalable, définit un profil sur le site http://www.pertinence.net/pin ou http://www.pertinence.net/pin/index.jsp?ui.lang=ar où PIN est testable en ligne sur Internet, respectivement en français et en arabe mais aussi en anglais et en espagnol.

2.1 Plate-forme de veille PIN : application à la langue arabe

L'objectif de Pertinence information Network (PIN) est de mettre à la disposition des utilisateurs une technologie utilisant des ressources linguistiques savamment codées au niveau informatique pour leur permettre d'exploiter au mieux l'information pertinente disponible sur les réseaux (Internet / Intranets / Extranets) à travers une alerte envoyée par courrier électronique. Nous donnons ici un exemple (Figure 1) qui montre comment on peut établir une veille sur Internet en explorant les sources d'information arabes (quotidiens en arabe) sur les mot-clés " العراق " et" الولايات المتحدة ", respectivement « Irak » et « Etats-Unis ». L'objectif principal est de passer moins de temps à rechercher et davantage à exploiter les résultats pertinents trouvés, en vue d'une utilisation optimale et profitable. Dans cette alerte, au moyen du lien « Résumé avancé », l'article est résumé dynamiquement en tenant compte de deux mots-clés « العراق » (Irak) et « الولايات المتحدة » (Etats-Unis) pour influencer la pertinence du résultat. Toutes les occurrences de « العراق » (Irak) et « الولايات المتحدة » (Etats-Unis) sont mises en relief dans le texte-source et le résumé de l'article. Cette fonctionnalité de **fouille de texte** (*application de text mining*) permet d'accéder en un clic à la première occurrence du terme reconnu et en utilisant la touche de tabulation du clavier, on peut explorer toutes les autres occurrences et connaître ainsi les différentes utilisations en contexte des deux mots clés « العراق » et « الولايات المتحدة »

Le Résumé statique (cf. « وكان من أبرز المتحدثين بالجلسة ... الفلسطينية والعراقية رغم أهميتهم. ») est lui réalisé par Pertinence Summarizer sans tenir compte des mots-clés « العراق » et « الولايات المتحدة » et est généré automatiquement à la volée (il comporte ici trois phrases). Il est possible de définir librement le nombre de phrases à générer lors de l'envoi de la Newsalerte. PIN bénéficie de la technologie de résumé automatique de texte de Pertinence Summarizer. Le résumeur professionnel « Pertinence Summarizer » se fonde exclusivement sur l'analyse linguistique du texte afin de l'évaluer sémantiquement en vue de la production d'un résumé pertinent. Pour ce faire, il utilise un concept de marqueurs linguistiques d'extraction, LEHMAM et BOUVET (2001). En matière de résumé automatique, le problème qui se pose est la pertinence des phrases en sortie, cela ouvre un très large champ à des recherches en linguistique textuelle mais aussi à des procédés informatiques intelligents qui peuvent améliorer le résumé

[3] http://www.pertinence.net/ps

automatique en sortie en terme d'efficience. Nous présentons ci-dessous quelques-uns de ces procédés, mis en branle dans Pertinence Summarizer, appliqués à des exemples en langue arabe.

1. Bases lexicales des domaines pouvant servir à spécialiser un système de résumé pour chacun de ces domaines. L'intégration de lexiques terminologiques d'un domaine peut améliorer la pertinence du résumé en terme de thématisation. Des fonctionnalités de fouille de texte permettant ensuite l'extraction des termes descripteurs d'un domaine et de leurs seuls synonymes (Figure 2; exemple, en français, pour le domaine du droit);
2. L'utilisateur peut indiquer ses propres mots ou expressions pour influencer le résumé automatique en afin de personnaliser le résumé
3. Aide à la cohérence en temps réel du résumé généré automatiquement

Figure 1: Pertinence Information Network - Newsalerte

3 Résumé automatique de texte avancé avec extraction de termes descripteurs et des synonymes correspondants

Avant de pouvoir illustrer Pertinence Summarizer d'exemples, nous allons commencer par présenter une fonctionnalité sophistiquée et unique en matière de résumé automatique, qui permet non seulement de résumer un document textuel mais aussi de reconnaître une **corrélation entre les termes-descripteurs d'un domaine et leurs synonymes** avec la possibilité d'extraire une liste de ces termes. Le processus de résumé mis en oeuvre par Pertinence Summarizer prend en charge les ressources linguistiques liées à ce domaine à travers le thésaurus intégré au logiciel. Le résultat produit est alors en relation avec le domaine sélectionné et provoque une pertinence encore plus personnalisée à ce domaine. Après traitement, tous les termes descripteurs du thésaurus reconnus par le logiciel sont dressés sous forme d'une liste avec la possibilité de naviguer dans le texte ou le résumé sur leurs seuls

synonymes (Figure 2). Cette fonctionnalité est très utile en matière de fouille de texte car elle permet d'accélérer la lecture (résumé automatique) tout en enrichissant le vocabulaire de l'utilisateur par l'acquisition des synonymes.

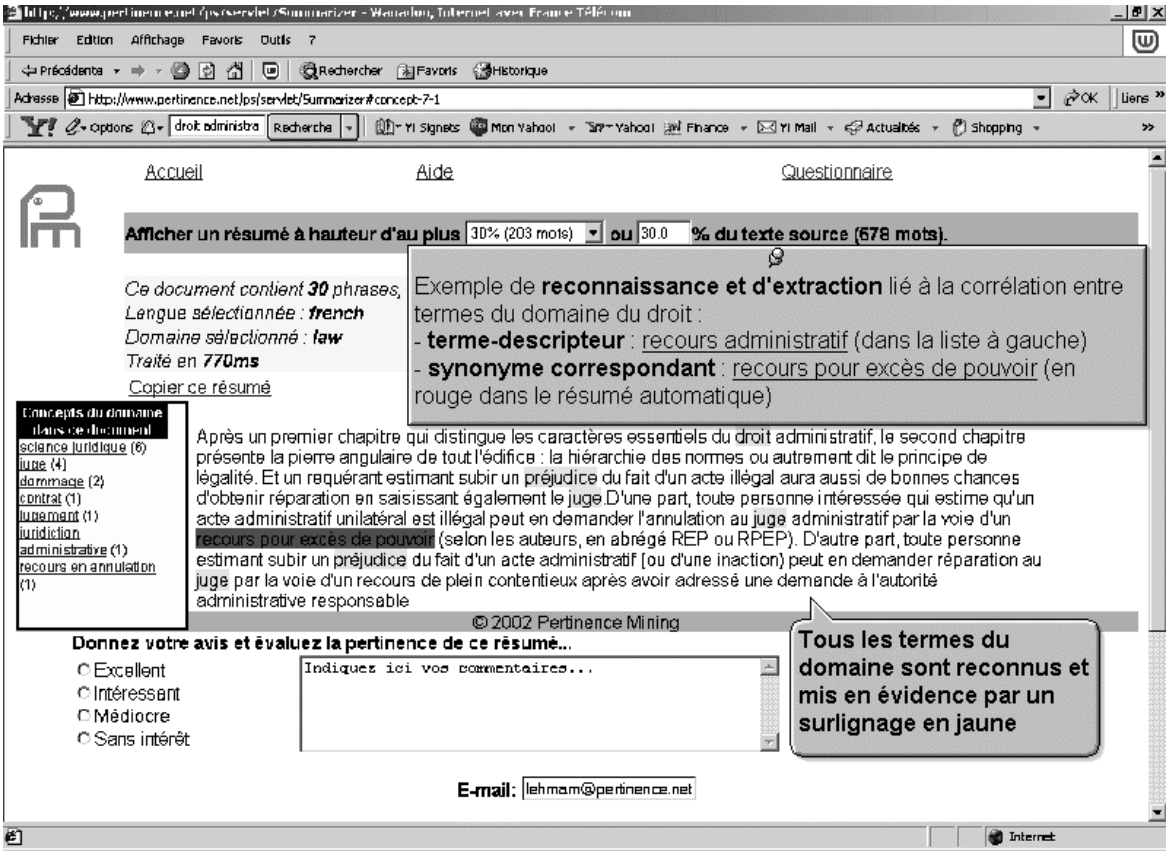

Figure 2: Résumé automatique et fouille de texte (domaine du droit)

3.1 Personnalisation du résumé automatique aux besoins de l'utilisateur et aide à la cohérence

L'utilisateur d'un système de résumé ne doit plus se suffire d'un simple résumé automatique mais doit pouvoir le plier à ses propres besoins. Les seuls systèmes de résumé pouvant être utilisés profitablement sont ceux fondés sur des méthodes linguistiques voire combinant la linguistique et la statistique, car il s'agit ici de matériel véhiculant du sens et non de simples chiffres sujets à des formules mathématiques. Les mathématiques trouvent ici leurs limites! L'utilisation de ces méthodes linguistiques combinées à la prise en charge dans le processus de résumé automatique des termes choisis par l'utilisateur donne en résultat des résumés intéressants. Que cela soit en traduction automatique ou en résumé automatique les résultats que peuvent produire les logiciels n'égaleront jamais ceux pouvant être produits par l'homme Par contre, ils lui permettent de gagner un temps précieux quand ces logiciels sont dotés de fonctionnalités prenant en compte les résultats de recherches théoriques en linguistique combinées à des procédés informatiques avancés. Dans le cas de Pertinence Summarizer, des fonctionnalités permettent de compléter en temps réel des phrases dont la référence anaphorique se trouve en amont ou en aval de la phrase extraite par le logiciel, et ceci en un simple clic de la souris. De plus, différents styles de présentation sont conçus de manière à

aider l'utilisateur à accéder efficacement à l'information pertinente tout en lui permettant ainsi, d'une manière pragmatique et de concert avec l'utilisateur, d'optimiser au mieux et interactivement la pertinence du résumé produit au niveau de la cohérence.

4 Conclusion

A travers Pertinence Information Network (PIN) et Pertinence Summarizer, les résultats que nous avons montrés, même s'ils restent les plus avancés[4] à ce jour par rapport aux systèmes commerciaux de résumé automatique, nécessitent des améliorations encore plus élaborées au niveau linguistique. Les phrases extraites sans reformulation donnent un résumé dont la cohésion n'est jamais complète mais au niveau informationnel les résultas restent cependant satisfaisants et bénéfiquement réutilisables car la méthode employée vise un effort dirigé vers l'extraction ciblée de l'information recherchée d'un texte en permettant ainsi un gain de temps significatif. Pour apporter des remèdes à ce manque, Pertinence Summarizer utilise, comme on l'a vu, des moyens pragmatiques qui permettent de retrouver instantanément le contexte de l'information extraite en utilisant un biais informatique permettant le rétablissement de la cohésion générale. On peut user de moyens issus de recherches linguistiques pour essayer de pallier à cette lacune, par exemple, en étudiant les schémas argumentatifs, DELANNOY (2001). Toutefois, la difficulté réside dans la capacité de formalisation de ces recherches en linguistique textuelle. Toute recherche en linguistique discursive, toute pertinente qu'elle soit, ne sera jamais utilisable si sa formalisation ne se prête pas à un codage informatique possible.

Références

LEHMAM A. (2004, à paraître), *Le résumé automatique des textes scientifiques et techniques, aspects linguistiques et computationnels*, Éditions de l'Harmattan, Paris.

LEHMAM A., BOUVET P. (2002), "Résumé de texte automatique : vers des solutions professionnelles", Journée d'Étude de l'ATALA "*Le résumé de texte automatique : solutions et perspectives*" organisée par le laboratoire LaLICC - FRE 2520 CNRS - Université Paris-Sorbonne et la société Pertinence Mining, Paris, ENST

LEHMAM A., BOUVET P. (2001), "Évaluation, rectification et pertinence du résumé automatique de texte pour une utilisation en réseaux Internet et Intranet " 3ème Colloque du Chapitre français de l'ISKO 2001 à l'Univ. de Paris X, Nanterre, "Filtrage et résumé automatique de l'information sur les réseaux" pp. 111-124, Paris

LEHMAM A. (1999), "Text structuration leading to an automatic summary system", *Information Processing & Management*, 35, pp. 181-191, Elsevier Ltd, New York,USA

DELANNOY J-F. (2001), "What are the points? What are the stances?", Workshop on Human language technology, Conference of the Association for Computational. Linguistics (ACL), Toulouse, France

[4] De nombreux témoignages en ce sens nous ont été envoyés par les utilisateurs de Pertinence Summarizer en ligne sur http://www.pertinence.net/ps