TerminoWeb : recherche et analyse d'information thématique

Caroline Barrière
ITI-CNR, Gatineau, Canada
caroline.barriere@nrc-cnrc.gc.ca

Résumé Notre démonstration porte sur le prototype TerminoWeb, une plateforme Web qui permet (1) la construction automatique d'un corpus thématique à partir d'une recherche de documents sur le Web, (2) l'extraction de termes du corpus, et (3) la recherche d'information définitionnelle sur ces termes en corpus. La plateforme intégrant les trois modules, elle aidera un langagier (terminologue, traducteur, rédacteur) à découvrir un nouveau domaine (thème) en facilitant la recherche et l'analyse de documents informatifs pertinents à ce domaine.

Abstract Our demonstration shows the TerminoWeb prototype, a Web platform which can (1) automatically assemble a thematic corpus from Web documents, (2) extract terms from that corpus, and (3) find definitional information in the corpus about terms of interest. As the platform integrates all three modules, it can help a language worker (terminologist, translator, writer) to explore a new domain (theme) as it facilitates the gathering and analysis of informative documents about that domain.

Mots-clés : information thématique, construction de corpus, extraction de termes, découverte de contextes définitionnels

Keywords: thematic information, corpus construction, term extraction, definitional contexts discovery

1 Introduction

Divers langagiers peuvent se retrouver en situation où ils ont besoin de maîtriser la terminologie d'un domaine auparavant peu connu. Il leur faut alors se renseigner sur le sujet, mais de façon efficace. Ils pourront lire des documents pertinents à la compréhension du domaine, et tenteront, à l'intérieur de ces documents, d'accéder rapidement à l'information importante, soit celle qui met en relief les termes importants du domaine en contextes définitionnels. Cette tâche n'est ni facile, ni précise. Tout de même, c'est à cette tâche que TerminoWeb[1] se voue pour être un outil d'aide aux langagiers. La version présente 2.0 de TerminoWeb accommode deux langues soient le français et l'anglais.

Les usagers envisagés sont multiples, soient : un terminologue devant générer un glossaire d'un domaine spécialisé, un traducteur devant comprendre les notions importantes d'un domaine dans lequel il a à traduire plusieurs documents, un spécialiste de domaine devant créer une ontologie de ce domaine, un journaliste devant maîtriser quelques concepts

[1] Le prototype TerminoWeb est disponible gratuitement pour essais à l'adresse http://terminoweb.iit.nrc.ca.

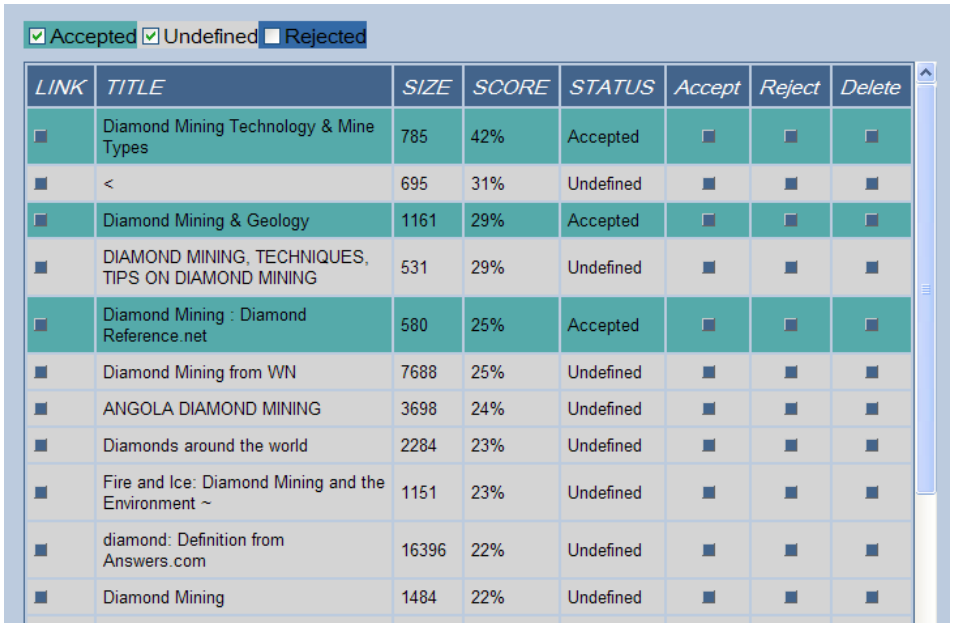
importants d'un sujet dans lequel il enquête, ou encore un enseignant de la langue spécialisée devant proposer un lexique de termes à apprendre aux étudiants pour un domaine précis.

2 Trois étapes vers l'analyse d'information thématique

Les trois modules principaux de TerminoWeb correspondent à trois étapes d'une tâche d'agglomération d'information thématique, soit (1) la construction d'un corpus spécialisé à partir de requêtes Web, (2) l'extraction de termes importants dans ce corpus, et (3) la recherche de contextes définitionnels entourant des termes choisis. Au-delà de ces modules, TerminoWeb permet aussi, entre autres : la gestion de corpora spécialisés (un usager peut conserver plusieurs corpora sur divers sujets), l'importation de textes d'une archive vers un corpus, la génération automatique de requêtes Web à partir d'un ensemble de mots-clés (voir Barrière 2009a), l'importation et exportation de listes de termes à étudier, la définition de nouveaux patrons linguistiques propres à la compréhension de domaines spécifiques (e.g. en médecine « may prevent », « may reduce the risk of »), la recherche de collocations et concordances (Barrière 2009b). Ces fonctionnalités intéressantes de TerminoWeb sont actives dans le prototype même si elles ne sont pas présentées ici par souci d'espace.

Étape 1 – *Création d'un corpus spécialisé*

Le module de création de corpus de TerminoWeb permet de recueillir des documents Web sur un sujet désiré et de les mettre en mémoire. L'usager entre une requête, par exemple *diamond mining,* et TerminoWeb lance la recherche de pages informatives et spécialisées, ces pages étant les plus susceptible de contenir des segments définitionnels dans le domaine d'intérêt.



LINK	TITLE	SIZE	SCORE	STATUS	Accept	Reject	Delete
■	Diamond Mining Technology & Mine Types	785	42%	Accepted	■	■	■
■	<	695	31%	Undefined	■	■	■
■	Diamond Mining & Geology	1161	29%	Accepted	■	■	■
■	DIAMOND MINING, TECHNIQUES, TIPS ON DIAMOND MINING	531	29%	Undefined	■	■	■
■	Diamond Mining : Diamond Reference.net	580	25%	Accepted	■	■	■
■	Diamond Mining from WN	7688	25%	Undefined	■	■	■
■	ANGOLA DIAMOND MINING	3698	24%	Undefined	■	■	■
■	Diamonds around the world	2284	23%	Undefined	■	■	■
■	Fire and Ice: Diamond Mining and the Environment ~	1151	23%	Undefined	■	■	■
■	diamond: Definition from Answers.com	16396	22%	Undefined	■	■	■
■	Diamond Mining	1484	22%	Undefined	■	■	■

Figure 1 – Ensemble de pages résultantes de la recherche « diamond mining »

Dans nos travaux précédents (Barrière et Agbago 2006), nous avons établi, de façon empirique, divers critères pour évaluer si une page est informative ou non. Le critère principal repose sur la présence de patrons de connaissances (Meyer 2001) qui sont des patrons linguistiques exprimant diverses relations sémantiques liées à la définition, telles l'hyperonymie (« est une sorte de », « est un type de »), la synonymie (« aussi appelé », « aussi connu sous le nom ») ou la méronymie (« est une partie de», « est composé de »). Une page

doit de plus être spécialisée et le critère pour cet aspect est la densité de termes connus du domaine s'y retrouvant. Ainsi, TerminoWeb lancera une requête au Yahoo API[2] et analysera X pages (e.g. 50 pages) pour leur assigner une note en fonction des critères établis et retournera les N pages (e.g. 10 pages) les plus informatives. Ces paramètres X (nombre de pages à analyser) et N (nombre de pages désirées) ont des valeurs par défaut qui peuvent être modifiées par l'usager.

La Figure 1 montre le résultat de la requête *diamond mining* tel que vu dans TerminoWeb. L'usager peut retrouver chaque document (en cliquant sur le lien dans la colonne Link) et décider d'accepter ou non le document dans le corpus (changement du statut de « Undefined » à « Accepted »). La colonne [score] indique la note obtenue par le document et la colonne [size] indique la taille du document en nombre de caractères.

Étape 2 *– Extraction de termes*

TerminoWeb permet l'extraction de termes en utilisant un algorithme inspiré de Smadja (1993). L'extraction peut se faire sur différents sous-ensembles de documents d'un corpus établi (le sous-ensemble de documents de statut « Accepted » par exemple). Dans l'exemple de l'exploitation du diamant, émergeront les termes: alluvial diamond (12), artisanal diamond mining (4), deposits (80), diamond deposits (20), diamond mine (67), extraction (14), gem (168), gravel (29), kimberlite (65), lamproite (13), mining technology (5), open-cast mining (4), open-pit (9), ore (427), placer diamond deposits (2), terrace gravels (3). La fréquence de chaque terme mise entre parenthèses est calculée sur le corpus entier.

Étape 3 *– Recherche de contextes définitionnels*

TERM	RELATION	CONTEXT
kimberlites	SYNONYMY	igneous rocks known as kimberlites and lamproites . Diamo
kimberlites	SYNONYMY	volcanic pipes...known as kimberlites, and kimberlite pipes.
kimberlite pipes	SYNONYMY	ical structures known as kimberlite pipes (above, right). Kimberli
kimberlite	SYNONYMY	igneous rock known as kimberlite . The diamond itself is
kimberlite	SYNONYMY	first type is known as kimberlite , and the second as lamp
kimberlite	SYNONYMY	ous rock known as either kimberlite or lamproite. [33] Th
kimberlite	SYNONYMY	enite , and magnetite . Kimberlite deposits are known as b
kimberlite	SYNONYMY	ical structures known as kimberlite pipes (above, right). Ki
kimberlite	HYPERONYMY	volcanic action carried kimberlite and other minerals from
kimberlites	DEFINITION	identify diamondiferous kimberlites. 70% interest in Aviat d
kimberlite pipes	CAUSE	m-quality diamonds. Many kimberlite pipes also produce alluvial
kimberlite	CAUSE	....Below-ground mining of kimberlite for diamond also requires
kimberlite	CAUSE	m-quality diamonds. Many kimberlite pipes also produce allu

Figure 2 – contextes définitionnels pour les termes *kimberlite(s)* et *kimberlite pipes*

TerminoWeb recherche des contextes potentiellement définitionnels contenant, en proximité des termes à l'étude, des patrons de connaissances préalablement définis et encodés dans le logiciel. Nous disons « potentiellement définitionnels » pour deux raisons : (1) les patrons

[2] Yahoo! fournit une interface de programmation Java pour l'utilisation de l'engin "Yahoo Search Engine" à des fins de recherche. Cette interface est appelée à partir du logiciel TerminoWeb.

étant parfois bruités, leur présence ne sera pas toujours indicative de la relation désirée et (2) la proximité entre un terme et un patron n'assure pas d'une relation entre eux. Malgré le bruit, nous préférons tout de même la simplicité de la proximité à la complexité d'une analyse syntaxique qui rendrait le traitement dépendant de la langue (et non sans bruit lui-même).

La figure 2 montre des contextes pour le thème de l'exploitation du diamant dans lequel les termes *kimberlite(s)* et *kimberlite pipes* seraient nouveaux pour le langagier. Chaque contexte peut être élargi (Figure 3), et un lien permet de retourner à la page Web d'où vient la phrase.

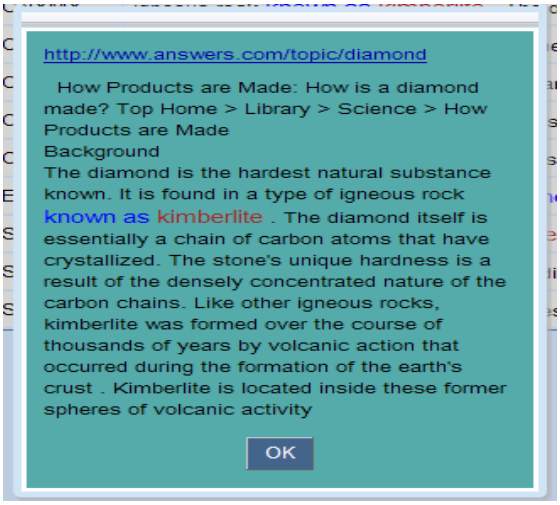


Figure 3 – Exemple d'un contexte définitionnel plus étendu

Conclusion

Les trois modules principaux de la plateforme TerminoWeb ont été brièvement décrits. Cette plateforme permet non seulement de chercher sur le Web des documents informatifs sur un thème à explorer, mais aussi de cibler les segments définitionnels dans ces documents pour permettre au langagier d'accéder plus rapidement à l'information pertinente et ainsi favoriser sa compréhension du domaine.

Références

BARRIÈRE, C., AGBAGO, A. (2006). TerminoWeb: a software environment for term study in rich contexts. *Proceedings of the International Conference on Terminology, Standardization and Technology Transfer*, Beijing, 103-113.

BARRIÈRE, C. (2009a). The Web as a source of informative background knowledge, *Proccedings of the Workshop "Beyond Translation Memories: New Tools for Translators", MT Summit 2009*, Ottawa, Canada.

BARRIÈRE, C. (2009b). Finding domain specific collocations and concordances on the Web, *Proceedings of the Workshop "Natural Language Processing methods and corpora in translation, lexicography and language learning", RANLP'2009*, Borovets, Bulgaria.

MEYER, I. (2001). Extracting knowledge-rich contexts for terminography, in D. Bourigault, C. Jacquemin, L'Homme M.C. (eds) *Recent Advances in Computational Terminology*, chapter 14, John Benjamins.

SMADJA F. (1993) Retrieving collocations from text: Xtract. *Computational Linguistics* 7(4): 143-177.