Classification automatique de documents issus du Web selon leur type de discours

Lorraine Goeuriot[1]

[1] LINA (CNRS FRE 2729) – Université de Nantes – France

Abstract

This paper summarizes researches on Web page automatic classification, according to the discourse. We first apply a stylistic analysis to our corpus, based on a contrastive analysis of medical domain documents, matching with scientific and popular science discourses. This analysis helps us to produce a typology for these two discourses. Finally, we apply learning algorithms to our learning corpus.

Résumé

Dans cet article, nous cherchons à classer automatiquement des documents issus du Web selon leur type de discours. Pour cela, nous effectuerons dans un premier temps une analyse stylistique sur notre corpus d'apprentissage, basée sur une étude contrastive de documents émanant des discours scientifiques et vulgarisés. Cette analyse donnera lieu à la création d'une typologie de ces deux types de discours. Des algorithmes d'apprentissage automatique seront ensuite appliqués à cette typologie afin de distinguer automatiquement les documents scientifiques des documents vulgarisés.

Mots-clés : analyse stylistique, classification de documents, type de discours

1. Introduction

L'objectif de notre travail est de classer automatiquement des documents issus du Web selon leur type de discours. Nous nous basons ici sur la notion de discours telle qu'elle est définie par (Ducrot et Schaeffer, 1999) : "Tout ensemble d'énoncés d'un énonciateur caractérisé par une unité globale de thème". Ce travail a été mené en trois phases : la construction d'un corpus d'apprentissage, l'analyse du corpus menant à la création d'une typologie et la mise en œuvre de cette typologie visant à apprendre un modèle de classification automatique.

2. Construction du corpus d'apprentissage

Le corpus exploité dans ce travail a été manuellement construit à partir de documents disponibles sur le Web. Notre objectif étant de disposer d'un corpus composé de documents relevant des deux discours scientifique et vulgarisation scientifique, il nous a paru plus cohérent que tous les documents partagent un même thème, mettant ainsi en évidence le contraste entre ces deux types de discours. Le domaine médical nous a paru dans un premier temps être le plus adapté à nos besoins, et très bien représenté sur le Web. Ne pouvant nous contenter d'un domaine aussi vaste, nous l'avons ensuite restreint à la thématique « Diabète et alimentation ». Ainsi, nous disposons d'un thème touchant un large public, nous garantissant une diversité dans les documents récoltés.

La démarche de constitution du corpus repose sur la recherche de pages Web correspondant à la thématique visée, puis sur la sélection des documents pertinents et enfin sur le classement de ces documents selon leur type de discours.

Trois ressources ont été utilisées lors de la construction : les moteurs de recherche, les portails médicaux et les pages de liens.

Les pages ont ensuite été conservées selon leur pertinence, puis classées. N'ayant pas pu faire appel à des spécialistes du domaine, nous sommes partis du principe suivant : un document scientifique est rédigé par des spécialistes, à destination de spécialistes. En ce qui concerne la vulgarisation scientifique, nous distinguons différents « degrés » de vulgarisation : les textes écrits par « le grand public », à destination de tous, et les textes écrits par des spécialistes à destination du « grand public ». Ces deux catégories ont été représentées dans le corpus, mais nous avons laissé une plus grande place aux documents écrits par des spécialistes, plus longs et plus riches en vocabulaire, au détriment des discussions de forums par exemple. La classification manuelle est donc basée sur ces définitions, recouvrant un certain nombre d'éléments dans les documents : le site contenant le texte, le vocabulaire utilisé, etc. Cependant, il est vrai que cette tâche de classification manuelle reste assez empirique, nous avons donc décidé de mettre de côté certains documents « ambigus » : les documents inclassables ou sur lesquels les avis divergeaient.
Le tableau suivant présente les principales caractéristiques du corpus, c'est-à-dire le nombre de documents et de mots pour les parties vulgarisée et scientifique.

	Nombre de documents	**Nombre de mots**
Partie scientifique	65	425 781
Partie vulgarisée	183	267 885
Total	248	693 666

3. Analyse du corpus

L'analyse du corpus vise à déterminer quels sont les facteurs et les caractéristiques des variations observables entre des textes émanant de différents discours. De ce type de travail d'analyse émerge une typologie de textes, ou de corpus, c'est-à-dire une classification et description des textes s'appuyant sur leurs caractéristiques internes. Notre démarche est déductive (Habert, 2000): en se basant sur un corpus dont les éléments sont préalablement classés, notre objectif est de créer une typologie des textes du corpus permettant de caractériser l'appartenance d'un texte à une des classes du corpus. Cette typologie est le fruit d'une *analyse contrastive* des éléments du corpus.
Contrairement à un grand nombre de travaux traitant de la classification de textes (Malrieu et Rastier, 2002 ; Biber, 1989), nos documents, puisqu'ils sont extraits du Web, présentent une structure propre dont nous ne pouvons négliger le contenu. Nous allons utiliser le texte ainsi que la structure du document afin de dégager une typologie contenant l'ensemble des éléments caractérisant leur type de discours.
Cette analyse stylistique, appelée analyse stylistique, consiste donc à dégager de l'observation des documents de notre corpus une série de caractéristiques distinctes susceptibles d'indiquer, selon leurs variations, l'appartenance d'un document au discours scientifique ou vulgarisé.
Nous distinguerons deux niveaux d'analyse des documents, inspirés des dimensions de (Sinclair, 1996) :
– Le niveau externe (dimension externe du document), concernant toutes les caractéristiques propres au contexte de création du document ;
– Le niveau interne (dimension interne du document), concernant les caractéristiques propres aux tâches communicatives du site et de son document. Ce niveau contient 3 catégories : les caractéristiques graphiques (apparence générale du site et de ses documents), structurelles (apparence du texte principal du document) et sémantico-discursives (caractéristiques stylistiques et linguistiques du document).
Les critères dégagés lors de cette analyse sont présentés dans la partie suivante sous la forme d'une liste non exhaustive de critères caractéristiques des documents scientifiques et

vulgarisés (en se guidant grâce aux travaux de (Biber, 1989 ; Karlgren, 1998)). Seules des observations sur un échantillon du corpus ont été faites, leur degré de discrimination ne sera évalué qu'une fois la reconnaissance de œs critères implémentée. Cette liste sera ensuite filtrée, afin de ne conserver que les critères opératoires.

3.1. Dimension externe

Cette dimension est composée de l'ensemble des critères impliqués dans le contexte de la création du site. Ces critères concernent en majeure partie le site Web contenant le document étudié.

Critères	Observations
Nom de domaine	Observation de patrons d'URL
Format de document	
Architecture du site	Structure interne d'un site Web (hiérarchie établie entre les différentes pages)
Taille du site	Taille de la globalité des éléments du site
Méta-informations	Balise `<META>`
Titre du document	Balise `<TITLE>`

3.2. Dimension interne (DI)

3.2.1. DI - caractéristiques graphiques

Critères	Observations
Techniques de mise en page	Cadres, tableaux, feuilles de style
Fonds	Couleur et/ou image de fond
Images	Photos, schémas, logos, etc.
Publicités	Liens publicitaires, avec ou sans image

3.2.2. DI - caractéristiques structurelles

Critères	Observations
Titre	Titre du texte
Mots clés, résumé	Informations concernant le contenu du texte (appartenant à une structure semblable à celle des articles de recherche)
Plan	Liste de liens ou ancres indiquant le plan du texte à suivre
Introduction / conclusion	
Bibliographie	Présence d'une liste d'ouvrages de référence
Apparence du texte principal	Paragraphes, typographie, longueur des phrases, listes, liens, citations, tableaux
Auteurs	Nom du(es) auteur(s)

3.4.1. DI - caractéristiques structurelles

Catégories	Critères	Observations
Caractéristiques pragmatiques	Dispositif énonciatif	Position du locuteur dans le discours (Charaudeau, 1992) ; éléments de l'analyse de la modalité (allocutive, élocutive, délocutive)
	Connecteurs pragmatiques	«mots qui ne sont pas destinés à apporter des informations, mais à marquer le rapport du locuteur à la situation » (Ducrot, 1980) (*mais, donc, alors que…*)
	Marqueurs de glose	Explication d 'une idée apr une autre, à l'aide de marqueurs (*à savoir, c'est-à-dire, par exemple…*)
Caractéristiques phrastiques	Type de phrase	Interrogatif, affirmatif, exclamatif
	Ponctuation	
	Pronoms	Présence de pronoms appartenant aux catégories suivantes : pronoms démonstratifs, possessifs, relatifs…
Caractéristiques lexicales	Vocabulaire spécialisé	Termes scientifiques et techniques
	Collocations textuelles	Cooccurrences privilégiées de mots pleins dans un contexte textuel
	Noms propres, acronymes	
	Mots inconnus	Néologismes ou fautes d'orthographe
	Caractères numériques	
	Symboles	
	Unités de mesure	
	Longueur des mots	

4. Mise en œuvre des critères

L'analyse stylistique du corpus nous a permis de créer une typologie caractérisant les discours scientifiques et vulgarisés, composée d'un ensemble de traits. Un système d'apprentissage automatique se basant sur la représentation des documents sous la forme d'une liste de caractéristiques, nous avons dû les implémenter. Avant cela, la liste des critères a dû être filtrée, afin de ne conserver que les critères opératoires.

4.1. Filtrage des critères

Un critère est jugé opératoire si son implémentation est possible et ne nécessite aucun traitement coûteux. L'un des objectifs fixés ici est de réaliser ce système de classification en n'utilisant que des critères simples et rapides, privilégiant le nombre de critères plutôt que la précision de chacun d'entre eux. Ainsi, parmi les critères non-opératoires se trouvent :

- L'architecture du site : l'intégralité du site doit être observée (il est parfois impossible de déterminer où se situe la racine d'un site) ;

- Le titre, l'introduction et la conclusion d'un texte : pas systématiquement présents, aucun standard de position, typographie, etc. (Hu et al. 2005).

4.2. Implémentation des critères

La plupart de ces critères consistent à effectuer une recherche de balise, chaîne de caractère

ou patron dans les fichiers. C'est pourquoi nous ne détaillerons ici que l'implémentation de certains critères de la catégorie sémantico-discursive. Les exemples dans un texte sont recherchés grâce à une liste prédéfinie de termes ou expressions utilisés lorsqu'on présente un exemple ("ex.", "notamment", "en particulier", etc.). C'est cette même technique qui est utilisée afin de compter les connecteurs pragmatiques ("donc", "alors que", "néanmoins", etc.), mais aussi les pronoms, reformulations et unités de mesure. L'analyse de la modalité (du dispositif énonciatif) a simplement consisté à observer les pronoms contenus dans les textes :

– Pronoms allocutifs : tu, vous ;
– Pronoms élocutifs : je, nous ;
– Pronoms délocutifs : il, elle...

Pour mesurer la densité terminologique, à l'instar de (Namer, 2005), nous avons évalué la présence de termes à celle de racines gréco-latines. Ainsi, nous avons créé une liste de préfixes et suffixes gréco-latins tirés du travail de (Béchade, 1992), que nous utilisons pour parcourir l'ensemble des textes de notre corpus.

Une fois l'ensemble de ces critères implémenté, nous obtenons une représentation sous forme de vecteurs de l'ensemble des documents, pouvant être utilisée pour appliquer au corpus des algorithmes d'apprentissage automatique.

4.3. Résultats obtenus

Selon (Sebastiani, 2002), les algorithmes les plus efficaces pour un apprentissage supervisé avec deux classes sur un corpus de taille limitée sont les machines à vecteurs de support, les arbres de décision ainsi que les classifieurs de Bayes. Nous avons choisi de nous limiter dans un premier temps aux deux premières techniques, en appliquant *SVMlight* (Joachims, 2002) et *C4.5* (Quinlan, 1993). La taille de notre corpus étant assez limitée, nous avons dû utiliser la méthode dite *par validation croisée* (Cornuéjols et Miclet, 2002), permettant d'apprendre sur une partie du corpus et de tester sur l'autre. Les résultats obtenus en moyenne sur les différents croisements sont présentés dans le tableau suivant.

		Précision	Rappel
SVMlight	SC	0,73	0,48
	VG	0,66	0,82
C4.5	SC	0,98	0,95
	VG	0,95	0,98

Les résultats obtenus grâce à ces classifieurs sont satisfaisants. En effet, C4.5 nous permet d'obtenir une moyenne de 96% de documents correctement classés. L'analyse stylistique, telle qu'elle est présentée ici, consiste à étudier les variations de style au sein d'un texte. Les résultats obtenus confirment la définition de (Karlgren, 1998) : le style correspond aux variations entre les différentes façons d'exprimer une même idée, variations relevant des choix de l'auteur lors de la rédaction. Ces choix concernent évidemment le texte en lui-même : au niveau syntaxique, lexical, sémantique, etc. ; mais aussi, dans le cadre de l'analyse des types de discours sur le Web les niveaux graphiques et structurels. Ces résultats obtenus montrent donc que les variations de l'ensemble des éléments de la typologie des discours élaborée ici permettent de caractériser les discours scientifiques et vulgarisés sur le Web.

5. Conclusion

En partant d'un corpus d'apprentissage composé de documents issus du Web préalablement classés selon leur type de discours, et en adoptant une démarche déductive, notre analyse stylistique a permis de créer une typologie des discours scientifique et vulgarisé dans le domaine médical, en utilisant des informations indépendantes de ce domaine. Les études effectuées sur l'analyse stylistique nous ont aussi éclairés sur les différentes méthodologies de création de typologies, et conforté dans l'idée qu'une typologie des discours du Web doit être le fruit d'une analyse contrastive des éléments d'un corpus préclassés.

Nous souhaitons étendre ce travail à d'autres langues, en particulier la langue japonaise sur laquelle nous travaillons actuellement. Ce travail pourrait enfin être appliqué à d'autres domaines scientifiques (l'écologie par exemple), touchant un large public et nous fournissant les mêmes types de discours.

Références

Biber D. (1989). A typology of English texts. *Linguistics*, vol. 27 :3-43.

Béchade H.-D. (1992). *Phonétique et morphologie du français moderne et contemporain*. Presses Universitaires de France.

Charaudeau P. (1992). *Grammaire du sens et de l'expression.* Hachette.

Cornuéjols A. and Miclet L. (2002). *Apprentissage artificiel : concepts et algorithmes.* Eyrolles.

Ducrot O. (1980). *Les mots du discours*. Minuit.

Ducrot O. and Schaeffer J.-M. (1999). *Nouveau dictionnaire encyclopédique des sciences du langage.* Seuil.

Habert B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? In Bilger, M. editor *Linguistique sur corpus : Etudes et réflexions*, vol. 31 :11-58.

Hu Y., Xin G., Song R., Hu G., Shi S., Cao Y. and Li H. (2005). Title extraction from bodies of html documents an dits application to web page retrieval. In Baezayates A., Ziviani N., Marchionini G., Moffat A. and Tait J. editors, *Proceedings of the 28th annual ACM SIGIR conference on research and development in Information Retrieval,* 523-548.

Joachims T. (2002). *Learning to classify text using Support Vector Machines.* Kluwer Academic Publishers.

Karlgren J. (1998). *Natural Language Information Retrieval.* Chapter Stylistic Experiments in Information Retrieval. Tomek, Kluwer.

Malrieu D. and Rastier F. (2002). Genres et variations morphosyntaxiques. In *Traitement Automatique des Langues (TAL).* Vol. 42 : 548-577.

Namer F. (2005). Morphosémantique pour l'appariement de termes dans le domaine médical : approche multilingue. In *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles (TALN).* 63-72.

Quinlan J. R. (1993). *C4.5 : Programs for machine learning.* Morgan Kaufmann Publishers.

Sebastiani F. (2002). Machine learning in automated text categorization. In *ACM computing surveys.*1-47

Sinclair J. (1996). Preliminary recommendations on text typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).