Classification de textes en domaines et en genres en combinant morphosyntaxe et lexique

Cleuziou, Guillaume [(1)], Poudat, Céline [(2)]

(1) LIFO – Université d'Orléans, B.P. 6759 F-45067 Orléans Cedex 2
guillaume.cleuziou@univ-orleans.fr
(2) ERTIM – INALCO, 2, rue de Lille, 75343 Paris Cedex 07
cpoudat@gmail.com

Résumé Nous présentons dans cet article le bilan de notre participation au 4e DÉfi Fouille de Textes 2008. L'étude porte sur la problématique de la classification textuelle en domaines et en genres qui représente un enjeu pour la Recherche d'Information (RI). Sa mise en œuvre nécessite notamment la sélection d'un ensemble de descripteurs adéquats. On considère généralement que les domaines sont corrélés au niveau du contenu (mots, termes, etc.) tandis que les genres sont discriminés au niveau morphosyntaxique. Malgré les bons résultats obtenus par ces choix méthodologiques, peu de travaux ont cherché à mesurer l'impact et la complémentarité des deux niveaux de description pour la classification. Le cadre pratique de ce défi permettra de compléter les premiers postulats formulés sur ce travail.

Abstract In this paper we present our contribution to the 4th *DÉfi Fouille de Textes* 2008. The challenge deals with topic and genre texts categorization which is of real interest for Information Retrieval researches. This task requires notably to select appropriate descriptors . In most categorization works, topics (or domains) are generally correlated to the content level (words, terms, bag of words, etc.) and genres to the morphosyntactic one. However, few studies have assessed the impact and the complementarity of the two description levels on genre and domain categorization. The practical framework of the DEFT challenge allows to complement the very first postulates we expressed on this research topic.

Mots-clés : Recherche d'Information, genre, domaine, classification, lexique, morphosyntaxe.

Keywords: Information Retrieval, genre, domain, classification, lexicon, morphosyntax.

1 Introduction

Les classifications textuelles en domaines et en genres représentent un enjeu pour la Recherche d'Information (RI), et leur mise en œuvre nécessite la sélection d'un ensemble de descripteurs adéquats. Dans les faits, domaines et genres sont généralement associés à des niveaux linguistiques différents. Quand il s'agit de classification thématique ou domaniale, les textes sont souvent réduits à l'état de "sacs de mots". Chaque document est alors décrit par le vocabulaire présent dans le corpus. Étant donné la taille de ce vocabulaire, une étape de réduction de l'espace de description est indispensable : sélection d'attributs par des mesures d'intérêt (mesure d'Information Mutuelle, Gain d'Information et mesure du chi2, etc.), reparamétrage de l'espace (LSI, pLSI) ou regroupement d'attributs. Ces formalismes d'indexation permettent d'obtenir des classifieurs performants, atteignant jusqu'à 90% de précision sur grands corpus (Hofmann, 1999, Dhillon et al., 2003). De la même manière, les classifications en genres à partir d'un jeu de variables morphosyntaxiques robuste sont à même d'obtenir de très bons résultats en matières de validation de typologies textuelles (Karlgren et Cutting, 1994, Kessler et al., 1997, Malrieu et Rastier, 2001).

Nous avons toutefois pu constater que la plupart des travaux recensés effectuent de la classification domaniale sur corpus génériquement homogènes (e.g. Reuters ou Newsgroup), et de la classification générique sur corpus discursivement[1] hétérogènes, ce qui augmente le pouvoir classificatoire des variables employées mais limite l'utilisation conjointe et l'évaluation de la portée des deux niveaux descriptifs. DEFT'08 propose ainsi un cadre et un corpus de travail audacieux qui cherche à dépasser cette opposition, bien que les deux « genres » contrastés relèvent de deux discours différents : le discours encyclopédique et le discours journalistique.

Nos travaux précédents (Poudat, Cleuziou, 2003, Poudat et al., 2006 et Cleuziou, Poudat, 2007) se sont attachés à la classification textuelle en domaines et genres au sein du discours scientifique. Dans cette perspective, nous avions utilisé et combiné deux types de traits : (i) des descripteurs lexicaux simples (substantifs les plus représentés) et (ii) un système de catégorisation morphosyntaxique adapté aux caractéristiques les plus saillantes des textes scientifiques (abréviations, connecteurs, modaux, indices de structuration des textes, etc.). Ce dernier jeu de descripteurs s'était avéré particulièrement discriminant lors des tâches de classification domaniale. Nous avions de plus eu recours à deux techniques d'apprentissage supervisée que sont les séparateurs à vaste marge (SVM) et les arbres de décision. La recherche d'une précision maximale justifie le choix de l'approche SVM puisque cette technique est actuellement la plus performante pour la tâche considérée (Dumais *et al.*, 1998); afin de mieux appréhender l'articulation des deux types de traits nous avions en parallèle étudié le résultat d'une méthode de classification de type arbre de décision qui présente l'avantage de fournir une explication du modèle appris.

Bien qu'il diffère substantiellement de nos corpus d'étude, nous avons tenté dans la mesure du possible d'adapter la méthodologie développée au corpus DEFT'08, à plusieurs exceptions près : (i) les délais du défi ne nous permettant malheureusement pas de réadapter de manière pertinente le système de descripteurs développé, nous avons recouru au système d'étiquetage

[1] Discours littéraire, juridique, scientifique, journalistique, etc. Les types de discours sont reliés à des pratiques sociales distinctes et organisent en leur sein les typologies génériques et domaniales. Le discours juridique inclut ainsi les genres de l'arrêt, du décret, de la loi, etc.

plus général du TreeTagger[2]. Nous proposons néanmoins une expérimentation additionnelle dans le présent article en utilisant le jeu de descripteurs précédent, et en comparant les résultats à ceux obtenus avec le TreeTagger ; (ii) lors de nos expériences précédentes, nous avions recouru aux substantifs les plus représentés, étant donné leur statut potentiel de *concepts* dans les textes, et par conséquent d'objets interprétables. Nous avions en effet mis l'accent sur l'intérêt d'une description raisonnée et interprétable des regroupements obtenus dans le processus de classification. L'objectif de DEFT'08 étant tout autre, et visant d'abord à obtenir le meilleur pourcentage de classification possible, ce sont les classes qui se sont avérées les plus efficaces que nous avons retenues, après quelques tests; (iii) dans cette même optique de recherche de performance nous avons enfin été naturellement conduit à privilégier un classifieur de type SVM.

Voici les différentes étapes de la méthodologie utilisée :

1. Etiquetage du corpus avec TreeTagger
2. Construction des dictionnaires de classification
    - extraction des descripteurs lexicaux (parties pleines du discours)
    - extraction des descripteurs morphosyntaxiques (construction de classes)
3. Apprentissage d'un modèle de séparation des classes au moyen d'un SVM linéaire

Le présent article reprend les différentes étapes de ce processus : après avoir documenté les descripteurs mobilisés par la tâche d'apprentissage (section 2), nous décrirons la méthode d'apprentissage SVM utilisée (section 3). La section 4 reprend les résultats obtenus, et nous présentons enfin les conclusions de notre expérience dans la section 5.

2 Descripteurs mobilisés

Nous avons mobilisé deux types de descripteurs : deux ensembles de descripteurs morphosyntaxiques, et un ensemble de descripteurs lexicaux simples. Ces ensembles de descripteurs que nous décrivons dans la suite seront fusionnés (additionnés) afin de constituer un ensemble plus grand de descripteurs utilisé pour la classification.

2.1 Descripteurs morphosyntaxiques – système d'étiquetage TreeTagger

Le corpus a d'abord été étiqueté à l'aide du TreeTagger, avec le jeu d'étiquettes morphosyntaxiques proposé par Achim Stein pour le français[3]. Il s'agit d'un système d'étiquetage robuste (33 catégories), qui permet au TreeTagger français d'atteindre un score de précision élevé (environ 92%). Cette robustesse entraîne naturellement des choix d'étiquetage discutables sur le plan linguistique : ainsi, les auxiliaires ne sont pas reconnus et la catégorie 'pronoms démonstratifs' inclut également les déterminants démonstratifs. A

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[3] http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html

fortiori, on sait que la ponctuation est discriminante en matière de classification, mais TreeTagger n'identifie précisément que les guillemets, les autres ponctuations étant indifféremment labellisées PUN. Malgré ces réserves, c'est à partir de cette annotation que nous avons construit nos dictionnaires d'apprentissage.

28 descripteurs ont au final été conservés (tableau 1) :

Distribution générale des lemmes	*Distribution des verbes*	*Distribution des pronoms*
% de SYM ou ABR dans le document	% de verbes VER:cond parmi les verbes du document	% de pronoms de type PRO (simple) parmi l'ensemble des pronoms du document
% de ADV ou KON	% de VER:futu	% de PRO:DEM
% de ADJ	% de VER:impe	% de PRO:IND
% de PRO	% de VER:impf	% de PRO:PER
% de VER	% de VER:infi	% de PRO:POS
% de DET	% de VER:pper	% de PRO:REL
% de NAM ou NOM	% de VER:ppre	
% de PRP	% de VER:pres	
% de PUN	% de VER:simp	
% de INT	% de VER:subi	
% de NUM	% de VER:subp	

Tableau 1 : Catégories morphosyntaxiques conservées

On obtient donc pour chaque document un ensemble de 28 descripteurs morphosyntaxiques : les 10 premiers traits rendent compte de la distribution des étiquettes rencontrées dans un document sur les 10 classes (1ère colonne du tableau 1); les 10 traits suivants précisent cette fois la distribution au sein de la classe d'étiquettes correspondant aux verbes; enfin les traits suivants précisent la classe des pronoms.

2.2 Descripteurs lexicaux

En ce qui concerne les descripteurs lexicaux, nous avons étudiés plusieurs sélections et plusieurs indexations afin de retenir la solution la plus performante en terme de classification par SVM sur les corpus d'entraînement (évaluations par validations croisées sur un sous ensemble de 1000 documents). La sélection des étiquettes NOM, ADJ, VER, NAM, ABR, SYM, INT, PRO et ADV s'est avérée être la plus performante aussi bien pour discriminer les domaines que pour discriminer les genres. Nous avons alors retenus tous les lemmes correspondant à ces étiquettes sans sélection sur leur nombre d'occurrences dans le corpus d'entraînement.

Concernant l'indexation nous avons comparé une indexation de type fréquence (où chaque document est représenté par le vecteur des fréquences des mots du vocabulaire apparaissant dans le document) avec une indexation pondérée par le tfxidf. De manière assez inattendue nous avons observé une influence négligeable du mode d'indexation, avec de plus un avantage

pour le mode le plus simple (fréquences) lorsqu'une différence de performance était observée. Ce dernier mode d'indexation a donc été retenu pour nos expérimentations finales.

Par la méthode retenue, 31389 descripteurs lexicaux ont alors été retenus pour la tâche 1 et 36935 pour la tâche 2.

2.3 Descripteurs morphosyntaxiques – système d'étiquetage spécifique

Le système d'annotation additionnel que nous avons mobilisé[4] comprend 129 étiquettes[5] au total. Bien qu'il soit originellement adapté aux spécificités du discours scientifique, il prend en compte les recommandations EAGLES, et résulte de l'examen attentif de plusieurs systèmes d'étiquetage (TreeTagger, WinBrill pour le français, développé par l'Inalf, Cordial, etc.). En ce sens, il est plus systématisé et plus complet que celui de TreeTagger et sa granularité est plus élevée.

Deux types d'observations sont ainsi fédérés : un ensemble de catégories morphosyntaxiques générales, ou « de langue », incluant les grandes parties du discours et leurs attributs traditionnels (nombre, temps et modes verbaux, etc.), un ensemble de variables spécifiques et supposées caractéristiques du discours scientifique (distinction des IL anaphorique/impersonnel, des connecteurs généralement étiquetés comme adverbes, annotation des indices de structuration de type *1.1.2.*, des éléments de langue étrangère, symboles etc.). Le système employé inclut donc différents niveaux d'observation linguistique, dans la mesure il combine des variables morphosyntaxiques et sémantiques.

3 Classification par SVM

Les SVM, en anglais *Support Vector Machine* (Machines à Points de Support ou encore Séparateurs à Vaste Marge) sont reconnus pour leurs performances inégalées dans l'application à la classification de textes (Dumais *et al.*, 1998). De manière simplifiée, cette méthode consiste à rechercher un hyperplan séparateur pour deux classes données de manière à maximiser la marge entre les exemples de chacune des deux classes. Les SVM présentent de plus l'intérêt de formaliser le problème d'optimisation à partir seulement des produits scalaires entre objets et ainsi de se prêter à l'utilisation d'un noyau. Cette dernière technique permet de plonger les objets dans un espace de dimension éventuellement plus grande favorisant ainsi la possibilité de trouver un bon séparateur.

Dans la tâche qui nous intéresse ici, les objets (documents) étant décrits dans un espace de dimension très importante (plusieurs dizaines de milliers) il est généralement inutile de chercher à l'augmenter. Nous avons d'abord confirmé empiriquement ce postula sur les tâches du défi pour choisir d'utilisé la version linéaire de la librairie LIBSVM[6].

---

[4] Précisément documenté sur http://www.revue-texto.net/Corpus/Publications/Poudat/Chapitre_2.pdf

[5] 163 si l'on inclut les étiquettes positionnelles de type [PREPOSITION :1st] / [PREPOSITION :2nd].

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

4 Résultats obtenus

Le tableau ci-dessous montre les résultats que nous avons obtenus avec notre approche à partir de l'annotation TreeTagger. De manière non surprenante, les équipes obtiennent les meilleurs résultats lorsqu'il s'agit de classifier les deux discours journalistique et encyclopédique (96%), tandis que les résultats sont plus mitigés en matière de classification thématique, Le Monde et Wikipédia étant généralistes par nature (81,1% en classification domaniale seule) :

Tâche	**F-score obtenu par notre équipe**	**F-score moyen des équipes**
Tâche 1 classification en genre	94%	**96%**
Tâche 1 classification en domaine	79%	**83%**
Tâche 2 classification en domaine	**82%**	81%

Tableau 2 : Résultats obtenus

Nos résultats s'avèrent en-deçà de la moyenne des équipes lorsqu'il s'agit de classer les genres (écarts de 2,3% et 3,6 % en dessous de la moyenne pour les tâches de classification en genres, et la tâche 1 de classification en domaine, qui inclut précisément l'information de genre).

Ces résultats peuvent surprendre, puisque notre hypothèse de départ consistait précisément à affirmer que le niveau morphosyntaxique avait un impact discriminatoire élevé en matière de classification en genres et discours. Rappelons toutefois (i) que la tâche de DEFT'08 est ici particulière, puisqu'il s'agit de discriminer deux genres, ou plutôt deux discours entre eux, alors que les études existantes prennent en considération un nombre plus important de genres et de discours, ce qui augmente la portée discriminatoire des observations ; (ii) que Wikipédia et Le Monde multiplient les domaines et les thèmes, ce qui laisse supposer l'existence de pratiques discursives internes aux domaines de spécialité, voire de sous-genres domaniaux. Soulignons également que Wikipédia est un genre plus émergent qu'établi, qui mime les discours et les genres existants, ce qui rend sa caractérisation fort complexe.

C'est peut-être en raison de ce dernier constat que nous obtenons finalement des résultats légèrement supérieurs à la moyenne en matière de classification domaniale, observation validée par la seconde expérimentation, qui modifie peu les résultats obtenus. Le Tableau 3 présente les résultats complémentaires obtenus en utilisant la description morphosyntaxique spécifique (cf. section 2.3) pour des raisons de temps les valeurs reportées correspondent à des estimations de précisions (et non de F-Score) obtenues par validations croisées sur des sous-ensembles de 1000 documents pour chaque tâche.

	Description simple			**Description spécifique**		
	Lexique seul	Morphosyn-taxe seule	Mixte	Lexique seul	Morphosyn-taxe seule	Mixte
Tâche 1 Genre	92%	74%	92%	92%	84%	93%
Tâche 1 domaine	80%	45%	82%	80%	54%	81%
Tâche 2 domaine	73%	36%	72%	74%	51%	75%

Tableau 3 : Résultats obtenus (en incluant notre expérience additionnelle)

Sur cette étude complémentaire, on note une amélioration significative du potentiel classificatoire de la description morphosyntaxique seule en utilisant la description spécifique plutôt que la description simple; cependant la description mixte ne semble pas tirer pas profit de cette amélioration. L'observation précédente suggère une analyse plus technique du phénomène : le classifieur SVM considère l'ensemble des descripteurs dans sa globalité et la sur-représentation des descripteurs lexicaux (plus de 30,000) par rapport aux descripteurs morphosyntaxique (une centaine au plus) revient quasiment à négliger cette dernière description. En revanche l'utilisation réaliste d'un classifieur suggèrerait de limiter la taille du vocabulaire (ou à produire un nombre limité de nouveaux traits) de manière à accélérer le traitement d'un nouveau document; on se ramènera alors à un meilleur équilibre entre les deux ensembles de descripteurs. Enfin si on souhaite proposer un classifieur intelligible on cherchera par exemple à produire des règles de décision (ce que ne permet pas l'approche SVM) et on utilisera par exemple un arbre de décision qui cette fois considère chaque descripteur indépendamment et recherche une combinaison de quelques descripteurs permettant de discriminer les classes : dans ce cas nous avons déjà pu observer un intérêt certain à augmenter l'offre de descripteurs en combinant lexique et morphosyntaxe (Cleuziou, Poudat, 2007).

5 Conclusion

Nous avons cherché à évaluer de manière expérimentale l'incidence des niveaux morphosyntaxique et lexical sur la classification en domaines et en genres dans le cadre pratique offert par le défi fouille de textes 2008.

Dans cette perspective, un ensemble de descripteurs morphosyntaxiques adapté aux caractéristiques du discours a été utilisé en parallèle d'un lexique extrait de manière traditionnelle à partir des corpus d'entraînement fournis. Le classifieur SVM a de plus été retenu pour ses performances reconnues pour la tâche considérée.

Les résultats que nous avons pu obtenir se situent dans la moyenne des performances de l'ensemble des équipes participantes. Pourtant d'une part l'originalité apportée par la description morphosyntaxique a peu influencé les modèles de classification appris et d'autre part l'approche globale utilisée reste relativement naïve (pas de sélection ni de pondération des descripteurs lexicaux). La première remarque s'explique par la nature même des corpus considérés et surtout plus techniquement par l'approche SVM utilisée. En revanche le fait qu'une approche simpliste nous permette d'obtenir des résultats tout à fait honorables tend à montrer que dans le cas de corpus volumineux les méthodologies d'indexation plus fines n'ont pas lieu d'être et s'inscrive plutôt dans un cadre de compensation d'information lorsque les corpus d'entraînement sont de taille plus restreinte.

Références

CLEUZIOU G., POUDAT C. (2007). On the impact of lexical and linguistic features in Genre- and Domain-Based Categorization. Actes de *CICLING-2007*. Lecture Notes in Computer Science, 599-610.

DHILLON I. S., MALLELA S., KUMAR R. (2003). A divisive information theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Researches*, vol. 3, 2003, MIT Press, 1265-1287.

DUMAIS S., PLATT J., HECKERMAN D., SAHAMI M. (1998). Inductive learning algorithms and representations for text categorization. Actes de CIKM ’98, ACM Press, 148-155.

HOFMANN T. (1999). Probabilistic Latent Semantic Indexing. Actes de *22nd Annual ACM Conference on Research and Development in Information Retrieval*, 50-57.

KARLGREN J., CUTTING D. (1994). Recognizing text genres with simple metrics using discriminant analysis. Actes de *COLING 94*, 1071-1075.

KESSLER B., NUNBERG G., SCHÜLTZE H. (1997). Automatic detection of text genre. Actes de *EACL’97*, 32-38.

MALRIEU D., RASTIER F. (2001). Genres et variations morphosyntaxiques. *Traitement Automatique des langues*, vol. 42, 2001/2, Hermes, Editions Lavoisier, 548-577.

POUDAT C., CLEUZIOU G. (2003). Genre and Domain processing in an Information Retrieval perspective. Actes de *3rd International Conference on Web Engineering, Lecture Notes in Computer Science*, 399-402.

POUDAT C., CLEUZIOU G., CLAVIER V. (2006). Catégorisation de textes en domaines et genres : complémentarité des indexations. *Document numérique* vol. 9 2006/1, Hermes, Editions Lavoisier, 61-76.