Un niveau de base pour la tâche 1 (corpus français et anglais) de DEFT'09

Yves Bestgen (1) et Guy Lories (2)

(1) PSOR/CECL – Université catholique de Louvain
Place du cardinal Mercier, 10 B-1348 Louvain-la-Neuve Belgique
yves.bestgen@psp.ucl.ac.be
(2) ECSA – Université catholique de Louvain
Place du cardinal Mercier, 10 B-1348 Louvain-la-Neuve Belgique
guy.lories@uclouvain.be

Résumé – Abstract

L'objectif de cette recherche était d'évaluer l'efficacité d'un classifieur simple de type SVM pour réaliser la première tâche de DEFT'09 sur les corpus français et anglais. Cette approche a été sélectionnée sur la base des informations disponibles à propos des principes qui ont gouverné l'affectation initiale des documents aux catégories objectives et subjectives et sur la base d'une analyse exploratoire des corpus. Une série de tentatives d'optimisation du classifieur n'ayant apporté que des gains négligeables, les performances obtenues peuvent être considérées comme des niveaux de base permettant de se faire une idée de la difficulté de la tâche.

The research goal was to assess the performance of a simple SVM classifier in task 1 of the DEFT09 competition on the french and english corpora. We were led to this approach by the information available regarding the initial categorization and by an exploratory analysis of the corpora. As various attempts to improve the classifier performance brought only negligible improvements the performance obtained can be considered as a base level assessment of task difficulty. A potentially original contribution stems from using discretized document length to bring about a slight performance improvement with the French corpus.

Mots-Clés – Keywords

Catégorisation de textes, Machines à support vectoriel, Discrétisation.
Text categorization, Support Vector Machines, Discretization.

1 Introduction

Le thème de DEFT'09, cinquième édition de la campagne d'évaluation en fouille de textes DEFT, est l'analyse multilingue d'opinion. Trois tâches étaient proposées dans trois langues : le français, l'anglais et l'italien. La première tâche se présentait comme un problème de détection du caractère objectif ou subjectif global d'un article de journal. La deuxième tâche visait à identifier les passages d'un texte qui sont subjectifs, par opposition à ceux qui sont objectifs. La troisième tâche exigeait l'identification du parti politique auquel appartient l'orateur d'une intervention dans le cadre de débats au parlement européen.

Nous avons choisi de nous concentrer sur la première tâche pour les corpus français et anglais. La sélection d'une procédure pour réaliser cette tâche dépend nécessairement de la manière dont la catégorisation de référence a été effectuée. Le tableau 1 reprend la totalité de l'information disponible à ce sujet. Une étape préliminaire d'analyse exploratoire des données était dès lors plus que

souhaitable. Les observations résultant de cette première étape sont présentées à la deuxième section de ce rapport. Sur la base de celles-ci, une technique classique, relativement opaque, mais bien connue pour son efficacité dans le cadre de la catégorisation supervisée de texte (Burges 1998 ; Joachims, 2002), a été choisie (section 3) et des essais ont été menés afin d'optimiser pour le corpus français une série de paramètres (section 4). La combinaison de paramètres sélectionnée est présentée à la section 5 et l'impact de chaque paramètre, considéré indépendamment, y est évalué sur le corpus de test. L'ensemble des résultats qui nous ont été transmis par les organisateurs est donné à la section 6. La conclusion synthétise ce que nous avons appris en participant à DEFT'09.

"Détection du caractère objectif/subjectif global d'un texte : *Le but de cette tâche est de pouvoir détecter si un texte est plutôt un texte d'opinion, subjectif, (comme une critique de film, ou un éditorial) ou plutôt un texte factuel, objectif, (comme une dépêche d'agence, ou des actualités). Nous nous en tenons ici strictement à l'explicite d'un texte, sans tenir compte de ce qui peut être sous-entendu, implicite."* *"Pour les tâches 1 et 3, les participants disposeront donc de références pouvant donner lieu à un apprentissage."* *"L'attribution des valeurs « objective » et « subjective » aux articles a été réalisée de manière différente selon les journaux* [suivent quelques exemples]"

Tableau 1 : Informations disponibles à propos de la catégorisation initiale, extraites du site web de DEFT'09.

2 Pré-traitement et analyse exploratoire

L'ensemble des traitements des textes a été mené en SAS (Statistical Analysis System, V6.12 sous MacOS 9 et V9.1 sous Windows 2000), section Base et Statistics, après prétraitement du corpus au moyen de TreeTagger (Schmidt, 1994). Il s'ensuit que, dans toutes les analyses rapportées ici, c'est la segmentation en termes (token) effectuée par Tree-Tagger qui a produit les descripteurs des documents. Il s'agit donc de mots ("a", "L'", "CNRS"), mais aussi de nombres ("2004", "3.5"), de symboles ("$", "&") et de signes de ponctuation (".", "?").

Les premières analyses ont porté sur la longueur des documents dans le corpus français. Il est immédiatement apparu que deux textes étaient anormalement courts (1 et 3 termes) et un autre anormalement long (plus de 90 000 termes, correspondant à une édition du journal). Il a été décidé de supprimer ces trois documents du corpus d'apprentissage, la catégorisation en objectif/subjectif d'au moins deux de ceux-ci étant sujette à caution.

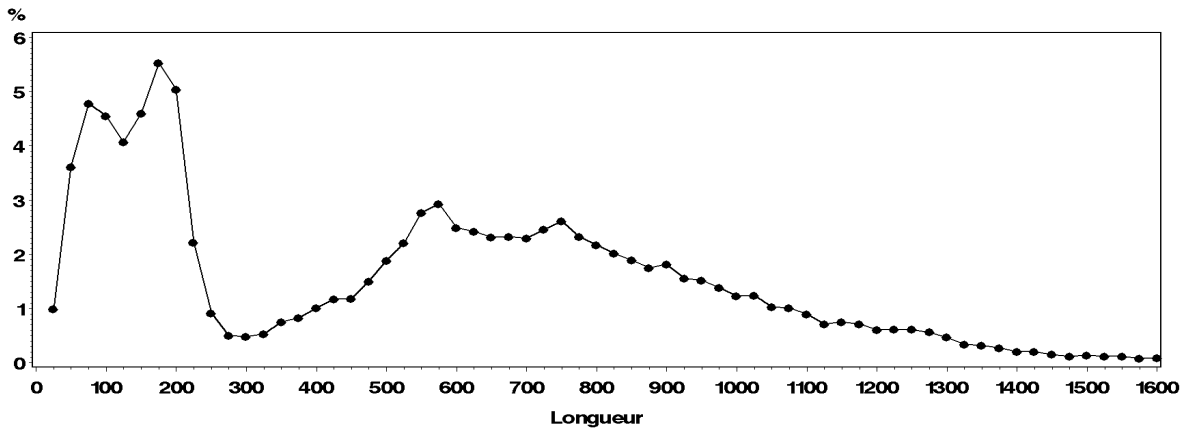

Figure 1 : Distribution des longueurs des documents pour le corpus français : pourcentage sur le nombre total de documents en ordonnée, longueur en nombre de termes en abscisse

La distribution[1] des longueurs des documents est donnée à la figure 1. Comme on peut le voir, cette distribution est nettement multimodale, ce qui n'est pas très étonnant puisque les longueurs des articles de journaux dépendent fréquemment de multiples contraintes tant éditoriales que matérielles. On note aussi une proportion relativement importante de textes courts (moins de 100 termes).

Une analyse détaillée des textes les plus courts montre que ceux-ci correspondent pour une part à des *erratas*, systématiquement considérés comme subjectifs et ce y compris lorsque l'erreur est factuelle et le document initial considéré comme objectif ainsi que l'atteste le tableau 1.

```
<doc id="1_fr:4528">
  <PROPRIETE valeur="SUBJECTIF" confiance="1" />
 <texte>
  <p>Dans l'article intitulé « Anschluss : l'Eglise d'Autriche fait son autocritique », le titre du livre d'Irene Harand publié en 1935 est Son combat, « réponse à Hitler », et non Mon combat, la réponse à Hitler, comme nous l'avons écrit par erreur.</p>
  <p>Par ailleurs, l'orthographe du cardinal Theodor Innitzer était erronée : il s'agit bien du cardinal Innitzer et non Innitzert.</p>
 </texte>

<doc id="1_fr:16181">
  <PROPRIETE valeur="OBJECTIF" confiance="1" />
 <texte>
  <p>Le palais de l'archevêché de Vienne accueille, samedi 12 mars - jour anniversaire de l'Anschluss, l'annexion par l'Allemagne nazie, en 1938 -, un événement que l'Eglise d'Autriche qualifie d' « historique ». Des religieux, artistes, écrivains et scientifiques doivent y lire à haute voix, pendant douze heures, le livre publié en 1935 par la militante catholique Irene Harand, l'une des rares à avoir dénoncé, à l'époque, l'antisémitisme et l'idéologie nationale-socialiste. Intitulé Mon combat, la réponse à Hitler, ce texte oublié vient d'être réimprimé avec un commentaire sans ambiguïté de l'archevêque de Vienne, le cardinal Christoph Schönborn : « Etre chrétien et antisémite sont deux positions inconciliables. »</p>
...
... Mais elle est aussi une critique indirecte des autorités ecclésiastiques sous le nazisme, coupables, au mieux, d'aveuglement, tel le cardinal Theodor Innitzert, au pire, de complicité avec le régime.
...
```

Tableau 1 : Document du type "Errata" catégorisé comme subjectif et article original catégorisé comme objectif

On notera aussi que des erreurs commises par d'autres entités que le journal lui-même sont considérées comme objectives (voir tableau 2).

```
<doc id="1_fr:4207">
  <PROPRIETE valeur="OBJECTIF" confiance="1" />
 <texte>
  <p>Une demande visant à ce que leur pays soit retiré de la liste des membres de la coalition pour la guerre en Irak a été formulée par les autorités slovènes. Le premier ministre, Anton Rop, avait réclamé, la semaine dernière, des explications à Washington, où le département d'Etat avait finalement reconnu qu'il avait cité par erreur la Slovénie comme faisant partie de la liste actuelle des 49 Etats qui soutiennent l'intervention armée.</p>
 </texte>
```

Tableau 2 : Document "Objectif" rapportant une erreur non commise par *Le Monde*

[1] Les distributions de longueur présentées dans ce rapport ont été obtenues en agrégeant les valeurs brutes par tranche de 25 termes et en censurant la distribution vers 1600 termes. Au-delà, les fréquences deviennent très faibles.

On trouve aussi parmi les textes très courts des légendes d'illustrations (tableau 3).

```
<doc id="1_fr:15946">
  <PROPRIETE valeur="SUBJECTIF" confiance="1" />
 <texte>
  <p>Paru dans « El Imparcial »cartoons@courrierinternational.com</p>
 </texte>
```

Tableau 3 : Document "Subjectif" donnant la référence d'une illustration

L'analyse de la longueur des documents pour l'anglais ne met pas en évidence de documents anormalement longs ou courts. La distribution des longueurs, donnée à la figure 2, se distingue de celle obtenue pour le corpus français (figure 1) par la présence d'un seul pic manifeste aux alentours de 300.

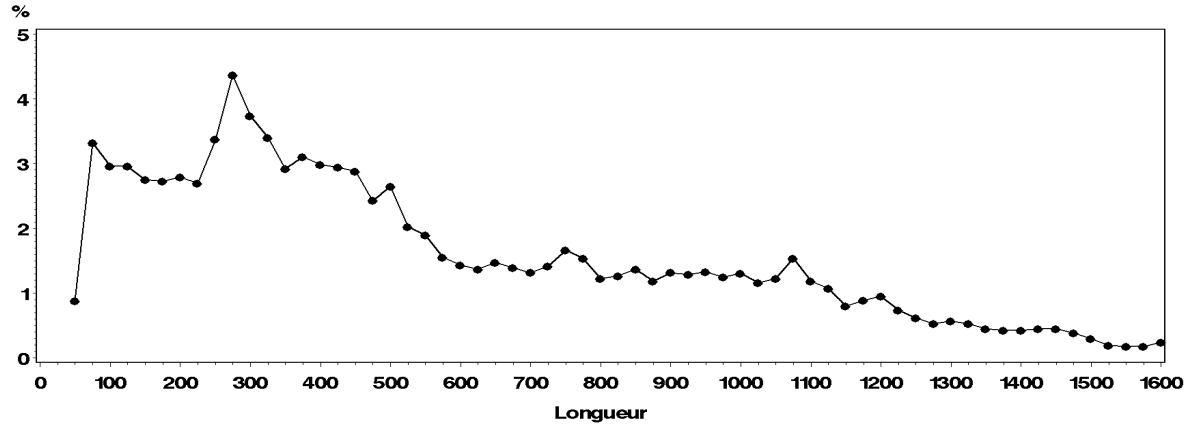

Figure 1 : Distribution des longueurs des documents pour le corpus anglais : pourcentage sur le nombre total de documents en ordonnée, longueur en nombre de termes en abscisse

Une analyse des textes courts du corpus anglais montre que ceux-ci correspondent en partie à des lettres de lecteurs qui sont une fois sur deux considérées comme objectives[2] (voir l'exemple donné dans le tableau 4). Cette affectation fréquente à la catégorie objective, qui ne se retrouve pas dans le corpus français, ne s'accorde pas aisément avec les conceptions courantes dans ce champ de recherche (Hurault-Plantet, 09.04.2009, dia n°9).

```
<doc id="1_en:832">
  <PROPRIETE valeur="OBJECTIF" confiance="1" />
 <texte>
  <p>Sir, I was astonished to read the comments by Professor Patrick Minford in Peter Marsh's article 'Salmon succeeds beer and sandwiches' (May 24). If an adviser to the Treasury can state, 'People from business are invariably a nuisance when it comes to talking about the economy', it explains much about the malign neglect and ignorance which our manufacturers have had to endure for many years.</p>
  <p>Does the professor believe that wealth is created from thin air, and where, if not from taxed wealth, does this bemused academic's salary come from?</p>
  <p>Campbell Dunford,</p>
  <p>chairman,</p>
 ...
```

Tableau 4 : Lettre d'un lecteur catégorisée comme objective

[2] Sur la base de l'analyse des 40 premiers documents du corpus d'apprentissage commençant par "<p>Sir, ".

Au vu de l'information disponible concernant les principes qui ont gouverné l'affectation initiale des documents aux deux catégories et des observations rapportées ci-dessus, il nous a semblé judicieux de nous fixer pour objectif d'évaluer le niveau de performance que peut atteindre une technique de catégorisation simple et relativement opaque (les machines à support vectoriel). L'intérêt de ce travail se limite donc à fournir un niveau de base auquel d'autres approches, plus informées, pourront être comparées.

3 Machines à support vectoriel

L'algorithme d'apprentissage utilisé est une machine à support vectoriel (SVM). De manière générale un tel algorithme apprend à classer un ensemble de vecteurs de R^n en deux catégories. Il s'applique cependant également à des vecteurs à composants binaires. Ici les vecteurs représentent un ensemble de propriétés des textes: longueur, présence ou fréquence éventuellement transformée de tel ou tel mot, appartenance à une catégorie déterminée par ailleurs etc... Il en existe une version linéaire et diverses versions « non linéaires » (à noyau).

L'algorithme *linéaire* catégorise les vecteurs en identifiant un hyper-plan qui sépare, si possible parfaitement, les exemples positifs des exemples négatifs. Il s'agit donc d'une technique qui peut évoquer les techniques linéaires traditionnelles comme l'analyse discriminante cependant l'hyper-plan est défini et calculé de manière particulière.

Dans le cas de *parfaite séparation*, l'hyper-plan de séparation passe entre les points positifs et négatifs les plus proches les uns des autres et il est choisi de sorte que la marge de séparation soit la plus grande possible. La figure 3 représente un tel cas en deux dimensions; la marge maximisée est la distance du plan aux points les plus proches, placés sur les droites en pointillé.

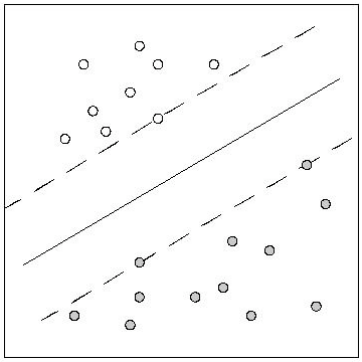

Figure 3 : Séparation parfaite en deux dimensions

Le plan ainsi défini peut être calculé à partir d'un sous-ensemble particulier de points appelé *ensemble de vecteurs de support* du plan. En réalité, le plan est défini par une somme pondérée de ces vecteurs de support (des coordonnées de ces exemples). On définit le plan en annulant l'équation 1 ou au moyen d'un seuil. Chaque vecteur x_i représente un point de support. Les y_i sont les valeurs à prédire et valent +1 ou -1. On voit que w est un vecteur de poids dont sont affectés au bout du compte les différents traits considérés; néanmoins, l'algorithme apprend en affectant chaque exemplaire de support (indicé par i) d'un poids a_i non nul.

$$\vec{w}\,\vec{x} = \sum_{i=1}^{n} a_i y_i (\vec{x}_i \vec{x})$$

Les poids sont choisis de manière à minimiser (à annuler en cas de parfaite séparation) le nombre d'erreurs de classification. La sortie de l'algorithme est constituée de ces coefficients pour chacun des vecteurs de support. Il permet de vérifier la qualité de l'apprentissage supervisé ainsi réalisé et de faire une prédiction pour un nouvel ensemble de points.

Le cas de *séparation imparfaite* peut évidemment, selon les domaines, apparaître plus ou moins fréquemment. Une solution plus flexible a donc été imaginée dans laquelle un point peut être éloigné arbitrairement du plan de séparation dans la direction qui aboutit à une classification correcte. La somme de ces éloignements est cependant soumise à optimisation. L'algorithme minimise alors une somme de deux termes l'un relatif à la marge et l'autre à la somme des ajustements consentis. Selon le poids C, plus ou moins grand, accordé à cette dernière, l'optimisation accorde donc une importance plus ou moins grande aux ajustements autorisés. Il s'en suit que pour des valeurs plus petites de C, le système produit plus d'erreurs de catégorisation durant l'apprentissage, c'est-à-dire s'ajuste moins aux particularités de l'échantillon et vice-versa. Un ajustement très précis aux particularités de l'échantillon d'apprentissage fait évidemment courir un plus grand risque de chute de performance lors de la généralisation à de nouvelles données. Le paramètre C est donc un élément important pour la généralisation ultérieure, tout gain de performance étant susceptible d'entraîner une perte de généralisation.

Un raffinement appelé *algorithme de transduction* peut être employé. Il consiste lors d'une épreuve de généralisation à utiliser l'information disponible dans les exemplaires du nouvel échantillon pour affiner l'apprentissage bien que les catégories de ces nouveaux exemplaires soient inconnues. Bien que leur classification ne soit pas disponible, il est clair que ces données fournissent une information sur l'occupation de l'espace. Cette information peut être utilisée pour éliminer des solutions qui ne s'adapteraient pas à cette occupation. Ceci allonge considérablement le temps de calcul, mais permet d'espérer un surcroît de précision.

Les versions *à noyau* de l'algorithme sont fondées exactement sur le même principe, mais ici une fonction, éventuellement non linéaire, (noyau) des produits vectoriels accumulés dans l'équation (1) est utilisée. Il peut être utile d'observer que dans l'équation 1 les poids sont déterminés sur la seule base des produits vectoriels entre exemplaires de support et exemplaires à classer. Une transformation de ces produits vectoriels suffit à projeter les exemplaires dans un espace différent. Ceci aboutit à identifier un hyperplan qui sépare non plus les exemplaires dans l'espace d'origine, mais, implicitement, leurs projections dans un espace différent. Il n'est pas considéré utile et il n'est donc pas courant d'utiliser ces techniques non-linéaires dans la classification de vecteurs représentant des textes. Nous nous sommes limités pour ce travail à la version linéaire.

4 Essais d'optimisation du classifieur pour le corpus français

L'ensemble des analyses rapportées ici a été réalisé au moyen du programme SVMLight V6.01 (Joachims, 2002). Afin de sélectionner les options de pré-traitement donnant lieu aux meilleures performances du classifieur, le corpus d'apprentissage français a été divisé en un corpus d'entraînement composé de 60% du corpus original et un corpus de test composé des 40% restants. Plusieurs répartitions aléatoires ont été effectuées. Les descripteurs initiaux correspondent à l'ensemble des termes (*token*) identifiés par TreeTagger dans les documents.

4.1 Paramètres classiques

Sur la base des procédures de pré-traitement courantes en traitement automatique du langage et de celles spécifiques au domaine de la catégorisation automatique, nous avons évalué les options suivantes :

Sélection et pré-traitement des descripteurs :

- Descripteurs : unigrammes, bigrammes et trigrammes.
- Seuil de fréquence minimale : 2, 3, 5, 10.
- Suppression de mots fonctionnels (articles, pronoms).
- Lemmatisation au moyen de Tree-Tagger.
- Sélection des descripteurs potentiellement les plus pertinents sur la base du test *t*, du test du *Chi-carré* et du test de *Wilcoxon-Mann-Whitney* (Paquot et Bestgen, 2009).

Attribution de valeurs numériques aux descripteurs :

- Pondération des fréquences des descripteurs dans les documents : fréquences brutes, binaire, log(freq+1) et LSA, la formule classique utilisée en analyse sémantique latente (Landauer et al., 1998 ; Piérard et Bestgen, 2006) qui combine des pondérations locale et globale au moyen de la formule suivante, dans laquelle f_{ij} fait référence à la fréquence brute du terme *j* dans le document *i* :

$$f_{ij}' = \frac{\log(f_{ij}+1)}{-\sum_j \frac{f_{ij}}{\sum_j f_{ij}} \log(\frac{f_{ij}}{\sum_j f_{ij}})}$$

- Normalisation des valeurs d'indice d'un document. Chaque valeur d'indice est divisée par la somme des valeurs pour ce document.

Paramètres de SVMLight

- Paramètre C.
- Algorithme de transduction.

On notera que certaines combinaisons de paramètres mènent à des difficultés, comme l'usage d'une pondération de type "fréquence" en l'absence de normalisation.

Étant donné le caractère partiel des analyses effectuées et le fait que les F-scores obtenus varient en fonction du rééchantillonnage effectué, nous ne rapportons ici que les informations générales qui en ont été extraites. Des données partielles, mais plus précises, basées sur le véritable corpus de test, sont données à la section suivante.

Ces analyses ont montré que l'utilisation d'un seuil de fréquence n'améliorait pas les analyses basées sur les unigrammes, mais bien celles sur les N-grammes, ce qui semble logique. Toutefois, ces analyses ont aussi montré que les bigrammes et les trigrammes pris indépendamment donnaient lieu à des performances moins bonnes que celles obtenues avec les unigrammes. Une combinaison des unigrammes, bigrammes et trigrammes n'est pas plus efficace que les unigrammes seuls.

La suppression des mots fonctionnels n'améliore pas les performances. La lemmatisation n'améliore pas non plus les résultats obtenus sur la base des unigrammes. Elle a un petit effet bénéfique pour les N-grammes, mais celui-ci est insuffisant pour égaler les performances des unigrammes non lemmatisés.

Les procédures de sélections des descripteurs testées n'ont pas permis d'améliorer les performances. Il faut cependant noter que nous n'avons probablement pas testé les indices les plus performants (Forman, 2003 ; mais voir (Gabrilovich, Markovitch, 2004 ; Joachims, 1998) pour une discussion de l'utilité de telles procédures).

L'analyse des différents schémas de pondération indique que la pondération dite « LSA » présentée ci-dessus est plus efficace que la pondération logarithmique et que l'absence de pondération. Ces essais ont aussi montré que la pondération LSA était plus efficace que TfIdf.

La normalisation des valeurs d'indice améliore nettement les performances, comme on pouvait s'y attendre en raison de la grande variabilité des longueurs des documents (Forman, 2003).

En ce qui concerne le paramètre C, nos essais indiquent qu'il a un impact important sur l'efficacité de l'apprentissage et qu'une valeur de 300 semble être optimale. L'apprentissage transductif donne lieu à une égalisation du nombre de documents mal classés dans les deux catégories, ce qui semble être légèrement bénéfique.

4.2 Prise en compte de la longueur des documents

Parmi les stratégies testées, la prise en compte de la longueur des documents mérite une attention toute particulière en raison de la nature même du corpus (articles de journaux) et de la tâche (catégorisation de référence effectuée, selon toute vraisemblance pour le corpus français, sur la base des rubriques dans lesquels les articles sont publiés).

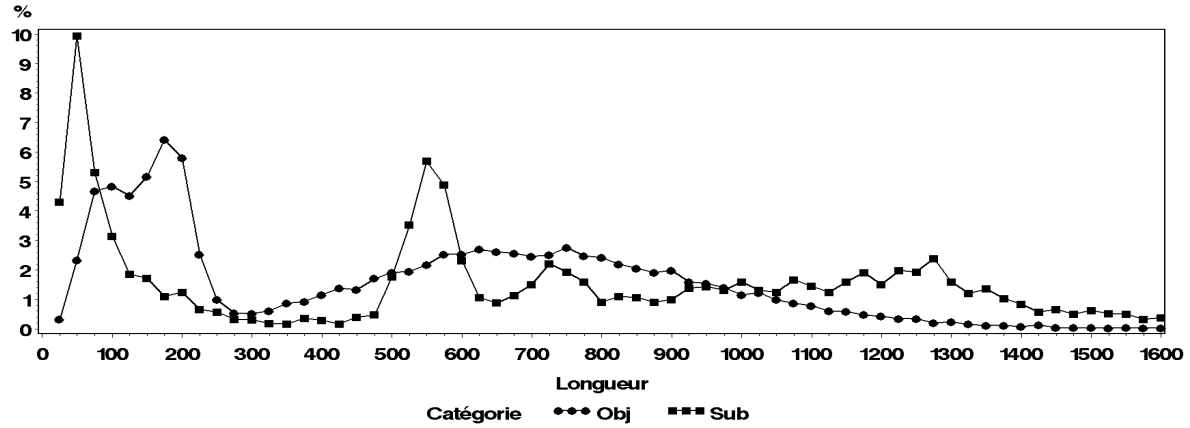

Figure 4 : Distribution des longueurs des documents pour le corpus français selon la catégorie : pourcentage sur le nombre de documents dans chaque catégorie en ordonnée, longueur en nombre de termes en abscisse

Les distributions des longueurs des documents selon la catégorie à laquelle ils ont été affectés sont très différentes comme l'indique la figure 4. Cette situation présente un défi intéressant pour l'emploi d'un algorithme comme SVM. En effet, celui-ci utilise une variable continue comme composant du vecteur représentant l'exemplaire et cette valeur entre dans les produits vectoriels de l'équation 1. Or, la comparaison des deux distributions montre que des plages disjointes de longueur correspondent à des taux de documents subjectifs très élevés. La longueur n'est donc pas liée de manière monotone à la probabilité que le document soit subjectif. Afin de permettre au classifieur de tirer idéalement parti de ces multiples zones, nous avons opté pour une discrétisation de la variable longueur (Liu, Setiono. 1997), représentée dès lors par un certain nombre de variables indicatrices. Discrétiser permet de transformer une variable "continue" en une série d'intervalles contigus qui sont employés pour recoder cette variable sous une forme discrète ou sous la forme d'une série de variables binaires. (Lustgarten et al., 2007) ont observé, dans un cadre d'analyse de données biomédicales, que la discrétisation ne permettait pas d'améliorer l'efficacité d'un classifieur SVM lorsqu'elle est appliquée à un très grand nombre de variables continues en raison du nombre de variables indicatrices ainsi créées. Pour réaliser celle-ci, plusieurs procédures sont envisageables comme :

- segmentation en intervalles de longueur constante, par exemple de 25 termes comme pour les graphiques,

- algorithme de type C4.5, tel qu'implémenté dans SAS Enterprise Mining,
- méthode de (Fayyad et Irani, 1993), basée sur une mesure d'entropie, telle qu'implémentée par exemple dans Weka.

On notera que la première approche est non supervisée alors que les deux autres le sont (Lustgarten et al, 2007). Nos analyses montrent que ces trois approches sont quasiment équivalentes les unes aux autres, les différences maximales observées lors des tests étant inférieures à 0.0025 point de F-score. Elles montrent aussi que la discrétisation de la longueur permet d'améliorer les performances du classifieur (voir la section suivante pour des détails). Il est cependant indispensable de noter que la longueur des documents est une variable très particulière puisqu'elle est liée de manière non monotone à la probabilité de classification subjective, ce qui d'ailleurs a mené à cette discrétisation. Une telle relation a, a priori, peu de chances de se produire avec beaucoup d'autres descripteurs. C'est ce que confirment une série d'analyses qui montrent que, parmi les autres descripteurs, seuls quelques signes de ponctuation montrent ce genre de profil et qu'ils apportent des gains négligeables.

La figure 5 montre qu'une discrétisation de la variable longueur semble nettement moins pertinente en anglais qu'en français (figure 4) pour distinguer les deux catégories. Force est donc de constater que l'intérêt potentiel de la procédure de discrétisation employée ici est limité à des problèmes de catégorisations très spécifiques.

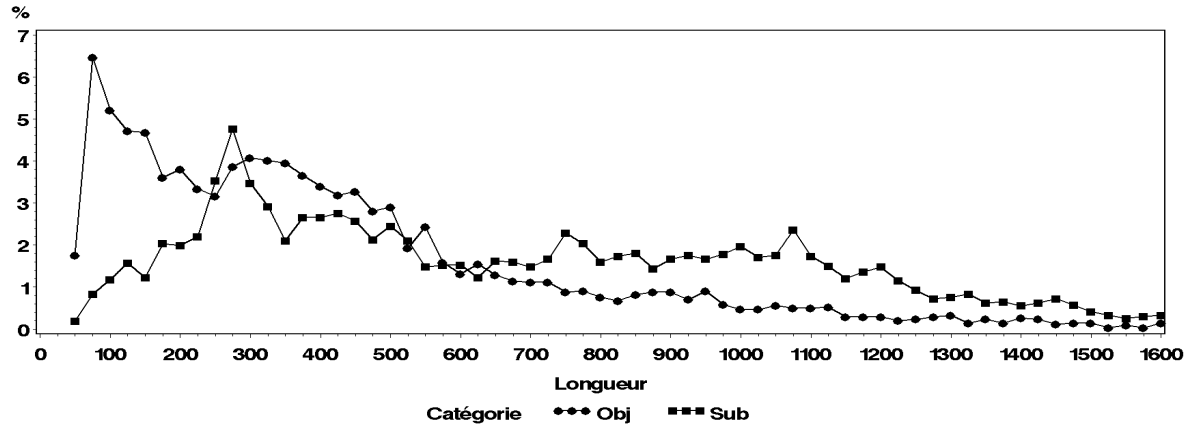

Figure 5 : Distribution des longueurs des documents pour le corpus anglais selon la catégorie : pourcentage sur le nombre de documents dans chaque catégorie en ordonnée, longueur en nombre de termes en abscisse

5 Analyse des effets spécifiques des paramètres

Les expérimentations brièvement rapportées ci-dessus ont conduit à sélectionner, comme optimaux pour le corpus français, les pré-traitements suivants : unigrammes, seuil de fréquence à 2, sans lemmatisation, pondération LSA, normalisation et longueur sous forme discrète. Cette solution, appliquée au corpus français et au corpus anglais de la tâche 1, a été soumise lors de la phase de test. Elle sert aussi de point de référence pour une analyse de modifications *indépendantes* de chacun de ces paramètres sur les performances du classifieur. Le paramètre C de SVMLight a été fixé à 300 et il a été décidé de ne pas utiliser l'algorithme de transduction qui accroît le temps calcul d'un facteur supérieur à 100.

La première ligne du tableau 5 donne les performances de la procédure que nous avons considérée comme optimale. Les lignes suivantes présentent les performances de procédures basées sur les mêmes paramètres que la procédure optimale à l'exception d'un seul. Celui-ci est le seul paramètre mentionné explicitement pour cette ligne. Une cellule vide, quelle que soit sa position dans le tableau, signale donc que la valeur considérée comme optimale de ce paramètre a été employée. Les lignes sont ordonnées en fonction du F-score pour le corpus français.

Lemme	Pondération	Normalisation	Seuil de fréquence	Longueur	F-score Fr	F-score En
Non	LSA	Oui	2	Discrète	0.9271	0.8513
			3		0.9269	0.8502
			5		0.9267	0.8488
			10		0.9260	0.8464
	Binaire				0.9254	0.8445
	Log				0.9209	0.8507
				Continue	0.9199	0.8491
				Aucune	0.9171	0.8485
Oui					0.9171	0.8456
	Fréquence				0.9131	0.8469
		Non			0.8969	0.8408

Tableau 5 : Résultats pour les corpus de test français et anglais en fonction des paramètres.

On note, en tout premier lieu, qu'aucun des paramètres évalués n'a un impact important sur les performances dans les deux langues. Pour l'anglais, la différence la plus importante est d'à peine 0.01 point de F-score. Pour le français, la normalisation semble nécessaire, les autres paramètres n'ayant qu'un impact inférieur à 0.015 point de F-score. C'est cette petitesse des écarts qui nous a conduits à présenter les F-score avec 4 décimales.

Dans ce paysage sans relief, on notera, néanmoins, que la prise en compte de la longueur sous une forme discrétisée apporte un gain de 0.007 point de F-score en français alors que le gain n'est que de 0.002 point de F-score en anglais. Même pour le français, le bénéfice est petit, mais il faut garder à l'esprit qu'il est obtenu par l'entremise de la discrétisation d'un seul descripteur alors que plus de 100 000 autres descripteurs étaient à la disposition du classifieur. Cette absence de relief confirme la possibilité d'utiliser les résultats comme une sorte de niveau de base.

La figure 6 permet de se faire une idée de la manière dont le classifieur a tiré profit de la discrétisation des longueurs des documents. L'ordonnée affichée à gauche sert de référence pour la courbe qui présente les rapports entre les fréquences des longueurs des documents pour les deux catégories obtenus au moyen de la formule suivante :

$$Rapport = 100 \times \frac{Freq_{Sub}}{Freq_{Sub} + Freq_{Obj}}$$

L'ordonnée à droite sert de référence pour la courbe qui présente les poids que le classifieur donne aux descripteurs de la longueur discrétisée, une valeur négative correspondant à un descripteur propre à la catégorie objective et une valeur positive à un descripteur typique de la catégorie subjective. Ces valeurs ont été obtenues au moyen de la procédure proposée par Joachims (http://www.cs.cornell.edu/People/tj/svm_light/svm_light_faq.htm). À titre de comparaison, le descripteur correspondant au mot *erreur* reçoit un poids de 25.87 qui peut être mis en relation avec la fréquence des erratas "subjectifs" dans la classe de longueur *25* ("*par erreur*", voir section 2). Le même mot, mais au pluriel, ne reçoit qu'un poids de 6.51. Comme les poids attribués à un descripteur dépendent des poids de l'ensemble des autres descripteurs, il est normal que la correspondance entre les deux courbes ne soit pas parfaite. Celle-ci est néanmoins suffisante pour montrer l'usage fait par le classifieur de l'information apportée par la discrétisation de la longueur des documents.

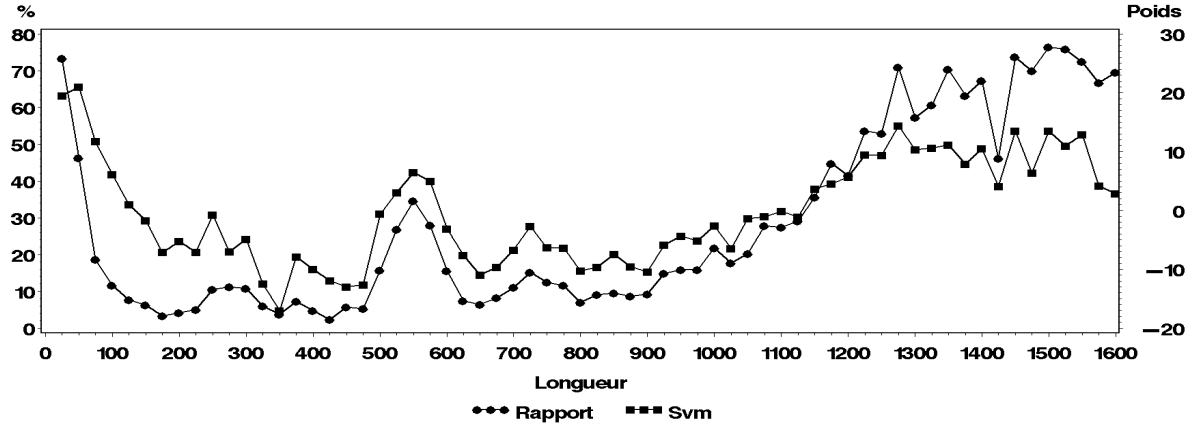

Figure 6 : Comparaison entre le rapport des fréquences des longueurs des documents pour les deux catégories et les poids des descripteurs correspondant à la discrétisation des longueurs (corpus français).

6 Résultats finaux

Les résultats finaux de cette campagne d'évaluation, tels qu'ils nous ont été transmis par les organisateurs, sont un F-score de 0.925 pour le corpus français et de 0.851 pour le corpus anglais. Étant donné que ces valeurs ont été obtenues au moyen d'un algorithme classique en catégorisation de textes et que la section 5 a montré que nos tentatives d'optimisation des paramètres n'avaient apporté que des gains négligeables, ces valeurs nous semblent pouvoir être considérées comme des niveaux de base permettant de se faire une idée de la difficulté de la tâche 1 de DEFT'09 pour les corpus français et anglais et pouvant servir, si nécessaire, de points de comparaison pour les autres participants. Le seul autre commentaire possible est que, contrairement à nos attentes, la procédure de transduction a donné lieu à une infime baisse de la performance pour le corpus français (0.002).

Remerciements

Yves Bestgen est chercheur qualifié du F.R.S-FNRS.

Références

Burges C.J.C. (1998), A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, Vol. 2, 1-47.

Fayyad U.M., Irani K.B. (1993), Multi-interval discretization of continuous-valued attributes for classification learning, Proceedings of the *13th International Joint Conference on Artificial Intelligence,* 1022-1027.

Forman G. (2003), An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research*, Vol 3, 1289-1305

Gabrilovich E., Markovitch S. (2004). Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5, Proceedings of the *21st International conference on machine learning*, 8 p.

Hurault-Plantet M. (2009), Détection des opinions dans les textes, *Séminaire INALCO*, 09/04/2009, (http://www.limsi.fr/Individu/mhp/seminaire/inalco-detection-opinion.pdf).

Joachims T. (1998). Text categorization with support vector machines: Learning with many relevant features, Proceedings of *ECML'98*, 137-142.

Joachims T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*, Dordrecht, Kluwer.

Landauer T.K., Foltz P.W., Laham D. (1998), An introduction to latent semantic analysis, *Discourse processes*, Vol. 25, 259-284.

Liu H, Setiono R. (1997). Feature selection via discretization. *Knowledge and data engineering*, Vol. 9, 642-645.

Lustgarten J.L., Gopalakrishnan V., Grover H., Visweswaran S., (2008), Improving classification performance with discretization on biomedical datasets. Proceedings of *AMIA 2008 Symposium*, 445-449.

Paquot M., Bestgen Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction, In Jucker A.H., Schreier D., Hundt M, (Eds), *Corpora: Pragmatics and discourse* (pp.243-265), Amsterdam: Rodopi.

Piérard S. Bestgen Y. (2006), Validation d'une méthodologie pour l'étude de deux types marqueurs de la segmentation dans un grand corpus de texte, *Traitement automatique des langues,* Vol. 47, 89-110.

Pomikalek J., Rehurek R. (2007), The Influence of preprocessing parameters on text categorization, *Proceedings of world academy of science, engineering and technology*, Vol. 21, 430-433.

Schmidt H. (1994). Probabilistic part-of-speech tagging using decision trees, Proceedings of the *International conference on new methods in language processing*, 9 p., (revised version).