TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Classification automatique de textes à partir de leur analyse syntaxico-sémantique

Jacques Chauché , Violaine Prince
Simon Jaillet, Maguelonne Teisseire
LIRMM-CNRS- Université Montpellier 2
161 rue Ada, 34392 Montpellier cedex 5
chauche@lirmm.fr, prince@lirmm.fr, jaillet@lirmm.fr, teisseir@lirmm.fr

Mots-clefs – Keywords

Analyse, Classification, Extraction d'information
Parsing, Categorization, Information Extraction

Résumé - Abstract

L'hypothèse soutenue dans cet article est que l'analyse de contenu, quand elle est réalisée par un analyseur syntaxique robuste avec calcul sémantique dans un modèle adéquat, est un outil de classification tout aussi performant que les méthodes statistiques. Pour étudier les possibilités de cette hypothèse en matière de classification, à l'aide de l'analyseur du Français, SYGMART, nous avons réalisé un projet en grandeur réelle avec une société qui propose des sélections d'articles en revue de presse. Cet article présente non seulement les résultats de cette étude (sur 4843 articles finalement sélectionnés), mais aussi cherche à montrer que l'analyse de contenu automatisée, quand elle est possible, est un moyen fiable de produire une catégorisation issue du sens (quand il est calculable), et pas simplement créée à partir d'une reconnaissance de "similarités" de surface.

This paper presents the assumption that discourse analysis, when perfomed by a robust parser backed up by an accurate semantic model, is a classification tool as efficient as statistical methods. To study the capabilities of discourse analysis in classification, we have used a parser for French, SYGMART, and applied it to a real project of press articles classification. This article presents the results of this research (on a corpus of 4843 texts), and tries to show that automatic discourse analysis, when possible, is an efficient way of classification through meaning discrimination, and not simply relying on surface similarities recognition.

.

1 Introduction

La classification automatique de textes est un domaine où la fouille de textes et les techniques statistiques produisent des résultats à partir des calculs de fréquence d'occurrence de termes extraits (Salton et al 1983) . On peut aussi leur adjoindre des méthodes d'apprentissage, incluant des modèles de régression , l'approche $k - nn$ (Yang et Liu 1999), des approches bayesiennes naïves, ou adjointes à des arbres de décision (Lewis et Ringuetee 1994). L'analyse syntaxico-sémantique était considérée, jusqu'à présent, comme pénalisante en raison des limitations des analyseurs eux-mêmes. L'idée fondamentale de notre travail est que l'analyse de contenu, quand elle est soutenue par un analyseur robuste[1] avec calcul sémantique dans un modèle adéquat, est un outil de classification tout aussi performant que les méthodes statistiques. En outre, elle est peu sensible à la qualité des corpus d'entraînement puisqu'elle se sert de ressources stables (des dictionnaires non variants), alors que les méthodes statistiques y sont très sensibles. Pour étudier les possibilités de l'analyse syntaxique et du calcul sémantique en classification, à l'aide de nos outils (décrits en section 2), nous avons réalisé un projet en grandeur réelle avec une société qui propose des sélections d'articles en revue de presse, après une veille sur l'ensemble des sources journalistiques possibles. Cette société doit classer plus de 5000 textes par jour et a rapidement cherché à automatiser la classification des textes obtenus. Elle a demandé à notre équipe d'étudier les capacités de classification des textes dans des catégories journalistiques quand on utilise l'analyseur SYGMART^{TM}. Cet article présente les résultats de cette étude. Les outils utilisés sont décrits dans la section 2, la méthode de catégorisation et la procédure sont proposées en section 3. Quelques résultats numériques sont présentés en section 4, et la conclusion (section 5) cherchera à instruire les mérites telle démarche.
Le principe de classification retenu et que expliciterons, est celui du **filtrage sémantique** de textes par des catégories représentées par des centroïdes, dans une méthode de catégorisation supervisée. L'originalité de la démarche est que l'outil d'analyse utilisé pour la classification n'est pas modifié par elle, et les algorithmes de classification se fondent sur la stabilité des centroïdes.

2 Outils de traitement automatique du langage appliqués au problème de la catégorisation

Les outils utilisés pour réaliser cette catégorisation sont un environnement complet d'analyse syntaxique et sémantique (SYGMART) et les vecteurs sémantiques.
L'analyseur SYGMART est fondé sur les algorithmes de Markov étendus aux arbres (**?**). Il a été prévu pour analyser tout langage dont on pourrait écrire la grammaire sous forme de transducteurs d'arbres. Pour le Français, une grammaire (3000 règles à ce jour) a été écrite, inspirée des travaux du linguiste J. Weissenborn. Associé à SYGMART, se trouve un dictionnaire des lexies (50000 entrées) possédant par ailleurs une représentation vectorielle pour la sémantique. La conjugaison de l'analyse syntaxique et de la représentation vectorielle permet d'affecter une "représentation sémantique" à des sous-textes, voire à des textes entiers.

[1] par robuste nous entendons capable de réaliser une analyse éventuellement partielle de toute phrase, même pour les phrases agrammaticales.

2.1 Les vecteurs sémantiques à la Roget

Dans les démarches inspirées de Salton, on définit un espace vectoriel à partir des mots les plus courants, où chaque texte est représenté par un vecteur $\vec{T}$ tel que, lorsqu'il est projeté sur la composante i vaut n_i, où n_i est le nombre d'occurrences du mot i dans T. Cet espace devrait varier chaque fois que l'on change de corpus de référence. Dans notre proposition, on projette la totalité des lexies du dictionnaire sur un espace défini à partir d'une famille de concepts "à la Roget" (Roget 1852). Pour le Français, les lexicologues du Larousse ont défini une famille de 873 concepts hiérarchisés en 4 niveaux (Larousse 1992). Sur un plan vectoriel, cela produit un espace à 873 dimensions que l'on admet comme étant de dimension donnée.[2].Notons que les approches à la "Roget" sont relativement nombreuses depuis quelques années, dans la littérature anglo-saxonne, (Yarowsky, 1992), (Ellman et Tait 1999). Pour le Français, elle a été proposée à l'origine par Chauché (Chauché 1990), mais on retrouve des utilisations vectorielles autres que saltoniennes dans (Besançon et Rajman 2002), et des approches avec thésaurus comme celles de Sinéqua.

2.1.1 Indexation des termes par les vecteurs

Larousse propose pour chaque terme une indexation sur des concepts parmi les 873 du thésaurus. Par exemple, pour le terme "autrefois" on trouve : *Autrefois* : $195.1, 201.3$ ce qui signifie que l'adverbe "autrefois" se projette sur les concepts 195 (PASSÉ) et 201 (ANCIENNETÉ). Les valeurs après le point (".") sont des indications morphologiques que l'on ne représentera pas ici. Le vecteur de "autrefois" se présentera de la manière suivante :
$12 \ldots\ldots\ldots\ldots\ldots\ldots 195$ **...** $201 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 873$
$(00 \ldots\ldots\ldots\ldots\ldots 0000..$**1**$..000..$**1** $0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 0)$
Le vecteur comprend des zéros sur toutes les composantes qui ne sont pas proposées comme signifiantes par le Larousse, et comprend un "1" sur les composantes dites d'indexation, c'est-à-dire celles qui permettent de définir le sens de ce terme.

2.1.2 Espace vectoriel lexical

On considère que tout terme t du dictionnaire est représenté par un vecteur $\vec{t}$ unique dans l'espace vectoriel considéré, que l'on nommera $\vec{\mathcal{V}}$. On suppose qu'il existe une application qui plonge l'espace lexical linguistique dans l'espace vectoriel engendré par la famille de concepts du thésaurus. Pour des besoins de calcul, si $\vec{t}$ est d'abord représenté de la manière indiquée ci-dessus pour "autrefois", en revanche, seule une version normée$\vec{t}_{nor}$ de ce vecteur est conservée dans l'espace. Comme on ne traite que de vecteurs normés, par convention, on écrira $\vec{t}$ pour désigner le vecteur normé du terme t.
Pour normer les vecteurs on introduit une norme euclidienne sur l'espace vectoriel sémantique $\vec{\mathcal{V}}$. En se référant aux propriétés des espaces vectoriels, on définit des lois de composition interne et externe, dont la somme normée, le produit par un scalaire (normé), et le produit vectoriel.
Somme normée : Soient deux vecteurs $\vec{t_1}$, et $\vec{t_2}$ représentant les vecteurs (normés) de deux termes t_1 et t_2.

$$\overrightarrow{(t_1 + t_2)_{nor}} = \frac{\vec{t_1}+\vec{t_2}}{\|\vec{t_1}+\vec{t_2}\|} \quad (1)$$

[2]Le choix de la représentation du thésaurus a été discuté dans deux précédents articles de l'équipe du LIRMM, dans cette même conférence en 2001 et 2002. Le lecteur se réfèrera aux actes idoines pour l' argumentation.

Remarque : la somme normée n'est pas associative : $\overrightarrow{(t_1 + t_2 + t_3)_{nor}}$ n'est pas égal à $(\overrightarrow{(t_1 + t_2)_{nor}} + \vec{t_3}))_{nor}$. Par convention, on ne retiendra comme opération de somme que la somme normée, et on omettra dorénavant l'indice 'nor'.
Distance "angulaire":La distance selon Salton, servant de mesure de similarité est calculée comme le cosinus de l'angle de deux vecteurs.

$$sim(\vec{t_1}, \vec{t_2}) = \cos \widehat{\vec{t_1}, \vec{t_2}} = \frac{\vec{t_1}.\vec{t_2}}{\|\vec{t_1} * \vec{t_2}\|} \tag{2}$$

où "." est le produit vectoriel classiquement défini. La distance que nous utilisons correspond à une mesure relative à l'angle $\widehat{\vec{t_1}, \vec{t_2}}$. Comme nous ramenons tous les angles considérés à l'espace $[0, \frac{\pi}{2}]$, alors la mesure que nous proposons se calcule par :

$$\delta(\vec{t_1}, \vec{t_2}) = 1 - \cos \widehat{\vec{t_1}, \vec{t_2}} \tag{3}$$

Remarques: Ramener les valeurs de δ à $[0, 1]$ est plus pratique que de mesurer des valeurs entre 0 et $1,67$ radiants. Lorsque deux vecteurs sont totalement divergents (intersection vide), leur angle est de $\frac{\pi}{2}$, et le cosinus vaut 0 : leur distance est maximale et vaut 1. Lorsque ces vecteurs sont très proches, leur angle tend vers 0 , le cosinus tend vers 1 et la distance, vers 0. Tous les vecteurs ont un angle forcément compris entre 0 et $\frac{\pi}{2}$, par construction, et appartiennent au même espace vectoriel.

2.1.3 Espace vectoriel sémantique

L'espace des points de $\vec{\mathcal{V}}$ étant beaucoup plus grand que le nombre d'entrées dans un dictionnaire D, l'espace vectoriel peut contenir une quantité de vecteurs qui ne sont pas ceux des termes de D. Si l'on admet une hypothèse de compositionalité en sémantique linguistique, selon laquelle le sens d'un ensemble de mots est une fonction des sens de chaque mot, on peut dire que $\vec{\mathcal{V}}$ définit un véritable espace sémantique, et pas seulement un espace sémantique lexical. Pour toute suite $x = w_1 w_2 \ldots . w_n$ de mots de D, de vecteurs respectifs $\vec{w_1}, \vec{w_2}, \ldots, \vec{w_n}$, il existe un vecteur $\vec{x}$ dans $\vec{\mathcal{V}}$, et il existe une fonction f_n de $\vec{\mathcal{V}}^n$ dans $\vec{\mathcal{V}}$ tels que : $\vec{x} = f_n(\vec{w_1}, \vec{w_2}, \ldots, \vec{w_n})$ et ce, pour tout n. Tout le problème est donc de définir les f_n telles qu'elles puissent être des images formelles des fonctions linguistiques prévalant pour l'obtention de la sémantique des ensembles ordonnés de mots que sont les phrases et, plus largement, les textes.

2.2 Modes de calcul des vecteurs sémantiques des textes

Le calcul du vecteur sémantique de tout segment de texte comprenant une suite de mots sera fondé sur une analyse syntaxique préalable qui permettra de pondérer par un scalaire le vecteur de chaque mot ou groupe de mots en fonction de son rôle syntaxique.

2.2.1 Vecteurs de groupe et de phrase

L'analyseur SYGMART, après avoir affecté des étiquettes aux différents termes d'une phrase, commence par reconnaître des groupes (verbaux, nominaux, prépositionnels). Le calcul sémantique d'un vecteur est donc attaché au groupe comme premier segment. *Les vecteurs de groupe sont des sommes normées des vecteurs de mots du groupe pondérés, ou des vecteurs de groupe qui composent le groupe*. C'est pourquoi, tout groupe de niveau i dans l'arbre d'analyse, s'écrit:

$$\vec{\gamma_i} = \frac{\sum_j \overrightarrow{(\lambda_j v_{j,i+1})_{nor}}}{\|\sum \overrightarrow{(\lambda_j v_{j,i+1})_{nor}}\|} \tag{4}$$

où les $v_{j,\vec{i+1}}$ désignent soit des vecteurs de mots, soit des vecteurs de sous-groupe d'un groupe, de niveau immédiatement inférieur ($i+1$). λ_j est une pondération du rôle syntaxique du mot (respectivement du sous-groupe) dans le groupe. λ_j est tel que si $\overrightarrow{v_{j,i+1}}$ a un rôle de gouverneur, alors λ_j est égal au double des pondérations des autres vecteurs.
Le vecteur d'une phrase est calculé récursivement comme celui d'un groupe de groupes. A chaque étape d'analyse, tout groupe incluant un autre groupe (exemple GV -> V GN : le calcul du vecteur de GV imposera d'abord d'avoir calculé celui de son GN complément). Le vecteur d'une phrase ϕ est donc celui du groupe de niveau 0 (ou racine). Au niveau de la phrase les pondérations sont calculées récursivement de façon à maintenir une atténuation exponentielle.

2.2.2 Vecteurs de textes et ensembles de textes

Bien qu'il existe, dans un texte, une articulation qui donne une importance relative à certaines portions par rapport à d'autres, dans un premier temps, nous avons considéré que les phrases d'un texte étaient équipotentes, et défini le vecteur de texte comme étant le **barycentre** des vecteurs de phrases. En pratique, l'application a montré que les introductions (attaques) et les conclusions (chutes) pouvaient jouer un rôle important. Les travaux de (Pery-Woodley 2000) ou de Nadine Lucas, sur l'articulation des textes, nous ont permis d'améliorer la couverture thématique des textes. Si le vecteur d'un texte T est calculé comme le barycentre des vecteurs de ses phrases, le vecteur d'un ensemble de textes est calculé comme le barycentre des vecteurs de textes, et est aussi un **centroïde**.

3 Application à la classification de textes

La projection sémantique de textes et d'ensembles de textes dans un espace permet de classer des textes par rapport à une direction vectorielle définie comme référence. A la demande de l'entreprise commanditaire, nous adoptons les catégories qu'elle a défini dans son référentiel "métier" et nous considérons une méthode de classification supervisée.

3.1 Constitution des vecteurs de chaque catégorie

Pour chaque catégorie K un ensemble E_K de textes est fourni comme étant le "représentant" de K. Pour cela, nous calculons, pour chaque K, son vecteur de référence qui correspond au centroïde de E_K. La représentation par centroïde est courante dans le domaine (e.g. (Eu-Hong et Karypis 2000)) mais diffère très fortement des nôtres par son mode d'obtention. Deux conditions apparaissent nécessaires pour notre méthode :
il faut , tout d'abord, que le nombre de textes soit suffisamment grand pour que le centroïde (le vecteur de référence) soit le plus précis possible.
En outre, le centroïde doit être stable, c'est-à-dire que la catégorie puisse avoir une direction vectorielle unique dans l'espace, qui ne fluctue plus, et qui puisse représenter une *tendance*.
On dit que le centroïde $\vec{\kappa}$ est **stable** si $\delta(\vec{\kappa}+\vec{T},\vec{T})$ tend fortement vers 0, quel que soit T, le texte considéré. L'avantage de la stabilisation du centroïde d'une catégorie K, représenté par $\vec{K}$, est que pour chaque nouveau texte, si on l'ajoute légitimement à E_K, c'est-à-dire qu'on le réinjecte dans le centroïde, il ne modifie rien, et sa comparaison avec $\vec{K}$ (avec une méthode de filtrage que nous allons expliciter) est fiable.

3.2 Classement d'un article dans une catégorie par filtrage

3.2.1 Vecteurs pour un article

En théorie, c'est le centroïde du texte qui représente le vecteur $\vec{T}$. Cependant, pour chaque texte, nous avons considéré un triplet composé de son centroïde, de son vecteur introduction $\overrightarrow{T_{intro}}$ et de son vecteur conclusion $\overrightarrow{T_{concl}}$. $\overrightarrow{T_{intro}}$ est calculé comme une somme normée de l'ensemble des phrases du texte mais avec une atténuation exponentielle des phrases en fonction de leur rang. Les scalaires affectés sont de la forme $\alpha * \frac{1}{i}$ où i est le rang de la phrase. $\overrightarrow{T_{concl}}$ est le vecteur symétrique de $\overrightarrow{T_{intro}}$. Les coefficients affectés aux phrases sont de la forme $\alpha*(1-\frac{1}{n-i})$ où n est le nombre total de phrases du texte. Pour classer un texte T dans une catégorie K il faut comparer son triplet ($\vec{T}$,$\overrightarrow{T_{intro}}$, $\overrightarrow{T_{concl}}$) avec le centroïde $\vec{K}$. Pour cela, nous avons envisagé plusieurs solutions, qui ont été appliquées l'une après l'autre dans une recherche d'amélioration de la classification.
La première est l'utilisation de la distance δ entre vecteurs : pour qu'elle soit suffisamment discriminante, il faut que les vecteurs des catégories soient bien différenciés (donc distants) dans $\vec{\mathcal{V}}$.
Deuxièmement, le calcul d'une "mesure de concordance" et son utilisation comme critère de classement: si la précédente solution est insuffisante, il faut utiliser le vecteur de catégorie comme filtre sémantique.
Enfin, l'utilisation de la distance entre vecteurs concordants : c'est le problème du choix préférentiel d'une catégorie par rapport à une autre en fonction de l'intensité de la concordance.

3.2.2 Mesure de concordance

Tout vecteur de $\vec{\mathcal{V}}$ a 873 composantes dont certaines n'ont que des intensités très faibles et sont donc peu significatives. Pour avoir une comparaison fiable et plus discriminante entre deux vecteurs, nous avons considéré qu'ils devaient comparables sur les concepts qu'ils "activent" le plus fortement, et jusqu'à quel point ils activent ces concepts. Pour cela, il était inutile de conserver pour le centroïde $\vec{K}$ toutes ses composantes et nous l'avons réduit, après l'avoir trié (par ordre décroissant d'intensité de ses composantes) à sa projection dans un espace plus restreint, à Nb dimensions, où $Nb < 873$. Expérimentalement, nous avons déterminé que Nb valait 250. Au-dessus, on conservait beaucoup de "bruit" dans la comparaison, et en dessous, on n'avait pas un vecteur assez précis. On appelle $\vec{K}_{tr}$ le vecteur de catégorie réduit trié. C'est lui qui va servir de filtre.
On procède de la même manière avec les vecteurs de texte et on produit $\vec{T}_{tr}$ (respectivement les vecteurs introduction, et conclusion triés. On ne proposera les formules ici que pour le vecteur de texte). Il est clair que les composantes de $\vec{K}_{tr}$ ne sont pas forcément les mêmes que celles de $\vec{T}_{tr}$. Mais deux solutions sont possibles : ou bien les deux vecteurs n'ont aucune composante (forte) commune, et auquel cas, cela se voit directement au calcul de la distance δ (elle devient très proche de 1) ou bien $\vec{K}_{tr}$ et $\vec{T}_{tr}$ ont des plus fortes composantes communes, et il est important de mesurer deux types d'écart: l'*écart de rang* et l'*écart en intensité*.
Ecart de rang : Soit i le rang d'une composante C_t du vecteur de référence $\vec{K}_{tr}$, et $\rho(i)$ le rang de cette même composante dans le vecteur $\vec{T}_{tr}$. La formule de l'écart est la suivante :

$$E(i,\rho(i)) = \frac{(i-\rho(i))^2}{(Nb^2 + (1+\frac{i}{2})} \tag{5}$$

Ecart en intensité : Non seulement le rang des composantes communes fortes est comparé, mais aussi leurs intensités respectives. Soit a_i l'intensité de la composante de rang i dans $\vec{K}_{tr}$,

et $b_{\rho(i)}$ l'intensité de cette même composante, qui a le rang $\rho(i)$ dans $\vec{T}_{tr}$. La formule de l'écart en intensité est la suivante:

$$I(i,\rho(i)) = \frac{\|a_i - b_{\rho(i)}\|}{Nb^2 + \frac{1+i}{2}} \tag{6}$$

Ces deux mesures sont ensuite utilisées dans la mesure de concordance P, dont la formule est:

$$P(\vec{K_{sort}}, \vec{T_{sort}}) = (\frac{\sum_{i=0}^{Nb-1} \frac{1}{1+E(i,\rho(i))*I(i,\rho(i))}}{Nb})^2 \tag{7}$$

$$I(i,\rho(i)) = \frac{\|a_i - b_{\rho(i)}\|}{Nb^2 + \frac{1+i}{2}} \tag{8}$$

Propriétés: P n'est pas une mesure de similarité classique, car on peut montrer que P n'est pas symétrique.Elle mesure l'adéquation entre deux vecteurs quand l'un des deux agit comme filtre. P fonctionne de manière "inverse" à la distance de type angulaire δ. En effet, P est élevée lors que δ tend vers 0. P est calculée aussi de la même façon pour les autres vecteurs du triplet.

3.2.3 Mesure de concordance et distance

Pour affecter ensuite un texte à une catégorie, il ne suffit pas de savoir si le vecteur de ce texte est suffisamment concordant avec le vecteur de la catégorie, car la mesure de concordance peut faire en sorte qu'il concorde avec plusieurs catégories. Il faut alors classer préférentiellement le texte dans une catégorie. Soit $\delta(\vec{K}_{tr}, \vec{T}_{tr})$ la distance "angulaire" entre le vecteur de texte et le centroïde de référence. Une nouvelle mesure de distance est proposée avec la formule ci-après où β est un coefficient permettant de renforcer l'importance de la concordance.

$$\triangle((\vec{K}_{tr}, \vec{T}_{tr}) = \frac{P((\vec{K}_{tr}, \vec{T}_{tr})) * \delta(\vec{K}_{tr}, \vec{T}_{tr})}{\beta * P((\vec{K}_{tr}, \vec{T}_{tr}) + (1-\beta)\delta(\vec{K}_{tr}, \vec{T}_{tr})} \tag{9}$$

$\triangle$ est aussi calculé pour les autres vecteurs du triplet du texte.

3.2.4 Vecteur de classement d'un texte

Pour tout texte T, s'il existe p catégories génériquement nommées K_i, on calcule un **vecteur de classement** $\overrightarrow{T_{class}}$ tel que la composante $i$ème de ce vecteur est égale à $P(\vec{K}_{tr(}i), \vec{T}_{tr})$ (respectivement à $\triangle(\vec{K}_{tr(}i), \vec{T}_{tr})$). Cela signifie que *tout texte T est classé* puisqu'on peut calculer la concordance par rapport à chaque catégorie. D'autre part, en triant son vecteur de classement ce qui donne le vecteur $\overrightarrow{T_{class,tr}}$, dans l'ordre décroissant des intensités de ses concordances, on obtient un vecteur dont les premières composantes correspondent aux catégories les plus concordantes.
On calcule alors le **vecteur des catégories du texte** T, qui est en fait le vecteur de dimension p et d'intensités K_i, par ordre d'importance. Il suffit de sélectionner la dimension du sous-vecteur de ce vecteur pour avoir le nombre de catégories, par ordre décroissant, avec lesquelles le texte concorde le mieux, ainsi que les numéros (ou noms) de ces catégories. On calcule de même le vecteur des catégories de l'introduction et de la conclusion du texte T.

3.3 Description de l'application

3.3.1 Les données : corpus et catégories

Le jeu d'essai comporte 4843 articles de presse en Français en provenance de plusieurs sources (agences de presse, journaux, autres) se répartissant en 37 catégories, représentant des rubriques journalistiques, et livrées sous forme d'une liste plate. (Theeramunkong et Lertnattee 2002) montrent qu'une liste plate pose intrinsèquement un problème de précision de la classification. Non seulement ces catégories n'étaient pas hiérarchisées, mais elles pouvaient avoir des recoupements entre elles.
Les catégories représentent des secteurs et des métiers (Banque, Logistique, Hôtellerie, Mode et Textile, Recherche, etc.). Les textes peuvent contenir de quelques phrases à quelques pages chacun. Comme certains articles pouvent appartenir à plus d'une catégorie, la multiplicité des affectations a permis d'établir 5026 liens entre articles et catégories. Les centroïdes ont été stabilisés sur un noyau à partir d'une centaine d'articles par catégorie.
Le noyau représente 2400 premiers articles du jeu d'essai, soit environ 50%. Aucun choix sémantique n'a prévalu à la détermination du noyau, à part le fait qu'il fallait au moins 30 articles par catégorie pour avoir une chance de stabiliser une tendance (nombre à partir duquel on peut raisonnablement avoir une gaussienne). Le noyau contient 2555 liens de classe (plusieurs articles étaient classés dans plusieurs catégories). Le reste des articles a été utilisé comme corpus de vérification.

3.3.2 Objectifs de la classification

Les mesures de classification communément invoquées sont le rappel et la précision du classement. La *précision* est traditionnellement définie par le nombre de liens correctement produits par rapport au nombre de liens produits (par le système). Le *rappel* est traditionnellement défini par le nombre de liens produits correctement (par le système) par rapport au nombre de liens produits par les experts humains. Comme nous calculons toujours le vecteur de classement d'un texte par rapport à toutes les catégories, la précision traditionnelle n'est pas très pertinente. En revanche, ce qui intéressait le commanditaire, c'était de retrouver , dans le vecteur des catégories d'un texte, la ou les catégories de classement proposée(s) par l'expert humain et dans quelle position. Cela pourrait correspondre à une notion de rappel (traditionnel). Cependant, dans la mesure où d'une part, plusieurs articles relevaient de classements multiples, et d'autres part, l'entreprise cherchait éventuellement à découvrir quelques classements inédits, nous avons défini une notion de **largeur** de recherche, dénotée par m, qui correspond au nombre de catégories parmi lesquelles on cherche à retrouver le classement de l'expert. Si on considère cette fois-ci le nombre de liens produits par le système égal à la largeur m, alors, ce que l'on mesure est effectivement une "précision", mais d'une facture un peu particulière. Soit cr le compteur des classements corrects. Soit $cat(exp, T)$ le(s) numéro(s)(ou nom de la catégorie(s)) affecté(s) par l'expert au texte T. L'algorithme de décompte des classements corrects (en figure 1) incrémente ce compteur si le lien est correct dans deux vecteurs des catégories au moins du triplet '(texte, introduction, conclusion)', dont le vecteur des catégories du texte. $cr(m)$ donne le nombre de classements corrects sur une largeur m, c'est-à-dire que $\frac{cr(m)}{m}$ fournit la précision. Si on devait calculer le rappel sur une largeur m on aurait dû avoir de la part des experts humains, un nombre de liens égal à au moins m par texte, or ce n'est pas le cas. On ne peut donc pas calculer un rappel traditionnel dans des conditions rigoureuses. C'est pourquoi nous définissons

Pour tout texte T, $\vec{V_T}$ son vecteur de catégorie, et $\vec{V_T}(i)$ la $i$ème composante de $\vec{V_T}$
Pour $i = 1$ à m faire
si $\vec{V_T}(i) = cat(exp, T)$ et $\vec{V_{T intro}}(i) = cat(exp, T)$
ou$\vec{V_T}(i) = cat(exp, T)$ et $\vec{V_{T concl}}(i) = cat(exp, T)$
alors $cr = cr + 1$

Figure 1: Décompte des classements corrects

une mesure $\pi(m)$ de la manière suivante. Soit $cr(m)$ le nombre de liens de classement corrects pour les articles du corpus. Soit n le nombre de liens de référence du corpus. $\pi(m) = \frac{cr(m)}{n}$.

4 Résultats

La distance δ entre deux catégories étant elle-même souvent inférieure à $0,01$ (un angle de quelques degrés seulement), la discrimination de cette seule mesure n'était pas suffisante. Le passage à la recherche de la concordance $P(\vec{K}_{tr(}i), \vec{T}_{tr})$ (respectivement les deux autres vecteurs du triplet), pour les valeurs des composantes du vecteur de classement, est de meilleure qualité car on a pu sélectionner les vecteurs concordants à une catégorie avec plus de netteté. C'est $\triangle((\vec{K}_{tr(}i), \vec{T}_{tr})$ (respectivement les deux autres vecteurs du triplet) qui a été finalement choisie car la plus probante. Grâce à elle, $\pi(1)$, qui est le "pire des cas" a quand même atteint 47%. Nos premiers essais ont donné les résultats fournis en figure 2. Les pourcentages ont été arrondis à l'unité la plus proche. *Remarques* : Il est difficile d'avoir des corpus dont la taille et le nombre

Valeurs de $\pi(m)$	Noyau	Corpus Vérification
$\pi(1)$, 1 catégorie	49%	42%
$\pi(2)$, 2 catégories	64%	58%
$\pi(3)$,3 catégories	75%	70%
Nombre de textes	2400	2443
Nombre de liens	2555	2471

Figure 2: Comparaison du noyau et du corpus de vérification

de liens sont exactement identiques. Nous avons fait au mieux pour avoir des nombres très proches . Les résultats montrent une différence relativement faible entre le noyau et le corpus de vérification (elle est au pire de 7%), c'est qui nou amène à dire que *la méthode est peu sensible à un entraînement*. Les essais faits avec des méthodes comme $k - nn$ ou comme SVM montrent des différences très nettes entre noyau d'entraînement et corpus de test (Jaillet et al. 2003). Les premiers résultats (que nous mentionnons juste) montrent pour $k - nn$ un rappel de $71,5\%$ pour le noyau, mais de $16,5\%$ sur le corpus de test. Ces résultats sont temporaires (non validés sur les mêmes corpus) mais relativement indicatifs. Le fait d'avoir doublé le nombre d'articles (au total) ne change pas significativement les valeurs de π. La "largeur" (nombre de catégories considérées) introduit une différence de plus grande ampleur. C'est pourquoi nous avons fourni (au commanditaire) un classement global de 4483 articles avec un ratio de 72% environ sur une largeur de trois catégories.

5 Conclusion

Dans cet article nous avons présenté une méthode de classification de documents fondée sur l'analyse syntaxique et le calcul sémantique à base de vecteurs d'indexation. Nous avons donné les règles de calcul sémantique, et les extensions de la représentation aux textes et ensembles de textes, dont certains peuvent être regroupés thématiquement. Ces règles sont fondées sur la capacité d'analyse syntaxique automatique qui n'est souvent pas complète en raison des ambiguïtés ou des formes inconnues qui émaillent les textes réels. SYGMART réalise toujours une analyse partielle, et l'on peut calculer les vecteurs de groupes, sinon de phrase. L'analyse partielle peut induire une mauvaise représentation aussi bien au niveau du centroïde de catégorie qu'au niveau des vecteurs de texte. C'est pourquoi, afin d'améliorer la classification, nous orientons notre recherche vers l'étude de l'impact du seuil d'analyse syntaxique sur la classification. Un passage complet des 4843 textes montre qu'environ $56,5\%$ d'entre eux possédaient un seuil d'analyse de de 30% (un tiers seulement des phrases de chaque texte sont analysées entièrement et correctement). Nous menons des expériences sur des sous-ensembles de textes pour la recherche d' un seuil d'analyse optimal sur un corpus complémentaire de 14000 textes.

Références

Besançon R., Rajman M. (2002) Validation de la notion de similarité textuelle dans un cadre multilingue. *Actes des JADT2002*. Pp.149-159.

Chauché J.(1984) Un outil d'analyse multi-dimensionnelle du discours. *Proc. of COLING-84*.

Chauché J. (1990) Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TA Information* vol 1/1, p 17-24.

Ellman J., Tait, J. (1999) Roget's thesaurus: An additional Knowledge Source for Textual CBR? *Proc. of 19th SGES Int. Conf. on Knowledge-Based and Applied AI*. Springer-Verlag.pp 204 – 217.

Eui-Hong H., Karypis, G. (2000) Centroid-Based Document Classification: Analysis and Experimental Results. *Proc. of PKDD*, p 424-431.

Jaillet S., Chauché J., Prince V., Teisseire M.(2003) Classification automatique de documents: la mesure de deux écarts. *Rapport de Recherche LIRMM* 18p.

Larousse.(1992) *Thésaurus Larousse - des idées aux mots, des mots aux idées.* Paris.

Lewis D.D., Ringuetee, M.(1994) A Comparison of Two Learning Algorithms for Text Categorization. *Proc. of 3rd An. Symp.on Document Analysis and Information Retrieval* Pp 81-93.

Pery-Woodley, M.P. (2000) Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. *Carnets de grammaire* N°8, 164 p, Université deToulouse-Le Mirail : ERSS.

Roget P.(1852) *Thesaurus of English Words and Phrases* Longman, London.

Salton G. , Fox E.A, Wu H. (1983) Extended Boolean Information retrieval. *Communications of the ACM* 26 (12). Pp. 1022-1036.

Theeramunkong T., Lertnattee V. (2002) Multi-Dimensional Text Classification. *Proc.of COLING2002*. Pp1002-1008.

Yang Y., Liu X.(1999)A Re-examination of Text Categorization Methods *Proc. of the 22nd ACM SIGIR Conference*, Pp 42-49.

Yarowsky D. (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proc. of COLING92*.