REPRÉSENTATION DE TEXTES À L'AIDE D'ÉTIQUETTES SÉMANTIQUES DANS LE CADRE DE LA CLASSIFICATION AUTOMATIQUE

CAMELIA IGNAT, FRANÇOIS ROUSSELOT

Abstract. This paper describes an algorithm for document representation in a reduced vectorial space by a process of feature extraction. The algorithm is evaluated in the context of the supervised classification of news articles. We are generating a document representation (profile) represented by semantic tags from a machine-readable dictionary. We are dealing with synonymy handled by thematic conflation, and polysemy for which we have developed a statistical method for word-sense disambiguation. We propose four variants for the profile generation depending on whether a recursive system is used or not, and whether a corrective factor for polysemous words is taken into account or not. We have evaluated 32 variants, depending on the algo rithm type and on three other parameters: grammatical category selection, 15% reduction of the profile, and a stop-list of semantic tags. Some parameters (like profile reduction) have low influence on the classifier performance and others (corrective factor for the ambiguous words, stop-list) improve the performance noticeably.

Mots-clé : représentation des documents; réduction de l'espace de représentation; catégorisation; classification des documents; traitement automatique; désambiguïsation statistique; évaluation.

1. PROBLÉMATIQUE

Pour faire face à l'augmentation grandissante de documentation au format électronique, la classification automatique de documents textes s'est particulièrement développée au cours de ces dernières années. Encore appelée catégorisation, celle-ci consiste à associer de manière automatique un nombre de documents déterminé à un ensemble de catégories pré-définies.

Selon Sebastiani (2004), la construction d'un système de classification repose sur trois étapes principales : la représentation des textes, dans une perspective d'indexation ; l'apprentissage, par entraînement d'un algorithme de classification (on parle de construction de classifieurs) ; et l'évaluation, déterminée en fonction de la finalité du système.

RRL, **LI**, *3–4*, p. 421–439, Bucureşti, 2006

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

La première phase consiste à représenter les textes sous une forme interprétable à la fois par l'algorithme sur lequel repose la catégorisation et par les classifieurs eux-mêmes une fois qu'ils ont étés construits. Celle-ci nécessite donc une définition précise des termes utilisés pour la représentation ainsi qu'une méthode performante permettant de déterminer le poids de chacun de ces termes. La manière de représenter un texte a en effet une très forte influence sur les systèmes de classification automatique. C'est à cette phase en particulier que nous nous sommes intéressés lors de notre travail.

Une des difficultés majeures de la catégorisation concerne la dimension extrêmement élevée de l'espace de représentation. Celui-ci se compose en effet d'un ensemble de termes uniques (mots ou phrases) dont la dimension peut atteindre plusieurs centaines de milliers pour une collection de textes relativement modérée, or seuls quelques traitements basés sur les réseaux neuronaux sont actuellement capables de traiter un si grand nombre de nœuds. Il est donc hautement souhaitable de réduire la dimension de l'espace d'origine, mais sans sacrifier pour autant la précision de la classification.

Notre travail présente une méthode originale permettant de construire une telle représentation des textes et des classifieurs. Pour générer cette description, nous nous basons sur l'ensemble des mots (plus précisément les lemmes) qui composent le document. Celui-ci est ainsi décomposé en une liste des lemmes qui représente le « sens » du texte. Ce système simple de représentation présente cependant plusieurs problèmes :

- La perte de l'information donnée par le contexte syntagmatique, nécessaire à la distinction des lemmes polysémiques (« prix » n'a pas le même sens dans « prix Goncourt », « grand prix » ou « prix d'une marchandise »).
- La présence de synonymes, considérés comme des lemmes différents même s'ils font référence au même concept (« mission » et « délégation » peuvent dénommer la même entité dans un article de journal).

Notre problématique s'axe donc sur la levée de ces deux difficultés issues de la représentation de textes. A cette fin, nous utilisons un dictionnaire associant une ou plusieurs étiquettes sémantiques (appelés *thèmes*) à chaque sens d'un lemme. Celles-ci indiquent la restriction d'emploi d'un lemme à un domaine particulier (informatique, littérature, économie,…). Développées par une équipe de linguistes pour France Telecom R&D, les thèmes sont plus raffinés que les marques de domaine présentes dans la plupart des dictionnaires classiques.

Afin de résoudre les problèmes de polysémie et de synonymie nous avons développé un algorithme de désambiguïsation basé sur les étiquettes sémantiques de ce dictionnaire permettant de représenter les textes par un vecteur pondéré dans l'espace constitué par ces-dernières. La description d'un document est ainsi largement « réduite », la perte relative d'information étant compensée par la dimension de l'espace de représentation final (environ 800 termes) suffisamment important pour permettre la réalisation d'un système de classification performant.

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

Cette méthode de représentation des textes dans l'espace des thèmes dépend de plusieurs paramètres dont nous nous proposons de déterminer l'influence sur l'efficacité de la classification.

2. REPRÉSENTATION DES DOCUMENTS POUR LA CLASSIFICATION AUTOMATIQUE

Au cours des dix dernières années, de nombreuses équipes de chercheurs issues de la recherche d'information ou de l'apprentissage automatique se sont en effet penchées sur les problèmes de classification automatique de textes. En raison du grand nombre d'applications possibles et de leur origine multidisciplinaire, ces systèmes sont extrêmement variés. Nous en présenterons dans les lignes suivantes quelques-uns en nous axant sur notre préoccupation principale : la réduction de l'espace de représentation.

2.1. Réduction de l'espace de représentation

Il existe, d'après Sebastiani (2004), deux manières principales de réduire la dimension de l'espace de représentation des textes :

- Par sélection des attributs : un score est associé à chaque attribut en fonction d'un algorithme chargé de déterminer son degré de pertinence pour un document donné, les attributs ayant les scores les plus faibles étant éliminés.
- Par extraction des attributs : un ensemble de nouveaux attributs extérieurs au document sont générés de manière à représenter ce document dans un espace indépendant dont le nombre d'attributs est plus restreint.

Nous inscrivons notre étude dans une perspective d'extraction des attributs.

2.2. Travaux existants

Les chercheurs travaillant en extraction d'information ont testé une vaste gamme de principes de représentation tels que l'utilisation des variantes de mots présents dans les textes originaux, de mots-clés déterminés par des linguistes, de citations et informations provenant des publications, ou encore de structures développées par les recherches en TALN (Lewis 1990). Au-delà de ces travaux empiriques, quelques tentatives ont également été réalisées dans le but de déterminer de manière théorique les propriétés, et les performances en recherche d'information de ces différents systèmes de représentation. Nous citerons comme travail le plus important dans cette direction celui de Salton « Term Discrimination Model » (Salton *et al.* 1975).

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

Plusieurs méthodes de reduction de l'espace de representation ont été développées : mathématiques, linguistiques ou probabilistes.

Parmi les méthodes mathématiques, on trouve l'analyse factorielle des correspondances (Benzécri 1993); ainsi que celle de « Latent Semantic Indexing » (Manning, Schütze 1999) qui prend un ensemble de documents représentés dans un espace important et le représente dans un espace plus petit.

Une étude de Yang, Pedersen (1997) porte encore sur l'évaluation et la comparaison des méthodes de sélection des attributs utilisées pour la réduction des espaces de très grande dimension appliquées aux problèmes de classification. Cherchant à répondre aux questions : « Quels sont les avantages et les inconvénients des méthodes de sélection des attributs appliquées en classification de textes ? », « Dans quelle mesure la sélection des attributs améliore la précision d'un système de classification ? » et « De quelle proportion le vocabulaire des documents peut-il être réduit sans qu'il y ait de perte d'information ? », ils évaluent et comparent cinq méthodes différentes.

Lewis (1990) présente également le résultat d'une étude portant sur l'amélioration de la représentation de textes basée sur le regroupement de syntagmes. La méthode repose sur l'utilisation d'un analyseur syntaxique pour créer un index des phrases. Ces syntagmes sont des attributs uniques déterminés selon plusieurs relations syntaxiques spécifiées dans le texte original (par exemple : verbe et nom principal du sujet, nom et adjectif qualificatif, etc…). Ceux-ci sont moins ambigus que les mots isolés et possèdent un sens plus précis, cependant, leurs propriétés statistiques sont plus faibles que celles des mots, en raison du grand nombre de phrases différentes et le grand nombre d'occurrence de phrases uniques. L'utilisation de syntagmes dans ces travaux n'a ainsi montré qu'une faible amélioration en extraction d'information.

Il existe enfin de nombreuses études comparant plusieurs méthodes de classification. Yang (1999) en présente quelques-unes dont certaines portent sur la représentation de textes comme moyen d'amélioration de la classification.

2.3. Quelques systèmes de classification

Les systèmes de classification développés aux cours de ces dernières années reposent le plus souvent sur des combinaisons des différents systèmes de représentation et techniques d'apprentissage. Lewis, Ringuette (1994) ont utilisé une méthode appelée « information gain measure » réduisant fortement le vocabulaire du document, dans une approche Baysienne naïve et en utilisant les arbres de décision pour réaliser une classification binaire. Wiener *et al*. (1995) ont utilisé la technique de l'information mutuelle et la méthode statistique du *chi-2* pour sélectionner les attributs dans une approche de réseaux neuronaux. Schutze *et al.* (1995) ont utilisé une analyse en composantes principales des documents pour déterminer leurs dimensions orthogonales dans l'espace des documents. Moulinier

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

et al. (1996) ont utilisé un algorithme inductif d'apprentissage afin d'obtenir des attributs sous une forme normale disjonctive dans le domaine de la classification des dépêches. Enfin, Lang (1995) a utilisé le principe de description minimale pour sélectionner les termes pour la classification de Netnews.

3. ALGORITHME DE GÉNÉRATION DE PROFIL ET CLASSIFICATION SUPERVISÉE

Dans le contexte de la classification automatique, notre travail porte sur la construction de classifieurs. Ceux-ci sont génèrés à l'aide d'un algorithme qui produit un profil de document et qui comporte une méthode de réduction de l'espace de représentation par extraction des attributs et une méthode de désambiguïsation statistique.

3.1. Principe général

Nous nous plaçons dans le cadre de la classification supervisée qui nécessite la connaissance a priori des caractéristiques des catégories prédéfinies dans laquelle on souhaite classifier les documents. Pour chaque catégorie prédéfinie nous disposons donc d'un ensemble de documents pré-classés qui sert à générer le profil de chaque catégorie. Nous utilisons alors le modèle vectoriel de représentation et le coefficient de similarité défini par le cosinus entre deux vecteurs pour déterminer la catégorie d'appartenance d'un document.

Comme nous l'avons présenté dans la partie « Problématique », notre approche part de la représentation classique d'un document dans l'espace vectoriel des lemmes et se propose de développer une alternative par l'utilisation d'un espace moins riche que nous l'appelons « espace des thèmes ». Pour chaque catégorie constituée d'un ensemble de documents, nous construisons alors un classifieur en générant la représentation vectorielle de l'ensemble des documents dans l'espace des thèmes. Pour chaque nouveau document à classifier nous générons un profil dans l'espace des thèmes que nous comparons avec chaque classifieur en calculant le coefficient de similarité (le cosinus entre les deux vecteurs) afin de déterminer la catégorie dont il est le plus proche. Si le profil du document est similaire au profil d'une catégorie on peut alors le classer dans celle-ci.

Nous avons donc besoin d'un corpus d'entraînement, constitué de documents pré-classés, pour construire les classifieurs, et d'un ensemble de test pour évaluer les performances de notre système.

La figure 1 représente le schéma (Moulinier 1996) des étapes nécessaire à la construction de nos classifieurs.

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

3.1.1. *Ressources*

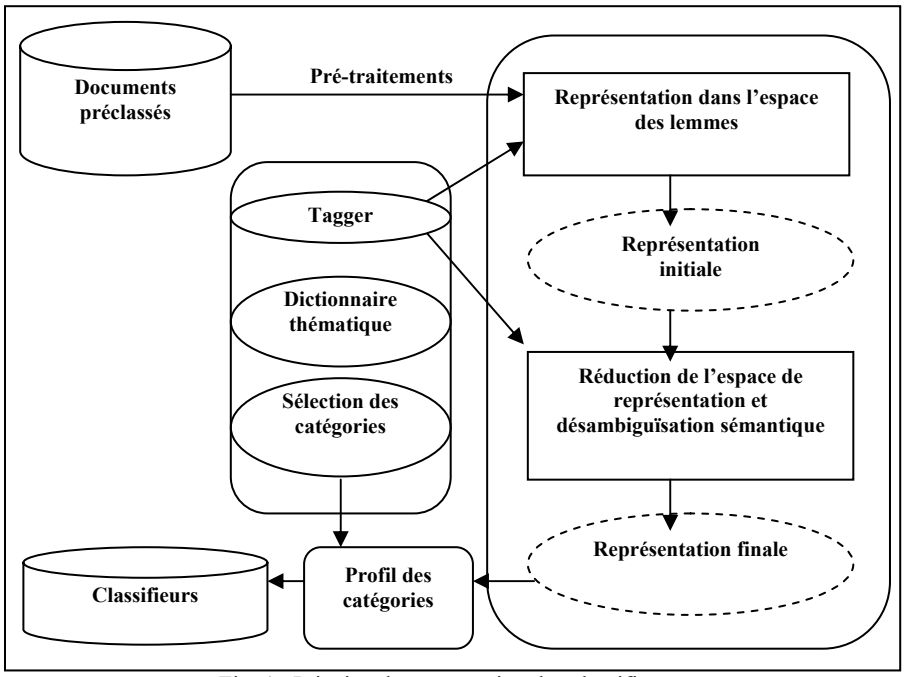


Fig. 1– Principe de construction des classifieurs

La ressource principale sur laquelle s'appuie notre travail, développée par France Telecom RD, est un dictionnaire électronique qui contient l'intégralité des mots de la langue française auxquels sont associés des définitions, des thèmes et des domaines. Chaque mot y est ainsi relié à un ou plusieurs concepts appartenant eux-même à un thème particulier. Ce dictionnaire thématique, réalisé par des linguistes, permet alors de relier directement l'ensemble des mots de la langue française à plus de 800 thèmes sélectionnés par ces derniers.

Ce découpage en thèmes constitue une partition de l'ensemble des sens des mots dans la mesure où un mot polysémique apparaît dans chaque thème pour lequel il possède un sens. Nous estimons, pour notre travail, que seuls les thèmes distincts font référence à des concepts différents alors que la répétition d'un même thème relève de la différenciation d'un sens dans des sous-sens. Le degré de polysémie d'un mot peut ainsi être exprimé par le nombre des thèmes distincts associés.

La construction des classifieurs et l'évaluation s'appuient sur les collections d'articles du journal *Le Monde* des années 2003 (corpus d'entraînement) et 2004 (corpus d'évaluation). Chaque année contient plus de vingt millions de mots répartis sur environ 30000 articles, soit un volume d'environ 120 mégaoctets de

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

données numériques. Cette collection se compose de 46 dossiers correspondant aux rubriques journalistiques habituelles (Sport, International, France, Société…), aux suppléments spécialisés ou dossiers liés à l'actualité.

3.1.2. *Construction des classifieurs*

La construction des classifieurs se fait en trois étapes :
(1) Les pré-traitements au cours desquels le format électronique d'origine du corpus est retranscrit au format XML, puis lemmatisé de manière à ce que chaque forme fléchie soit associée à son lemme et sa catégorie grammaticale. Nous utilisons pour cela le TreeTagger[1] [Schmid (1994)], un analyseur morpho-syntaxique développé à l'Université de Stuttgart, disponible gratuitement et entraîné pour le français.
(2) L'extraction des lemmes (représentation dans l'espace des lemmes) où toutes les paires [lemme / description morphosyntaxique] sont associées à leur fréquence d'occurrence ainsi que la liste des thèmes distincts qui lui sont attachés. Le regroupement par paires [lemme // description morphosyntaxique] se justifie par la présence de lemmes homographes qui possèdent des catégories grammaticales différentes (par exemple « linguistique » pris soit comme nom, soit comme adjectif), mais cette distinction est également utile dans le cas où l'on souhaite réaliser une sélection des catégories morpho-syntaxiques à traiter. Dans notre traitement nous ne gardons que les catégories grammaticales qui sont porteuses d'information sémantique : les noms, les verbes, les adjectifs et les adverbes.
(3) Le calcul du profil de chaque catégorie (représentation dans l'espace des thèmes) où l'algorithme de représentation calcule le poids de chaque thème afin de représenter l'ensemble des documents d'une catégorie donnée dans l'espace vectoriel des thèmes.

3.2. Algorithme de génération des profils

Comme nous l'avons déjà présenté, notre contribution porte sur le développement d'un algorithme permettant d'affiner la représentation des documents dans l'espace des thèmes de manière à réduire les différentes ambiguïtés sémantiques. Pour cela, nous nous appuyons sur l'hypothèse « un seul sens par document » pour la désambiguïsation sémantique et posons trois prémisses de base. Nous analysons ensuite les différents paramètres constitutifs de celui-ci pour optimiser son efficacité.

3.2.1. *Désambiguïsation sémantique*

La désambiguïsation sémantique est une tâche importante qui trouve des applications dans les domaines de l'analyse syntaxique, de la traduction

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

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

automatique ou de la recherche documentaire (Ide, Veronis 1998). Nous nous sommes intéressés, dans notre étude, à la désambiguïsation dans le cadre spécifique de la classification des documents.

L'hypothèse « un seul sens par document » consiste à poser qu'il y a de fortes chances pour qu'un mot ambigu ait le même sens que le sens général d'un document bien structuré. En fonction de la méthode utilisée pour valider cette hypothèse, Gale *et al.* (1992) ont montré que celle-ci est valable dans 98% des cas alors de Krovetz (1998), qui utilise WordNet comme outil, trouve que seuls 33% des mots du corpus Semcor possède plus d'un sens au sein d'un même texte, sans doute en raison de l'affinement des sens dans WordNet. Dans la même perspective, l'hypothèse « un seul thème par document » suppose qu'un mot en accord avec un texte en partage le thème. Magnini *et al.* (2002) ont ainsi réalisé une étude utilisant les domaines de WordNet qui prouve la validité de cette hypothèse, montrant ainsi que le domaine d'un document est un attribut essentiel à la caractérisation de textes.

Si ces travaux de Magnini confirment notre hypothèse de travail, notre tâche est toutefois légèrement différente puisque nous ne cherchons pas à déterminer le sens du mot particulier, mais sa probabilité de prendre un certain sens en fonction du contexte global du document auquel il appartient. Pour la classification, nous considérons ainsi la probabilité d'occurrence de chaque thème pour une collection de texte. Pour cela, nous parlons de désambiguïsation statistique endogène, c'est à dire induite par la structure du document.

3.2.2. *Description de l'algorithme*

La première prémisse sur laquelle repose notre algorithme, est la suivante:

(P1) La fréquence de chaque thème est spécifique à un document donné (nous l'appelons poids du thème).

Ce qui signifie que chaque document possède une représentation unique dans l'espace des thèmes caractérisée par le vecteur pondéré des thèmes

$$\vec{S}^{(doc)}_{thèmes} = \left(S^{(doc)}_{thème_1},\ S^{(doc)}_{thème_2},\ \ldots,\ S^{(doc)}_{thème_n} \right)$$

qui représente la fréquence respective d'apparition de chaque thème dans l'ensemble du document.

L'occurrence d'un thème correspond à l'occurrence de l'ensemble des lemmes qui appartient à ce thème. Ainsi, pour un lemme non ambigu le poids de son thème sera la fréquence d'occurrence du lemme. Pour les lemmes polysémiques, qui ont plusieurs thèmes associés nous tenons compte de la probabilité d'occurrence de chaque thème dans l'ensemble du document grâce à notre deuxième prémisse qui permet de calculer le poids de chaque thème :

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

(P2) La probabilité d'association de chaque lemme polysémique à un thème particulier est proportionnelle au poids de ce thème (soit à sa fréquence d'apparition dans le document).

Si $\sum_{1\le j\le da_k}^{(mot\ k)} S_{thème\ j}^{(doc)}$ est la somme des poids des themes d'un mot motk présent dans le document, dont le degré d'ambiguïté est da_k, alors la probabilité d'associer le thème $thème_i$ au mot_k (où le $thème_i$ est un des thèmes du mot_k), est donnée par la formule :

$$P_{thème_i}(mot_k) = \frac{S_{thème\ i}^{(doc)}}{\sum_{1\le j\le da_k}^{(mot\ k)} S_{thème\ j}^{(doc)}} \tag{1}$$

$i \in \{1,...,\dim(E_{mots})\}$, $k \in \{1,...,\dim(E_{thèmes})\}$

Le poids d'un thème est alors dépendant de la contribution de chaque mot du document associé à ce thème et les coordonnées du vecteur pondéré des thèmes se calculent par la formule :

$$S_{thème\ i}^{(doc)} = \sum_{1\le k\le p} Occ(mot\ k)\cdot \frac{S_{thème\ i}^{(doc)}}{\sum_{1\le j\le da_k}^{(mot\ k)} S_{thème\ j}^{(doc)}} \tag{2}$$

pour $\forall i \in \{1,\ldots,n\}$ où Occ(mot k) représente le nombre d'occurrence du mot k dans le document et p est le nombre des mots distincts du document.

La solution de ce système ne pouvant être calculée directement, nous déterminons $S_{thème\ i}^{(doc)}$ en recherchant le point fixe 2 de l'équation (2) qui est du type :

$$S_{thème\ i}^{(doc)} = F\left(S_{thème\ i}^{(doc)}\right).$$

Il y existe plusieurs méthodes pour déterminer la solution d'une telle équation. Nous avons choisis d'approcher la solution de ce système par la méthode des approximations successives. Nous partons pour cela d'une approximation initiale S_0, puis calculons une suite d'approximations successives S_1, S_2, …, S_{N-1}, S_N par la formule $S_N = F(S_{N-1})$. Le point fixe est atteint quand $S_{N-1} = F(S_{N-1}) \equiv S_N$, c'est-à-dire quand deux valeurs successives sont presque égales (leur différence est presque nulle $|S_N - S_{N-1}| < \varepsilon$, ou ε est la précision désirée).

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

En supposant que le système possède une solution, il est possible d'en calculer facilement une valeur approchée par la méthode itérative de point fixe, grâce à la formule suivante :

$$S^{N}_{thème\ i} = \sum_{1 \le k \le p} Occ(mot\ k) \cdot \frac{S^{N-1}_{thème\ i}}{\sum_{1 \le j \le da_k} S^{N-1}_{thème\ j}} \quad (3)$$

où $S^{0}_{thème\ i} = 1, \forall i \in \{1, ..., n\}$.

Cette relation permet pour la N-ième itération de calculer les poids des thèmes en utilisant les poids du pas N-1 et, lorsque la différence $S^{N}_{thème\ i} - S^{N-1}_{thème\ i}$ devient presque nulle, on obtient une solution approchée de $S^{(doc)}_{thème\ i}$.

Remarque : les poids respectifs des thèmes au premier pas d'itération nous donne les résultats d'un système dans lequel les thèmes d'un lemme ambigu sont équiprobables, soit :

$$S^{1}_{thème\ i} = \sum_{1 \le k \le p} \frac{Occ(mot\ k)}{da_k}$$

où da_k est le degré d'ambiguïté du mot_k.
Utiliser ces poids est alors équivalent à l'emploi de la prémisse suivante :

« Pour un lemme polysémique, la probabilité d'appartenir à un thème est proportionnelle à 1/*da*, où *da* est le degré d'ambiguïté du lemme ».

En considérant que le nombre d'itérations peut être un paramètre de notre algorithme nous avons également évalué notre système par rapport à ce cas particulier.

3.2.3. *Facteur de correction*

Selon l'algorithme exposé précédemment un mot polysémique qui apparaît très souvent dans un texte peut avoir une contribution égale à un mot moins fréquent et non ambigu. Or, les mots polysémiques introduisent parfois, dans le profil d'un document, des thèmes qui ne sont pas connexes au sujet et qui conservent un poids relativement important malgré leur réduction par la méthode itérative. Pour réduire ce biais et donner plus de poids aux mots non ambigus (mots spécifiques à un domaine précis, dans la plus part des cas) nous proposons d'introduire un facteur de correction pour chaque mots, défini en fonction du degré d'ambiguïté da_k :

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

$$mot_k \mapsto f_{corr}(da_k) \quad \forall k \in \{1, \ldots, p\}$$

Pour renforcer la spécificité d'un corpus, nous proposons ainsi une deuxième version de notre algorithme qui prend en compte le degré d'ambiguïté des lemmes contenus dans le corpus, par l'équation :

$$S^{(doc)}_{thème\ i} = \sum_{1 \le k \le p} f_{corr}(da_k) \cdot Occ_{(mot\ k)} \cdot \frac{S^{(doc)}_{thème\ i}}{\sum_{1 \le j \le d_a} S^{(doc)}_{thème\ j}}$$

Nous accordons alors plus d'importance aux mots mono-thématiques (avec un degré d'ambiguïté réduit), en formulant l'hypothèse que ceux-ci sont porteurs de plus d'information sur un domaine précis que les mots fortement ambigus. Cela peut-être traduit par la prémisse :

(P3) Le pouvoir de différentiation d'un lemme est inversement proportionnel à son degré d'ambiguïté.

Nous modélisons cette hypothèse en associant à chaque lemme une fonction décroissante dépendante du degré d'ambiguïté $f_{corr}(da_k)$, définie par :

$f_{corr} : N^* \to R$: $f_{corr}(da_k) = \frac{1}{\ln(CORR + da_k)}$ où da_k est le degré d'ambiguïté de mot_k et *CORR* un paramètre d'optimisation.

Le choix de ce paramètre était un des buts de nos recherches précédentes. Intuitivement, *CORR* devrait donc être le plus petit possible, en laissant le poids discriminant entièrement sur les lemmes non ambigus. Cependant, comme nous l'avons démontré dans une publication précédente (Tufiş *et al.* 2000) ce n'est pas toujours le cas. Une des explications étant que les lemmes mono-thématiques sont rares. En effet, dans cette étude nos classifieurs prenaient leurs décisions en se basant sur des tirages aléatoires parmi les documents à classifier, empêchant de tels lemmes d'apparaître dans les échantillons. Sur la base de ces résultats, nous étudions, dans notre présente série d'expérimentations, l'algorithme qui utilise le facteur le correction ayant pour paramètre *CORR* = 0,1.

3.2. Paramétrage

Il existe plusieurs facteurs capables d'entraîner des variations conséquentes pour la représentation des documents dans l'espace des thèmes.

Nous présentons ici les quatre paramètres que nous avons choisis d'analyser de manière exhaustive. Ceux-ci portent sur le choix des catégories grammaticales,

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

le choix de l'algorithme (itératif ou non, avec ou sans le facteur de correction), la réduction du nombre d'attributs (thèmes) constitutifs du profil (par un seuil sur la somme des poids ou par l'utilisation d'une « stop-liste » des thèmes). En jouant sur ces facteurs, nous espérons accroître les performances de notre algorithme et déterminer ses conditions optimales d'utilisation en fonction du type de corpus à classer.

Chaque type de paramètre fournit une méthode différente destinée à réduire l'espace de représentation d'un document et s'encadre soit dans la réduction par sélection d'attributs, soit dans la réduction par extraction d'attributs (voir paragraphe **2.1.**).

(1) Le **choix des catégories grammaticales** permet de réduire l'espace des lemmes par sélection initiale des attributs. Il est possible de choisir les catégories grammaticales sur lesquelles on se base pour la génération des profils. En fonction de la nature du document dont on réalise le profil (texte descriptif, roman, article de journal) le sens général peut être porté par différentes catégories grammaticales.

(2) Le **choix de l'algorithme** appartient aux méthodes de réduction par extraction des attributs, car l'espace des lemmes est projeté dans l'espace des thèmes. Comme nous l'avons décrit dans la section précédente, il existe 4 manières différentes de générer un profil de corpus en modifiant le système (avec un seule itération ou récursif) et l'algorithme (avec ou sans coefficient de correction).

Nous chercherons à travers les différentes variations de l'algorithme à mettre en évidence l'apport de ses capacités de désambiguïsation. Il est probable que la version de l'algorithme la plus performante soit dépendante du domaine auquel appartient le document dont on réalise le profil selon le degré de spécialisation et de l'homogénéité du corpus auquel il appartient. En fonction de l'ambiguïté des termes qu'il contient, un document sera plus ou moins « sensible » à l'algorithme de désambiguïsation et au coefficient correctif. L'algorithme de désambiguïsation est sans doute également sensible au dictionnaire utilisé puisque c'est celui-ci qui fournit l'ensemble des sens d'un mot. En effet, la méthode itérative amplifiant la différenciation des sens d'un mot, certaines différenciations « subjectives » des sens peuvent influencer les résultats.

(3) La **réduction du nombre de thèmes** est une réduction par la méthode de « sélection des attributs » appliquée à l'espace final.

La technique de réduction du profil consiste à déterminer le nombre de thèmes qui représente un certain pourcentage des thèmes présents dans le corpus. Nous considérerons ici les thèmes dont la somme des poids représente 85% de la somme totale. L'expérience nous a conduit à utiliser ce chiffre de 85%, de manière à éliminer les thèmes les moins importants tout en conservant la spécificité du corpus. Dès lors, nous utiliserons ce chiffre lorsque nous procèderons à la réduction du profil. Cette réduction permet d'augmenter la rapidité de classification, mais leur réduction peut affecter la performance de la classification due à la perte d'information qu'elle implique.

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

(4) L'**utilisation d'une stop-liste** de thèmes réduit également par « sélection des attributs » l'espace final des thèmes et concerne l'élimination des thèmes généraux récurrents du profil du document, une fois celui-ci créé.

Dans le cas de la collection du journal *Le Monde* par exemple, on remarque que des thèmes comme « *temporalité* », « *calendrier* », « *jour* » ou « *nombre* » se trouvent systématiquement en tête des attributs constitutifs des différents profils des dossiers, quel que soit le domaine auquel appartient l'article dont on réalise le profil.

Les thèmes communs renforcent la similarité entre articles provenant de domaines différents. Il existe plusieurs manière de réduire ou d'éliminer l'importance de ces thèmes : soit en établissant, a priori, une liste des thèmes généraux à supprimer (appelée stop-liste), soit en rajoutant une pénalité supplémentaire à leur poids, comme le facteur *idf* (inverse document frequency) (Salton 1989) qui représente l'inverse de la fréquence relative d'un thème dans les autres documents du corpus. Nous proposons pour notre étude une solution qui consiste à utiliser une stop-liste des thèmes construite grâce au calcul de leur facteur *idf*. Pour éviter d'avoir à générer les profils de tous les articles de la collection nous avons constitué un « mini - corpus » par tirage aléatoire de 1000 articles provenant du corpus « Le Monde 2003 ». C'est cet échantillon que nous avons utilisé pour calculer le facteur *idf* des thèmes. Ce facteur nous permet alors de sélectionner automatiquement les thèmes communs à tous les documents de l'échantillon.

4. EXPÉRIMENTATION ET ÉVALUATION

Comme indiqué précédemment, l'implémentation de notre algorithme s'est faite en deux temps : par la construction des classifieurs (phase d'apprentissage) tout d'abord, puis par leur évaluation sur un ensemble d'articles à classer (phase de test).

4.1. Construction des classifieurs

À fin de construire les 32 classifieurs possibles pour chaque catégorie – variations de 5 paramètres binaires – nous avons extrait aléatoirement 400 articles de la collection du Monde 2003.

4.1.1. *Le choix des catégories*

Les dossiers de la collection étant très hétérogènes par leur contenu et par leur taille, nous avons été obligés d'en réaliser une sélection sur la base de deux critères.

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

(1) **La dimension**, soit le nombre d'articles contenu dans une catégorie, a constitué notre première préoccupation. Il convient en effet de déterminer la taille minimale nécessaire à l'obtention du point fixe de notre algorithme. Nous avons ainsi étudié la convergence du profil de chaque catégorie en ajoutant aléatoirement les articles un par un et en calculant la similarité des profils entre les pas N et N+1. Une forte convergence étant obtenu au delà d'une centaine de pas, nous avons choisi de ne retenir que les catégories de plus de 400 articles pour construire nos classifieurs.
(2) **Le sujet**, soit le thème de la catégorie, a été notre deuxième facteur de sélection. Nous n'avons conservé que les catégories spécialisées de manière à éviter les superpositions de sujets similaires, tels que « Economie », « Entreprise » et « Argent ».
Les six catégories finalement retenues sont « Culture », « Economie », « International », « Entreprises », « Sciences » et « Sports ».

4.1.2. *Détection de seuil*

La classification d'un article se fait par le calcul de la similarité (cosinus des vecteurs) de son profil avec celui de chacun des six classifieurs calculés précédemment. A chaque document d_j correspond ainsi pour chaque catégorie C_i un score de similarité compris entre 0 et 1 $CSV_i(d_j)$ encore appelé « *Categorization Status Value* » [Sebastiani (2002)] qui détermine la proximité d'un document et d'une catégorie. La classification nécessite alors la détermination d'un seuil au dessus duquel un document d_j sera classé dans la catégorie C_i. Mathématiquement nous pouvons écrire :

$$CSV_i\left(d_j\right) \geq \tau_i \Leftrightarrow d_j \in C_i$$

$\forall_i = \{1,2,...,6\}$ et $\forall_j = \{1,...,n\}$, où n est le nombre de documents à classer.

Il existe plusieures possibilités de déterminer ce seuil qui influence directement les résultats. Nous avons retenu deux méthodes de calcul, en prenant comme référence la mesure F_1, qui combine la précision et le rappel:
(1) **Le seuil** CSV pour lequel chaque seuil est déterminé en choisissant la valeur minimale qui produit les meilleurs résultats pour un échantillon donné – cent documents dans notre étude. Cette méthode se situe dans une approche multi-classes où le document à classer est associé à toutes les catégories qui possèdent un score de similarité supérieur au seuil établi.
(2) **Le seuil fixe** ou « **k-per-doc** » qui consiste à associer à chaque document les k premières catégories dont les scores sont les plus grands. Nous avons considéré pour nos évaluations le cas particulier où k=1 (« **1-per-doc** ») où chaque article appartient à la catégorie dont la similarité est le plus élevé. Cette méthode se place ainsi dans une approche mono-classe, pour laquelle un article ne peut-être classé que dans une seule catégorie.

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

Le choix de la méthode de détection de seuil est dépendant de l'application envisagée. Nous considérons ainsi le même classifieur dans deux contexte d'évaluation différents: mono-classe ou « 1-per-doc » et multi-classes ou « CSV-seuil ».

4.2. Évaluation

L'ensemble de nos traitements nous a conduit à un grand nombre de résultats (2 méthodes de détection de seuil * 6 catégories * 32 classifieurs) dont nous avons réalisé l'évaluation de la manière suivante.

4.2.1. *Mesures d'évaluation*

Pour chaque catégorie et chaque classifieur nous avons calculé la précision, le rappel, et la mesure F_1 (leur moyenne harmonique). Nous utilisons cette mesure pour comparer les performances de nos classifieurs en fonction des algorithmes et des paramètres utilisés [Lewis (1992)].

Pour évaluer la performance globale du système nous nous sommes basés sur deux mesures : (**micro-moyenne** micro-averaging) où les catégories sont considérées proportionnellement à leur nombre d'exemples positifs et **macro-moyenne** macro-averaging) où les catégories sont toutes également considérées [Tague (1981)] (voir tableau 1). La micro-moyenne regroupe en effet toutes les décisions individuelles dans un tableau de contingence global et nous donne le rappel et la précision pour le tableau « global » alors que la macro-moyenne calcule ces décisions pour chaque catégorie avant de faire la moyenne de ces valeurs. La première favorise ainsi les classifieurs capables de mieux gérer les catégories les plus peuplées alors que la seconde attire l'attention sur les catégories les moins fréquentes.

Tableau 1

Les mesures globales d'évaluation : précision et rappel en micro-moyenne et macro-moyenne, où VP=nb. de vrai-positifs, FP=nb. de faux-positifs, FN=nb. de faux-negatifs et m=nb. de classes.

	Micro-moyenne	**Macro-moyenne**
Précision (π)	$\frac{\sum_{i=1}^{m} VP_i}{\sum_{i=1}^{m} VP_i + FP_i}$	$\frac{\sum_{i=1}^{t} \pi_i}{m} = \frac{\sum_{i=1}^{m} \frac{VP_i}{VP_i + FP_i}}{m}$
Rappel (ρ)	$\frac{\sum_{i=1}^{m} VP_i}{\sum_{i=1}^{m} VP_i + FN_i}$	$\frac{\sum_{i=1}^{t} \rho_i}{m} = \frac{\sum_{i=1}^{m} \frac{VP_i}{VP_i + FN_i}}{m}$

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

4.2. Résultats

En raison du grand nombre de classifieurs que nous avons évalués, l'étude des résultats est relativement complexe. Nous commencerons donc par une présentation générale pour nos deux types de seuils avant d'étudier les paramètres plus en détail.

Pour l'évaluation « 1-per-doc », précision et rappel sont monotones vis à vis de la mesure F_1, le meilleur classifieur possède ainsi le meilleur rappel et la meilleure précision. Les valeurs de la mesure d'évaluation de différents classifieurs sont comprises entre 0.76 pour le meilleur et 0.64 pour le moins bon ce qui signifie que pour cette méthode, notre classification est efficace.

Pour l'évaluation « CSV seuil » par contre, nous observons que le rappel n'est pas monotone pour la précision et F_1. Le meilleur classifieur déterminé par la précision n'est pas le même que celui déterminé par le rappel. La performance des classifieurs est également différente puisque le meilleur possède une mesure F_1 de 0.62, ce qui reste bon, mais le plus mauvais une mesure de 0.32. Cela signifie donc que la variation des paramètres est beaucoup plus importante dans le cas de l'évaluation par méthode CSV.

4.2.1. *Facteur de correction*

Dans les deux types d'évaluation, nous avons remarqué que les cinq meilleurs classifieurs sont issus de la variante de l'algorithme utilisant un coefficient de correction pour les mots polysémiques. Pour le seuil « 1-per-doc », l'algorithme avec facteur de correction est systématiquement mieux classé que l'autre. Cela signifie également que cet algorithme fournit des classifieurs ayant un rappel plus élevé. Cette tendance n'est pas confirmée pour la détermination du seuil par la méthode CSV. Dans ce cas, la précision l'emporte sur le rappel, empêchant la construction de classifieurs performants. Mais les meilleurs résultats sont également obtenus avec le facteur de correction dans la plupart des cas.

4.2.2. *Stop-liste*

La stop-liste exerce une véritable influence sur le résultat de la classification. Avec le « seuil-CSV », elle est le facteur le plus important de distinction et le deuxième dans le cas « 1-per-doc », après le facteur de correction.

4.2.3. *Catégories grammaticales*

Pour les catégories grammaticales, les différences sont moins nettes, néanmoins, par utilisation du « seuil-CSV », on remarque que la sélection des noms seuls nous donne de meilleurs classifieurs. L'influence est étroitement dépendante du domaine thématique (pour le dossier « International » par exemple,

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

cette remarque n'est pas valable) En classification « 1-per-doc », il semble que la différence soit encore plus nette entre les différents dossiers (« Sports » a les meilleures performances pour les noms et « Société » et « International » pour toutes les catégories).

4.2.4. *Réduction du profil*

Ce paramètre n'a pas une très grande influence sur les résultats. Cela nous permet ainsi de réduire les profils de manière à diminuer le temps de calcul nécessaire à la classification.

4.2.5. *Algorithme simple vs. Algorithme récursif*

Il est relativement difficile de réaliser une distinction claire entre les performances de l'algorithme simple et celle de l'algorithme récursif. On remarque en effet que les classifieurs sont souvent regroupés par paires (rec/it1) mais sans qu'il y ait de hiérarchie systématique. Cela est sans doute dû au manque de pertinence de notre type d'évaluation pour distinguer les caractéristiques spécifiques de l'un et de l'autre.

5. CONCLUSIONS ET PERSPECTIVES

L'évaluation « 1-per-doc » nous permet de valider notre algorithme, puisque le meilleur classifieur a un rappel de 80% et une précision de 75% et il semble tout à fait justifié de vouloir l'utiliser dans un système de classification automatique. Toutefois, si cette évaluation nous donne les meilleurs résultats, le calcul du seuil par la méthode « CSV-seuil » nous apporte beaucoup plus d'informations sur les paramètres de construction des classifieurs et permet un paramétrage et une classification plus fins. Ainsi la variante de l'algorithme appliquant une pénalité aux mots polysémiques est nettement plus efficace et permet de mettre en évidence la spécificité des documents.

Pour aller plus loin, il serait intéressant de comparer les performances de notre système de classification construit autour de notre algorithme avec un système « baseline », qui représente les documents dans l'espace des lemmes pondérés par le facteur tf*idf.

Les classifieurs présentés dépendant du contenu et de la structure du dictionnaire qui fournit les thèmes, il serait également utile de comparer deux dictionnaires électroniques et plus précisément les distinctions de sens proposés pour l'ensemble des entrées. De même, pour un dictionnaire donné la précision d'un classifieur donné dépend de l'homogénéité du corpus sur lequel on effectue la classification. Notre algorithme pourrait alors se révéler être une méthode pour comparer l'homogénéité des corpus.

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

Il serait encore possible de développer un système d'indexation contrôlée dans lequel nous remplacerions les thèmes par des descripteurs issus d'un thésaurus. Par une méthode d'apprentissage (voir Pouliquen *et al.* 2003) nous associerions à chaque mot une liste pondérée des descripteurs. Nous pourrions alors appliquer notre algorithme de génération de profil pour chaque nouveau document en utilisant la liste des descripteurs associés aux mots. Nous disposerions ainsi d'une liste pondérée des descripteurs du document.

Finalement, il serait également intéressant d'effectuer une évaluation de notre algorithme en utilisant un autre dictionnaire fournissant les étiquettes thématiques pour chaque mot. Les domaines issus de WordNet (Magnini, Cavaglia 2000), par exemple, ouvriraient la perspective d'une approche multilingue pour nos applications.

BIBLIOGRAPHIE

Benzécri J.-P., 1993, *Pratique de l'analyse des donnees - tome.1 Analyse des correspondances*, Paris, Dunod.

Gale W., K. Church, D. Yarowsky, 1992, "One sense per discourse", in *Proceedings of the 4th ARPA Workshop on Speech and NLP*, 33–237.

Ide, N., J. Veronis, 1998, "Introduction to a special issue on word desambiguation: the state of art", in *Special issue of Computational Linguistics on Word Sense Desambiguation*, 1–40.

Krovetz, R, 1998, "More than one sense per discourse", *Technical Report*, Princeton, NEC Research Institute.

Lang, K., 1995, "Newsweeder: Learning to filter netnews", in *Proceedings of the twelfth International Conference on Machine Learning.*

Lewis, D. D., 1990, "Text representation for text classification", in: P. S. Jacobs (ed.), *Text-Based Intelligent systems: Current Research in Text Analysis, Information Extraction, and Retrieval.*

Lewis, D. D., 1990, "Representation Quality in text Classification: An Introduction and experiment", in *Proceedings of a Workshop in Speech and Natural Language*, Hidden Valley, Pennsylvania, Morgan Fafmann, San Mateo, CA, 288–295.

Lewis, D. D., 1992, "Feature selection and feature extraction for text categorization", in *Proceedings of Speech and Natural Language Workshop*, 212–217.

Lewis, D. D, M. Ringuette, 1994, "Comparation of two learning algorithms for text categorization", in *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*.

Magnini, B., G. Cavaglia, 2002, "Integrating Subject Field Codes into WordNet", in *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece.

Magnini, B., C. Strapparava, G. Pezzulo, A. Gliozzo, 2002, "The Role of Domain Information in Word Sense Disambiguation", *Special issue of the Journal of Natural Language Engineering*.

Manning, C. D., H. Schütze, 1999, *Fundations of Statistical Natural Langage Processing*, Cambridge, The MIT Press Cambridge.

Moulinier, I., 1996, "A Framework for Comparing Text Categorization Approaches", in *AAAI Spring Symposium on Machine Learning and Information Access*, Stanford University.

BDD-A242 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)

Moulinier, I., G. Raskinis, J. Ganascia., 1996, "Text categorization: a symbolic approach", in *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval*.

Pouliquen, B., R. Steinberger, C. Ignat, 2003, "Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus", in *Proceedings of the Workshop Ontologies and Information Extraction, EUROLAN*.

Salton, G., 1989, *Automatic text processing*, Addison-Wesley Publishing Company, Inc.

Salton G., C.S.Yang, C.T. Yu, 1975, "A theory of text importance in automatic text analysis", *Journal of the American Society for Information Science*, January-February, 33–44.

Schutze, H., D. Hull, J. Pederson, 1995, "A comparison of classifiers and document representaions for the routing systems", *Proceedings of SIGIR-95*, p. 229–237.

Schmid, H., "Probabilistic part-of-speech tagging using decision trees", in *Proceedings of the International Conference New Methods in Langage Processing.*

Sebastiani, F., 2004, "Text categorization, Text Mining and its Applications", in A. Zanassi (ed.).

Sebastiani, F., 2002, "Machine learning in automated text categorization", *ACM Computing Surveys*, 34, 1, 1–47.

Tague, J., 1981, "The Pragmatics of Information Retrieval Experimentation", in *Information Retrieval Experimentation*, Butterworth, London, 59–102.

Tufiş, D., C. Popescu (Ignat), R. Roşu, 2000, "Automatic Classification of Documents by Random Sampling", *Publishing House Proceedings of the Romanian Academy*, Series A, 1, 2.

Wiener, E, J, Pedersen, A.S. Weigend, 1995, "A neural network approach to topic spotting", in *Proceedings of the Fourth Annual Symposium on Document analysis and Information retrieval,* SDAIR'95.

Yang, Y., 1999, "An evaluation of statistical approaches to text categorization", *Journal of Information retrieval*, 1, 1/2, Kluwer Academic Publishers, 67-88.

Yang, Y., J. Pederson, 1997, "A comparative study on feature selection in text categorization", in *Proceedings of ICML'97*, 412–420.

BDD-A242 © 2006 Editura Academiei

Provided by Diacronia.ro for IP 216.73.217.142 (2026-11-14 12:59:30 UTC)