Classification supervisée hybride par motifs lexicaux étendus et classificateurs SVM

Laurent Kevers, Amin Mantrach, Cédrick Fairon, Hugues Bersini, Marco Saerens

Classification supervisée hybride par motifs lexicaux étendus et classificateurs SVM

Version corrigée et mise à jour de l'article JADT 2010 – Revised and updated version of JADT 2010 article

L. Kevers[1], A. Mantrach[2], C. Fairon[1], H. Bersini[2], M. Saerens[1,2]

[1]CENTAL & ISYS - Université catholique de Louvain (UCL) - Louvain-la-Neuve - Belgique
[2]IRIDIA - Université Libre de Bruxelles (ULB) - Bruxelles - Belgique

Abstract

We present the comparison and combination of two different semi automatic classification methods: an original linguistic based analysis, named *extended lexical patterns* (*motifs lexicaux étendus*, MLE) and a machine learning approach (SVM). Classification is supervised because of the use of a thesaurus containing the definition of categories. First, both systems are used and evaluated separately on the same real dataset: law and parliament documents in French. Quite unexpectedly, MLE performs as well as a *state-of-the-art* method such as SVM. During the second step, the combined method gave a performance improvement which underlines the complementarities of both basis methods.

Résumé

Dans le contexte de l'indexation semi-automatique de textes, nous présentons la comparaison et la combinaison de deux méthodes de classification mettant en œuvre des approches différentes : analyse par une méthode originale à forte composante linguistique que nous nommons *motifs lexicaux étendus* (MLE) d'une part et apprentissage artificiel SVM d'autre part. La classification est de type supervisée car elle exploite un ensemble de catégories définies par un thésaurus documentaire. Dans un premier temps, les deux systèmes sont appliqués et évalués séparément sur un même jeu de données réelles, des textes de type législatif et parlementaire en français. De manière quelque peu inattendue, la méthode MLE permet d'atteindre des performances tout à fait compétitives par rapport à la technique *state-of-the-art* que constitue SVM. Les méthodes sont ensuite combinées dans le but d'obtenir une performance finale supérieure aux performances individuelles. Le gain obtenu indique une complémentarité entre les deux méthodes.

Mots-clés : classification supervisée, thésaurus, transducteurs, SVM, méthode hybride.

1. Introduction

Cet article s'inscrit dans la problématique de l'accès à l'information. On constate que, confronté à une masse documentaire importante, il est compliqué d'apparier de manière optimale une requête formulée par un utilisateur avec les documents les plus pertinents.

C'est dans ce contexte que l'on a vu surgir, depuis plusieurs années, de nombreuses technologies destinées à faciliter l'accès aux bases de données documentaires. La classification et l'indexation automatique de documents en font partie. Nous nous intéressons plus spécifiquement au cas *supervisé* dans lequel les catégories, ou classes, sont déterminées et définies *a priori*. La tâche de classification[1] revient alors à attribuer, pour tout nouveau document à insérer dans la base documentaire, une ou plusieurs de ces classes.

Dans les entreprises, la classification s'effectue assez fréquemment de manière complètement manuelle, ce qui lui apporte une grande précision. La cohérence de l'indexation peut cependant être diminuée suite à la variabilité des décisions prises par les différents documentalistes sur une longue période. D'autre part, la quantité de documents est souvent

[1] Par la suite, nous utiliserons indifféremment les termes classification, catégorisation et indexation.

énorme et rend le processus manuel inabordable pour beaucoup d'organisations. Il est donc parfois remplacé par des méthodes d'indexation automatiques, moins précises mais beaucoup plus rapides et cohérentes. Si ces techniques sont adéquates pour des applications *temps réel* ou pour analyser un très grand nombre de documents, elles ne constituent pas toujours une solution idéale dans le contexte de systèmes d'information nécessitant une haute précision.

Dans cet article, nous décrivons une méthode semi-automatique dont le but est d'améliorer l'efficacité et la cohérence de l'indexation manuelle en suggérant de manière automatique à un documentaliste une liste de catégories ou de mots clés potentiels. L'approche adoptée consiste à exécuter deux méthodes de classification différentes, l'une exploitant des critères plutôt linguistiques et l'autre mettant en œuvre des méthodes statistiques basées sur des techniques d'apprentissage artificiel. Ces deux approches se distinguent l'une de l'autre par différents points qui les rendent complémentaires, ce qui permet des les combiner afin d'obtenir une performance finale supérieure à leurs performances respectives.

Dans cet article, la section 2 est consacrée à une brève présentation des données ayant servi de base à notre travail. La première méthode de classification, qui effectue une analyse au moyen de ce que nous avons appelé des *motifs lexicaux étendus* (MLE) (Kevers, 2009), est ensuite présentée en section 3.1. Celle-ci est suivie, au point 3.2, par un exposé de la seconde méthode, basée sur une technique d'apprentissage artificiel *state-of-the-art* : SVM (*Support Vector Machine* ou *machines à vecteurs de supports*) (Cortes et Vapnik, 1995; Sebastiani, 2002). Après un bref commentaire des premiers résultats (section 3.3), les possibilités de combinaison sont ensuite décrites à la section 3.3 suivies par les conclusions en section 4.

2. Présentation des données

La classification supervisée s'appuie sur une ressource qui définit l'ensemble des catégories ou classes attribuables aux documents. Ici, il s'agit d'un *thésaurus*, c'est-à-dire un *vocabulaire contrôlé* qui regroupe, décrit et définit un ensemble de concepts relatifs à un certain domaine[2]. Un concept correspond à une classe ou catégorie et est représenté par un terme principal appelé *descripteur*. Il peut être relié à plusieurs *non descripteurs* ou *synonymes* par une relation *used-for* (UF). Les concepts sont organisés hiérarchiquement à l'aide des relations *broader-than* (BT) et *narrower-than* (NT). La relation *related-term* (RT) permet de définir une similarité sémantique entre concepts. Les grands thésaurus peuvent être fragmentés en *microthésaurus* couvrant chacun un sous thème particulier. Plusieurs normes internationales, dont ISO 2788 (1986) et AFNOR Z47-100 (1981), définissent plus longuement les thésaurus.

Les documents et le thésaurus utilisés sont issus d'une base documentaire d'une grande organisation. Les textes relèvent du domaine législatif et parlementaire. Le thésaurus a été spécialement conçu pour l'indexation manuelle des documents au sein de cette organisation.

Le thésaurus contient 2.514 descripteurs et 2.362 synonymes. Les descripteurs sont répartis en 47 microthésaurus. Le nombre maximal de niveaux hiérarchiques est 6, mais la profondeur s'établit le plus fréquemment entre 2 et 4. Les expressions composées sont bien représentées : 66,59% des descripteurs (1.674 sur 2.514) et 61,85% des synonymes (1.461 sur 2.362).

Notre corpus compte 12.734 documents XML contenant 32.953.724 mots[3]. La taille moyenne d'un document se situe par conséquent à 2.588 mots. Le titre du document est délimité à

[2] A titre d'exemples, citons le thésaurus du Parlement de la Communauté européenne, Eurovoc (http://europa.eu/eurovoc/), ou encore le thésaurus de l'Organisation des Nations Unies pour l'alimentation et l'agriculture, Agrovoc (http://www.fao.org/aims/ag_intro.htm).

[3] Cette mesure approximative du texte brut (pas de balises XML) a été obtenue à l'aide de la commande *wc*.

l'aide de balises particulières. Pour chaque document, on dispose des descripteurs assignés manuellement par des documentalistes professionnels en situation réelle. Ces informations servent de référence pour l'évaluation. Le nombre de descripteurs attribués varie entre 1 et 37, avec une moyenne de 1,92 (1,62 microthésaurus différents). Certaines catégories ne sont représentées par aucun document et d'autres, au contraire, sont utilisées de manière très soutenue. 669 catégories ne sont jamais utilisées et le descripteur le plus fréquent est lié à 412 documents. En moyenne, une catégorie est utilisée pour l'indexation de 9,71 documents.

Pour cet article, nous nous limitons aux 47 catégories générales (microthésaurus), sans tenir compte des catégories plus fines. Cette généralisation du problème de départ est dictée par le manque de données pour certaines classes, ce qui rend difficile l'entraînement de classificateurs SVM pour celles-ci. D'autre part, certains documents étant multilingues, nous avons restreint le corpus aux 11.157 textes exclusivement rédigés en français.

3. Méthodes de classification

3.1. Analyse par motifs lexicaux étendus (MLE)

Le point de vue défendu par cette approche est que l'appartenance d'un texte à une catégorie thématique se matérialise dans le document par l'utilisation d'un certain nombre de mots particuliers. Dès lors, si les catégories sont correctement définies[4], on peut raisonnablement penser qu'il est possible de trouver une intersection suffisante entre les lexiques du document et du thésaurus pour décider de manière automatique de leur assignation.

Nous nous appuyons également sur l'hypothèse selon laquelle une expression composée a généralement un sens très précis et constitue souvent un bon candidat en tant que concept descripteur du document. L'observation du lexique d'une langue telle que le français nous montre que les concepts complexes sont souvent exprimés à l'aide d'expressions composées. Le terme « allocations » est par exemple souvent délaissé au profit d'une forme composée telle que « allocations de chômage » ou « allocations familiales ». De plus, les expressions composées sont souvent moins polysémiques (Yarowsky, 1993). Par conséquent, notre système se doit de prendre en compte les unités polylexicales.

3.1.1. Principe général

La méthode consiste à exploiter le thésaurus en créant automatiquement, à l'aide des termes descripteurs et de leurs synonymes, une ressource d'extraction comparable à un ensemble d'expressions régulières améliorées, ce que nous nommons des motifs *lexicaux étendus* (MLE) (Kevers, 2009). La ressource est appliquée aux textes afin de retrouver un maximum d'expressions *pertinentes* dérivées du thésaurus, en délaissant les expressions ne possédant *a priori* pas de pouvoir classifiant. Ces motifs utilisent des unités lexicales issues directement du thésaurus (favorise la précision) et des éléments plus généraux (lemmes, méta-étiquettes, *etc*.) qui permettent une certaine variation par rapport aux dénominations de départ (favorise la couverture). Les expressions extraites servent ensuite de base pour la sélection des classes.

3.1.2. Travaux apparentés

L'utilisation d'un thésaurus pour améliorer ou guider la classification n'est pas une approche des plus répandue. Certains travaux y sont cependant apparentés. *KEA++* (Medelyan et Witten, 2006) est un système agissant en deux phases : l'extraction de mots clés et leur

[4] Elles doivent posséder un terme descripteur principal et le maximum de non-descripteurs (synonymes).

filtrage à l'aide d'un thésaurus sont suivis par une étape d'apprentissage artificiel capable de mettre en œuvre différents types d'algorithmes. Pouliquen et al. (2003) présentent une méthode statistique et associative de classification de documents dans *Eurovoc* qui se base sur différentes mesures de similarité. Cette étude a été menée sur diverses langues dont l'anglais, l'espagnol et le français. Névéol et al. (2005) évaluent deux systèmes hybrides – *MTI* pour l'anglais, *MAIF* pour le français – d'indexation de documents médicaux dans *MeSH*[5]. Ces systèmes combinent à la fois une approche de type *sac de mots* et une approche plus statistique (*PubMed Related Citations* pour l'anglais, une mesure de similarité exploitant la méthode *k-Nearest Neighbour* pour le français). Toujours dans le domaine médical et en français, Pereira et al. (2008) étudient avec *F-MTI* les possibilités d'assignation de descripteurs *MeSH* à l'aide de plusieurs terminologies. La technique d'analyse repose à nouveau sur un algorithme de *sac de mots*. Enfin, c'est Névéol et al. (2006) qui s'approche le plus de notre méthode. Ce système d'indexation de documents en français repose sur *MeSH* et exploite, comme nous, une série de transducteurs, mais ceux-ci sont cependant construits manuellement et non automatiquement.

3.1.3. Processus de classification

Notre système de classification crée une version du thésaurus sous la forme de transducteurs. Cette opération est unique et ne fait pas partie du processus de classification répété pour chaque document. Elle est facilement réalisable, malgré le nombre élevé d'éléments que peut contenir un thésaurus, car elle est complètement automatique.

Les transducteurs sont générés dans un format compatible avec le logiciel de traitement de corpus *Unitex*[6] (Paumier, 2008). Chaque catégorie est représentée par un automate contenant une transduction qui renseigne le code de la catégorie. Un certain nombre de traitements sont nécessaires afin de passer d'un transducteur uniquement capable de reconnaître les expressions contenues dans le thésaurus (*motif lexical strict*) à un transducteur qui permet de retrouver une famille d'expressions proches ou reliées (*motif lexical étendu*).

Le premier traitement vise à couvrir les variations relatives à l'emploi dans les textes du pluriel ou du singulier et est atteint au moyen d'une étape de *lemmatisation*. Cette étape est réalisée à l'aide du *TreeTagger*[7] (Schmid, 1994). Par exemple, le transducteur basé sur l'expression « taux d'intérêt légal », issue du thésaurus, est capable de retrouver les formes « taux d'intérêt**s** légal », « taux d'intérêt lég**aux** », ou encore « taux d'intérêt**s** lég**aux** »[8].

Le deuxième traitement consiste, comme dans de nombreux travaux en recherche d'information, à éliminer les *mots vides* (*stopwords)*. Ils sont en réalité remplacés par une méta-étiquette (`<TOKEN>`). Cela permet d'améliorer la reconnaissance d'expressions dans lesquelles un mot peut être remplacé par un autre. C'est par exemple le cas pour « contrôle **de** chômeurs », « contrôle **du** chômeur » et « contrôle **des** chômeurs »[9].

Le troisième traitement apporté est la possibilité d'insertion, entre chaque mot, d'un terme facultatif. Cette extension permet d'aller plus loin dans la reconnaissance d'expressions similaires. En effet, il est assez courant qu'une expression complexe puisse être simplifiée ou

[5] Medical Subject Headings : http://www.nlm.nih.gov/mesh/.

[6] http://www-igm.univ-mlv.fr/unitex/.

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

[8] Le processus de lemmatisation peut conduire à un phénomène de surgénération. Dans un contexte de reconnaissance, cela ne représente cependant pas un problème. Au contraire, cela permet de reconnaître les occurrences mal orthographiées ou s'éloignant de la norme.

[9] Ce cas n'est pas couvert par la lemmatisation car le lemme de « des » est « un ».

qu'elle soit modulée par des adjectifs. Par exemple, « agence de protection de l'environnement » peut être retrouvé sous la forme de « agence <ADJ> de protection de l'environnement », avec un adjectif tel que « fédérale », « régionale », « belge », *etc*.

Enfin, nous avons exploité une base de données toponymique afin d'ajouter aux noms de pays les formes adjectivales et les gentilés correspondants («Italie» : <italien> et <Italien>).

D'autres stratégies d'extension de la ressource d'extraction ont encore été testées sans être retenues car peu ou pas performantes : la détection des entités nommées ainsi que la transformation de groupes nominaux en groupes verbaux à l'aide de *Verbaction*[10], un lexique de noms d'actions morphologiquement apparentés à des verbes. Ce mécanisme aurait permis de retrouver les formes telles que « <traiter> l'eau » à partir de « traitement de l'eau ». Ces deux extensions sont largement tributaires des ressources sous-jacentes. Leur amélioration pourrait donc éventuellement nous amener à en réexaminer l'utilisation.

Les transducteurs générés sont rassemblés en un transducteur principal. La figure 1 illustre un transducteur généré automatiquement.

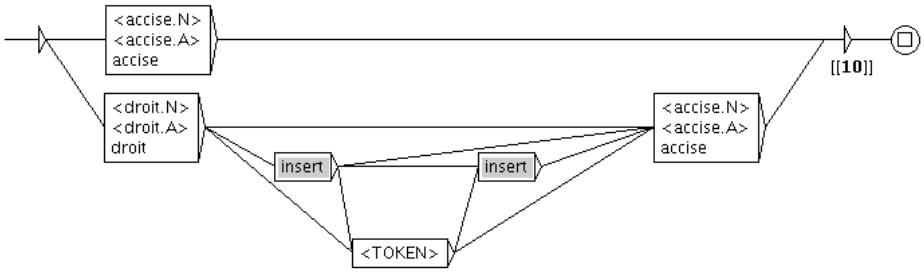

Figure 1: Transducteur pour la classe 10 (descripteur : « accise », synonyme : « droit d'accise »).

Une fois cette ressource à disposition, son application à un texte peut être effectuée. Ce dernier doit cependant subir une série de prétraitements. Outre les classiques normalisation, tokenisation et application de dictionnaires électroniques, il est intéressant d'implémenter une désambiguïsation ciblée afin d'éviter certaines erreurs récurrentes et souvent typiques du thème des textes. Citons l'exemple de « art. 2 » (abréviation de « article 2 ») qui est de manière systématique interprété comme relié à la catégorie ART (arts plastiques, *etc*.) du thésaurus. La désambiguïsation peut être réalisée à l'aide d'un transducteur de remplacement qui substitue les motifs posant problème par une expression non ambigüe.

L'application des transducteurs à un texte résulte en la production d'une liste d'expressions et de termes pertinents, accompagnés des catégories auxquels ils sont reliés (*cf*. figure 2).

```
0 12 @000101024.xml@                193 193 batiments[[MT191]]
14 16 <title>                       235 235 controlees[[MT992]]
53 53 aeroport[[MT111]]             264 270 personnel de le aeroport[[MT111]]
57 57 bruxelles[[MT991]]            274 274 bruxelles[[MT991]]
60 63 </title>                      295 295 ministre[[MT124]]
77 77 president[[MT157]]            299 299 transports[[MT111]]
113 113 ministre[[MT124]]           348 348 aeroport[[MT111]]
117 117 transports[[MT111]]         356 356 livre[[MT133]]
124 124 armee[[MT122]]              360 360 marchandises[[MT192]]
124 124 armee[[MT102]]              385 385 ministre[[MT124]]
140 140 aeroport[[MT111]]           420 420 president[[MT157]]
144 144 bruxelles[[MT991]]          446 446 deputee[[MT124]]
```

Figure 2 : Liste de mots ou d'expressions retrouvés à l'aide des transducteurs dans un texte. Le code de catégorie (ici des microthésaurus) est inclus entre crochets.

[10] http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=hathout&subURL=verbaction/main.html

Pour chaque expression, une pondération de type *TF-IDF* (Salton et al., 1983) est calculée. Plusieurs éléments viennent moduler cette valeur, principalement la présence d'expressions polylexicales et la présence dans le titre du document. Les poids calculés par expression sont ensuite additionnés afin d'obtenir une valeur globale par catégorie. Les catégories les plus significatives pour un document sont sélectionnées à l'aide d'une fonction de seuil dynamique (nombre variable de catégories) dans cette liste pondérée. Une description plus complète de ce système, incluant une présentation technique plus détaillée, se trouve dans Kevers (2009).

3.1.4. Résultats

L'évaluation a été réalisée par rapport à l'indexation manuelle renseignée dans les documents originaux. Les mesures classiques ont été utilisées et appliquées à chaque document : *rappel*, *précision* et *f-mesure* (F_1). La moyenne sur l'ensemble des documents est calculée de manière *macroscopique*, c'est-à-dire en considérant les trois mesures document par document avant de réaliser leurs moyennes respectives.

La fonction de seuil dynamique peut être appliquée de manière plus ou moins stricte (21 niveaux possibles) sur les listes de catégories. Pour chaque niveau, les performances sont évaluées sur l'ensemble des documents et une valeur moyenne est calculée. Cette méthode nous permet de faire varier le rappel et la précision et de rechercher une approximation de la solution optimale. Le meilleur résultat obtenu, en termes de f-mesure, se situe au 11ème niveau de seuil et donne une valeur égale à 61,02, soit un rappel de 64,79% et une précision de 66,05%. Le nombre moyen de catégories proposées est 1,61. Une évaluation plus détaillée est rapportée dans Kevers (2009).

3.2. Analyse par classificateurs SVM

Les SVM, ou *machines à vecteurs de supports* (Cortes et Vapnik, 1995), fait référence à une technique déjà largement exploitée et ayant fait ses preuves, notamment en ce qui concerne la catégorisation automatique de textes (Joachims, 1998 ; Sebastiani, 2002). Le principe général repose sur la recherche d'un séparateur qui regroupe les échantillons en deux classes. Le séparateur est disposé de manière à maximiser la marge avec l'échantillon le plus proche de chaque classe. Les échantillons sont représentés par un ensemble de caractéristiques qui prend la forme d'un vecteur. La marge est délimitée par deux vecteurs qui sont nommés *vecteurs de supports*. Une fois entrainé, le modèle ainsi construit permet de décider de l'appartenance à une classe plutôt qu'à une autre pour tout nouveau document soumis au système.

3.2.1. Adaptation au cas multiclasses

A l'origine, SVM est destiné à résoudre les problèmes de choix entre deux classes. Or, nous sommes confrontés à un cas *multiclasses*. Plusieurs approches existent cependant pour étendre le modèle SVM à un tel problème. Nous nous proposons d'en évaluer deux.

La première approche, nommée *un contre tous* (SVM1-T), consiste à créer un classificateur par classe (Duan et Keerthi, 2005). Chaque classificateur est entraîné à différencier les documents qui appartiennent à une classe et ceux qui n'y appartiennent pas. Concrètement, cela signifie que pour 47 classes nous avons 47 classificateurs à estimer. Chaque classificateur est un classificateur binaire entraîné sur tout le corpus annoté. En termes de temps d'entraînement, il faudra donc entraîner autant de fois le système sur l'entièreté du corpus que nous avons de classes différentes. La classification d'un nouveau document est réalisée en le soumettant à l'ensemble des classificateurs. Chacun fournit un score d'appartenance à la classe qui correspond à la distance par rapport à la marge. Etant dans un problème *multilabels*, le résultat est constitué par une liste ordonnée des classes, basée sur les scores obtenus.

La seconde approche porte le nom de *un contre un* (SVM1-1) et consiste à opposer deux à deux les différentes classes (Duan et Keerthi, 2005). Chaque classificateur est alors entraîné à différencier deux classes spécifiques. Concrètement, cela signifie que pour 47 classes nous aurons 1081 (47*46/2) paires de classes et donc 1081 classificateurs. Ces classificateurs binaires ne sont entraînés que sur les documents des deux classes qu'ils discriminent. Lorsqu'un nouveau document doit être classifié, celui-ci est soumis à l'ensemble des 1081 classificateurs. Pour chaque classe, le nombre de duels gagnés contre les autres classes est calculé. La classe qui aura emporté le plus de duels sera celle qui aura la plus haute place dans la liste ordonnée finale. Celle-ci contient donc toujours l'ensemble des 47 classes.

Notons que si le nombre de classificateurs à entraîner pour le SVM1-1 est largement supérieur, ils ne doivent cependant être entraînés que sur les documents appartenant à deux classes spécifiques. Avec SVM1-T, chaque apprentissage se fait sur l'ensemble du corpus.

3.2.2. Des textes aux vecteurs supports

Nous avons choisi de représenter les textes bruts selon un moyen souvent utilisé en recherche d'information : le *sac de mots* (Berry et Browne, 1996). Dans cette représentation, on extrait de chaque document d_i un vecteur de termes dans lequel chaque composante indique la fréquence du terme j (pour j compris entre 1 et N) dans le document i : $d_i = (f_{i,1},f_{i,2},...,f_{i,N})$. Cependant, plutôt que d'utiliser la fréquence des termes, nous avons opté pour une mesure de type *TF-IDF* (Salton et al., 1983).

Le nombre de composantes N de chaque vecteur correspond à la taille du vocabulaire d'un texte. Cette taille est souvent assez élevée. Afin de la réduire, deux prétraitements spécifiques à la langue française ont été effectués. Premièrement, on réalise la suppression des *mots vides*. Ceux-ci n'ayant aucun pouvoir classifiant, on peut les éliminer sans altérer le lexique caractéristique du thème du document. Ensuite, la *racinisation* (*stemming*) permet également de rassembler sous une même unité lexicale une famille de termes (« Italie », « italien(s) » et « italienne(s) » peuvent par exemple être tous rassemblés sous la racine « itali »). Ces deux traitements ont été réalisés à l'aide de la librairie *open source Galilei*[11].

Une fois les vecteurs constitués, la classification proprement dite peut être effectuée. Pour les deux approches présentées (*un contre tous* et *un contre un*), nous avons utilisé la bibliothèque *svm_light*[12]. Le script *svm_learn* nous a permis d'apprendre les classificateurs à partir des ensembles d'apprentissage. Celui-ci a été utilisé en mode classification avec un noyau linéaire. Les autres options ont été laissées en mode par défaut. Le script *svm_classify* a été utilisé pour classifier les documents de test.

3.2.3. Résultats

Nous avons tout d'abord comparé les classificateurs SVM en mode *un contre tous* (SVM1-T) et en mode *un contre un* (SVM1-1). Pour ce faire, nous avons effectué dix tirages aléatoires des documents, et pour chacun une validation croisée à dix plis. L'hyperparamètre C[13] du SVM a été calibré en interne lors du premier pli, la valeur retenue a alors été utilisée durant toute la phase d'apprentissage. Afin que les résultats soient comparables, les mêmes tirages ont été utilisés pour les deux modèles et toutes les expérimentations (SVM1-T et SVM1-1). Lors de chaque pli, nous obtenons une liste ordonnée de classes pour chaque document.

[11] http://www.galilei.ulb.ac.be/.

[12] http://svmlight.joachims.org/.

[13] Constante qui permet de régler le compromis entre le nombre d'erreurs de classement et la largeur de la marge.

Comme pour la première méthode de classification (*cf.* section 3.1), les mesures de *rappel* et de *précision* sont calculées sur la base de cette liste et par rapport aux annotations manuelles originales. L'évaluation de ces mesures est effectuée à plusieurs reprises en faisant varier le nombre de catégories finalement retenues dans la liste. Pour ce faire, on utilise une fonction de seuil qui ne retient que les *N* (de 1 à 47) meilleures catégories. Pour chaque valeur de *N*, on calcule la moyenne du rappel et de la précision sur les dix plis.

Les résultats comparatifs entre le modèle SVM1-T et SVM1-1 sont présentés à la figure 3. Le modèle SVM1-1 permet d'obtenir, à rappel équivalent, une précision moyenne de l'ordre de 20 à 25% supérieure. Par conséquent, c'est le modèle SVM1-1 qui a été sélectionné pour participer à la classification hybride.

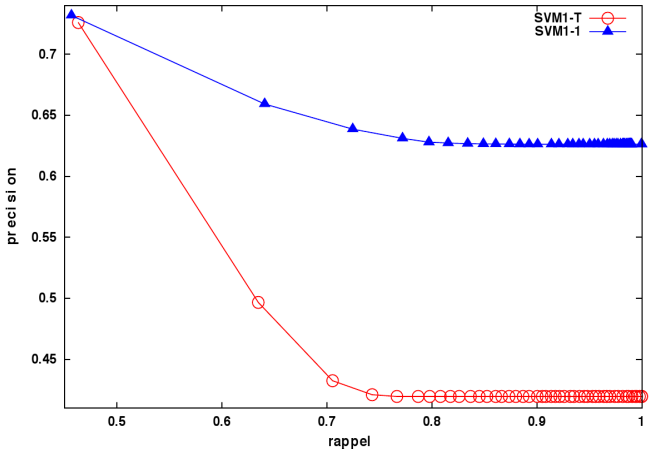

Figure 3 : Courbes précision-rappel pour les méthodes SVM1-T et SVM1-1 moyennées sur 10 tirages aléatoires de l'ensemble des documents en mode validation croisée à dix plis.

Afin de pouvoir comparer SVM1-1 avec la méthode MLE (*cf.* section 3.1), nous avons calculé les valeurs de *rappel*, *précision* et *f-mesure* en appliquant la méthode de seuil dynamique, telle qu'utilisée pour la méthode MLE. Le meilleur résultat obtenu, en termes de f-mesure, se situe au 1er niveau de seuil et donne une valeur égale à 59,15, soit un rappel de 53,93% et une précision de 72,90%. Le nombre moyen de catégories proposées est 1,05.

3.3. Discussion des résultats

De manière quelque peu inattendue, MLE atteint des résultats qui égalent ou surpassent, ceux de la méthode *state-of-the-art* que constitue SVM. La méthode MLE se distingue par son mode d'analyse très transparent qui s'oppose à l'aspect partiellement *boîte noire* de SVM. L'exploitation du thésaurus comme ressource d'extraction et de classification permet aussi d'éviter le recours à une phase d'apprentissage, ce qui offre la possibilité d'utiliser la méthode MLE en l'absence de tout corpus annoté. Le temps de mise en œuvre et la dépendance à la langue et au domaine d'application sont cependant plus importants qu'avec SVM. Ces constatations plaident pour une complémentarité des deux méthodes.

3.4. Combinaison des deux méthodes

3.4.1. Principes

Les deux méthodes de classification présentent certaines caractéristiques différentes et sont donc susceptibles de fournir des résultats complémentaires. Tout d'abord, les deux systèmes mettent en œuvre des approches qui exploitent des *indices* qui ne sont pas exactement de même nature. Plus concrètement, avant l'application d'une fonction de seuil, les deux

méthodes fournissent des listes de résultats de longueur différentes : SVM renvoie toujours une liste de 47 catégories tandis que MLE livre un résultat de longueur variable. Dans ces listes, la répartition des poids de catégories est aussi fort différente. Avec SVM, ils décroissent de manière très progressive, alors que la méthode MLE donne parfois lieu à de brusques sauts dans la pondération. Les premières catégories ont ainsi souvent un poids beaucoup plus élevé que le reste de la liste.

La combinaison des résultats des deux méthodes de classification, se fait par fusion des listes de classes pondérées et ordonnées. Deux *modes* de combinaison sont envisagés : le premier correspond à l'union des résultats et le second à leur intersection. En ce qui concerne le *procédé* de combinaison, à nouveau, deux approches différentes ont été envisagées.

La première approche exploite les listes de catégories déjà restreintes à l'aide de la fonction de seuil appliquée de manière optimale pour chaque méthode de classification séparément. La fusion des deux listes revient donc à prendre le meilleur résultat obtenu par chaque méthode (*cf*. sections 3.1 et 3.2) et à effectuer leur union ou leur intersection.

La seconde approche consiste à fusionner les listes complètes des deux classificateurs avant d'appliquer la fonction de seuil dynamique pour obtenir la sélection finale. Les poids fournis par les deux listes ont été normalisés afin de toujours manipuler des valeurs comprises entre 0 et 1. Lors de l'union ou l'intersection des deux listes, pour chaque document et chaque catégorie, les deux valeurs sont additionnées en appliquant un facteur de pondération qui permet de donner plus ou moins d'importance à chaque méthode. Lors des différents tests, nous avons fait varier ces multiplicateurs entre 0,1 et 0,9 (avec des sauts de 0,1) en prenant soin que la somme des deux multiplicateurs soit toujours égale à 1.

3.4.2. Résultats

La synthèse des résultats obtenus pour les méthodes MLE et SVM, ainsi que pour les différentes approches et modes de combinaison de ces deux méthodes, est reprise au tableau 1. Pour l'approche 2 (fusion des listes complètes avant application du seuil), nous ne rapportons que le meilleur résultat parmi les combinaisons possibles de multiplicateurs.

On constate que, à l'exception de l'intersection des listes réduites par application d'un seuil, les combinaisons des deux approches débouchent sur des résultats supérieurs. La meilleure performance en termes de f-mesure est 66,08 (pour un rappel de 67,70% et une précision de 73,70%) et est atteinte en réalisant l'union des listes complètes fournies par les deux méthodes avant application de la fonction de seuil. Cela représente une augmentation de 5,06 par rapport à la méthode MLE et de 6,93 par rapport à SVM.

En principe, l'intersection favorise plutôt la précision au détriment du rappel, alors que l'union provoque les variations inverses. Pour l'approche 1, on constate que l'intersection des listes restreintes optimales donne des résultats en net recul par rapport aux deux méthodes de base. Cette détérioration peut s'expliquer par le faible nombre de catégories fournies par MLE (1,61) et SVM (1,05), qui lors de l'intersection, atteint un niveau très faible (0,67). Certains documents ne reçoivent donc pas de suggestion de catégorie ce qui pénalise le rappel. Le choix d'une unique mauvaise catégorie est également très lourd en termes de précision. L'union, avec une augmentation du rappel (à 76,16%) par rapport aux deux méthodes de base et une précision se stabilisant (à 64,96%) légèrement en dessous de MLE, se comporte comme prévu et permet d'atteindre une f-mesure plus élevée (65,13). La forte augmentation du rappel prouve que les catégories correctes comprises dans les deux listes sont en partie différentes.

L'approche 2, suite à la pondération, modifie les poids initiaux des catégories et donne l'opportunité à celles-ci de se réorganiser avant application du seuil. On remarque que les meilleurs résultats sont obtenus à l'aide d'une pondération forte de la méthode SVM. L'ordre et les poids attribués par cette méthode ont donc une grande importance sur le résultat final. L'union et l'intersection atteignent un niveau similaire de f-mesure, supérieur aux méthodes de bases. C'est une nouvelle fois l'union qui réalise la meilleure performance (66,08), alors que l'intersection suit de très près (66,01). Les résultats présentent un bon niveau de précision, ce qui est dû à la forte pondération de SVM. Comme prévu, l'union favorise plutôt le rappel (ici, peu élevé en raison du faible nombre de catégories présentées, soit 1,48) et l'intersection, la précision (en partie grâce à l'effet *filtre* de MLE).

Méthode	Rapport M1/M2	Rappel	Précision	F-mesure	Nb. catégories
MLE (M1)	n/a	64,79	66,05	61,02	1,61
SVM (M2)	n/a	53,93	72,90	59,15	1,05
Approche 1 : fusion des listes *optimales* de M1 et M2					
M1 UNION M2	n/a	76,16	64,96	65,13	1,99
M1 INTER M2	n/a	42,57	57,24	46,81	0,67
Approche 2 : fusion pondérée des listes *complètes* de M1 et M2 et application du seuil					
M1 UNION M2	0,1 / 0,9	67,70	73,70	**66,08**	1,48
M1 INTER M2	0,1 / 0,9	70,20	71,31	66,01	1,62

Tableau 1: Synthèse des résultats obtenus pour les méthodes 1 (MLE) et 2 (SVM), ainsi que pour les différents approches et modes de combinaison.

Ces mesures doivent être mises en perspective. Dans l'optique de l'indexation semi-automatique, nous pourrions augmenter le nombre de catégories proposées au documentaliste afin d'améliorer le rappel. Avec la meilleure méthode combinée, nous pourrions ainsi proposer en moyenne 3,46 catégories pour atteindre un rappel de 87,16% (précision de 50,06%). En acceptant de laisser chuter la précision à 33,50%, et en suggérant en moyenne 4,77 catégories, le rappel pourrait même augmenter jusqu'à 91,72%. La mise en avant du rappel s'accompagne d'un renversement progressif de la pondération vers la méthode MLE (0,3/0,7 dans un premier temps et 0,6/0,4 ensuite), qui démontre donc son apport sur ce point.

D'autre part, il faut également tenir compte du fait que notre évaluation a été effectuée par rapport à une indexation manuelle réalisée, pour chaque document, par une seule personne. Or, Van Slype (1987) montre que la cohérence de l'indexation d'un même document par deux documentalistes se situe entre 50% et 80%. De même, Pouliquen et al. (2003) rapportent un *accord inter-annotateur* allant de 78% à 87%. Certaines catégories étant parfois assez proches[14], on comprend aisément qu'une certaine variation existe dans les indexations des documentalistes. Dès lors, étant donné ce désaccord, on peut considérer que notre système peut difficilement atteindre 100% s'il est évalué par rapport au jugement d'un seul individu.

Nous avons également évalué, document par document, si la meilleure méthode combinée (l'union pour l'approche 2) offre une f-mesure plus élevée que les deux méthodes de base. Pour MLE, on constate au tableau 2 que les résultats restent inchangés pour 56,31% des documents, et que 15,03% subissent une détérioration de la f-mesure (en moyenne –39,21) alors que 28,65% bénéficient d'une meilleure analyse (en moyenne +38,26). Le résultat est

[14] A titre d'exemples, le thésaurus contient les microthésaurus « éducation » et « enseignement », « travail », « emploi » et « activité professionnelle » ou encore « politique internationale », « relations extérieures » et « organisations internationales ».

donc meilleur ou inchangé pour 84,96% des documents. En ce qui concerne la méthode SVM (*cf.* tableau 3), le nombre et la répartition des documents concernés par des variations sont assez semblables. On note cependant une proportion un peu plus grande de documents sans changement (61,78%). Les variations de performances sont un peu plus importantes, surtout à la hausse (+49,86 dans 24,74% des cas) mais aussi à la baisse (–40,06 dans 13,48% des cas). Au total, les résultats sont meilleurs ou inchangés pour 86,52% des textes.

Variation F-mesure	**Positive (>)**					**Egale (=)**			**Négative (<)**				
Variation Rappel	>			=	<	>	=	<	>	=	<		
Variation Précision	>	=	<	>	>	<	=	>	<	<	>	=	<
Nb. docs.	1.116	243	47	1.781	10	3	6.257	23	1	541	136	446	553
%	10,00	2,18	0,42	15,96	0,09	0,03	56,08	0,21	0,01	4,85	1,22	4,00	4,96
Total	**3.197 (28,65%)**					**6.283 (56,31%)**			**1.677 (15,03%)**				
Variation moyenne F-mesure	58,09	31,27	14,15	27,57	10,90	n/a	n/a	n/a	9,53	27,98	15,22	31,00	63,06
	38,26					**n/a**			**39,31**				

Tableau 2 : Analyse des variations de performance (> : augmentation, < : diminution, = : identique) entre la méthode 1 (MLE) et l'approche de combinaison 2, union.

Variation F-mesure	**Positive (>)**					**Egale (=)**			**Négative (<)**				
Variation Rappel	>			=	<	>	=	<	>	=	<		
Variation Précision	>	=	<	>	>	<	=	>	<	<	>	=	<
Nb. docs.	1.551	880	199	130	0	11	6.882	0	1	1.193	9	36	265
%	13,90	7,89	1,78	1,16	0	0,10	61,68	0	0,01	10,69	0,08	0,32	2,37
Total	**2.760 (24,74%)**					**6.893 (61,78%)**			**1.504 (13,48%)**				
Variation moyenne F-mesure	65,96	32,47	14,56	29,37	0	n/a	n/a	n/a	5,00	30,81	14,55	31,55	83,85
	49,86					**n/a**			**40,06**				

Tableau 3 : Analyse des variations de performance (> : augmentation, < : diminution, = : identique) entre la méthode 2 (SVM) et l'approche de combinaison 2, union.

4. Conclusion

Nous avons présenté deux méthodes de classification textuelle supervisée : l'une exploitant une méthode originale qui prend la forme de transducteurs et est nommée *motifs lexicaux étendus* (MLE), et l'autre mettant en œuvre des techniques d'apprentissage artificiel de type SVM. Ces deux méthodes ont donné des premiers résultats intéressants se situant, en termes de f-mesure, entre 59,15 (SVM) et 61,02 (MLE). Les deux méthodes présentant des caractéristiques assez différentes, nous avons montré que la combinaison des deux méthodes permettait d'améliorer la performance finale du système. Nous avons en effet obtenu des gains significatifs en atteignant une f-mesure de 66,08 (+5,06 pour MLE et +6,93 pour SVM).

Diverses perspectives de développements sont encore envisagées pour améliorer les deux méthodes de base. Pour MLE, citons entre autres la recherche automatique de synonymes pour l'enrichissement du thésaurus, l'amélioration des ressources pour les entités nommées, *etc*. Les deux méthodes pourraient également tirer parti de l'information de hiérarchie fournie par le thésaurus, par exemple en s'inspirant du travail de Mencia et Furnkranz (2008).

Remerciements

Ce travail a été effectué dans le cadre du projet STRATEGO (Wist n° 616442) financé par la Région wallonne (Belgique). Nous tenons également à remercier IRIS, notre partenaire industriel dans ce projet, pour leur précieuse collaboration.

Références

AFNOR (1981). Règles d'établissement des thésaurus monolingues. NF Z47-100.

Berry, M. W., and Browne, M. (1996). *Understanding Search Engines : Mathematical Model and Text Retrieval.* SIAM.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. In *Machine Learning*, 20(3), p. 273-297.

Duan, K.-B., and Keerthi, S. (2005). Which Is the Best Mutliclass SVM Method? An Empirical Study. In *Sixth International Workshop on Multiple Classifier Systems*. p. 278-285.

ISO (1986). Guidelines for the establishment and development of monolingual thesauri. ISO 2788.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, p. 137-142.

Kevers L. (2009). Indexation semi-automatique de textes : thésaurus et transducteurs. In *Actes de CORIA09 (Sixième Conférence Francophone en Recherche d'Information et Applications)*, Presqu'Île de Giens, France, p. 151-167.

Medelyan O. and Witten I. H. (2006). Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, ACM, Chapel Hill, NC, USA, p. 296-297.

Mencia, E. L., and Furnkranz, J. (2008). Algorithm for Large-Scale Problems in the Legal Domain. In *Proceedings of the European Conference on Machine Learning*, p. 50-65.

Névéol A., Mork J. G., Aronson A. R. and Darmoni S. J. (2005). Evaluation of French and English MeSH Indexing Systems with a Parallel Corpus. In *AMIA Annual Symposium Proceedings*, vol. 2005, p. 565-569.

Névéol A., Rogozan A. and Darmoni S. (2006). Automatic indexing of online health resources for a French quality controlled gateway. In *Information Processing and Management : an International Journal*, vol. 42 (3), p. 695-709.

Paumier S. (2008). Unitex 2.0 User Manual. http://www-igm.univ-mlv.fr/unitex/manuel.html.

Pereira S., Neveol A., Kerdelhué G., Serrot E., Joubert M. and Darmoni S. J. (2008). Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue. In *AMIA Annual Symposium Proceedings*, vol. 2008, p. 586-590.

Pouliquen B., Steinberger R. and Ignat C. (2003). Automatic annotation of multilingual text collections with a conceptual thesaurus. In *Proceedings of EUROLAN'2003*, Bucharest, Romania, p. 9-28.

Salton G., Fox E. A. and Wu H. (1983). Extended Boolean information retrieval. In *Communications of the ACM*, vol. 26 (11), p. 1022-1036.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, p. 44-49.

Sebastiani, F. (2002). Machine learning in automated text categorization. In *ACM Computing Surveys* vol. 34.1, p. 1-47.

Van Slype G. (1987). *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires*. Systèmes d'Information et de Documentation. Les éditions d'organisation.

Yarowsky D. (1993). One sense per collocation. In *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, Princeton, New Jersey, p. 266-271.