Indexation semi-automatique de textes : thésaurus et transducteurs

Laurent Kevers

Centre de traitement automatique du langage (CENTAL)
Université catholique de Louvain (UCL)
Collège Erasme - Place Blaise Pascal, 1
1348 Louvain-la-Neuve - Belgium
laurent.kevers@uclouvain.be

RÉSUMÉ. *Cet article présente une méthode de classification ne nécessitant pas de phase d'apprentissage. Son but est d'améliorer l'indexation manuelle des documents textuels, une opération souvent menée au sein de certains systèmes d'information requérant un niveau de précision élevé. Le système, qui apporte une aide à l'indexeur humain, est semi-automatique. Par analogie à la terminologie utilisée en apprentissage automatique, la méthode est dite supervisée car elle exploite une définition préalable des catégories d'indexation. Un vocabulaire contrôlé, par exemple un thésaurus, est utilisé comme la ressource de base servant à la génération automatique de transducteurs (ou automates). L'application de ceux-ci à un texte permet d'extraire un nombre limité d'expressions pertinentes, chacune accompagnée d'au moins un code de catégorie dont l'analyse finale permet la classification du document. Nos tests sur un corpus de textes en français ont permis d'obtenir une f-measure située entre 0,51 et 0,64.*

ABSTRACT. *This article presents a classification method without any learning stage. It can help to improve the manual indexation process of textual documents traditionally conducted in some high precision information systems. The described system is defined as semi-automatic as it will help the human indexing. By analogy with machine learning terminology, this method can be qualified as supervised as it uses a priori defined indexing categories. A controlled vocabulary, e.g. a thesaurus, is used as the main resource to automatically generate a set of transducers (or automata). The extraction of a document's significant phrases, each one comming with at least one corresponding class code, is obtained when using these transducers on the text. The final classification is obtained after analysis of phrases and codes. Testing results on a french text corpus are comprised between 0.51 and 0.64 for f-measure.*

MOTS-CLÉS : *catégorisation, classification supervisée, indexation, thésaurus, transducteur*

KEYWORDS: *categorization, supervised classification, indexing, thesaurus, transducer*

1. Introduction

Traditionnellement, l'indexation de textes au sein de grandes bases de données documentaires est réalisée à l'aide de mots clés, souvent issus d'un vocabulaire contrôlé. Cette indexation s'effectue généralement manuellement, ce qui apporte une grande précision au système. La cohérence de l'indexation peut cependant être diminuée par la variabilité des décisions au cours du temps et en fonction des différents indexeurs. D'autre part, la quantité de documents à traiter augmente rapidement et rend le processus manuel inabordable pour beaucoup d'organisations. Il est dès lors souvent 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 lorsqu'un très grand nombre de documents doit être analysé, elles ne constituent pas la solution idéale dans le contexte de systèmes d'information nécessitant une précision élevée.

Nous décrivons une méthode ayant une visée applicative réelle et concrète. Le but est d'améliorer l'efficacité et la cohérence de l'indexation manuelle en suggérant à l'indexeur une liste de catégories ou de mots clés potentiels automatiquement construite. L'approche adoptée est centrée sur l'utilisation d'une ressource de base - un thésaurus définissant les catégories - pour générer automatiquement des automates de reconnaissance, ou plus précisément des transducteurs (*cf.* section 4.3). Ceux-ci permettent d'exprimer des patrons contraints dépassant de loin les possibilités de la simple expression régulière (*cf.* section 6.1). L'application des transducteurs à un texte permet d'extraire un nombre limité d'expressions *pertinentes*, chacune accompagnée d'au moins un code de catégorie dont l'analyse finale permet la classification du document (*cf.* sections 6.2 à 6.4). L'indexation humaine se résume alors à la consultation sommaire du document - titre, résumé et éventuellement premier paragraphe - et à la sélection d'un ou plusieurs éléments dans la liste proposée et non plus dans le thésaurus complet. Afin de conserver ce gain de temps, il est évidemment très important que l'indexeur n'ait pas à consulter le texte en entier pour faire son choix. Par conséquent, la liste doit contenir un maximum de mots clés *plausibles* quitte à y inclure certaines propositions non pertinentes. Cette méthode semi-automatique à l'avantage d'améliorer la rapidité et la cohérence de l'indexation grâce à l'analyse automatique tout en conservant une précision élevée garantie par la validation humaine. L'approche, sans apprentissage, ne nécessite pas de données annotées manuellement et est fonctionnelle dès le premier document. Cette caractéristique ouvre des perspectives en terme d'interaction avec d'autres méthodes requérant un apprentissage.

Cet article introduit quelques concepts sur la classification et les thésaurus. Les hypothèses, les prérequis et les choix d'implémentation sont ensuite présentés. Ils sont suivis par un aperçu des travaux précédemment menés sur cette thématique. L'analyse de texte et la classification sont ensuite détaillées, avant de clôturer avec les résultats, perspectives et conclusions.

2. Classification

L'attribution d'index issus d'un vocabulaire contrôlé à un document est comparable au processus de classification de textes, domaine largement couvert par les techniques d'apprentissage automatique. Bien que notre méthode n'en fasse pas partie, nous allons brièvement en présenter la terminologie afin de pouvoir nous situer par rapport à ces travaux.

La classification de textes peut être divisée en deux activités distinctes : le clustering et la catégorisation. Le *clustering* regroupe les documents en ensembles de textes similaires sur la base du seul contenu de ces documents. La *catégorisation* exploite une ressource extérieure, les catégories à attribuer. Le clustering et la catégorisation sont aussi qualifiées respectivement de classification *non supervisée* et *supervisée*. Notre système s'apparente à la classification *supervisée* que nous appelerons désormais indifféremment classification ou catégorisation. Dans notre cas, le résultat attendu est un liste de catégories accompagnées de poids, ce qui serait qualifié de classification *soft* et *multi labels* en apprentissage automatique.

3. Thésaurus

Nous avons mentionné dans la section précédente que la classification s'appuie sur une ressource qui définit l'ensemble des catégories attribuables. Cette ressource peut être aussi simple qu'une liste de termes ou prendre la forme plus élaborée d'une ontologie ou d'un thésaurus. C'est souvent ce dernier qui est utilisé car il représente un bon compromis entre puissance descriptive et complexité acceptable du point de vue du développement et de la maintenance.

Un thésaurus est un vocabulaire contrôlé qui regroupe un ensemble de concepts relatifs à un certain domaine. Il constitue un moyen de décrire ce domaine, d'en définir les concepts et de fixer la terminologie utilisée par un groupe de personnes. Un concept est représenté par un terme principal appelé *descripteur* qui 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 un lien de similarité entre deux concepts. Les grands thésaurus peuvent être fragmentés en *microthésaurus* couvrant chacun un sous thème particulier. Plusieurs normes internationales dont (ISO, 1986) et (AFNOR, 1981) définissent plus précisément les thésaurus.

La portée d'un thésaurus peut être très large, par exemple Eurovoc[1], ou au contraire très spécialisée, tel que Agrovoc[2]. Ces deux thésaurus comptent un grand nombre de

1. Thésaurus du Parlement de la Communauté européenne, couvre une grande diversité de domaines, mais toujours en rapport avec le travail parlementaire : http ://europa.eu/eurovoc/
2. Thésaurus de l'Organisationn des Nations Unies pour l'alimentation et l'agriculture, se concentre sur l'agriculture : http ://www.fao.org/aims/ag_intro.htm

niveaux hiérarchiques et de descripteurs[3]. De nombreuses organisations se contentent cependant de vocabulaires de tailles plus modestes. Van Slype (Van Slype, 1987) préconise l'usage de 500 à 1500 descripteurs pour des bases de données ayant un accroissement de 10.000 documents par an et de 3000 à 6000 descripteurs si la base s'étend jusqu'à 100.000 documents par an.

De nombreuses applications en traitement automatique du langage peuvent tirer parti d'une ressource telle qu'un thésaurus (Da Sylva, 2006). Par exemple, l'expansion de requêtes pour les moteurs de recherche, la désambiguisation lexicale ou encore la traduction automatique. En ce qui nous concerne, nous allons utiliser le thésaurus comme base du processus de classification : chaque descripteur, identifié par son code et accompagné par ses synonymes, constituera une catégorie.

4. Hypothèses, prérequis et choix d'implémentation

4.1. *Hypothèses*

Nous pensons que l'appartenance d'un texte à une catégorie thématique se matérialise dans le document par l'utilisation d'un certain nombre de mots. Dès lors, si les catégories sont correctement définies (elles doivent posséder un terme descripteur principal et de préférence autant de non-descripteurs qu'il existe de synonymes), il est possible de trouver une intersection suffisante entre le vocabulaire du document et la définition des catégories à sélectionner pour décider de manière automatique de leur assignation.

Nous adhérons également au principe qu'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. On utilise par exemple rarement le terme *allocations* seul, mais plutôt dans une forme composée telle que *allocations de chômage* ou *allocations familliale*. De plus, les expressions composées sont souvent moins polysémiques (Yarowsky, 1993). Par conséquent, notre système ne doit pas se limiter aux mots simples et doit prendre en compte les unités polylexicales.

A partir de ces hypothèses, nous voulons montrer qu'il est possible de mettre en œuvre une analyse performante ne requérant pas de phase d'apprentissage pour la classification/indexation de textes. Cette analyse est basée sur des principes simples et utilise une ressource lexicale et sémantique telle qu'un thésaurus. Celui-ci doit être de qualité, c'est-à-dire que les descripteurs soient correctement organisés, qu'ils soient au moins accompagnés de leurs synonymes les plus courants et qu'ils ne soient pas trop abstraits.

3. Eurovoc : 6,645 pour chaque langue ; Agrovoc : 28,718 en anglais uniquement

4.2. *Prérequis*

L'implémentation fait appel à diverses techniques de traitement automatique du langage exploitant des ressources linguistiques. Le principal prérequis pour l'utilisation de cette méthode est l'existence d'une description des catégories utilisées pour l'indexation, tel qu'un thésaurus. Cette limitation est peu contraignante en pratique puisque de nombreuses organisations utilisent ce genre de ressource.

4.3. *Choix d'implémentation : extraction ou assignation de mots clés*

La première approche, l'*extraction de mots clés*, part du texte. Elle consiste à extraire des mots simples ou composés porteurs de sens. L'extraction s'appuie sur des critères lexicaux et grammaticaux. L'appartenance au thésaurus des termes extraits d'un texte permet de dériver les catégories potentiellement attribuables à celui-ci. Beaucoup de ces termes sont cependant comparés sans succès au thésaurus.

La seconde approche, l'*assignation de mots clés*, démarre du thésaurus et consiste à créer une ressource d'extraction à l'aide des termes descripteurs, accompagnés de leur synonymes. Chaque descripteur correspond à une catégorie. Cette ressource est appliquée aux textes afin de retrouver un maximum d'expressions dites *pertinentes*, c'est-à-dire ayant *a priori* un intérêt pour la classification car dérivées du thésaurus, en délaissant les expressions ne possédant *a priori* pas de pouvoir classifiant.

Pour extraire les candidats, ne retenir que ceux qui sont potentiellement intéressants et enfin les confronter au thésaurus, l'extraction de mots clés nécessite un grand nombre d'heuristiques et de règles, difficiles à développer et à maintenir. Au contraire, l'assignation réalisée à partir d'une ressource dérivée du thésaurus constitue une méthode plus simple. Elle se limite à analyser un ensemble restreint d'expressions *pertinentes*, ce qui est plus efficace. Nous avons donc choisi d'utiliser la seconde approche.

La ressource d'extraction est constituée par des transducteurs et est capable de reconnaître des unités formées d'un nombre variable et non limité de tokens. Le formalisme des automates et transducteurs a été choisi car il dépasse de loin la simple recherche *mot à mot*. Il est utilisé pour un grand nombre d'analyses effectuées en traitement automatique du langage (Crochemore *et al.*, 1994). Il nous permet d'exprimer le contenu du thésaurus au moyen d'expressions régulières complexes, de codes grammaticaux et sémantiques[4], d'y insérer certaines contraintes sur le contexte lexico-syntaxique ou encore d'autoriser des insertions facultatives à certains endroits. La manière dont ces possibilités sont exploitées pour adapter le thésaurus en une ressource d'extraction la plus performante possible est exposée plus en détail à la section 6.1 et un exemple de graphe[5] est présenté à la figure 1. L'utilisation des transducteurs

4. Après application des dictionnaires électroniques.
5. Les transducteurs sont représentés graphiquement sous la forme de graphes. Un graphe contient toujours un état initial et un état final et peut être accompagné d'un ensemble de chemins concurents constitués de transitions étiquettées ainsi que des sorties (ou transductions).

sur un texte résulte en une liste d'expressions, chacune accompagnée d'une indication d'appartenance à une ou plusieurs catégories du thésaurus. La pondération de ces expressions permet ensuite de trier les résultats présentés à l'indexeur humain.

5. Travaux apparentés

Comme nous l'avons déjà mentionné, la classification (semi-) automatique est un domaine dans lequel les techniques d'apprentissage automatique sont souvent appliquées. Ces techniques ne nous intéressent pas directement dans le cadre de ce travail. Le lecteur intéressé pourra en trouver une introduction en consultant une référence telle que (Sebastiani, 2002), (Moens, 2006), ou encore (Baeza-Yates *et al.*, 1999).

Bien que l'usage d'un thésaurus pour améliorer ou guider la classification ne soit pas l'approche la plus répandue, certains travaux y sont cependant apparentés. KEA++ (Medelyan *et al.*, 2006) est un système agissant en deux phases : l'extraction de mots clés et leur filtrage à l'aide d'un thésaurus sont suivis par une étape d'apprentissage capable de mettre en œuvre différents types d'algorithmes. Pouliquen *et al.* (Pouliquen *et al.*, 2006) 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.* (Névéol *et al.*, 2005) évaluent deux systèmes hybrides - MTI pour l'anglais, MAIF pour le français - d'indexation de documents médicaux dans MeSH[6]. Ces systèmes combinent à la fois une approche de traitement automatique du langage 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.* (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.* (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 une série de transducteurs. Il existe cependant une différence de taille puisque ceux-ci sont construits manuellement en collaboration avec des experts alors que nous proposons de les générer automatiquement.

6. Analyse du texte et classification

Notre système de classification nécessite la production d'une version du thésaurus sous la forme de transducteurs, mieux adaptés au traitement automatique des textes (*cf.* section 6.1). Cette opération est unique et ne fait pas à proprement parler du processus de classification répété pour chaque document. Elle se doit d'être complètement automatique en raison du nombre élevé d'éléments que peut contenir un thésaurus.

6. Medical Subject Headings : http ://www.nlm.nih.gov/mesh/

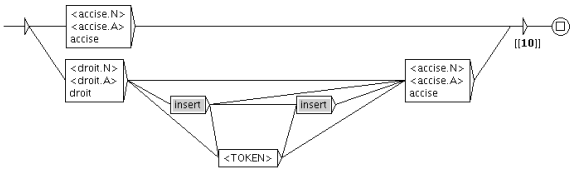

Figure 1. *Transducteur lemmatisé (descripteur : accise ; synonyme : droit d'accise ; catégorie de code 10).*

Une fois cette ressource à disposition, son application aux textes (*cf.* section 6.2) donne une liste d'expressions et de termes *pertinents*, accompagnés des catégories auxquels ils sont reliés. Pour chaque expression, une pondération est calculée selon plusieurs critères (*cf.* section 6.3). Les poids sont ensuite additionnés de manière à obtenir une valeur globale pour chaque catégorie représentée dans le texte. Cette liste pondérée permet finalement de sélectionner les catégories les plus significatives pour le document (*cf.* section 6.4).

6.1. *Du thésaurus aux transducteurs*

Les transducteurs sont automatiquement générés à partir du thésaurus dans un format compatible avec le logiciel de traitement de corpus Unitex[7] (Paumier, 2008). Chaque catégorie est représentée par un automate contenant une transduction qui renseigne le code de la catégorie. Les transducteurs générés sont rassemblés en un transducteur principal. La figure 1 illustre un transducteur généré automatiquement.

Nous avons mené deux types d'expériences différentes. Pour la première, les catégories sont réduites aux grandes divisions matérialisées par les microthésaurus. Un transducteur contient l'ensemble des descripteurs et synonymes reliés à un microthésaurus particulier. Ce regroupement constitue une sorte de généralisation ou simplification qui permet de réduire le nombre de catégories. La seconde expérience, conserve toutes les catégories. Un transducteur est donc généré par descripteur. Dans ce cas nous obtenons de petits transducteurs en très grand nombre alors que le premier cas débouche sur des transducteurs plus volumineux mais en nombre plus restreint. Les paragraphes suivants détaillent les quatre principaux traitements nécessaires à la production des patrons qui seront ensuite retranscrits sous la forme de transducteurs.

Le premier traitement est une étape de généralisation dont le but est d'étendre la couverture aux variations possibles d'une expression, telle que le passage du singulier au pluriel. Par exemple, à partir de l'expression *taux d'intérêt légal* issue du thésaurus, nous désirons aussi retrouver les formes *taux d'intérêt**s** légal*, *taux d'intérêt lég**aux***, ou

7. http ://www-igm.univ-mlv.fr/ unitex/

encore *taux d'intérêt**s*** *lég**aux***[8]. Deux techniques permettent d'atteindre ce résultat : le stemming et la lemmatisation. Le *stemming* consiste en l'extraction d'un préfixe correspondant à la racine d'un mot. Nous avons utilisé l'implémentation Snowball[9] de l'algorithme de Porter (Porter, 1997). Cette approche produit des transducteurs principalement composés d'expressions régulières. La *lemmatisation* permet de relier une forme fléchie à sa forme canonique (ou lemme). Les patrons générés font alors directement référence aux lemmes et non plus à des formes fléchies, ce qui permet de tirer parti de la puissance des dictionnaires électroniques disponibles dans Unitex. Pour obtenir le lemme, nous avons utilisé Treetagger[10], un étiqueteur morpho-syntaxique multilingue (Schmid, 1994). C'est cette dernière approche qui a finalement été choisie en raison des meilleurs résultats préliminaires obtenus. De plus, l'utilisation d'Unitex nous a permis de constater que le temps d'exécution des transducteurs constitués d'expressions régulières est de loin plus élevé que celui obtenu avec les transducteurs lemmatisés.

Le deuxième traitement consiste, comme dans de nombreux travaux en recherche d'information, à éliminer les *stopwords* (*mots vides*). Ils ont en réalité été 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 dans *contrôle **de** chômeurs*, *contrôle **du** chômeur* et *contrôle **des** chômeurs*[11].

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 possède une forme complète et une forme simplifiée ou qu'elle soit modulée par des adjectifs. C'est par exemple le cas pour *agence de protection de l'environnement* qui peut être retrouvé sous la forme d'*agence **<ADJ>** de proctection de l'environnement*, avec <ADJ> tel que *fédérale*, *régionale*, *belge*...

Enfin, le quatrième traitement concerne certaines exceptions qui doivent être prises en compte. Par exemple l'acronyme *CAS* qui correspond à *Caisse d'allocations sociales* est ambigu avec le nom *cas*. La casse ne peut être prise en compte car certains thésaurus sont encodés complètement en majuscules, ce qui empêche l'exploitation de ce critère. Insérer cet acronyme tel quel conduirait à reconnaitre sa présence à chaque occurence du nom commun, ce qui fausserait fortement l'analyse. Dans ce cas, nous nous limitons donc à utiliser l'acronyme dans sa version avec points (*C.A.S.*). Autre exception, suite au traitement des *stopwords*, certains patrons se résumeraient normalement à une seul méta-étiquette telle que `<TOKEN>`. Comme cela conduirait à la reconnaissance de toutes les unités du texte, nous gardons alors la forme d'origine lemmatisée avec la restriction supplémentaire de la présence d'un déterminant.

---

8. De telles ressources ont une tendance à la surgénération. Dans un contexte de reconnaissance, cela ne représente pas un problème. Au contraire, cela permet de reconnaître les occurences mal orthographiées ou s'éloignant de la norme.
9. http ://snowball.tartarus.org/
10. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
11. Ce cas n'est pas couvert par la lemmatisation car le lemme de *des* est *un*

Forme	Thésaurus	Substitut
art.	ART (arts plastiques)	*article*
livre (section de texte)	LIVRE (ouvrage)	*livreSection*
titre (section de texte)	TITRE (finance)	*titreSection*
au titre de ...	TITRE (finance)	
à titre ...	TITRE (finance)	
à juste titre	TITRE (finance)	
j'ai (nous avons) l'honneur de ...	HONNEUR	*je*

Tableau 1. *Normalisation des mots ambigus d'un texte.*

6.2. *Application des transducteurs aux documents et catégorisation*

Avant l'application des transducteurs générés à partir du thésaurus aux documents, une étape de prétraitement des textes est nécessaire. Lors de la création des transducteurs et du traitement des *stopwords*, les formes élidées telles que *l'* ont été remplacée par une méta-étiquette, par exemple `<TOKEN>`. Or Unitex crée, pour cette forme *l'*, deux tokens : *l* et *'*. Cette forme n'est donc plus reconnaissable par une seule étiquette `<TOKEN>` mais bien par deux. Afin d'éviter ce problème, un transducteur de prétraitement remplace toutes les formes élidées par une forme complète correspondante, par exemple *le* pour *l'*.

Une procédure de désambiguïsation ciblée est également souhaitable afin d'éviter certaines erreurs récurentes. L'expression *art. 2* (article 2) est par exemple interprétée comme reliée à la catégorie `ART` (arts plastiques, etc.) du thésaurus. D'autres cas sont repris au tableau 1. L'idéal est de réaliser une étude exhaustive des termes posant un problème d'ambiguïté dans le thésaurus. Evidemment, cette tâche n'est pas complètement automatisable et est spécifique à un thésaurus et à une langue en particulier. Afin de minimiser l'effort nécessaire, on peut cependant envisager de mener cette étude lors de la construction même du thésaurus, qui mobilise de toutes façons les compétences de spécialistes en terminologie. Pour les thésaurus existants, il est nécessaire de mettre au point une méthode de détection de la polysémie permettant de repérer les cas problématiques et requérant une intervention. Ce point fait partie des développements futurs que nous envisageons.

Finalement, d'autres tâches de prétraitement plus classiques sont réalisées. Le texte est désaccentué, car certains thésaurus sont complètement capitalisés et non accentués, ce qui est le cas de celui utilisé pour nos expériences. La suite du processus est réalisé au moyen d'Unitex : tokenisation, application des dictionnaires et application des transducteurs issus du thésaurus. Le résultat obtenu se présente sous la forme d'un index de mots ou d'expressions tel qu'illustré à la figure 2[12].

12. Les deux premières colonnes indiquent les numéros des tokens délimitant l'expression

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 retrouvées à l'aide des transducteurs pour un texte. Le code de catégorie (ici des microthésaurus) est inclus entre crochets.*

6.3. *Pondération*

Sur la base de la liste construite après application des transducteurs au texte (*cf.* Figure 2), un poids est calculé pour chaque expression et ensuite globalement pour chaque catégorie. Cette pondération est basée sur une mesure de fréquence mais d'autres critères sont aussi pris en compte. Ils sont implémentés par des multiplicateurs appliqués au poids initial. La recherche de leurs valeurs optimales a été effectuée empiriquement avec un nombre limité de valeurs[13].

La valeur de base pour la pondération est TF.IDF (*term frequency-inverse document frequency*). Cette mesure est couramment utilisée pour évaluer le poids d'un terme par rapport à un corpus donné. Ce score de base sera éventuellement modifié pour déterminer le score final d'une expression. Les formules appliquées sont :

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \qquad idf_i = log\frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

où n_{ij} est la fréquence d'un terme i dans le document d_j, $|D|$ étant le nombre de documents dans le corpus et $|\{d_j : t_i \in d_j\}|$ le nombre de documents dans lesquels le terme i est présent. La valeur finale du TF.IDF est obtenue par : $tf.idf_{ij} = tf_{ij} * idf_i$ Le but de cette mesure est de donner plus d'importance aux mots très fréquents dans un document, mais rares à l'échelle du corpus. Chaque expression de la liste obtient donc un poids TF.IDF. Les valeurs IDF sont précalculées sur le corpus en appliquant les transducteurs de reconnaissance issus du thésaurus. Cette méthode peut être perçue comme un biais, mais il s'agit d'une approximation raisonnable des scores IDF qui seraient graduellement construits lors du traitement des mêmes documents en situation réelle.

Le deuxième critère est basé sur le fait que les informations importantes apparaissent souvent au début du document, c'est-à-dire principalement dans le titre et le résumé s'il existe. Nous avons donc introduit un multiplicateur qui est appliqué au

13. Pour chaque multiplicateur, les valeurs testées sont 1, 2, 5, 10, 20, 50 et 100. Toutes les combinaisons ont été testées.

score de base (TF.IDF) si l'expression se situe dans le titre. Pour nos expériences, ce multiplicateur a été fixé empiriquement à 100.

Le troisième critère exploite l'intérêt particulier soulevé pour les expressions composées. Bien que cette caractéristique soit déjà indirectement prise en compte dans la mesure du TF.IDF, nous avons prévu un multiplicateur supplémentaire pour en augmenter le score. Sa valeur, 2, a été fixée de manière empirique.

Le quatrième et dernier critère envisagé concerne les entités nommées. Celles-ci sont détectées à l'aide de transducteurs spécifiques lors de la phase de prétraitement du texte. Le multiplicateur relié à ce critère a été fixé empiriquement à 2.

Après l'application de ces multiplicateurs, un score final est calculé pour chaque catégorie représentée dans le document afin d'obtenir une liste ordonnée des meilleurs propositions.

6.4. *Définition d'un seuil de sélection*

La liste pondérée obtenue peut être assez longue et les différences de poids importantes. Nous désirons donc réduire cette liste afin de ne garder que les candidats les plus probables. Cette sélection est opérée au moyen d'un seuil.

La première méthode (*k-first*) consiste simplement à conserver les k premières catégories correspondant aux meilleurs scores. Deux autres méthodes consistent à calculer la moyenne arithmétique des poids de catégories (*averaged pivot*) ou la moitié du score maximal (*middle pivot*). Ces valeurs centrales constituent un premier seuil. A partir de celui-ci, d'autres valeurs plus grandes ou plus petites sont ensuite obtenues par sauts de taille fixe. Pour obtenir x seuils plus élevés, on augmente ainsi x fois la valeur centrale de $\frac{val.max. - val.centrale}{x}$. De même, on diminue à x reprises la valeur centrale de $\frac{val.centrale}{x}$ pour obtenir x valeurs de seuil inférieures. Nous avons fixé x à 10, ce qui donne 21 valeurs de seuil au total. Notons que la première méthode produit toujours k propositions par point alors que les deux autres en retournent un nombre variable. Le but final est de déterminer quel type de seuil serait le plus approprié dans un environnement applicatif réel.

7. Expériences et résultats

Nos tests ont été réalisés sur un corpus de textes en français. La méthode peut être adaptée assez aisément pour une autre langue à partir du moment où les ressources suivantes sont disponibles : un dictionnaire électronique général, une liste de *stopwords*, un étiqueteur morpho-syntaxique ainsi que le graphe de prétraitement de l'ambiguïté (non indispensable, mais préférable). Bien entendu, un thésaurus approprié au corpus est toujours nécessaire. Nous n'avons pas à ce jour mené nos tests sur une autre langue que le français. Il s'agit cependant d'une de nos priorités pour le futur.

7.1. *Description du corpus et du thésaurus*

Les documents et le thésaurus proviennent de la base documentaire d'une organisation actuellement en activité. Les textes sont des documents relevant du domaine législatif et parlementaire. Le thésaurus a été spécialement conçu pour l'indexation de ces documents au sein de cette organisation.

Le thésaurus contient 2514 descripteurs et 2362 synonymes. Les decripteurs sont répartis en 47 microthésaurus. Le nombre de niveaux hiérarchiques monte jusqu'à 6, mais s'établit plus fréquemment entre 2 et 4. Les expressions composées sont bien représentées : 66,59% des descripteurs (1674 sur 2514) et 61.85% des synonymes (1461 sur 2362).

Notre corpus de test compte 12.734 fichiers XML contenant 32.953.724 mots[14]). La taille moyenne d'un document se situe par conséquent à 2588 mots. Le titre du document est délimité à l'aide de balises particulières. Pour chaque document, on dispose des catégories assignées manuellement par des indexeurs professionnels en situation réelle. Ces informations nous serviront de référence pour l'évaluation. Le nombre de descripteurs attribués varie entre 1 et 37, la valeur moyenne étant de 1,92. Dans ce corpus, certaines catégories du thésaurus 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 utilisées par 9,71 documents.

7.2. *Description des expériences*

Nous avons utilisé le système décrit dans cet article sur l'ensemble des documents à notre disposition. Un test préliminaire a permis d'évaluer les performances de classification à l'aide de transducteurs ne contenant que les formes telles qu'elles apparaissent dans le thésaurus, sans aucune transformation. Le poids de chaque mot est uniquement fonction de sa fréquence d'apparition, sans intervention d'un facteur IDF ou de multiplicateurs. Deux expériences ont ensuite été menées : l'une avec les transducteurs générés au niveau hiérarchique des microthésaurus (47 catégories possibles), et l'autre à l'aide des transducteurs générés pour tous les descripteurs (2514 catégories possibles). Quelques catégories ont été interdites de sélection. C'est par exemple le cas pour le microthésaurus - et pour toutes les catégories qu'il contient - concernant les expressions temporelles. Celles-ci sont en effet souvent seulement constituées d'une indication d'année, ce qui en fait un élément très ambigu. Nous proposons plutôt d'exploiter ces informations lors d'une analyse séparée. Cette tâche, dont nous ne nous sommes pas occupé, pourra constituer une extension de ce travail.

---

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

7.3. *Mesures*

Pour évaluer nos résultats, nous avons employé les mesures classiques de précision (P), de rappel (R) et de f-measure (F).

$$P = \frac{Syst_{OK}}{Syst_{TOT}} \quad R = \frac{Syst_{OK}}{Man_{OK}} \quad F = \frac{2 * P * R}{P + R}$$

où $Syst_{OK}$ est le nombre de catégories correctement proposées par le système, Man_{OK} est le nombre de catégories attribuées manuellement par l'indexeur humain et $Syst_{TOT}$ est le nombre total de catégories proposées par le système. La f-measure est une combinaison à proportion égale de la précision et du rappel.

Nous avons choisi de calculer les résultats globaux du système selon une approche microscopique. La précision, le rappel et la f-measure sont donc calculés pour chaque document et les mesures finales sont obtenues au moyen d'une moyenne arithmétique de ces valeurs. Ce choix est motivé par l'application visée qui consiste à traiter les documents un à un et non globalement.

7.4. *Résultats et perspectives*

Les résultats que nous rapportons doivent être interprétés avec précautions, car il s'agit d'une évaluation effectuée par rapport à une indexation manuelle réalisée, pour chaque document, par une seule personne. Van Slype (Van Slype, 1987) montre que la cohérence de l'indexation d'un même document par deux indexeurs se situe entre 50% et 80%. De même, Pouliquen *et al.* (Pouliquen *et al.*, 2006) rapportent un accord inter-annotateur allant de 78% à 87%. Etant donné ce désaccord, on peut considérer que notre système peut difficilement atteindre les 100% s'il est évalué par rapport à l'annotation d'une seule personne. Une évaluation plus correcte devrait être effectuée en comparant nos résultats avec les indexations de plusieurs personnes. Une autre possibilité consisterait à faire vérifier à la main les propositions de notre système afin de déterminer parmi les *mauvais* descripteurs proposés lesquels auraient pu être sélectionnés par une autre personne et lesquels auraient été jugés totalement inappropriés. Malheureusement, ce type d'évaluation nécessite une mobilisation de spécialistes qu'il est souvent difficile d'obtenir et nous ne pouvons garantir que nous pourrons mener ce type d'évaluation dans le futur.

Pour les différents tests, les trois méthodes de sélection par seuil ont été calculées. Le test préliminaire a été effectué sur l'ensemble des 2514 catégories. La recherche des expressions d'origine non modifiées et dont la fréquence n'a pas été pondérée a abouti à une f-measure maximale de 23,83% (rappel=31,65% et précision=19,11%). Le meilleur rappel atteint se situe à 52,80% mais il est accompagné par une précision très faible (6,91%). En ce qui concerne les deux expériences principales, trois points ont été mis en évidence et sont repris dans le tableau 2 : celui qui obtient la meilleure f-measure, celui qui obtient le meilleur rappel pour une précision *acceptable* d'environ 30% et enfin celui qui obtient le rappel maximal.

	47 catégories			2514 catégories		
	k-first	averaged pivot	middle pivot	k-first	averaged pivot	middle pivot
Meilleure f-measure						
Nbr. de cat.	2	1,8	**1,9**	2	1,9	**2,3**
F-measure	0,5743	0,6362	**0,6431**	0,4427	0,5066	**0,5117**
Rappel	0,6789	0,6555	**0,6785**	0,4990	0,5009	**0,5296**
Précision	0,4976	0,6180	**0,6113**	0,3978	0,5123	**0,4949**
Meilleur rappel avec précision à +/- 30%						
Nbr. de cat.	4	6,4	**5,6**	3	**10,1**	4
F-measure	0,4523	0,4516	**0,4714**	0,4004	**0,4141**	0,4799
Rappel	0,8119	0,8630	**0,8587**	0,5610	**0,6291**	0,5876
Précision	0,3135	0,3058	**0,3248**	0,3113	**0,3086**	0,4056
Meilleur rappel						
Nbr. de cat.	21	15,1	**15,1**	21	38,8	**38,8**
F-measure	0,2424	0,2354	**0.2354**	0,1694	0,1450	**0,1450**
Rappel	0,9077	0,9101	**0,9101**	0,6890	0,7086	**0,7086**
Précision	0,1399	0,1352	**0,1352**	0,0966	0,0807	**0,0807**

Tableau 2. *Résultats des test de classification.*

On remarque que la méthode *k-first* est significativement moins bonne que pour les deux autres. L'inconvénient présenté par cette méthode est de toujours proposer le même nombre de descripteurs, quel que soit le texte. Les autres méthodes, basées sur une valeur moyenne, adaptent automatiquement le nombre de propositions retournées en fonction du nombre total de descripteurs dans la liste complète, et surtout en fonction de leurs scores. Ces méthodes dynamiques sont bien entendu plus adaptées étant donné le nombre variable de catégories attribuées par les indexeurs humains. Comme ces deux méthodes donnent des résultats relativement similaires et sauf indication contraire, nous n'allons détailler les résultats que par rapport à la méthode *middle pivot*.

Dans le cas de la classification sur les 47 microthésaurus, les meilleurs résultats en terme de f-measure sont obtenus avec une valeur de 64,31%. La rappel obtenu est 67,85% pour une précision de 61,13%. Le nombre moyen de catégories proposées est 1,9 (sur 47 catégories possibles). Pour l'application visée, nous sommes intéressés de savoir quel rappel nous pouvons obtenir en acceptant une précision moindre. Un rappel de 85,87% est atteint en conservant une précision *acceptable* de 32,48% (f-measure : 47,14%). En moyenne, l'indexeur humain disposerait de 5,6 catégories. Enfin, le meilleur taux de rappel obtenu se situe à 91,01% pour une précision de 13.52%, une f-measure de 23.54% et un nombre moyen de catégories proposées de 15,1.

La classification sur l'ensemble des 2514 catégories donne des résultats moins élevés. Cela s'explique aisément par le nombre bien plus important de catégories. Par la généralisation qu'elle implique, la classification dans les 47 microthésaurus permet

d'éviter une série d'erreurs telles que la classification dans une catégorie sœur ou dans une catégorie mère/fille. La meilleure f-measure est obtenue à 51,17%. Le rappel se situe alors à 52,96% et la précision atteint 49,49%. Le nombre moyen de catégories proposées est 2,3 (sur 2514 catégories possibles). Pour rappel, l'attribution de descripteurs dans notre corpus varie entre 1 et 37, la valeur moyenne étant de 1,92. La méthode de sélection par seuil *middle pivot* ne nous a pas fourni de valeur de précision avoisinant les 30%. La méthode *averaged pivot* indique par contre que pour une précision *acceptable* de 30,86% , le rappel peut atteindre 62,91% (f-measure : 41,41%). Le nombre moyen de catégories proposées à l'indexeur humain est de 10,1. Enfin, le meilleur taux de rappel obtenu se situe à 70,86% pour une précision de 8,07%, une f-measure de 14,5% et un nombre moyen de catégories proposées de 38,8.

Deux causes sont principalement responsables des catégories non trouvées : l'absence de synonymes dans le thésaurus et l'utilisation, dans le texte, de termes trop ou pas assez concrets en regard du thésaurus. La polysémie de certains termes génère quant à elle du bruit.

A titre de comparaison, nous avons testé KEA++ sur le même jeu de données. Les résultats obtenus sont assez décevants, surtout en terme de précision[15]. Ces chiffres sont à prendre avec précaution car ils sont très éloignés de ceux rapportés par les auteurs[16] (Medelyan *et al.*, 2005).

L'absence d'apprentissage est un point intéressant de notre méthode, qui permet de la positionner comme une solution adéquate lorsque le nombre de documents disponibles pour l'apprentissage n'est pas suffisant. Il est également imaginable de l'utiliser en tant que processus d'amorçage permettant la production de l'ensemble de documents annotés nécessaires aux algorithmes d'apprentissage. Nous pensons qu'il est également possible d'exploiter cette méthode seule ou en combinaison avec d'autres techniques. Un système hybride pourait être imaginé selon deux modes différents. Tout d'abord, notre méthode peut produire un nombre restreint d'expressions *pertinentes* pouvant être ensuite exploitées par des techniques d'apprentissage. C'est une approche qui ressemble fort à celle présentée par KEA++. La seconde possibilité est l'analyse en parallèle avec d'autres systèmes de classification, conduisant à une combinaison finale des résultats obtenus par les diverses méthodes. Ces deux modes de collaboration ont été sommairement testés. Les premiers résultats ne montrent pas d'amélioration significative lors de l'utilisation d'algorithmes SVM suite à notre système, alors que la combinaison en parallèle des deux techniques semble pouvoir mener à un gain de l'ordre d'environ 5%. Ces possibilités de développement constitueront un axe important pour notre travail futur.

Ces résultats sont donc encourageants, d'autant que plusieurs évolutions doivent encore être apportées. Parmi les point importants à développer, nous avons déjà cité la mise au point d'une méthode d'analyse de l'ambiguïté pour le thésaurus et son extension automatique à l'aide de diverses ressources. D'autres améliorations de la méthode

15. Pour un rappel compris entre 0,48 et 0,53, la précision n'a atteint que 0,10.
16. Rappel=0,53 ; Précision=0.48 ; F-measure=0,47

sont encore envisageables : prise en compte de la structure hiérarchique du thésaurus, reconnaissance des expressions temporelles, correction orthographique, amélioration de la pondération et des méthodes de sélection par seuil. Enfin, il nous semble indispensable de tester ce système à la fois sur un autre corpus et sur une langue différente.

8. Conclusion

Nous avons présenté une méthode semi-automatique permettant d'améliorer la rapidité et la cohérence de l'indexation manuelle de textes. Cette méthode est basée, comme pour le processus manuel, sur un thésaurus qui décrit le domaine d'indexation. Ce thésaurus a été converti automatiquement sous la forme de transducteurs qui, appliqués aux textes, repèrent les expressions *pertinentes* et y adjoignent la catégorie du thésaurus correspondante. Après pondération des expressions, on obtient finalement pour chaque document, un score total pour chaque catégorie représentée dans la liste. Un système de sélection par seuil permet ensuite de réduire le nombre de catégories proposées en ne gardant que les plus probables. L'indexeur humain peut alors faire son choix parmi cette liste réduite. L'évaluation a été effectuée sur des données en français : un ensemble de textes indexés manuellement par des indexeurs humains à l'aide d'un thésaurus ad-hoc. Les résultats encourageants obtenus (une f-measure située entre 0,51 et 0,64 selon les tests), les possibilités de développements futurs, l'absence d'apprentissage et la possibilité de démarrer une analyse sur un ensemble restreint de documents permettent d'envisager ce système en tant que méthode principale d'indexation, mais aussi en tant que méthode préliminaire ou parallèle à d'autres méthodes.

Remerciements

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

9. Bibliographie

AFNOR, « Règles d'établissement des thésaurus monolingues », December, 1981. NF Z47-100.

Baeza-Yates R., Ribeiro-Neto B., *Modern Information Retrieval*, 1st edn, Addison Wesley, May, 1999.

Crochemore M., Rytter W., *Text Algorithms*, Oxford University Press, October, 1994.

Da Sylva L., « Thésaurus et systèmes de traitement automatique de la langue », *Documentation et bibliothèques*, vol. 52, p. 149-156, 2006.

ISO, « Guidelines for the establishment and development of monolingual thesauri », 1986. ISO 2788.

Medelyan O., Witten I. H., « Thesaurus-Based Index Term Extraction for Agricultural Documents », *6th Agricultural Ontology Service (AOS) workshop at EFITA/WCCA 2005*, Vila Real, Portugal, 2005.

Medelyan O., Witten I. H., « Thesaurus based automatic keyphrase indexing », *6th ACM/IEEE-CS joint conference on Digital libraries*, ACM, Chapel Hill, NC, USA, p. 296-297, 2006.

Moens M.-F., *Information Extraction : Algorithms and Prospects in a Retrieval Context*, 1 edn, Springer, October, 2006.

Névéol A., Mork J. G., Aronson A. R., Darmoni S. J., « Evaluation of French and English MeSH Indexing Systems with a Parallel Corpus », *AMIA Annual Symposium Proceedings*, vol. 2005, p. 565-569, 2005. PMC1560460.

Névéol A., Rogozan A., Darmoni S., « Automatic indexing of online health resources for a French quality controlled gateway », *Inf. Process. Manage.*, vol. 42, p. 695-709, 2006.

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

Pereira S., Neveol A., Kerdelhué G., Serrot E., Joubert M., Darmoni S. J., « Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue », *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*p. 586-90, 2008. PMID : 18998933.

Porter M. F., *An algorithm for suffix stripping*, Morgan Kaufmann Publishers Inc., p. 313-316, 1997.

Pouliquen B., Steinberger R., Ignat C., « Automatic annotation of multilingual text collections with a conceptual thesaurus », *cs/0609059*, September, 2006. Proceedings of the Workshop 'Ontologies and Information Extraction' at the Summer School 'The Semantic Web and Language Technology - Its Potential and Practicalities' (EUROLAN'2003), pp 9-28. Bucharest, Romania, 28 July - 8 August 2003.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », Manchester, UK, 1994.

Sebastiani F., « Machine learning in automated text categorization », *ACM Computing Surveys*, vol. 34, p. 1-47, 2002.

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

Yarowsky D., « One sense per collocation », Association for Computational Linguistics, Princeton, New Jersey, p. 266-271, 1993.