Une mesure de proximité et une méthode de regroupement pour l'aide à l'acquisition d'ontologies spécialisées

Guillaume Cleuziou , Sylvie Billot , Stanislas Lew ,
Lionel Martin , Christel Vrain

Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
Université d'Orléans
Rue Léonard de Vinci - 45067 ORLEANS Cedex 2
prénom.nom@univ-orleans.fr

Résumé. Cet article traite du regroupement d'unités textuelles dans une perspective d'aide à l'élaboration d'ontologies spécialisées. Le travail présenté s'inscrit dans le cadre du projet BIOTIM. Nous nous concentrons ici sur l'une des étapes de construction semi-automatique d'une ontologie qui consiste à structurer un ensemble d'unités textuelles caractéristiques en classes susceptibles de représenter les concepts du domaine. L'approche que nous proposons s'appuie sur la définition d'une nouvelle mesure non-symétrique permettant d'évaluer la proximité entre lemmes, en utilisant leurs contextes d'apparition dans les documents. En complément de cette mesure, nous présentons un algorithme de classification non-supervisée adapté à la problématique et aux données traitées. Les premières expérimentations présentées sur les données botaniques laissent percevoir des résultats pertinents pouvant être utilisés pour assister l'expert dans la détermination et la structuration des concepts du domaine.

1 Introduction

L'exploitation de données textuelles issues de fonds scientifiques est un objectif de recherche ambitieux dans le domaine de la gestion et de l'acquisition des connaissances. Une des premières étapes pour la mise en place d'un système d'information est la construction d'une ontologie du domaine. Dans cette étude, nous abordons le problème de construction d'une ontologie spécialisée avec une approche mixte (ou semi-automatique). Pour cela, nous nous intéressons à l'étape d'extraction automatique de classes terminologiques susceptibles d'être ensuite validées comme concepts puis structurées par un expert du domaine, l'embryon d'ontologie résultant devant par la suite être enrichi automatiquement.

La tâche de regroupement de mots peut être envisagée de différentes manières (selon l'application visée, les connaissances disponibles sur le domaine ou les traitements possibles). Les études proposées dans ce domaine s'intéressent généralement à l'une des deux étapes suivantes : la définition d'une mesure de proximité entre mots et/ou la proposition d'une méthode de regroupement efficace.

Il existe de nombreuses mesures destinées à évaluer la proximité sémantique entre des mots. On peut classer ces mesures en trois grandes catégories : statistiques, syntaxiques ou

utilisant une base de connaissances. Les mesures statistiques proposées se fondent le plus souvent sur l'étude des cooccurrences de mots dans les textes ou parties de textes en utilisant l'hypothèse de Harris et al. (1989) selon laquelle deux mots sémantiquement proches apparaissent souvent dans des contextes similaires. Ces contextes d'utilisation peuvent être plus précisément repérés en identifiant la syntaxe des phrases. Par exemple Bouaud et al. (1997) analysent les relations de type *nom-adjectif* extraites de syntagmes nominaux et évaluent la proximité entre deux noms en comparant les ensembles de modifieurs (adjectifs) associés. Enfin, la proximité sémantique entre deux mots peut être appréhendée relativement à une base de connaissances structurée, comme par exemple un thésaurus ou une ontologie pré-existante dans le domaine. Les travaux de Rada et Bicknell (1989); Wu et Palmer (1994); Lin et Kondadadi (2001) consistent ainsi à rechercher dans le thésaurus WordNet la position relative des mots dans la hiérarchie de concepts.

Les travaux dans le domaine du regroupement offrent également un assez large éventail de choix d'algorithmes pour organiser un ensemble de mots en classes *via* une mesure de proximité sur cet ensemble. Les méthodes génériques de regroupement (par exemple k-moyennes (MacQueen, 1967), classification ascendante hiérarchique (Sneath et Sokal, 1973)) restent les plus utilisées malgré quelques propositions récentes d'approches plus adaptées (Lelu, 1993; Turenne, 2000; Lin et Kondadadi, 2001; Pantel et Lin, 2002; Cleuziou et al., 2004).

L'orientation que nous proposons dans cette étude est fondée sur la définition d'une nouvelle mesure de proximité utilisant les informations syntaxiques contenues dans les documents d'un corpus spécialisé. Cette mesure est couplée avec une méthode de regroupement agglomératif hiérarchique, adaptée aux besoins de l'étude.

L'article est organisé comme suit : la section 2 présente le projet BIOTIM ainsi que certaines notions fondamentales du domaine de recherche. Les deux sections suivantes sont destinées respectivement à l'étude des proximités entre mots (section 3) et à la proposition d'une méthode de regroupement adaptée (section 4). Cette dernière partie présente également les premiers résultats expérimentaux sur un corpus de botanique. Une synthèse des avancées proposées dans cet article ainsi qu'une discussion sur les nombreuses perspectives de ce travail sont présentées dans la dernière partie.

2 Contexte de l'étude

2.1 Le projet BIOTIM

L'étude menée s'inscrit dans le cadre du projet BIOTIM[1] dont l'objectif est de concevoir des méthodes génériques d'analyse automatique de masses de données regroupant textes et images dans le domaine de la biodiversité. Nous nous intéressons, pour notre part, à la construction semi-automatique d'une ontologie textuelle du domaine à partir de corpus botaniques.

La complémentarité des équipes associées au projet BIOTIM (Traitement du Langage Naturel, Apprentissage, experts du domaine, etc.), permet d'assurer un traitement adapté aux particularités des données. L'utilisation de termes spécialisés, la structure complexe des phrases rencontrées dans le corpus (longues descriptions, souvent sans verbe) et la masse importante de données à traiter sont autant de spécificités à prendre en compte.

[1] ACI "masse de données" : http ://www-rocq.inria.fr/imedia/biotim/

Le choix a été fait de ne pas laisser à l'expert la difficile tâche d'identifier seul les concepts du domaine. Il nous a semblé préférable de l'assister pour cette étape stratégique en proposant des embryons de concepts potentiels, émergeant directement et automatiquement des corpus. Ainsi, le travail de l'expert consistera à juger si un groupe de mots peut traduire ou non un concept du domaine.

Dans la suite de l'étude nous utilisons le corpus de la "Flore du Cameroun", composé de 37 volumes et commercialisé par l'Herbier National Camerounais. Chaque volume a fait l'objet d'une procédure de numérisation, à l'origine de quelques erreurs d'OCR (*Optical Character Recognition*).

2.2 Ontologies et dépendances syntaxiques

La chaîne de traitements[2] effectuée pour extraire un ensemble de termes à partir du texte brut est détaillée dans Rousse et de la Clergerie (2005). Les sorties de ce traitement linguistique sont des termes de la forme *Nom-Adjectif* ou *Nom-(Prép.(Dét.))-Nom*. Au total, près de 35 000 termes construits sur une base de plus de 12 000 lemmes (noms et adjectifs) ont ainsi été extraits sur le corpus de la "Flore du Cameroun". On dénombre par exemple 102 termes différents contenant le lemme "foliole" ; parmi les plus fréquents dans le corpus on peut citer les termes suivants : "foliole terminal", "foliole oblong", "foliole à sommet", "paire de foliole" ou encore "feuille à foliole".

On appelle contexte d'apparition d'un lemme m_i, une structure textuelle de l'une des formes suivantes : *Nom-(Prép.(Dét.))-*$\sim$, *Adjectif-*$\sim$, $\sim$*-(Prép.(Dét.))-Nom* ou encore $\sim$*-Adjectif* dans lesquels m_i peut se substituer à $\sim$ pour former un terme.

Nous présentons ci-dessous le principe de rapprochement de ces lemmes à partir de l'analyse de leurs contextes d'apparition dans les textes.

2.2.1 Modélisation en graphes

Partant de l'hypothèse de Harris, nous utilisons les "dépendances syntaxiques" entre lemmes à l'intérieur des termes pour construire un graphe dont les sommets correspondent aux lemmes présents dans les termes. L'existence d'une arête entre deux sommets indique que les deux lemmes associés partagent des contextes identiques (Bouaud et al., 1997).

Considérons par exemple les termes "arbre à feuille" et "arbre à foliole" ; le contexte "arbre à $\sim$" est partagé par les deux lemmes "feuille" et "foliole", favorisant ainsi leur liaison dans le graphe. Réciproquement on note que "$\sim$ à feuille" et "$\sim$ à foliole" correspondent à deux contextes d'apparition pour le lemme "arbre".

Nous présentons en figure 1, un exemple de sous-graphe obtenu sur le corpus botanique. Il est d'usage, pour cette modélisation en graphes, de recourir à un seuil afin de ne retenir que les dépendances dites non artificielles[3]. Ce seuil doit être choisi relativement à l'importance du corpus utilisé. Le graphe ainsi obtenu donne une cartographie globale du corpus et met en évidence la fonction de certains lemmes dans le domaine considéré. Par l'examen de certaines

[2] Segmentation, étiquetage morpho-syntaxique, lemmatisation, extraction terminologique.

[3] Les dépendances sont qualifiées d'artificielles lorsqu'elles mettent en jeu très peu de contextes, notamment à cause de l'emploi de mots "vides".

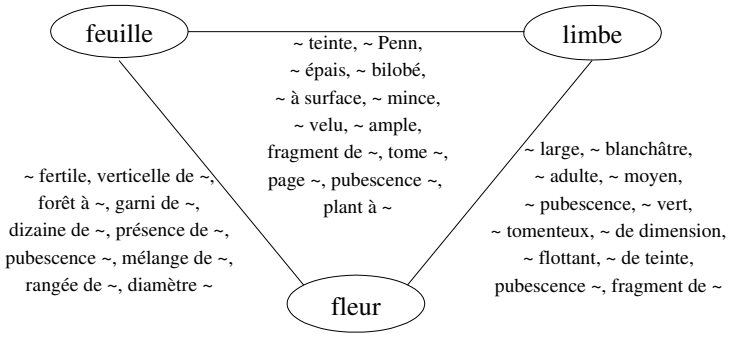

FIG. 1 – *Exemple de sous-graphe représentant les dépendances syntaxiques entre lemmes dans le domaine de la botanique.*

classes de sous-graphes (cliques, composantes connexes, etc.) on peut alors tenter de faire émerger des embryons de catégories sémantiques.

Sur le corpus botanique que nous utilisons un traitement similaire à celui proposé par Bouaud et al. (1997) (un seuil minimum de 10 contextes partagés est requis pour placer une arête entre deux sommets) conduit aux mêmes observations, à savoir la présence dans le graphe d'une composante connexe de taille importante accompagnée de très petites composantes connexes assez pertinentes d'un point de vue sémantique.

2.2.2 Modélisation numérique

Parallèlement à cette modélisation en graphes, une approche numérique est possible. Le Moigno et al. (2002) introduit alors plusieurs coefficients dérivés des graphes précédents :

Le coefficient a correspond au nombre de contextes partagés par deux lemmes (par exemple a(feuille,fleur)=10 d'après le graphe de la figure 1).

La productivité d'un lemme, notée $prod(m)$, correspond au nombre de contextes différents dans lesquels ce lemme apparaît. De manière analogue, la productivité d'un contexte, $prod(c)$ correspond au nombre de lemmes différents apparaissant dans ce contexte. Par exemple, une analyse du lemme "foliole" sur le corpus botanique montre que 102 termes différents contiennent ce lemme ; "foliole" apparaît alors dans 102 contextes distincts ($prod$(foliole) $= 102$). Inversement, seuls les lemmes "sommet", "nervation", "marge" et "pétiole" apparaissent dans le contexte "foliole à $\sim$" ($prod$(foliole à $\sim$) $= 4$).

Le coefficient *prox* utilise cette notion de productivité. Ce coefficient formalise l'intuition que si un contexte est très productif sa contribution dans le rapprochement de deux mots est plus faible que celle d'un contexte peu productif. Soit $\mathcal{C}_i$ (resp. $\mathcal{C}_j$) l'ensemble des contextes d'apparition du lemme m_i (resp. m_j), $prox$ est défini par

$$prox(m_i, m_j) = \sum_{c \in \mathcal{C}_i \cap \mathcal{C}_j} \frac{1}{\sqrt{prod(c)}}$$

Le coefficient J (non symétrique) tente enfin de formaliser le déséquilibre pouvant exister entre un mot très productif et un autre peu productif :

$$J(m_i, m_j) = \frac{a(m_i, m_j)}{prod(m_i)}$$

$J(m_i, m_j)$ sera d'autant plus élevé que m_i partage beaucoup de ses contextes avec m_j.

La formalisation numérique entraîne nécessairement une perte d'information : par exemple on ne retient que le nombre de contextes partagés par deux lemmes et non la liste de ces contextes. Cependant nous montrerons qu'il est possible de tenir compte de ce dernier aspect dans le processus de regroupement. Dans la suite, nous nous attachons à définir une mesure globale de proximité entre deux lemmes, définie à partir des différentes notions précédentes.

3 Une mesure de proximité non symétrique

Le Moigno et al. (2002) ont introduit, *via* le coefficient J, la notion de déséquilibre à propos de la proximité entre deux mots. L'idée est alors de considérer à la fois ce qui rapproche deux mots (leurs contextes partagés) et ce qui les différencie (leurs contextes propres).

Considérons par exemple les deux mots "pétale" et "fleur". On observe sur le corpus les caractéristiques suivantes : a(fleur,pétale)=54, $prod$(fleur)=284 et $prod$(pétale)=196. Nos connaissances générales dans le domaine nous permettent de dire qu'un "pétale" est une partie d'une "fleur". La notion de "pétale" est donc sémantiquement très dépendante de celle de "fleur" tandis que l'inverse n'est pas vrai. La seule donnée du coefficient a ne permet pas d'observer cette propriété tandis que l'information supplémentaire apportée par les productivités respectives des deux mots le permet : "pétale" partage plus de 27% de ses contextes avec "fleur" alors que "fleur" n'en partage que 19% avec "pétale".

Cette vision relative du nombre de contextes partagés permet de faire émerger des dissymétries dans les proximités et nous encourage alors à proposer une mesure qui tienne compte de ces deux informations (nombre de contextes partagés et non partagés) mais également du fait que la proximité entre deux mots n'est pas nécessairement une notion symétrique.

De même que $prox$ est une extension du coefficient a, nous définissons le coefficient α par extension du coefficient J, en introduisant la notion de productivité sur les contextes.
Soient m_i et m_j deux lemmes, $\mathcal{C}_i$ et $\mathcal{C}_j$ les contextes d'apparition associés :

$$\alpha(m_i, m_j) = \frac{\sum_{c \in \mathcal{C}_i \cap \mathcal{C}_j} \frac{1}{\sqrt{prod(c)}}}{\sum_{c \in \mathcal{C}_i} \frac{1}{\sqrt{prod(c)}}}$$

Le coefficient J concerne l'aspect quantitatif de la proportion de contextes partagés relativement à la productivité d'un mot tandis que le coefficient α introduit une dimension qualitative en considérant, en plus, la qualité de ces contextes à travers leur productivité. Ce coefficient sera donc d'autant plus élevé que les coefficients partagés par les deux mots sont peu productifs.

Nous présentons ci-dessous les propriétés vérifiées par α (pour tout couple de mots (m_i, m_j) :

a. $\alpha(m_i, m_j) \in [0, 1]$ (0 si $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ et 1 si $\mathcal{C}_i \quad \mathcal{C}_j$),
b. $\alpha(m_i, m_j)$ augmente avec le nombre de contextes communs aux deux lemmes,
c. $\alpha(m_i, m_j)$ augmente lorsque la productivité des contextes communs diminue,
d. $\alpha(m_i, m_j)$ diminue lorsque le nombre de contextes propres à m_i ($\mathcal{C}_i \setminus \mathcal{C}_j$) augmente,
e. $\alpha(m_i, m_j)$ diminue lorsque la productivité des contextes propres à m_i diminue.

Nous définissons ensuite la proximité p entre deux mots par le produit des termes α et $prox$:

$$p(m_i, m_j) = \alpha(m_i, m_j).prox(m_i, m_j)$$

La mesure p présente alors la caractéristique d'être une mesure non-symétrique (pour cette raison on ne peut parler de similarité) et telle que $p(m_i, m_j) \in [0, prox(m_i, m_j)]$. Au sens de p, un mot m_i sera d'autant plus proche d'un mot m_j que :

1. le nombre de contextes partagés par les deux termes est élevé,
2. les contextes partagés sont peu productifs,
3. il y a une forte proportion des contextes d'apparition de m_i qui sont partagés par m_j,
4. les contextes propres à m_i sont productifs.

Nous avons effectué une étude comparative des mesures a, $prox$ et p en observant les listes ordonnées des Plus Proches Voisins Mutuels[4] (PPVMs) obtenus par chaque mesure. Cette évaluation a révélé une quantité de PPVMs correspondant à des relations sémantiques de synonymie et d'antonymie plus importante pour $prox$ et p que pour le coefficient a. De plus, l'ordonnancement des PPVMs est apparu davantage pertinent avec la mesure p.

Les résultats obtenus sont encourageants pour la suite du processus de construction de classes terminologiques. Nous nous attachons dans ce qui suit à présenter une méthode agglomérative adaptée au regroupement de mots d'après leurs contextes d'apparition dans les textes.

4 Le regroupement de mots

La problématique générale du regroupement (ou *clustering*) consiste à organiser un ensemble d'objets en groupes de façon à ce que deux objets similaires se retrouvent dans un même groupe et deux objets dissimilaires dans des groupes distincts. De nombreuses stratégies ont été proposées pour répondre à cette problématique (Jain et al., 1999), comme par exemple les approches par partitionnement (k-moyennes), les algorithmes hiérarchiques (agglomératifs ou divisifs), les méthodes utilisant des mélanges de densités de probabilité, des découpages en grilles, etc.

La plupart des travaux présentés dans le domaine du regroupement de données textuelles (chaînes graphiques, lemmes, termes, mots-clés, documents, etc.) se focalisent davantage sur

[4]Deux mots m_i et m_j constituent une paire de PPVMs si m_i (resp. m_j) a pour plus proche voisin m_j (resp. m_i) selon la mesure considérée.

le sens à donner à la notion de proximité que sur l'algorithme permettant de regrouper les unités textuelles considérées. Certaines études proposent cependant des approches originales afin de regrouper des objets textuels en tenant compte de leurs spécificités telles que la polysémie d'un mot ou l'aspect multi-thématique d'un document (Lelu, 1993; Pantel et Lin, 2002; Cleuziou et al., 2004). Malgré ces travaux récents et marginaux, l'étape de regroupement reste généralement réalisée par les méthodes classiques (algorithme des k-moyennes ou approche agglomérative) car simples et maîtrisées par les utilisateurs.

4.1 Une méthode de regroupement adaptée

Dans notre travail, nous proposons d'adapter le processus de regroupement aux données dont nous disposons (lemmes et contextes associés) ainsi qu'à la tâche visée (aide à l'élaboration d'une ontologie). L'approche présentée ici est une adaptation de l'algorithme agglomératif hiérarchique du lien moyen. Ce dernier procède par fusions successives des deux plus proches groupes[5], en partant des feuilles (un objet par groupe) et aboutissant à une racine (tous les objets dans un même groupe). Ce type d'approche présente l'avantage de conserver une trace de l'élaboration des groupes à travers l'arbre hiérarchique (ou dendrogramme) construit. En revanche, un problème récurrent pour cette méthode est la recherche des groupes pertinents parmi l'ensemble des nœuds de l'arbre.

Pour cela nous choisissons d'interdire l'agglomération autour d'un groupe lorsque cette fusion conduit à un groupe conceptuellement non pertinent (cf. définition 4.1 ci-dessous). La structure ainsi obtenue est une *hiérarchie partielle*, soit un ensemble de (petits) dendrogrammes (cf. définition 4.2 ci-dessous).

Définition 4.1 *Soient P un groupe constitué des lemmes $\{m_1, \ldots, m_n\}$ et $\mathcal{C}_1, \ldots, \mathcal{C}_n$ les ensembles de contextes d'apparition associés à chacun d'eux, P est* **conceptuellement non pertinent** *si il n'existe aucun contexte dans lequel apparaissent l'ensemble des lemmes de P :*

$$\bigcap_{i=1\ldots n} \mathcal{C}_i = \emptyset$$

Définition 4.2 *Soit H un ensemble de parties non-vides sur un ensemble d'objets X, H est une* **hiérarchie partielle** *si les propriétés suivantes sont vérifiées :*

i) *pour tout* $x_i \in X,\ \{x_i\} \in H$,
ii) *pour tout* $h, h' \in H,\ h \cap h' \in \{\emptyset, h, h'\}$,
iii) *pour tout* $h \in H,\ \cup\{h' \in H\ :\ h' \quad h\} \in \{h, \emptyset\}$.

L'ajout de la propriété "$X \in H$" permet de se ramener à la définition formelle classique d'une hiérarchie (complète).

L'algorithme agglomératif hiérarchique adapté au regroupement de mots relativement à leurs contextes d'apparition est le suivant : initialement chaque lemme constitue un groupe à lui seul (feuille) auquel est associé une caractérisation (ensemble des contextes d'apparition du lemme) ; à chaque itération, on recherche parmi les fusions possibles (respect de la contrainte

[5] Par la méthode du lien moyen la proximité entre deux groupes est obtenue en effectuant la moyenne des proximités entre deux objets de groupes différents.

de pertinence) celle qui permet d'agglomérer les deux groupes les plus proches selon la mesure de proximité spécifiée. De cette fusion résulte un nouveau groupe auquel est associée une nouvelle caractérisation "mère", intersection des deux caractérisations "filles". La matrice des proximités entre groupes est mise à jour. Lorsque la contrainte de pertinence interdit toute fusion, l'agglomération est terminée et l'algorithme retourne l'ensemble des groupes constitués, les arbres hiérarchiques et les caractérisations associées. On pourra choisir de ne considérer par la suite que les items associés à des groupes de taille supérieure à 1.

L'ajout d'une contrainte de pertinence est essentiel dans cet algorithme. L'utilisation du lien moyen pour évaluer la proximité entre deux groupes P_i et P_j permet de considérer tous les couples de lemmes dans $P_i \quad P_j$ (contrairement aux liens simple et complet). Cependant cette information seule n'apporte aucune garantie quant à l'existence d'une propriété commune à l'ensemble des objets des deux groupes. Cette "propriété commune" est pourtant indispensable pour définir un concept. La contrainte de pertinence apporte cette garantie ; les lemmes d'un groupe apparaissent tous dans un même contexte au minimum. De plus la caractérisation d'un groupe aidera l'expert à proposer une étiquette au concept associé.

4.2 Application aux données botaniques

Nous avons testé l'algorithme de regroupement sur les lemmes extraits du corpus botanique. Parmi les 12 000 lemmes repérés, nous avons sélectionné ceux partageant au moins trois contextes avec un autre lemme, restreignant ainsi à 2 024 la quantité de données à traiter.

La mesure de proximité utilisée est la mesure p, en choisissant comme proximité pour deux lemmes donnés m_i et m_j, la plus petite des deux valeurs possibles par p : $\min\{p(m_i, m_j), p(m_j, m_i)\}$.

Quelques groupes obtenus sont présentés dans les figures 2 et 3. Les dendrogrammes proposés sont ceux mettant en jeux les 10 premières fusions (itérations de l'algorithme). Ces arbres hiérarchiques sont représentatifs de l'ensemble des résultats obtenus. On peut les organiser en trois catégories en fonction des termes mis en jeu : les termes spécifiques au domaine, les termes génériques et enfin ceux relevant des abréviations des noms propres ou des mots étrangers.

Les arbres mettant en jeu des termes spécifiques au domaine (figures 2 et 3, arbres c, d, e, f et j) sont difficiles à évaluer pour des lecteurs non-experts du domaine. On peut malgré tout appréhender la sémantique globale de certains groupes : par exemple le groupe f est la représentation textuelle du concept "aspect du limbe"[6]. D'autres concepts spécifiques émergent de l'analyse de l'ensemble des résultats, par exemple la "forme du sépale" ou plus généralement du lobe (*linéaire-lancéolé, ovale-lancéolé, ensiforme*), la "forme d'une foliole ou d'un lobe" (*deltoïde, linéaire, ovale, rhomboïde, falciforme, cunéiforme, polymorphe*), l'"apparence" que peut prendre une espèce végétale (*plante, herbe, liane, arbrisseau*), etc.

Les arbres contenant des termes génériques sont cette fois plus faciles à évaluer (figure 3, arbres g, h et i). Leur analyse vient confirmer l'impression de qualité puisque les groupes observés peuvent effectivement être associés à des concepts du domaine :

– g est une représentation textuelle du concept de "couleur d'écorce",

[6] *Limbe* : partie élargie d'une feuille ou d'un pétale.

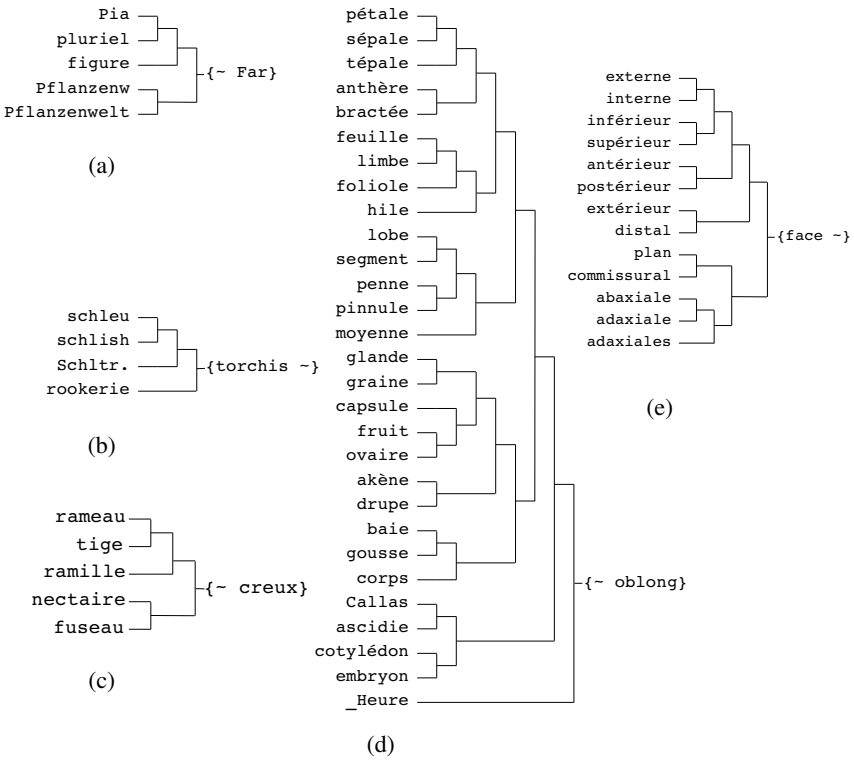

FIG. 2 – *Exemple de groupes obtenus. Les dendrogrammes (a) à (e) contiennent respectivement les itérations 1 à 5 de l'algorithme.*

- h est une représentation textuelle du concept de "variantes de couleurs" (en particulier pour les teintes noir, marron et jaune),
- i est une représentation textuelle du concept d' "unité de mesure" (en particulier pour indiquer la hauteur des végétaux).

Ces concepts peuvent notamment être identifiés plus facilement grâce à la caractérisation proposée. Parmi les résultats non présentés ici, on retrouve d'autres concepts simples tels que : la "forme" associée à un élément d'une plante (*bec, dôme, languette, ruban, gouttière*), les "points cardinaux" (*nord, sud, ouest, Ouest*), les "mois" du calendrier (*janvier, mai, août, novembre, décembre*), etc.

Enfin, les arbres de la dernière catégorie (figure 2, arbres a et b), sont composés en grande partie de termes correspondant à des abréviations des noms propres ou des mots étrangers. Les termes de ce type pourront être supprimés en sélectionnant dans les documents, uniquement les parties descriptives de plantes (travail en cours de réalisation).

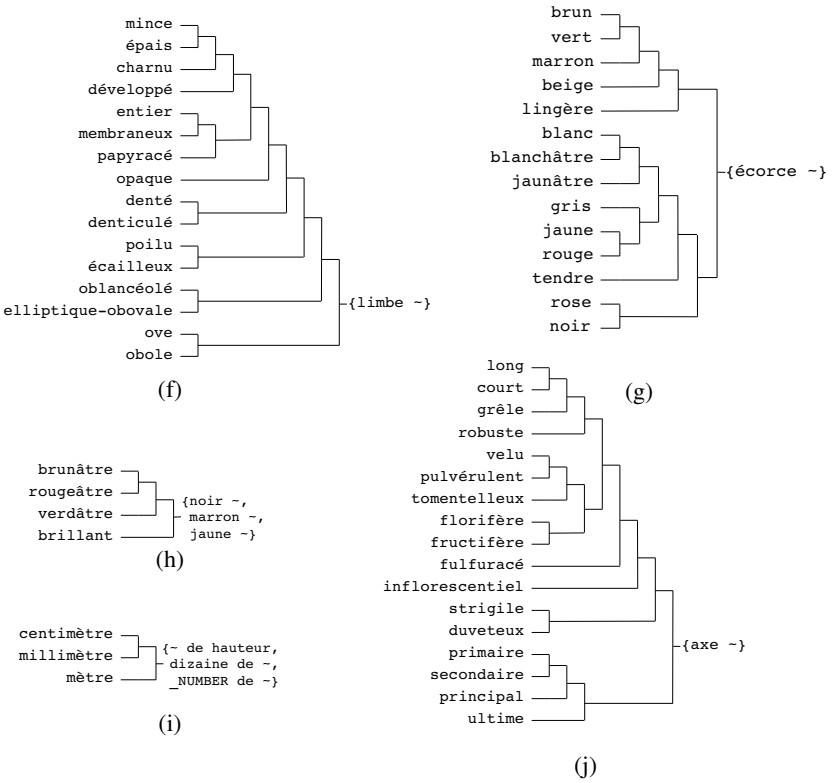

FIG. 3 – *Exemple de groupes obtenus. Les dendrogrammes (f) à (j) contiennent respectivement les itérations 6 à 10 de l'algorithme.*

Pour effectuer une synthèse de nos premières analyses, nous pouvons conclure à la pertinence globale des groupes obtenus et noter l'aide précieuse apportée par la caractérisation associée à chaque groupe. Ce résultat est imputable pour partie à la mesure de proximité proposée mais également à l'adaptation de l'algorithme de classification.

5 Conclusion et perspectives

Notre étude s'est focalisée sur la tâche de regroupement de mots dans un domaine de recherche plus vaste qu'est la construction semi-automatique d'ontologies spécialisées. Nous avons défini une nouvelle mesure de proximité entre mots d'une part, et proposé une méthode de regroupement adaptée d'autre part.

La mesure de proximité présentée se place dans la lignée des mesures utilisant les dépendances syntaxiques. Contrairement aux précédentes propositions, nous ne considérons pas la proximité comme une notion symétrique. L'algorithme de regroupement utilisé est une adaptation des méthodes de classifications ascendantes hiérarchiques. Plutôt que d'aboutir à un arbre hiérarchique complet, c'est un dendrogramme partiel qui est élaboré par l'ajout d'une contrainte de cohérence sur les fusions effectuées. Finalement, tous les objets initiaux ne sont pas nécessairement utilisés, les groupes obtenus sont de petite taille et complétés par des informations structurelles (dendrogramme) et conceptuelle (caractérisation).

Dans le cadre du projet BIOTIM, ce travail a été appliqué sur des textes du domaine de la botanique. Les résultats obtenus avec la mesure de proximité mettent en évidence des paires de lemmes pertinentes. Ce résultat est confirmé par l'analyse des groupes finalement obtenus en utilisant cette mesure couplée à la méthode de classification proposée. En effet la plupart de ces groupes semble correspondre à des concepts du domaine.

Nous envisageons actuellement d'améliorer la qualité des groupes obtenus en travaillant selon deux axes :

- réduire l'impureté par le renforcement des contraintes de cohérence appliquées aux fusions (extraction d'ensembles de contextes fréquents),
- proposer des groupes exhaustifs en autorisant la réutilisation de lemmes déjà agglomérés pour définir de nouveaux concepts (approches pyramidales).

Parallèlement à ces perspectives, nous tâcherons de mettre en évidence les fortes analogies qui existent entre les mesures de proximité étudiées dans cet article et les indices proposés dans le cadre d'approches plutôt probabilistes utilisant les cooccurrences : information mutuelle, coefficient de Dice, etc. (Cleuziou et al., 2003).

Enfin nous développerons une interface de validation destinée aux experts du domaine. Cette interface permettra d'une part d'aider les experts à construire l'ontologie en intervenant sur le processus de classification (validation/structuration des concepts) et d'autre part d'évaluer quantitativement la pertinence de la méthodologie proposée.

Références

Bouaud, J., B. Habert, A. Nazarenko, et P. Zweigenbaum (1997). Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles. In *Ingéniérie de la connaissance*, Roskoff, pp. 207–223.

Cleuziou, G., V. Clavier, et L. Martin (2003). Une méthode de regroupement de mots fondée sur la recherche de cliques dans un graphe de cooccurrences. In L. ENSAIS (Ed.), *5èmes rencontres Terminologie et Intelligence Artificielle*, Strasbourg, France, pp. 179–182. Poster.

Cleuziou, G., L. Martin, et C. Vrain (2004). PoBOC : an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data. In R. López de Mántaras and L. Saitta, IOS Press (Ed.), *Proceedings of the 16th European Conference on Artificial Intelligence*, Valencia, Spain, pp. 440–444.

Harris, Z., M. Gottfried, T. Ryckman, P. Mattick, A. Daladier, T. N. Harris, et S. Harris (1989). *The form of Information in Science : Analysis of an immunology sublanguage*. Dordrecht : Kluwer Adademic Publishers.

Jain, A. K., M. N. Murty, et P. J. Flynn (1999). Data clustering : a review. *ACM Computing Surveys 31*(3), 264–323.

Le Moigno, S., J. Charlet, D. Bourigault, P. Degoulet, et M. Jaulent (2002). Terminology extraction from text to build an ontology in surgical intensive care. In *Proceedings of the AMIA Annual Symposium*, San Antonio, Texas, pp. 9–13.

Lelu, A. (1993). Modèles neuronaux pour l'analyse de données documentaires et textuelles. Thèse de doctorat, Université de Paris VI.

Lin, K. I. et R. Kondadadi (2001). A Word-Based Soft Clustering Algorithm for Documents. In *Proceedings of 16th International Conference on Computers and Their Applications*, Seattle, Washington.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability*, Volume 1, Berkeley, pp. 281–297. University of California Press.

Pantel, P. et D. Lin (2002). Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Edmonton, Alberta, Canada, pp. 613–619. ACM Press.

Rada, R. et E. Bicknell (1989). Ranking documents with a thesaurus. *Journal of the American Society for Information Science 40*(5), 304–310.

Rousse, G. et E. de la Clergerie (2005). Analyse automatique de documents botaniques : le projet biotim. In *6èmes rencontres Terminologie et Intelligence Artificielle*, Rouen, France.

Sneath, P. H. A. et R. R. Sokal (1973). Numerical Taxonomy - The Principles and Practice of Numerical Classification. San Francisco, W. H. Freeman and Compagny.

Turenne, N. (2000). Apprentissage statistique pour l'extraction de concepts à partir de textres. application au filtrage d'informations textuelles. Thèse de doctorat. ENSAIS, Université Louis-Pasteur Strasbourg.

Wu, Z. et M. Palmer (1994). Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, Las Cruces, New Mexico, pp. 133–138.

Summary

In this paper, we study the problem of clustering textual units in the framework of helping an expert to build a specialized ontology. This work has been achieved in the context of a French project, called BIOTIM, handling botany corpora. Building an ontology, either automatically or semi-automatically is a difficult task. We focus on one of the main steps of that process, namely structuring the textual units occurring in the texts into classes, likely to represent concepts of the domain. The approach that we propose relies on the definition of a new non-symmetrical measure for evaluating the semantic proximity between lemma, taking into account the contexts in which they occur in the documents. Moreover, we present a non-supervised classification algorithm designed for the task at hand and that kind of data. The first experiments performed on botanical data have given relevant results.