Conférence TIA-2005, Rouen, 4 et 5 avril 2005

Apprentissage par analogie pour la structuration de terminologie - Utilisation comparée de ressources endogènes et exogènes

Vincent Claveau et Marie-Claude L'Homme

OLST – Université de Montréal
C.P. 6128 succ. Centre-Ville, Montréal, QC, H3C 3J7 Canada
{vincent.claveau,mc.lhomme@umontreal.ca}
http://www.olst.umontreal.ca

Résumé

Cet article présente une méthode originale pour détecter en corpus spécialisé des couples de termes morphologiquement liés et prédire le lien sémantique qui les unit dans le domaine étudié. Ces liens sémantiques, modélisés à l'aide de fonctions lexicales, permettent ainsi de structurer une terminologie du domaine. La méthode exposée repose sur une technique d'apprentissage artificiel par analogie qui permet de confronter efficacement des couples de mots inconnus à des exemples de couples de termes dont le lien sémantique est connu. Elle tire également parti d'un système d'extraction de termes qui permet d'éviter la détection de liens non pertinents dans le domaine.

Cette approche est évaluée dans le domaine de l'informatique ; les résultats montrent que l'approche simple que nous proposons est très performante. Deux expériences sont notamment menées, l'une utilisant des exemples issus du domaine, l'autre, des exemples tirés d'une base généraliste. La comparaison des résultats de chacune d'elles permet ainsi d'évaluer quantitativement l'intérêt de telles ressources généralistes pour ce type de tâches et apporte ainsi une contribution chiffrée au débat opposant l'utilisation ressources endogènes et exogènes en terminologie computationnelle.

Mots-clés : Terminologie, structuration, relations sémantiques, fonctions lexicales, apprentissage artificiel, analogies

1. Introduction

Dans le domaine de la terminologie computationnelle, en marge des recherches menées en extraction de termes, de plus en plus de travaux soulignent l'importance de la structuration de terminologie. Cette structuration nécessite de découvrir les liens, notamment sémantiques, entre les unités terminologiques et d'éventuellement les étiqueter. La découverte de tels liens se fait le plus souvent soit par des méthodes « externes », soit par des méthodes « internes » (Grabar & Zweigenbaum, 2002) ; un état de l'art de quelques unes de ces méthodes parmi les plus ré-

centes est proposé par Daille *et al.* (2004). Les méthodes externes s'appuient sur des indices trouvés en corpus dans le contexte des termes pour découvrir les relations sémantiques qu'ils entretiennent (Claveau & L'Homme, 2004, *inter alia*). Les méthodes internes s'appuient quant à elles sur la forme des termes pour faire ces mêmes associations. Parmi ces dernières méthodes, certaines exploitent des connaissances externes riches (Namer & Zweigenbaum, 2004), ou des règles morphologiques données a priori (Daille, 2003), ce qui rend ces méthodes peu portables d'un domaine ou d'une langue à l'autre. D'autres travaux utilisent au contraire peu d'informations et exploitent au mieux des données existantes, comme des thésaurus (Zweigenbaum & Grabar, 2000) ou des corpus (Zweigenbaum & Grabar, 2003), mais cherchent uniquement à constituer des familles morphologiques sans distinguer les relations sémantiques que chaque mot entretient avec les autres.

Dans cet article, nous proposons une méthode interne originale portant sur les liens morphologiques entre termes, proche des travaux de Zweigenbaum & Grabar (2003). Cependant, dans notre cas, notre but n'est pas seulement de découvrir si un terme est morphologiquement lié à un autre, mais aussi de prédire quelle relation sémantique précise les lie. Ce travail repose sur deux hypothèses qui servent de fils conducteurs à cet article :

1. les corpus spécialisés contiennent des dérivations morphologiques régulières indicatrices de relations sémantiques régulières ;
2. ces liens morpho-sémantiques sont pour une grande part propres au domaine d'étude.

Pour découvrir en corpus les termes morphologiquement liés et prédire la relation sémantique qui les unit, nous proposons d'utiliser une technique d'apprentissage simple et originale en conjonction avec un système d'extraction de termes. Cette technique d'apprentissage supervisée (*i.e.* nécessitant des exemples, voir section 3) nous permet de rendre compte des particularités de cette tâche de classification et ne nécessite aucune connaissance externe autre que les exemples. Conformément à notre première hypothèse, cette approche est évaluée dans le domaine spécialisé de l'informatique, en français.

Par ailleurs, pour mesurer la portée de notre seconde hypothèse, nous présentons deux expériences, l'une (expérience 1 par la suite) utilisant des exemples du domaine traité (ressources endogènes), l'autre (expérience 2) utilisant des exemples issus de la langue générale (ressources exogènes). La comparaison des résultats de ces deux expériences doit ainsi nous permettre de mesurer clairement l'intérêt des ressources endogènes par rapport aux ressources exogènes pour ce type de tâche de terminologie computationnelle.

Nous présentons tout d'abord le cadre de ce travail, et plus particulièrement la façon dont les liens sémantiques entre termes morphologiquement proches sont modélisés. Ensuite, nous détaillons le fonctionnement de notre système, et plus précisément l'utilisation qui est faite de l'apprentissage par analogie. Nous décrivons enfin la méthodologie d'évaluation et les résultats de nos deux expériences.

2. Cadre linguistique

Le travail entrepris ici a pour but d'assister les terminologues travaillant au développement d'un dictionnaire français spécialisé du domaine de l'informatique, le DiCoInfo. Ce dictionnaire est développé en suivant une approche lexico-sémantique de l'analyse terminologique (L'Homme, 2004) et repose essentiellement sur l'emploi des fonctions lexicales (Mel'čuk *et al.*, 1984 1999) – FL par la suite – pour représenter les relations sémantiques entre termes.

De nombreux liens sémantiques sont encodés au sein du DiCoInfo. On trouve ainsi des liens syntagmatiques, *i.e.* exprimés par des collocations. Par exemple, les verbes *enregistrer, formater, défragmenter* et l'adjectif *externe* se trouvent dans la fiche décrivant le terme *disque dur*. On trouve également des relations paradigmatiques telles que l'hyperonymie, la synonymie, l'antonymie, les relations actancielles. Les FL sont utilisées pour rendre compte de manière uniforme et systématique de ces diverses relations. Un exemple des FL encodées pour une entrée du DiCoInfo est donné en annexe A.

Le travail présenté dans cet article n'est concerné que par un sous-ensemble de relations sémantiques. Elles peuvent être syntagmatiques ou paradigmatiques, mais impliquent toutes des couples de termes morphologiquement liés. Le tableau 1 donne quelques exemples de ces relations avec la FL correspondante.

FL	**Mot clé**	**Valeur**	**Explication de la FL**
A_1	*résider*	*résident*	l'agent a ou est + le sens du mot clé
A_2	*infecter*	*infecté*	le patient est + le sens du mot clé
$Able_1$	*interagir*	*interactif*	l'agent peut être + le sens du mot clé
$Able_2$	*programmer*	*programmable*	le patient peut être + le sens du mot clé
Anti	*installer*	*désinstaller*	antonymie
De_nouveau	*compiler*	*recompiler*	de nouveau
$Caus_1Func_0$	*imprimé*	*imprimer*	l'agent crée le mot clé
$Caus_1Oper_2$	*partition*	*partitionner*	l'agent cause que le patient ait un mot clé
CausPred	*valide*	*valider*	qqn ou qqch. rend mot clé
$Fact_1$	*pirate*	*pirater*	le mot clé intervient sur le patient
$Labreal_{12}$	*navigateur*	*naviguer*	l'agent agit sur le patient en utilisant le mot clé
S_0	*formater*	*formatage*	le nom a le même sens que le mot clé
S_{agent}	*programme*	*programmeur*	l'agent typique du mot clé
$S_{instrument}$	*éditer*	*éditeur*	l'instrument typique du mot clé
S_{res}	*programmer*	*programme*	le résultat typique du mot clé

TAB. 1 – Exemples de relations sémantiques entre termes

Il est important de noter que les FL servent à représenter les relations sémantiques sans considération de similarité formelle (la ressemblance morphologique est de fait considérée comme accidentelle dans ce cadre théorique). Cependant, comme nous l'avançons dans notre première hypothèse, en nous restreignant à des domaines de spécialité particuliers, nous considérons que des liens morphologiques réguliers trahissent également des liens sémantiques réguliers.

D'autres travaux ont montré que la similitude formelle – même si elle ne révèle par la totalité de la structure terminologique d'un domaine – permet de mettre au jour des relations terminologiques importantes et ce, dans une variété de domaines :

- domaine médical (Zweigenbaum & Grabar, 2000) : *acide, acido, acidité, acidurie, acidémie, acidophile, acidocitose* ;
- domaine agro-alimentaire (Daille, 2003) : *solubilisation micellaire* $\Rightarrow$ *insolubilisation micellaire* ; *plume de canard* $\Rightarrow$ *plumard de canard* ; *filetage de saumon* $\Rightarrow$ *filet de saumon* ;
- domaine des affaires (Binon *et al.*, 2000) : *promotion, promo, promoteur, promotrice, promouvoir, promotionner*.

3. Technique d'acquisition

3.1. Apprentissage par analogie

La méthode d'apprentissage sous-tendant notre approche est basée sur l'analogie. Une excellente présentation de l'analogie et de son utilisation dans un cadre d'apprentissage artificiel se trouve dans Lepage (2003), dont nous reprenons ci-après les notations. Des travaux en traitement automatique des langues (morphologie, syntaxe, traduction automatique...) s'appuient directement sur ce concept (Lepage, 2003 ; Lepage, 2004).

Formellement, une analogie peut être représentée par la proposition A : B $\doteq$ C : D, qui signifie « A est à B ce que C est à D » ; c'est-à-dire que le couple A-B est en analogie avec le couple C-D. Ce cadre est parfaitement adapté à notre tâche dans laquelle de telles analogies peuvent être dérivées de couples de termes morphologiquement liés. Par exemple, on peut postuler l'analogie suivante :

$$\textit{connecteur} : \textit{connecter} \doteq \textit{éditeur} : \textit{éditer},$$

et si l'on sait par ailleurs que l'on a la relation sémantique $S_{\text{instrument}}$(*connecter*) = *connecteur*, on peut alors prédire qu'il existe le même lien (c'est-à-dire la même FL) entre *éditeur* et *éditer*, et donc que $S_{\text{instrument}}$(*éditer*) = *éditeur*. C'est ce type d'analogies entre couples dont la FL est connue et couples inconnus qui doit nous permettre de prédire les liens sémantiques existant au sein de ces derniers.

D'un point de vue de l'apprentissage artificiel, cette approche a plusieurs aspects très intéressants. Tout d'abord, c'est une technique essentiellement supervisée, sans saut inductif[1], puisque c'est un cas particulier de *case-based learning* (Kolodner, 1993) dans lequel les instances sont des couples de mots. Nous devons donc fournir à cette technique des exemples de couples de termes en relation morphologique et sémantique. Ensuite, cette approche permet de rendre compte simplement de contraintes imposées par notre tâche. Ainsi, le nombre de classes considérées, c'est-à-dire le nombre de FL pouvant décrire les liens entre dérivés morphologiques, est assez grand (on parle d'apprentissage multi-classe) et dépendant des exemples fournis. Enfin, un couple de termes en dérivation morphologique peut partager plusieurs relations sémantiques différentes ; la technique d'apprentissage utilisée doit donc permettre qu'une instance relève de plusieurs classes. Ces différentes contraintes sont parfaitement intégrées par notre approche par analogie, alors qu'elles excluent une grande majorité des techniques d'apprentissage plus communes.

3.2. Constitution des données d'apprentissage

Pour identifier les analogies morphologiques, nous avons besoin d'exemples de termes morphologiquement et sémantiquement liés et annotés avec leur FL. Pour notre première expérience, les exemples que nous utilisons sont propres au domaine que nous traitons. Pour les récupérer, nous utilisons les entrées existantes du DiCoInfo (*cf.* annexe A). Des couples de termes en relations morphologiques et sémantiques sont automatiquement extraits des fiches dictionnairiques en cherchant, parmi tous les liens encodés pour une entrée, ceux dont les termes reliés

[1] Dans ce type de technique, le saut inductif est repoussé au moment de la comparaison d'une nouvelle instance avec les exemples connus ; c'est pourquoi ces techniques sont aussi parfois qualifiées de *lazy learning*. Le saut inductif est donc mis en œuvre par la fonction de similarité utilisée pour faire cette comparaison (*cf.* sous-section 3.3).

sont proches en terme de distance d'édition ou de plus longue sous-chaîne commune. Ce sont ainsi près de 900 exemples (couples de termes avec leur FL) qui sont collectés de cette manière à partir des 1700 entrées que compte actuellement le DiCoInfo.

Pour notre seconde expérience, les exemples sont tirés d'une base de données lexicales généraliste, le DiCo (Dictionnaire des Collocations, voir l'extrait donné en annexe B). Cette base française est développée à l'OLST par une équipe dirigée par Igor Mel'čuk et Alain Polguère (Polguère, 2000) ; l'extrait que nous utilisons compte environ 700 entrées (lexies). Le DiCo a constitué un modèle pour l'élaboration du DiCoInfo ; de ce fait, le formatage de leurs entrées respectives est très similaire, et les relations sémantiques y sont notamment encodées de manière identique à l'aide du formalisme des FL. Pour collecter les exemples de cette base, nous procédons de manière parfaitement similaire à précédemment. Environ 379 exemples sont ainsi récupérés.

Néanmoins, outre le fait que le DiCoInfo porte sur un domaine spécialisé et le DiCo, sur la langue générale, les ressources diffèrent sur certains plans. Nous signalons ici quelques différences importantes du point de vue des expériences décrites dans cet article. D'une part, la plupart des entrées du DiCo sont de nature nominale, alors que le DiCoInfo accorde une grande place aux verbes et aux adjectifs (on y trouve également quelques adverbes). D'autre part, certains sens réguliers dans le domaine de l'informatique sont décrits au moyen de FL qui n'apparaissent pas dans le DiCo.

3.3. Analogie entre termes morphologiquement liés

Le préalable essentiel à l'utilisation effective de l'apprentissage par analogie est la définition de la notion de similarité qui permet de statuer que deux paires de propositions – dans notre cas deux couples de lemmes – sont en analogie. La notion de similarité que nous utilisons, notée Sim, est simple mais adaptée au français, de même qu'à de nombreuses autres langues dans lesquelles la dérivation est principalement obtenue par préfixation et suffixation (voir également Lepage (2003) pour un exemple d'utilisation de l'analogie dans un cadre de dérivation par infixation).

Notons $\mathrm{lcss}(\mathrm{X},\mathrm{Y})$ la plus longue sous-chaîne commune à deux chaînes de caractères X et Y (*e.g.* lcss(*installer*, *désinstallation*) = *install*), $\mathrm{X} +_{suf} \mathrm{Y}$ la concaténation du suffixe Y à X, $\mathrm{X} -_{suf} \mathrm{Y}$ la soustraction du suffixe Y à X, $\mathrm{X} +_{pre} \mathrm{Y}$ la concaténation dur préfixe Y à X, $\mathrm{X} -_{pre} \mathrm{Y}$ la soustraction du préfixe Y à X. La mesure de similarité Sim est définie de la manière suivante ; si on a deux couples de mot m_1-m_2, m_3-m_4 :

$$\text{Sim}(m_1\text{-}m_2, m_3\text{-}m_4) = 1 \quad \text{si} \quad \begin{cases} m_1 = \mathrm{lcss}(m_1, m_2) +_{pre} \mathrm{Pre}_1 +_{suf} \mathrm{Suf}_1, \text{ et} \\ m_2 = \mathrm{lcss}(m_1, m_2) +_{pre} \mathrm{Pre}_2 +_{suf} \mathrm{Suf}_2, \text{ et} \\ m_3 = \mathrm{lcss}(m_3, m_4) +_{pre} \mathrm{Pre}_1 +_{suf} \mathrm{Suf}_1, \text{ et} \\ m_4 = \mathrm{lcss}(m_3, m_4) +_{pre} \mathrm{Pre}_2 +_{suf} \mathrm{Suf}_2 \end{cases}$$

Sim(m_1-m_2, m_3-m_4) = 0 sinon

où $\mathrm{Pre_i}$ et $\mathrm{Suf_i}$ sont des chaînes de caractères quelconques. Intuitivement, Sim vérifie que, pour passer de m_3 à m_4, la même suite de préfixation et de suffixation que pour passer de m_1 à m_2 est nécessaire. Si Sim(m_1-m_2, m_3-m_4) = 1, cela signifie que l'analogie $m_1 : m_2 \doteq m_3 : m_4$ est vérifiée et donc que les relations sémantiques (*i.e.* les FL) entre m_1 et m_2 sont les mêmes qu'entre m_3 et m_4.

Notre processus de détection de dérivation morphologique et d'annotation de relation sé-

mantique consiste ainsi à vérifier si un couple de lemmes inconnu est en analogie avec un ou plusieurs de nos exemples. Si c'est le cas, le couple inconnu est annoté avec la ou les mêmes FL que les exemples analogues. En pratique, pour réaliser efficacement cette comparaison avec les exemples, on construit des règles de dérivation reflétant la façon dont Sim est calculée, c'est-à-dire la suite d'opérations nécessaires pour aller d'un terme à un autre dans un couple exemple. Considérons par exemple le couple *programmation–programmer*, sachant que l'on a la FL V_0(*programmation*) = *programmer*, la règle suivante va être apprise :

$V_0(m_1) = m_2$ si $m_1 -_{suf}$ "ation" $+_{suf}$ "er" $= m_2$

Tout nouveau couple compatible avec cette règle est en analogie avec *programmation–programmer* et est donc étiqueté avec la FL V_0. À l'inverse, puisque nous savons aussi que l'on a S_0(*programmer*) = *programmation*, on infère également une règle :

$S_0(m_1) = m_2$ si $m_1 -_{suf}$ "er" $+_{suf}$ "ation" $= m_2$

De la même manière, à partir de l'exemple AntiAble_2(*activer*) = *désactivable*, la règle suivante est construite :

$\text{AntiAble}_2(m_1) = m_2$ si $m_1 -_{suf}$ "er" $+_{suf}$ "able" $+_{pre}$ "dés" $= m_2$

Un certain nombre de couples exemples produisent des règles parfaitement identiques (même suite d'opérations et même FL) ; une seule règle est alors conservée pour la suite.

Finalement, ce sont 402 règles qui sont ainsi obtenues des exemples du DiCoInfo pour l'expérience 1, permettant d'identifier 67 FL différentes, et 279 règles, tirées des exemples du DiCo, pour l'expérience 2, permettant d'identifier 66 FL différentes. Un couple quelconque de mots compatible avec l'une de ces règles est donc en analogie avec l'un de nos exemples et peut être étiqueté avec la même FL que ce dernier.

3.4. Utilisation du système d'extraction de termes TERMOSTAT

Conjointement à la technique d'apprentissage décrite ci-avant, nous utilisons un système d'extraction de termes sur corpus appelé TERMOSTAT (Drouin, 2003). Contrairement à beaucoup d'autres, ce système a la particularité de pouvoir extraire, en plus des termes complexes, des termes simples. Pour ce faire, TERMOSTAT calcule un indice appelé « coefficient de spécificité » (plus simplement spécificité par la suite) pour chacun des mots apparaissant dans le corpus du domaine spécialisé traité, en comparant sa fréquence (plus précisément une mesure dérivée de sa fréquence) à celle qu'il a dans un corpus de langue générale. Plus un mot a une spécificité élevée, plus il a de probabilité d'être un terme du domaine. À l'inverse, un mot ayant une spécificité négative appartient certainement au langage général.

Le corpus que nous avons utilisé pour ces expériences est un corpus en français contenant environ 1 million de mots. Il est composé de près de 100 textes tirés de livres ou de sites Web, tous publiés entre 1996 et 2004, traitant de différents sous-domaines de l'informatique (réseau, administration système, webcams, utilisation de Linux...). Ce corpus spécialisé est confronté à un corpus généraliste du journal *Le Monde*. Les résultats détaillés des performances de TERMOSTAT pour l'extraction de termes simples sont présentés par Lemay *et al.* (2005).

Dans le cadre de notre application, TERMOSTAT, en nous fournissant des mots étant probablement des termes, sert à éviter d'obtenir des couples non pertinents pour le domaine étudié. En effet, il est possible d'éviter des associations erronées comme *application–appliquer* (dans laquelle *appliquer* est morphologiquement relié à *application* mais sans lien sémantique dans le domaine de l'informatique), puisque le mot *appliquer* n'a pas un haut coefficient de spécificité. Ainsi, pour extraire des couples morphologiquement liés, pertinents dans le domaine étudié, et les annoter avec leur(s) FL, les 402 règles obtenues pour l'expérience 1 ou les 279 de l'expé-

rience 2 sont appliquées à chaque couple de mots du corpus d'informatique ayant une spécificité supérieure à un certain seuil.

Notons qu'outre l'utilisation de TERMOSTAT pour nous focaliser sur des termes du domaine, l'exploitation de ses résultats a aussi un effet bénéfique sur la complexité en temps de calcul de notre approche. En effet, la détection d'analogies implique de tester tous les couples possibles de lemmes du corpus avec nos règles ; la complexité est donc en $O(n^2)$ avec n le nombre de lemmes du corpus. Se restreindre aux lemmes ayant une spécificité supérieure à un certain seuil permet de garder n petit et de diminuer considérablement le coût calculatoire de cette recherche d'analogies.

4. Évaluation

4.1. Constitution du jeu de test

Pour évaluer la couverture et la précision des résultats obtenus par notre technique, nous constituons un jeu de test contenant des couples de termes du domaine morphologiquement liés et annotés avec leur FL. La première étape consiste à sélectionner au hasard plus de 220 mots dans la liste des lemmes du corpus d'informatique. Ensuite, pour chacun de ces mots test, on constitue des couples en recherchant tous les mots morphologiquement liés apparaissant dans le corpus. Cependant, seuls sont retenus les couples répondant aux deux conditions suivantes : les deux mots du couple sont des termes du domaine et les deux mots partagent un lien sémantique pertinent dans le domaine. Ceci signifie qu'un couple comme *découvrir–découverte* n'est pas considéré intéressant puisqu'aucun de ses constituants n'est un terme, de même que le couple *référentiel–référencer* puisqu'il n'y a pas de lien sémantique entre ces deux mots dans le domaine de l'informatique. Finalement, on assigne sa ou ses FL à chaque couple retenu.

Le tableau 2 donne quelques informations sur le jeu de test ainsi construit. Bien entendu, pour éviter de biaiser les résultats, aucun des termes retenus dans ce jeu de test n'a servi lors de nos phases d'apprentissage.

Nombre total de mots test	222
Nombre total de couples	469
Nombre total de liens différents (FL)	50

TAB. 2 – Informations sur le jeu de test

4.2. Mesures de performances et résultats

Pour évaluer les performances de notre système et comparer les résultats de nos deux expériences, nous utilisons l'approche standard de calculs des taux de rappel R et de précision P. La qualité globale du système est donnée par une unique mesure, la f-mesure (la moyenne harmonique du taux de rappel et de précision), définie par : $f = \frac{2*P*R}{P+R}$.

Pour chacune des deux expériences, le processus d'évaluation est le suivant : on applique les règles apprises (en utilisant le jeu d'exemples idoine) à tous les couples de lemmes du corpus possibles tels que les deux lemmes aient une spécificité supérieure à un certain seuil et qu'au moins l'un d'eux soit un des 220 mots test. Un couple validé par une des règles est en analogie avec l'un des exemples et reçoit la même FL. La liste de couples ainsi annotés est comparée

avec celle obtenue manuellement (*cf.* section précédente) ; les taux de rappel et de précision ainsi que la f-mesure sont calculés. Ce processus d'évaluation est ainsi répété pour différents seuils de spécificité pour évaluer l'influence de ce paramètre.

La variation de R, P et f en fonction du seuil de spécificité est présentée respectivement pour l'expérience 1 et 2 dans les figures 1 et 2. Le tableau 3 présente quant à lui les résultats obtenus pour le seuil maximisant la f-mesure.

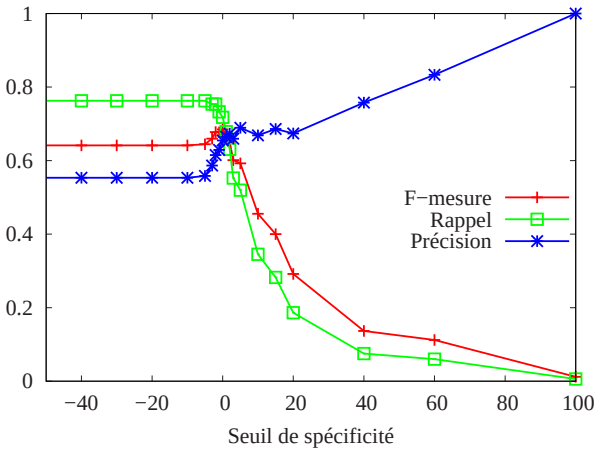

FIG. 1 – Expérience 1 : variation de R, P et f selon le seuil de spécificité

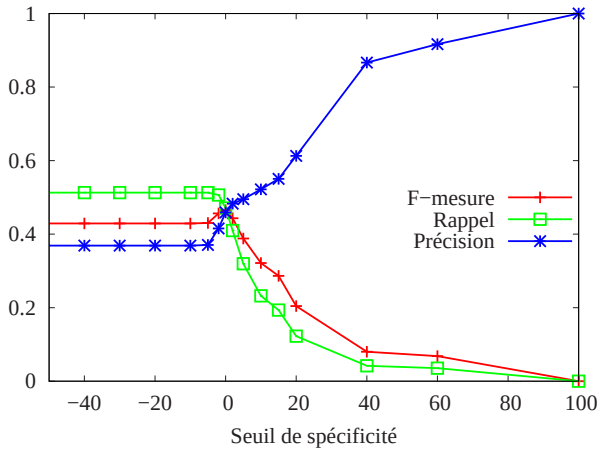

FIG. 2 – Expérience 1 : variation de R, P et f selon le seuil de spécificité

	Seuil de spécificité	**f-mesure**	**Rappel**	**Précision**
Expérience 1	0	0.6848	71.77%	65.48%
Expérience 2	0	0.4650	47.10%	45.91%

TAB. 3 – F-mesure, rappel et précision au seuil optimal

4.3. Discussion des résultats

À l'examen des données précédentes, il apparaît que les résultats de l'expérience 1 sont très bons, aussi bien en terme de rappel qu'en terme de précision, ce qui est surprenant au vu de la simplicité de notre approche. De plus, comme attendu, on obtient un meilleur compromis rappel/précision (*i.e.* une meilleure f-mesure) en se concentrant sur les spécificités positives qu'en examinant la liste complète des lemmes du corpus ; ces résultats sont donc cohérents avec la façon dont fonctionne l'extracteur de termes TERMOSTAT. Par ailleurs, ces très bons résultats confirment clairement le bien-fondé de notre première hypothèse.

Comme on pouvait également s'y attendre, les résultats obtenus dans l'expérience 2 sont moins bons que ceux de l'expérience 1. À tout seuil de spécificité, les taux de rappel et de précision sont nettement inférieurs à ceux obtenus précédemment, mais ils sont cependant assez élevés pour être exploitables ensuite par un terminologue. On note aussi que cette fois encore, le seuil optimal (en terme de f-mesure), égal à 0, est cohérent avec le fonctionnement de TERMOSTAT.

Ces résultats plus faibles s'expliquent par l'inadéquation partielle des données du DiCo pour rendre compte de phénomènes propres au domaine de l'informatique. Ainsi, la relation sémantique existant entre *donnée* et *méta-donnée*, assez fréquente en informatique (*méta-index*, *méta-*

programmation, *méta-tag*...), est absente du DiCo, tout comme la FL De_manière_automatique (*détection–auto-détection*, *extractible–auto-extractible*). À l'inverse, certaines configurations morphologiques issues de la langue générale, comme A_1(*indignation*) = *indigné* (l'agent est + le sens du mot clé) génèrent des règles d'analogie (ici, $A_1(m_1) = m_2$ si $m_1 -_{suf}$ "ation" $+_{suf}$ "é" $= m_2$) produisant du bruit dans le domaine de l'informatique où une telle configuration sera plutôt révélatrice de la FL A_2 (le patient est + sens du mot clé) comme dans *programmation–programmé* ou *administration–administré*.

Dans les deux expériences, deux sortes d'erreurs sont faites par notre technique. Il s'agit d'une part des faux positifs, c'est-à-dire de couples détectés à tort comme étant liés sémantiquement. Ces cas sont principalement dus à la détection de couples valides mais associés à une mauvaise FL. Par exemple, beaucoup d'erreurs sont causées par les mots en *-eur* qui peuvent être des instruments typiques, comme *éditeur*, ou agent, comme *programmeur*, du verbe dont ils sont dérivés. Le deuxième type d'erreurs est l'absence de détection d'un couple valide, c'est-à-dire les faux négatifs. Ceux-ci sont principalement dus à l'absence de l'un des deux termes dans la liste des spécificités, ou bien à des configurations morphologiques rares n'apparaissant pas dans nos exemples, qu'ils soient tirés du DiCoInfo ou du DiCo. C'est le cas par exemple de la FL S_0Inter(*connecter*) = *interconnexion*, apparaissant dans le jeu de test.

Finalement, les résultats peuvent être présentés au terminologue sous forme de graphes dont un extrait est proposé en figure 3. Cet extrait a été produit à l'aide des règles d'analogie issues de l'expérience 1 ; les FL erronées y apparaissent en pointillés longs et les FL manquantes en pointillés courts.

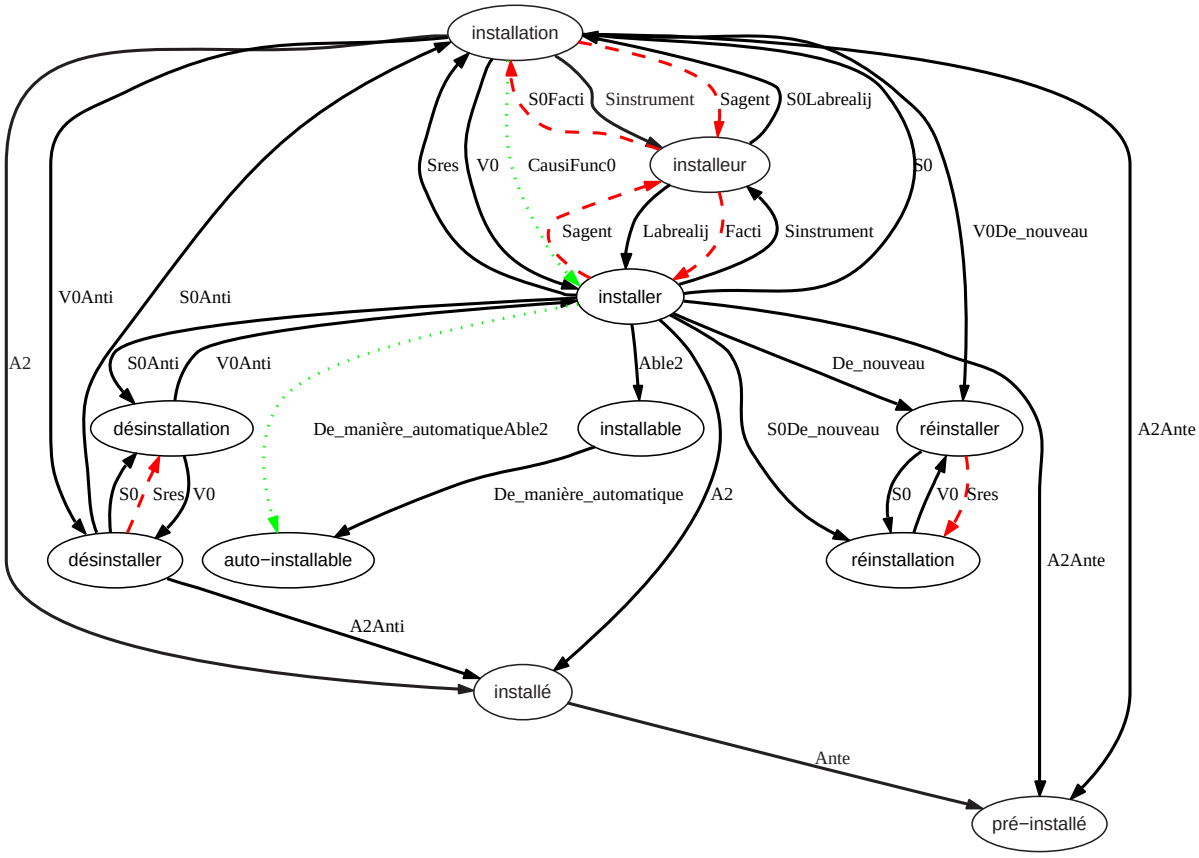

FIG. 3 – Graphe des relations sémantiques de la famille morphologique *installer*

5. Conclusion et perspectives

Cet article présente une technique simple mais originale visant à détecter les mots morphologiquement dérivés et à étiqueter les relations sémantiques existant entre eux à l'aide de Fonctions Lexicales (FL). Cette technique utilise une approche particulière de l'apprentissage artificiel, l'apprentissage par analogie, et exploite les résultats d'un système d'extraction de termes. S'appuyant uniquement sur des exemples de couples de termes morphologiquement et sémantiquement liés, elle ne nécessite aucune connaissance morphologique externe (comme des règles ou des bases de données morphologiques) et peut donc être utilisée pour n'importe quel domaine pourvu que quelques relations sémantiques entre dérivés morphologiques aient été collectées pour servir d'exemples. Les résultats obtenus à partir d'exemples tirés du domaine sont particulièrement bons, aussi bien en terme de rappel qu'en terme de précision de relations sémantiques trouvées. Les résultats issus de la deuxième expérience, utilisant des exemples tirés d'une base lexicale de la langue générale, sont de moins bonne qualité, mais certainement suffisants pour faciliter le travail d'un terminologue.

À travers les deux expériences décrites, nous avons aussi confirmé notre première hypothèse selon laquelle une proximité morphologique indique souvent une proximité sémantique, notamment dans des domaines spécialisés. Par ailleurs, en ce qui concerne notre deuxième hypothèse, la comparaison des résultats de ces deux expériences souligne l'intérêt d'utiliser des ressources endogènes pour ce type de tâches, celles-ci permettant d'obtenir des résultats bien meilleurs qu'en utilisant des ressources généralistes. Cette comparaison se veut donc un apport chiffré dans le débat toujours actif concernant l'utilisation de telles ressources généralistes et/ou exogènes pour des applications dans des domaines spécialisés.

Plusieurs perspectives sont envisagées à ce travail. Nous prévoyons notamment d'intégrer à cette technique d'autres approches de l'acquisition de relations sémantiques sur corpus (Claveau & L'Homme, 2004). Certaines des erreurs fréquentes exposées en section 4.3, notamment celles concernant des FL syntagmatiques, pourraient ainsi être résolues en s'appuyant notamment sur la relation syntaxique entre le mot-clé et son collocatif. D'un point de vue applicatif, nous projetons par ailleurs d'utiliser cette méthode sur un corpus d'informatique anglais pour assister le développement de la version anglaise du DiCoInfo.

Remerciements

Les auteurs tiennent à remercier Léonie Demers-Dion pour l'examen des données nécessaires au jeu de test et Alain Polguère pour la mise à disposition des données du DiCo et son aide précieuse sur l'utilisation des FL.

Références

BINON J., VERLINDE S., DYCK J. V. & BERTELS A. (2000). *Dictionnaire d'apprentissage du français des affaires*. Paris: Didier.

CLAVEAU V. & L'HOMME M.-C. (2004). Discovering specific semantic relationships between nouns and verb in a specialized french corpus. In *Proceedings of the 3rd International Workshop on Computational Terminology, CompuTerm'04*, Genève, Suisse.

DAILLE B. (2003). Conceptual structuring through term variation. In *Workshop on Multiword Expressions. Analysis, Acquisition and Treatment. Proceedings of ACL 2003*, Sapporo, Japon.

B. DAILLE, K. KAGEURA, H. NAKAGAWA & L.-F. CHIEN, Eds. (2004). *Terminology. Special issue on Recent Trends in Computational Terminology*, volume 10. Amsterdam/Philadelphie: John Benjamins Publishing Company.

DROUIN P. (2003). Term-extraction using non-technical corpora as point of leverage. *Terminology*, **9**(1).

GRABAR N. & ZWEIGENBAUM P. (2002). Lexically-based structuring ; some inherent limits. In *Proceeding of the 2nd workshop on Computational Terminology, CompuTerm'02*, Taipei, Taiwan.

J. L. KOLODNER, Ed. (1993). *Machine Learning, special issue on Case-Based Reasoning*, volume 10. Dordrecht: Kluwer Academic Publishers,.

LEMAY C., L'HOMME M.-C. & DROUIN P. (2005). Two methods for extracting specific single-word terms from specialized corpora: Experimentation and evaluation. *International Journal of Corpus Linguistics*, **10**(2).

LEPAGE Y. (2003). *De l'analogie ; rendant compte de la communication en linguistique*. Thèse d'habilitation (HDR), Université de Grenoble 1, Grenoble, France.

LEPAGE Y. (2004). Lower and higher estimates of the number of "true analogies" between sentences contained in a large multilingual corpus. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*, Genève, Suisse.

L'HOMME M.-C. (2004). A lexico-semantic approach to the structuring of terminology. In *Proceedings of the 3rd International Workshop on Computational Terminology, CompuTerm'04*, Genève, Suisse.

MEL'ČUK I., ARBATCHEWSKY-JUMARIE N., ELNITSKY L., IORDANSKAJA L., LESSARD A., DAGENAIS L., LEFEBVRE M.-N., MANTHA S. & POLGUÈRE A. (1984–1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques*, volumes I-IV. Montréal: Les Presses de l'Université de Montréal.

NAMER F. & ZWEIGENBAUM P. (2004). Acquiring meaning for french medical terminology: contribution of morpho-semantics. In *Proceedings of the Conference Medinfo 2004*, San-Francisco, États-Unis.

POLGUÈRE A. (2000). Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for french. In *Proceedings of EURALEX 2000*, p. 517–527, Stuttgart, Allemagne.

ZWEIGENBAUM P. & GRABAR N. (2000). Liens morphologiques et structuration de terminologie. In *Ingénierie des connaissances*, p. 325–334, Toulouse, France.

ZWEIGENBAUM P. & GRABAR N. (2003). Learning medical words from medical corpora. In *Proceedings of the 9th Conference on Artificial Intelligence in Medecine, AIME'03*, Protaras, Chypre.

Annexes

Annexe A

Le tableau suivant présente les fonctions lexicales encodées dans l'entrée ADMINISTRATEUR 1 dans la base terminologique de l'informatique DiCoInfo. Certaines entrées du DiCoInfo sont consultables à partir de l'URL http://www.olst.umontreal.ca/dicoinfo.

ADMINISTRATEUR 1, n. m. *administrateur de patient{réseau 1, système 1}*		
FL	**Valeur**	**Explication**
$Syn_{\cap}$	*client 1*	Intersection de sens
$S_{patient}$	*réseau 1*	Nom du patient
$S_{patient}$	*système 1*	Nom du patient
$Fact_1$	*l'~administre 1 le patient*	Le mot-clé intervient sur le patient
$Fact_1$	*l'~gère 1 le patient*	Le mot-clé intervient sur le patient
S_0Fact_1	*administration 1a du patient par l'~*	Nom pour le mot-clé intervient sur le patient
S_0Fact_1	*gestion 1a du patient par l'~*	Nom pour le mot-clé intervient sur le patient

Annexe B

Le tableau suivant présente les fonctions lexicales standard encodées pour la lexie ABCÈS 1 dans le Dictionnaire des Collocations (DiCo) (Polguère, 2000).

ABCÈS 1, n. m. *~ DE être animé X SUR partie-2→corps Y(X)*		
FL	**Valeur**	**Explication**
Magn	*énorme, gros*	Gros
$IncepFunc_2$	*se forme*	A. se développe
$Oper_{12}$	*souffrir, avoir*	[X] avoir un A.
QSyn	*bouton, phlegmon, furoncle, chancre, bubon*	→
$LiquFunc_0$	*drainer, vider, percer, ouvrir, inciser, crever, débrider*	[Qqn.] vider un A.
IcepPredPlus	*mûrir*	A. se développe
$FinFunc_0$	*se vider, s'épancher, se percer, perce, s'ouvrir, crever*	A. s'ouvre