Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français

Nordine Fourour
Institut de Recherche en Informatique de Nantes - Université de Nantes
2, chemin de la Houssinière - BP 92208 - 44322 Nantes Cedex 3, France
fourour@irin.univ-nantes.fr

Mots-clefs – Keywords

Entités nommées, reconnaissance incrémentielle, apprentissage, surcomposition référentielle
Name Entities, incremental recognition, learning processing, referential composition

Résumé - Abstract

Cet article présente une étude des conflits engendrés par la reconnaissance des entités nommées (EN) pour le français, ainsi que quelques indices pour les résoudre. Cette reconnaissance est réalisée par le système Nemesis, dont les spécifications ont été élaborées conséquemment à une étude en corpus. Nemesis se base sur des règles de grammaire, exploite des lexiques spécialisés et comporte un module d'apprentissage. Les performances atteintes par Nemesis, sur les anthroponymes et les toponymes, sont de 90 % pour le rappel et 95 % pour la précision.

This paper presents an investigation of the conflicts generated by the recognition of the French proper names (PN), and some indications to solve them. This recognition is carried out by the Nemesis system, whose specifications have been elaborated through corpus investigation. Nemesis is grammar-rule based, uses specialised lexicons, and includes a learning module. The system performance, evaluated on the categories composing the anthroponym and the toponym classes, achieves 95 % precision and 90 % recall.

1 Introduction

La reconnaissance des noms propres français est un problème qui se pose dans les différents domaines du traitement automatique de la langue naturelle (TALN) : veille technologique, indexation de textes ou traduction (Daille & Morin, 2000). Cette reconnaissance a été convenablement réalisée en extraction d'information (EI) pour des textes journalistiques anglais (précision et rappel supérieurs à 90 %) (MUC-7, 1998).

Pour le français, comme pour l'anglais (Wacholder *et al.*, 1997), cette reconnaissance se heurte aux problèmes d'ambiguïté liés aux majuscules en début de phrase (Mikheev, 1999) et à la localisation des limites à droite du nom propre : modification adjectivale et attachement des prépositions et des coordinations, possibilité que certains noms propres soient composés en quasi totalité de mots en minuscules. En plus de cette ambiguïté, se posent les problèmes de surcomposition : une EN complexe peut contenir une EN d'une autre catégorie référentielle (e.g. *Guerre d'Algérie*, *Université de Nantes*). Pour traiter ces problèmes, nous avons étudié les conflits qu'ils engendraient sur la reconnaissance des EN et proposons des heuristiques pour les traiter.

Après une présentation de l'architecture logicielle de Nemesis (cf. section 2) précisant le rôle des lexiques et le formalisme retenu pour la conception des règles, nous nous penchons sur la désambiguïsation des conflits posés par l'identification des bornes des entités nommées (EN) et proposons deux méthodes pour l'inférence de nouvelles règles (cf. section 3). Enfin, nous présentons nos conclusions et les perspectives qu'ouvre notre travail (cf. section 4).

2 Architecture logicielle

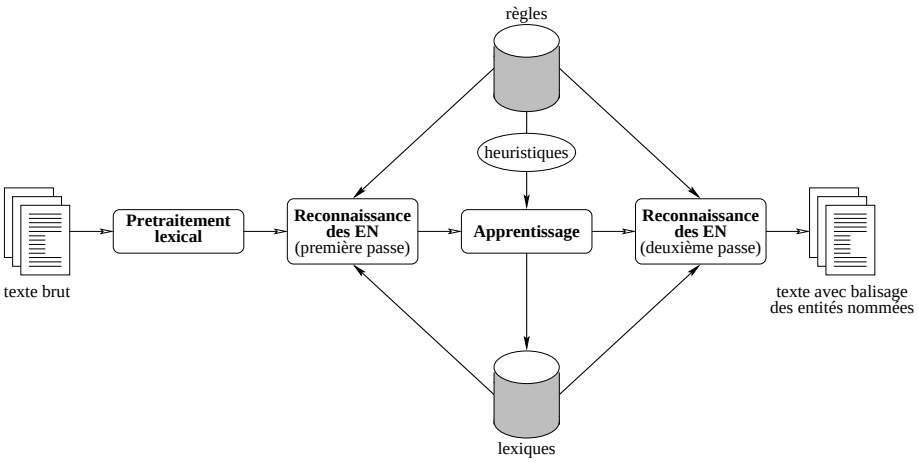

Figure 1: Architecture du système Nemesis

Les spécifications de notre système de reconnaissance des entités nommées ont été réalisées à la suite d'une étude en corpus (Daille *et al.*, 2000) et s'appuient sur des critères graphiques et référentiels. Après avoir montré, dans cette étude, les limites de la catégorisation à quatre classes exploitée dans le cadre des conférences MUC (MUC-7, 1998), nous avons présenté une typologie générale la plus complète possible, basée sur celle de Grass (2000).

Nemesis, élaboré conséquemment à cette étude, est un système qui permet l'identification des bornes des EN, ainsi que leur catégorisation selon cette typologie. Son architecture, présentée à la figure 1, se compose de quatre modules qui effectuent un traitement séquentiel immédiat des données : prétraitement lexical, première reconnaissance des entités nommées, apprentissage et seconde reconnaissance des entités nommées.

2.1 Prétraitement lexical

Il s'agit d'un processus à deux phases : segmentation du texte en phrases et en formes, puis association des sigles et de leur forme étendue, étape basée sur les travaux de Morin (1999). Wolinski *et al.* (1995) et Wacholder *et al.* (1997) utilisent l'association entre les sigles et leur forme étendue, mais uniquement pour les coréférences en ce qui concerne Wacholder *et al.* (1997).

2.2 Première reconnaissance

2.2.1 Projection des lexiques

La projection a lieu en trois étapes : 1) passage du texte en fichier inverse (Salton & McGill, 1983) pour limiter les accès disque ; 2) projection : les étiquettes sémantiques liées aux lexiques sont associées aux différentes formes du texte ; 3) étiquetage des mots commençant par une majuscule et absents des lexiques par *NP*.

Il a été démontré que l'utilisation de lexiques spécialisés était la base de tout système de reconnaissance des noms propres (McDonald, 1994; Wakao *et al.*, 1996). Nos lexiques ont été construits soit manuellement, soit automatiquement, en exploitant des ressources textuelles (pages *Web*, etc.). Les éléments composant ces lexiques peuvent tenir un ou plusieurs rôles : `EN` (l'élément est une entité nommée connue : *OMS*, *Alexandre*, *Canal+*), `mot déclencheur` (l'élément fait partie de l'entité nommée : *Fédération*, *Boulevard*), `contexte` (l'élément appartient au contexte gauche immédiat de l'EN, mais ne fait pas partie de celle-ci : *philosophe*, *français*), `fin d'EN` (l'élément est la dernière forme composant l'entité nommée : *football*, *régional*), `élément d'EN` (il s'agit de tout les éléments lexicaux pouvant faire partie de l'EN, mais sans en permettre la délimitation ou la catégorisation).

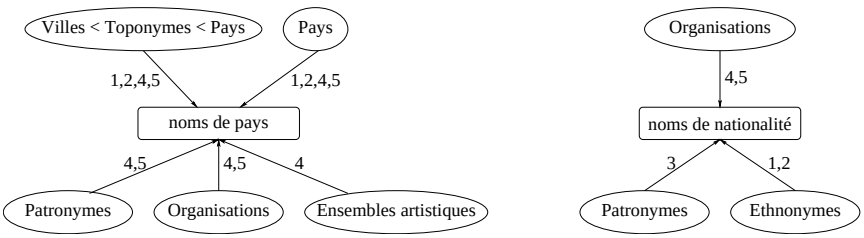

Figure 2: Deux lexiques et leur rôles selon les catégories référentielles

Nous pouvons donc assigner des rôles à nos lexiques, en fonction des catégories référentielles dans lesquelles ils sont utilisés. Cette assignation peut être visualisée sous deux angles : en prenant comme point central, soit une catégorie référentielle (e.g. la reconnaissance des patronymes utilise les éléments du lexique des noms de pays comme `fin d'EN` ou `élément`

d'EN (cf. figure 3)), soit un lexique (les éléments du lexique des noms de pays sont utilisés uniquement comme fin d'EN pour la reconnaissance des ensembles artistiques (cf. figure 2)).

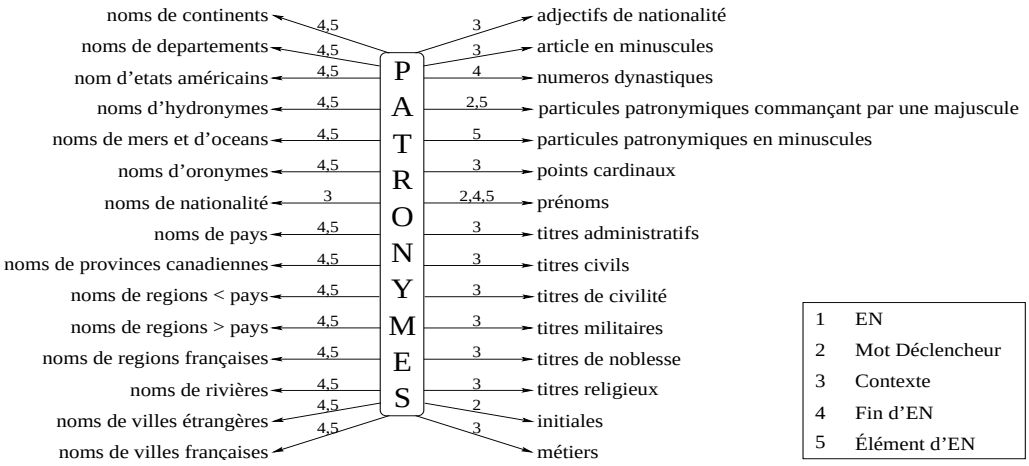

Figure 3: Rôles des lexiques pour la reconnaissance des patronymes

Chaque catégorie référentielle utilise un nombre réduit de lexiques (cf. tableau 1). Sur nos 45 lexiques, regroupant 60 761 éléments, seuls les patronymes et les organisations utilisent plus de 10 lexiques. Ce nombre plus important s'explique par la grande variété de mots pouvant composer les EN de ces deux catégories.

ANTHROPONYMES					
Patronymes	29	Prénoms	1	Ethnonymes	3
Organisations	45	Ensembles artistiques	6		
TOPONYMES					
Toponymes > Pays	5	Pays	1	Villes < Toponymes < Pays	10
Villes	2	Microtoponymes	2	Hydronymes	4
Oronymes	3	Rues	2	Édifices	5

Table 1: Nombre de lexiques utilisés par catégorie référentielle

2.2.2 Application des règles

Une fois la projection des lexiques réalisée, les règles lexico-sémantiques de réécriture sont appliquées, afin de permettre une première reconnaissance des entités nommées. Ces règles s'appuient essentiellement sur l'évidence interne définie par McDonald (1994) et utilisent des patrons basés sur les étiquettes sémantiques correspondant aux lexiques. Voici le formalisme retenu pour la conception de ces règles :

- X, une règle de réécriture ;
- $P_X \rightarrow catégorie_X$, la forme générale de X ;
- P_X, le patron de X ;
- $catégorie_X$, la catégorie référentielle de la balise à poser par X (Patronyme, Pays, etc.) ;
- $[w_X^1, w_X^n]$, un intervalle discret, représentant P_X, avec w_X^1 et w_X^n les éléments de début et de fin (resp.) du patron ;
- w_X^i, le $i^{ème}$ élément de P_X, pouvant être :
  forme une forme quelconque (e.g. la forme *équipe* peut être directement recherchée) ;

balise une étiquette sémantique référant à une forme appartenant à un lexique (e.g. un nom de pays, un métier, un mot clé d'organisation, etc.) ;

np une étiquette référant à une forme n'appartenant pas à un lexique, mais commençant par une majuscule.

Au niveau de l'implémentation (sous forme d'expressions régulières), nous distinguons les **balises** et les **np** des **formes** en faisant précéder les premiers d'un dollar. À chaque w_X^i peut être associé un quantificateur : `?` (0 ou 1 fois), `+` (1 fois ou plus), `*` (0 fois ou plus). La partie de P_X à baliser est mise entre crochets. Certaines règles sont « factorisées » : si plusieurs règles ont exactement la même forme à un w_X^i prêt, elles sont unifiées en remplaçant le w_X^i différent par une variable représentant l'un ou l'autre ($A\ B\ C \rightarrow Cat_1$ et $A\ B\ D \rightarrow Cat_1$ donnent $A\ B\ CouD \rightarrow Cat_1$).

À chaque w_X^i peut être associé un rôle. Les premiers éléments des patrons regroupent les formes possédant les rôles de `contexte` ou de `mot déclencheur`, alors que les derniers ont plutôt pour rôle `fin d'EN`. Ces derniers sont donc moins « fiables », car ils ne permettent pas de catégoriser les EN, ni d'en identifier la présence, mais simplement d'en définir la limite à droite. Quant aux éléments de type `élément d'EN`, ils sont encore moins fiables de par leur nature. Nous avons conçu 53 règles de cette forme, comme :

`$Clé_hydro $Article_min+ [$NP+]` → `Hydronyme`
(*rives de la Kamogawa* ou *eaux du Yangtze Kiang*).

`$Titre_militaire $Adj_nationalité? [$Prénom* $NP]` → `Patronyme`
(*commandant Massoud* ou *général bosno-serbe Ratko Mladic*).

2.3 Apprentissage et seconde reconnaissance

La mise à jour des lexiques a été abordée par certains systèmes (Poibeau, 1999; Cucchiarelli *et al.*, 1998), mais pose encore un grand nombre de problèmes. La mise en place d'un tel module vise deux objectifs : la résolution de certaines coréférences et l'identification de nouvelles entités.

Pour améliorer les performances de notre système, nous avons mis au point une méthode basée sur des heuristiques, afin de créer automatiquement de nouveaux lexiques (Fourour, 2001).Ces heuristiques sont au nombre de 22 pour les patronymes et 3 pour les toponymes.

Après avoir été ainsi obtenus, ces lexiques sont de nouveau projetés sur le corpus (c.f. section précédente).

La mise en place de ce processus apporte une amélioration non négligeable aux performances de Nemesis (+0,6 % pour la précision et +5,3 % pour le rappel) tout en étant très fiable (Fourour, 2001).

3 Conflits

Les problèmes de conflits engendrés par la reconnaissance des entités nommées sont récurrents dans les systèmes que nous avons rencontrés. Poibeau (1999), s'il définit trois heuristiques pour l'application de ses règles de réécriture, ne gère pas la résolution des conflits. Trouilleux (1997)

lui gère, *a priori*, quelques cas d'ambiguïté (noms de lieu qui peuvent entrer dans la composition de patronymes ou d'organisations) et de composition (pour les organisations). Friburger (2001) définit, *a priori*, un ordre à l'application de ses transducteurs, afin de traiter le cas de l'ambiguïté de composition entre les noms de personne et d'organisation.

À partir des règles lexico-sémantiques de réécritures élaborées, nous avons étudié les problèmes que leur application peut engendrer : au niveau conceptuel tout d'abord (sur les patrons de règles), puis, au niveau expérimental (sur les entités nommées reconnues).

3.1 Étude conceptuelle

Nous avons analysé l'ensemble de nos règles, afin d'identifier les différentes façons dont elles pouvaient entrer en conflit et comment réduire *a priori* la réalisation de ces conflits sur les EN reconnues (ambiguïtés sur l'identification ou la catégorisation).

Prenons comme exemple les deux règles suivantes :

$$[w_1^1, w_1^{n_1}] \rightarrow catégorie_1 \quad (1)$$
$$[w_2^1, w_2^{n_2}] \rightarrow catégorie_2 \quad (2)$$

3.1.1 Chevauchement de patrons

Définition 1 *Il y a un conflit de chevauchement de patrons de (2) sur (1)* $\Longleftrightarrow \exists\ i,j \in \mathbb{N}^2$ *(*$1 < i \leq n_1$ *et* $1 \leq j < n_2$*) tels que* $[w_1^i, w_1^{n_1}] = [w_2^1, w_2^j]$.

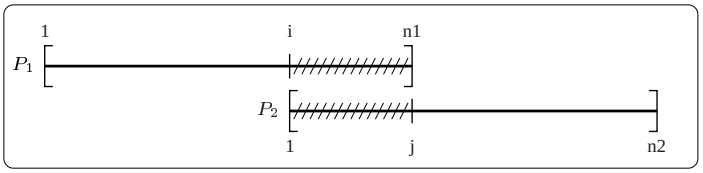

Figure 4: Chevauchement de (2) sur (1)

Les conflits de ce type (cf. figure 4) apparaissent à 43 occasions (moins de 1,6 % des conflits possibles par calcul combinatoire) et concernent 26 règles différentes (sur 53 règles). Pour chacune de ces 43 occurrences, le chevauchement n'a lieu que sur un seul élément des patrons (cas où $i = n_1$ et $j = 1$) qui est de type `mot déclencheur` ou `EN` pour (2) et `fin d'EN` ou `EN` pour (1). Par conséquent, les conflits de chevauchement de patrons n'ont que peu de risques de se concrétiser sur les entités nommées reconnues lors de l'application des règles de réécriture, il semble peu probable de rencontrer en corpus une suite de formes qui corresponde à la suite d'éléments de patron $(w_1^1, \ldots, w_1^{n_1} = w_2^1, \ldots, w_2^{n_2})$: ${}_1$(*Europe de l'* ${}_2$(*Est*)$_1$ *de l'Afrique*)$_2$ ou ${}_1$(*XV de* ${}_2$(*France*)$_1$ *méridionale*)$_2$.

Au niveau de la conception de ces règles, nous ne gérons pas encore la résolution de conflit, mais pour limiter leur apparition nous réordonnons les règles en terme de priorité : lorsqu'un conflit de chevauchement de patrons est identifié entre (1) et (2), nous choisirons d'appliquer en priorité la règle qui possède le plus grand nombre d'éléments de types `contexte` et `mot déclencheur`, car ils en font *a priori* la règle la plus fiable[1] (cf. section 2.2.2). Poibeau (1999)

[1] Un poids plus fort pourrait également être donné aux lexiques utilisés pour une unique catégorie.

donne la priorité aux règles les plus longues, sans se soucier des éléments qui les composent, et, dans le cas de deux règles de même longueur, produit un résultat aléatoire.

3.1.2 Inclusion de patrons

Définition 2 *Il y a un conflit d'inclusion de patrons de (2) dans (1)* $\Longleftrightarrow [w_2^1, w_2^{n_2}] \subseteq [w_1^1, w_1^{n_1}]$.

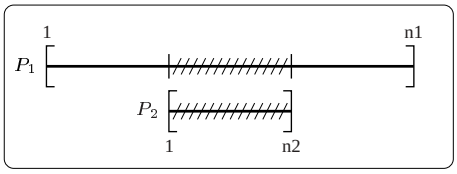

Figure 5: Inclusion de (2) dans (1)

De tels conflits, dont les différents cas sont schématisés à la figure 5, se présentent à 35 reprises (moins de 1,3 % des conflits possibles) et concernent 28 règles différentes. Ils correspondent au cas où le patron d'une règle est totalement inclus dans celui d'une autre règle. Contrairement aux conflits de chevauchement, les conflits d'inclusion vont certainement se concrétiser sur les entités nommées reconnues lors de l'application des règles de réécriture sur corpus. En effet, si P_1 est associé à une série de formes, P_2, qui est inclus dans P_1, sera forcément associé à une série de formes incluse dans la première.

Là encore, la constatation de tels conflits peut nous permettre de réordonner les règles en terme de priorité : s'il existe un conflit d'inclusion de patrons de (2) dans (1), nous choisirons de privilégier (1) sur (2). En effet, comme cette règle a un patron qui possède plus d'éléments, notamment ceux de types `contexte` et `mot déclencheur`, la règle (1) est la plus spécifique et possède un niveau de fiabilité plus grand. Prenons les patrons ${}_2$(`$Prénom* $NP+`)$_2$ et ${}_1$(`$Métier $Adj_nationalité? [$Prénom* $NP+]`)$_1$ qui correspondraient, par exemple, à la suite de formes ${}_1$(*philosophe français* ${}_2$(*René Descartes*)$_2$)$_1$. Le patron de la règle (1) est le plus fiable du fait des éléments de type `contexte` (*philosophe* et *français*). Nous privilégions donc par défaut les entités nommées maximales, tout comme Friburger (2001).

3.2 Étude expérimentale

Cette étude porte sur les anthroponymes et les toponymes, dont l'ensemble des catégories représente plus de 80 % des EN présentes dans nos corpus (Daille *et al.*, 2000). Elle a été réalisée sur un corpus extrait du Monde, composé de 13 368 mots et comportant environ 250 anthroponymes et 80 toponymes. L'observation des conflits qui apparaissent lors de la reconnaissance des EN en corpus nous permet de vérifier si nos règles sont correctement ordonnées. Elle peut éventuellement permettre la mise au point d'heuristiques pour la désambiguïsation des conflits. Parmi ces conflits, nous introduisons un nouveau type (accolement), qui n'existe pas au niveau des règles, mais s'observe lors de leur application sur le corpus.

3.2.1 Chevauchement

Nous n'avons relevé aucun conflit de chevauchement sur les anthroponymes et les toponymes reconnus. Si ce type de conflit apparaissait, il témoignerait très probablement d'un problème

lors de la conception d'une, voire de deux règles. Dans ce cas, différentes solutions s'offriraient alors à nous : 1) éliminer l'une des deux règles voire les deux ; 2) inverser la priorité sur ces deux règles ; 3) fusionner les deux règles ; 4) créer une troisième règle, prioritaire sur les deux premières.

3.2.2 Inclusion

Les conflits d'inclusion de patrons, eux, se concrétisent pendant la phase d'application des règles de réécriture sur un corpus : $_1$(*Université de* $_2$(*Nantes*)$_2$)$_1$, $_1$(*Guerre d'*$_2$(*Algérie*)$_2$)$_1$, $_1$(*Anatole* $_2$(*France*)$_2$)$_1$. Dans ce cas, nous gardons de préférence l'EN maximale, car c'est la plus « fiable ». En effet, les formes qui la composent ont nécessité une correspondance avec un plus grand nombre d'éléments lexicaux de type `mot déclencheur` (*Université*, *Guerre*, *Anatole*).

Par conséquent, la présence de tels conflits parmi les EN d'un corpus ne pose pas de problème d'ambiguïté, mais peut indiquer la présence de surcompositions référentielles, comme dans *Université de Nantes* ou *Guerre d'Algérie*. Pour prendre en compte cette éventualité, une solution serait de se pencher sur l'EN la plus petite : si elle apparaît seule dans le texte, nous considérons que nous nous trouvons devant une surcomposition référentielle.

3.2.3 Accolement

Ce type de conflit n'a de sens qu'en regard des entités nommées reconnues lors de l'application des règles. Il correspond au cas où deux entités nommées sont identifiées, l'une se situant immédiatement à la suite de l'autre. Il ne s'agit pas réellement d'un conflit dans la mesure où les deux EN sont distinctes. Cependant, il peut être intéressant de repérer ces cas, afin de voir si nous n'avons pas commis une erreur. Jusqu'à présent, dans nos corpus, ce cas n'apparaît que pour dix paires d'anthroponymes, soit 3,8 % des EN de cette classe. Il s'agit à chaque fois d'un ethnonyme suivi immédiatement d'un patronyme comme dans *les* $_1$(*Français*)$_1$ $_2$(*Voltaire*)$_2$, *Rousseau et Baudelaire* ou *le* $_1$(*Brésilien*)$_1$ $_2$(*Ricardo Bofill*)$_2$. Dans cette circonstance, il n'y a pas d'erreur ; cependant, nous pourrions avoir affaire à une entité nommée plus complète dont nous aurions mal identifié les limites : $_1$(*Fédération*)$_1$ $_2$(*Française de football*)$_2$ qui devrait être une seule EN, $_1$(*Caisse Primaire d'*)$_1$ $_2$(*Assurance Maladie dans le rouge*)$_2$ qui devrait être regroupée en une seule EN privée de la partie *dans le rouge*[2]. Dans de tels cas, il faudrait étudier la présence de chacune des deux EN séparément dans le texte (cf. tableau 2). Les solutions à apporter peuvent être affinées en ne retenant pas systématiquement l'EN la plus complète, mais éventuellement une sous-partie de celle-ci (comme pour satisfaire à l'exemple précédent).

		Présence de l'EN 1 seule	
		oui	**non**
Présence de l'EN 2 seule	**oui**	prendre les deux séparément	prendre l'EN 2 ou l'EN complète
	non	prendre l'EN 1 ou l'EN complète	prendre l'EN complète

Table 2: Désambiguïsation de l'accolement d'EN

[2]Ces exemples restent hypothétiques, car il n'est notamment pas possible à Nemesis de reconnaître une EN terminant par un déterminant (*d'*).

La découverte de ce type de conflits et les solutions retenues pour les résoudre pourront naturellement nous conduire à modifier nos règles dans les mêmes mesures que pour les conflits de chevauchement de patrons (cf. section 3.2.1) : si l'EN complète est retenue, nous pourrons éventuellement généraliser cette décision et fusionner les deux règles ou en créer une troisième prioritaire sur les deux autres.

3.3 Inférence de nouvelles règles

Dès lors que des situations de conflits sont identifiées entre les EN reconnues après une première application des règles de réécriture, il nous est possible d'en inférer de nouvelles :

processus automatique en créant une règle résultant de la fusion de deux règles. Nous introduisons un opérateur de fusion $\oplus$, inspiré de l'opérateur de fusion de règles morphologiques de Mikheev (1997) :

$$(1) \oplus (2) = (w_1^1, \ldots, w_1^i, w_2^j, \ldots, w_2^{n2}) \rightarrow catégorie_1$$
avec les contraintes $(w_1^{i+1}, \ldots, w_1^{n1}) = (w_2^1, \ldots, w_2^{j-1})$ ou $(i = n_1$ et $j = 1)$

Cet opérateur pouvant être utilisé dans le cas d'un chevauchement ou d'un accolement d'EN, dont nous ne retrouverions pas de présence individuellement ;

processus supervisé en pointant les situations de conflit et en proposant un certain nombre de règles qui pourraient permettre de les résoudre. Dans ce cas, pourraient être proposées différentes solutions pour chaque conflit et une généralisation de la solution retenue par l'utilisateur.

Cette deuxième méthode paraît être la plus efficace et comporter le moins de risques, dans la mesure où nous ne sommes pas du tout sûr des règles inférées par la première méthode et que la deuxième nous permet d'en proposer un plus grand nombre. De plus, l'apport de cette approche est de laisser une plus grande marge de manœuvre à l'utilisateur et de l'intégrer dans le processus de reconnaissance : il n'a plus l'impression d'avoir affaire à une boîte noire, il peut interagir avec le système.

Toutes les remarques faites en ce qui concerne la réévaluation des priorités sur les règles, suivant les formes de conflits qu'elles peuvent engendrer dans leur conception ou leur application, pourront nous permettre de dégager un algorithme d'insertion pour les nouvelles règles induites par un processus d'apprentissage.

4 Conclusions et perspectives

Nous avons présenté l'architecture logicielle de Nemesis, un système de reconnaissance des entités nommées pour le français. Nous nous sommes plus précisément intéressés à l'organisation des lexiques, au formalisme des règles de réécriture et aux conflits que leur application pouvait engendrer. Les résultats, sur l'ensemble des deux classes reconnues, sont de 95 % pour la précision et 90 % pour le rappel. Ces résultats regroupent ceux de Fourour (2001) sur les anthroponymes et une évaluation pour les toponymes avec les mêmes modalités et sur le même corpus (Le Monde, 13 368 mots, 528 anthroponymes et 137 toponymes).

Pour améliorer Nemesis, notre étude des conflits (cf. section 3) devra nous amener à mettre en place un véritable module de désambiguïsation. Parallèlement, elle devra nous permettre d'inférer de nouvelles règles (cf. 3.3) durant la première reconnaissance et de concevoir un algorithme pour les insérer parmi les règles déjà existantes. En outre, il nous faudra implémenter la reconnaissance des autres catégories référentielles de notre typologie (comme certaines sous-classes d'ergonymes et de praxonymes), grâce notamment à des règles de surcomposition référentielle.

Références

CUCCHIARELLI A., LUZI D. & PAOLA V. (1998). Using corpus evidence for automatic gazetteer extension. In *Proceedings of LREC'98*, p. 83–89.

DAILLE B., FOUROUR N. & MORIN E. (2000). Catégorisation des noms propres : une étude en corpus. *Cahiers de Grammaire*, **25**, 115–129.

DAILLE B. & MORIN E. (2000). Reconnaissance automatique des noms propres de la langue écrite : Les récentes réalisations. *Traitement automatique des langues*, **41**(3), 601–621.

FOUROUR N. (2001). Identification et catégorisation automatiques des anthroponymes du français. In *Actes, TALN-Récital 2001*, volume 1, p. 441–450.

FRIBURGER N. (2001). Élaboration d'une cascade de transducteurs pour l'extraction de motifs. In *Actes, TALN-Récital 2001*, volume 1, p. 183–192.

GRASS T. (2000). Typologie et traductibilité des noms propres de l'allemand vers le français. *Traitement automatique des langues*, **41**(3), 643–670.

MCDONALD D. D. (1994). Internal and external evidence in the identification and semantic categorization of proper names. In *Corpus Processing for Lexical Acquisition*, chapter 2, p. 61–76.

MIKHEEV A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics*, **23**(3), 405–423.

MIKHEEV A. (1999). A knowledge-free method for capitalized word disambiguation. In *Proceedings of the 37th Annual Meeting of the ACL*, p. 159–166, University of California, Maryland.

MORIN E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse en informatique, Institut de Recherche en Informatique de Nantes.

MUC-7 (1998). *Proceedings of the 7th seven Message Understanding Conference (MUC-7)*.

POIBEAU T. (1999). Repérage des entités nommées : un enjeu pour les systèmes de veille. In *Actes des troisièmes rencontres de Terminologie et Intelligence Artificielle (TIA'99)*, volume 19, p. 43–51.

SALTON G. & MCGILL M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

TROUILLEUX F. (1997). Identification et classement automatique des noms propres dans des textes français. DEA linguistique, logique et informatique, Université Blaise-Pascal Clermont II.

WACHOLDER N., RAVIN Y. & CHOI M. (1997). Disambiguation of proper names in text. In *Procedings of ANLP'97*, p. 202–208.

WAKAO T., GAIZAUSKAS R. & WILKS Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. In *Proceedings of COLING'96*, volume 1, p. 418–423.

WOLINSKI F., VICHOT F. & DILLET B. (1995). Automatic processing of proper names in texts. In *Proceedings of EACL'95*, p. 23–30.