Repérage des entités nommées : un enjeu pour les systèmes de veille(1)

Cet article présente un système de repérage et d'acquisition d'entités nommées pour le français, c'est-à-dire de repérage des noms de personnes, de géographie, de société, les organisations et les dates. Le système utilise des techniques originales pour l'acquisition d'entités nouvelles et de règles servant à reconnaître ces entités. Il permet de développer rapidement des ressources pour un nouveau domaine et atteindre ainsi des performances quasi comparables à celles de systèmes nécessitant davantage de connaissances *a priori*.

Termes clés : Linguistique sur corpus spécialisés ; outils et applications.

1 Introduction

Le développement de la documentation électronique a révélé le besoin de nouveaux accès au texte. C'est dans ce cadre que l'on a assisté récemment au développement d'analyseurs terminologiques permettant de fournir des bases de connaissances et des index à partir de textes techniques ou spécialisés (Bourigault *et al.*, 1996). De son côté, la veille économique ou stratégique a besoin de pouvoir repérer rapidement les noms d'entreprises et de dirigeants figurant dans les textes. Cette tâche est accomplie par le repérage de ce qui est souvent qualifié d'« entités nommées », c'est-à-dire les noms de personnes, de géographie, de société, les organisations et les dates (Appelt *et al.* 1995) (Gaizauskas *et al.* 1995) (Mani *et al.* 1996) (Wacholder 1996). On voit ainsi apparaître des systèmes d'aide à la décision, permettant aux opérateurs de faire des choix en fonction d'éléments mis en évidence à même le texte, de manière automatique.

Cet article présente un système de repérage et d'acquisition d'entités nommées pour le français. Le système utilise pour partie des techniques classiques pour la reconnaissance d'entités, techniques issues des systèmes d'extraction d'information et d'acquisition de terminologie. Il met également en jeu des techniques d'analyse particulières pour repérer et typer les mots inconnus puis proposer automatiquement des candidats potentiels au statut d'entité nommée. Le système surligne ces séquences à même le texte en utilisant un système de balisage de type hypertexte.

Nous situons d'abord le repérage des entités nommées parmi les nouveaux modes d'accès au texte puis nous présentons le système développé. Celui-ci est basé sur des ressources limitées *a priori* mais possède un module d'acquisition et d'enrichissement interactif de ses données par analyse de corpus. Un modèle de marquage au moyen de balises XML permet un accès assisté au texte. Enfin, une expérience portant sur un corpus issu du journal *Le Monde* donne lieu à une évaluation du processus d'acquisition et du repérage d'entités.

2 Analyse de textes pour l'aide à la décision

Le repérage des entités nommées permet une analyse rapide du texte facilitant la prise de décision. Cette tâche se situe dans la mouvance des nouveaux courants de recherche sur les moyens d'accès au texte.

2.1 L'extraction d'information comme nouveau moyen d'accès au texte

L'extraction d'information désigne l'activité qui consiste à remplir automatiquement une banque de données à partir de textes en langage naturel (Appelt *et al.* 1993) (Pazienza 1997). Il ne s'agit donc pas simplement de filtrage de documents, où le système renvoie un ensemble de textes pertinents par rapport à une

(1) Je remercie Adeline Nazarenko ainsi que les trois relecteurs anonymes de TIA pour leurs commentaires qui ont permis de notablement améliorer cet article. Ma réflexion a également été enrichie de conversations avec Benoît Habert (Limsi), Christian Jacquemin (Limsi) et Célestin Sedogbo (Thomson-CSF).

question. L'extraction met en œuvre une analyse fine du texte pour produire du factuel (remplir un formulaire prédéfini, en anglais *template*) et apporter des réponses précises aux questions des utilisateurs plutôt que du texte brut (Wilks 1997).

Les systèmes d'extraction d'information ont connu un fort développement depuis la fin des années 80 sous l'impulsion des conférences américaines MUC *(Messages Understanding Conferences)*. On assiste toutefois depuis quelques années à un infléchissement dans l'évolution de ce type de système. Une analyse strictement locale ne permet pas d'analyser correctement certains phénomènes linguistiques comme la résolution des anaphores. La portabilité des systèmes est également limitée. L'élaboration d'un système d'extraction passe par la définition de patrons identifiant une information donnée qui pourra ainsi être reconnue dans les textes. Il s'agit d'un travail long et fastidieux qu'il faut refaire pour toute application portant sur un domaine nouveau.

L'extraction d'information pose également question quant au statut à attribuer à l'information extraite. Le but traditionnel de l'extraction, que nous mentionnons ci-dessus, consiste à remplir une base de données à partir de textes en langage naturel. Mais les taux de réussite, qui varient de 50 à 90 % suivant la complexité de la tâche, ne permettent pas de considérer l'information extraite comme certifiée (Hirschman 1997). Diverses expériences qui ont pu être faites au contact de professionnels de l'information montrent un besoin de retour au texte, de vérification de l'information. Souvent, l'expert a besoin du texte original parce qu'il mène à partir du document une démarche d'interprétation particulière. La demande ne va donc pas tant vers des systèmes d'extraction que vers des systèmes de marquage d'éléments textuels pertinents aidant l'interprétation des textes.

On assiste ainsi, essentiellement depuis MUC-6 (1995), à un retour du corpus dans le cadre des systèmes d'extraction. Les systèmes doivent à la fois être centrés sur l'utilisateur (qui définit ce qu'il souhaite extraire) et sur le corpus (qui est sollicité de manière interactive par l'utilisateur). Ce rapport nouveau au texte a été constaté en amont pour la définition de patrons qui se base alors sur une analyse minutieuse du corpus et sur des procédures d'apprentissage automatique. Il est aussi nécessaire, en aval, de permettre les allers et venues entre l'information extraite et le corpus, pour fournir des systèmes d'aide à l'interprétation et à la décision. Par exemple, dans le domaine de la veille économique, mettre en évidence le nom de personnes et d'entreprises dans le texte permet à l'analyste de décider immédiatement si le texte est intéressant ou non pour son activité.

Les technologies d'extraction d'information s'insèrent donc naturellement dans le cadre des nouveaux moyens d'accès au texte. Ces systèmes permettent une navigation au moyen d'hyperliens et un accès à partir de bases de connaissances et d'index structurés (Assadi 1998). Le système d'extraction que nous présentons porte sur le repérage des entités nommées. Ce type d'analyse se révèle indispensable dans certains secteurs comme la veille économique et stratégique, où l'expert doit déterminer dans les dix secondes si le texte est intéressant ou non. Ces entités, en fournissant des indications sur les noms de personnes et d'entreprise en jeu, sur les dates et les lieux, permettent un jugement rapide sur la pertinence et l'importance d'un document.

2.2 Les systèmes d'extraction pour le français

Le projet européen *Écran* (LE-2110, 1996-98) a permis de développer des systèmes d'extraction complets pour plusieurs langues, notamment l'anglais et le français (Basili *et al.* 1998) (Poibeau 1998). C'est dans ce cadre qu'avait été écrit une première version d'un module de repérage des entités nommées du français, dont certains résultats ont été réutilisés lors du développement de notre système. Le système *Écran* souffre toutefois de plusieurs limitations importantes. En particulier, aucune analyse fine des mots inconnus n'est faite, ce qui provoque un nombre d'erreurs relativement important lors du repérage des entités nommées. De plus, aucun processus d'apprentissage n'avait été prévu dans le module développé pour *Écran* : il est dès lors nécessaire de procéder à un travail manuel important pour enrichir les dictionnaires de base lors du développement d'une application portant sur un nouveau domaine.

Écran est, à notre connaissance, le seul système d'extraction complet développé à ce jour pour le français et s'intéressant à des textes écrits[2]. La société Lexiquest a développé une grammaire des noms de personnes et de société. Ce système est en partie paramétrable (insertion de lexiques particuliers) mais il ne permet pas un développement incrémental des ressources. Plusieurs systèmes

(2) *Écran* permet d'extraire des textes non seulement des entités nommées mais aussi d'autres éléments comme les relations entre entités à partir d'une analyse minutieuse du verbe. Le projet *Exibum* (Kosseim & Lapalme, 98) vise à développer un système d'extraction mais ne comporte pas de module d'analyse des entités nommées.

d'analyse locale ont été développés à partir de lexique-grammaire. Maurel (1989) a ainsi développé un système de repérage des dates par automates et tables d'acceptabilité. Plus récemment, Belleil (1997) a présenté un système de repérage des toponymes français et Sénellart (1998) un système de repérage des noms de ministre à partir du journal *Le Monde*. Ces approches sont essentiellement basées sur des dictionnaires exhaustifs du sous-langage concerné. Sénellart propose une méthode interactive d'acquisition à partir d'automates, mais l'approche vise avant tout à constituer un dictionnaire exhaustif sur un domaine restreint.

Le projet *Xicop* (3) (eXtraction d'Information de COrpus de Parole), actuellement en cours de développement au Limsi/CNRS (Orsay), est un système de repérage d'entités nommées pour des corpus audio. S'il partage certains des principes adoptés ici (ensemble de règles permettant le repérage d'entités basé sur l'analyse d'« amorces » et de mots inconnus), en revanche il ne peut s'appuyer sur la distinction minuscule / majuscule et doit supporter des données bruitées. Les techniques que nous présentons sont davantage liées à l'analyse de documents écrits.

3 Un système de repérage d'entités nommées

L'architecture que nous adoptons est héritée des systèmes d'extraction d'information ayant participé à MUC. Ceux-ci procèdent par analyse locale progressivement étendue par un ensemble d'automates à état finis. On parle alors d'automates en cascade (Appelt *et al*. 1993).

(3) http://genesis.limsi.fr/~xicop.

3.1 Première étape : étiquetage lexical

Dans un premier temps, le système procède à la fois à un découpage du texte en unités minimales et à un étiquetage lexical :

- Les nombres sont analysés suivant des critères formels (un nombre est composé de chiffres et, éventuellement, d'un tiret, d'un point ou d'une virgule) ;
- Un dictionnaire de noms propres permet d'identifier et de typer un certain nombre de noms de personnes et de lieux. La reconnaissance est souvent partielle à ce stade (le système peut par exemple avoir reconnu le prénom mais pas le nom). Le dictionnaire est particulièrement développé au niveau des listes semi-fermées comme les prénoms, beaucoup moins pour les noms ;
- Un dictionnaire d'amorces permet de reconnaître certaines séquences importantes pour la suite comme *M.* (pour *Monsieur*) ou *SA* (pour *Société anonyme*) ;
- Un algorithme permet de traiter et de normaliser les sigles comme *I.B.M* ou *I.B.M.* (avec ou sans point à la fin). En revanche, le signe *IBM*, écrit sans point, est considéré comme un mot inconnu s'il n'a pas été stocké préalablement dans le dictionnaire de noms propres ;
- Les mots inconnus sont étiquetés en tant que tels au moyen d'un dictionnaire général de la langue. Ils reçoivent une étiquette différente suivant qu'ils sont en majuscules ou en minuscules, au début de phrase ou non ;
- Enfin, parmi les mots figurant dans le dictionnaire de la langue générale, ceux qui commencent par une majuscule sont étiquetés. Une distinction est faite suivant que le mot se trouve en début de phrase ou non.

Cette analyse est effectuée au moyen d'un arbre de décision. Les choix que doit faire le système s'appliquent dans l'ordre des points énumérés ci-dessus. Une fois cette étape franchie, le texte est enrichi de balises entourant les mots repérés. Nous donnons ci-dessous un texte étiqueté par l'analyseur. Chaque mot reconnu est entouré de l'étiquette qui convient.

<COMPANY> Fiat </COMPANY> possède <NUMBER> 90 </NUMBER> % de <PERSON> <COMPANY>Ferrari </COMPANY></PERSON>.

<COMPANY> Fiat </COMPANY> contrôle désormais <NUMBER> 90 </NUMBER> % du capital de <PERSON> <COMPANY> Ferrari </COMPANY></PERSON>, a annoncé le <NUMBER> 7 </NUMBER> <DATE type=MONTH> septembre </DATE> le groupe automobile <NATIONALITY> italien </NATIONALITY>. <COMPANY> Fiat </COMPANY> qui détenait déjà <NUMBER> 50 </NUMBER> % de la firme de <LOCATION> Modène </LOCATION>, précise qu'il a racheté «ces mois derniers» les <NUMBER> 40 </NUMBER> % qui appartenaient à <PERSON> Enzo </PERSON> <PERSON><COMPANY> Ferrari </COMPANY></PERSON>. L'opération s'est donc déroulée avant le décès du «<UNKNOWN> commandatore </UNKNOWN>», le <NUMBER> 14 </NUMBER> <DATE type=MONTH> août </DATE>. Les <NUMBER> 10 </NUMBER> % restants appartiennent au fils adoptif d' <PERSON> Enzo </PERSON> <PERSON><COMPANY> Ferrari </COMPANY></PERSON>, <TR_PERSON> M. </TR_PERSON> <PERSON>Piero </PERSON> <UFIRSTUNKNOWN> Lardi </UFIRSTUNKNOWN>. - (<UCASEUNKNOWN>AFP <UCASEUNKNOWN>.)

Dans l'exemple précédent apparaissent les principales étiquettes utilisées par le système, c'est-à-dire des noms de personnes (<PERSON>) et de société (<COMPANY>), des nombres (<NUMBER>), des dates (<DATE type=MONTH>), des amorces de noms de personnes (<TR_PERSON>) et des mots inconnus (<UNKNOWN>, <UFIRSTUNKNOWN>, <UCASEUNKNOWN> suivant que le mot est inconnu, qu'il est inconnu et que sa première lettre est en majuscule ou qu'il est inconnu et entièrement en majuscules). On remarque l'ambiguïté de *Ferrari*, qui peut être considéré soit comme un nom de personne, soit comme un nom de société. En conséquence, *Ferrari* est doublement étiqueté: <PERSON><COMPANY> Ferrari </PERSON></COMPANY>.

À la suite de cette étape, le fonctionnement du système ne repose plus que sur l'analyse des suites d'étiquettes, indépendamment des formes effectivement attestées dans le texte.

3.2 Deuxième étape: reconnaissance de séquences pertinentes

Une grammaire des entités nommées permet ensuite de repérer parmi les suites d'éléments repérés à l'étape précédente celles qui sont susceptibles de former une entité.

Cette grammaire est écrite sous forme de règles de réécriture avec comme partie droite une étiquette syntagmatique et comme partie gauche une expression régulière. Les règles sont compilées et s'appliquent en respectant certaines heuristiques:

- Les règles les plus longues (celles qui contiennent le plus d'éléments) s'appliquent les premières,
- Une règle ne peut plus s'appliquer à l'intérieur d'une séquence précédemment reconnue (autrement dit, une séquence reconnue forme par la suite un îlot inanalysable).
- Si deux règles de même longueur peuvent s'appliquer, le résultat est aléatoire (il dépend de l'ordre dans lequel les règles ont été compilées).

Ce principe, en évitant d'avoir à gérer des conflits, permet d'assurer la robustesse du système.

Nous donnons à la suite un exemple de règle. TR_PERSON désigne une amorce de nom de personne (« M.», «Mme»,…), PERSON? un nom ou un prénom optionnel et UFIRSTUNKNOWN+ un ou plusieurs mots inconnus dont l'initiale est en majuscule:

```
// M. Piero Lardi
// M. Lardi
TR_PERSON PERSON?
UFIRSTUNKNOWN+ ==>
PERSON
```

On voit ici que, même si le mot *Lardi* n'est pas connu du système, il est possible d'inférer, d'après cette règle, qu'il s'agit d'un nom de personne et compléter la base de noms propres. La séquence *M. Piero Lardi* forme par la suite un tout étiqueté <PERSON> qui ne peut plus être décomposé. L'évaluation tend à prouver que des règles simples déterminent avec un bon taux de réussite les éléments faisant partie de l'entité et son type (Sénellart 1998).

Il va de soi que le système fonctionne d'autant mieux qu'il a des listes de noms propres relativement complètes dès le début. En l'absence de telles listes, le système pourra faire des prédictions sur les mots inconnus apparaissant dans des règles spécifiques, comme la règle ci-dessus qui permet d'inférer que le mot inconnu représente un nom de personne. En revanche, la tâche sera plus longue pour les mots inconnus isolés, puisque ceux-ci se retrouvent dans des listes très hétérogènes (avec environ 20 % d'éléments pertinents seulement). L'avantage de notre système est d'être robuste et de permettre un fonctionnement en «mode dégradé», c'est-à-dire avec peu de connaissances préalables. Les performances en mode dégradé pourront toutefois être faibles et dépendent étroitement du domaine et des données disponibles (cf. la partie «évaluation», où le système sur un nouveau domaine ne couvre tout d'abord que 20 % des entités présentes).

Un système de préférence est mis en place pour résoudre les ambiguïtés comme pour le mot *Ferrari*. Les règles permettant de regrouper le plus grand nombre d'éléments s'appliquent d'abord. Cette procédure se fait en tenant compte des ambiguïtés et résout la plupart des cas. Par exemple dans le cas de <PERSON> Enzo </PERSON> <PERSON> <COMPANY> Ferrari </COMPANY></PERSON>, l'étiquette <COMPANY> ne sera pas retenue du fait qu'une règle permet de reconnaître *Enzo Ferrari* comme un nom de personne. Quand le contexte ne permet pas de désambiguïser une expression, un système de préférence permet de choisir l'étiquette la plus probable (*Ferrari* est classé comme étant préférentiellement un nom de société) ou, en l'absence d'élément répertorié permettant de décider, le système choisit une étiquette de façon aléatoire. On obtient alors le résultat suivant:

<COMPANY> Fiat </COMPANY> possède 90 % de <COMPANY> Ferrari </COMPANY>.

<COMPANY> Fiat </COMPANY> contrôle désormais 90 % du capital de <COMPANY> Ferrari </COMPANY>, a annoncé le <DATE> 7 septembre </DATE> le groupe automobile <NATIONALITY> italien </NATIONALITY>. <COMPANY> Fiat </COMPANY> qui détenait déjà

50 % de la firme de <LOCATION> Modène </LOCATION>, précise qu'il a racheté «ces mois derniers» les 40 % qui appartenaient à <PERSON> Enzo Ferrari </PERSON>. L'opération s'est donc déroulée avant le décès du « <UNKNOWN> commandatore </UNKNOWN>», le <DATE> 14 août </DATE>. Les 10% restants appartiennent au fils adoptif d'<PERSON> Enzo Ferrari </PERSON>, <PERSON> M. Piero Lardi </PERSON>. - (<UCASEUNKNOWN> AFP </UCASEUNKNOWN>.)

La sortie de cette étape d'analyse est un texte enrichi de nouvelles étiquettes indiquant les entités reconnues et leur type. Dans l'exemple ci-dessus, nous simplifions en effaçant, pour des raisons de lisibilité, une partie des étiquettes posées lors de l'étape précédente. En fait, chaque niveau de marquage rajoute son jeu d'étiquettes sur le texte sans supprimer les étiquettes du niveau précédent. L'utilisateur doit préciser les éléments qu'il souhaite voir apparaître dans l'interface(4).

3.3 Regroupement d'entités co-référentes

Une dernière passe permet de repérer de nouvelles entités, notamment parmi les mots inconnus et ceux qui n'ont pu être typés par une des règles de la grammaire. Ce niveau opère de façon simple en gardant simplement la mémoire des derniers éléments trouvés. On part du principe que les textes analysés ont une certaine cohérence et que celle-ci passe par la reprise des mêmes éléments au sein de périodes dans le texte. Nous entendons ici par période une suite de phrases ou de paragraphes marqués par la reprise d'éléments linguistiques (pronoms et autres anaphores, déictiques, etc.) ou thématiques communs.

<COMPANY id=1> Fiat </COMPANY> possède 90 % de <COMPANY id=2> Ferrari </COMPANY>.

<COMPANY id=1> Fiat </COMPANY> contrôle désormais 90 % du capital de <COMPANY id=2> Ferrari </COMPANY>, a annoncé le <DATE id=3> 7 septembre </DATE> le groupe automobile <NATIONALITY id=4> italien </NATIONALITY>. <COMPANY id=1> Fiat </COMPANY> qui détenait déjà 50% de la firme de <LOCATION id=5> Modène </LOCATION>, précise qu'il a racheté «ces mois derniers» les 40 % qui appartenaient à <PERSON id=6> Enzo Ferrari </PERSON>. L'opération s'est donc déroulée avant le décès du «<UNKNOWN> commandatore </UNKNOWN>», le <DATE id=7> 14 août </DATE>. Les 10 % restants appartiennent au fils adoptif d' <PERSON id=6> Enzo Ferrari </PERSON>, <PERSON id=8> M. Piero Lardi </PERSON>. - (<UCASEUNKNOWN> AFP </UCASEUNKNOWN>.)

Le texte que nous avons pris en exemple ne pose pas de problème particulier pour l'analyse des éléments co-référents. L'ambiguïté sur *Ferrari* en tant que nom de personne ou nom de société a été levée au niveau précédent. Ici, il suffit au système de s'assurer que les éléments co-référents sont de même type. Cette étape de l'analyse permet malgré tout de mettre en évidence, parmi les occurrences d'une chaîne de caractères ambiguë, celles qui réfèrent à un même objet. Ainsi, *Lardi*, qui était à l'origine un mot inconnu commençant par une majuscule (<UPPERUNKNOWN>) peut à l'issue du traitement être typé comme <PERSON> et être inséré automatiquement dans le dictionnaire.

Les pronoms et autres reprises anaphoriques ne sont pas pris en compte par le système à l'heure actuelle. Il ne permet pas non plus de relier des éléments comme *Fiat* et *la firme de Milan*: ceci serait possible mais nécessiterait une base de données gérant les synonymes. Enfin, les entités complexes comme le «fils adoptif d'Enzo Ferrari » ne sont pas analysées.

C'est également à ce niveau que certains mots inconnus peuvent être répertoriés et classés en fonction de leur contexte d'apparition. Un mot inconnu peut être étiqueté et typé correctement s'il peut être mis en rapport avec une entité déjà étiquetée. L'algorithme d'analyse est relativement simple: on co-indice un mot inconnu avec une entité connue si et seulement si elle possède une chaîne de caractère pertinente en commun. Les amorces (« trigger words ») permettant de reconnaître les entités ne sont pas considérées comme des chaînes de caractères pertinentes (<TR_PERSON> par exemple), à l'inverse d'éléments déjà étiquetés comme <PERSON> ou <DATE>.

Dans le cas présent, *AFP* a pu être repéré comme mot inconnu en majuscule. *Commandatore* est repéré comme mot inconnu sans majuscule mais aucun élément du contexte ne permet d'aller plus loin, c'est-à-dire de les typer plus précisément. Si les mots isolés sont ici pertinents (le «commandatore » est le surnom de *Enzo Ferrari*, *l'AFP* est une société), ce n'est évidemment pas toujours le cas.

(4) Le document XML, pour être visualisable dans un navigateur Web, doit au préalable être transformé en document HTML. Le système utilise une feuille de style associée qui permet de ne visualiser qu'une partie des éléments étiquetés.

3.4 Marquage hypertexte des séquences repérées

Les éléments repérés sont marqués au moyen de balises XML. XML est un langage permettant de définir une structure de document au moyen de balises. Ce type de marquage, où l'on peut définir ses propres balises, permet un repérage aisé des entités grâce à un langage de description normalisé. Par ailleurs, XML possèdera à terme ses propres « parseurs » et ne nécessitera pas, en principe, le développement de visualiseurs spécifiques.

Une DTD (Définition de type de document) est définie pour rendre directement compte du marquage vu précédemment. Une DTD permet de définir une grammaire décrivant une classe de documents. La DTD définie pour l'application contient deux blocs principaux d'instructions. Le premier consiste en un marquage minimal rendant compte de la forme du document, essentiellement pour préserver le découpage en paragraphes. Le deuxième bloc rend compte des étiquettes que nous avons vues précédemment pour le repérage des entités nommées proprement dites. La grammaire est stockée sous la forme d'un fichier de règles de réécriture classiques pour des raisons de lisibilité pour les personnes amenées à développer les ressources. Elle est équivalente et conforme à la DTD développée conjointement, au sens où la grammaire est conforme à la hiérarchie des balises définie dans la DTD. Pour l'instant, la cohérence entre la DTD et les règles est maintenue à la main, du fait de la stabilité et de la simplicité de la DTD. Il serait envisageable d'intégrer au système un parseur XML pour assurer cette cohérence de manière automatique.

En l'absence de visualiseurs de document XML au moment de l'implémentation, le système génère *in fine* un document au format HTML 4, avec feuille de style séparée. Le résultat est alors visible dans n'importe quel navigateur Web. Actuellement les entités apparaissent en surbrillance dans le texte original. Le système sera prochainement étendu pour pouvoir réagir de manière dynamique. Il sera alors possible de faire apparaître toutes les occurrences d'une entité donnée quelle que soit sa forme linguistique (entités co-référentes). Une fiche synthétique sera aussi accessible par simple clic sur l'entité afin d'avoir un système interactif, à base d'hyperliens.

4 Expérimentation et évaluation sur un corpus issu du journal *Le Monde*

Dans cette section, nous détaillons le protocole d'expérimentation et les résultats de l'évaluation qui a été menée.

4.1 Constitution du corpus

Un corpus d'évaluation a été constitué à partir de textes concernant les affaires internationales, tirés des archives du journal *Le Monde*. Les textes ont été sélectionnés en faisant une simple requête sur la base avec le mot clé *Affaires internationales*. Ces textes font usage d'un grand nombre de noms d'entités et constituent donc un excellent corpus d'évaluation. Celui-ci est distinct du corpus concernant l'actualité économique qui avait préalablement été utilisé lors du développement du système. Il s'agit, de plus, d'une source fréquemment utilisée en veille stratégique. Pour l'évaluation, nous avons isolé 25 textes d'environ 20 000 mots au total.

4.2 Évaluation du processus d'acquisition de noms d'entités

Nous avons évalué la couverture et la pertinence des mots inconnus relevés automatiquement par l'analyseur à partir d'un dictionnaire noyau de *Écran*. À l'instar des candidats termes en terminologie, on obtient des candidats entités, c'est-à-dire des groupes nominaux candidats au statut d'entité (de la même façon que le système Lexter permet d'obtenir des candidats termes (Bourigault *et al.* 1996). Ces candidats sont réparties en classes plus ou moins homogènes. Les éléments inconnus du dictionnaire qui apparaissent dans des séquences reconnues par les règles de repérage des entités nommées (étape 2, *cf.* point 3.2) forment une classe pertinente à plus de 95 %, comme en témoigne le tableau suivant :

Abdel	Baas
Abdullah	Baden-Baden
AFP	Bagdad
Ahmadi	Baker
Akaba	Balladur
Al	Bassorah
Amalric	Beaucé
Andréani	Bérégovoy
Andreotti	Bin
AP	Blum
Arafat	Bréhier
Aramco	Brent
Arens	Bush
Aziz	Bush-Baker

Mots inconnus ayant été repérés par au moins une règle de la grammaire et proposés comme candidats entités

Cette classe comporte aussi bien des mots correspondant à des noms de personnes qu'à des noms de lieux ou d'organisation. Il n'a pas été jugé utile de pousser plus loin le typage proposé, même si cela devrait être possible en fonction de la règle ayant permis de reconnaître un mot donné. La vérification du type proposé pour

un élément donné risque en effet de se révéler aussi coûteuse qu'un typage purement manuel.

Les classes de mots inconnus apparaissant en dehors de tout schéma identifié sont moins pertinentes. Elles permettent cependant de compléter la couverture du dictionnaire général actuel. Une évaluation manuelle révèle environ 20 % de candidats pertinents pour les entités nommées, 40 % de mots absents du dictionnaire général que l'on peut ainsi compléter et 40 % de segments divers qui ont échappé à l'analyse (notamment les séquences avec tiret comme *a-t-il*, *semble-t-il* ou *quasi-inconditionnel*, avec emploi incorrect du tiret devant l'adjectif).

Les règles de grammaires sont écrites de manière incrémentale, de manière à progressivement couvrir la majeure partie des éléments pertinents. Le développement de la grammaire est interactif: l'introduction de nouveaux éléments dans les dictionnaires permet de reconnaître de manière partielle de nouvelles entités. Leur pleine reconnaissance requiert de nouvelles règles qui à leur tour font apparaître de nouvelles entités... Par exemple, la règle

TR_PERSON PERSON? UFIRSTUNKNOWN+ ==> PERSON

ne permet de reconnaître que partiellement *M. Frederik de Klerk*. Il est alors nécessaire de créer une nouvelle règle de type

TR_PERSON PERSON? PREP UFIRSTUNKNOWN+ ==> PERSON.

Mais cette règle, à son tour, ne permet de reconnaître que partiellement un nom tel que *M. Jean de la Guerivière*. La règle précédente est alors modifiée comme suit:

TR_PERSON PERSON? PREP DET? UFIRSTUNKNOWN+ ==> PERSON.

On peut choisir de rendre aussi la préposition PREP optionnelle pour fusionner la règle avec la première mentionnée ci-dessus. En pratique, on évite d'écrire des règles avec trop d'éléments optionnels pour des raisons de lisibilité. Toute règle doit par ailleurs comporter un élément lexical non optionnel pour éviter de reconnaître une amorce ou une préposition de manière isolée.

Les performances du système sont de l'ordre de 20 % d'entités reconnues au début de l'expérience. Après 3 itérations (analyse du corpus, insertion des mots inconnus repérés par le système dans les différents dictionnaires, ajouts de règles, nouvelle analyse du corpus), un taux de reconnaissance d'environ 90 % est atteint grâce la méthode d'enrichissement incrémental proposée (5). Ces performances ont été atteintes en environ 3 heures de travail sur le corpus d'entraînement.

4.3 Évaluation du système

Une partie du corpus a été réservée pour l'évaluation du système, une fois celui-ci mis au point. Ce corpus a été étiqueté manuellement d'un côté et traité par le système de l'autre. On a enfin procédé à une comparaison des résultats du système avec ceux obtenus par l'analyste humain, en l'occurrence le concepteur du système (6).

Les règles écrites sous la forme d'expressions régulières lors de l'étape précédente sont compilées en un ensemble de règles simples. Dans le cadre de notre expérimentation, le système dispose d'une quarantaine de règles utilisant des métacaractères qui génèrent près de 200 règles simples. On garantit ainsi une analyse de complexité linéaire. Un corpus de 126 Ko a été traité en moins de 30 secondes sur un PC (K6 à 200 MHz), produisant comme résultat un fichier au format XML de 470 Ko.

Le taux de reconnaissance est d'environ 80 % des entités du texte correctement analysées. Les règles sont suffisamment contraintes pour avoir un bruit quasi nul. Les principales causes de reconnaissance partielle ou de silence sont les suivantes:

- Incomplétude de la grammaire ou du dictionnaire (l'expression *M. Valéry Giscard* a été reconnue alors que c'est l'expression *M. Valéry Giscard d'Estaing* qui figurait dans le texte);
- Transformations ayant échappé à l'analyse (*Charette Hervé* au lieu de *Charette Hervé de*, introduire une règle pour reconnaître ce type de transformation reviendrait à introduire énormément de bruit, puisqu'on reconnaîtrait beaucoup de noms de personne suivi de la préposition *de*);
- Orthographe approximative (*Langelier Jean Pierre* pour *Langelier Jean-Pierre*);
- Mot fortement ambigu (*Le Monde* qui dans le cas présent est un nom propre, mais qui est difficilement analysable hors contexte).

Ces résultats se situent légèrement en dessous des scores

(5) L'évaluation de la couverture du système sur le corpus d'entraînement a été faite pour donner un ordre d'idée. Le protocole d'évaluation que nous exposons dans la partie suivante (comparaison des résultats du système avec un étiquetage manuel) a été mis en œuvre, quant à elle, sur la partie du corpus qui n'avait pas servi au développement des ressources du système.

(6) En toute rigueur, un étiquetage par un utilisateur extérieur serait souhaitable.

obtenus par les systèmes anglo-saxons participant à MUC, qui oscillent entre 85 et 95 %. Mais le but de notre système est avant tout d'obtenir un score honorable après un temps d'adaptation limité (ici environ 3 heures), contrairement aux systèmes MUC qui nécessitent un travail manuel important pour développer les ressources les plus complètes possibles. Notons enfin que le processus d'acquisition fonctionne grâce à un système de règles. L'utilisateur peut ainsi facilement contrôler l'activité du système, contrairement aux résultats obtenus par des méthodes statistiques (Cucchiarelli, 98).

Le bon résultat que nous obtenons sur le corpus *Le Monde* devrait être comparé avec des résultats provenant d'autres corpus. Il est certain qu'il y a une corrélation assez forte entre le résultat et le genre textuel, le style ou l'auteur. Ainsi, le repérage des noms propres dans le journal *Le Monde* est facilité par le caractère quasi systématique du *M.* ou *M^me^* qui précède le nom. Il n'en irait pas de même dans un corpus issu du journal *Libération*, qui n'en fait pas un usage aussi systématique. Un travail d'adaptation en fonction du corpus est donc nécessaire.

5 Conclusion

Nous avons présenté un système d'extraction d'entités nommées fonctionnant à partir de ressources limitées. Un processus d'acquisition permet de proposer des éléments susceptibles d'entrer dans la formation de nouvelles entités et d'enrichir semi-automatiquement le dictionnaire. Ce processus d'acquisition permet aussi d'étendre progressivement la grammaire décrivant les entités. Le système développé est donc extrêmement portable d'un domaine à l'autre, voire d'une langue à l'autre. On a montré qu'on obtenait des résultats légèrement inférieurs à ceux des systèmes anglo-saxons ayant participé à MUC en un temps de développement limité. En injectant davantage de connaissances dans notre système, il est possible d'obtenir des résultats comparables.

Notre système se classe dans la famille des outils d'aide à l'accès à la documentation électronique. Il fait partie d'un ensemble plus important d'outils d'extraction d'information à partir de textes en cours de développement. Le but de ce système est de fournir à l'utilisateur les moyens de définir sa requête ainsi que des outils d'exploration interactive de corpus. Nous nous situons également dans la perspective d'une nouvelle ergonomie linguistique en laissant à l'utilisateur le soin de faire les choix qui lui incombent, par rapport à la tâche qu'il cherche à accomplir. Dans ce cadre, le repérage des entités nommées est une aide à la décision, qui doit permettre à l'analyste chargé d'effectuer une veille sur un domaine donné de déterminer rapidement si un document est pertinent ou non. Des applications sont en cours dans le domaine de la veille économique, où les mêmes noms de société et de personnes reviennent régulièrement et sont particulièrement représentatifs.

Poibeau, Thierry,
Thomson-CSF/LCR,
LIPN,
Université Paris-Nord,
Villetaneuse.

Bibliographie

Appelt (D.E.), Hobbs (J.), Bear (J.), Israel (D.), Kameyana (M.) & Tyson (M.), 1993, «FASTUS: a finite-state processor for information extraction from real-world text», dans *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambéry, p. 1172-1178.

Appelt (D.E.), Bear (J.), Hobbs (J.), Israel (D.), Kameyana (M.), Kehler (A.), Martin (D.), Myers (K.) & Tyson (M.), 1995, *The FASTUS name recognition grammar*, Rapport Interne, SRI international.

Assadi (H.), 1998, *Constructions d'ontologies à partir de textes techniques*, Thèse de l'université Paris 6.

Basili (R.), Catizone (R.), Pazienza (M.T.), Stevenson (M.), Velardi (P.), Vindigni (M.) & Wilks (Y.), 1998, «An Empirical approach to lexical tuning», Acte du workshop «Adapting lexical and corpus resources to sublanguages and applications», dans *First International Conference on Resources and Evaluation*, Grenade.

Belleil (C.), 1997, *Reconnaissance, typage et traitement des coréférences des toponymes français et de leurs gentilés par dictionnaire électronique relationnel*, thèse de doctorat en informatique, Université de Nantes.

Bourigault (D.), Gonzalez-Mullier (I.) & Gros (C.), 1996, «LEXTER, a Natural Language Processing Tool for Terminology Extraction», dans *Proceedings Euralex'96*, Göteborg.

Cucchiarelli (A.), Luzi (D.) & Velardi (P.), 1998, «Using corpus evidence for automatic gazetteer extension», dans *First International Conference on Resources and Evaluation*, Grenade, p. 83-88.

Gaizauskas (R.), Wakao (T.), Humphreys (K.), Cunningham (H.) & Wilks (Y.), 1995, «University of Sheffield: description of the LaSIE system as used for MUC-6», dans *Proceedings of the sixth Message Understanding Conference*, Morgan Kaufmann Publishers, Los Altos, CA, p. 207-220.

Hirschman (L.), 1997, «Language Understanding Evaluations: A Case Study of MUC and ATIS», dans *Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology* (Sheffield).

Kosseim (L.) & Lapalme (G.), 1998, «EXIBUM: un système expérimental d'extraction bilingue», dans *Actes Rifra'98* (Sfax), p. 129-140.

Mani (I.), McMillian (R.), Luperfoy (S.), Lusher (E.) & Laskowski (S.), 1996, « Identifying unknown proper names in newswire text », dans Pustejovsky (J.) & Boguraev (B.), dir., *Corpus processing for lexical acquisition*, MIT Press, Cambridge, MA.

Maurel (D.), 1989, *Reconnaissance des séquences de mots par automate, adverbes de date du français*, Thèse de Doctorat en Informatique, Université Paris 7.

MUC-6, 1995, *Proceedings of the sixth Message Understanding Conference*, Morgan Kaufmann Publisher.

Pazienza (M.T.), dir., 1997, *Information extraction (a multidisciplinary approach to an emerging information technology), International Summer School SCIE'97 (Frascati 14-18 juil. 1997)*, Springer Verlag (Lecture Notes in Computer Science).

Poibeau (T.), 1998, « Extraction d'information : adaptation lexicale et calcul dynamique du sens », dans *Actes Rifra'98* (Sfax), p. 141-153.

Sénellart (J.), 1998, « Locating noun phrases with finite state transducers », dans *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, Montréal, p. 1212-1217.

Wacholder (N.), Ravin (Y.) & Choi (M.), 1996, « Disambiguation of proper names in text », dans *Proceedings of the fifth Applied Natural Language Conference*, Washington, DC.

Wilks (Y.), 1997, « Information Extraction as a core language technology », dans Pazienza (M.T.), dir., *Information Extraction*, Springer Verlag.