TALN 2010, Montréal, 19–23 juillet 2010

Reconnaissance d'entités nommées : enrichissement d'un système à base de connaissances à partir de techniques de fouille de textes

Damien Nouvel Arnaud Soulet Jean-Yves Antoine
Nathalie Friburger Denis Maurel
Université François Rabelais Tours, LI
Antenne Universitaire de Blois, 3 place Jean Jaurès, F-41000 Blois, France
{prénom.nom}@univ-tours.fr

Résumé. Dans cet article, nous présentons et analysons les résultats du système de reconnaissance d'entités nommées CasEN lors de sa participation à la campagne d'évaluation Ester2. Nous identifions quelles ont été les difficultés pour notre système, essentiellement : les mots hors-vocabulaire, la métonymie, les frontières des entités nommées. Puis nous proposons une approche pour améliorer les performances de systèmes à base de connaissances, en utilisant des techniques exhaustives de fouille de données séquentielles afin d'extraire des motifs qui représentent les structures linguistiques en jeu lors de la reconnaissance d'entités nommées. Enfin, nous décrivons l'expérimentation menée à cet effet, donnons les résultats obtenus à ce jour et en faisons une première analyse.

Abstract. In this paper, we present and analyze the results obtained by our named entity recognition system, CasEN, during the Ester2 evaluation campaign. We identify on what difficulties our system was the most challenged, which mainly are : out-of-vocabulary words, metonymy and detection of the boundaries of named entities. Next, we propose a direction which may help us for improving performances of our system, by using exhaustive hierarchical and sequential data mining algorithms. This approach aims at extracting patterns corresponding to useful linguistic constructs for recognizing named entities. Finaly, we describe our experiments, give the results we currently obtain and analyze those results.

Mots-clés : Reconnaissance d'Entités Nommées, Séquences Hiérarchiques, Motifs, Ester2.

Keywords: Named Entity Recognition, Hierarchical Sequences, Patterns, Ester2.

1 Introduction

La campagne d'évaluation Ester2, organisée par l'AFCP[1] et la DGA[2], a porté sur la transcription, la segmentation et l'extraction d'informations de flux de parole en langue française radiodiffusés (Galliano *et al.*, 2009). La tâche d'extraction d'information portait sur la reconnaissance d'Entités Nommées (EN) dans les transcriptions de ces flux (manuelles ou issues de systèmes de reconnaissance de parole). Les EN détectées devaient être catégorisées selon sept catégories : personnes (*pers*), lieux (*loc*), organisations (*org*), productions humaines (*prod*), montants (*amount*), mesures de temps (*time*) et fonctions (*fonc*). Cette typologie

[1] Association Francophone de la Communication Parlée
[2] Direction Générale de l'Armement

a été sous-divisée en 38 catégories fine (qui n'ont pas été évaluées). La mesure des performances était le « Slot Error Rate » (SER) (Makhoul *et al.*, 1999) ; la précision, le rappel et la f-mesure étaient aussi fournis. Sept systèmes (dont le nôtre, CasEN) ont participé à cette campagne, reposant sur des méthodes variées : apprentissage par CRF, systèmes à base de règles, avec analyses syntaxiques de surface ou profondes.

Dans cet article, nous analysons les résultats de la campagne et les performances de notre système. Nous présentons ensuite une technique de fouille de textes expérimentée dans le but de compléter de manière semi-automatique la base de connaissance du système.

2 Le système CasEN dans la campagne Ester2

CasEN est développé sur la plateforme CasSys (Friburger, 2002) d'analyse de textes par cascades de transducteurs. Reposant sur la plateforme Unitex[3], CasSys applique les transducteurs dans un ordre prédéfini, pour détecter des îlots de certitude (Abney, 1991), tout en réduisant progressivement l'espace de recherche. CasSys peut être utilisé pour toute tâche d'analyse de texte, CasEN en est une application à la reconnaissance d'EN, initialement dédiée aux textes écrits (Friburger & Maurel, 2004; Friburger, 2006). Elle utilise des transducteurs pour implémenter des motifs de surface qui reconnaissent des EN dans des textes. Le système a été adapté à la langue parlée pour l'annotation du corpus Eslo dans le cadre du projet ANR VariLing (Maurel *et al.*, 2009), puis pour notre participation à Ester2.

Le tableau 1 présente les résultats officiels de la campagne Ester2 (Galliano *et al.*, 2009) pour la tâche de reconnaissance des EN. Les systèmes centrés connaissances effectuant une analyse syntaxique profonde obtiennent les meilleurs résultats pour la transcription de référence. Pour la transcription automatique (ASR, fournie par le LIMSI), une approche à base d'apprentissage l'emporte.

	Référence			**ASR**
Participant (approche)	**SER**	**P**	**R**	**SER**
LIA (CRF)	23,9	86,4	71,8	43,4 (-19,5)
LIMSI (syntaxe surface)	30,9	81,1	70,9	45,3 (-14,4)
LINA (syntaxe surface)	37,1	80,7	55,4	54,0 (-16,9)
LI Tours (syntaxe surface)	*33,7*	*79,3*	*65,8*	*50,7 (-17,0)*
LSIS (CRF)	35	82,6	73	55,3 (-20,3)
Synapse (syntaxe profonde)	9,9	93	89,3	44,9 (-35,0)
Xerox (syntaxe profonde)	9,8	93,6	91,5	44,6 (-34,8)

Tableau 1 – *Reconnaissance d'EN lors d'Ester2 sur les transcription de référence (SER, Précision, Rappel) et automatiques (ASR)*

Les performances de CasEN (*LI Tours*) sont proches de celles des autres systèmes non industriels (à l'exception de celui du LIA), ce qui est rassurant pour un système initialement développé pour l'écrit. La dégradation des performances de CasEN sur les transcriptions automatiques est satisfaisante : en cela notre système s'avère être relativement tolérant aux erreurs de transcription, probablement grâce à la robustesse des analyses syntaxiques partielles et à la faible dépendance des motifs aux lexiques.

Afin d'analyser le comportement du système, nous avons caractérisé chaque erreur de CasEN (1180 erreurs pour 2512 EN) par sa localisation, le type d'erreur (suppression, insertion, catégorie erronée, erreur de frontière, etc.) et la règle de la convention Ester2 concernée. A cette occasion, nous avons remarqué un nombre non-négligeable d'incohérences dans l'annotation manuelle : de fausses erreurs d'insertions (EN omises dans la référence mais correctement détectées par CasEN), des annotations de référence qui ne respectent pas les règles spécifiées dans le guide d'annotation, etc. Au final, nous observons une réduction

[3]http ://www-igm.univ-mlv.fr/˜unitex/

de presque 10% du SER, après avoir corrigé manuellement la référence.

Par ailleurs, avec 7 types d'EN, Ester2 introduit une catégorisation plus fine que celles mise en place lors des campagnes antérieures, rendant la catégorisation plus subtile (ce qui explique en partie les difficultés qu'ont rencontrées les annotateurs humains). Certaines difficultés sont intéressantes, car elles relèvent du phénomène de métonymie (Markert & Hahn, 2002). Dans d'autres cas, on peut au contraire estimer que des sous-catégorisations artificielles ont eu une influence directe sur les résultats (Nouvel *et al.*, 2010).

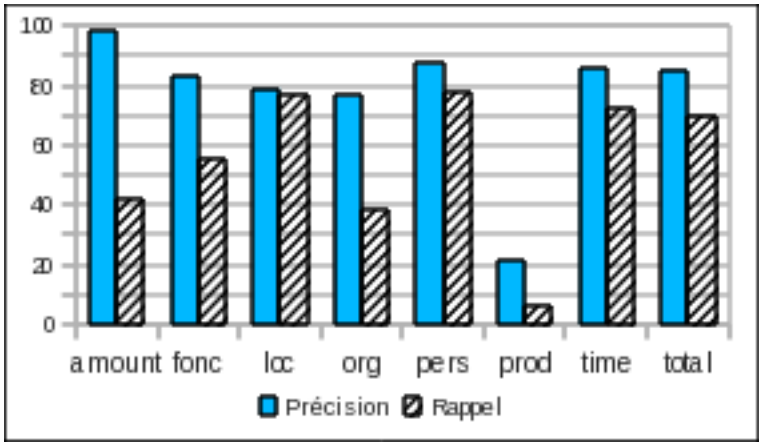

Figure 1 – *Résultats de CasEN par type d'entités nommées*

La figure 1 présente les performances de CasEN par catégories. Globalement satisfaisante, la précision est médiocre pour la catégorie *prod* (hétérogène, elle a aussi gêné les autres participants). Le rappel varie significativement d'une catégorie à une autre. Nos difficultés sont en partie dues au relatif manque de couverture de notre lexique de noms propres (notamment pour ceux d'origine africaine dans les enregistrements de la radio *Africa1*). Face à ce problème (Charton & Torres-Moreno, 2009) propose l'extraction d'EN d'encyclopédies. Il n'en reste pas moins que dans ce type d'application, le problème des mots hors-vocabulaire sera toujours présent, notamment du fait de la dépendance des EN au contexte du document à analyser (Favre *et al.*, 2005).

Nous avons estimé la part de chaque transducteur (qui implémente une famille de motifs) dans le taux d'erreurs du système (Nouvel *et al.*, 2010). Huit transducteurs génèrent 54% des erreurs, mais ils reconnaissent 60% des EN : nous ne remarquons pas de transducteurs défaillants. Les transducteurs qui reconnaissent des lieux font des erreurs liées à des usages métonymiques (par ex. « Egypte » pour une équipe sportive, « Paris » pour le gouvernement français). Ceux qui reconnaissent des personnes ont des difficultés avec les EN imbriquées : certaines fonctions s'étendent sur plusieurs mots et CasEN ne parvient pas à en détecter la frontière droite. Ceci concerne également les expressions de temps ou les organisations (par ex. « [chambre régionale des comptes d'Ile-de-France] [...] »). Pour surmonter ce problème, nous envisageons de parenthéser l'énoncé à l'aide d'un « chunker » (Antoine *et al.*, 2008), afin de fournir une segmentation utile et consistante.

3 Application de la fouille de texte à la reconnaissance d'EN

Une cause majeure d'erreurs du système CasEN est donc l'insuffisance de la couverture de son dictionnaire lors de la recherche des EN. Ce constat nous a conduit à développer des techniques semi-automatiques d'extension des bases de connaissances à l'aide de techniques de fouille de textes. Notre objectif est d'extraire des motifs qui, à partir des lemmes ou des catégories morphosyntaxiques, caractérisent des « structures linguistiques » pertinentes pour la reconnaissance d'EN. Par exemple, le motif « *DET président ADJ NP NP* » est approprié pour décrire une EN telle que « *le président américain Barack Obama* ». Les éléments constituant les séquences, nommés items, pourront donc être issus de niveaux hiérarchiques différents. L'objectif étant, à terme, d'enrichir la base de transducteurs de CasEN à l'aide de ces motifs.

3.1 Méthode d'extraction des règles séquentielles hiérarchiques

Le processus d'extraction de règles se décompose en trois phases : transformation du corpus en séquences hiérarchiques ; extraction et groupement de séquences fréquentes ; sélection des meilleures règles.

Le corpus est segmenté en phrases, grâce aux tours de parole et aux étiquettes morphosyntaxiques : chacune sera considérée comme une séquence. Les annotations d'EN du corpus de référence sont délimitées, pour un *type* donné, par des balises "<*type*BEG>" et "<*type*END>". Nous utilisons l'étiqueteur TreeTagger (Schmid, 1994) pour transformer le corpus de référence, en associant à chaque terme sa catégorie (POS_1), sa sous-catégorie (POS_2) morphosyntaxiques, et son lemme. Ainsi, le corpus est représenté comme suites d'items, chaque item appartenant à une taxonomie, nous les notons POS_1/POS_2/LEMME. Enfin, nous omettons les lemmes pour ce qui est reconnu par TreeTagger comme un nom propre (POS_1 = "NAM"), afin d'éviter (en première approche) de construire des motifs reposant sur des éléments lexicaux.

Considérons l'exemple suivant : « `Le nouveau <PERSBEG> président Barack Obama <PERSEND> est arrivé à <LOCBEG> Moscou <LOCEND>. Il y a vu le nouveau <PERSBEG> président Dimitri Medvedev <PERSEND>[...]` »
Il est transformé en séquences d'items, dont la première est : « `DET/ART/le ADJ/nouveau <PERSBEG> NOM/président NAM NAM <PERSEND> VER/pres/être VER/pper/arriver PRP/à <LOCBEG> NAM <LOCEND>` ».

La seconde étape extrait, à travers la taxonomie, toutes les séquences d'items fréquentes (selon un seuil donné) (Mannila *et al.*, 1997), qui contiennent au moins une balise d'EN. La recherche étant exhaustive et tenant compte d'une hiérarchie, l'extraction génère une grande richesse de descriptions, mais aussi de nombreux motifs redondants. Nous groupons les motifs de même longueur et de même fréquence, lorsque l'un est généralisation de l'autre. Par exemple, le motif « ADJ/nouveau <PERSBEG> NOM » est plus général que « ADJ/nouveau <PERSBEG> NOM/président » : s'ils ont mêmes fréquences, ils couvrent alors les mêmes occurrences et un seul des deux suffirait, à priori, à reconnaître la forme d'EN qu'ils couvrent. Nous avons évalué les motifs obtenus avec ou sans ce groupement et n'avons pas observé de différence significative des performances. Voici des groupes que l'on obtiendrait sur l'exemple précédent :

Groupe A :
```
DET/ART ADJ <PERSBEG> NOM/président NAM NAM <PERSEND>
DET ADJ/nouveau <PERSBEG> NOM NAM NAM <PERSEND>
DET/ART/le ADJ/nouveau <PERSBEG> NOM/président NAM NAM <PERSEND>
```
Groupe B :
```
PRP/à <LOCBEG> NAM <LOCEND>
PRP <LOCBEG> NAM <LOCEND>
```

La dernière étape est la sélection des motifs selon leur confiance : la proportion de séquences, parmi celles où le motif a été observé, qui comportent effectivement les annotations d'EN à détecter. Par groupe, nous sélectionnons le motif qui a la meilleure confiance, si celle-ci dépasse un seuil donné. Ainsi, nous obtenons les motifs qui, à priori, conduiraient, selon une bonne probabilité (la confiance), à la découverte d'une annotation d'EN. Ceci peut-être mis en parallèle, en fouille de données, aux règles d'association. Pour notre exemple, nous conservons les motifs suivants :

Règle sélectionnée pour le groupe A : `DET ADJ NOM/président NAM NAM =>DET ADJ <PERSBEG> NOM/président NAM NAM <PERSEND>`
Règle sélectionnée pour le groupe B : `PRP/à NAM => PRP/à <LOCBEG> NAM <LOCEND>`

3.2 Utilisation des règles pour la reconnaissance des EN

Les règles que nous obtenons par cette méthode nous permettent de déterminer dans quels contextes apparaissent une ou plusieurs annotations d'EN. Cependant, l'objectif est d'obtenir des paires d'annotations (ouvrantes et fermantes), correctement ordonnées et sans chevauchements, qui délimitent des EN. A cet effet, nous mettons en œuvre, pour le moment, une stratégie simple, qui consiste à ne conserver que les paires d'annotations à l'empan le plus large possible dans une fenêtre considérée.

3.3 Résultats et améliorations envisagées

La mise en application donne de nombreux motifs, qui sont souvent des mélanges de lemmes et de catégories morphosyntaxiques : ceci nous conforte dans l'idée qu'il est utile de s'appuyer sur une taxonomie pour reconnaître une EN. La figure 2 présente la dispersion précision /rappel des motifs, extraits puis évalués sur Ester2 en validation croisée (12 groupes). Nous obtenons des motifs d'une bonne précision, mais le rappel demeure assez faible. Comme attendu, baisser le seuil de fréquence ou de confiance amène à une sélection plus large de motifs, faisant baisser la précision et augmenter le rappel. La courbe à confiance 0,6 nous paraît être la plus intéressante en termes de compromis précision / rappel pour l'extraction des motifs. Mais ce paramètre est spécifique au corpus utilisé, à la méthode d'extraction et d'utilisation des règles de reconnaissance d'EN. Nous expérimentons actuellement une extraction sur un corpus élargi (Ester2 + Eslo, 250 Kmots).

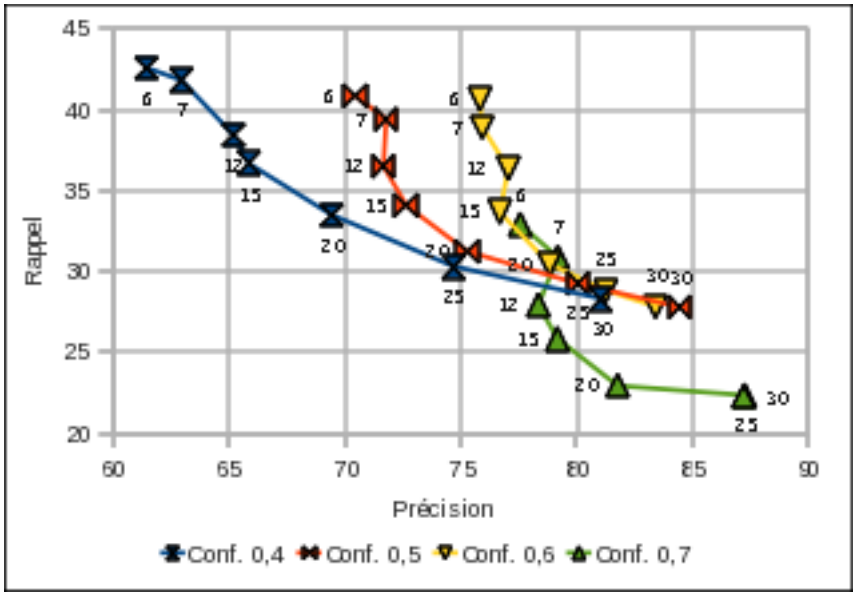

Figure 2 – *Dispersion précision / rappel pour l'évaluation des motifs à diverses confiances et selon le seuil fréquence*

Nous envisageons des améliorations sur ce procédé d'extraction de motifs. D'une part, afin d'éviter la redondance entre motifs extraits, nous cherchons à mieux les grouper, indépendamment de leur taille, à fréquence « proche ». Par ailleurs, nous remarquons que certains motifs gagneraient à contenir des éléments optionnels ou disjonctifs (« *président* », « *ministre* », « *gouverneur* », etc.), ce que nous comptons autoriser dans les motifs. L'objectif étant de constituer des groupes cohérents de motifs, qui correspondent à des structures linguistiques à confronter aux transducteurs implémentés dans CasEN.

4 Conclusions et perspectives

Nous avons présenté les résultats du système CasEN dans la campagne d'évaluation Ester2, que nous avons analysés pour déterminer les faiblesses réelles de notre système. Celles-ci se répartissent essentiellement en trois catégories : résolution de la métonymie, délimitation des frontières, couverture du système.

La direction que nous avons prise, plus particulièrement pour le manque de couverture, a été d'adopter une approche semi-automatique, systématique et exhaustive, afin de trouver les structures linguistiques correspondant à des EN. Par des techniques de fouille de données, nous pensons apporter de nouveaux éléments, jusque là non implémentés par CasEN, qui nous donneront des pistes pour améliorer le système.

Nous cherchons actuellement à estimer les gains à espérer en intégrant les motifs trouvés par fouille de textes dans CasEN, afin d'évaluer les perspectives que nous ouvrent cette approche. Par ailleurs, nous menons des expériences pour déterminer la dépendance de cette extraction au corpus, autant du point de vue du domaine et des thématiques que des dates d'enregistrement du corpus.

5 Remerciements

Ce travail a été réalisé dans le cadre des projets Variling (ANR-06-CORP-023), Epac (ANR-00-MDCA-006-03), créés par l'Agence Nationale de la Recherche (ANR) et FEDER *Région Centre*.

Références

ABNEY S. P. (1991). *Parsing by Chunks*, In *Principle-Based Parsing*, p. 257–278. Kluwer.

ANTOINE J.-Y., MOKRANE A. & FRIBURGER N. (2008). Automatic rich annotation of large corpus of conversational transcribed speech : the chunking task of the epac project. In *LREC'08*.

CHARTON E. & TORRES-MORENO J. M. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. In *TALN'2009*.

FAVRE B., BÉCHET F. & NOCERA P. (2005). Robust named entity extraction from large spoken archives. In *HLT/EMNLP'05*.

FRIBURGER N. (2002). *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques*. PhD thesis, Université François-Rabelais Tours, France.

FRIBURGER N. (2006). Linguistique et reconnaissance automatique des noms propres. *Meta : Translators' Journal*, **51-4**, 637–650.

FRIBURGER N. & MAUREL D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Sciences*, **313**, 93–104.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech'09*, p. 2583–2586.

MAKHOUL J., KUBALA F., SCHWARTZ R. & WEISCHEDEL R. (1999). Performance measures for information extraction. In *DARPA Broadcast News Workshop*, p. 249–252.

MANNILA H., TOIVONEN H. & VERKAMO A. I. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, **1**, 259–289.

MARKERT K. & HAHN U. (2002). Understanding metonymies in discourse. *AI*, **135**, 145–198.

MAUREL D., FRIBURGER N. & ESHKOL I. (2009). Who are you, you who speak ? In *LTC'09*.

NOUVEL D., ANTOINE J.-Y., FRIBURGER N. & MAUREL D. (2010). An analysis of the performances of the casen named entities recognition system in the ester2 evaluation campaign. In *LREC'10*.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *NEMLP'94*, p. 44–49.