Fouille de règles d'annotation pour la reconnaissance d'entités nommées

Damien Nouvel — Jean-Yves Antoine — Nathalie Friburger — Arnaud Soulet

Université François Rabelais Tours - Laboratoire d'informatique
damien.nouvel@limsi.fr,
{jean-yves.antoine, nathalie.friburger, arnaud.soulet}@univ-tours.fr

RÉSUMÉ. *Comme pour de nombreuses autres problématiques TAL, la reconnaissance d'entités nommées met en jeu aussi bien des systèmes à base de connaissances que des systèmes guidés par les données. Dans cet article, nous proposons une approche médiane par l'adaptation de méthodes issues de l'extraction de connaissances. Notre système, mXS, intègre des techniques de fouille séquentielle hiérarchique pour la détection des entités nommées. Le système adopte une démarche centrée sur les données pour extraire des motifs symboliques. Il repose par ailleurs sur une stratégie originale qui consiste à rechercher séparément le début et la fin des entités. Cette approche présente l'intérêt de conserver une certaine robustesse par rapport aux bruit et disfluences. Elle est adaptée au cadre applicatif visé par le système : la détection d'entités nommées au sein de flux de parole conversationnelle transcrite automatiquement. À ce titre, mXS a participé à la campagne d'évaluation ETAPE où il a présenté de bons résultats. Cet article présente le fonctionnement de mXS et ses performances sur les jeux de données issus de deux campagnes d'évaluation francophones (ESTER 2 et ETAPE).*

ABSTRACT. *Like many NLP tasks, the question of Named Entity Recognition can be adressed either using a symbolic or a data-centered approach. In this paper, we present a hybrid approach which consists in the adaptation of data mining techniques. Our system, mXS, relies on a sequential hierarchical text mining techniques. It implements a data-centered approach to extract symbolic patterns. Besides, mXS relies on an original strategy of recognition which consists in detecting separately the begining and the ending of entities. This strategy is robust on noisy data, especially when speech disfluences or recognition errors occur. Our system has participated to the ETAPE French-speaking evaluation campaign over conversational speech. This paper describes mXS and reports results obtained on reference data (ESTER 2 and ETAPE).*

MOTS-CLÉS : *reconnaissance d'entités nommées, fouille de données, règles d'annotation.*

KEYWORDS: *named entity recognition, data mining, annotation rules.*

1. Introduction

Le développement des technologies de l'information et de la communication a modifié en profondeur la manière dont nous manipulons les connaissances. Outre l'accroissement indéniable des volumes de données mis à disposition des professionnels comme du grand public, ceux-ci correspondent à des modalités et des types de contenus de plus en plus variés : texte, image, audio, vidéo, SMS, *tweet*, etc. Cette adoption massive des technologies numériques nécessite de résoudre de nombreuses problématiques. Parmi celles-ci, notre travail se focalise sur la reconnaissance d'éléments plus particulièrement sollicités au sein de données langagières : les entités nommées (EN).

En effet, les travaux menés en traitement automatique des langues (TAL) ont porté une attention particulière aux noms propres de personnes, de lieux et d'organisations, ces éléments semblant être utiles à diverses tâches comme, par exemple, la recherche d'information. La communauté scientifique s'est penchée sur cette question dès les années 90 pour ce qui est des textes écrits électroniques. Plus récemment, les progrès de la reconnaissance automatique de la parole ont ouvert ces problématiques aux transcriptions orales, en particulier de contenus radio ou télédiffusés. Par ailleurs, au gré des besoins, ces travaux ont été étendus aux dates, aux expressions numériques, aux marques ou aux fonctions, pour recouvrir progressivement un spectre large mais également assez flou d'expressions linguistiques (Nadeau et Sekine, 2007). Ainsi, les visées référentielles ou applicatives ont dirigé les travaux sur le sujet (Ehrmann, 2008 ; Grouin *et al.*, 2011). À l'heure actuelle, il ne semble pas exister de définition qui fasse consensus pour caractériser en intention ces éléments du langage (Nouvel, 2012). Ainsi, la complexité de la tâche s'est accrue, à la fois par l'élargissement de la gamme des éléments d'intérêt et par les difficultés à traiter certaines modalités, en particulier orales.

Comme souvent en TAL, la reconnaissance d'entités nommées (REN) a donné lieu à des systèmes qui reposent sur des approches orientées connaissances ou sur des approches guidées par les données. Les premières, souvent à base de transducteurs (Favre *et al.*, 2005 ; Brun et Ehrmann, 2010 ; Maurel *et al.*, 2011) ont généralement pour elles une grande précision mais nécessitent un coût de développement important, qui se traduit le plus souvent par une couverture et donc un rappel perfectibles. À l'opposé, les secondes, par utilisation d'apprentissage automatique comme les CRF (Zidouni *et al.*, 2009 ; Raymond et Fayolle, 2010 ; Raymond, 2013 ; Savary *et al.*, 2010), permettent d'obtenir de bonnes performances avec un coût d'entrée plus limité, à partir du moment où l'on dispose déjà d'un corpus d'apprentissage. L'aspect « boîte noire » des algorithmes d'apprentissage peut cependant gêner la recherche d'une amélioration de leurs performances. En revanche, ils sont réputés être plus couvrants (moins de silence) et présenter une dégradation plus graduelle de leurs performances sur des données bruitées.

Ces constats ont été vérifiés par de nombreuses campagnes d'évaluation, telle la campagne d'évaluation francophone ESTER 2 (Galliano *et al.*, 2009) portant sur le traitement de transcriptions de la parole pour des émissions de radio. Effectivement,

il a été montré que les meilleurs systèmes travaillant sur des transcriptions exactes étaient à base de connaissances, alors que les tests sur des sorties de reconnaissance de la parole ont été dominés par un système orienté données. Les travaux que nous présentons dans cet article ont été menés sur les données de la campagne ESTER 2, puis, ultérieurement, dans le cadre de la campagne ETAPE qui lui a succédé (tâches similaires, mais avec d'autres données, notamment des débats télévisés).

Par ailleurs, la généralisation récente de techniques de fouille de données à de nombreuses problématiques a démontré la possibilité d'extraire des connaissances à partir des données. En particulier, la fouille du Web et l'exploitation d'encyclopédies en ligne afin d'enrichir des lexiques (Bunescu et Pasca, 2006 ; Etzioni *et al.*, 2005 ; Béchet et Roche, 2010) ont connu un grand succès. La fouille de données tournée vers l'extraction automatique de motifs utiles au TAL pour des tâches dédiées fait l'objet de nombreux travaux récents (Sun et Grishman, 2010 ; Ezzat, 2010 ; Charton *et al.*, 2011) et semble porteuse de nouvelles perspectives pour la mise au point de systèmes qui soient à la fois performants et dont la mécanique soit plus directement accessible.

Dans ce contexte, l'approche pour la REN présentée dans cet article est novatrice, puisqu'elle consiste à employer et adapter des méthodes issues du domaine de l'extraction de connaissances. Les travaux les plus similaires (Freitag et Kushmerick, 2000 ; Plantevit *et al.*, 2009 ; Snow *et al.*, 2004), s'ils extraient des motifs pour la REN, se penchent peu sur la manière d'utiliser ces motifs. Plus précisément, nous avons développé un système de REN, `mXS`, qui intègre des techniques de fouille séquentielle hiérarchique de données adaptées à l'annotation des EN. Ces travaux présentent plusieurs originalités du point de vue du TAL :

– ils constituent un moyen terme entre les approches orientées données et orientées connaissances puisqu'ils reposent sur la recherche, à partir de données d'apprentissage, de motifs pour la reconnaissance des EN. Nous sommes donc en présence d'une technique centrée sur les données permettant l'extraction de connaissances symboliques et parfaitement interprétables ;

– ils reposent sur une stratégie originale de détection des EN, qui consiste à rechercher séparément le début et la fin des entités, en s'appuyant sur l'environnement immédiat des marqueurs d'annotation correspondants. Cette approche présente l'intérêt de pouvoir conserver une certaine robustesse en cas de disfluence au sein de l'entité nommée, ou d'erreur de reconnaissance ayant un impact sur une partie de l'entité ;

– enfin, nous présentons dans cet article des expériences d'hybridation entre notre système d'extraction de connaissances (`mXS`) et un système à base de transducteurs (`CasEN`) développé également par notre équipe.

L'approche que nous proposons a participé à la campagne ETAPE. Les résultats que nous avons obtenus au cours de cette campagne soutiennent la comparaison avec les autres systèmes participants. C'est en particulier le cas lorsque le système est hybridé avec un système à base de connaissances.

En première partie, nous revenons sur la description des ressources qu'utilisent nos recherches, à savoir les corpus ESTER 2 et ETAPE. Nous présentons en deuxième partie l'approche que nous adoptons, en nous focalisant sur les originalités de la méthode d'extraction de motifs séquentiels hiérarchiques pour l'annotation. En troisième partie, nous détaillons l'application de cette approche pour la REN en décrivant quelques spécificités pour l'implémentation du système, ses performances dans diverses configurations et sa comparaison avec d'autres approches. Les résultats montrent que nous obtenons des performances au niveau de l'état de l'art de la REN sur le français parlé spontané.

2. Données d'expérimentation

Les travaux qui sont présentés dans cet article ont été réalisés dans le contexte des campagnes d'évaluation ESTER 2 et ETAPE auxquelles a participé le laboratoire LI[1], avec le système `CasEN` (pour ESTER 2) puis avec les deux systèmes `CasEN` et `mXS` (pour ETAPE). Cette section décrit les jeux de données correspondant à ces campagnes.

2.1. *Corpus ESTER 2*

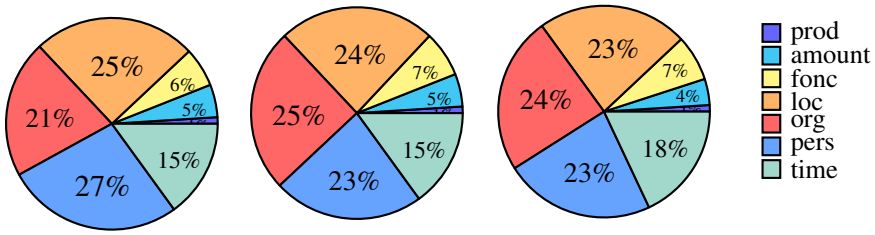
Les campagnes d'évaluation ESTER, organisées par l'AFCP[2] et la DGA[3], ont porté sur la transcription, la segmentation et l'extraction d'informations de flux radiodiffusés de la parole en langue française (Galliano *et al.*, 2009). La tâche d'extraction d'information portait sur la reconnaissance d'EN dans les transcriptions de ces flux (manuelles ou issues de systèmes de reconnaissance de la parole). Les EN détectées devaient être catégorisées selon sept catégories : personnes (pers), lieux (loc), organisations (org), productions humaines (prod), montants (amount), dates et heures (time) et fonctions (fonc). Cette typologie a été sous-divisée en trente-huit catégories fines mais l'évaluation n'a pas été menée à ce niveau de détail et les résultats que nous présentons dans cet article se limitent donc au niveau supérieur de la typologie. Le tableau 1 donne les caractéristiques des principales sous-parties du corpus, et la figure 1 donne la répartition de ces types au sein de ces sous-parties. Nous remarquons que le corpus est assez équilibré, avec quatre types dominants : pers, org, loc et time. Le nombre d'EN rapporté au nombre de tokens du corpus est de 6,3 %.

Réalisé avec des moyens limités, le corpus de référence ESTER 2 n'était pas exempt d'erreurs d'annotation. En particulier, certaines annotations ne respectaient pas le guide d'annotation, ce qui a posé un problème concernant les données de référence lors de la campagne entre les systèmes centrés sur les données, qui privilégiaient l'annotation effective du corpus, et les systèmes à base de connaissances reposant sur

1. Laboratoire d'informatique de l'université de Tours.
2. Association francophone de la communication parlée.
3. Direction générale de l'armement.

Corpus	Sources (nombre de fichiers)	Tokens	Énoncés	EN
ESTER2-Train	France Inter (47), RFI (29), RTM (103), France Info (13), France Culture (1), France Classique (1)	1 269 327	44 211	80 227
ESTER2-Dev	Africa1 (9), France Inter (5), RFI (2), TVME (4)	73 386	2 491	5 326
ESTER2-Test	Africa1 (9), France Inter (6), RFI (7), TVME (4)	87 165	2 983	5 875
Total	240 enregistrements	1 429 878	49 685	91 428

Tableau 1. *Caractéristiques pour chaque partie d'ESTER 2*



(a) ESTER2-Train (b) ESTER2-Dev (c) ESTER2-Test

Figure 1. *Répartition des types d'EN pour chaque partie d'ESTER 2*

le guide. Lors de la campagne, il a été finalement décidé d'appuyer l'évaluation sur la référence effective. Toutefois, pour atteindre une évaluation plus juste, nous avons repris après coup une partie du corpus pour corriger les infractions évidentes au guide d'annotation (Nouvel *et al.*, 2010) : nous avons ainsi créé une sous-partie du corpus de test corrigé (39 707 tokens, 2 792 EN) sur lequel porteront nos évaluations.

2.2. *Corpus ETAPE*

Soutenue par l'ANR, la campagne d'évaluation ETAPE a fait suite à ESTER 2 et a été réalisée en interaction avec le programme de recherche Quaero, en particulier pour tout ce qui concerne l'annotation des corpus (Rosset *et al.*, 2011). Cette campagne a porté sur le traitement d'émissions radiodiffusées et télévisuelles, donc orales et en partie spontanées. En particulier, les émissions de débats représentent des difficultés supplémentaires pour le TAL : les conversations ne sont pas toujours calmes et, régulièrement, les interlocuteurs se coupent la parole, parlent en même temps, hésitent, etc., ce qui pose des problèmes autant pour la reconnaissance de la parole que pour les traitements ultérieurs. Le corpus ETAPE intègre ESTER 2 (dont l'annotation a été mise en conformité avec le nouveau guide d'annotation) ainsi que d'autres sources de données (corpus EPAC).

Comme pour ESTER 2, les corpus annotés présentent des disfluences orales susceptibles de gêner l'analyse. De même, la campagne a porté sur la transcription manuelle des dialogues oraux mais également sur des sorties de systèmes automatiques de reconnaissance vocale.

Corpus	Sources (nombre de fichiers)	Tokens	Énoncés	EN
ETAPE-Train	BFMTV (5), France Inter (16), LCP (23)	355 975	14 989	46 259
ETAPE-Dev	BFMTV (1), France Inter (6), LCP (6), TV8 (2)	115 530	5 724	14 112
ETAPE-Test	BFMTV (1), France Inter (6), LCP (5), TV8 (2)	123 221	6 770	13 055
Total	73 enregistrements	594 726	27 483	73 426
ETAPE-Quaero	France Classique (1), France Culture (1), France Inter (62), France Info (13), RFI (14), RTM (97)	1 596 427	43 828	279 797

Tableau 2. *Caractéristiques pour chaque partie d'ETAPE*

Le corpus ETAPE est divisé en plusieurs parties, décrites dans le tableau 2. Au moment où les travaux ici décrits ont été amorcés, le corpus ETAPE-Test n'était pas disponible (campagne en cours) : les expériences pour mettre au point le système ont été réalisées avec ETAPE-Train pour fouiller les données (extraction des règles et paramétrage du modèle) et ETAPE-Dev pour les évaluations. Depuis, des évaluations préliminaires ont été réalisées et nous sommes en mesure de présenter les performances de notre approche comparativement à d'autres systèmes[4].

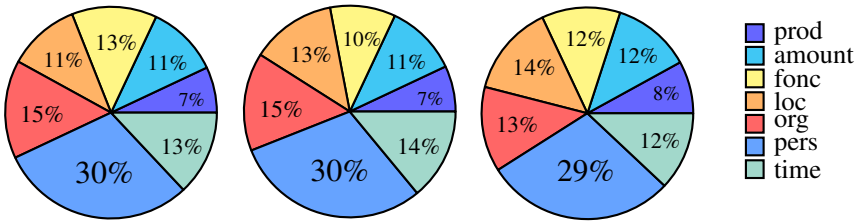
L'annotation ETAPE est plus fine que celle utilisée lors de la compagne ESTER 2. À la base, ETAPE reprend les types d'EN définis dans ESTER 2. Les types primaires[5] d'EN rencontrés dans le corpus ETAPE sont les personnes (pers), les fonctions (fonc), les organisations (org), les lieux (loc), les productions humaines (prod), les dates et les heures (time), les montants (amount) et les événements[6] (event). Comme nous le voyons en figure 2, nous retrouvons une répartition des classes assez équivalentes à celle du corpus ESTER 2, avec toutefois des proportions plus importantes de montants (amount) et de personnes (pers). Ceci est en partie lié à l'extension des EN aux expressions formées autour de noms communs (par exemple, « internautes » comme pers.coll). Les types principaux sont décomposés en trente-quatre sous-types, sur lesquels a porté l'évaluation.

Dans la continuité des travaux menés dans le cadre du projet Quaero, ETAPE envisage une analyse plus fine qu'ESTER 2 de la structure des EN. Ainsi, en plus des EN à

4. Merci à Olivier Galibert (LNE), Matthieu Carré (ELDA) et Guillaume Gravier (IRISA) pour avoir mis ces évaluations à notre disposition.

5. Premier niveau de la typologie.

6. Très marginale dans les annotations, nous écartons cette catégorie de cet article.



(a) ETAPE-Train (b) ETAPE-Dev (c) ETAPE-Test

Figure 2. *Répartition des types d'EN pour chaque partie d'ETAPE*

proprement parler, sont distingués les « composants » de ces EN. Ces derniers peuvent être génériques (nom, valeur, objet) ou spécifiques à certains types (jour, mois, etc. pour une date). Par exemple, le type « qualificateur » permet de décrire certaines finesses de réalisation ou de référenciation liées à une entité. Ces éléments permettent de mieux représenter les entités lors de leur annotation (Rosset *et al.*, 2011) et sont susceptibles de caractériser des informations importantes pour la recherche d'information. Le nombre d'EN rapporté au nombre de tokens du corpus est de 12,3 %, dont 4,8 % pour les entités et 7,5 % pour les composants, ce qui montre l'importance de ces derniers pour ce protocole d'annotation.

L'annotation ETAPE étant bien plus fine que celle considérée dans la campagne ESTER 2, on conçoit aisément que la tâche était plus difficile pour les systèmes participants. La comparaison des performances observées sur ESTER 2 et ETAPE est donc délicate.

3. Explorer les données pour extraire des règles d'annotation

Pour la fouille de données, les corpus dont nous disposons vont nous permettre de rechercher des motifs pour la REN. Cette partie présente l'approche que nous avons élaborée à cet effet. Nous décrivons en premier lieu les traitements appliqués aux données afin de les mettre à disposition de la fouille de données. Puis nous présentons notre proposition originale d'exploration des données comme motifs séquentiels hiérarchiques. Enfin, nous indiquons comment ces motifs peuvent être paramétrés au sein d'un modèle à régression logistique pour annoter automatiquement des textes.

3.1. *Mise à contribution des ressources pour enrichir les données*

Un des problèmes principaux auxquels doivent faire face les systèmes symboliques de REN est le coût humain de développement d'une base de connaissances (par exemple, les transducteurs et les lexiques) pour cette tâche. Ce constat nous conduit à étudier l'utilisation de techniques de fouille de données pour extraire automatique-

ment de tels motifs. Dans le cas présent, notre objectif est de tirer parti d'analyses préliminaires, dans le sens où les motifs peuvent utiliser l'item lexical, son lemme ou sa catégorie morphosyntaxique (partie du discours). Par exemple, à partir de la séquence *'le Centre Pompidou'* et d'autres équivalentes, nous souhaiterions extraire un motif du type `'DET Centre NP'`, où `'DET'` et `'NP'` sont respectivement les parties du discours des déterminants et des noms propres. L'utilisation d'une hiérarchie sur les tokens nous permet d'extraire des motifs généralisés, tout en limitant la combinatoire inhérente à ce type de recherche dans les données. Ces enrichissements peuvent être ambigus : à un token donné peuvent être associées plusieurs catégories.

Afin d'illustrer l'intuition de notre approche, considérons l'exemple suivant : *'Pierre a visité le Centre Pompidou'*, dans lequel le mot « Pompidou » est ambigu lorsque l'on ne tient pas compte de son contexte, et peut correspondre à un bâtiment (*'Centre Pompidou'* catégorisé `'BAT'`) ou à une célébrité (*'Pompidou'* catégorisé `'CELEB'`).

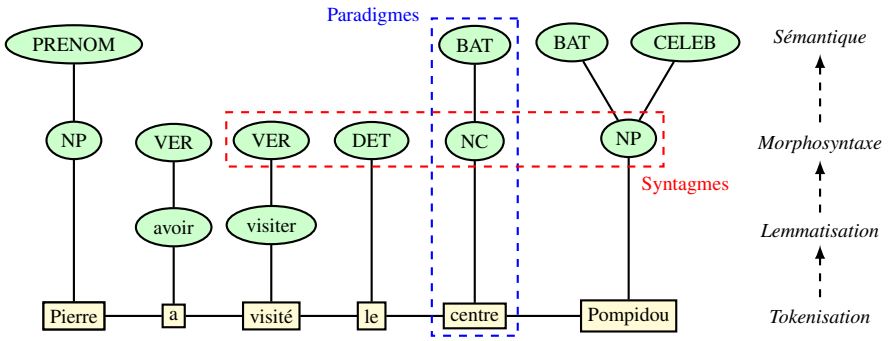


Figure 3. *Représentation des textes pour la fouille*

La figure 3 illustre la représentation obtenue. Notre objectif sera alors d'explorer la combinatoire de telles représentations, simultanément sur l'axe syntagmatique (concaténation d'items) et paradigmatique (catégories attribuées à un ou plusieurs items), afin de déterminer des motifs séquentiels d'intérêt pour la REN s'appuyant sur divers niveaux de la hiérarchie. Pour notre exemple, nous obtiendrions alors la séquence enrichie suivante (où $\oplus$ marque l'ambiguïté par disjonction exclusive) :

`'PRENOM/NP/Pierre VER/avoir/a VER/visiter/visité DET/le BAT/NC/Centre BAT/NP/Pompidou`$\oplus$`CELEB/NP/Pompidou'`

Cette hiérarchie sur les tokens nous permet ainsi de réaliser la fouille de données sur un langage qui est à la fois formellement défini et flexible selon la tâche visée. Il n'y a pas de contraintes *a priori* sur la profondeur ou la largeur des hiérarchies, ce qui nous permettra de moduler à volonté l'axe paradigmatique selon les éléments observés et la tâche d'annotation à réaliser. L'axe syntagmatique sera construit par concaténation d'éléments. Précisons que la hiérarchie discutée ici concerne un token

(ou un segment comme nous le verrons plus loin) et ne doit pas être confondue avec les relations de généralisation entre motifs, même si elle l'engendre.

Nous allons maintenant décrire les modules de prétraitement qui sont mis en œuvre pour réaliser cet enrichissement hiérarchique des données.

3.1.1. *Morphosyntaxe*

Nous utilisons `TreeTagger` (Schmid, 1994) afin de réaliser conjointement la tokenisation, la lemmatisation et l'étiquetage en parties du discours du texte. Ainsi, un token pourra être graduellement généralisé à son lemme, à son étiquette morphosyntaxique ou à sa partie du discours. Les ambiguïtés sont prises en compte : par exemple, le verbe ambigu *'suis'* sera représenté comme l'item `'VER/suivre/suis⊕VER/être/suis'` et pourra donc être généralisé comme `'VER/suivre'`, `'VER/être'` ou `'VER'`. De plus, quelques adaptations supplémentaires sont apparues pertinentes :

– *déterminants* : les déterminants définis (*'le'*, *'la'*, *'les'*, *'l"*) sont sous-catégorisés en `'DET/DEF'` (pour faciliter la détection des descriptions définies) ;

– *prépositions* : la sous-catégorie de TreeTagger `'PRP:det'` (*'au'*, *'du'*, *'des'*) forme une catégorie `'PRPDET'` ;

– *nombres* : pour mieux distinguer les dates et les montants, les nombres sont sous-catégorisés selon leur nombre de chiffres[7] ;

– *noms propres et abréviations* : ces deux catégories se généralisent en `'NAMABR'` ;

– *verbes* : les sous-catégories relatives au mode et temps du verbe sont supprimées.

Ainsi, l'énoncé *'Je suis en France'* donnera, avant utilisation des lexiques, la séquence suivante :

```
'PRO/PER/je/Je VER/suivre/suis⊕VER/être/suis PRP/en/en
NAMABR/NAM/France/France'
```

Pour la recherche des motifs, nous souhaitons faire abstraction des variations flexionnelles et omettons le token lui-même, pour ne conserver que le lemme comme élément le plus spécifique d'un motif. Ainsi, dans les motifs, l'élément `'VER/suivre/suivrions'` ne pourra apparaître, seules ses généralisations `'VER/suivre'` et `'VER'` pourront être des éléments des motifs. Pour l'exemple précédent, nous obtenons maintenant la séquence :

```
'PRO/PER/je VER/suivre⊕VER/être PRP/en NAMABR/NAM/France'
```

Nous avons ainsi adapté et modulé la description des tokens selon l'intérêt qu'ils présentent pour la REN. Si `mXS` devait être adapté à d'autres tâches, il suffirait d'approfondir la description de certaines parties de la hérarchie selon leur pertinence *a*

7. Ce nombre de chiffres est précisé s'il est inférieur ou égal à quatre, et le préfixe (PREF) est utilisé dans ce dernier cas : `'NUM/DIGITS:MANY'`, `'NUM/DIGITS:4/PREF:20'` ..., `'NUM/DIGITS:1'`).

priori pour la reconnaissance d'autres phénomènes du langage. Nous nous distinguons ici des approches orientées données traditionnelles, pour lesquelles les prétraitements prennent rarement la forme de hiérarchie et fournissent souvent des traits en nombre fixe, afin d'alimenter des matrices.

3.1.2. *Lexiques*

La principale ressource lexicale que nous exploitons est le dictionnaire Prolex (Tran et Maurel, 2006), qui contient un grand nombre de noms propres finement catégorisés (villes, pays, hydronymes, chanteurs, groupes musicaux, partis politiques, etc.) et leurs dérivés. Cette ressource apporte beaucoup pour la REN, en particulier pour la reconnaissance des personnes et des lieux. En deuxième lieu, système à base de transducteurs `CasEN` nous permet d'exporter des motifs linguistiques lexicalisés sous forme d'entrées lexicales supplémentaires, par linéarisation des automates. Ceux-ci nous permettent de disposer d'expressions linguistiques variées liées à la REN (syntagmes nominaux constitutifs d'EN ou contextes discriminants). Enfin, nous avons créé manuellement quelques listes complémentaires pour améliorer la reconnaissance des fonctions, lieux, organisations, montants, dates et heures.

Au total, ces ressources contiennent 221 547 expressions distinctes qui produisent 443 112 catégorisations sémantiques[8]. De nombreuses entrées sont dédiées à la reconnaissance des personnes et des lieux, tandis qu'il y en a comparativement moins pour les organisations. Nous utilisons ces ressources telles quelles afin de produire un enrichissement, qui pourra dans de nombreux cas être sémantiquement ambigu. Ainsi, l'énoncé *'En 1970, Pompidou a été à Washington'* sera enrichi de la manière suivante (nous factorisons les suffixes de l'opérateur $\oplus$ pour plus de lisibilité) :

```
'TIME-MOD-PRE/PRP/en NUM/DIGITS:4/PREF:19/1970 PUN/,
(IND⊕TOPO⊕POL⊕VILLE)/NAMABR/NAM/Pompidou
VER/avoir VER/être PRP/à
(IND⊕TOPO⊕ORG-LOC-GOV⊕PREN⊕VILLE)/NAMABR/NAM/Washington'
```

À l'image des variations flexionnelles, nous considérons que les noms propres, étant une classe ouverte, n'ont pas vocation à être utilisés au sein des règles d'annotation. Une fois que les noms propres ont donné lieu à des enrichissements sémantiques, nous omettons les items lexicaux eux-mêmes pour extraire des règles d'annotation qui ne reposent que sur les enrichissements morphosyntaxiques ou lexicaux. Ce qui nous donne, pour notre exemple :

```
'TIME-MOD-PRE/PRP/en NUM/DIGITS:4/PREF:19/1970 PUN/,
(IND⊕TOPO⊕POL⊕VILLE)/NAMABR/NAM
VER/avoir VER/être PRP/à
(IND⊕TOPO⊕ORG-LOC-GOV⊕PREN⊕VILLE)/NAMABR/NAM'
```

8. Il y a donc en moyenne un peu plus de deux catégorisations sémantiques par entrée lexicale, mais ce chiffre est assez inégalement réparti : certaines entrées ont de nombreuses catégories, comme les célébrités, tandis que beaucoup d'entrées n'en ont qu'une seule.

3.2. *Exploration de règles d'annotation de segments*

Les données enrichies ont vocation à être fouillées afin d'y rechercher des motifs séquentiels d'intérêt (Fischer *et al.*, 2005 ; Cellier et Charnois, 2010) pour la REN. Dans un premier temps, nous allons décrire formellement le processus de fouille mis en œuvre pour extraire de telles règles. Nous verrons par la suite qu'il est nécessaire de filtrer les règles extraites pour constituer un système opérationnel.

Le langage obtenu après enrichissement des données par la morphosyntaxe et les lexiques est noté Σ_r. Or nous savons qu'il sera possible d'en généraliser les éléments : à partir de Σ_r, la prise en considération de toutes les généralisations nous permet de former le langage Σ_p, tel que $\Sigma_r \subset \Sigma_p$. Par ces généralisations, nous établissons un ordre partiel, noté $\geq_h$ au sein de Σ_p. Par exemple, `'NAMABR'` $\geq_h$ `'NAMABR/NAM'` $\geq_h$ `'NAMABR/NAM/France'`.

Dans le cadre des campagnes ESTER 2 ou ETAPE, les données d'apprentissage à notre disposition sont des corpus enrichis qui intègrent une délimation et un typage des entités. Ceux-ci sont réalisés par des balises typées (marqueurs) de début et de fin d'entité. Par exemple, le texte *'En 1970, Pompidou a été à Washington'* sera annoté de la manière suivante : *'En <date> 1970 </date>, <pers> Pompidou </pers> a été à <loc> Washington </loc>'*. Nous orientons la fouille de données par la présence de ces marqueurs au sein des textes. Ainsi, ces marqueurs sont conservés comme items pour opérer la fouille et nous distinguons, au sein de Σ_r, ces marqueurs comme un sous-alphabet Σ_m. Ceci vient en contraste avec le format traditionnel, dit BIO[9] qui attribue une étiquette à chaque mot du texte. L'objectif est de faire opérer la fouille sur une représentation similaire à celle qu'utilisent les transducteurs.

Afin de mener la fouille, nous nous dotons en premier lieu d'une relation qui comptabilise les occurences d'un item dans les données. Les enrichissements étant ambigus, ceci est pris en compte pour définir la couverture des items sur les données.

Couverture d'un item sur une donnée enrichie : soit un item $p \in \Sigma_r$ et une donnée enrichie $i = i_1 \oplus i_2 \oplus i_3 \ldots i_n$ telle que, pour tout j, $i_j \in \Sigma_r$, alors p *couvre* i, noté $p \geq_{ci} i$ s'il existe au moins un $k \in [1, n]$ tel que $p = i_k$.

Nous vérifions par exemple que l'item `'TOPO/Washington'` couvre bien `'CELEB/Washington`$\oplus$`TOPO/Washington'`. Il s'agit maintenant de déterminer comment la concaténation d'items couvre la concaténation de données. Nous émettons une proposition originale à ce sujet (Nouvel, 2012), qui consiste à couvrir les données par *segments*. Intuitivement, nous souhaitons qu'un élément de motif puisse couvrir indifféremment un ou plusieurs tokens. Par exemple, le trait sémantique pour une célébrité `'CELEB'` doit pouvoir couvrir indifféremment *'Valery Giscard d'Estaing'*, *'Georges Pompidou'* ou *'Zola'*. Ainsi, les motifs couvrent les données par « segments ».

9. *Begin, Inside, Outside.*

Motif de segments : un motif de segments $P = p_1 p_2 \dots p_n$ est une concaténation d'items de Σ_p telle que, pour tout $j \in [1, n-1]$ et $k \in [j+1, n]$ tels que tout $l \in [j+1, k-1]$ vérifie $p_l \in \Sigma_m$, alors $p_j \not\geq_h p_k$ et $p_k \not\geq_h p_j$.

Cette définition stipule que, pour ces motifs, deux items contigus (ou séparés par des marqueurs) ne peuvent être identiques ou parents l'un de l'autre. Ceci est requis pour satisfaire la propriété d'antimonotonie, ce qui permettra d'extraire des motifs fréquents. Nous définissons maintenant la couverture pour un tel motif.

Couverture d'un motif de segments sur des données : soit une séquence de la base de données enrichie $I = i_1 i_2 \dots i_m \in {\Sigma_r}^*$ et un motif de segments $P = p_1 p_2 \dots p_n \in {\Sigma_r}^*$ tels que $m \geq n$, alors P *couvre les segments de* I, noté $P \geq_{c^+} I$, s'il existe une fonction *discrète croissante surjective* $S()$ définie de $[1, m]$ vers $[1, n]$ telle que, pour tout $j \in [1, m]$, alors $p_{S(j)} \geq_{ci} i_j$.

Par exemple, `'A B E'` $\geq_{c^+}$ `'A B⊕C B⊕D E'` où l'item `'B'` couvre à la fois `'B⊕C'` et `'B⊕D'`. Cette formulation permet de décrire les données par *segments contigus*, qui pourront se situer à divers niveaux de la hiérarchie.

Afin d'explorer des motifs généralisés, nous reprenons les travaux en fouille de données sur le sujet (Srikant et Agrawal, 1996) que nous appliquons aux motifs de segments. Nous commençons par exploiter les hiérarchies.

Généralisation hiérarchique entre motifs de segments : soit deux motifs de segments $P = p_1 p_2 \dots p_n \in {\Sigma_p}^*$ et $Q = q_1 q_2 \dots q_m \in {\Sigma_p}^*$ tels que $m \geq n$, alors P *généralise hiérarchiquement les segments de* Q, noté $P \geq_{h^+} Q$, s'il existe une fonction *discrète croissante surjective* $S()$ définie de $[1, m]$ vers $[1, n]$ telle que, pour tout $j \in [1, m]$, alors $p_{S(j)} \geq_h q_j$.

Par exemple, `'A B C'` $\geq_{h^+}$ `'A B C/G C/H'` $\geq_{h^+}$ `'A B/D B/E B/F C/G C/H'`.

En plus de la généralisation selon la hiérarchie, un motif peut être généralisé par suppression d'items à sa gauche ou à sa droite.

Généralisation par affixation entre motifs : soit deux motifs $P = p_1 p_2 \dots p_n \in {\Sigma_p}^*$ et $Q = q_1 q_2 \dots q_p \in {\Sigma_p}^*$, alors P *généralise par affixation* Q, noté $P \geq_a Q$, si $p \geq n$ et s'il existe au moins un $k \in [0, p-n]$ tel que, pour tout $j \in [1, n]$, alors $q_{j+k} = p_j$.

Par exemple, `'B'` $\geq_a$ `'A B'` $\geq_a$ `'A B C'`.

Ces généralisations s'appliquent aux données du langage ${\Sigma_r}^*$, au sein duquel les marqueurs sont considérés comme des items particuliers. Nous ajoutons à leur sujet un mécanisme similaire à la transduction[10], que nous modélisons aussi comme une forme de généralisation. Ainsi, pour la fouille de données, supprimer un marqueur dans un texte est vu comme une forme particulière de généralisation.

10. Sans entrer dans les détails, la transduction permet d'insérer des symboles (lettres ou mots) au sein d'une séquence de symboles.

Généralisation sur marqueurs entre motifs : soit deux motifs $P = p_1 p_2 \dots p_n \in {\Sigma_p}^*$ et $Q = q_1 q_2 \dots q_p \in {\Sigma_p}^*$, alors P *généralise sur marqueurs* Q, noté $P \geq_m Q$, si $p \geq n$ et s'il existe une fonction *discrète strictement croissante surjective* $C()$ définie de $[1, n]$ vers $[1, p]$ telle que, pour tout $j \in [1, n]$, alors $p_j = q_{C(j)}$ et, pour tout $k \in [1, p]$ tel que $k \notin \{C(j), j \in [1, n]\}$, alors $q_k \in \Sigma_m$.

Par exemple, `'A B C'` $\geq_m$ `'A <loc> B C'` $\geq_m$ `'<pers> A </pers> <loc> B </loc> C'` : l'ajout des marqueurs spécialise le motif `'A B C'`.

En combinant ces trois relations de généralisation, un motif de segments est plus général qu'un autre si ses éléments sont plus élevés dans la hiérarchie, s'il est plus court ou s'il contient moins de marqueurs.

Généralisation entre motifs de segments : soit deux motifs $P \in {\Sigma_p}^*$ et $Q \in {\Sigma_p}^*$, alors P *généralise* Q, noté $P \geq_{g^+} Q$, s'il existe $R \in {\Sigma_p}^*$ et $S \in {\Sigma_p}^*$ tels que $P \geq_a R$, $R \geq_m S$ et $S \geq_{h^+} Q$.

Par exemple, `'A <pers> B </pers>'` $\geq_{g^+}$ `'A <pers> B/D B/E </pers> <loc> C </loc>'`. Par conséquent, si le motif `'A <pers> B/D B/E </pers> <loc> C </loc>'` couvre une séquence du texte, sa généralisation `'A <pers> B </pers>'` la couvrira également.

Pour résumer, ces généralisations permettent de rechercher des motifs dans lesquels apparaissent les marqueurs d'EN, modulo les éventuelles répétitions d'items (segments) et la présence d'autres marqueurs. Pour illustration, au sein de l'énoncé *'Le <fonc> président </fonc> <pers> Georges Pompidou </pers> débattait souvent.'*, nous relevons, par relations de couverture et de généralisation, une occurrence pour les motifs `'NOM/président <pers> CELEB </pers>'` et `'CELEB </pers> VERB/débattre'`.

Nous introduisons enfin la notion de règle d'annotation, comme un motif de segments particulier.

Règle d'annotation de segments : un motif $P \in {\Sigma_p}^*$ est une règle d'annotation de segments ssi P contient au moins un élément de Σ_r et un élément de Σ_m.

Les règles d'annotation sont alors des motifs qui contiennent nécessairement un élément du langage naturel et une balise d'annotation (ou marqueur). Notons qu'à ce stade les règles d'annotation contiennent un nombre indéterminé de marqueurs. Il conviendra de filtrer au besoin lors de l'extraction des motifs ou de s'assurer que l'on utilise ces règles de manière adéquate afin de produire une annotation valide.

3.3. *Filtrage et extraction de règles d'annotation partielles*

Comme la taille du langage ${\Sigma_p}^*$ est très importante, l'utilisation de critères de filtrage est nécessaire pour retenir les règles d'annotation les plus pertinentes vis-à-vis du corpus. Pour réaliser cela, nous proposons d'adapter le cadre des règles d'association (Agrawal *et al.*, 1993), bien connu en extraction de connaissances, à l'exploration

des règles d'annotation que nous avons définies. Nous définissons la fréquence et la confiance pour une règle d'annotation dans l'objectif d'éliminer les règles peu intéressantes.

Pour commencencer, la fréquence $Freq(P, \mathcal{D})$ (ou support lorsque cette valeur est utilisée comme seuil) correspond au nombre d'occurrences d'un motif au sein du corpus $\mathcal{D}$. Ces occurrences sont comptabilisées à l'aide de la couverture telle que définie ci-dessus. Plus la fréquence est élevée, plus la règle « s'applique ». Bien entendu, certaines règles d'annotation peuvent parfois s'appliquer à tort. Nous définissons alors la confiance d'une règle d'annotation P comme la proportion de phrases où la règle est appliquée avec justesse :

$$Conf(P, \mathcal{D}) = \frac{Freq(P, \mathcal{D})}{Freq(Ret_m(P), \mathcal{D})}$$

où $Ret_m(P)$ correspond au motif P sans les marqueurs. Par exemple, pour la règle `'NOM/président <pers> CELEB </pers>'`, il s'agira de rapporter sa fréquence à celle du motif `'NOM/président CELEB'`, afin de déterminer dans quelle proportion de cas le motif annote `'CELEB'` comme une célébrité.

Fixer des seuils pour ces critères nous permettra de limiter le nombre de motifs extraits. Cependant, même en fixant des seuils de support et de confiance sélectifs, les règles d'annotation peuvent être trop nombreuses à cause des combinaisons possibles au travers de la hiérarchie. Afin de contenir cette abondance de règles, nous proposons de grouper les règles, puis d'éliminer celles qui ne sont pas informatives, à l'instar de Pasquier *et al.* (1999). L'idée forte est que deux motifs qui couvrent les mêmes exemples sont redondants car ils appartiennent à la même classe d'équivalence.

Équivalence de motifs au regard d'une base de données : soit P et Q deux motifs et $\mathcal{D}$ une base de données, alors P est *équivalent à Q au regard de $\mathcal{D}$*, notée $P \equiv_{\mathcal{D}} Q$, si $P \geq_g Q$ ou $Q \geq_g P$ et $Freq(P, \mathcal{D}) = Freq(Q, \mathcal{D})$

Dans la suite, plutôt que d'extraire toutes les règles d'une même classe d'équivalence, nous nous contenterons des motifs les plus spécifiques (maximaux). Ce principe est illustré dans la figure 4. Le diagramme de Hasse présente des motifs extraits à partir des séquences `'A/E B/F C/G'`, `'A/E B/F C/H'` et `'A/E B/F D/G'`, à fréquence 2 et 3. Par classe d'équivalence, nous sélectionnons les motifs maximaux (motifs encadrés).

Cette extraction de règles d'annotation nous fournit de nombreux motifs qui sont corrélés aux marqueurs d'annotation, donc aux frontières (début ou fin) des entités nommées. À ce stade, il est possible d'utiliser ces règles de diverses manières. D'une part, elles permettent d'observer les mécanismes qui sont liés à la REN de manière objective. Nous pouvons également les utiliser directement pour implémenter un système pour la REN. Enfin, notons qu'il est aussi possible, à partir des motifs, d'alimenter un système à base de connaissances.

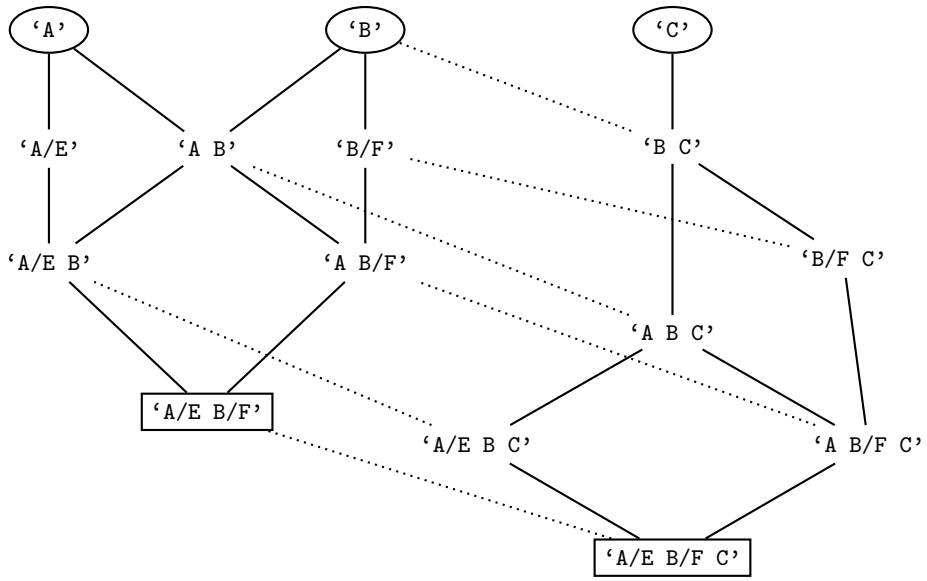


Figure 4. *Classes d'équivalence des motifs*

3.4. *Reconnaissance d'entités nommées à l'aide des règles extraites*

Les règles extraites peuvent être utilisées afin de réaliser l'annotation de textes. À cet effet, rappelons que l'application des règles d'annotation ne produit pas nécessairement une annotation complète ou valide. Certaines règles n'insèrent qu'un seul marqueur, et l'application des règles disponibles offre de nombreuses possibilités d'annotations. Nous adoptons ici une approche orientée données, qui a pour objectif de placer indépendamment les marqueurs au sein des textes. Contrôler que l'annotation produite est valide en termes de balises est réalisé en dernière étape, en même temps que sont sélectionnées les solutions les plus vraisemblables.

Nous avons pu vérifier lors de nos expériences (Nouvel, 2012) que l'utilisation de règles par un modèle probabiliste était pertinente. Ainsi, pour une position i d'une séquence, les marqueurs proposés par des règles $\mathcal{R}_i$ sont considérés comme autant de fonctions caractéristiques d'un modèle à régression logistique (Berger *et al.*, 1996) visant à estimer la probabilité d'insérer des marqueurs M_i à cette même position :

$$P(M_i = m_1 \dots m_k | \mathcal{R}_i) = \frac{exp\left(\sum_{T \in \mathcal{R}_i} \lambda_{T,m_1 \dots m_k}\right)}{Z(\mathcal{R}_i)}$$

où les paramètres $\lambda_{T,m_1 \dots m_k}$ correspondent aux poids des diverses règles dans l'estimation de la probabilité d'une séquence de marqueurs $m_1 \dots m_k$ donnée et où le

dénominateur $Z(\mathcal{R}_i)$ est un facteur de normalisation. L'ajustement de ces paramètres est réalisée à l'aide de l'outil maxent[11].

Enfin, lorsque les probabilités de séquences de marqueurs $P(M_i)$ sont estimées, il convient de les utiliser afin de déterminer quelle annotation est la plus vraisemblable. À cet effet, nous faisons une hypothèse d'indépendance entre marqueurs insérés au sein d'un énoncé :

$$P(M_1 M_2 \dots M_n) \approx \prod_{j=1}^{n} P(M_j)$$

Notons ici que, parmi les marqueurs qu'il est possible d'insérer aux diverses positions d'un énoncé, un nombre restreint de combinaisons forment une annotation valide (selon le guide d'annotation). Généralement, outre le fait qu'une entité nommée est nécessairement reconnue par deux marqueurs (un ouvrant et un fermant) et que les chevauchements sont interdits, il peut aussi s'agir de stipuler quelles imbrications sont tolérées. Dans les cas que nous aurons à prendre en compte, il sera aisé de déterminer si une séquence de marqueurs M_i à insérer à la position i, lorsque l'on connaît les marqueurs précédemment positionnés $M_1 \dots M_{i-1}$, amorce une annotation qui pourra être valide. L'espace de recherche pourra ainsi être exploré en examinant séquentiellement les marqueurs à insérer. Trouver les annotations valides les plus vraisemblables sera résolu grâce à des techniques de programmation dynamique.

Le système ainsi conçu cherche à exploiter simultanément les avantages des approches orientées connaissances (ou systèmes « symboliques ») comme les transducteurs de `CasEN` qui sont utilisés comme ressources lexicales et des approches orientées données (ou apprentissage automatique, ou encore systèmes « statistiques ») par paramétrage automatique du modèle à l'aide de la régression logistique. La fouille de données nous permet d'extraire des motifs utiles à la reconnaissance d'entités nommées. Enfin, l'approche voit la reconnaissance d'entités nommées comme une tâche d'insertion de balises d'annotation, plutôt que comme une catégorisation de mots.

3.5. *Hybridation avec le système CasEN*

Le laboratoire LI développe depuis plusieurs années un système à base de connaissances dédié à la REN[12]. `CasEN` fonctionne avec l'outil CasSys (Friburger et Maurel, 2011), qui permet l'analyse de textes à l'aide de cascades de transducteurs. Reposant sur la plate-forme Unitex[13], CasSys applique des transducteurs dans un ordre prédéfini, pour détecter des îlots de certitude, tout en réduisant progressivement l'espace de recherche (Abney, 1991). La cascade de transducteurs `CasEN`[14] en est une applica-

11. http ://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html
12. Ce système a également participé aux campagnes d'évaluation ESTER 2 et ETAPE.
13. http ://www-igm.univ-mlv.fr/∼unitex/
14. http ://tln.li.univ-tours.fr/Tln_CasEN.html

tion à la REN, initialement dédiée aux textes écrits et développée sur des textes journalistiques (Friburger et Maurel, 2004 ; Friburger, 2002). L'implémentation de motifs de surface détectant la structure des EN ou leur contexte permet de réaliser la reconnaissance des EN.

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), ce qui a ensuite permis une participation à ESTER 2 (Nouvel *et al.*, 2010) et ETAPE.

Nous avons cherché à utiliser conjointement les systèmes `mXS` et `CasEN`. Pour ce faire, nous tenons compte de la sortie des transducteurs de la même manière que les règles d'annotation : les frontières détectées par `CasEN` sont autant de fonctions caractéristiques, utilisées par la régression logistique pour estimer la vraisemblance des séquences de marqueurs. Nous avons vérifié que cette technique d'hybridation était satisfaisante, face à d'autres possibilités (Nouvel *et al.*, 2012). En complément à `mXS`, c'est par cette hybridation que nous parvenons à tisser des liens entre approches à base de connaissances (transducteurs) et orientées données (règles d'annotation).

4. Expérimentations et performances

Cette section décrit l'application de l'approche pour les jeux de données dont nous disposons. Nous y détaillons quelques considérations spécifiques à propos de l'implémentation du système, puis nous en rapportons les performances.

4.1. *Spécificités d'implémentation du système*

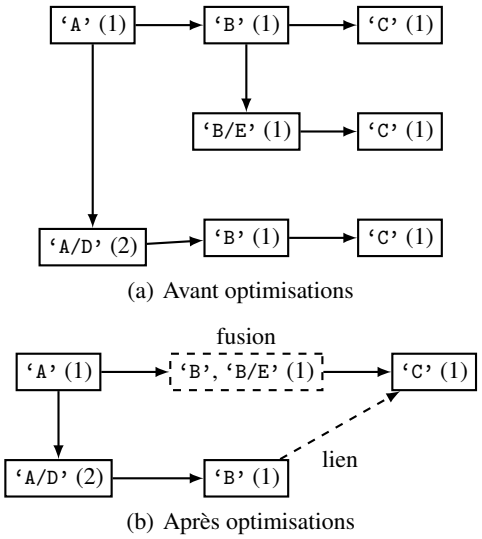
Pour extraire les règles d'annotation, nous utilisons une approche par niveaux, qui repose sur la propriété d'*antimonotonie* (Agrawal et Srikant, 1995 ; Pei *et al.*, 2004). Cette propriété stipule qu'aucun motif $P = p_1 \dots p_n$ ne peut être plus fréquent que $p_1 \dots p_{n-1}$ ou $p_2 \dots p_n$. Ainsi, l'exploration des données itère sur la taille des motifs : à partir d'un niveau n, tous les motifs *candidats* de taille $n + 1$ sont générés, leurs fréquences sont relevées dans les données puis confrontées au seuil requis afin d'écarter ceux qui n'ont pas une fréquence suffisante. Cette approche nous épargne de représenter et de tester tous les motifs de $(\Sigma_p)^*$, mais il sera alors crucial de pouvoir, à chaque itération, parcourir très rapidement la base de données afin d'y relever les occurrences des motifs pour un niveau donné.

Cette approche par niveaux est implémentée par utilisation d'un arbre des préfixes pour représenter les motifs fréquents. En extension aux techniques traditionnelles pour son implémentation, l'utilisation de hiérarchies nous permet d'optimiser la structure de l'arbre afin d'en limiter la taille et le temps de parcours (Wang et Han, 2004 ; Qian *et al.*, 2010 ; Bonchi et Lucchese, 2005). Effectivement, les propritétés des hiérarchies nous permettent d'établir qu'un nœud de l'arbre peut stocker plusieurs motifs lorsque ceux-ci ont exactement les mêmes occurrences dans les données. Or, pour deux motifs, nous pouvons affirmer qu'ils ont les mêmes occurrences dans les données s'ils sont

une généralisation l'un de l'autre et s'ils ont la même fréquence. Selon ce principe, deux optimisations se sont avérées simples et efficaces :

– **fusion de suffixes** : deux nœuds feuilles de l'arbre d'un même parent qui ont la même fréquence et dont l'un est généralisation de l'autre peuvent être fusionnées ;

– **lien de préfixes** : deux nœuds de parents distincts qui ont la même fréquence et dont l'un est généralisation de l'autre auront les mêmes nœuds enfants.



(a) Avant optimisations

(b) Après optimisations

Figure 5. *Optimisations de l'abre des préfixes*

La figure 5 illustre ces techniques sur un exemple minimal. Les nœuds feuilles des motifs `'A B'` et `'A B/E'` peuvent être fusionnés. De la même manière, les nœuds feuilles des motifs `'A B'` et `'A/D B'` peuvent être liés au même nœud enfant `'C'`.

4.2. *Résultats de l'extraction*

Nous procédons à l'extraction des règles d'annotation selon divers seuils de fréquence et de confiance. Nos expériences (Nouvel, 2012) ont montré l'intérêt d'extraire de nombreux motifs, y compris ceux qui sont faiblement corrélés aux entités nommées. La figure 6 donne le spectre des motifs selon la fréquence et la confiance, sur le corpus ETAPE. Alors que celle-ci nous confirme que le nombre de motifs croît fortement lorsque la fréquence est abaissée, nous constatons une égale répartition des motifs selon la confiance, ce qui était moins attendu.

Ainsi, nous voyons qu'il sera difficile pour `mXS` d'explorer exhaustivement les règles avec un seuil de fréquence bas. En revanche, il est possible d'utiliser un large

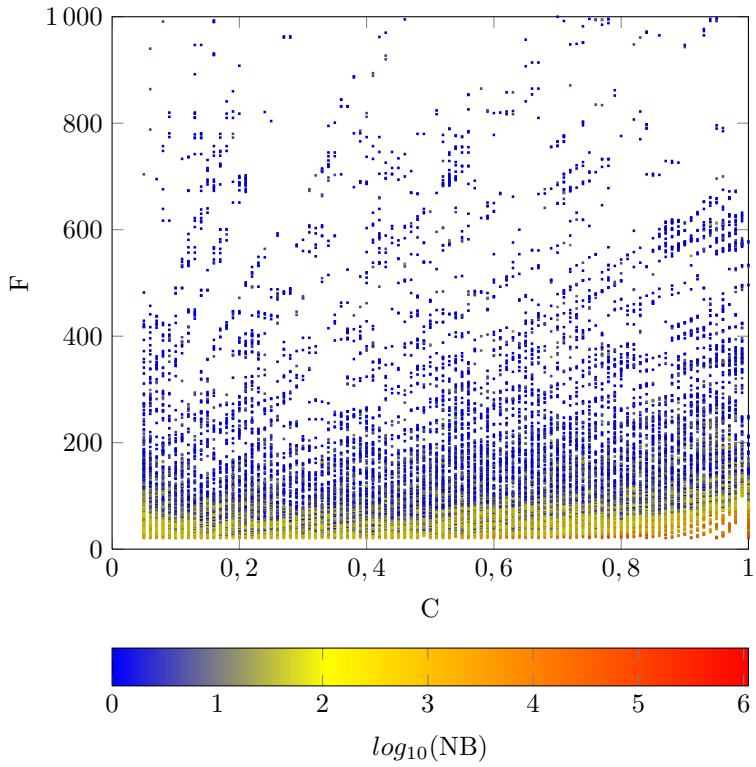
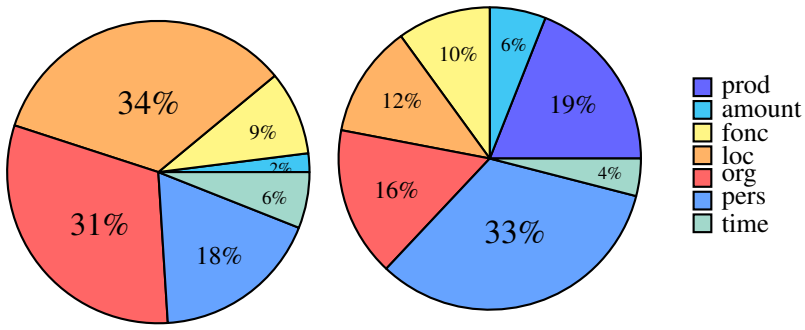


Figure 6. *Nombre de règles extraites (NB) selon la fréquence (F) et la confiance (C)*

spectre de confiance, ce qui permettra de bénéficier de nombreux indices, même lorsqu'ils sont faiblement corrélés aux marqueurs d'EN.

Nous examinons en figure 7 la répartition du nombre de règles d'annotation selon les types d'EN. De manière générale, les proportions sont corrélées aux nombres d'EN dans les corpus concernés. Nous voyons que les proportions de motifs pour les `pers` sont bien plus importantes pour ETAPE que pour ESTER 2, ce qui est dû à la prise en compte d'expressions construites autour de noms communs. Nous constatons dans les deux cas une assez nette sous-représentation des types `time` et `amount` : il semble qu'il y ait objectivement, avec l'approche que nous mettons en œuvre, moins de descripteurs pour ces types. Nous en déduisons qu'ils sont relativement homogènes dans les données. Le type `prod`, inversement, est surreprésenté au sein des motifs extraits d'ETAPE (il est quasiment absent d'ESTER 2) et nous faisons l'hypothèse que ce type est assez hétérogène.



(a) ESTER 2 (b) ETAPE

Figure 7. *Répartition des types d'EN au sein des motifs*

4.3. *Métriques d'évaluation des performances*

Dans les sections qui suivent, et à l'image des campagnes d'évaluation sur lesquelles nous reportons nos résultats, notre principale métrique d'évaluation sera le SER (*Slot Error Rate*) (Makhoul *et al.*, 1999). Celui-ci est réalisé par une recherche de correspondances entre les entités de la référence et de l'hypothèse, en considérant les erreurs suivantes (Galibert *et al.*, 2011) :

– suppressions (D) : l'entité n'est présente que dans la référence ;

– insertion (I) : l'entité n'est présente que dans l'hypothèse ;

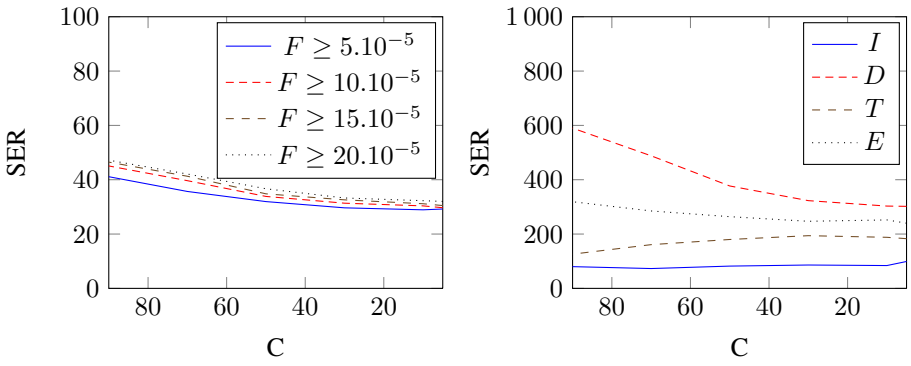
– type (T) : une même entité est présente dans la référence et dans l'hypothèse, mais elles n'ont pas le même type ;

– extension (E) : une même entité est présente dans la référence et dans l'hypothèse, mais elles n'ont pas les mêmes frontières.

Notons qu'il est possible d'avoir à la fois une erreur de type et d'extension, que nous notons ici TE. Ce nombre d'erreurs est pondéré et normalisé selon le nombre d'entités de la référence $\#R$:

$$SER = \frac{D + I + TE + 0,5(T + E)}{\#R}$$

Lorsqu'il y a nécessité de prendre une décision entre plusieurs affectations des erreurs d'un système (par exemple si une entité en hypothèse en chevauche plusieurs en référence), le score le moins pénalisant sera systématiquement choisi. En plus de cette mesure, nous tenons compte de la précision, du rappel et de la f-mesure pour analyser le comportement du système. Le SER a l'avantage de moduler les erreurs, en tenant compte de la capacité de certains sytèmes à reconnaître partiellement les entités les plus difficiles.



(a) Performances (b) Erreurs, $F \geq 5.10^{-5}$

Figure 8. *Performances (SER), erreurs d'insertion (I), de suppression (D), de type (T), d'extension (E) selon la fréquence relative (F) et la confiance (C) avec l'approche avec motifs de segments* `Logit+Segs` *sur ESTER 2*

4.4. *Performances sur ESTER 2*

Le système `mXS` a été développé pour une participation à la campagne d'évaluation ETAPE. Nous avons utilisé dans un premier temps le jeu de données issu de la campagne ESTER 2 pour étudier l'influence des paramètres de filtrage du modèle (fréquence, confiance) sur les résultats. La figure 8 en présente les principaux enseignements. Nous y voyons la confirmation de l'intérêt d'extraire des règles largement, en particulier pour un faible seuil de confiance. Nous constatons également que notre système commet plus d'erreurs de suppression (silence), même si elles se réduisent à mesure que le seuil de confiance est abaissé. Afin de mieux situer le niveau de performance atteint, nous le comparons avec un système CRF. Pour ce dernier, nous utilisons l'outil wapiti[15] (Lavergne *et al.*, 2010), alimentés par les mêmes éléments que pour l'approche à base de motifs (*feature* pour le token enrichi, *features* pour la catégorie morphosyntaxique, celle du token précédent et celle du token suivant, *features* sur les catégories morphosyntaxiques détaillées, *features* binaires pour chaque trait issu des ressources lexicales). Ainsi, les configuration sont les suivantes :

– `mXS` : système tel que décrit ci-dessus ;

– `mXS-Dicos` : identique à `mXS`, mais les ressources lexicales sont désactivées ;

– `CRF` : approche CRF ;

– `CasEN` : système `CasEN`, dans une version améliorée depuis ESTER 2 ;

– `mXS+CasEN` : identique à `mXS`, avec l'ajout des EN reconnues par `CasEN` comme niveau supplémentaire de la hiérarchie au-dessus des ressources lexicales.

15. http ://wapiti.limsi.fr

Approche	SER	I	D	T	E	P	R	Fm
mXS	**27,97**	81	301	184	221	83,51	70,85	0,77
mXS-Dicos	**33,85**	97	353	322	181	75,56	65,90	0,70
CRF	**24,64**	72	229	183	200	84,97	77,90	0,81
CasEN	**28,58**	42	343	165	260	85,21	69,78	0,77
mXS+CasEN	**24,64**	79	243	189	191	83,21	74,40	0,79

Tableau 3. *Performances (SER), erreurs d'insertion (I), de suppression (D), de type (T), d'extension (E) et précision (P), rappel (R), f-mesure (Fm) des approches*

Le tableau 3 récapitule les paramètres des approches et le détail des performances obtenues. Sans surprise, on constate que le système mXS présente une moins bonne précision que le système symbolique CasEN. La couverture du système, évaluée par le rappel, est, quant à elle, supérieure à celle de CasEN. Cette domination en termes de rappel reste toutefois assez limitée (70,85 % contre 69,78 %), ce qui suggère que la base de connnaissances développée au cours du temps pour CasEN est couvrante. Au final, les systèmes mXS et CasEN ont donc des performances assez proches (F-mesure identique et SER proche). Leurs points forts et leurs points faibles semblent toutefois complémentaires, puisque l'hybridation des deux systèmes se traduit par une augmentation significative des performances, qui place le système au niveau du CRF.

Une analyse plus fine des erreurs montre en particulier que l'hybridation par mXS induit certes l'insertion d'erreurs par rapport au système à bonne précision qu'est CasEN, mais permet une réduction significative du nombre d'entités oubliées. Cette réduction concerne aussi bien les erreurs de suppression qu'auraient faites mXS et CasEN, preuve que l'hybridation autorise de nouvelles détections de bonne qualité. En comparant le système hybride à CasEN, le nombre d'erreurs d'insertion ajoutées (37) est ainsi trois fois inférieur aux suppressions évitées (100). Nous remarquons enfin que les erreurs d'extension (frontières) sont en baisse sensible (260 pour CasEN contre 191 pour le système hybride, mais mXS se comportait déjà mieux que CasEN sur ce point). Ces résultats suggèrent que la détection séparée des marques de début et de fin d'entité telle que réalisée par mXS est susceptible d'améliorer la détection des frontières (en particulier terminale) des entités.

4.5. *Performances sur ETAPE*

Sur le corpus ETAPE, le système affiche, de manière générale, un comportement assez similaire. Nous menons des évaluations séparées des types primaires d'EN et de composants, que nous comparons avec l'évaluation globale dans le tableau 4. Les performances globales sont moins bonnes (ceci est dû à la difficulté plus importante pour ce corpus, et a été constaté par tous les systèmes participant à cette campagne). Il apparaît que les EN sont moins bien reconnues que leurs composants, en particulier du point de vue des erreurs de substitution : leur empan plus large et leur type relevant plus de la sémantique semblent les rendre plus difficiles à identifier.

Types	SER	I	D	S	P	R	Fm
Entités	**38,9**	6,9	25,4	12,3	76,4	62,3	68,6
Composants	**33,0**	4,2	25,0	6,5	86,4	68,5	76,4
Test	**35,9**	5,6	24,2	10,8	79,8	64,9	71,6

Tableau 4. *Performances (SER), erreurs d'insertion (I), de suppression (D), de substitution (S) et précision (P), rappel (R), f-mesure (Fm) sur les types primaires et sur toutes les annotations d'ETAPE*

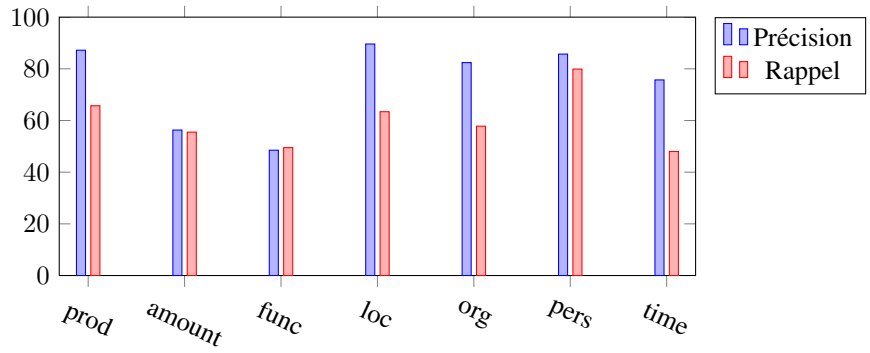
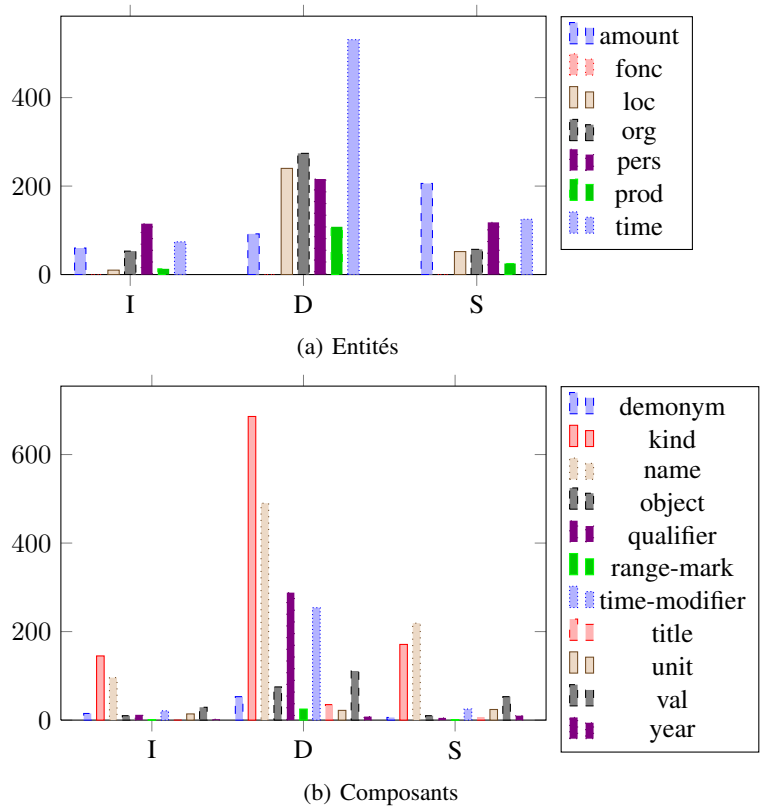


Figure 9. *Performances par types primaires (ETAPE)*

La figure 9 donne la précision et le rappel lorsque l'on réalise l'évaluation uniquement sur les types primaires d'entités d'ETAPE (en faisant abstraction des sous-types), et la figure 10 les nombres d'erreurs correspondants. Nous notons la difficulté à détecter les dates et heures, ce qui se traduit par un rappel relativement faible et un nombre d'erreurs conséquent. Nous remarquons aussi que les montants et les fonctions restent difficiles à reconnaître correctement. Les difficultés pourraient être liées au fait que ces EN correspondent souvent à des descriptions définies.

L'adjudication de la campagne d'évaluation ETAPE n'était pas achevée à l'heure de la rédaction de cet article, mais les résultats sont à paraître prochainement (Galibert *et al.*, 2014). Nous avons cependant été autorisés à reporter au tableau 5 les performances anonymes des systèmes. Les SER présentés sont donnés sur les transcriptions manuelles et sur les sorties de différents systèmes de reconnaissance : le *Rover* est un système de vote majoritaire des divers systèmes ASR, ces derniers étant présentés par colonnes selon leur WER[16]. Pour les systèmes REN en compétition, nous distinguons les systèmes orientés connaissances (OC) et ceux qui font appel à des champs aléatoires conditionnels (CRF) ou une combinaison de CRF et de grammaires probabilistes (CRF+PCFG).

16. *Word Error Rate*, où pour WXX le taux d'erreurs est XX.



(a) Entités

(b) Composants

Figure 10. *Erreurs par types primaires d'EN et de composants sur ESTER 2*

Part.	**Type**	**Man**	**Rover**	**W23**	**W24**	**W25**	**W30**	**W35**
1	OC	84,8	98,1	100,7	94,2	98,9	98,4	100,9
2	OC	172,0	147,4	178,8	160,4	168,0	163,9	168,2
3	CRF	**33,8**	**57,2**	**59,3**	64,7	**62,0**	**61,7**	**71,8**
4	OC	55,6	88,0	98,8	76,8	92,8	94,9	99,6
5	CRF	43,6	69,7	73,8	72,1	73,7	74,8	86,0
6	CRF+PCFG	na	79,2	79,5	66,8	80,8	80,0	87,0
7	CRF+PCFG	na	67,8	68,4	67,6	70,9	69,9	85,2
8	CRF+PCFG	36,4	na	na	na	na	na	na
9	CRF	62,8	75,8	79,2	76,9	79,8	80,5	90,5
10	OC	42,9	65,0	69,9	66,3	70,5	69,9	87,0
11	OC	49,3	na	na	68,4	na	na	na
mXS		41,0	63,7	67,5	**64,1**	69,1	68,6	80,4

Tableau 5. *SER de la campagne ETAPE par système sur les transcriptions avant adjudication (manuel : Man, transcription automatiques : Rover et WXX)*

Parmi les autres systèmes participants, le système 3 utilise des CRF (binarisés, un par type) (Raymond, 2013), le système évalué en lignes 6, 7 et 8 (dans trois configurations différentes) utilise un CRF pour les composants et un PCFG pour reconstituer les entités. De manière générale, `mXS` affiche de bonnes performances (entre la première et la troisième position). Les taux d'erreurs élevés sont liés à la difficulté de la tâche (parole spontanée, imbrications, typologie fine). Sans surprise, les performances sont dégradées sur les données bruitées par la reconnaissance de la parole. Dans ce cas, nous voyons que `mXS` résiste bien aux erreurs de reconnaissance de la parole.

Afin de mieux déterminer quels sont les avantages et les inconvénients des systèmes, en particulier lorsqu'il s'agit de les comparer avec des systèmes de type CRF qui ne sont pas conçus pour réaliser des annotations imbriquées, nous sélectionnons les annotations à réaliser, soit avant la phase d'apprentissage, soit au moment de l'évaluation, selon les configurations suivantes :

– **all** : toutes annotations (Wapiti-join étant la simple concaténation des labels BIO pour chaque niveau d'annotation) ;

– **XXX/outer** : annotations les plus larges (EN) pour l'évaluation ;

– **XXX/inner** : annotations les moins larges (COMP) pour l'évaluation ;

– **outer/XXX** : annotations les plus larges (EN) dès l'apprentissage ;

– **inner/XXX** : annotations les moins larges (COMP) dès l'apprentissage.

Le tableau 6 donne les performances des systèmes dans ces configurations. De manière générale, nous constatons que le système hybride obtient les meilleures performances, même s'il n'atteint pas les performances du meilleur système de la campagne ETAPE. Comme pour ESTER 2, nous remarquons que le système `mXS` seul est en retrait par son silence important (67 % de rappel), mais, en hybridation `mXS+CasEN`, il obtient très nettement les meilleures performances.

Dans les expériences EN et COMP, nous constatons les bonnes performances obtenues par `mXS` en comparaison avec les modèles CRF lorsque les annotations sont sélectionnées dès l'apprentisssage. En particulier, `mXS` obtient des performances équivalentes (42 % contre 42,4 % pour le CRF) lorsqu'il s'agit d'annoter les entités les plus larges. Ceci conforte l'idée que la reconnaissance séparée du début ou de la fin d'une EN donne de bons résultats sur les entités d'empan large. En revanche, nous constatons que lorsque l'apprentissage tient compte de toutes les annotations, ceci pénalise le système qui commet bien plus d'insertions (14 % contre 11,3 %).

Ces résultats viennent confirmer l'idée qu'il est possible de reconnaître des EN par recherche des marqueurs, sans nécessairement avoir à catégoriser tous les tokens d'une telle expression. Pour l'imbrication d'annotations, nous sommes confrontés à la même problématique que pour un système CRF, à savoir la difficulté à définir un modèle qui reconnaisse simultanément plusieurs annotations sans dégrader ses performances. Cependant, nos résultats pour la campagne ETAPE montrent que chaque approche peut apporter selon que l'on considère les entités ou leurs composants : il serait probablement avantageux de combiner leurs sorties.

		SER%	I%	D%	S%	P	R	Fm
all	mXS	**39,1**	11,5	22	10,3	75,7	67,7	71,5
	mXS+CasEN	**34,6**	11,5	18,1	9,2	77,8	72,7	75,1
	Wapiti-join	**37,3**	12,7	18,4	11,4	74,5	70,3	72,3
EN	outer/mXS	**42**	11,3	23,5	13,1	72,1	63,3	67,4
	outer/mXS+CasEN	**37,5**	11,7	19,4	12,4	73,9	68,2	70,9
	outer/Wapiti	**42,4**	14,1	18,9	17,6	66,7	63,5	65,1
	mXS/outer	**44**	14	21,7	15,4	68,2	63	65,5
	mXS+CasEN/outer	**39,3**	13,2	19	13,6	71,6	67,4	69,4
COMP	inner/mXS	**36,7**	14,5	17,1	9,1	75,8	73,8	74,8
	inner/mXS+CasEN	**31,8**	13,6	13,9	7,7	78,6	78,4	78,5
	inner/Wapiti	**36**	13,2	17,5	9,4	76,4	73,1	74,7
	mXS/inner	**37,1**	11	21,9	7,5	79,2	70,6	74,6
	mXS+CasEN/inner	**33,8**	11,5	18,4	7	80,1	74,6	77,3

Tableau 6. *Comparatif des erreurs d'insertion (I), de suppression (D), de substitution (S) et précision (P), rappel (R), f-mesure (Fm) de la sélection des entités lors de l'apprentissage (inner/XXX, outer/XXX) ou lors de l'évaluation (XXX/inner XXX/outer)*

5. Conclusion

Les travaux présentés dans cet article portent sur la reconnaissance des entités nommées à l'aide de techniques de fouille de données. Nous formalisons une approche originale, qui extrait des motifs séquentiels hiérarchiques par segments comme règles d'annotation. Celle-ci est vue comme un moyen terme entre les systèmes orientés connaissances et les systèmes guidés par les données. L'observation des motifs extraits nous permet d'étudier les types d'entités nommées à reconnaître. Les règles sont utilisées par une approche à régression logistique, qui détecte séparément le début ou la fin des entités nommées pour réaliser les annotations. Dans le cadre de campagnes d'évaluation, le système mXS obtient de très bonnes performances pour la reconnaissance des entités nommées sur des transcription de la parole en contexte difficile.

Les expériences que nous avons menées montrent qu'il est possible de combiner avantageusement les approches orientées connaissances et celles orientées données. Ceci passe par la définition d'un mode de représentation commun : notre choix s'est porté sur la recherche de balises d'annotation, ce qui se rapproche des mécanismes de transduction. Nous envisageons maintenant de mieux déterminer dans quelle mesure ce choix pèse sur les performances observées. Par ailleurs, la mise en œuvre de techniques de fouille nous permet de nous appuyer sur les données afin d'en extraire des motifs et d'interagir avec les connaissances au sein d'un système unifié. Les pistes que nous envisageons comme perspectives à ces travaux sont nombreuses, parmi celles-ci figurent des comparaisons plus poussées de notre proposition avec les approches orientées connaissances ou guidées par les données, l'expérimentation du système pour d'autres tâches d'annotation, ou encore l'exploitation des motifs de segments comme ressources pour les bases de connaissances.

Remerciements

Ce travail a été réalisé dans le cadre des projets ESTER 2 (AFCP, DGA, ELDA) et ETAPE (ANR-09-CORD-009). Merci en particulier à Olivier Galibert (LNE), Matthieu Carré (ELDA) et Guillaume Gravier (IRISA) pour avoir mis le corpus de test et les outils d'évaluation à notre disposition. Merci à l'INRIA et au LIMSI pour avoir permis la rédaction de cet article.

6. Bibliographie

Abney S. P., *Parsing by Chunks*, Springer, p. 257-278, 1991.

Agrawal R., Imielinski T., Swami A. N., « Mining Association Rules between Sets of Items in Large Databases », *ACM SIGMOD International Conference on Management of Data (MOD'93)*, p. 207-216, 1993.

Agrawal R., Srikant R., « Mining Sequential Patterns », *International Conference on Data Engineering (ICDE'95)*, p. 3-14, 1995.

Berger A. L., Pietra S. A. D., Pietra V. J. D., « A Maximum Entropy approach to Natural Language Processing », *Computational Linguistics*, vol. 22, p. 39-71, 1996.

Bonchi F., Lucchese C., « Pushing Tougher Constraints in Frequent Pattern Mining », *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'05)*, p. 114-124, 2005.

Brun C., Ehrmann M., « Un système de détection d'entités nommées adapté pour la campagne d'évaluation ESTER 2 », *Traitement automatique des langues naturelles (TALN'10)*, 2010.

Bunescu R. C., Pasca M., « Using Encyclopedic Knowledge for Named entity Disambiguation », *Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, 2006.

Béchet N., Roche M., « How to Expand Dictionaries with Web-Mining Techniques, CogALex Workshop », *International Conference on Computational Linguistics (COLING'10)*, Beijing, China, p. 33-37, 2010.

Cellier P., Charnois T., « Fouille de données séquentielles d'itemsets pour l'apprentissage de patrons linguistiques », *Traitement automatique des langues naturelles (TALN'10)*, 2010.

Charton E., Gagnon M., Ozell B., « Génération automatique de motifs de détection d'entités nommées en utilisant des contenus encyclopédiques », *Traitement automatique des langues naturelles (TALN'11)*, 2011.

Ehrmann M., Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation, PhD thesis, Université Paris VII, France, 2008.

Etzioni O., Cafarella M., Downey D., Popescu A.-M., Shaked T., Soderland S., Weld D. S., Yates A., « Unsupervised named-entity extraction from the Web : An experimental study », *Artificial Intelligence*, vol. 165, p. 91-134, 2005.

Ezzat M., « Acquisition de grammaires locales pour l'extraction de relations entre entités nommées », *Rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues (RECITAL'10)*, 2010.

Favre B., Béchet F., Nocera P., « Robust Named Entity Extraction from Large Spoken Archives », *Joint Conference on Human Language Technology Conference and Empirical Methods in Natural Language Processing (HLT/EMNLP'05)*, 2005.

Fischer J., Heun V., Kramer S., « Fast Frequent String Mining Using Suffix Arrays », *5th IEEE International Conference on Data Mining (ICDM'05)*, p. 609-612, 2005.

Freitag D., Kushmerick N., « Boosted wrapper induction », *European Conference on Artificial Intelligence (ECAI'00) - Workshop on Machine Learning for Information Extraction*, Berlin, Germany, 2000.

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

Friburger N., Maurel D., « Finite-state transducer cascades to extract named entities in texts », *Theoretical Computer Sciences (TCS)*, vol. 313, p. 93-104, 2004.

Friburger N., Maurel D., « Writing and ordering FS cascades for NLP Tasks using Unitex », *Workshop on Finite-State Methods and Natural Language Processing (FSMNLP'11)*, 2011.

Galibert O., Leixa J., Adda G., Choukri K., Gravier G., « The ETAPE speech processing evaluation », *Proc of LREC*, ELRA, Reykjavik, Iceland, 2014.

Galibert O., Rosset S., Grouin C., Zweigenbaum P., Quintard L., « Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions », *International Joint Conference on Natural Language Processing (IJCNLP'11)*, 2011.

Galliano S., Gravier G., Chaubard L., « The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts », *International Speech Communication Association (INTERSPEECH'09)*, 2009.

Grouin C., Rosset S., Zweigenbaum P., Fort K., Galibert O., Quintard L., « Proposal for an Extension of Traditional Named Entities : From Guidelines to Evaluation, an Overview », *Conference of the Association for Computational Linguistics (ACL'11) - Fifth Linguistic Annotation Workshop (LAW-V)*, p. 92-100, 2011.

Lavergne T., Cappé O., Yvon F., « Practical Very Large Scale CRFs », *Annual Meeting of the Association for Computational Linguistics (ACL'10)*, p. 504-513, 2010.

Makhoul J., Kubala F., Schwartz R., Weischedel R., « Performance measures for information extraction », *DARPA Broadcast News Workshop*, p. 249-252, 1999.

Maurel D., Friburger N., Eshkol I., « Who are you, you who speak ? », *Language & Technology Conference (LTC'09)*, 2009.

Maurel D., Friburger N., Nouvel D., Eshkol-Taravella I., Antoine J.-Y., « Cascade de transducteurs : Applications autour des entités nommées », *Traitement automatique des langues (TAL)*, 2011.

Nadeau D., Sekine S., « A survey of named entity recognition and classification », *Linguisticae Investigationes*, vol. 30, p. 3-26, 2007.

Nouvel D., Reconnaissance des entités nommées par exploration de règles d'annotation, PhD thesis, Université François Rabelais Tours, 2012.

Nouvel D., Antoine J.-Y., Friburger N., Maurel D., « An Analysis of the Performances of the CasEN Named Entities Recognition System in the ESTER 2 Evaluation Campaign », *International Language Resources and Evaluation (LREC'10)*, 2010.

Nouvel D., Antoine J.-Y., Friburger N., Soulet A., « Coupling Knowledge-Based and Data-Driven Systems for Named Entity Recognition », *Innovative hybrid approaches to the processing of textual data (HYBRID'12, EACL Workshop)*, 2012.

Pasquier N., Bastide Y., Taouil R., Lakhal L., « Efficient Mining of Association Rules Using Closed Itemset Lattices », *INF. SYST.*, vol. 24, n° 1, p. 25-46, 1999.

Pei J., Han J., Mortazavi-Asl B., Wang J., Pinto H., Chen Q., Dayal U., Hsu M.-C., « Mining Sequential Patterns by Pattern-Growth : The PrefixSpan Approach », *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, p. 1424-1440, 2004.

Plantevit M., Charnois T., Klema J., Rigotti C., Cremilleux B., « Combining sequence and itemset mining to discover named entities in biomedical texts : a new type of pattern », *International Journal of Data Mining, Modelling and Management (IJDMM)*, vol. 1, p. 119-148, 2009.

Qian X., Zhang Q., Huang X., Wu L., « 2D Trie for Fast Parsing », *International Conference on Computational Linguistics (COLING'10)*, Beijing, China, p. 904-912, 2010.

Raymond C., « Robust Tree-Structured Named Entities Recognition from Speech », *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'13)*, 2013.

Raymond C., Fayolle J., « Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement », *Traitement automatique des langues naturelles (TALN'10)*, 2010.

Rosset S., Grouin C., Zweigenbaum P., Entité nommées structurées : guide d'annotation Quaero, Technical report, LIMSI (2011-04), 2011.

Savary A., Waszczuk J., Przepiórkowski A., « Towards the Annotation of Named Entities in the National Corpus of Polish », *International Language Resources and Evaluation (LREC'10)*, 2010.

Schmid H., « Probabilistic POS Tagging Using Decision Trees », *New Meth. in Lang. Proc. (NEMLP'94)*, 1994.

Snow R., Jurafsky D., Ng A. Y., « Learning syntactic patterns for automatic hypernym discovery », *Advances in Neural Information Processing Systems 17*, 2004.

Srikant R., Agrawal R., « Mining Sequential Patterns : Generalizations and Performance Improvements », *International Conference on Extending Database Technology (EDBT'96)*, p. 3-17, 1996.

Sun A., Grishman R., « Semi-supervised Semantic Pattern Discovery with Guidance from Unsupervised Pattern Clusters », *International Conference on Computational Linguistics (COLING'10)*, Beijing, China, p. 1194-1202, 2010.

Tran M., Maurel D., « Prolexbase - Un dictionnaire relationnel multilingue de noms propres », *Traitement automatique des langues (TAL)*, vol. 47-3, p. 115-139, 2006.

Wang J., Han J., « BIDE : Efficient Mining of Frequent Closed Sequences », *International Conference on Data Engineering (ICDE'04)*, 2004.

Zidouni A., Glotin H., Quafafou M., « Recherche d'entités nommées dans les journaux radiophoniques par contextes hiérarchique et syntaxique », *Conférence en Recherche d'Informations et Applications (CORIA'09)*, p. 421-432, 2009.