Annotation d'Entités Nommées par Extraction de Règles de Transduction

Damien Nouvel *,** Arnaud Soulet *,***

*Université François Rabelais Tours, Laboratoire d'Informatique
** damien.nouvel@univ-tours.fr,
** arnaud.soulet@univ-tours.fr,

Résumé. La reconnaissance d'entités nommées est une problématique majoritairement traitée par des modèles spécifiés à l'aide de règles ou par apprentissage numérique. Les premiers ont le désavantage d'être coûteux à développer pour obtenir une couverture satisfaisante, les seconds sont souvent difficiles à interpréter par des experts (linguistes). Dans cet article, nous présentons une approche, dont l'objectif est d'extraire des règles symboliques discriminantes qu'un humain puisse consulter. A partir d'un corpus de référence, nous extrayons des règles de transduction, dont seules les plus informatives sont retenues. Elles sont ensuite appliquées pour effectuer une annotation : à cet effet, un algorithme recherche parmi les annotations possibles celles de meilleure qualité en termes de couverture et de probabilité. Nous présentons les résultats expérimentaux et discutons de l'intérêt et des perspectives de notre approche.

1 Introduction

Parmi les tâches d'extraction d'information, la Reconnaissance d'Entités Nommées (REN) consiste à reconnaître (rechercher et catégoriser) toutes les Entités Nommées (EN) d'un texte : les expressions *univoques* et *référentiellement autonomes* (Ehrmann, 2008). Par simplification, ces unités correspondent intuitivement aux noms propres : personnes, lieux et organisations. En pratique seront aussi recherchées les expressions numériques et les expressions de temps, considérées comme "descriptions définies".

La REN est une tâche étudiée depuis une quinzaine d'années. Initialement symboliques, les approches sont aujourd'hui majoritairement tournées vers des modèles numériques par recherche de traits discriminants à l'aide d'apprentissage automatique (e.g., Chaînes de Markov, CRF, SVM). Cependant, ces méthodes permettent difficilement de capitaliser la connaissance modélisée.

Malgré ces récents développements, les dernières campagnes d'évaluation en français montrent que les systèmes symboliques demeurent plus performants, à condition d'utiliser des bases de connaissances suffisamment riches. Ces dernières sont généralement constituées de lexiques (recensant les formes connues de noms propres) et de règles pour reconnaître des entités nommées (spécifiées par des grammaires). Entre autres, les transducteurs sont des expressions régulières permettant d'insérer dans un

texte des balises qui délimitent et catégorisent les EN. Alors, le défi est d'énumérer les transducteurs afin d'obtenir un bon rappel (être aussi couvrant que possible) tout en restant précis (ne pas générer de faux positifs) à travers les catégories d'EN.

Cet article focalise sur la découverte de transducteurs à partir d'un corpus (d'apprentissage) annoté en entités nommées, pour enrichir notre base de règles. Plus précisément, notre première contribution est une formalisation de l'extraction de règles de transduction basées sur des motifs morpho-syntaxiques. Ces derniers s'appuient sur une analyse linguistique de surface du corpus (étiquetage grammatical et lemmatisation), ce qui permet de découvrir des règles par généralisation. Pour éviter de générer trop de règles, nous ne retenons que les règles de transduction *informatives*. Ensuite, l'objectif est d'exploiter cette base de règles afin d'évaluer les annotations EN qu'elles peuvent produire. Dans ce contexte, notre contribution est la proposition d'un algorithme qui effectue une annotation couvrante à partir de l'intégralité des règles (dont celles qui ne détectent qu'une borne d'EN) meilleure qu'une annotation par application des règles consistantes (qui détectent les deux bornes d'une catégorie d'EN), selon leur confiance.

La section 2 situe notre approche par rapport à la littérature. Nous présentons à la section 3 le formalisme qui définit la notion de règle morpho-syntaxique de transduction informative. Dans la section 4, nous présentons deux stratégies qui exploitent la base de règles pour réaliser une annotation. Puis nous présentons des résultats chiffrés pour l'extraction des règles et pour l'évaluation des annotations produites en section 5.

2 Etat de l'art

La théorie des EN et la tâche de REN sont présentées extensivement dans Ehrmann (2008) ou Nadeau (2007). L'application de méthodes d'apprentissage automatiques à la REN a été expérimentée à l'aide de modèles à maximum d'entropie (Borthwick et al., 1998), de HMM (Favre et al., 2005) de CRF (McCallum, 2003; Zidouni et al., 2009), de SVM (Bunescu et Pasca, 2006; Nadeau, 2007). Ces modèles sont assez performants et assez peu coûteux à élaborer, à partir du moment où l'on dispose de corpus d'apprentissage. Néanmoins, ils permettent difficilement d'enrichir une base de connaissances.

Les approches en fouille de donnée sont originellement axées sur l'extraction de motifs fréquents (Mannila et Toivonen, 1997) et séquentiels (Agrawal et Srikant, 1995) au sein d'une base de données. Nous nous inspirons également des approches à base de chaînes (Fischer et al., 2005). L'extraction de motifs morpho-syntaxique est exploitée par Cellier et Charnois (2010) pour la recherche d'expressions qualificatives. De manière générale et théorique, la relation de la fouille de textes avec les grammaires, automates et expression régulière est présentée dans Mendes et Antunes (2009) et Parekh et Honavar (2000).

La fouille de données pour la REN a notamment été expérimentée par Plantevit et al. (2009), qui appliquent la recherche de séquences fréquentes à la détection d'entités dans le domaine biomédical, ainsi que par Budi et Bressan (2007) qui recherchent ce type d'entité (en indonésien) par extraction de règles d'association. Nous nous inscrivons dans cette approche, tout en autorisant la généralisation des motifs (grâce à une hiérarchie morpho-syntaxique) et en extrayant de surcroît des règles partielles qui peuvent détecter une ou plusieurs bornes d'EN.

3 Règles morpho-syntaxiques de transduction

3.1 Du langage du corpus au langage morpho-syntaxique

Langages de motifs Dans le but d'utiliser des méthodes de fouille de données pour le traitement du langage, nous représentons le jeu de données comme ensemble de transactions, chaque transaction correspondant à une phrase. Ceci nous permet de ne pas extraire de motifs qui chevaucheraient plusieurs phrases, ce qui a peu de sens d'un point de vue linguistique.

Nous considérons des motifs fondés sur la notion de chaîne (Fischer et al., 2005). Soit un alphabet $\mathcal{A}$, le langage $\mathcal{L}_{\mathcal{A}}$ désigne l'ensemble des chaînes $a_1 a_2 \dots a_n$ où $a_i \in \mathcal{A}$ pour tout $i \in \{1, \dots, n\}$. Dans la suite, nous utilisons deux alphabets : celui des token (ou mots) du corpus $\mathcal{W}$ et celui des marques $\mathcal{M}$. Ainsi, le corpus annoté, dénoté par $\mathcal{D}$, correspond à un multi-ensemble de motifs de $\mathcal{L}_{\mathcal{I}}$ où $\mathcal{I} = \mathcal{W} \cup \mathcal{M}$. Le tableau 1 donne un exemple d'un tel corpus où $\mathcal{W} = \{\text{Le}, \text{nouveau}, \text{président}, \dots\}$ et $\mathcal{M} = \{\texttt{<pers>}, \texttt{</pers>}, \texttt{<loc>}, \dots\}$. Dans cette section, nous ne formulons pas de contrainte préalable sur la bonne formation des marques : cette problématique sera abordée dans la section 4 dans le cadre de l'annotation.

$\mathcal{D}$

Trans.	Motifs de $\mathcal{L}_{\mathcal{I}}$
t_1	Le nouveau <pers> président Barack Obama </pers> est arrivé à <loc> Moscou </loc>.
t_2	Il y a vu l'ancienne <pers> chancelière Michelle Bachelet </pers>.
t_3	Le <pers> président Dimitri Medvedev </pers> n'était pas sur la belle <loc> place Vladimir Lenine </loc>.

TAB. 1 – *Exemple jouet de corpus annoté*

Afin de prendre en compte les catégories morpho-syntaxiques, nous souhaitons extraire des motifs fondés sur une extension du langage $\mathcal{L}_{\mathcal{W}}$. Plus précisément, l'alphabet $\mathcal{W}^*$ complète les mots $\mathcal{W}$ avec un ensemble de descripteurs correspondant à des catégories (et des sous-catégories) morpho-syntaxiques et à des lemmes. Cet alphabet $\mathcal{W}^*$ est muni d'une hiérarchie $\leq$ i.e., une relation d'ordre partiel où pour tout $x \leq w$ et $y \leq w$, alors $x \leq y$ ou $y \leq x$. Ainsi, $w \leq w'$ signifie que w est plus général que w'. Typiquement, avec le corpus du tableau 1, on a $\mathcal{W}^* = \mathcal{W} \cup \{\text{DET}, \text{ART}, \text{ADJ}, \text{NOM}, \dots\}$ où DET $<$ ART $<$ le. Au final, le *langage morpho-syntaxique* $\mathcal{L}_{\mathcal{I}^*}$ correspond à toutes les chaînes reposant sur les items avec hiérarchies $\mathcal{W}^*$ et des marques $\mathcal{M}$ i.e., $\mathcal{I}^* = \mathcal{W}^* \cup \mathcal{M}$.

Spécialisation et support Intuitivement, le motif « <pers> NOM » est plus général que le motif « <pers> NOM NAM » car il est plus court (où « NOM » correspond à un nom commun et « NAM » à ce qui est détecté par l'étiqueteur comme un nom propre). De même, le motif « <pers> NOM » est plus général que le motif « <pers> président » car la catégorie « NOM » est plus générale que le mot « président ». Afin de formaliser cette notion de spécialisation, nous introduisons une forme standard pour comparer les motifs morpho-syntaxiques entre eux. Pour cela, nous définissons la *marque vide* $\emptyset_{\mathcal{M}}$

qui est incluse dans n'importe quel ensemble de marques i.e., $\emptyset_{\mathcal{M}} \subseteq M$ où $M \in \mathcal{L}_{\mathcal{M}}$. Nous pouvons alors définir la notion de chaîne alternée :

Définition 1 (Chaîne alternée) *La chaîne alternée $B_0 w_1 B_1 w_2 \dots B_{n-1} w_n B_n$ du motif morpho-syntaxique $S \in \mathcal{L}_{\mathcal{I}^*}$ est l'unique décomposition de S telle que :*
- *$w_i \in \mathcal{W}^*$ pour tout $i \in \{1, \dots, n\}$,*
- *$B_i \in \mathcal{L}_{\mathcal{M}} \cup \emptyset_{\mathcal{M}}$ pour tout $i \in \{0, \dots, n\}$ et*
- *la chaîne $B_0 w_1 B_1 w_2 \dots B_{n-1} w_n B_n$ est égale à S si on omet les $B_i = \emptyset_{\mathcal{M}}$.*

Par exemple, les chaînes alternées de « <pers> NOM » et « <pers> NOM NAM » sont respectivement « <pers> NOM $\emptyset_{\mathcal{M}}$ » et « <pers> NOM $\emptyset_{\mathcal{M}}$ NAM $\emptyset_{\mathcal{M}}$ » . Nous pouvons comparer ces deux chaînes alternées de la manière suivante :

Définition 2 (Spécialisation) *Un motif morpho-syntaxique S est plus général qu'un motif morpho-syntaxique S', noté $S \preceq S'$, ssi il existe $k \in \{0, .., m-n\}$ tel que $w_i \leq w'_{i+k}$ pour tout $i \in \{1, \dots, n\}$ et $B_i \subseteq B'_{i+k}$ pour tout $i \in \{0, \dots, n\}$ où $B_0 w_1 B_1 w_2 \dots B_{n-1} w_n B_n$ est la chaîne alternée de S et $B'_0 w'_1 B_1 w'_2 \dots B'_{m-1} w'_m B'_m$ est la chaîne alternée de S'.*

La définition 2 stipule qu'un motif S est plus général qu'un motif S' lorsque chaque mot et chaque ensemble de marques de la chaîne alternée de S sont plus généraux que ceux d'une sous-chaîne alternée de S'. Nous pouvons vérifier que « <pers> NOM $\preceq$ <pers> NOM NAM » et « <pers> NOM $\preceq$ <pers> président ».

Comme le langage $\mathcal{L}_{\mathcal{I}^*}$ englobe tous les langages $\mathcal{L}_{\mathcal{W}}$, $\mathcal{L}_{\mathcal{M}}$, $\mathcal{L}_{\mathcal{I}}$ et $\mathcal{L}_{\mathcal{W}^*}$, la relation de spécialisation $\preceq$ définie sur $\mathcal{L}_{\mathcal{I}^*}$ est utilisée pour comparer entre eux n'importe quels motifs de ces langages. Nous pouvons ainsi compter le nombre de fois où un motif morpho-syntaxique est observé dans une phrase du corpus :

Définition 3 (Support avec marques) *Le support (avec marques) d'un motif $S \in \mathcal{L}_{\mathcal{I}^*}$ dans un jeu de données $\mathcal{D}$, dénoté par $Supp(S, \mathcal{D})$, est son nombre d'occurrences.*

Par exemple, les motifs « <pers> NOM » et « <pers> président » possèdent respectivement un support de 3 et de 2 dans le jeu de données $\mathcal{D}$. Le support des motifs décroît suivant leur spécialisation, i.e. pour tout motif $S \preceq S'$, on a $Supp(S, \mathcal{D}) \geq Supp(S', \mathcal{D})$. Cette propriété est utilisée pour extraire les motifs morpho-syntaxiques fréquents en utilisant un algorithme par niveaux (Mannila et Toivonen, 1997).

3.2 Règles de transduction

Une règle morpho-syntaxique de transduction est un motif qui contient à la fois au moins un élément de $\mathcal{W}^*$ et au moins une marque de $\mathcal{M}$:

Définition 4 (Règle morpho-syntaxique de transduction) *Une règle morpho-syntaxique de transduction est un motif de $\mathcal{R} = \mathcal{L}_{\mathcal{I}^*} \backslash (\mathcal{L}_{\mathcal{W}^*} \cup \mathcal{L}_{\mathcal{M}})$.*

Par exemple, le motif « <pers> président » est une règle de transduction : l'utilisation de cette règle dans un corpus non-annoté permettra d'ajouter la marque <pers> lorsque

le mot « président » sera observé (cf. la section 4). Plus généralement, la règle de transduction r peut s'appliquer là où les tokens (ou les catégories) sont observé(e)s dans le corpus. Cette partie de la règle est appelée le *motif sans marques*. Le motif sans marques de $S \in \mathcal{L}_{\mathcal{I}^*}$, dénoté par $\widetilde{S}$, est le motif le plus spécifique de $\mathcal{L}_{\mathcal{W}^*}$ qui soit plus général que S : $\widetilde{S} = \max_{\preceq}\{W \in \mathcal{L}_{\mathcal{W}^*} | W \preceq S\}$. Cela correspond au motif $\widetilde{S}$ dont on a supprimé les marques, pour lequel nous définissons le *support sans marques* :

Définition 5 (Support sans marques) *Le support sans marques d'un motif $S \in \mathcal{L}_{\mathcal{I}^*}$ dans un jeu de données $\mathcal{D}$ est $\widetilde{Supp}(S, \mathcal{D}) = Supp(\widetilde{S}, \widetilde{\mathcal{D}})$ où $\widetilde{\mathcal{D}} = \{\widetilde{d} | d \in \mathcal{D}\}$.*

Par exemple, la règle « `<pers>` NOM » a un support sans marques de 4 : c'est le nombre d'occurrences du motif « NOM » (antécédent) dont celle qui se trouve dans t_3 (mais qui ne contient pas le conséquent). Ces deux notions combinées permettent d'estimer la fiabilité de la règle par sa *confiance*, la probabilité conditionnelle d'observer les marques au sein d'une phrase lorsqu'elle contient le motif morpho-syntaxique :

Définition 6 (Confiance) *La confiance d'une règle morpho-syntaxique $r \in \mathcal{R}$ est :*

$$Conf(r, \mathcal{D}) = \frac{Supp(r, \mathcal{D})}{\widetilde{Supp}(r, \mathcal{D})}$$

Par exemple, la confiance des règles « `<pers>` NOM » et « `<pers>` président » sont respectivement de 3/4 et 2/2. La section 5 reviendra sur le rôle du support et de la confiance pour contrôler respectivement le rappel et la précision de notre approche.

3.3 Règles de transduction informatives

Même en fixant des seuils de support et confiance sélectifs, les règles morpho-syntaxiques de transduction peuvent être trop nombreuses. Par exemple, 39 règles résultent de l'extraction des règles de support supérieur à 2 et de confiance supérieure à 2/3 avec le corpus du tableau 1. Cette profusion de règles découle des combinaisons possibles au travers de la hiérarchie. Le tableau 2 (a) illustre cette problématique en détaillant un ensemble de règles qui diffèrent les unes des autres par l'ajout d'un item.

	$\mathcal{R}$	$\mathcal{IR}$
r_1	`<pers>` président	×
r_2	`<pers>` président NAM	
r_3	`<pers>` président NAM NAM	
r_4	`<pers>` président NAM NAM `</pers>`	×

(a) Détail d'un groupe avec $Supp = 2$ et $Conf = 1$

Supp.	Conf.	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{IR}\rvert$
2	1	5	2
3	3/4	6	3
2	2/3	24	2
2	2/3	4	1
	Total :	39	8

(b) Règles regroupées

TAB. 2 – *Impact des règles de transduction informatives*

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 Bastide et al. (2000).

Par exemple, la règle r_2 du tableau 2 n'apporte aucune information par rapport à r_1 car son motif sans marques n'est pas plus couvrant et son marquage n'est pas plus important. A l'inverse, la règle r_4 est informative car son application permet l'ajout d'une marque supplémentaire par rapport à r_1. Nous généralisons maintenant cette intuition avec la définition suivante :

Définition 7 (Règle morpho-syntaxique de transduction informative) *Une règle morpho-syntaxique de transduction r est informative ssi il n'existe pas de règle r' de même support et confiance telle que $\widetilde{r'} \prec \widetilde{r}$ et $|r'|_{\mathcal{M}} = |r|_{\mathcal{M}}$ ou $\widetilde{r'} = \widetilde{r}$ et $|r'|_{\mathcal{M}} > |r|_{\mathcal{M}}$.*

La définition 7 signifie qu'une règle est informative si aucune règle de même support et de même confiance ne possède (i) un motif sans marques plus général conduisant à un marquage similaire ou (ii) un motif sans marques identique conduisant à un marquage plus important. Par exemple, r_2 n'est pas informative car elle enfreint (i) à cause de r_1. La règle r_3 n'est pas informative car elle enfreint (ii) à cause de r_4. Le tableau 2 (b) illustre l'apport significatif des règles de transduction informatives sur notre exemple. On constate que l'ensemble des règles informatives, dénoté par $\mathcal{IR}$, se limite à 8 règles contre 39 avec la collection complète. Ce résultat s'observe à plus grande échelle avec les expérimentations sur les données réelles de la section 5.

4 Application des règles pour l'annotation

4.1 Critères pour une solution d'annotation

Une fois les règles de transduction extraites, nous les exploitons, comme règles d'annotation, en les appliquant sur des textes. Pour ce faire, nous nous basons sur la même représentation que lors de la phase d'extraction : un texte est segmenté en transactions (phrases), dont les items sont des tokens : c'est un multi-ensemble de $\mathcal{L}_{\mathcal{W}^*}$.

Pour appliquer les règles de transduction $\mathcal{R}$, diverses stratégies peuvent être mises en œuvre. L'extraction, exhaustive, nous fournit de nombreuses règles qui insèrent une ou plusieurs marques d'entités nommées. Pour notre problématique REN, l'objectif du marquage est d'obtenir une annotation, qui soit **consistante** : chaque marque (ou balise) ouvrante (suivie de tokens) doit nécessairement être suivie d'une marque fermante *de même catégorie* (e.g. après `<loc>`, la prochaine marque doît être `</loc>`).

Par ailleurs, nous souhaitons mettre l'accent sur deux critères de qualité :

- **couverture** : choisir l'annotation qui apporte un maximum d'information,
- **confiance / précision** : choisir l'annotation qui soit la plus probable possible.

Le premier critère peut-être comptabilisé par le nombre de marques introduites. Pour le second, nous considérons en première approximation que la vraisemblance d'une annotation correspond au produit des confiances des marques insérées.

4.2 Application des règles d'annotation

Règles complètes, par ordre de confiance Une première stratégie, intuitive, consiste à ne sélectionner que les règles qui créent une annotation consistante (avec une marque de début et de fin d'EN de même catégorie), appelées règles d'annotations

complètes, à les trier selon leur confiance, puis à les appliquer tant que la portion de texte qu'elles annotent n'a pas déjà été annotée. L'application ordonnée des règles nous conduit alors à une solution qui est toujours consistante et donne priorité à la confiance (par ordre d'application des règles) puis à la couverture (ajouter des annotations tant que c'est possible). Nous considérons cet algorithme comme notre "baseline".

Règles complètes et règles partielles Lorsque l'on considère toutes les règles pour effectuer l'annotation, alors la consistance n'est plus garantie : nous disposons de beaucoup de règles partielles, qui n'apportent qu'une partie d'information (début ou fin d'EN), mais dont l'annotation pourra être rendue consistante par un marquage ultérieur. Notre idée est donc d'insérer toutes les marques possibles sur une séquence donnée, en notant leur confiance, puis de sélectionner celles qui apporteront la meilleure annotation en termes de couverture, puis de confiance.

Algorithm 1 Sélection de l'annotation la plus couvrante puis la plus confiance

Entrée: une séquence $S = s_1 \dots s_n \in \mathcal{L}_{\mathcal{I}^*}$: des tokens $s_i \in \mathcal{L}_{\mathcal{W}^*}$ et des marques $s_i \in \mathcal{L}_{\mathcal{M}}$ dont nous connaissons le type $s_i.type = \{Ouvrante, Fermante\}$, la catégorie EN $s_i.catégorie$ et la confiance a priori $s_i.confiance$
Sortie: une séquence $S \in \mathcal{L}_{\mathcal{I}^*}$: qui maximise la couverture, puis la confiance

```
    // HypCons est l'hypothese consistante, HypPart sont les hypothèses partielles
    // L'opérateur ≫ compare deux hypothèses en termes de couverture et de confiance
  tant que s_i faire
    tant que s_i ∈ L_{W*} faire
      ∀Hypothèse ∈ HypCons ∪ HypPart : Hypothèse ← Hypothèse + s_i
    fait
    Marques ← ∅
    tant que s_i ∈ L_M faire
      si Marques[s_i.type][s_i.catégorie].confiance < s_i.confiance alors
        Marques[s_i.type][s_i.catégorie] ← s_i
      fin si
    fait
    pour chaque Marque ∈ Marques[Fermante] telle que
    HypPart[Marque.catégorie] + Marque ≫ HypCons faire
      HypCons ← HypPart[Marque.catégorie] + Marque
    fin pour
    pour chaque Marque ∈ Marques[Ouvrante] telle que
    HypCons + Marque ≫ HypPart[Marque.catégorie] faire
      HypPart[Marque.catégorie] ← HypCons + Marque
    fin pour
  fait
  renvoyer HypCons
```

Nous pouvons procéder par phrase (séquence). Malgré tout, l'espace de recherche pour une séquence donnée est exponentielle pour le nombre de marques insérées. Ceci peut-être problématique dans des contextes où les phrases sont difficilement délimitées, notamment à l'oral (dans notre cas, des émissions radio). Nous résolvons ceci grâce à un algorithme de programmation dynamique qui parcourt les solutions possibles en lar-

geur et ne maintient en mémoire que les hypothèses qui pourront devenir des solutions optimales. A cet effet, nous subordonnons le critère de confiance à celui de couverture : l'objectif est d'obtenir une annotation qui contienne le plus de marques possibles, tout en restant consistante. Avec cette contrainte, l'algorithme peut maintenir $n + 1$ hypothèses distinctes lorsqu'il y a n catégories distinctes à considérer : une hypothèse non-consistante ("ouverte") pour chaque catégorie et une hypothèse consistante.

Pour l'algorithme 1 présenté ci-dessus, nous considérons une séquence sur laquelle tous les transducteurs ont été appliqués sur toutes les portions de textes possibles : les marques sont toutes insérées et mélangées à la séquence. Nous introduisons l'opérateur $+$, qui permet d'ajouter un token ou une marque à une hypothèse, tout en mettant à jour son nombre de marques et sa confiance ; l'opérateur $\gg$ qui compare deux hypothèses et sélectionne celle qui dispose du plus grand nombre de marques, ou la plus confiante en cas d'égalité.

L'algorithme fonctionne par lots : il parcourt la séquence et accumule les tokens, jusqu'à tomber sur une ou plusieurs marque(s). Pour une série de marques rencontrées, il enregistre alors, pour chaque type (ouvrante / fermante) et pour chaque catégorie, la meilleure probabilité. Puis il considère chaque marque fermante : si elle permet de compléter une hypothèse ouverte (non-consistante) et si l'hypothèse (fermée) obtenue par ajout de cette marque est meilleure que l'hypothèse consistante courante il la met à jour. Puis il réalise la même opération pour chaque marque ouvrante, mettant alors à jour les hypothèse non-consistantes (ouvertes) s'il en trouve des meilleures.

5 Cas d'étude : reconnaissance d'entités nommées sur le corpus Ester2 et résultats expérimentaux

5.1 Préparation des données

Le corpus Ester2 a été constitué lors d'une campagne d'évaluation organisée par l'AFCP [1] et la DGA [2], qui a porté sur la transcription, la segmentation et l'extraction d'informations de flux de parole français radiodiffusés (Galliano et al., 2009). L'extraction d'information portait sur la REN dans les transcriptions de ces flux. Les EN détectées étaient à catégoriser en : personnes, lieux, organisations, produits, montants, temps ou fonctions. Nous disposons du corpus de référence, réalisé par des annotateurs humains. Notre système symbolique, CasEN, a participé à cette campagne, nous en avons analysé en détail les résultats (Nouvel et al., 2010).

Nous disposons également du corpus Eslo (Maurel et al., 2009), transcription de conversations lors d'enquêtes sociologiques menées dans la région d'Orléans. Celles-ci ont été annotées en EN selon les mêmes catégories que lors de la campagne Ester2. Pour une sous-partie du corpus, l'annotation a été réalisée manuellement (Eslo - man) et pour le reste, le corpus a été préannoté par un système REN, puis corrigé manuellement (Eslo - pre). Ce corpus comporte une proportion plus faible d'EN qu'Ester2.

Pour établir la hiérarchie des catégories morpho-syntaxiques, nous utilisons TreeTagger (Schmid, 1994) qui réalise un étiquetage robuste et rapide, et lemmatise les

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

mots (le lemme est la forme *normale*, sans déclinaisons). Cet étiqueteur donne des résultats supérieurs, en f-mesure, à 90 sur du texte écrit. L'analyse morpho-syntaxique du TreeTagger détermine le lemme pour chaque token (item), puis lui attribue une catégorie morpho-syntaxique au sein d'un ensemble d'arbres (forêt). Les noms propres sont distingués des noms communs (notamment à l'aide de lexiques) en tant que *NAM*. Comme nous ne souhaitons pas extraire de motifs reposant sur des éléments lexicaux, nous ne conservons pour ces items que la catégorie morpho-syntaxique.

5.2 Extraction des règles de transduction

Nous cherchons dans ces corpus des motifs selon la méthode exposée en section 3. L'extraction des règles informatives utilise une structure de données sous forme de trie (arbre de préfixes) et un algorithme par niveaux (qui extrait des motifs de taille croissante). Le groupement de motifs et leur sélection sous forme de règles est réalisée en fin de traitement. Le tableau 3 présente les caractéristiques de la fouille réalisée sur ces corpus, sur une machine cadencée à 2.4GHz disposant de 4Go de RAM.

Corpus	Tokens	EN	F.	C.	Motifs	Groupes	Règles	Temps
Ester2	40 167	2 798	10	0.5	2 270	975	1 119	0' 4"
			5	0.5	28 047	2 747	3 673	0' 5"
			3	0.3	458 875	7 448	12 653	0' 19"
Eslo - man	165 987	2 303	10	0.5	3 167	1 111	1 287	0' 27"
			7	0.3	51 339	2 424	3 682	0' 52"
Eslo - pre	894 564	14 402	30	0.5	36 236	3 672	4 887	5' 35"
			20	0.3	5 489 632	7 906	11 937	7' 10"

TAB. 3 – *Extraction sur les corpus, à seuils de fréquence (F.) et confiance (C.)*

Nous remarquons que lorsque nous diminuons le seuil de fréquence (sur Ester2, de 10 à 3), le nombre de motifs augmente très largement (multiplié par 200), cependant la méthode de groupement nous permet de conserver une base de règles de taille raisonnable (multipliée par 10). Par ailleurs, effectuer la fouille sur un corpus de référence plus large ne nous apporte par une grande quantité de nouveaux motifs : nous supposons que l'utilisation de motifs morpho-syntaxiques produit effectivement des règles de transduction généralisées qui capturent la redondance au sein du corpus.

5.3 Annotation à partir des règles extraites

Pour tester nos algorithmes d'annotation, nous utilisons le corpus Ester2, pour lequel nous disposons d'un outil d'évaluation qui fournit le SER (Makhoul et al., 1999) (nombre d'erreurs par entité, à minimiser), la précision, le rappel et la f-mesure (calculés sur les catégories des tokens). L'évaluation est réalisée sur douze fichiers, des enregistrements d'approximativement même durée. Nous effectuons une validation croisée à douze plis : l'extraction est réalisée sur onze fichiers, l'application des règles d'annotation et l'évaluation sont effectuées sur le douzième fichier, ceci douze fois de suite. En premier lieu, nous faisons cette expérimentation en annotant avec les règles complètes à divers seuil de fréquence et de confiance, afin de déterminer les paramètre optimaux.

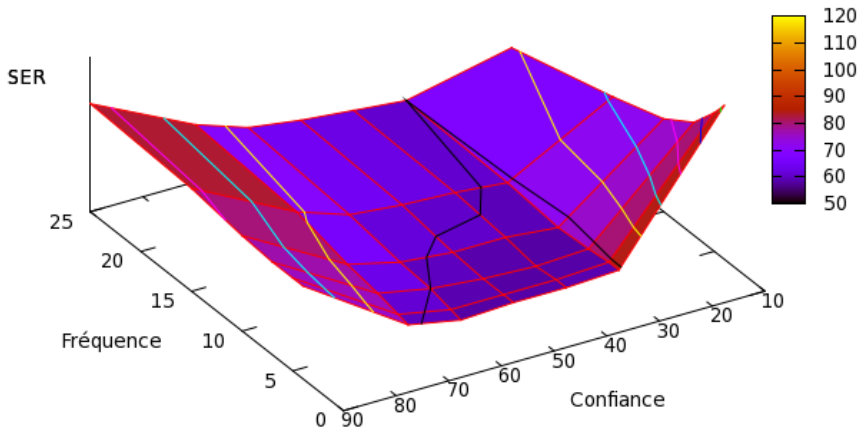

FIG. 1 – *Evaluation sur Ester2, Règles complètes*

La figure 1 présente les résultats en SER selon la fréquence et la confiance. De manière générale, lorsque l'on diminue le seuil de fréquence, la performance s'améliore. Nous voyons qu'un optimum est atteint pour un seuil de fréquence de 3 et un seuil de confiance compris entre 40% et 60%. Notons également que plus la fréquence est basse, plus il faut augmenter la confiance pour atteindre l'optimum local. Ces résultats montrent qu'une annotation peut être réalisée à partir de motifs morpho-syntaxiques extraits d'un corpus et que certaines règles de transduction sont de bonne qualité.

Paramètres		**Règles complètes**				**Règles partielles**			
Freq.	**Conf.**	**P..**	**R.**	**F.**	**SER**	**P..**	**R.**	**F.**	**SER**
3	40	67,66	50,19	0,58	55,14	62,91	57,29	0,6	52,96
3	45	70,66	48,37	0,57	**54,82**	65,91	54,76	0,6	**51,99**
3	60	76,61	45,52	0,57	55,66	72,53	50,62	0,6	53,66
5	40	69,63	45,87	0,55	56,75	65,41	52,08	0,58	54,09
5	45	72,01	44,39	0,55	56,91	67,52	49,91	0,57	54,45
5	60	78,9	41,49	0,54	57,72	76,73	45,05	0,57	56,42
9	40	73,9	40,89	0,53	57,78	69,78	46,41	0,56	54,9
9	45	76,14	40,4	0,53	57,58	73,62	44,92	0,56	55,42
9	60	81,49	37,41	0,51	60,92	80,04	40,72	0,54	58,62

TAB. 4 – *Evaluation sur Ester2, (Précision P., Rappel R., F-mesure F.)*

Le tableau 4 présente les résultats pour l'application des règles complètes et partielles. Le SER et la F-mesure sont proches, avec une différence relativement constante en faveur de la version utilisant les règles partielles. La précision est meilleure avec les règles complètes : consistantes, elles introduisent moins de faux positifs. Les règles partielles donnent un meilleur rappel : de nombreuses EN, non-détectées par des règles consistantes peuvent l'être par combinaison de règles partielles. Le fait de rechercher séparément des annotations de début ou de fin d'EN a du sens, il devient envisageable

de réaliser une extraction de motifs morpho-syntaxique qui ne repère pas systématiquement une EN entière, mais uniquement sa frontière gauche ou droite.

Si notre évaluation vise à déterminer si les règles extraites pourront potentiellement venir enrichir un système symbolique (et sous quel format), nous remarquons cependant que les performances sont assez éloignées des meilleurs systèmes de REN. Les systèmes évalués lors de la campagne Ester2, se situent en SER de 9 à 37 (f-mesure de 94 à 65). La comparaison est cependant difficile : notre système se veut minimal et utilise très peu de ressources et de traitements linguistiques (lexique, analyse syntaxique, parenthésage...). Nous cherchons avant tout à déterminer comment extraire des règles (grammaires) pertinentes pour la reconnaissance d'entités nommées.

6 Conclusion et perspectives

Cet article présente une approche en reconnaissance d'entités nommées, qui consiste à découvrir par fouille de données des motifs morpho-syntaxiques fortement corrélés aux annotations les délimitant, dont nous ne conservons que les plus informatifs sous forme de règles. Cet apprentissage nous permet d'extraire des *règles de transduction informatives*, que nous sommes en mesure d'appliquer à d'autres textes.

Nous apportons quelques éléments à travers cette recherche à des problématiques liées à l'extraction de connaissances et à la reconnaissance d'entités nommées. Nous montrons qu'il est possible d'extraire des motifs morpho-syntaxique sous forme de règles, dont la qualité semble suffisamment bonne pour venir enrichir un système symbolique. Par ailleurs, nous montrons que la tâche d'annotation peut-être considéré comme une recherche indépendante des bornes gauches ou droites d'une entité.

Ces résultats nous encouragent à poursuivre notre travail dans plusieurs directions. D'une part, afin d'enrichir la base de connaissances, nous souhaitons obtenir des règles plus denses (avec des alternatives et éléments optionnels), qui pourraient être plus couvrantes, en restant précises. D'autre part, la stratégie d'annotation à partir de règles partielles faisant de nombreuses simplifications, nous chercherons à mieux tirer parti de l'adéquation entre la séquence observée et la base de règles pour réaliser l'annotation. Enfin, nous envisageons d'expérimenter cette approche à d'autres langues.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *9th International Conference on Data Engineering (ICDE'95)*, pp. 3–14.

Bastide, Y., N. Pasquier, R. Taouil, G. Stumme, et L. Lakhal (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *Comp. Logic*, pp. 972–986.

Borthwick, A., J. Sterling, E. Agichtein, et R. Grishman (1998). Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Work. on Very Large Corpora*.

Budi, I. et S. Bressan (2007). Application of association rules mining to named entity recognition and co-reference resolution for the indonesian language. In *IJBIDM'07*, Volume 2.

Bunescu, R. C. et M. Pasca (2006). Using encyclopedic knowledge for named entity disambiguation. In *Conference of the European Chapter of the Ass. for Comp. Ling. (EACL'06)*.

Cellier, P. et T. Charnois (2010). Fouille de données séquentielles d'itemsets pour l'apprentissage de patrons linguistiques. In *Traitement Automatique du Langage Naturel (TALN'10)*.

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

Favre, B., F. Béchet, et P. Nocera (2005). Robust named entity extraction from large spoken archives. In *HLT/EMNLP'05*.

Fischer, J., V. Heun, et S. Kramer (2005). Fast frequent string mining using suffix arrays. In *5th IEEE International Conference on Data Mining (ICDM'05)*, pp. 609–612.

Galliano, S., G. Gravier, et L. Chaubard (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *INTERSPEECH'09*.

Makhoul, J., F. Kubala, R. Schwartz, et R. Weischedel (1999). Performance measures for information extraction. In *DARPA Broadcast News Workshop*, pp. 249–252.

Mannila, H. et H. Toivonen (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery 1*(3), 241–258.

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

McCallum, A. (2003). Efficiently inducing features of conditional random fields. In *Conference on Uncertainty in Artificial Intelligence (UAI'03)*, pp. 403–410.

Mendes, A. C. et C. Antunes (2009). Pattern mining with natural language processing : An exploratory approach. In *MDLM'09*, pp. 266–279.

Nadeau, D. (2007). *Semi-Supervised Named Entity Recognition : Learning to Recognize 100 Entity Types with Little Supervision*. Ph. D. thesis, University of Ottawa, Canada.

Nouvel, D., J.-Y. Antoine, N. Friburger, et D. Maurel (2010). An analysis of the performances of the casen named entities recognition system in the ester2 evaluation campaign. In *International Language Resources and Evaluation (LREC'2010)*.

Parekh, R. et V. Honavar (2000). *Grammar Inference, Automata Induction, and Language Acquisition*, Chapter 29, pp. 727–764.

Plantevit, M., T. Charnois, J. Klema, C. Rigotti, et B. Cremilleux (2009). 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 1*, 119–148.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference in New Methods for Language Processing (NEMLP'94)*, pp. 44–49.

Zidouni, A., H. Glotin, et M. Quafafou (2009). Recherche d'entités nommées dans les journaux radiophoniques par contextes hiérarchique et syntaxique. In *CORIA'09*, pp. 421–432.

Summary

Recognizing named entities is a task that is mainly processed by systems that are specified using rules or that are learned. In this paper, we introduce an approach aiming at extracting symbolic and discriminative rules that may be reviewed by humans. We are given a reference corpus, from which we extract informative transducer rules. Then an algorithm searches covering and probable solutions for annotating. We report experimental results and discuss advantages and perspectives of this approach.