Recherche d'Entités Nommées dans les Journaux Radiophoniques par Contextes Hiérarchique et Syntaxique

Azeddine Zidouni*, Hervé Glotin, Mohamed Quafafou***

** Laboratoire LSIS, Univ. Aix-Marseille 2.*
{azeddine.zidouni,mohamed.quafafou}@univmed.fr

*** Laboratoire LSIS, Univ. Sud Toulon Var.*
glotin@univ-tln.fr

RÉSUMÉ. *Ce papier présente une approche pour la recherche d'entités nommées dans des transcriptions radiophoniques. Nous allons utiliser les structures des entités nommées afin d'améliorer le taux de leur reconnaissance. En effet, l'espace des entités peut être représenté par une structure hiérarchique (arbre). Ainsi, un concept peut être vu comme un noeud dans l'arbre, et une entité comme un parcours dans la structure de l'espace. Nous allons montrer l'apport de cette représentation en utilisant le modèle des Champs Aléatoires Conditionnels (CAC). La comparaison de notre approche avec la méthode des Modèles de Markov Cachés (MMC) montre une amélioration de la reconnaissance en utilisant les CAC Combinés. Nous montrons également l'impact de l'utilisation des informations a priori dans le processus en incluant les informations syntaxiques des transcriptions comme nouveau contexte.*

ABSTRACT. *This paper focuses on the role of structures in the Named Entity retrieval inside audio transcription. We consider the transcription documents strucure that guides the parsing process, and from which we deduce an optimal hierarchical structure of the space of concepts. Therefore, a concept is represented by a node or any sub-path in this hierarchy. We show the interest of such structure in the recognition of the Named-Entities using the Conditional Random Fields (CRF). The comparison of our approach to the Hiden Markov Model (HMM) method shows an important improvement of recognition using Combining CRFs. We also show the impact of the Part-of-Speech tagging (POS tagging) in the prediction quality.*

MOTS-CLÉS : *Reconnaissance d'entités nommées, Champs conditionnels aléatoires, Structures hierarchiques, Recherche d'information.*

KEYWORDS: *Named Entity Research, CRFs, Hierarchical Structure, Information Retrieval.*

1. Introduction

Les langues naturelles sont vivantes, elles traduisent une certaine évolution des pensées. Cette liaison entre la pensée et la langue rend le Traitement Automatique de la Langue Naturelle (TALN) très complexe. La langue offre de nombreuses façons d'exprimer une idée et une multitude de formes pour la représenter, plusieurs sens peuvent être associés à un mot selon son contexte d'utilisation. Les modèles de recherche d'information basés sur la sémantique tentent de représenter le fond de l'information de façon indépendante de la forme (ex. la syntaxe dans le texte). Cette tâche est d'autant plus complexe qu'il n'existe pas de bijection entre les mots et les sens associés. Les *Entités Nommées* (EN) sont des entités du monde réel, dont la forme linguistique est une représentation directe qui varie selon son contexte. L'extraction des EN dans un contexte journalistique représente une tâche importante dans la chaîne de l'analyse sémantique. Une EN représente une description conceptuelle qui fait référence à un objet[1] dont la représentation linguistique est unique. Dans le domaine de l'indexation sémantique, la représentation conceptuelle d'un objet prime sur sa représentation linguistique. En effet, avoir la description d'un objet dans une phrase nous permettra d'identifier plus facilement les objets porteurs de sens dans son contexte.

Dans les traitements textuels, on distingue deux types d'EN : les entités singulières, et les entités composées. Les entités singulières représentent généralement des noms propres, la présence de majuscules est un bon indicateur pour ce type. Les entités composées : le séquencement d'un ensemble de mots, insignifiants à l'origine, peut faire référence à un concept particulier (organismes, dates, etc.). Cette dernière rend la tâche encore plus difficile, car en plus du problème de la reconnaissance d'EN, s'ajoute le problème de la segmentation. En effet, on cherche à attribuer à chaque mot une étiquette signifiant l'existence d'une EN. Or, une étiquette peut s'étendre sur plusieurs mots. Plusieurs approches sont proposées afin de palier cette contrainte (modèles markoviens). La particularité du traitement des EN dans le cadre radiophonique réside dans la spontaniété de la parole, les entités différant d'une source à une autre. En effet, chaque source (radio) implique différents intervenants avec des origines différentes (des vocabulaires différents). Ces différences rendent la tâche de création de modèles génériques de prédiction complexe. En outre, les EN possèdent un cycle de vie (elles ont une forte apparition pendant une durée puis elles disparaissent). D'où la nécessité de construire des modèles qui se basent sur le contexte d'apparition des EN.

L'article est structuré comme suit. La section 2 décrit le problème de la *Recherche d'Entités Nommées* (REN) en utilisant les modèles graphiques. Dans la section 3, nous allons définir la structure d'arbre des EN extraite à partir du corpus de données, et expliquer le principe de la méthode utilisée pour la REN. La section 4 présente les résultats expérimentaux obtenus sur le corpus d'ESTER (Sylvain Galliano *et al.,* n.d.) en considérant les précisions des prédictions ainsi que le temps pris pour la construction des modèles. Nous allons conclure en section 6.

1. Une entité du monde réel qui est représenté par un ou plusieurs mots qui portent un sens particulier.

2. Etiquetage de données séquentielles

Un modèle séquentiel est un modèle graphique dans lequel les liaisons entre certaines variables sont spécifiques. Les modèles graphiques nous permettent d'attribuer des classes prédéfinies aux variables (problème de classification), mais le vrai point fort de ces approches est leur capacité de modéliser plusieurs variables interdépendantes (Sutton *et al.,* 2006). Les modèles de séquences supposent avoir les variables d'entrées X sous une forme séquentielle. Le problème d'étiquetage de séquences peut être formalisé de la manière suivante : Etant donnée $X = \langle x_1, x_2, \ldots x_n \rangle$ une séquence de données d'observation (données d'entrée), trouver la séquence d'états $Y = \langle y_1, y_2, \ldots y_n \rangle$ (séquence d'étiquettes associées aux données d'entrée) qui maximise la probabilité conditionnelle $P(Y|X)$:

$$Y = argmax P(Y|X). \quad [1]$$

Dans ce qui suit, nous donnons les différents modèles markoviens et les fondements théoriques des *Champs Aléatoires Conditionnels* (CAC).

2.1. *Les Modèles de Markov Cachés*

Dans un problème d'étiquetage de données séquentielles, comme la reconnaissance des EN, une première approche consiste à attribuer à chaque mot de la séquence, indépendamment des autres, une étiquette de l'ensemble Y. Ce type de traitement suppose que toutes les variables de sortie sont indépendantes (étiquettes). Ainsi, les EN de deux mots voisins sont indépendantes. Cela peut engendrer quelques lacunes comme par exemple : si on attribue au mot $Paris$ l'étiquette $location$, dans le cas où il est suivi du mot $match$ (paris match), l'étiquette sera $organisation$. Pour pallier à cet inconvénient, les *Modèles de Markov Cachés* (MMC) considèrent les variables de sortie comme une chaîne linéaire. En effet, un MMC modélise une séquence d'observations $X = \{x_t\}_{t=1}^{T}$ en supposant qu'il existe une séquence d'états $Y = \{y_t\}_{t=1}^{T}$ formée à partir des états finis des variables de sortie. Pour calculer la probabilité jointe $P(x, y)$, un MMC considère deux indépendances : (a) Chaque état y_t dépend uniquement de son état prédécesseur y_{t-1}, il est indépendant de tous les états $y_1, y_2 \ldots y_{t-2}$ (certaines approches considèrent un ensemble fini de prédécesseurs), (b) chaque observation x_t dépend uniquement de son état courant y_t. De cette manière, la probabilité jointe de la séquence d'états Y et de la séquence d'observation X est factorisée comme suit :

$$P(x, y) = p(y_1)p(x_1|y_1) \prod_{t=2}^{T} p(y_t \mid y_{t-1})p(x_t \mid y_t), \quad [2]$$

avec comme distribution de transition d'états $p(y_t \mid y_{t-1})$, $p(x_t \mid y_t)$ est la distribution d'observation, $p(y_1)$ est la distribution de l'état initial.

Les MMC sont des modèles génératifs, qui assignent une probabilité jointe à la paire de séquences (observation, variable de sortie). Pour définir la probabilité jointe, les modèles génératifs énumèrent toutes les séquences d'observation possibles. Pour cela, elles doivent construire des modèles qui contiennent beaucoup de dépendances caractéristiques[2] entre les variables d'observation. Or, le problème d'inférence dans ce type de modèles difficilement traitable (vu le nombre de dépendances dans les problèmes de TALN, les approches existantes appliquent des approximations). D'où l'apparition des modèles discriminatifs.

2.2. *Les Champs Markoviens Aléatoires*

Dans un Champ Markovien Aléatoire (CMA), le système de voisinage est déterminé uniquement par un ensemble d'arêtes dans le graphe.

Définition 2.2.1 *Soit $G = (V, E)$ un graphe, avec V l'ensemble des sommets, et E l'ensemble des arêtes. Le système de voisinage dans G est déterminé uniquement par un ensemble d'arêtes de E ; le voisinage $N(v)$ du noeud $v \in V$ est définit par $N(v) = \{u \in V | < v, u > \in E\}$. Soit $X_V = \langle X_v | v \in V \rangle$ un vecteur aléatoire avec X_v est une variable aléatoire associée au sommet v. La distribution de probabilité strictement positive $P(x_v) = P(X_v = x_v)$ est dite champ aléatoire.*

Une distribution de probabilités P dans un graphe G est dite Champ Markovien Aléatoire si pour toute configuration x_v et pour tout sommet $v \in V$, on a :

$$P(x_v | x_{V-v}) = P(x_v | x_{N(v)}). \quad [3]$$

La probabilité conditionnelle de x_v connaissant toutes les autres probabilités des sommets du graphe (x_{V-v}) n'est rien d'autre que la probabilité du même sommet connaissant les probabilités de ses voisins $x_{N(v)}$. Donc la probabilité d'un état v ne dépend que des probabilité de son voisinage $N(v)$. Cette méthode de calcul qui ne se base que sur le voisinage peut s'avérer intéressante pour beaucoup de problèmes. Cependant, elle présente quelques lacunes dans les tâches séquentielles. L'exemple de la figure 1, présenté dans (McCallum *et al.,* 2003b), illuste un modèle à états finis conçu pour différencier les deux séquences rib et rob. En supposant que la séquence d'observation est rib, la première observation r coïncide avec deux états (1 et 4). Ainsi les probabilités sont identiques pour les deux transitions ($P = 0.50$). Etant donné que les transitions sont conditionnées par les observations, les états n'ayant qu'un seul état suivant ignorent l'observation (les états 1 et 4). Ainsi les deux séquences ri et ro seront équivalentes, indépendamment de la séquence observée. De même pour rib et rob. D'autant si le mot rob est plus fréquent, il sera favorisé par rapport à rib, et donc ce dernier ne sera jamais reconnu.

2. Par exemple, la dépendance entre la capitalisation des mots et leurs suffixes

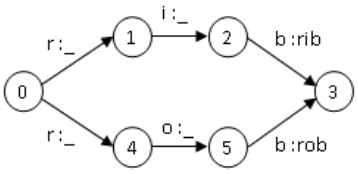

Figure 1. *Graphe de séquences illustrant l'effet de biais : les séquences RIB et ROB sont équivalentes car elles sont conditionnées seulement par les observations.*

Nous allons exposer dans la section suivante l'approche des CAC qui prend en compte cet effet de biais.

2.3. *Les Champs Aléatoires Conditionnels Séquentiels*

La puissance des modèles graphiques réside dans leur capacité à modéliser plusieurs variables indépendantes. L'étiquetage textuel est un cas particulier des modèles graphiques où les variables d'entrées sont présentées sous une forme séquentielle (séquence de mots). Pour cette raison, nous allons étudier un cas particulier des CAC qui est le CAC-Séquentiel (*Linear-Chain CRFs*). Les Champs Aléatoires Conditionnels Séquentiels (*Linear-Chain Conditional Random Fields*) (Lafferty *et al.,* 2001) est un cadre probabiliste discriminant utilisé pour la segmentation et l'étiquetage des données séquentielles (une segmentation peut être vue comme un étiquetage). L'avantage que présente CAC par rapport aux modèles markoviens classiques est la prise en compte du problème du biais des étiquettes (l'exemple illustré dans la figure 1). En effet, les transitions des états ne dépendent pas seulement que des états mis en cause dans la transition (les états voisins), mais aussi des états du modèle global(McCallum *et al.,* 2003a). D'autre part, CAC peut prendre plusieurs paramètres en entrée (autres caractéristiques des éléments d'entrée, comme la position syntaxique des mots dans un cadre textuel), ce qui permet l'utilisation de plusieurs niveaux hiérarchiques d'étiquetage. Cette propriété nous permet d'utiliser des information annexes afin de mieux décrire les données d'entrée. La probabilité conditionnelle $P(Y|X)$ est exprimée sous une forme exponentielle en CAC, comme le montre la formule suivante :

$$P(Y|X) = \frac{1}{Z_X} exp(\sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k f_k(y_i, x_i, v(x_i), X)), \qquad [4]$$

où $\{f_k(y_i, x_i, v(x_i), X)\}_{k=1}^{K}$ est un ensemble de fonctions aléatoires, appelées *fonctions caractéristiques*. Elles sont généralement des fonctions binaires de la présence des caractéristiques en fonction de l'étiquette candidate y_i, du mot (x_i), de son voisinage $v(x_i)$, et de sa séquence (X). Dans ce cas, le voisinage $v(x_i)$ ne se limite pas seulement aux mots voisins mais il peut inclure toutes les caractéristiques annexes

des mots. λ_k est un paramètre de pondération associé à chaque fonction caractéristique f_k. Z_X est un facteur de normalisation, c'est la somme de toutes les séquences candidates :

$$Z_X = \sum_{l \in L(X)} exp(\sum_{i=1}^{l} \sum_{k=1}^{K} \lambda_k f_k(y_i, x_i, v(x_i), X)), \quad [5]$$

avec $L(X)$ l'ensemble de toutes les sollutions possibles de X. Le problème d'estimation de paramètres (phase d'apprentissage) est de déterminer le vecteur $\theta = \langle \lambda_1, \lambda_2 \ldots \lambda_s \rangle$ à partir des données d'apprentissage $D = \{(x^i, y^i)\}_{i=1}^{N}$ avec une distribution empirique[3]. Dans (Lafferty *et al.,* 2001), l'auteur présente un algorithme qui maximise la fonction objective de log-vraisemblance $\ell(\theta)$. Le but étant de trouver le meilleur vecteur ℓ qui caractérise le mieux les données d'apprentissage :

$$\ell(\theta) = \sum_{i=1}^{N} log P_\theta(y^i | x^i). \quad [6]$$

Dans le cas où la probabilité conditionnelle porte sur une séquence d'éléments (comme c'est le cas dans les problèmes du TALN), l'algorithme de $Viterbi$ peut être appliqué pour maximiser cette fonction, comme c'est le cas dans l'implémentation que nous utilisons.

3. Etiquetage par le contexte

3.1. *Le corpus ESTER*

Nous avons utilisé dans nos expérimentations le corpus de données annoté de la compagne ESTER (Sylvain Galliano *et al.,* n.d.). La campagne ESTER[4] a pour but de mesurer les performances des systèmes de transcription, recherche d'information, et de compréhension d'émissions radiophoniques francophones. Les transcriptions du corpus ESTER sont enrichies par un ensemble d'informations annexes, comme le découpage automatique en tours de paroles, le marquage des EN. L'évaluation de la qualité de ces informations annexes est une tâche importante qui va permettre la mesure des performances d'un système d'indexation complet. Les données, composées de journaux et d'émissions radiophoniques, sont segmentées en sections, chaque section

3. La densité empirique d'une variable à valeurs discrètes est simplement constituée de la proportion des observations prenant chaque valeur.

4. campagne d'Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques

est dédiée à une thématique définie, qui implique des intervenants et des invités. Le corpus fourni se compose de 90 heures de radio en français, transcrites et annotées manuellement. Ce corpus est divisé en trois parties : la première, sert à l'apprentissage des modèles (*appr.*) ; la seconde, ensemble de développement (*dev.*), sert à l'ajustement des paramètres des modèles ; la troisième est dédiée aux tests et à l'évaluation des performances (*test*). La table 1 illustre la répartition des sources de données de la compagne ESTER.

Source	appr.	dev.	test
France Inter	32h	2h	1h
RFI-FM	21h	1h	2h
France Info	8h	1h	1h
RTM	21h	1h	1h
Total	82h	5h	5h

date	appr.	dev.	test
1998	18h	1h	1h
1999	4h	1h	-
2000	13h	1h	1h
2003	47h	2h	3h
Total	82h	5h	5h

Tableau 1. *distribution des données par sources et par dates.*

Nous Nous somme intéressé à la tâche d'extraction de l'information dans la compagne d'ESTER (Sylvain Galliano *et al.,* n.d.). Cette tâche consiste en l'annotation des EN à paritr des transcriptions manuelles des émissions radiophoniques, une tâche indispensable pour l'analyse sémantique des données.

3.2. *Le contexte hierarchique des entités*

L'étiquetage utilisé pour l'identification des EN est une description à plusieurs niveaux. En effet, chaque entité y est représentée par un concept y_1 ou plusieurs concepts $y_1, y_2, \ldots y_k$. Ainsi, nous avons $y = y_1.y_2 \ldots y_k$ où chaque concept y_i est subsumé par le concept y_{i-1} pour $i \in \{2, 3, \ldots k\}$ et le concept y_1 est subsumé par le concept le plus général dans notre représentation $Entity$. En conséquence, chaque concept est un noeud dans la hièrarchie des concepts, et chaque EN est représentée par un chemin dans la structure (Figure 2). Dans les annotations d'ESTER, le nombre maximal de niveaux est de 3. De ce fait, une étiquette est de la forme $y = y_1.y_2.y_3$ avec $y_1 \in Niveau(N_1)$, $y_2 \in Niveau(N_2)$, et $y_3 \in Niveau(N_3)$. Par exemple pour *Michael*, on associe l'étiquette *pers.hum*, où *pers* correspond au concept le plus général qui est personne ($pers \in Niveau(N_1)$), *hum* est le concept le plus spécifique qui est humain ($hum \in Niveau(N_2)$). Cette forme d'identification, détaillée, rend la tâche de reconnaissance plus difficile et complexe. Un système de reconnaissance de qualité acceptable peut reconnaître les concepts généraux, mais il y a moins de chances qu'il reconnaisse toute la chaîne. La simple application des CAC est de considérer chaque étiquette y comme une chaîne de concaténation des trois niveaux. L'inconvénient avec cette approche est qu'elle considère toutes les étiquettes indépendantes. En conséquence, le nombre d'étiquettes est plus important (elle contient toutes les combinaisons possibles des trois niveaux) et nécessite un grand nombre de données d'apprentissage pour construire un modèle qui caractérise le mieux toutes les

étiquettes. Notre approche consiste à construire un modèle de prédiction pour chaque niveau de conceptualisation. Chaque modèle M_j avec $j \in \{N_1, N_2, N_3\}$ est identifié dans un domaine non ambigu $D_j = \left\{\left(x^i, y_j^i\right)\right\}_{i=1}^{N}$ où N est l'ensemble des mots d'apprentissage. Chaque modèle M_k va nous fournir la prédiction y_j du mot x pour le niveau de conceptualisation j. La prédiction finale du mot x est alors la concaténation des trois prédictions : $y = y_1.y_2.y_3$. En utilisant la structure hiérarchique des concepts définie précédemment, nous pouvons vérifier la validité de la prédiction y. En effet, si y ne représente pas un chemin valide[5] dans la structure, des approximations peuvent êtres effectuées afin de valider cette prédiction. Pour illustrer ce cas de figure, donnons l'exemple suivant.

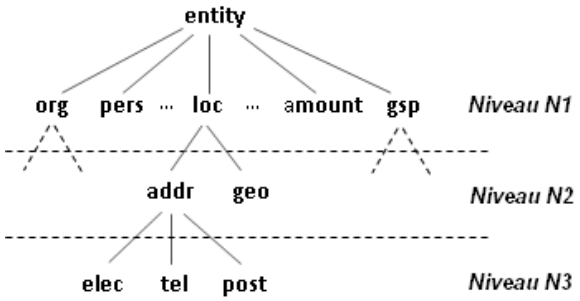

Figure 2. *Les niveaux hiérarchiques des EN*

3.2.1. *Exemple : approximation hiérarchique*

La figure 3 illustre l'application d'un étiquetage par niveaux pour une phrase qui se compose de 8 mots $\langle m_1, m_2, m_3, m_4, m_5, m_6, m_7, m_8\rangle$. L'étiquetage en niveau N_1 nous donne 3 labels N_1^1, N_1^2 et N_1^3 (figure 3.1). La figure 3.2 montre l'étiquetage en niveau N_2, et la figure 3.3 l'étiquetage en N_3. L'étiquetage final (complet) est la concaténation des étiquettes des 3 niveaux (figure 3.4). Dans ce dernier, on remarque que le mot m_4 est étiqueté en N_2 et N_3 mais non étiqueté en niveau N_1. Dans ce cas une amélioration peut être apportée par approximation de concepts en attribuant au mot m_4 l'étiquette N_1^X associée à la branche $N_1^X.N_2^3.N_3^2$ dans l'arbre des concepts. Notre approche consiste alors à construire un modèle pour chaque niveau indépendamment des autres. Cela implique une diminution considérable de la complexité ainsi qu'une augmentation des fréquences d'apparition des étiquettes.

3.3. *Les modèles enrichies*

La phase d'apprentissage des CAC permet l'utilisation de plusieurs informations pour caractériser le voisinage du mot ($v(i)$). En utilisant cette propriété, on peut ap-

5. Un chemin valide dans l'arbre est un chemin qui a comme début la racine de l'arbre, et une feuille comme fin.

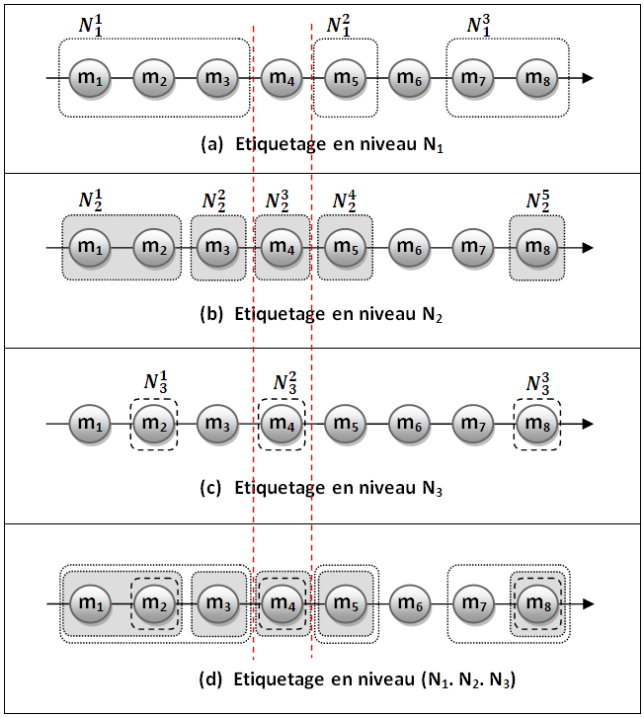

Figure 3. *Exemple d'utilisation de la hiérarchie pour l'amélioration de la prédiction*

pliquer un apprentissage par combinaison de niveaux. Ainsi, nous allons construire un modèle pour chaque niveau sachant les prédictions des autres niveaux. Le principe consiste à construire le modèle simple M_j puis les modèles combinés M_j^{comb} pour chaque niveau (étape 2 dans la figure 4). Le modèle combiné d'un niveau j utilise les prédictions fournies par les modèles simples $s/s \neq j$ comme données d'entrée. Dans la phase de test, on utilise les modèles simples pour génèrer les prédictions associées à chaque niveau (étape 3 dans la figure 4). Ces prédictions seront ainsi utilisées comme des données d'entrées pour les modèles combinés (étape 4 dans la figure 4). Par exemple, pour avoir la prédiction du 3ème niveau, on utilise comme connaissance a priori les prédictions des niveaux 1 et 2 générés par les modèles M_1 et M_2. Cette démarche nous permet de raffiner les prédictions pour chaque niveau, car on dispose de plus d'information a priori.

3.4. *Le contexte syntaxique*

Les CAC construisent des modèles qui caractérisent les données fournies en entrée. Chaque élément de ces données peut être définie par une ou plusieurs caractéristiques

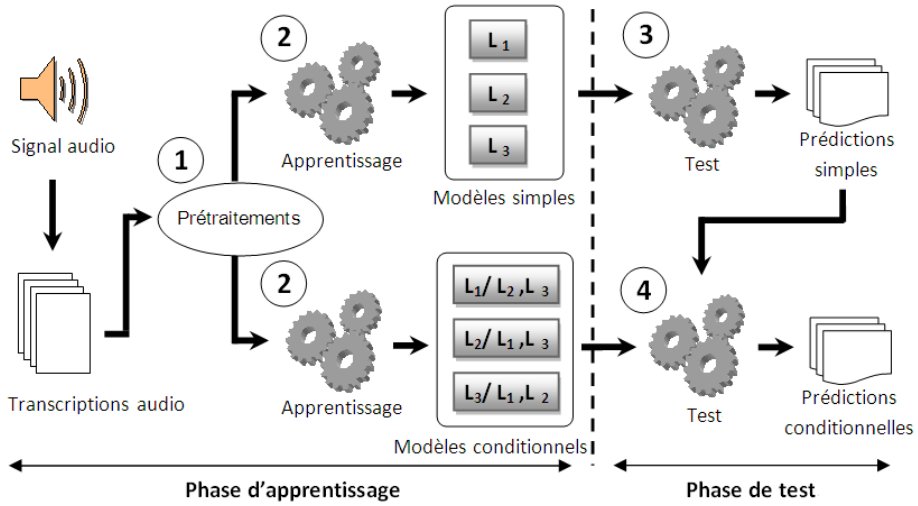

Figure 4. *Processus d'apprentissage conditionnel*

(appelées aussi attributs). Ces attributs peuvent ainsi être utilisés pour améliorer la qualité de prédiction des modèles(Sutton *et al.,* 2005). Dans le processus de REN, on peut identifier un mot simplement par sa représentation textuelle. Mais on peut aussi inclure d'autres informations annexes. Dans notre méthode nous avons inclus les étiquettes syntaxiques des mots[6]. En effet, on construit des modèles qui prennent en compte le contexte syntaxique dans la prédiction des mots. Pour effectuer les annotations syntaxiques des mots, nous avons utilisé le logiciel TreeTagger (Schmid, 1994). Ses performances atteignent une précision de 96% sur le corpus de test proposé par ses concepteurs.

4. Expérimentations et résultats

Pour tester les performances de notre approche, nous avons utilisé le corpus annoté d'ESTER. Etant dans un cadre de parole spontanée, la notion de segmentation en phrase est inexistante. En conséquence, nous avons utilisé une segmentation en tours de parole entre les intervenants. Nous avons utilisé le logiciel CRF++[7] avec une fenêtre de voisinage de 10 mots. Nous avons utilisé comme mesure de performance le *rappel*, la *précision*, et la *F(1)-mesure*. Nous avons fait en premier lieu une comparaison entre les modèles simples M_j et les modèles combinés M_j^{comb} pour chaque niveau $j \in \{1, 2, 3\}$ (figure 5). On remarque une amélioration de la F(1)-mesure de

6. En anglais : Part Of Speech tagging (POS tagging)
7. http ://crfpp.sourceforge.net/

1 à 3% dans les modèles combinés. A cet effet, nous avons considéré les prédictions combinées dans le résultat final.

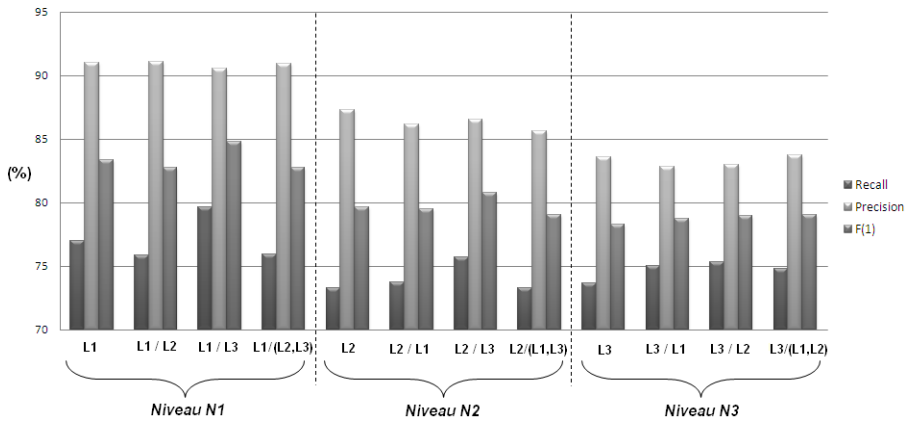

Figure 5. *Les résultats des CAC simples et combinés pour chaque niveau de conceptualisation*

Nous avons comparé les prédictions des différentes méthodes graphiques (MMC, CAC Séquentiels) avec les prédictions de notre approche (CAC Combinés). Les résultats illustratifs sont présentés dans le tableau 2. On remarque que les CAC nous donnent une meilleure précision par rapport au modèle markovien. Les CAC combinés apportent une fine amélioration la F(1)-mesure de 1% par rapport aux CAC séquentiels, dû aux approximations de concepts. L'intégration du contexte syntaxique améliore significativement le rappel (il passe de 61% à 74%). Ces résultats montrent que l'utilisation du contexte dans le processus de REN donne une meilleure qualité et un meilleur taux de prédiction. Par ailleurs, l'utilisation de l'apprentissage par niveaux devise par 5 le temps nécessaire pour la construction des modèles de prédiction (phase d'apprentissage). Ceci est dû principalement à la diminution de la complexité du corpus : la complexité passe de $O\left((N_1)^2 \times (N_2)^2 \times (N_3)^2\right)$ à $O\,(N_1)^2 + O\,(N_2)^2 + O\,(N_1)^2$ avec (N_j) la taille du domaine de sortie du niveau j.

Mesures	Rappel	Précision	F(1)-Mesure	Temps CPU
MMC	57, 9	69, 8	**63,0**	239h
CAC Classiques	61, 0	85, 8	**71,3**	274h
CAC Combinés	61, 4	86, 5	**72,2**	58h
CAC Classiques + Syntaxe	74, 0	83, 8	**78,5**	290h
CAC Combinés + Syntaxe	75, 9	84, 2	**79,9**	59h

Tableau 2. *Résultats des prédictions, les temps sont donnés pour l'apprentissage des 82h de temps de parole.*

5. Conclusion

La REN est une tâche centrale dans la chaîne des traitements sémantiques des transcriptions radiophoniques. Nous avons exposé les performances des méthodes graphiques dans la résolution du problème de classification, et plus particulièrement le problème d'étiquetage des données séquentielles. L'utilisation des structures des étiquettes dans le processus d'apprentissage apporte un gain dans la qualité d'étiquetage, car elle apporte une connaissance a priori des données. Nous avons montré que dans le domaine de l'indexation sémantique, la représentation conceptuelle surpasse la représentation linguistique des entités. En effet, l'utilisation de la hiérarchie des concepts améliore la classification et permet l'utilisation d'heuristiques d'approximation de concepts. Cette approche diminue considérablement le temps d'apprentissage. L'utilisation du contexte syntaxique des transcriptions apporte une connaissance supplémentaire sur données et améliore nettement le rappel. Nous allons, dans nos travaux futurs, combiner les connaissances extraites de plusieurs corpus et faire les correspondances entre les hiérarchies de concepts.

6. Bibliographie

Lafferty J. D., McCallum A., Pereira F. C. N., « Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data », *ICML '01 : Proceedings of the Eighteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 282-289, 2001.

McCallum A., Freitag D., Pereira F. C. N., « Maximum Entropy Markov Models for Information Extraction and Segmentation », *ICML '00 : Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 591-598, 2000.

McCallum A., Li W., « Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons », *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, Association for Computational Linguistics, Morristown, NJ, USA, p. 188-191, 2003a.

McCallum A., Rohanimanesh K., Sutton C., « Dynamic Conditional Random Fields for Jointly Labeling Multiple Sequences », *Workshop on Syntax, Semantics, Statistics*, 2003b.

Schmid H., « Part-of-speech tagging using decision trees », *Proceedings of International Conference on New Methods in Language Processing*, 1994.

Sutton C., McCallum A., « Joint Parsing and Semantic Role Labeling », *In Conference on Natural Language Learning (CoNLL)*, 2005.

Sutton C., Mccallum A., « An Introduction to Conditional Random Fields for Relational Learning », *in* , L. Getoor, , B. Taskar (eds), *Introduction to Statistical Relational Learning*, MIT Press, 2006.

Sylvain Galliano Edouard Geoffrois J.-F. B. G. G. D. M., Choukri K., « Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News », *Language Resources and Evaluation Conference (LREC 06)*, n.d.