# Compléments à l'extraction de descripteurs structurels et sémantiques

## Sommaire


**4.1 Segmentation en phrases par étiquetage de séquence . . . . . 70**
4.1.1 Conditional Random Fields . . . . . 71
4.1.2 Traits acoustiques et linguistiques . . . . . 73
4.1.3 Performances . . . . . 74
4.1.4 Améliorations envisagées . . . . . 77
**4.2 Extraction d'entités nommées dans le flux de parole . . . . . 77**
4.2.1 Introduction . . . . . 78
4.2.2 Coopération avec le processus de transcription . . . . . 80
4.2.3 Performances . . . . . 85
4.2.4 Limites . . . . . 90
**4.3 Conclusion . . . . . 90**


Le chapitre 2 a présenté les nombreuses méthodes de recherche d'information parlée et il s'avère que ce type d'information nécessite une extraction spécifique de descripteurs sémantiques à partir de l'acoustique. Par la suite, le chapitre 3 a introduit les différentes tâches de structuration et leur mise en œuvre dans la chaîne de structuration Speeral. L'objectif de ces travaux est de faciliter l'accès à l'information audio à l'aide du résumé de parole et les éléments de structuration présentés au chapitre précédent ne sont pas suffisants pour obtenir un résumé de qualité. Nous nous concentrons maintenant sur la présentation de deux compléments à la structuration pour le résumé automatique de parole. Tout d'abord, une segmentation en phrases de qualité est nécessaire pour résumer la parole avec une approche par extraction. En effet, du point de vue de l'utilisateur, cet aspect de la forme d'un résumé parlé est déterminant car une coupure inopportune au milieu d'une phrase peut fortement dégrader la compréhension. La méthode proposée pour la segmentation en phrases s'appuie sur un étiquetage

de séquence dans le cadre des *Conditional Random Fields* (section 4.1). Le second point de contribution réside dans l'extraction d'entités nommées dans le flux de parole. Ces entités liées au domaine (personnes, organisations, lieux...) dirigent la projection dans l'espace sémantique lors de la génération du résumé. L'approche développée pour cette tâche consiste en une recherche des entités nommées dans l'ensemble des hypothèses de transcription au lieu d'être restreinte à la meilleure hypothèse (section 4.2).

## 4.1 Segmentation en phrases par étiquetage de séquence

Il a été remarqué dans la section 2.2 que la segmentation en phrases demandait une attention particulière dans le cadre du résumé de parole par extraction (Rappelons que Mrozinski et al. (2006) ont observé une forte réduction de la qualité des résumés de parole fondés sur une segmentation automatique par rapport à une segmentation manuelle).

Dans la littérature, le problème de segmentation en phrases est généralement reformulé en un problème d'identification de frontières de phrases (étiquetage de séquence). La transcription automatique est employée pour générer une suite de mots et des frontières (événement binaire $B$) sont recherchées entre les mots. La décision est généralement issue d'une combinaison de paramètres prosodiques (événement $S$) et linguistiques (événement $L$). Trouver des frontières de phrases est loin d'être facile, en attestent par exemple Stevenson et Gaizauskas (2000), qui évaluent les performances d'annotateurs humains sur la reponctuation d'un texte, et qui observent qu'il est beaucoup plus facile de reponctuer un flux de mots contenant les majuscules d'origine ($F_1$-mesure de 0.95) qu'en l'absence de ces marqueurs ($F_1$-mesure de 0.80), comme dans le cas d'une transcription automatique.

La majorité des approches sont fondées sur des modèles probabilistes tentant de prédire la séquence $B$ en fonction de $S$ et $L$. Gotoh et Renals (2000) constituent un modèle pour chacune des modalités ($S$ et $L$) sur des ensembles de données séparés. La probabilité linguistique $P(B, L)$ qu'une frontière de phrase précède un mot est modélisée à partir de données textuelles disponibles en masse ; l'implication de la prosodie $P(B, S)$ est modélisée à partir des durées de pauses sur un corpus acoustique de plus petite taille. Les deux modèles sont fusionnés grâce à une heuristique[1]. Shriberg et al. (2000) étudient les différentes caractéristiques prosodiques en profondeur : les pauses, le rythme phonétique ou syllabique, la pente de fréquence fondamentale ($f_0$) et sa continuité, les sauts de $f_0$, l'écart à la moyenne de la $f_0$, et la qualité de voix. Les valeurs sont fonction du locuteur ou d'un locuteur moyen lorsque les données sont insuffisantes. En plus de ces paramètres, la décision repose sur la durée des phrases et les changements de locuteurs (segmentation manuelle en locuteurs). Un arbre de décision donne une sélection des paramètres les plus pertinents et ces derniers servent à construire un modèle de séquence génératif. Les paramètres les plus efficaces sur des données radio-diffusées semblent être les pauses et les changements de locuteurs. Liu et al. (2005) continuent ces

[1] $P(B, L, S) \simeq P(B, S)^{\alpha} P(B, L)$, $\alpha > 10$ donnant les meilleurs résultats.

travaux en comparant des approches HMM, maximum d'entropie et CRF pour l'étiquetage de la séquence : ce dernier modèle s'avérant être le plus efficace (une fusion des trois apporte un gain complémentaire). Il est intéressant de noter que la décision prosodique sur la frontière est prise avant l'inclusion dans le modèle de séquence. Des travaux similaires de (Kim et al., 2004) intègrent des arbres de décision avec un système de détection de difficultés de prononciation.

La tâche de détection de frontières de phrase (en anglais, *Sentence Unit Boundary Detection*, SUBD) a été évaluée lors des éditions 2002 à 2004 des campagnes *Rich Transcription* « automne » (RT-fall), organisées par NIST. Les données de référence reposent sur un guide d'annotation (Strassel, 2003)[2] précisant que la notion de phrase à l'oral (nommée « unité syntagmatique ») est différente de l'écrit. Les différences sont avant tout grammaticales ; les unités sont classées selon leur type (déclarations, questions, éléments phatiques et unités incomplètes). La mesure de performance NIST est le taux d'erreur sur les frontières (nombre de frontières oubliées, ajoutées ou de mauvais type, divisé par le nombre de frontières dans la référence : équation 4.1 dans laquelle $\mathrm{nb}(\cdot)$ est le cardinal d'un ensemble de frontières).

$$SB_{err} = \frac{\mathrm{nb}(\text{oubli}) + \mathrm{nb}(\text{ajout}) + \mathrm{nb}(\text{mauvais type})}{\mathrm{nb}(\text{référence})} \tag{4.1}$$

Sur des données radio-diffusées, Liu et al. (2005) aboutissent à un taux d'erreur de 0.54 (sans prendre en compte les erreurs de type). Cette valeur correspond à une $F_1$-mesure d'environ 0.70, proche des performances annoncées par les autres auteurs.

La détection de frontières de phrases que nous avons mise en place pour le résumé de parole est similaire à l'approche de Liu et al. (2005). En restant dans le cadre de l'étiquetage bi-classe de la séquence de mots, nous appliquons un modèle CRF sur des caractéristiques prosodiques et linguistiques. Ces dernières sont issues de la chaîne de structuration Speeral. Les frontières de phrases sont recherchées dans les émissions de radio en français de la campagne ESTER.

### 4.1.1 Conditional Random Fields

#### Définition

*Conditional Random Fields* (CRF, Lafferty et al., 2001) est un cadre probabiliste discriminant pour l'étiquetage de séquences. Au lieu de modéliser la probabilité jointe d'apparition des séquences d'observation et des séquences d'étiquettes comme le fait une approche générative telle que HMM, CRF repose sur la probabilité conditionnelle de l'étiquetage sachant l'ensemble de la séquence. Les méthodes à maximum d'entropie de Markov (MEMM) recherchent aussi à maximiser cette probabilité conditionnelle, mais de façon locale. Ceci pose des problèmes au niveau des hypothèses partielles débouchant sur un petit nombre de successeurs car ils sont systématiquement préférés

[2] disponible en ligne sur http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V5.0.pdf, visité en novembre 2006

aux chemins de plus grande entropie. Cet effet est décrit sous le nom d'effet du biais des étiquettes par Lafferty et al. (2001).

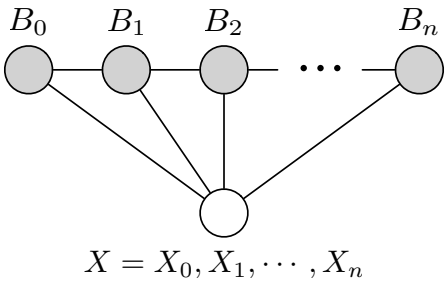


**FIG. 4.1:** *Détection de frontières de phrases par modélisation CRF. La séquence d'événements représentant la présence ou l'absence de frontières ($B = B_0, \cdots, B_n$) est globalement conditionnée par la séquence d'observations phonétiques et linguistiques ($X = X_0, \cdots, X_n$).*

Appliquons CRF à une tâche de segmentation en phrases : $B$ est une séquence d'étiquettes ($B = 1$ pour une frontière de phrase, $B = 0$ pour une absence de frontière) ; $X$ est une séquence d'observations prosodiques et linguistiques. Le modèle conditionne la séquence $B$ sur l'ensemble de la séquence $X$ (figure 4.1). La meilleure hypothèse d'étiquetage est celle qui maximise la probabilité $P(B|X)$. Cette probabilité est estimée par une distribution de forme exponentielle satisfaisant des caractéristiques sur des données d'apprentissage (équation 4.2).

$$\begin{aligned} \hat{B} &\operatorname*{argmax}_B P(B|X) \\ P(B|X) &\simeq \frac{1}{Z(X)} e^{\sum_k \lambda_k f_k(B,X)} \\ Z(X) &= \sum_B e^{\sum_k \lambda_k f_k(B,X)} \\ f_k(B,X) &\geq 0 \end{aligned} \tag{4.2}$$

Dans cette équation, les $\lambda_k$ sont les paramètres du modèle ; $Z(X)$ sert à la normalisation de la distribution ; $f_k(B,X)$ sont les fonctions caractéristiques sur les arcs et les sommets du modèle graphique associé au problème. Ces fonctions sont des relations entres les $B_i$ et $X$ et entre des $B_i$ voisins.

L'inférence des paramètres $\lambda_i$ se fait par maximisation de la vraisemblance conditionnelle sur un ensemble de données étiquetées. Le maximum de cette fonction log-concave est découvert par des méthodes de maximisation classiques, comme *Generalized Iterative Scaling* (GIS, Darroch et Ratcliff, 1972), *Improved Iterative Scaling* (IIS, Della Pietra et al., 1997), ou *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (LBFGS, Liu et Nocedal, 1989), qui s'avère être la plus rapide. Ces méthodes sont comparées dans (Malouf, 2002). La dépendance des étiquettes sur l'ensemble de la séquence d'observation rend l'apprentissage beaucoup plus coûteux que pour un maximum d'entropie local classique. L'étiquetage d'une séquence nouvelle se fait par programmation dynamique.

#### La boîte à outils CRF++

L'ensemble de nos expériences sur la détection de frontières de phrases repose sur CRF++[3], une boîte à outils pour l'étiquetage de séquences fondée sur CRF. CRF++ implémente un apprentissage dont l'optimisation repose sur une méthode de quasi-Newton (LBFGS) et un décodage grâce à l'algorithme Viterbi. Cette boîte à outils a été utilisée avec succès pour de nombreuses tâches de traitement automatique du langage naturel comme la désambiguïsation sémantique, la décomposition en groupes grammaticaux, l'étiquetage morpho-syntaxique ou encore la détection d'entités nommées (Kudo et al., 2004).

### 4.1.2 Traits acoustiques et linguistiques

Nous suivons les approches classiques pour la segmentation en phrases en recherchant des frontières potentielles uniquement entre les mots et en fixant l'événement $B = 1$ si une frontière a précédé un mot et $B = 0$ dans le cas contraire. La prédiction de la présence d'une frontière de phrase avant un mot dépend de caractéristiques linguistiques et acoustiques que nous allons décrire (voir table 4.1). Au niveau linguistique, les mots et leurs catégories morpho-syntaxiques modélisent les phénomènes grammaticaux de la séquence. La catégorie morpho-syntaxique des mots est trouvée grâce à lia_tagg[4]. Cet étiqueteur repose sur un dictionnaire d'étiquettes possibles par mots et effectue l'étiquetage dans un cadre HMM. Alors que certains couples syntaxiques, comme « le déterminant et le nom », qui ne doivent pas être séparés par une frontière de phrase, sont plutôt bien capturés par cette modélisation, d'autres groupes comme « le verbe et son complément » sont moins faciles à détecter sans une modélisation plus approfondie de la grammaire. Si les éléments linguistiques sont utiles pour reponctuer un texte, ils peuvent être faussés par les erreurs de transcription, d'étiquetage morpho-syntaxique et le manque relatif de grammaire de la langue parlée. Pour y remédier, il faut associer des caractères acoustiques aux indices linguistiques, comme les changements de locuteur et quelques éléments de prosodie. Les changements de locuteurs sont issus, comme la séquence de mots, de la chaîne de transcription et employés tels quels sans prendre en compte les identités retrouvées. En terme de prosodie, les pauses sont explorées à deux niveaux : avant le mot et à l'intérieur du mot pour essayer d'éviter de prendre les hésitations pour des fins de phrase. De plus, comme il est difficile de profiter des informations apportées par la courbe de fréquence fondamentale ($f0$), nous utilisons seulement sa pente globale, sur trois horizons temporels différents (le mot, une fenêtre allant de 4 secondes avant le début du mot jusqu'à sa fin et une fenêtre allant de 8 secondes avant le début du mot jusqu'à sa fin). Bien que cette approche ne soit pas optimale, elle permet tout de même de modéliser les grands phénomènes macro-prosodiques de la phrase. Toutefois, certaines caractéristiques sont perdues, comme les effets du rythme prosodique ou syllabique connus pour ralentir en fin de phrase. Les

[3]Disponible sur http://chasen.org/~taku/software/CRF++, visité en août 2006.

[4]Étiqueteur morpho-syntaxiaque du LIA, disponible sous licence GPL, sur http://www.univ-avignon.fr/chercheurs/bechet/download_fred.html, visité en octobre 2006.

frontières des mots de la référence sont extraites grâce à un alignement forcé sur le signal en utilisant un outil dérivé du système de transcription (gvalign).

| Mot | Étiquette | P1 | P2 | Loc. | F1 | F2 | F3 | Ponct. |
|---|---|---|---|---|---|---|---|---|
| avait | V3S | C0 | C0 | SPK | C0 | C0 | C0 | point |
| le | DETMS | C0 | C0 | n | C0 | C0 | C0 | x |
| salut | NMS | C0 | C4 | n | C4 | C0 | C0 | x |
| à | XSOC | C0 | C0 | n | C0 | C8 | C3 | x |
| tous | AINDMP | C0 | C0 | n | C4 | C0 | C0 | x |
| ceux | PDEMMP | C0 | C0 | n | C5 | C4 | C0 | x |
| en | PREP | C0 | C0 | n | C5 | C5 | C0 | x |
| bonne | AFS | C0 | C0 | n | C2 | C4 | C0 | point |
| journée | NFS | C0 | C0 | n | C5 | C3 | C0 | x |
| euh | ADV | C0 | C0 | n | C4 | C8 | C1 | x |

**TAB. 4.1:** *Exemple des paramètres extraits pour la segmentation en phrases. Au niveau linguistique : le mot de la transcription et son étiquette morpho-syntaxique. Au niveau prosodique : la durée de pause avant le mot (P1) et à l'intérieur du mot (P2), un éventuel changement de locuteur avant le mot (Loc.), la pente de F0 à divers horizons temporels (F1=le mot, F2=-4s, F3=-8s). La ponctuation qui précède le mot est prédite grâce à ces paramètres (Ponct.). Les valeurs numériques sont quantifiées uniformément selon les classes C0 à C9 (sur une fenêtre glissante de 300 valeurs, avec un jeu de classes par paramètre).*

CRF++ facilite la génération des fonctions caractéristiques en utilisant des patrons de conjonction d'événements de $X$ et $B$. Dans notre implémentation, une frontière de phrase potentielle est conditionnée par des séquences $n$-grammes de chaque type de caractères linguistiques et acoustiques autour du mot à étiqueter et par la conjonction des séquences précédentes (illustrées par la figure 4.2). La boîte à outils est cependant limitée dans sa version actuelle à des caractéristiques symboliques. Cette limitation implique la quantification des valeurs continues comme la durée des pauses ou la pente de fréquence fondamentale. La quantification se fait sur une fenêtre glissante en utilisant une répartition uniforme en $n$ classes[5]. Cette approche permet de normaliser les valeurs lors de changements de locuteurs et d'environnement.

### 4.1.3 Performances

Les performances en segmentation en phrases sont calculées sur la base du nombre de frontières bien placées par rapport au nombre de frontières erronées, en rappel, précision, et $f$-mesure (un exemple est donné par la table 4.2). Les expériences sont réalisées sur le corpus ESTER qui n'a malheureusement pas fait l'objet de directives d'annotation pour les frontières de phrases. Le guide d'annotation précise que « la ponctuation est facultative, mais peut être utilisée pour faciliter la tâche de transcription ». Cette dernière varie donc beaucoup d'un annotateur à l'autre ; les phrases peuvent être

[5] $n = 10$ dans les expériences qui suivent. La fenêtre glissante fait 300 valeurs. Ces valeurs sont fixées empiriquement, mais ne semblent pas avoir un impact important sur les performances.

| Mot | Étiquette | P1 | P2 | Loc. | F1 | F2 | F3 | Ponct. |
|---|---|---|---|---|---|---|---|---|
| avait | V3S | C0 | C0 | SPK | C0 | C0 | C0 | point |
| le | DETMS | C0 | C0 | n | C0 | C0 | C0 | x |
| salut | NMS | C0 | C4 | n | C4 | C0 | C0 | x |
| à | XSOC | C0 | C0 | n | C0 | C8 | C3 | x |
| tous | AINDMP | C0 | C0 | n | C4 | C0 | C0 | x |
| ceux | PDEMMP | C0 | C0 | n | C5 | C4 | C0 | x |
| en | PREP | C0 | C0 | n | C5 | C5 | C0 | x |
| bonne | AFS | C0 | C0 | n | C2 | C4 | C0 | point |
| journée | NFS | C0 | C0 | n | C5 | C3 | C0 | x |
| euh | ADV | C0 | C0 | n | C4 | C8 | C1 | x |

**FIG. 4.2:** *Illustration des groupes de paramètres utilisés pour la prédiction de la présence ou absence d'une frontière de phrase. En plus de ces événements, le modèle prend en compte les unigrammes dans une fenêtre de deux mots autour du mot courant et la conjonction de chacun des événements précédents sur l'ensemble de la séquence. Les données sont celles de la figure 4.1*

très longues, jusqu'à faire un tour de parole complet, contenant un grand nombre de virgules, alors que dans d'autres cas, chaque pause du locuteur a été annotée par une fin de phrase. Ce problème de fiabilité du corpus implique une nécessaire prudence dans l'interprétation des résultats d'évaluation.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Référence | * | * | * | p | * | p | * | * | * | * | * | p |
| Hypothèse | * | p | * | * | * | p | * | * | * | p | * | p |

**TAB. 4.2:** *Performances de la segmentation en phrases pour un exemple fictif. « p » représente une frontière de phrase et « * » une absence de frontière. Il y a 3 frontières à trouver dans la référence, 4 frontières ont été trouvées dans l'hypothèse, dont 2 bien placées. La précision est de $P = 2/4 = 0.5$, le rappel est de $R = 2/3 = 0.66$ et la $F_1$-mesure est de $F_1 = 2 * PR/(P + R) = 0.57$. Le taux d'erreur NIST est égal au nombre d'erreurs ($hyp_i \neq ref_i$) par rapport au nombre de frontères à trouver : $SB_{err} = 3/3 = 100\%$.*

Il est intéressant de noter que les journalistes des radios francophones du corpus ont tendance à utiliser une architecture prosodique très spéciale qui détériore la cohérence des événements caractérisant une frontière de phrase. En effet, pour captiver l'attention de l'auditeur, les journalistes reprennent leur souffle en milieu de phrase, pour provoquer un effet « d'attente ». Cet effet diminue la cohérence de l'annotation par l'insertion de pauses. Ces pauses ont les caractéristiques acoustiques d'une fin de phrase et les caractéristiques linguistiques d'un milieu de phrase.

#### Comparatif en structuration automatique et manuelle

Les données d'entraînement utilisées dans ces expériences correspondent aux 80 heures d'entraînement (environ 874000 mots) du corpus ESTER, alors que les performances sont rapportées pour les 10 heures de la partie développement du corpus (en-

| Données | Rappel | Précision | $F_1$-mesure |
|---|---|---|---|
| *Étiquetage : points* | | | |
| M+M | 0.42 | 0.80 | 0.55 |
| M+A | 0.34 | 0.84 | 0.49 |
| A+M | 0.62 | 0.74 | 0.67 |
| A+A | 0.61 | 0.77 | 0.68 |
| *Étiquetage : points et virgules* | | | |
| M+M | 0.41 | 0.64 | 0.50 |
| M+A | 0.30 | 0.72 | 0.42 |
| A+M | 0.50 | 0.65 | 0.56 |
| A+A | 0.49 | 0.74 | 0.59 |
| *Étiquetage : points et virgules fusionnés* | | | |
| M+M | 0.55 | 0.78 | 0.64 |
| M+A | 0.37 | 0.81 | 0.51 |
| A+M | 0.59 | 0.70 | 0.64 |
| A+A | 0.55 | 0.81 | 0.66 |

**TAB. 4.3:** *Performances de la segmentation en phrases selon le type d'étiquetage recherché et les données utilisées en apprentissage et en test. Par exemple, « M+A » représente un apprentissage sur les données extraites à la main (M) et un test sur les données transcrites et segmentées automatiquement (A).*

viron 88000 mots). La table 4.3 présente des comparatifs entre l'utilisation de données structurées automatiquement ou manuellement en apprentissage et en test, pour les tâches d'étiquetage sur les points («.») comme frontières de phrases, les points et les virgules sous forme d'un problème 3-classes («.», «,» et ∅) et la fusion des points et des virgules («,»=«.»).

Globalement, les tests sur les données structurées manuellement montrent que l'approche admet un faible rappel et une forte précision sur les frontières retrouvées. La différence est moins prononcée lors de l'utilisation de données structurées automatiquement lors du test. De plus, la méthode la plus performante consiste en l'utilisation de données structurées automatiquement en apprentissage et en test. En revanche, les données de référence mènent à de moins bonnes performances générales. Il semblerait que ceci soit dû à une différence dans la notion de pause entre l'algorithme d'alignement automatique et le système de transcription. Pour ce qui est des différents étiquetages possibles, étant donné qu'aucun guide d'annotation en frontières de phrases n'a été fourni lors de la création des données de référence, nous avons essayé de réduire les incohérences virgule-point en fusionnant ces 2 types de frontières et en les annotant séparément. Les performances ne sont néanmoins jamais au niveau de celles obtenues par l'annotation des « points ».

Finalement, nous déduisons de ces résultats que l'approche permet d'établir des performances de l'ordre de ce qui est donné dans la littérature (une $f_1$-mesure d'environ 0.70). De plus, il est bon de noter que la méthode a une bonne précision et une tendance à sous-générer les frontières de phrases. Ce type de comportement est béné-

fique pour le résumé automatique car le type d'erreur le plus pénalisant dans ce cadre reste l'insertion de frontières de phrases là où elles n'ont pas lieu d'être.

### 4.1.4 Améliorations envisagées

Nous avons proposé une détection des frontières de phrases par étiquetage d'une séquence d'« inter-mots » à l'aide de CRF. L'approche peut être améliorée en utilisant des caractéristiques continues (et non symboliques) — en prenant en compte les scores de confiance du système de transcription — et en calculant une courbe de $f0$ plus fine, normalisée pour chaque locuteur. Une des limitations de CRF est que cette approche ne peut tenir compte de paramètres au niveau global de la phrase, comme sa longueur ou sa cohérence syntaxique et sémantique. Une solution à ce problème peut être semi-CRF (Sarawagi et Cohen, 2005) qui tente de remettre en cause de l'hypothèse Markovienne[6] du processus sur un segment temporel de taille raisonnable de l'ordre de la phrase. D'autres pistes doivent être envisagées, comme un test sur la fiabilité du corpus afin de détecter et d'écarter les phrases mal annotées, ou une intégration complète de la segmentation en phrases dans la transcription du contenu parlé pour retarder la prise de décision sur les frontières de mots.

## 4.2 Extraction d'entités nommées dans le flux de parole

Les entités nommées sont des entités du monde « réel », dont la forme linguistique est une représentation directe dénuée d'ambiguïté. Notamment, lorsqu'une de ces entités se retrouve dans le discours de plusieurs personnes, il est considéré que ces différentes références ont le même antécédent. Bien que cette affirmation soit loin d'être vraie dans le cas général, les types d'entités recherchés doivent s'en approcher le plus possible. Par exemple, « une table » est un concept qui se réfère à un objet dans un contexte donné. Dans un autre contexte, le locuteur se référera généralement à une autre entité. En revanche, dans un domaine journalistique, les noms propres se réfèrent à des objets considérés comme uniques, dont la forme linguistique peut être séparée de son contexte sans rendre la référence ambiguë. Ce type de comportement est très intéressant dans le cadre de l'analyse sémantique indispensable pour le résumé car la projection depuis la linguistique devient transparente.

Dans le cadre de l'extraction de descripteurs sémantique de journaux radio diffusés, les entité nommées sont étendues à certaines quantités fortement porteuses d'information dans ce domaine. Les entités recherchées sont de deux types : entités uniques basées sur des noms propres (personnes, lieux, organisations...) et entités basées sur des séquences de noms communs (dates, quantités monétaires, distances...). Les majuscules des noms propres sont de bons indicateurs de la présence d'entités du premier type et les valeurs numériques sont de bons indicateurs de la présence du second type d'entité.

[6] L'hypothèse Markovienne est vérifiée pour un processus si et seulement si la distribution conditionnelle de probabilité des états futurs, étant donné l'instant présent, ne dépend que de ce même état présent et pas des états passés.

### 4.2.1 Introduction

Les deux approches majeures pour l'extraction d'entités nommées sont la création de grammaires spécifiques au domaine, à base de règles et de listes de mots, et l'étiquetage de séquence par apprentissage. Ce dernier est le plus efficace lorsque la séquence à étiqueter est bruitée et que des données d'apprentissage sont disponibles. Dans ce cadre, la séquence d'étiquettes est mise en correspondance avec la séquence de mots (les observations) en utilisant le formalisme *Begin Inside Outside* (BIO). Lorsqu'une étiquette s'étale sur plusieurs mots, elle est subdivisée en une méta-étiquette par mot, selon la position du mot dans l'étiquette. La méta-étiquette *Begin* (B) correspond à un mot en début d'entité, la méta-étiquette *Inside* (I) correspond à un mot en milieu ou fin d'entité et la méta-étiquette *Outside* (O) correspond à un mot à l'extérieur d'une entité. La table 4.4 donne un exemple de correspondance BIO. Celle-ci permet de s'affranchir du problème de segmentation et de se concentrer sur le problème d'étiquetage.

| ← **org** → | | | ← **pers** → | | | | ← **time** → | |
|---|---|---|---|---|---|---|---|---|
| B-org | I-org | O | B-pers | I-pers | O | O | B-time | I-time |
| *France* | *Inter* | bonjour | *Nicolas* | *Stoufflet* | il | est | *sept* | *heures* |

**TAB. 4.4:** *Illustration de la correspondance Begin Inside Outside (BIO) pour transformer le problème d'étiquetage en entités nommées en un problème de classification de séquence.Begin (B) correspond à un mot en début d'entité, Inside (I) à un mot au milieu ou en fin d'entité et Outside (O) à un mot à l'extérieur d'une entité.*

Dans un cadre probabiliste, le problème est formalisé de la façon suivante : pour une séquence de mots $W = w_0, \cdots, w_n$ donnée, la séquence d'étiquettes $L = l_0, \cdots, l_n$ la plus probable est recherchée (elle maximise la probabilité $P(L|W)$, équation 4.3).

$$\hat{L} = \underset{L}{\operatorname{argmax}}\, P(L|W) \tag{4.3}$$

La probabilité d'une séquence peut être exprimée en factorisant la probabilité à posteriori et sera approximée en ne prenant en compte qu'un contexte fixe pour déterminer l'étiquette associée à un mot (à travers $p(l_n|ctx_n)$, équation 4.4).

$$\begin{aligned} P(L|W) &= \prod_n P(l_n|l_0, \cdots, l_{n-1}, w_0, \cdots, w_n) \\ P(L|W) &\simeq \prod_n p(l_n|ctx_n) \end{aligned} \tag{4.4}$$

Bender et al. (2003) et Miller et al. (2000) appliquent une méthode à base de maximum d'entropie pour estimer $P(L|W)$. Dans ce cadre proposé par Berger et al. (1996)[7] pour le traitement de la langue naturelle, la probabilité d'une étiquette en contexte est exprimée sous forme d'une distribution satisfaisant des contraintes sur les données

[7]Une boîte à outils est disponible sur `http://maxent.sourceforge.net/`, visité en novembre 2006.

d'apprentissage $f(\cdot)$ (équation 4.5).

$$p(l_n|ctx_n) = \frac{1}{Z(ctx_n)} \prod_f e^{\lambda_f f(ctx_n, l_n)} \tag{4.5}$$
$$Z(ctx_n) = \sum_l \prod_f e^{\lambda_f f(ctx_n, l)}$$

Dans cette formulation, les $f(\cdot)$ sont des fonctions binaires de la présence de caractéristiques. $\lambda_f$ est le paramètre associé à chaque caractéristique. $Z(\cdot)$ est un facteur de normalisation dépendant uniquement du contexte. Le contexte est généralement constitué de mots autour du mot courant et des $m$ étiquettes attribuées aux mots précédents. Les caractéristiques sont fondées sur des propriétés des mots pour assurer la généralisation du procédé, comme la casse (capitalisé, majuscule, minuscule), la présence de caractères spécifiques (chiffres, ponctuations, tirets, virgule, dollar), la catégorie morpho-syntaxique (adjectif, nom, verbe) et l'appartenance à des listes de mots (prénom, ville, société...).

Les paramètres $\lambda_f$ sont appris sur un corpus d'entraînement étiqueté manuellement par maximisation de la vraisemblance conditionnelle des modèles produits en utilisant l'algorithme *Global Iterative Scaling* (GIS, Darroch et Ratcliff, 1972) ou des algorithmes d'optimisation convexe. L'algorithme Viterbi est utilisé pour déterminer la séquence la plus probable sachant que la fonction de normalisation $Z(\cdot)$ peut être ignorée (car elle est indépendante de l'étiquette). Pour éviter le sur-apprentissage, les paramètres des modèles sont contraints par une distribution *a priori* de type gaussienne (Chen, 1999). Chieu et Ng (2002) proposent d'ajouter des statistiques globales à cet étiquetage local afin de prendre en compte les affinités entre les entités et leur contexte d'utilisation. Collins et Singer (1999) appliquent une technique de co-apprentissage pour trouver les attributs efficaces d'extraction d'entités nommées sur un corpus partiellement étiqueté. Le principe est d'utiliser deux méthodes d'étiquetage faiblement couplées ; les étiquettes trouvées par la première méthode associées à une bonne confiance, sont utilisées comme référence pour la seconde méthode, puis le processus est itéré alternativement sur chaque méthode jusqu'à l'étiquetage complet du corpus. D'autres algorithmes obtiennent d'excellentes performances sur cette tâche, comme CRF (McCallum et Li, 2003), ou SVM (Kazama et al., 2002), en utilisant une formulation similaire du problème.

Les performances de ces approches sont directement liées à la quantité de données d'apprentissage disponibles. Haghighi et Klein (2006) n'utilisent pas de données d'apprentissage, mais un corpus non étiqueté et un petit nombre d'exemples d'étiquetages. La distribution globale des voisinages des exemples connus permet d'inférer l'étiquetage du reste du corpus. Cette méthode obtient environ 80% des performances d'une méthode avec des données complètement étiquetées (sur une tâche d'étiquetage morpho-syntaxique).

Contrairement à la recherche d'information sur un contenu parlé qui profite de la redondance des documents, l'annotation en entités nommées est beaucoup plus sensible aux erreurs de transcription. Par exemple, Kubala et al. (1998) étudient l'application

d'Identifinder (Bikel et al., 1997) sur un corpus de parole journalistique et notent que le taux d'erreur de mots de la transcription a un effet direct sur celui de l'annotation, les noms propres étant les plus touchés. Une première idée afin de prendre en compte les erreurs de transcriptions est d'adapter l'algorithme de détection pour qu'il autorise des insertions et des délétions (les substitutions sont des délétions suivies d'insertions, Grishman, 1998; Gotoh et Renals, 1999). Une autre proposition est de travailler directement sur les graphes d'hypothèses du système de transcription de la parole (Horlock et King, 2003; Béchet et al., 2004) et ainsi retrouver des entités bien formées parmi les hypothèses les plus probables. Enfin, Acero et al. (2004) adaptent le modèle de langage du système de transcription à la tâche finale en utilisant une grammaire probabiliste. Cette approche a le désavantage de nécessiter beaucoup de données d'apprentissage et ne peut s'appliquer qu'à des cas très particuliers (Wang et Acero, 2003). Pour éviter que les noms propres soient considérés comme des mots inconnus, Allauzen (2003) met à jour le lexique du système de transcription avec des mots de corpus textuels similaires au corpus de parole annoté.

L'évaluation de l'étiquetage en entités nommées est réalisé en utilisant principalement deux mesures : la $F_\beta$-mesure et le Slot Error Rate. La $F_\beta$-mesure est composée du rappel $R$ (nombre d'entités justes par rapport au nombre d'entités à trouver) et de la précision $P$ (nombre d'entités justes par rapport au nombre d'entités trouvées). La $F$-mesure est présentée en 3.2.2. Le *Slot Error Rate* (SER) se veut plus précis car il caractérise mieux les erreurs et les pénalise de façon plus fine. La formulation du SER est donnée par l'équation 4.6, dans laquelle $I$ est le nombre d'insertions, $D$ le nombre de délétions, $S$ le nombre de substitutions, $R$ le nombre d'entités de la référence et $\alpha_i$ le poids associé chaque type d'erreur.

$$SER = \sum_{e \in \{I,D,S\}} \frac{\alpha_e e}{Nb_{Ref}} \tag{4.6}$$

SER autorise une pondération des substitutions en fonction de leur origine : erreur sur le type de l'entité, son contenu, ou sa portée. La réalisation manuelle de références (et donc de corpus d'apprentissage) pose de nombreux problèmes car les classes peuvent être redondantes et l'annotation ambiguë (à cause d'une interprétation différente des règles d'annotation par les annotateurs). Les phénomènes d'anaphore et de métonymie entrent en jeu et bien définir les règles d'annotation prend une grande importance afin d'en conserver la cohérence. Ces problèmes sont abordés dans la convention d'annotation des entités nommées de la campagne ESTER (Le Meur et al., 2004).

### 4.2.2 Coopération avec le processus de transcription

La reconnaissance d'entités nommées dans un flux de parole en utilisant un système de transcription de parole est intrinsèquement limitée par le vocabulaire que ce dernier peut reconnaître et par les erreurs qu'il peut commettre en le générant. Plus précisément, les entités nommées sont généralement constituées de noms propres intéressants pour leur fréquence élevée dans un contexte local par rapport à leur rareté globale. Ces

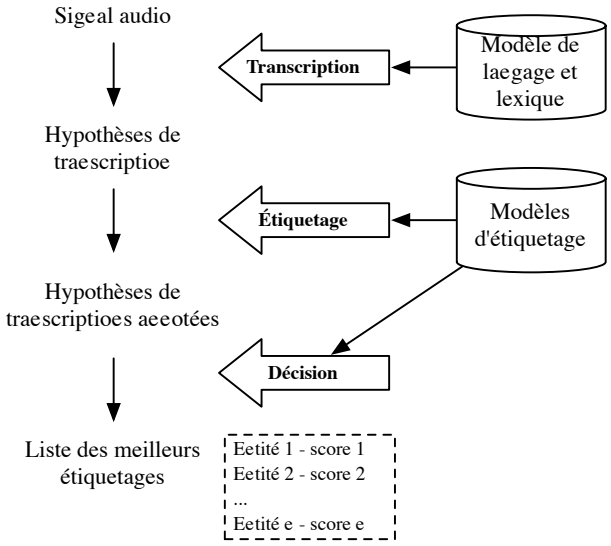


**FIG. 4.3:** *Illustration de l'extraction d'entités nommées dans un flux de parole. L'approche est fortement couplée au moteur de transcription, travaillant sur les treillis d'hypothèses de phrases. Un premier module détermine les annotations possibles de toutes les hypothèses de phrases ; le module de décision extrait les meilleurs entités en fonction des besoins de l'application visée.*

noms propres ont moins de chances d'apparaître dans le lexique du système de transcription et leurs probabilités d'apparition sont faibles dans les modèles grammaticaux du dit système. Les entités sont relativement mal reconnues ou tout simplement non reconnues. L'approche proposée consiste à diminuer l'influence des entités mal reconnues en recherchant des entités nommées directement dans le graphe des hypothèses générées par le reconnaisseur de parole. Le second problème, lié au taux de mots inconnus du lexique doit être résolu avant ou pendant la phase de transcription, car des approches *a posteriori* impliqueraient la localisation des frontières du mot inconnu avec précision.

### Modèle

L'objectif est de trouver le meilleur étiquetage $\hat{L}$ connaissant l'acoustique $A$. Il est intéressant de faire intervenir la séquence de mots $W$, portée par l'acoustique (équation 4.7) :

$$\begin{aligned} \hat{L} &= argmax_L P(L|A) \\ \hat{L} &\simeq argmax_{L,W} P(L,W|A) \\ \hat{L} &\simeq argmax_{L,W} P(A|W)P(W)P(L|W) \end{aligned} \tag{4.7}$$

Cette formulation laisse apparaître la probabilité de la séquence de mots issue de la transcription à travers $P(A|W)P(W)$ et la probabilité d'un étiquetage sur la séquence de

mots à travers $P(L|W)$. Ceci nous permet de modéliser les entités nommées séparément en utilisant un modèle appris sur des données textuelles sans nécessiter explicitement la connaissance du comportement des entités dans l'espace acoustique. Généralement, en reconnaissance automatique de la parole, la probabilité de la séquence entière est approximée en considérant que les dépendances entre éléments de la séquence sont limitées à un contexte proche (Markov). Cette approximation permet de résoudre le problème de maximisation en utilisant la programmation dynamique et, surtout, de représenter l'espace de recherche sous forme d'un graphe. Si la même approximation est faite pour la modélisation de l'étiquetage, il est possible de générer les hypothèses d'étiquetage par composition du modèle représenté sous la forme d'un transducteur pondéré ($\Omega_{P(L|W)}$, équation 4.8), avec le treillis de mots (accepteur pondéré) issu du reconnaisseur de parole ($\Omega_{P(A|W)P(W)}$).

$$\Omega_{P(L|A)} = \Omega_{P(A|W)P(W)} \oplus \Omega_{P(L|W)} \tag{4.8}$$

L'étiquetage d'entités nommées dans un flux de parole doit tenir compte des erreurs faites lors de la transcription. En effet, si l'étiquetage textuel utilisé fait entièrement confiance à la séquence de mots, une simple substitution peut faire apparaître ou disparaître une entité. Un bon compromis entre généralisation et précision doit être trouvé pour que l'algorithme soit robuste aux erreurs. Le modèle que nous proposons est constitué de deux composantes aux rôles bien distincts : une grammaire non contextuelle $\Omega_G$ dont le rôle est de contrôler la généralisation et un modèle $n$-gramme $\Omega_N$ qui permet de désambiguïser les cas incertains et de prendre en compte le contexte (équation 4.9).

$$\Omega_{P(L|W)} = \Omega_G \oplus \Omega_N \tag{4.9}$$

**Grammaire**

Dans le cadre d'une transcription à vocabulaire limité, aucun mot n'est inconnu lors de la phase d'étiquetage. Il devient possible d'attribuer à chacun des mots du lexique une classe ou un ensemble de classes spécifique aux entités dans lesquelles les mots peuvent apparaître. Par exemple, les mots ayant pour classe Prénom participent majoritairement à l'étiquetage des entités personne, mais peuvent faire partie de n'importe quelle entité impliquant un nom propre, comme une entité adresse.

Les entités nommées auxquelles nous nous intéressons sont généralement composées de noms propres mais peuvent également être formés par des séquences spécifiques comme des numéros de téléphone ou des dates. Ces dernières entités sont efficacement décrites par une grammaire car elles impliquent généralement des quantités numériques et des unités de mesure.

La grammaire proposée dans ces travaux est une grammaire non-contextuelle, régulière à droite, composée d'un ensemble de règles de réécriture impliquant des symboles terminaux $t$ (les mots du lexique) et des symboles non terminaux $NT$ (incluant les classes de mots et les entités nommées). Elle peut être exprimée sous forme du graphe

de l'ensemble des séquences reconnues. L'équation[8] 4.10 définit le transducteur $\Omega_G$ comme un ensemble de règles de réécriture.

$$\Omega_G = \{NT \leftarrow (NT|t)+\} \tag{4.10}$$

Pour être fonctionnelle, cette grammaire se voit imposer des contraintes précises : en tant que transducteur, elle doit prendre en entrée n'importe quelle séquence de mots et générer en sortie tous les étiquetages possibles, dont la séquence non étiquetée. Pour cela, l'axiome $S$ correspond à une alternative répétable de n'importe quelle entité $NE$ et d'un non-terminal spécial, appelé mange-mots $B_g$, qui accepte l'ensemble du vocabulaire (équation 4.11).

$$\begin{aligned} S &\leftarrow (B_g|NE)* \\ B_g &\leftarrow t+ \end{aligned} \tag{4.11}$$

Les étiquettes sont insérées lors de la transduction en ajoutant une transition, vide ($\varepsilon$) en entrée, génératrice d'un début d'étiquette $\langle tag \rangle$ avant le non-terminal correspondant à l'entité ; les fins d'étiquettes $\langle /tag \rangle$ sont insérées de la même façon après l'entité (équation 4.12). Des exemples de règles sont donnés dans la figure 4.4.

$$NE \leftarrow (\varepsilon \Rightarrow \langle tag \rangle)t + (\varepsilon \Rightarrow \langle /tag \rangle) \tag{4.12}$$

### Modèle N-grammes

La grammaire précédente permet de déterminer tous les étiquetages valides d'un treillis de mot, mais elle n'est pas capable de faire un choix entre ces étiquetages. Notamment, il faut choisir entre les différents étiquetages possibles d'une séquence (dont l'absence d'étiquetage).

Un modèle N-gramme (équation 4.13, $l_i$ est l'étiquette associée au mot $w_i$) approximant $P(L|W)$ a l'avantage de pouvoir être instancié sous forme d'un transducteur capable de probabiliser l'espace d'hypothèses. Pour que ce transducteur ne soit pas de taille exponentielle par rapport à la taille du vocabulaire, il est approximé grâce aux travaux de Mohri et al. (2002).

$$\begin{aligned} P(L|W) &= \frac{P(L,W)}{P(W)} \\ P(L,W) &= \prod_{l_i,w_i} P(l_i, w_i | l_{i-1}, w_{i-1}, \cdots, l_{i-n}, w_{i-n}) \end{aligned} \tag{4.13}$$

[8] Notations : « $\cdot \leftarrow \cdot$ » représente une règle de réécriture d'une partie droite par une partie gauche ; « $(\cdot)$ » est un regroupement ; « $\cdot+$ » est un opérateur de répétition (au moins une fois) ; « $\cdot*$ » est un opérateur de répétition (zéro ou plusieurs fois) ; « $\cdot|\cdot$ » représente l'alternative entre deux éléments ; $(a \Rightarrow b)$ sont les symboles d'entrée ($a$) et de sortie ($b$) portés par une transition d'un transducteur ; lorsque l'entrée et la sortie sont identiques, $(a \Rightarrow a)$ est remplacé par $a$ pour clarifier la notation ; $\varepsilon$ est une transition vide ne consommant pas de symbole d'entrée et/ou de sortie.

```
PERS_HUM:
($LEFT_CONTEXT_TITLE)? <pers.hum> $PERSON_NAME </pers.hum>
<pers.hum> $PERSON_NAME </pers.hum> ($RIGHT_CONTEXT_TITLE)?
$RELATIVE_TYPE de <pers.hum> $PERSON_NAME </pers.hum>

PERSON_NAME:
($FIRST_NAME)? ($FIRST_NAME)? (de|de la|du|le|des|les)?
     $FAMILY_NAME
$FIRST_NAME ($FIRST_NAME)?

LEFT_CONTEXT_TITLE:
monsieur|madame|mademoiselle|chef|président|présidente
     |responsable|chancelier|roi|reine|premier ministre
     |ministre|docteur

RIGHT_CONTEXT_TITLE:
(le|la)? (chef|président|présidente|responsable|chancelier
     |roi|reine|premier ministre|ministre|docteur)

RELATIVE_TYPE:
le (petit|beau)? fils|le (grand|beau)? père|la (petite|belle)?
     fille|la (grand|belle)? mère|le cousin|la cousine
     |l'oncle|la tante|la (demi|belle)? soeur|le
     (demi|beau)? frère
```

**FIG. 4.4:** *Exemple de règles contextuelles pour l'étiquetage des entités de type Personne, sous-type Humaine (pers.hum). La définition d'un non terminal débute par son nom en majuscules suivi de deux points ; une règle est définie par ligne ; dans la partie droite des règles, un non terminal est précédé du signe dollar ; le signe « | » représente une alternative ; un point d'interrogation désigne une partie facultative ; et les parenthèses facilitent le regroupement ; les débuts et fins d'étiquette sont exprimés par des balises ouvrantes et fermantes.*

Un modèle discriminant comme CRF pourrait certainement conduire à de meilleures performances, mais il est difficile à l'heure actuelle d'instancier ce modèle de façon efficace sous forme de transducteur. L'étiquetage en entités nommées est un problème de segmentation en plus d'être un problème d'étiquetage. Les étiquettes doivent avoir la même granularité que les mots, ce qui nécessite de mettre en place un étiquetage *Begin Inside Outside* (BIO), qui produit des sous-classes différentes pour les mots en début (B), milieu (I) ou hors entité (O).

#### Mélange avec les hypothèses de transcription

La méthode proposée pour l'annotation d'entités nommées dans le treillis d'hypothèses de transcription a l'avantage de permettre la mise en place de diverses techniques de fusion des hypothèses de transcription et d'étiquetage. Les modèles peuvent

être fusionnés en imposant un facteur de normalisation $\alpha$ entre les composantes issues de la transcription et de l'étiquetage (équation 4.14).

$$P(L|A) \simeq \prod_i p(a_i|w_i)p(w_i)p(l_i|w_i)^{\alpha} \tag{4.14}$$

L'hyperparamètre $\alpha$ est alors déterminé empiriquement en utilisant un corpus d'apprentissage. Ce type de fusion est appliqué entre les espaces acoustiques et linguistiques dans un système de transcription et permet de minimiser le taux d'erreur de mots moyen. Il n'est pas intuitif dans le cas de l'étiquetage car il s'agit de mélanger des événements de nature différente, dont les distributions de probabilités sont inférées dans des conditions et sur des données généralement différentes.

Un autre type de fusion utilise la probabilité *a posteriori* de transcription du *support* de chaque hypothèse d'étiquetage (équation 4.15). Le *support* d'une hypothèse d'étiquetage est formé du sous-graphe de mots de l'espace d'hypothèses qui produit une séquence d'étiquettes donnée. Cette probabilité est souvent utilisée comme mesure de confiance dans les tâches fondées sur la reconnaissance de la parole (Falavigna et al., 2002).

$$P(L|A) \simeq \frac{\sum_{W_L} P(A|W_L)P(W_L)}{\sum P(A|W)P(W)} P(L|W) \tag{4.15}$$

Enfin, lorsque la tâche permet de construire un modèle du type d'entité attendu, il est possible de filtrer le graphe d'hypothèses afin de construire la liste des $n$-meilleures séquences de mots correspondant à chaque type d'entité. La transcription n'est plus dans ce cas guidée par la maximisation de la probabilité d'une hypothèse, mais par la tâche post-transcription elle-même.

### 4.2.3 Performances

Cette section est dédiée à l'évaluation de la méthode proposée sur les données de la campagne ESTER.

#### Système de comparaison Lingpipe

Lingpipe[9], un système d'étiquetage en entités nommées optimisé pour l'annotation du texte, a été choisi pour établir des comparaisons avec l'approche utilisant le graphe d'hypothèses de transcription. Ce système offre l'avantage d'utiliser une modélisation HMM, proche de celle présentée dans ces travaux (équation 4.16, dans laquelle $L$ est

[9]Disponible sur http://www.alias-i.com/lingpipe/ visité en août 2006

une séquence d'étiquettes $l_i$, $W$ est une séquence de mots $w_i$).

$$\begin{aligned} \hat{L} &= \underset{L}{\operatorname{argmax}}\, P(L, W) \\ P(L, W) &= \prod_{n=0}^{N} P(w_n, l_n | w_{n-2}, w_{n-1}, l_{n-1}) \\ P(w_n, l_n | w_{n-2}, w_{n-1}, l_{n-1}) &= P(l_n | w_{n-2}, w_{n-1}, l_{n-1}) P(w_n | w_{n-2}, w_{n-1}, l_{n-1}, l_n) \end{aligned} \tag{4.16}$$

Lingpipe utilise une sous-classification BIO pour différencier les début, milieu et fin d'entité. Cette notation permet l'utilisation de modèles communs pour chaque entité pour remplacer les étiquettes conditionnant les probabilités. Finalement, un repli de Witten-Bell (Witten et Bell, 1991) et un lissage par une loi uniforme affinent la qualité du modèle.

Étant donné que les corpus disponibles pour l'apprentissage des paramètres de l'extracteur d'entités nommées ne sont pas très fournis, un processus de généralisation sur les mots inconnus est nécessaire. Lingpipe effectue cette généralisation sur les mots dont la fréquence est faible dans le corpus d'apprentissage. Ils sont remplacés par des classes qui prennent en compte leurs caractéristiques morphologiques (table 4.5).

| Classe | Description |
|---|---|
| 1-DIG | composé d'un seul chiffre |
| 2-DIG | composé d'exactement 2 chiffres |
| 3-DIG | composé d'exactement 3 chiffres |
| 4-DIG | composé d'exactement 4 chiffres |
| 5+-DIG | composé de 5 chiffres ou plus |
| DIG-LET | composé de chiffres et lettres |
| DIG– | composé de chiffres et tirets |
| DIG-/ | composé de chiffres et barres oblique |
| DIG-, | composé de chiffres et virgules |
| DIG-. | composé de chiffres et points |
| 1-LET-UP | composé d'une seul lettre majuscule |
| 1-LET-LOW | composé d'une seule lettre minuscule |
| LET-UP | composé uniquement de lettres majuscules |
| LET-LOW | composé uniquement de lettres minuscules |
| LET-CAP | commence par une majuscule suivie par des minuscules |
| LET-MIX | contient des majuscules et des minuscules |
| PUNC- | ponctuations |
| OTHER | tout le reste |

**TAB. 4.5:** *Les classes morphologiques utilisées par Lingpipe pour généraliser les mots à faible fréquence. La classification est appliquée dans l'ordre des règles.*

Lingpipe est utilisé dans de nombreux travaux car il est facile à mettre en œuvre et couvre de nombreuses tâches. Il est utilisé en question/réponse (Chen et al., 2004; Neumann et Sacaleanu, 2004), en résumé automatique (Schilder et al., 2006), et en résolution

d'anaphores (Vlachos et al., 2006). Ses performances ont été évaluées sur CoNLL 2002, où une $f_1$-mesure de 0.77 lui permet d'atteindre la troisième[10] place sur 14 participants.

### Évaluation ESTER

| Type | Corpus Test | | Corpus Dév. | |
|---|---|---|---|---|
| | Nb. | % | Nb. | % |
| Personne | 1662 | 25.2 | 1689 | 27.2 |
| Lieu | 166 | 2.5 | 155 | 2.5 |
| Organisation | 1001 | 15.1 | 839 | 13.5 |
| GSP | 1794 | 27.2 | 1624 | 26.1 |
| Quantité | 501 | 7.5 | 337 | 5.5 |
| Temps | 1071 | 16.3 | 1245 | 20.1 |
| Produit | 286 | 4.3 | 212 | 3.5 |
| Construction | 125 | 1.9 | 99 | 1.6 |
| Total | 6606 | 100.0 | 6200 | 100.0 |

**TAB. 4.6:** *Distribution des types d'entités dans le corpus de test ESTER. Les entités les plus nombreuses sont les groupes géo-socio-politiques (GSP), les noms de personnes, les références temporelles et les organisations.*

Une tâche expérimentale de reconnaissance des entités nommées a été introduite dans la campagne d'évaluation ESTER. Elle permet de mesurer les performances des systèmes d'extraction d'entités nommées à partir d'un flux de parole. Cette campagne est la seule organisée à l'heure actuelle sur le français parlé. Les types d'entités à reconnaître sont les suivants (leur distribution est détaillée dans la table 4.6) :

1. Personne : humaine, fictive, animal familier ;
2. Lieu : géographique, voie de communication, adresse physique et électronique, numéro de téléphone ;
3. Organisation : politique, commerciale, à but non lucratif ;
4. Groupe géo-socio-politique (GSP) : clan, famille, nation, région administrative ;
5. Quantité : durée, devise, longueur, température, âge, poids, vitesse ;
6. Temps : relatif, absolu, heure ;
7. Produit : œuvre artistique, journal, récompense, véhicule ;
8. Construction : bâtiment, monument.

L'évaluation présente de nombreuses difficultés dues à la nature du média, la diversité des classes et la qualité de l'annotation qui en résulte. D'une part, les malformations du discours oral, comme les hésitations, les reprises ou les confusions, ont un impact sur les entités nommées et doivent être annotées. D'autre part, la classification choisie introduit des ambiguïtés sur le choix de la classe, entre par exemple le temps (il y a 5

[10]Détails sur http://www.alias-i.com/lingpipe/demos/tutorial/ne/read-me.html, visité en septembre 2006.

heures) et la durée (pendant 5 heures), le lieu (place de la Tour Eiffel) et la construction (la Tour Eiffel). Le phénomène est accentué par les métonymies fréquentes qui nécessitent la séparation du rôle et de la nature de l'entité. En effet, un nom de pays utilisé pour représenter son peuple (la France part en vacances) doit-il être annoté comme GSP, ou comme Organisation ? Même si tous ces cas particuliers sont décrits et pris en compte dans le guide d'annotation de la campagne, ils ont néanmoins tendance à abaisser la qualité de l'annotation car les annotateurs ne sont pas toujours cohérents entre eux pour sur l'interprétation du guide. Au niveau des métonymies par exemple, les modèles construits sur le contexte d'utilisation des entités doivent prendre en compte l'ambiguïté constante des classes. Un dernier problème affecte plus particulièrement la qualité de la transcription au niveau des entités nommées. En effet, un décalage temporel de 6 mois différencie les données d'apprentissage de celles de test. Ce décalage implique que les entités dont parle l'actualité ont changé et sont moins bien reconnues par les systèmes de transcription avec le risque non moindre que des noms propres fréquents n'existent pas dans le lexique du reconnaisseur.

| **Données** | **Système** | **Corpus Dév.** | | | **Corpus Test** | | |
|---|---|---|---|---|---|---|---|
| | | SER | $f_1$ | WER | SER | $f_1$ | WER |
| Réf. | $S_{texte}$ | 21 | 0.84 | 0.0 | 27 | 0.79 | 0.0 |
| | $S_{audio}$ | 22 | 0.84 | | 34 | 0.74 | |
| Trans. | $S_{texte}$ | 42 | 0.72 | 21.2 | 55 | 0.63 | 26.4 |
| | $S_{audio}$ | 41 | 0.73 | | 54 | 0.63 | |

**TAB. 4.7:** *Comparatif des performances du système Lingpipe ($S_{texte}$) et de l'approche proposée ($S_{audio}$) sur la tâche d'extraction d'entités nommées ESTER. Les systèmes sont comparés sur les transcriptions de référence (Ref.) et sur les sorties du système de transcription (Trans.), en terme de Slot Error Rate (SER) et $f_1$-mesure ($f_1$). Les performances sont comparées entre le corpus de développement (Dév.) et le corpus de test (Test). L'augmentation du taux d'erreur de mots (WER) entre ces deux corpus implique une forte diminution des performances en détection des entités nommées. Cette diminution est expliquée par une différence de 6 mois entre les corpus qui induit un changement des thèmes d'actualités et donc des entités nommées concernées. On observe toutefois que le système adapté à l'audio est au moins aussi bon que Lingpipe sur les donnée transcrites.*

Les performances du système proposé et de Lingpipe selon diverses conditions d'expérimentation sont proposées dans la table 4.7. Les mesures de performances sont le *Slot Error Rate* (SER) et la $F_1$-mesure. Lingpipe est un système optimisé pour l'annotation de texte et il obtient logiquement de meilleures performances sur la transcription de référence. Par contre, l'approche proposée est meilleure dans le cas où la transcription provient de la chaîne de traitement Speeral, avec un taux d'erreur de mots (WER) de plus de 20%. Néanmoins, l'écart entre l'utilisation des données de référence et la transcription automatique est important : cette relation est de l'ordre de 1.9 points de $f_1$-mesure perdus (et 1.1 points de SER en plus) pour 1% de taux d'erreur de mots supplémentaire. Cette observation peut être comparée à celle de Miller et al. (2000) selon laquelle $\Delta\text{WER} = -0.7\Delta f_1$-mesure sur l'évaluation Hub-4 de NIST (Przybocki et al., 1998). La différence s'explique par l'inadaptation du système spécialisé sur la parole sur des données propres (avec 2.4 points de $f_1$-mesure perdus par point de WER sur le corpus de test). La table 4.8 recense les taux d'erreur (SER) pour chaque type d'en-

tité nommée. Les groupes géo-socio-politiques (GSP) sont bien reconnus et nombreux dans le corpus (en général des noms de pays). Par contre, les catégories moins nombreuses, comme les Produits et les Constructions connaissent des taux d'erreur élevés. Ces performances sont explicables par l'absence de caractères morphologiques spécifiques à ces catégories et la diversité des formes rencontrées. Une comparaison entre l'annotation de la transcription de référence et l'annotation des sorties du système de transcription montre que les entités nommées contenant des noms propres sont plutôt mal reconnues. Cette chute s'explique par une trop forte généralisation sur les erreurs de transcription et la mauvaise modélisation des noms propres dans les modèles de transcription.

Bien que cette évaluation reste une tâche expérimentale[11] d'ESTER, il faut noter que notre approche est la plus performante parmi les trois participants à la tâche d'extraction d'entités nommées. De plus, comparé à une approche ne remettant pas en cause la transcription, il est possible de rechercher explicitement des entités nommées dans le graphe d'hypothèses de transcription lorsque la tâche visée apporte des informations sur le type d'entité recherchée.

| **Type** | **Corpus Dév.** | | **Corpus Test** | |
|---|---|---|---|---|
| | Trans. | Réf. | Trans. | Réf. |
| Personne | 43.9 | 21.7 | 69.1 | 32.3 |
| Lieu | 60.5 | 44.9 | 67.2 | 55.2 |
| Organisation | 46.8 | 32.6 | 67.9 | 50.9 |
| GSP | 26.4 | 9.6 | 36.5 | 11.1 |
| Quantité | 54.8 | 36.7 | 68.2 | 49.7 |
| Temps | 33.6 | 20.3 | 51.9 | 37.9 |
| Produit | 80.5 | 56.5 | 86.3 | 72.8 |
| Construction | 70.6 | 56.9 | 91.9 | 65.6 |

**TAB. 4.8:** *Performances du système proposé sur la tâche d'étiquetage en entités nommées d'ESTER. Les résultats sont donné en Slot Error Rate (SER), sur les données de développement (Dév.) et de test (Test), selon que l'annotation est réalisée sur la transcription de référence (Réf.) ou les sorties du système Speeral (Trans.). On observe que les entités les mieux reconnues sont les GSP et que les entités à base de noms propres subissent le plus gros impact lors de la transcription.*

Toujours selon la table 4.7, il existe une forte différence de performances entre le corpus de développement et le corpus de test. La perte de 5% au niveau du taux d'erreur de mots se traduit par une perte de plus de 10% en SER sur l'annotation. Cette baisse de performances s'explique par un décalage temporel de 6 mois entre les deux corpus. Les changements de thèmes d'actualité et des noms propres les plus fréquents sont les principales explications de ce phénomène. Par exemple, les données de test couvrent la libération de George Malbruno et Christian Chesnaut, otages des forces irakiennes, personnes dont parlent très peu les médias dans le corpus de développement. Afin de pallier ce genre de problème, il est essentiel de créer un modèle de l'actualité correspondant à l'époque du corpus de test et d'introduire ces informations dans les processus

[11] La qualité de l'annotation de référence et le protocole d'évaluation ont évolué jusqu'à la dernière minute.

de transcription et d'annotation.

### 4.2.4 Limites

L'étiquetage des entités nommées d'un flux audio permet de favoriser les zones à forte densité d'informations dans un espace informatif et de présenter à l'utilisateur des indicateurs qualifiés pertinents sur les entités impliquées dans un sujet ou un thème. Néanmoins, sans une résolution complète des coréférences (déterminer que *il*, *le président Américain*, *Mr Bush*, *Georges Bush*, *Georges W. Bush* et *Georges Walker Bush* font tous référence à la même personne), il est impossible de raisonner sur le contenu informatif afin d'en déduire des informations précises sur le fond. De plus, dans une application de recherche d'information, des personnes génériques (les pompiers...) sont peut être aussi importantes que les personnes nommées.

Il faut aussi remarquer que les algorithmes à base d'apprentissage souvent utilisés pour l'annotation de séquences nécessitent une forte qualité des références construites manuellement. En effet, les définitions fines des catégories d'objets à retrouver mènent à des interprétations différentes par chaque annotateur et à des corpus d'autant plus ambigus. Un exemple simple est l'annotation de *La tour Eiffel* comme lieu géographique, bâtiment ou personne (Gustave Eiffel). Bien que le guide d'annotation ESTER essaye de contourner et limiter ces problèmes, leur nombre reste élevé et rend l'évaluation de l'annotation en entités nommées compliquée et laborieuse.

Le modèle choisi dans ces travaux a l'avantage d'être modulaire et de s'appuyer sur des outils existants. Toutefois, on peut imaginer l'instanciation d'autres modèles restant dans le même cadre (automates à états finis), comme les CRF, utilisés dans ces travaux pour trouver des frontières de phrases.

La coopération entre la transcription et l'extraction de descripteurs sémantiques ne s'arrête pas à un espace de recherche commun. Nous avons essayé dans (Favre et al., 2005) de mettre à jour les modèles de langage du système de transcription en utilisant des données externes sélectionnées pour leur recouvrement avec les entités retrouvées. Bien qu'aucune amélioration du WER n'a été observée, cette approche améliore de 10% la quantité d'entités potentiellement extraites (car présentes dans les hypothèses de transcription).

## 4.3 Conclusion

Nous venons d'illustrer l'extraction de descripteurs structurels et sémantiques à travers la segmentation en phrases et la détection d'entités nommées. La section 4.1 a présenté une méthode de segmentation en phrases fondée sur CRF par l'interaction entre des paramètres prosodiques et linguistiques. Cette segmentation admet une $F_1$-mesure de 0.70 dans des conditions réelles sur des émissions radiophoniques en français. La section 4.2, pour sa part, s'est concentrée sur une plus grande collaboration entre les processus de transcription et d'annotation en entités nommées. Une $F_1$-mesure de 0.63

a pu être obtenue sur les mêmes données, sans toutefois améliorer de beaucoup un système optimisé pour le texte.

De nombreux autres descripteurs peuvent être extraits par des méthodes similaires. Toutefois, ces tâches nécessitent une définition rigoureuse pour obtenir des données d'apprentissage les plus cohérentes possibles. L'amélioration de telles méthodes passera alors par une amélioration des algorithmes d'apprentissage impliqués, par une amélioration des paramètres extraits (limitation de la variabilité de la parole), et, surtout, par l'augmentation de la quantité de données d'apprentissage. Le prochain challenge est de s'affranchir de chacune de ces contraintes, en recherchant un modèle unique pour une structuration non-supervisée à toutes les granularités, facilitant l'extraction de concepts représentatifs d'une sémantique à l'échelle humaine.