TALN 2010, Montréal, 19–23 juillet 2010

Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement

Christian Raymond[1, 2] Julien Fayolle[1]
(1) Université Européenne de Bretagne, INRIA,IRISA, UMR 6074, France
(2) INSA de Rennes, 20 Avenue des buttes de coesme, Rennes, France
prénom.nom@irisa.fr

Résumé. Les transcriptions automatiques de parole constituent une ressource importante, mais souvent bruitée, pour décrire des documents multimédia contenant de la parole (*e.g.* journaux télévisés). En vue d'améliorer la recherche documentaire, une étape d'extraction d'information à caractère sémantique, précédant l'indexation, permet de faire face au problème des transcriptions imparfaites. Parmis ces contenus informatifs, on compte les entités nommées (*e.g.* noms de personnes) dont l'extraction est l'objet de ce travail. Les méthodes traditionnelles de reconnaissance basées sur une définition manuelle de grammaires formelles donnent de bons résultats sur du texte ou des transcriptions propres manuellement produites, mais leurs performances se trouvent fortement affectées lorsqu'elles sont appliquées sur des transcriptions automatiques. Nous présentons, ici, trois méthodes pour la reconnaissance d'entités nommées basées sur des algorithmes d'apprentissage automatique : les champs conditionnels aléatoires, les machines à de support, et les transducteurs à états finis. Nous présentons également une méthode pour rendre consistantes les données d'entrainement lorsqu'elles sont annotées suivant des conventions légèrement différentes. Les résultats montrent que les systèmes d'étiquetage obtenus sont parmi les plus robustes sur les données d'évaluation de la campagne ESTER 2 dans les conditions où la transcription automatique est particulièrement bruitée.

Abstract. Automatic speech transcripts are an important, but noisy, ressource to index spoken multimedia documents (*e.g.* broadcast news). In order to improve both indexation and information retrieval, extracting semantic information from these erroneous transcripts is an interesting challenge. Among these meaningful contents, there are named entities (*e.g.* names of persons) which are the subject of this work. Traditional named entity taggers are based on manual and formal grammars. They obtain correct performance on text or clean manual speech transcripts, but they have a lack of robustness when applied on automatic transcripts. We are introducing, in this work, three methods for named entity recognition based on machine learning algorithms, namely conditional random fields, support vector machines, and finite-state transducers. We are also introducing a method to make consistant the training data when they are annotated with slightly different conventions. We show that our tagger systems are among the most robust when applied to the evaluation data of the French ESTER 2 campaign in the most difficult conditions where transcripts are particularly noisy.

Mots-clés : étiqueteur d'entités nommées, transcription automatique de parole, apprentissage automatique, champs conditionnels aléatoires, machines à vecteurs de support, transducteurs à états finis.

Keywords: named entity tagger, automatic speech recognition transcripts, machine learning, conditionnal random fields, support vector machines, finite-state transducers.

1 Introduction

La transcription de flux audio (enregistrements radiophoniques, de réunions, de journaux télévisés, *etc.*) est un enjeu important pour le domaine de l'archivage et de la recherche d'information. L'extraction automatique de contenus à valeur ajoutée à partir de ces transcriptions devient un axe de recherche primordial afin d'utiliser et d'exploiter le maximum d'information contenu dans le flux audio. Parmi ces contenus à valeur ajoutée, sont souvent considérées les entités nommées. La plupart des systèmes d'étiquetage en entités nommées utilisent des méthodes symboliques à base de grammaires formelles, éventuellement complétées par des connaissances *a priori* (*e.g.* listes de prénoms, de villes ou de pays). Dans les grandes campagnes d'évaluation, ces systèmes implémentés manuellement obtiennent les meilleurs résultats sur le texte propre (texte ou transcription manuelle de parole) (voir [1] pour la campagne ESTER 2). Lorsque la reconnaissance d'entités nommées se fait sur des transcriptions automatiques de parole, le problème gagne en difficulté car contrairement aux documents textuels, les documents transcrits automatiquement ne sont pas structurés (ni casse, ni ponctuation) et certains mots transcrits sont erronés : le taux d'erreur de mots peut varier de 5% à plus de 50% selon le document et les conditions de transcriptions. Dans ces conditions, les systèmes symboliques sont généralement moins robustes que des étiqueteurs basés sur des méthodes d'apprentissage automatique, notamment car elles sont capables d'extraire de ces données des règles de décision qu'un expert humain n'aurait pu appréhender. Guidés par cette notion de robustesse face aux transcriptions automatiques, nous présentons trois systèmes d'étiquetage basés sur différents algorithmes de classification automatique qui ont déjà fait leurs preuves dans la tâche de reconnaissance en entités nommées (voir respectivement [8, 5, 3]), un système à base de champs conditionnels aléatoires (CRF), un à base de machines à vecteurs de support (SVM), et un à base de transducteurs à états finis (FST). Dans la partie 2, nous présentons l'approche générale utilisée en reconnaissance d'entités nommées par des méthodes à base d'apprentissage automatique. Seront ensuite présentées dans la partie 3, les méthodes d'apprentissage utilisées dans ce travail. Enfin, la partie 4 présentera les données d'évaluations et la partie 5 les expériences effectuées ainsi que les résultats obtenus.

2 Reconnaissance d'entités nommées

La reconnaissance d'entités nommées consiste à rechercher des objets textuels (*i.e.* un mot, ou un groupe de mot) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, *etc.* C'est un problème typique d'étiquetage de séquences dont le but est, pour une séquence donnée, de trouver la séquence d'étiquettes correspondante la plus probable. Dans notre cas, la séquence donnée est une séquence de mots issus de la transcription de parole, et la séquence d'étiquettes recherchée est la séquence d'entités nommées correspondante. Pour résoudre le double problème de la segmentation et de l'étiquetage (*i.e.* trouver l'entité ainsi que ses frontières dans la séquence de mots), l'encodage BIO (pour begin, inside, outside) est traditionnellement utilisé. Un indicateur B, I ou O est ajouté à l'étiquette afin d'identifier si le mot correspondant est au début, à l'intérieur ou à l'extérieur de l'entité nommée. Un exemple est donné table 1.

Les méthodes à base d'apprentissage automatique utilisent des données annotées (en général par des experts humains) pour construire automatiquement des règles de décision à partir d'un ensemble de descripteurs. Les mots de la transcription sont déjà un premier niveau de description. Afin de construire des systèmes plus performants, d'autres niveaux sont généralement envisagés.

mots	ici	jacques	doutisoro	lomé	africa	numéro	un
étiquettes	O	pers-B	pers-I	loc-B	org-B	org-I	org-I
entités	null	pers		loc		org	

TABLE 1 – Exemple d'étiquetage de séquences appliqué à la reconnaissance d'entités nommées. On y retrouve les mots de la transcription automatique à étiqueter, les étiquettes trouvées suivant l'encodage BIO et les entités nommées correspondantes.

Par exemple, le résultat d'un étiqueteur morpho-syntaxique peut être utilisé pour construire des règles à portée plus générale et ainsi améliorer le rappel du système, ou bien des connaissances *a priori* fortes peuvent être intégrées pour améliorer la précision ainsi que le rappel.

niveau	**type**	**exemple**
premier	MOT : mot	"Jacques"
second	MS : étiquette morpho-syntaxique	"NPMS", nom propre masculin singulier
	AP : classe connu *a priori*	"VILLE"
	MI : mot "important"	"numéro"

TABLE 2 – Niveaux de description

Ici, deux niveaux de description (table 2) sont utilisés. Le premier est directement composé des mots (MOT) de la transcription et le second peut être des trois types suivants :
- MS : résultat d'un étiquetage morpho-syntaxique [4],
- AP : classe de généralisation correspondant à des connaissances connues *a priori*, *i.e.* listes de pays, de villes, de gentilés, d'unités de mesure,
- MI : mot "important" dont l'information mutuelle partagée avec son étiquette d'entité nommée est supérieure à zéro (*i.e.* mot supposé plus discriminant qu'une étiquette morpho-syntaxique) et qui apparaît au moins trente fois dans le corpus d'apprentissage (*i.e.* mot à capacité de généralisation suffisante).

Comme illustré sur la figure 1, l'étiquette courante est estimée à partir des descripteurs (mots et classes) situés dans la fenêtre locale $[-2, +2]$ entourant la position 0 de décision. On y retrouve les trois types de classe du second niveau de description, à savoir les étiquettes morpho-syntaxiques en rouge, les classes connues *a priori* en bleu, et les mots "importants" en vert.

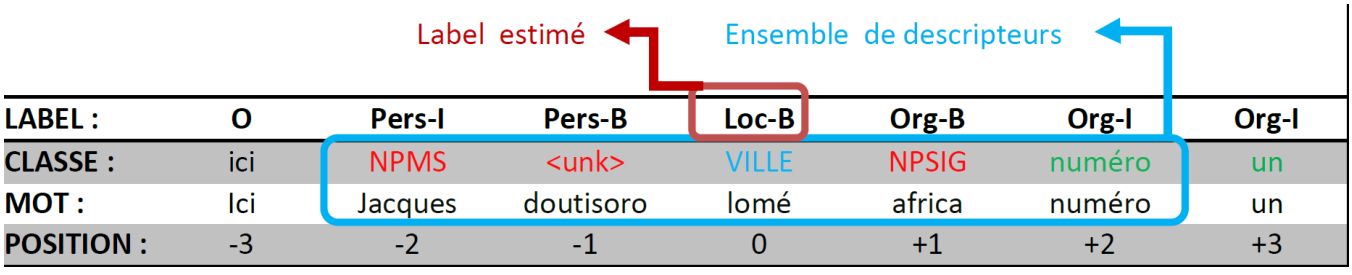

FIGURE 1 – Exemple d'étiquetage en entités nommées à partir des descripteurs de premier et second niveaux

3 Algorithmes d'apprentissage automatique

3.1 Machines à vecteurs de support

Les machines à vecteurs de support introduites par Vapnik [11], couramment abrégées en SVM sont des classifieurs discriminants à large marge. Les SVM sont au départ des classifieurs binaires qui représentent les échantillons à classer sous la forme d'un vecteur dont chaque composante représente la contribution d'un paramètre à un exemple. Par exemple, pour une tâche de classification de documents les vecteurs représentant chaque document pourraient avoir la taille du vocabulaire associé aux documents et chaque composante du vecteur serait nulle ou non nulle selon que le mot correspondant est absent ou non du document en question. Le principe est alors de déterminer l'hyperplan séparateur optimal entre les deux classes (si le problème est linéairement séparable), celui qui maximise la marge entre les échantillons et l'hyperplan séparateur. La marge est la distance entre l'hyperplan et les échantillons le plus proches : appelés "vecteurs de support". Si la méthode ne fonctionne que si le problème est linéairement séparable, grâce aux fonctions noyaux, les SVM sont capables de considérer le problème dans un espace de dimension plus élevé dans lequel il existe probablement un séparateur linéaire.

Deux méthodes principales ont été proposées pour étendre la classification binaire au cas où l'on a M classes :

– La méthode **one-versus-all** consiste à construire M classifieurs binaires en attribuant le label 1 aux échantillons de l'une des classes et le label -1 à toutes les autres. En phase de test, le classifieur donnant la valeur de confiance (*e.g.* la marge) la plus élevée remporte le vote.
– La méthode **one-versus-one** consiste à construire $M(M-1)/2$ classifieurs binaires en confrontant chacune des M classes. En phase de test, l'échantillon à classer est analysé par chaque classifieur et un vote majoritaire permet de déterminer sa classe.

Bien que les SVM permettent l'utilisation de paramètres très variés, contrairement aux algorithmes spécifiquement connus pour l'étiquetage séquentiel, ils ne peuvent prendre de décision globale sur la séquence car chaque étiquette de la séquence est vue indépendamment des autres. Toutefois, certaines heuristiques peuvent être implémentées, par l'exemple l'ajout d'un paramètre de classification qui serait la décision précédente dans la séquence. YAMCHA, un système basé sur cette approche, a obtenu les meilleurs résultats dans la tâche de chunking et BaseNP chunking de CoNLL2000 [6] et a été choisi pour l'implémentation de l'étiqueteur SVM.

Dans ce travail, le vecteur de chaque exemple est composé des couples mots ou/et classes associés à leur position par rapport à la position de décision dans un intervalle local $[-2, +2]$ (figure 1).

3.2 Transducteurs à états finis

L'approche à base de transducteurs à états finis est une approche générative stochastique basée sur le calcul de la probabilité jointe entre la séquence d'observations (mots) et la séquence d'étiquettes (entités nommées). Cette approche est particulièrement appropriée pour traiter des transcriptions de parole [3] puisqu'elle est basée sur le même paradigme traditionnellement utilisée dans les systèmes de reconnaissance automatique de la parole. Plus formellement, notons $\mathbf{e} = e_1, e_2, \ldots, e_N$ la séquence d'étiquettes associées à la séquence de mots $\mathbf{m} = m_1, m_2, \ldots, m_N$ produite par un système de reconnaissance automatique de la parole. Le processus d'étiquetage consiste à trouver la séquence d'étiquettes maximisant la

probabilité *a posteriori* $p(\mathbf{e}|A)$, où A représente les observations acoustiques extraites du signal de parole. Pour résoudre ce problème, il est commode de faire intervenir des connaissances supplémentaires tels que des classes d'équivalence (présentées dans la partie 2). Notons $\mathbf{c} = c_1, c_2, \ldots, c_N$ la séquence de classes. Ainsi, trouver la meilleure séquence d'étiquettes $\hat{\mathbf{e}}$ étant données les observations acoustiques A se formule par :

$$\begin{aligned}
p(\hat{\mathbf{e}}|A) &= \underset{\mathbf{e}}{\operatorname{argmax}} \sum_{\mathbf{c}} \sum_{\mathbf{m}} p(\mathbf{m}, \mathbf{c}, \mathbf{e}|A) \\
&= \underset{\mathbf{e}}{\operatorname{argmax}} \sum_{\mathbf{c}} \sum_{\mathbf{m}} p(A|\mathbf{m}, \mathbf{c}, \mathbf{e}) p(\mathbf{m}, \mathbf{c}, \mathbf{e}) \\
&\approx \underset{\mathbf{m},\mathbf{c},\mathbf{e}}{\operatorname{argmax}} p(A|\mathbf{m}, \mathbf{c}, \mathbf{e}) p(\mathbf{m}, \mathbf{c}, \mathbf{e}) \\
\text{où} & \\
p(A|\mathbf{m}, \mathbf{c}, \mathbf{e}) &\approx p(A|\mathbf{m}) \\
p(\mathbf{m}, \mathbf{c}, \mathbf{e}) &= p(\mathbf{m}|\mathbf{c}, \mathbf{e}) p(\mathbf{c}, \mathbf{e}) \\
\text{alors} & \\
p(\hat{\mathbf{e}}|A) &\approx \underset{\mathbf{m},\mathbf{c},\mathbf{e}}{\operatorname{argmax}} p(A|\mathbf{m}) p(\mathbf{m}|\mathbf{c}, \mathbf{e}) p(\mathbf{c}, \mathbf{e}) \qquad (1)
\end{aligned}$$

La probabilité $p(A|\mathbf{m})$ est estimée par le modèle acoustique du système de reconnaissance automatique de la parole. $p(\mathbf{m}|\mathbf{c}, \mathbf{e})$ est la probabilité d'une séquence de mots sachant le couple classe/étiquette et $p(\mathbf{c}, \mathbf{e})$ la probabilité jointe estimée sur les couples classe/étiquette.

$p(\mathbf{m}|\mathbf{c}, \mathbf{e})$ est estimée de la manière suivante :

$$p(\mathbf{m}|\mathbf{c}, \mathbf{e}) = \prod_{i=1}^{N} \frac{cooccurence(m_i, c_i, e_i)}{cooccurence(c_i, e_i)} \qquad (2)$$

Les problèmes de segmentation et de classification sont résolus simultanément à travers l'utilisation de l'encodage BIO (table 1). À chaque mot m_i est alors associée l'étiquette t_i, $t_i = \{e_i\text{-[B|I]}, \mathrm{O}\}$. La probabilité jointe est alors calculée par :

$$p(\mathbf{c}, \mathbf{e}) = p\{(t_1, c_1)(t_2, c_2) \ldots (t_n, c_n)\} = p(\mathbf{t}, \mathbf{c})$$

Cette probabilité est estimée par un modèle *N*-gramme du troisième ordre :

$$\begin{aligned}
p(\mathbf{c}, \mathbf{e}) &= \prod_{n=1}^{N} p(c_n, t_n|h_n) \qquad (3) \\
\text{avec } h_n &= (c_{n-1}, t_{n-1}), (c_{n-2}, t_{n-2})
\end{aligned}$$

Le processus d'étiquetage est généralement effectué pour une séquence de mots **m** fixée (*i.e.* la meilleure hypothèse de transcription automatique). L'originalité de cette approche est de pouvoir s'intégrer directement dans le processus de reconnaissance automatique de la parole. En effet, elle permet de réaliser l'étiquetage directement sur des graphes de mots, à condition que ceux-ci soient encodés comme des automates à états finis. L'implémentation de cette approche a été réalisé avec la librairie AT&T [9].

La meilleure séquence de couples mot/étiquette est le meilleur chemin dans le transducteur λ_{m2e} obtenu par composition de trois transducteurs : $\lambda_{m2e} = \lambda_m \circ \lambda_{m2ce} \circ \lambda_{ce}$

Les trois transducteurs sont définis de la manière suivante :

1. λ_m est la représentation de l'entrée à étiqueter sous forme d'automate à états finis (hypothèse ou graphe de mots généré par le moteur de reconnaissance de la parole avec les scores acoustiques, $p(A|\mathbf{m})$ dans la formule 1). Dans les expériences suivantes, dans le but de rester comparable avec les autres méthodes, λ_m encode la meilleure hypothèse de reconnaissance ;
2. λ_{m2ce} fait l'association entre les mots et leurs couples classe/entité, les classes peuvent être le résultat d'un étiquetage morpho-syntaxique ou/et des classes représentant des connaissances *a priori* sur les mots (*e.g.* liste de pays, de ville) ou/et les mots eux-mêmes. Le transducteur possède alors en entrée les mots et en sortie les couples classe/entité. Les scores associés aux transitions encodent $p(\mathbf{m}|\mathbf{c},\mathbf{e})$ dans la formule 1 et sont calculés selon 2 ;
3. λ_{ce} encode le modèle estimant la probabilité jointe étiquette/entité décrite dans la formule 4.

3.3 Champs Conditionnels Aléatoires

Les champs conditionnels aléatoires, introduits par [7], possèdent les avantages des modèles génératifs et discriminants. Comme les classifieurs discriminants, ils peuvent manipuler un grand nombre de descripteurs et comme les modèles génératifs, ils intègrent des dépendances entre les étiquettes de sortie et prennent une décision globale sur la séquence. Cependant, ils ne sont pas facilement intégrables avec le système de reconnaissance automatique de la parole (*e.g.* analyse d'un graphe de mots).

Un champ conditionnel aléatoire est défini par un graphe de dépendances et un ensemble de fonctions f_k auxquelles sont associées des poids λ_k. La probabilité conditionnelle d'une annotation, séquence d'étiquettes **e** étant donné l'observation O (*i.e.* mots, étiquettes morpho-syntaxiques) est donnée par :

$$\begin{aligned} p(\mathbf{e}|O) &= \frac{1}{Z(O)} exp(\sum_{c\in C}\sum_{k}\lambda_k f_k(e_c, O, c)) \\ avec & \\ Z(O) &= \sum_{e} exp(\sum_{c\in C}\sum_{k}\lambda_k f_k(e_c, O, c)) \end{aligned}$$

Les connaissances, les descripteurs (lexicaux, sémantique, *etc.*), et les relations entre concepts sont encodés dans le modèle à travers ces fonctions. Ces fonctions binaires retournent 1 s'il y a correspondance ou 0 sinon. Elles prennent en paramètre les valeurs prises par les variables aléatoires (e_c) de la clique (c) sur laquelle elles s'appliquent, ainsi que la totalité de l'observation O. Les poids λ_k associés à chacune de ces fonctions sont les paramètres du modèle estimés lors de la phase d'apprentissage. Dans ce travail, le graphe des dépendances modélise des dépendances du premier ordre, et les cliques seront alors composées de deux variables aléatoires, celle à la position courante et à la position précédente. Les expériences ont été réalisées à l'aide de l'outil libre CRF++ [1].

Dans nos expériences, les fonctions encodent tous les unigrammes, bigrammes et trigrammes construits sur les couples symbole/position dans une fenêtre $[-2, +2]$ autour de la position de décision.

1. Disponible sur Internet à l'adresse `http://crfpp.sourceforge.net/`

4 Conditions expérimentales

4.1 Corpus et tâches ESTER 2

Le corpus ESTER 2 relatif aux entités nommées se composent de 72 heures d'émissions radiophoniques francophones (France-Inter, France Info, RFI, RTM, France Culture, Radio Classique) manuellement transcrites et annotées en entités nommées suivant les conventions des deux campagnes ESTER [Ester] qui sont légèrement différentes. La première campagne comporte un jeu de 30 types d'entités nommées réparties en 9 catégories principales (personne, organisation, groupe géo-socio-politique, lieu, bâtiment et construction humaine, production humaine, date et heure, montant, inconnue), alors que la seconde possède un jeu de 37 types d'entités nommées réparties en 7 catégories principales (personne, fonction, organisation, lieu, production humaine, date et heure, montant). Le tableau 3 détaille la composition des données utilisées dans ce travail.

corpus	**nombre d'heures**	**source**
entrainement	60h	apprentissage ESTER 1
	6h	développement ESTER 2
test	6h	test ESTER 2

TABLE 3 – Décomposition du corpus ESTER 2 pour la tâche d'annotation en entités nommées

La campagne ESTER 2 comporte deux tâches de reconnaissance d'entités nommées qui consistent à reconnaître les entités nommées, d'une part, dans la transcription manuelle (**man**) du corpus de test et, d'autre part, dans les trois transcriptions automatiques (**aut**) du corpus de test dont les taux d'erreur de mots sont 12.11%, 17.83% et 26.09%. On se place, ici, dans le cas où la transcription automatique est la plus bruitée (*i.e.* taux d'erreur de mots de 26.09%).

4.2 Mesure des performances

Les performances pour la reconnaissance d'entités nommées sont ici évaluées en terme de *slot error rate* (SER) utilisé dans la campagne ESTER 2 [Ester]. Le SER fournit un taux d'erreur sur l'ensemble des entités nommées de référence (R) pour lequel on distingue les erreurs d'insertion (I), de suppression (D) et de substitution (S). Dans le cas de la substitution, on distingue les erreurs de type (T), d'extension (E), de type et d'extension (TE), ou multiples (M) où plusieurs hypothèses correspondent à une entité de référence. Pour évaluer la mise au point de nos systèmes, nous utilisons un premier SER défini par :

$$SER_1 = \frac{\#I + \#D + \#S}{\#R}$$

Dans le cadre de la campagne ESTER 2, chaque type d'erreur est pondéré par un coefficient suivant son importance. Il est défini par :

$$SER_2 = \frac{\alpha_I.\#I + \alpha_D.\#D + \alpha_T.\#T + \alpha_E.\#E + \alpha_{TE}.\#TE + \alpha_M.\#M}{\#R}$$

avec $(\alpha_I, \alpha_D, \alpha_T, \alpha_E, \alpha_{TE}, \alpha_M) = (1, 1, 0.5, 0.5, 0.7, 0.7)$.

La F-mesure en entités nommées est aussi utilisé pour mesurer les performances lors la mise au point de nos systèmes. Elle correspond à la moyenne harmonique entre la précision et le rappel en entités nommées.

Le SER_1 et la F-mesure permettront de comparer les différents systèmes combinés aux différents descripteurs utilisés, tandis que le SER_2 permettra de s'évaluer par rapport aux meilleurs systèmes de la campagne ESTER 2.

5 Résultats

5.1 Apport des différents descripteurs

Afin de mesurer l'apport des différents descripteurs, on teste quatre cas de système (MOT, MOT+MS, MOT+MS+AP, et MOT+MS+AP+MI), utilisant progressivement les informations décrites dans la table 2.

descripteurs	MOT	MOT+MS	MOT+MS+AP	MOT+MS+AP+MI
transcription	man	man	man	man
FST	32.3 (0.77)	31.9 (0.77)	32.2 (0.77)	30.9 (0.78)
SVM	35.1 (0.77)	29.4 (0.80)	29.1 (0.81)	28.9 (0.81)
CRF	41.7 (0.72)	29.8 (0.79)	28.4 (0.80)	28.1 (0.80)

TABLE 4 – Performances des étiqueteurs en SER_1 (F-mesure) suivant différents descripteurs

Les résultats (tableau 4) nous montrent l'influence positive de l'ajout du deuxième niveau de description. On le voit notamment pour les méthodes discriminantes (SVM et CRF) où le gain est significatif lorsqu'on ajoute des informations morpho-syntaxiques qui permettent de déduire des règles plus généralisatrices que dans le cas des mots seuls. Il est intéressant de noter que, pour le modèle FST, l'ajout de connaissances *a priori* dégrade les performances. On peut aussi supposer que l'inconsistance des annotations (différentes suivant les deux campagnes ESTER) du corpus d'apprentissage perturbe les classifieurs, ce qui sera l'objet de la partie suivante.

5.2 Correction des données d'entraînement

En apprentissage automatique (AA), la quantité de données d'apprentissage est une notion cruciale. C'est pourquoi certains concepteurs privilégient l'enrichissement des connaissances au détriment de leur aspect qualitatif. Or les méthodes d'AA y sont très sensibles. De plus, dans un contexte de transcription de parole, où la robustesse est une priorité et où les mots à analyser sont limités au lexique défini par le système de reconnaissance de la parole, la qualité de ces données est primordiale. Comme précisé dans la partie 4.1, l'ensemble utilisé pour l'entraînement des classifieurs est le corpus d'apprentissage annoté dans le cadre de la campagne ESTER 1 ainsi que le corpus de développement annoté, suivant des conventions légèrement différentes, dans le cadre de ESTER 2. Les systèmes obtenus sont bien sûr plus performants en utilisant conjointement les deux corpora plutôt que séparément. Néanmoins, l'incohérence des annotations affecte les performances de ces systèmes [10]. Il n'est pas rare de se retrouver dans ce genre de situation et nous proposons une méthode pour harmoniser, et rendre les annotations cohérentes. L'idée est de conserver le système de description le plus performant (MOT+MS+AP+MI) pour le corpus le plus fiable (corpus de développement (DEV), annoté suivant les mêmes conventions que le test ESTER 2). Le corpus d'apprentissage (APP), quant à lui, est décrit avec les mots ainsi que les étiquettes morpho-syntaxiques (MOT+MS). Le premier niveau de description, composé des mots, va permettre de générer des règles faibles, peu généralisantes. Le deuxième niveau de description, composé de (MOT+MS) ou de MOT+MS+AP+MI selon la partie

considérée des données d'entraînement, va permettre de produire des règles plus fortes. En utilisant le système le plus performant (CRF), un modèle est appris sur la concaténation des deux corpora. Àu corpus d'entraînement sont de nouveau associés les niveaux de description les plus efficaces MOT+MS+AP+MI sur toutes les données. Il est ensuite ré-annoté automatiquement avec le modèle précédent. Lors de la ré-annotation, les annotations de la partie DEV seront reproduites. Sur la partie APP, les règles fortes apprises sur la partie APP ne pourront plus s'appliquer du fait de la modification du second niveau de description, seules les règles fortes apprises sur la partie DEV vont s'appliquer. Les règles faibles apprises sur la partie APP vont permettre de régénérer les annotations sur la partie APP (et donc de conserver la connaissance incluse dans APP) sauf en cas de contradiction avec les règles fortes apprises sur le DEV qui dans ce cas vont l'emporter pour proposer une nouvelle annotation plus conforme à la partie DEV. C'est cette nouvelle annotation qui servira de référence à l'entraînement de tous les systèmes.

Le tableau 5 montre les performances des systèmes entraînés avec ce nouveau jeu d'annotations. On constate une amélioration significative (jusqu'à 5 points absolu ou environ 20% relatif de gain) des résultats par rapport à ceux de la partie précédente.

descripteurs	MOT	MOT+MS	MOT+MS+AP	MOT+MS+AP+MI
transcription	man	man	man	man
FST	27.3 (0.81)	29.6 (0.80)	29.1 (0.80)	26.6 (0.82)
SVM	32.4 (0.79)	27.4 (0.82)	26.9 (0.83)	26.6 (0.83)
CRF	36.2 (0.76)	24.8 (0.83)	23.4 (0.84)	22.8 (0.84)

TABLE 5 – Performances des étiqueteurs en SER_1 (F-mesure) suivant différents descripteurs avec correction automatique du corpus d'entraînement

5.3 Évaluation ESTER 2

Nous évaluons ici les performances de nos systèmes sur le jeu de données ESTER 2 en les comparant aux meilleurs systèmes de la campagne ESTER 2. Le meilleur système ESTER 2 sur la transcription manuelle est basé sur des règles de grammaires formelles, noté **ref-man**, tandis que le meilleur système ESTER 2 sur la transcription automatique la plus bruitée (taux d'erreur de mots de 26.09%) est à base d'apprentissage automatique proche de notre approche à base de CRF, noté **ref-aut**. Un post-traitement simple est appliqué sur la sortie des trois systèmes. Il consiste à appliquer quelques règles d'imbrications d'entités nommées et de correction de segmentation afin de mieux correspondre aux conventions d'annotation d'ESTER 2.

système	FST	SVM	CRF	oracle(SVM+CRF)	oracle(FST+SVM+CRF)	ref-man	ref-aut
man	27.89	28.06	22.79	/	/	9.80	23.91
aut	59.44	59.83	53.49	50.40	45.80	66.22	56.79

TABLE 6 – Performances des étiqueteurs en SER_2 (évaluation ESTER 2)

Les trois systèmes que nous présentons obtiennent tous des résultats inférieurs à 60% en SER_2, ce qui les classerait premier, troisième et quatrième de la campagne ESTER 2 sur les transcriptions les plus bruitées. Notre système à base de CRF, proche du meilleur système **ref-aut**, obtient des performances sensiblement meilleures grâce à notre méthode d'amélioration de la qualité des annotations (partie 5.2) et ce bien qu'aucune ressource autre que celles annotées durant ESTER n'ai été utilisées contrairement au système participant. Outre les avantages respectifs de chaque méthode, l'analyse directe des graphes de

mots pour les FST et la précision des méthodes discriminantes (CRF et SVM), l'intérêt de proposer trois systèmes ayant une vue différente du problème permettra d'améliorer dans de futurs travaux la robustesse de l'étiquetage par des stratégies de fusion. Le tableau 6 illustre à travers les taux oracles (taux d'erreur minimal que ferait une stratégie qui prend toujours la bonne décision) le gain potentiel d'une telle stratégie.

6 Conclusion

Nous avons présenté dans ce travail trois méthodes pour la reconnaissance d'entités nommées à partir de transcriptions automatiques de parole ainsi qu'une méthode permettant d'améliorer automatiquement la qualité des annotations d'un corpus dont les annotations ont été effectuées suivant des conventions légèrement différentes. Nous avons montré que les systèmes d'étiquetage obtenus sont parmi les plus robustes sur les données d'évaluation de la campagne ESTER 2 dans les conditions où la transcription automatique est particulièrement bruitée (taux d'erreur de mots de 26.09%). Le système à base de CRF obtient les meilleures performances. Bien que moins performante, la méthode à base de SVM offre tout de même une distribution des erreurs différentes. La méthode à base de FST possède l'avantage de pouvoir être couplée efficacement avec les systèmes de reconnaissance de la parole et sera évaluée prochainement dans ces conditions. Enfin, l'évaluation oracle montre que les trois systèmes offrent des résultats complémentaires que nous comptons combiner grâce à des méthodes de fusion pour améliorer les performances.

Références

[1] BRUN C. & EHRMANN M. (2009). Adaptation of a named entity recognition system for the ester 2 evaluation campaign. In *IEEE NLP-KE*, Dalian, Chine.

[Ester] ESTER. Conventions et plans d'évaluation des campagnes ester. Disponible sur Internet à l'adresse `http://www.afcp-parole.org/ester/docs.html`.

[3] FAVRE B., BÉCHET F. & NOCÉRA P. (2005). Robust named entity extraction from spoken archives. In *HLT-EMNLP'05*.

[4] HUET S., GRAVIER G. & SÉBILLOT P. (2008). Morphosyntactic resources for automatic speech recognition. In *LREC'08*, Marrakech, Maroc.

[5] ISOZAKI H. & KAZAWA H. (2002). Efficient support vector classifiers for named entity recognition. In *COLING*.

[6] KUDO T. & MATSUMOTO Y. (2001). Chunking with support vector machines. In *NAACL'01*, p. 1–8.

[7] LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *ICML'01*, p. 282–289.

[8] MCCALLUM A. & LI W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *CoNLL-2003*, p. 188–191.

[9] MOHRI M., PEREIRA F. & RILEY M. (1997). *AT&T FSM Library - Finite State Machine Library*. Rapport interne, AT&T.

[10] RAYMOND C. & RICCARDI G. (2007). Generative and discriminative algorithms for spoken language understanding. In *Interspeech*, Anvers, Belgique.

[11] VAPNIK V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc.