Modèles discriminants pour la prédiction d'erreur dans les réseaux de confusion

Alexandre Allauzen

LIMSI/CNRS, Univ Paris-Sud, F-91405, Orsay, France

allauzen@limsi.fr

ABSTRACT

In this article, error detection for broadcast news transcription system is addressed in a post-processing stage. To estimate the probability of errors, we introduce the use of linear-chain conditionnal random fields based on features extracted from confusion networks. The linear-chain is a discriminative alternative to hidden Markov models for sequence classification. The linear chain configuration is experimented with both real valued and binarized features showing a slight impact of binarization on classification performances. To improve our models, the linear chain is then augmented to include dependencies to adjacent feature vectors. Our best model yields to an absolute reduction of the classification error rate of 9% to be compared with the standard ASR output (from 13.9% to 4.7%) and 6% to be compared to a logistic regression model trained in same conditions.

Keywords: Reconnaissance automatique de la parole, détection d'erreur, champs aléatoires conditionnels (CRF)

1. INTRODUCTION

Ces dernières années les campagnes internationales d'évaluation NIST ont tenté de montrer que la qualité des transcriptions automatiques de la parole devenaient suffisante pour proposer un accès pertinent aux documents audiovisuels. Ainsi en 2000, la campagne *SDR (Spoken Document Retrieval)* s'est achevée sur le constat qu'en matière d'indexation automatique de documents audiovisuels, l'état de l'art est un système de reconnaissance automatique de la parole (RAP), allié à des techniques de recherches d'information [2].

Ainsi, la transcription automatique de la bande sonore devient l'accès premier au contenu audiovisuel mais les erreurs qu'elle peut contenir en détermine la pertinence. En particulier, les zones d'erreurs sont sources de bruits et d'informations perdues dans une perspectives d'indexation. De plus ces erreurs rendent difficile l'application de post-traitement comme l'extraction d'information. Il est donc important de repérer ces zones afin de pouvoir appliquer une stratégie appropriée comme par exemple : un décodage phonétique si les erreurs sont dues à des mots hors vocabulaire ; ou une adaptation des modèles du système de RAP pour tenter de corriger les erreurs.

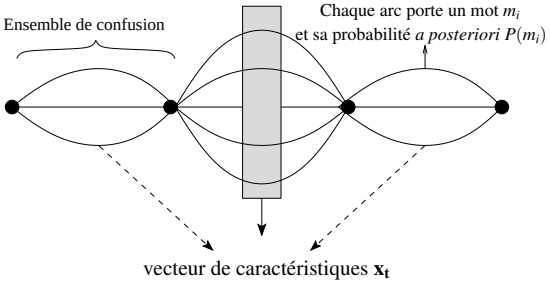
Une première approche pour appréhender les erreurs de reconnaissance est de modéliser les erreurs afin qu'elles soient prises en compte par le système de RAP. Cette approche consiste à introduire dans les connaissances du système un ou plusieurs modèles de l'inconnu [5] (*filler model*). L'autre approche consiste à estimer une mesure de confiance à partir de caractéristiques extraites lors du décodage et de la sortie du système de RAP. Une comparaison de ces deux approches a montré qu'il était plus efficace d'estimer un score de confiance [5].

De nombreux articles traitent de l'estimation d'une mesure de confiance à partir des différentes représentations de l'espace de recherche d'un système de RAP. Les premiers travaux ont proposé d'extraire des caractéristiques de la meilleure hypothèse afin d'estimer un score de confiance grâce à la régression logistique [4] ou aux réseaux de neurones [12]. D'autres travaux se sont également intéressés au treillis de mots [8] avec des méthodes similaires. L'utilisation du réseau de confusion est plus récente et semble prometteuse : la régression SVM peut être utilisée comme un autre estimateur des probabilités *a posteriori* des mots [6], ou les forêts aléatoires peuvent servir de classifieur [13]. Dans [7], le lecteur pourra trouver une vue d'ensemble plus large du domaine.

Dans un des articles fondateurs [4], les auteurs introduisaient l'usage d'apprentissage discriminant, la régression logistique, afin d'estimer la probabilité qu'un mot de l'hypothèse soit une erreur. Ces travaux ont été récemment réactualisés et adaptés aux réseaux de confusions [1]. Cependant, la régression logistique effectue une classification hors contexte contrairement par exemple aux modèles de Markov Cachés (MMC). L'absence de prise en compte du contexte par l'ensemble des travaux sur le sujet est problématique car les erreurs de reconnaissances peuvent s'enchaîner et intervenir par zone.

Dans cette article nous proposons donc d'explorer les champs aléatoires conditionnels (*Conditional Randam Fields ou CRF*) pour la détection d'erreurs en post-traitement dans les réseaux de confusion. Comme représenté à la figure 1, un vecteur de caractéristiques est extrait pour chaque ensemble de confusion. Un réseau de confusion est alors converti en une séquence de vecteurs d'observations à laquelle va être associée une séquence de classes indiquant la présence potentielle d'une erreur. Les expériences sont menées sur le corpus d'émissions radiophoniques de l'évaluation ESTER avec le système de transcription automatique temps-réel du LIMSI développé lors de cette campagne. Par la suite, nous décrirons donc tout d'abord les caractéristiques extraites des réseaux de confusion, nous expliciterons ensuite l'application des CRF à la détection d'erreur et le paradigme d'évaluation. Puis nous décrirons les données et le système de RAP utilisés, avant de présenter les résultats expérimentaux.

FIG. 1: Représentation d'un réseau de confusion et du processus d'extraction de caractéristiques



2. CARACTÉRISATION DES ERREURS

La plupart des systèmes de RAP reposent sur une modélisation statistique du processus de génération de la parole. L'objectif est de déterminer la séquence de mots minimisant l'espérance du taux d'erreur sur les mots. Ainsi, la probabilité *a posteriori* est calculée pour chaque mot hypothèse lors de la conversion du treillis de mots en réseau de confusion [10]. Comme représenté à la figure 1, un réseau de confusion est un graphe linéaire où la complexité d'un treillis de mots est réduite à une séquence d'ensembles de confusion (EC). Chaque EC est un ensemble d'hypothèses de mots temporellement parallèles avec leurs probabilités *a posteriori* associées. La meilleure hypothèse de reconnaissance s'extrait en prenant dans chaque EC le mot le plus probable.

2.1. Sorties des systèmes de RAP

Un système de reconnaissance peut produire différents types de sorties pouvant servir de base à l'estimation de mesure de confiance. La probabilité *a posteriori* associée à la meilleure hypothèse semble la plus évidente, cependant les probabilités *a posteriori* sont en général surestimées et ne suffisent pas comme mesure de confiance[1]. Pour ce travail, nous avons choisi d'utiliser les réseaux de confusion. Cette représentation est plus riche que la meilleure hypothèse et ne contient pas les redondances d'une liste des *n-best* hypothèses. Comparé au treillis de mots, le réseau de confusion est plus compact et son utilisation plus efficace, puisque les probabilités *a posteriori* tiennent comptent du meilleur chemin dans le treillis de mots mais aussi des chemins alternatifs.

2.2. Caractéristiques

Comme représenté à la figure 1, un vecteur de caractéristiques $\mathbf{x_t}$ est extrait pour chaque EC. Une sélection des caractéristiques est effectuée à partir des travaux [13, 6, 1], elles sont toutes extraites directement du réseau de confusion. Les caractéristiques à valeurs réelles utilisées sont la probabilité *a posteriori* de l'hypothèse émise, et l'entropie locale de l'EC définie par :

$$H = -\sum_{i=1}^{N} P(m_i) log(P(m_i)),$$

avec N le nombre d'hypothèses parallèles de l'EC et $P(m_i)$ la probabilité *a posteriori* associée à la i-ème hypothèse de mot m_i de l'EC. À celles-ci s'ajoutent des indicateurs binaires : l'identité du mot hypothèse, des marqueurs des débuts et fins de segment, et des indicateurs de la présence d'un arc nul[2], d'une respiration ou d'une hésitation comme meilleure hypothèse.

3. MODÈLE DISCRIMINANT POUR LA DÉTECTION DES ERREURS

Considérons un réseau de confusion comme une séquence d'EC représentée par une séquence de vecteurs d'observation $\mathbf{x} = (\mathbf{x_t})$. Attribuer un mesure de confiance à chaque élément de $\mathbf{x}$ peut être formalisé comme un problème de classification de séquence. À $\mathbf{x}$ est associée une séquence d'étiquettes de même longueur $\mathbf{y} = (y_t)$; y_t est la classe associée à l'EC représenté par $\mathbf{x_t}$, elle indique si la meilleure hypothèse est correcte ou non. La détection d'erreurs est une tâche plus simple de classification binaire, qui peut être déduite par l'application d'un seuil de décision de 0.5 sur la mesure de confiance.

3.1. Champs aléatoires conditionnels (CRF)

Les champs conditionnels aléatoires (CRF) sont des modèles discriminants ayant pour objectif de définir une distribution de probabilités sur une séquence de classes $\mathbf{y}$ étant donnée une séquence d'observation $\mathbf{x}$ [9]. Pour un problème à 2 classes, les CRF sont l'extension de la régression logistique à la classification de séquences. Dans cet article nous utiliserons un cas particulier des CRF : les chaînes linéaires, l'équivalent discriminant des MMC.

Soient $\mathbf{x}$ une séquence de vecteurs aléatoires, la séquence d'étiquettes associées $\mathbf{y}$, et $\Lambda = (\lambda_k) \in \mathbb{R}^K$ un vecteur de paramètres pondérant un ensemble de fonctions caractéristiques à valeurs réelles $(f_k(y, y', \mathbf{x_t}))_{k=1}^K$. Une chaîne linéaire est une distribution $p(\mathbf{y}|\mathbf{x})$ prenant la forme :

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\Big\{ \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \Big\} \quad (1)$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp\Big\{ \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \Big\}, \quad (2)$$

avec $Z(\mathbf{x})$ la fonction de partition permettant de normaliser la distribution. Les fonctions caractéristiques permettent d'encoder différents types de relation entre les étiquettes et le vecteur d'observation : par exemple la fonction $\log p(y_t|y_{t-1})$ est l'équivalent des probabilités de transition typiques d'un MMC ; des caractéristiques binaires comme l'identité de la meilleure hypothèse : $f_k(y_t, y_{t-1}, \mathbf{x}_t) = 1$ si la meilleure hypothèse est le mot m, 0 sinon. Ainsi les caractéristiques décrites au paragraphe 2.2 sont mises sous forme de fonction caractéristiques. L'apprentissage des CRF consiste à déterminer les coefficients Λ afin de maximiser la vraisemblance conditionnelle des données d'entraînement [9].

[1] Les raisons de cette surestimation sont d'une part que les vraisemblances sont normalisées par rapport au treillis de mots et non par rapport à la totalité de l'espace de recherche, d'autre part la possibilité de mots hors vocabulaire fait que l'espace de recherche peut être incomplet.

[2] Lors de la conversion du treillis de mots en réseau de confusion, chaque mot hypothèse se voit assigné une position pour des raisons d'alignement. Les omissions sont représentées par les arcs nuls.

3.2. Évaluation

Les résultats sont évalués en termes de taux d'erreur de classification (*Classification Error Rate* ou CER). Comme résultat de départ, faisons l'hypothèse que le système de RAP ne se trompe pas. Dans ce cas, le CER du système de reconnaissance est égale au taux d'erreur sur les mots (*Word Error Rate* ou *WER*) à la différence près que les omissions du système de RAP ne sont pas comptées. Le CER est défini comme :

$$CER = \frac{\textit{nombre d'EC classés correctement}}{\textit{nombre d'EC}}$$

Les omissions ne sont pas prises en compte puisque l'objectif est d'étiqueter correctement les hypothèses émises par le système de RAP. Pour évaluer un étiquetage de séquence les deux autres mesures classiquement utilisées sont le rappel et la précision pour chaque étiquette. Les résultats obtenus avec la régression logistique [1] entraînée dans les mêmes conditions proposent un point de comparaison avec un classifieur semblable mais hors contexte.

4. CORPUS ET SYSTÈME DE RECONNAISSANCE

Les expériences décrites dans cet article utilisent les données de test de la campagne d'évaluation ESTER qui contiennent 10 heures d'émissions radio d'actualité en français[3]. Ces émissions proviennent de 4 chaînes de radio différentes présentes dans les données officielles d'entraînement, d'une chaîne de radio sans données d'entraînement transcrites et d'une chaîne inconnue. Les données ont été collectées entre octobre et décembre 2004 et contiennent environ 107k mots après normalisation.

Pour générer les réseaux de confusion, le système français de transcription temps-réel de document audio du LIMSI est utilisé. Ce système a été développé lors de la campagne ESTER, et sa description complète peut être trouvée dans [3]. Nous en résumons ici les grandes lignes. Il comporte deux traitements : la segmentation du flux audio et la reconnaissance de la parole. Le processus de segmentation permet de diviser le flux audio continu en segments acoustiques homogènes étiquetés en genre, en largeur spectrale (téléphone, studio), et en locuteur. Le système de RAP utilise un vocabulaire de 65k mots, des MMC pour la modélisation acoustique et des modèles n-grammes interpolés pour la modélisation du langage. Le décodage est effectué avec un modèle de langage bigramme en deux passes. Chaque passe génère un treillis de mots qui est étendu avec un modèle de langage quadrigramme. Le dernier treillis de mots est converti en réseau de confusion.

5. EXPÉRIMENTATIONS

Nous avons utilisé la librairie *GRMM* [11] pour l'entraînement et l'inférence des CRF. Les réseaux de confusion sont d'abord générés pour tout le corpus de test ESTER puis alignés avec les transcriptions de référence grâce à l'algorithme de programmation dynamique. Si des alignements sont ex-æquo, la sélection est faite en appliquant une contrainte supplémentaire sur la distance de Levenshtein entre les mots. Avant l'alignement, les règles de normalisation usuelles pour évaluer les systèmes RAP sont appliquées à l'hypothèse et à la référence (conversion en minuscule, conversion de mots composés dans une forme commune, ...). Cependant ses règles sont modifiées afin de préserver la segmentation en mots induite par le système de RAP. Le WER est alors différent de l'officiel.

[3] Plus de détails peuvent être trouvés sur le site Web de l'évaluation : *http ://www.afcp-parole.org/ester/index.html*

		CER	
		chaîne linéaire	contexte
	régression log.	11.5	10.6
	binaires	8.5	5.8
quantif.	entropie+ *post.*	6.0	5.5
réelle	entropie+*post.*	6.1	4.7

TAB. 1: Résultats en terme de taux d'erreur de classification (CER) : pour les caractéristiques réelles le cas où elles sont quantifiées "*quantif.*" ou non "*réelle*" ; les formes de CRF évaluées sont la "chaîne linéaire" et sa version ("*contexte*") augmenté. "*post.*" désigne la probabilité *a posteriori* de la meilleure hypothèse.

Pour la détection d'erreur, une séquence d'étiquettes est associée à un réseau de confusion ; chaque étiquette indique si la meilleure hypothèse de l'EC est correcte ou non. Certains EC sont écartés de la tâche de classification car leurs meilleures hypothèses n'interviennent pas dans l'évaluation classique d'un système de RAP : arc nul ou marqueur d'hésitation et de respiration. Cependant ces EC peuvent intervenir dans la construction des vecteurs de caractéristiques. Le corpus obtenus contient environ 1400 réseaux de confusion et un total de 105k EC. Dix sous-corpus sont crées pour la validation croisées qui a été utilisée systématiquement pour les résultats de cet article.

Le tableau 1 regroupe l'ensemble des résultats en terme de CER. À titre de comparaison, le CER du système de RAP est de 13.9%. Ce taux d'erreur peut être réduit à 11.5% ou 10.6% selon les configurations en utilisant la régression logistique [1] entraînée dans les mêmes conditions.

5.1. Caractéristiques binaires ou réelles

Regardons la première colonne du tableau 1 qui donne les résultats pour un CRF de type chaîne linéaire. La première expérience consiste à entraîner une chaîne linéaire en utilisant uniquement les caractéristiques binaires décrites au paragraphe 2.2. Ainsi l'ensemble restreint des caractéristiques binaires permet au CRF d'améliorer nettement les résultats obtenus avec la régression logistique, en réduisant le CER de 11.5% à 8.5%.

En traitement automatique du langage, les caractéristiques sont la plupart du temps discrètes et peuvent se mettre sous la forme d'attributs binaires. En traitement automatique de la parole, des caractéristiques peuvent être à valeur réelle comme dans notre cas la probabilité *a posteriori* et l'entropie locale (définis au paragraphe 2.2). En théorie, les fonctions caractéristiques peuvent être à valeur réelle. En pratique, d'une part cela dépend de l'implémentation des CRF utilisées, et d'autre part, une normalisation des caractéristiques est utile. L'autre approche la plus souvent utilisée contourne le problème en quantifiant les valeurs réelles. Dans cet article nous évaluons les deux approches.

Étant données les distribution des probabilités *a posteriori* et de l'entropie locale, une quantification linéaire est in-

adaptée. Les pas de quantification sont donc calculés à partir de l'histogramme des valeurs de manière à ce que chaque quantum contienne la même quantité de données. Par la suite, 20 quantums sont utilisés pour quantifier les caractéristiques réelles. Les résultats montrent d'une part le gain significatif à ajouter les deux caractéristiques réelles et d'autre part l'absence de différence significative entre l'utilisation des valeurs quantifiées ou réelles sachant que seulement 2 caractéristiques réelles sont introduites.

5.2. Impact des dépendances

Dans la seconde colonne du tableau 1 (contexte), les CRF utilisés étendent les dépendances aux observations de la chaîne linéaire à la totalité du contexte : les fonctions caractéristiques introduites à l'équation 1 sont désormais de la forme $f_k(y_t, y_{t-1}, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1})$. Cette extension du contexte permet une réduction significative du CER : de 9% relatif pour les caractéristiques quantifiées et 30% relatif pour les réelles. L'impact de la quantification est ici plus important car en ajoutant le contexte le nombre de caractéristiques réelles est démultiplié. D'ailleurs, avec la prise en compte du contexte, la différence devient notable (20% relatif) en faveur des caractéristiques réelles.

Le meilleur classifieur obtient donc un CER de 4.7%, soit une diminution d'environ 9% absolu par rapport à la sortie du système de RAP et 6% par rapport à la régression logistique. Alors que le taux de rappel du système de RAP pour les mots correct est de 100%, la détection d'erreur obtient un rappel 98.7% soit une faible dégradation. Cependant le taux de précision est nettement amélioré en passant de 86.1% à 96.5%. Enfin le taux d'EC corrects classés comme erreurs quantifie la dégradation en terme de taux d'erreur sur les mots que peut engendrer l'utilisation du classifieur. Ce taux est de 3% pour le meilleur classifieur.

6. CONCLUSION

Dans cette article, nous avons introduit l'utilisation des champs aléatoires conditionnels afin de prédire les erreurs dans les réseaux de confusion. Par rapport aux travaux précédents utilisant les SVM ou la régression logistique, ces modèles discriminants permettent de classer des séquences d'observation à partir d'un ensemble de fonctions caractéristiques pouvant être à valeurs binaires ou réelles. Ces fonctions établissent le lien entre les observations extraites des réseaux de confusions et la séquence d'étiquettes à déterminer. Les réseaux de confusion ont été générés avec le système temps-réel de transcription d'émission radiophonique du LIMSI pour les données de test des évaluations francophones ESTER.

Dans ce travail nous avons étudié les chaînes linéaires qui sont l'équivalent discriminant des modèles de Markov cachés. Les résultats montrent l'opportunité de détecter les erreurs en séquence et l'impact relatif de la quantification des fonctions caractéristiques. Puis les dépendances ont été étendues en prenant en compte les vecteurs d'observation adjacents permettant de réduire significativement le taux d'erreur de classification : jusqu'à 30% relatif par rapport à une chaîne linéaire standard.

Le meilleur classifieur réduit le taux d'erreur de classification à 4.7%, soit un gain absolu de 9% par rapport à la sortie du système de transcription et de 6% par rapport à la régression logistique. Ces résultats confirment donc d'une part l'importance de la prise en compte du contexte pour détecter les erreurs de reconnaissance, et d'autre part, que les champs aléatoires conditionnels sont appropriés à la tâche.

7. REMERCIEMENTS

L'auteur tient à remercier Charles Sutton pour son aide indispensable à propos de *GRMM*, ainsi que Jean-Luc Gauvain et François Yvon pour leurs conseils.

RÉFÉRENCES

[1] Alexandre Allauzen. Error detection in confusion network. In *Proc. of InterSpeech*, Antwerp, September 2007.

[2] J. Garofolo, G. Auzanne, and E. Voorhees. The trec spoken document retrieval track : A success story. In *Proceedings of the 8th Text Retrieval Conference TREC-8*, pages 107–130, November 1999.

[3] Jean-Luc Gauvain, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Veronique Gendner, Lori Lamel, and Holger Schwenk. Where Are We in Transcribing French Broadcast News ? In *Proc. of InterSpeech*, Lisbon, September 2005.

[4] L. Gillick, Y. Ito, and J. Young. A probabilistic approach to confidence estimation and evaluation. In *Proc.of the 1997 IEEE ICASSP*, volume 2, pages 879–882, Munich, Germany, 1997.

[5] T. Hazen and I. Bazzi. A comparison and combination of methods for OOV word detection and word confidence scoring. In *Proc. of the 2001 IEEE ICASSP*, Salt Lake City, USA, May 2001.

[6] Dustin Hillard and Mari Ostendorf. Compensating for word posterior estimation bias in confusion networks. In *Proc. of the 2006 IEEE ICASSP*, Toulouse, France, May 2006.

[7] Mauclair Julie. *Mesures de confiance en traitement automatique de la parole*. PhD thesis, Université du Maine, 2006.

[8] Thomas Kemp and Thomas Schaaf. Estimating confidence using word lattices. In *Proc. of Eurospeech*, pages 827–830, Rhodes, Greece, 1997.

[9] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[10] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition : Word error minimization and other applications of confusion networks. *Computer, Speech and Language*, 14(4) :373–400, 2000.

[11] Charles Sutton. Grmm : A graphical models toolkit. http ://mallet.cs.umass.edu, 2006.

[12] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke. Neural-network based measures of confidence for word recognition. In *Proc. of the 1997 IEEE ICASSP*, pages 887–890, Germany, 1997.

[13] Jian Xue and Yunxin Zhao. Random forests-based confidence annotation using novel features from confusion network. In *Proc. of the 2006 IEEE ICASSP*, Toulouse, France, May 2006.