Étude pour l'amélioration d'un système d'identification nommée du locuteur *

Vincent Jousse[§†], *Christine Jacquin*[§], *Sylvain Meignier*[†], *Yannick Estève*[†], *Béatrice Daille*[§]

LINA[§] - Laboratoire d'informatique de Nantes Atlantique - Nantes
LIUM[†] - Laboratoire d'informatique de l'Université du Maine - Le Mans
prenom.nom@univ-nantes.fr ou prenom.nom@univ-lemans.fr

ABSTRACT

Automatic speaker segmentation and classification produce generic labels rather than the true identity of the speakers. The proposed approach is based on the use of a semantic classification tree using lexical rules to extract the true identity of the speakers from the transcription. In this paper, experiments are carried out on French broadcast news from ESTER 2005 to evaluate this approach, focusing on the impact of the various combinations of automatic vs. manual transcription with automatic vs. manual speaker segmentation/classification. We also study the errors generated by the system.
Keywords: speaker identification, rich transcription, real name extraction.

1. Introduction

De très grandes collections de données audio ont besoin d'être indexées pour faciliter la recherche et l'accès à l'information. Les annotations manuelles ont un coût élevé, particulièrement si les besoins portent sur des informations comme le thème, des mots clés ou le nom des locuteurs. Les systèmes automatiques de transcriptions enrichies permettent de réduire les coûts. Toutefois, les erreurs plus ou moins importantes engendrées par ces systèmes peuvent pénaliser l'exploitation des transcriptions.

La transcription enrichie inclut une étape de segmentation en tours de parole du document qui sont ensuite classés par locuteurs. Chaque classe est identifiée par un label générique permettant de déterminer qui parle à chaque instant du document. Cependant, la vraie identité (nom, prénom) n'est pas disponible et rend impossible la recherche des interventions de *"Monsieur Untel"*.

Actuellement, il existe deux approches principales permettant d'attribuer sa vraie identité à un locuteur. La première se fonde sur l'analyse de l'acoustique à partir de méthodes issues de la reconnaissance du locuteur. Cette méthode nécessite des exemples de voix de chaque locuteur cible à identifier. La collecte de ces échantillons peut être coûteuse et difficile à faire évoluer pour un système amené à traiter des collections de taille importante qui sont susceptibles d'être complétées quotidiennement.

La seconde approche extrait les identités de locuteur de la transcription. Une des méthodes consiste à analyser la séquence de mots utilisée par chaque locuteur cible pour caractériser sa manière de parler [1]. Cette méthode a le même désavantage que la méthode acoustique : elle nécessite des connaissances et des données transcrites pour chaque locuteur cible. D'autres méthodes proposent d'extraire l'identité directement de la transcription [2, 3, 4]. Ces méthodes peuvent être utilisées uniquement si les locuteurs s'annoncent ou sont présentés. Ces méthodes sont bien adaptées aux enregistrements radiophoniques où le passage de parole est généralement fait en nommant le locuteur. Elles le sont moins pour les enregistrements de réunion, par exemple.

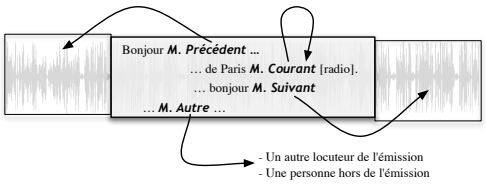

Fig. 1: Principe de base des systèmes d'identification nommée

Le système proposé dans cet article est apparenté à cette dernière catégorie de méthodes. Il repose sur un arbre de classification sémantique, couplé avec un système de détection d'entités nommées. Le système a été évalué sur des enregistrements radiophoniques en français de la campagne d'évaluation ESTER. Une analyse détaillée des erreurs montre qu'elles proviennent de faiblesses d'étiquetages en entités nommées et de manques d'exemples lors de l'apprentissage de l'arbre de classification sémantique.

2. Identification nommée du locuteur

2.1. Principes généraux

Les méthodes [2, 3, 4] partent toutes de la même source de données, à savoir la transcription du signal audio, transcription qui peut être effectuée manuellement ou automatiquement. Ces transcriptions permettent d'obtenir nombre d'informations exploitables pour l'identification nommée. Le texte, bien sûr, mais aussi les différents tours de parole qui sont regroupés en classes de locuteur identifiées par des labels anonymes. Le système repose sur la détection des noms de locuteur et des autres entités nommées dans cette transcription. Une fois ces noms détectés, il faut ensuite pouvoir les affecter aux différents tours de parole, puis aux classes de locuteur.

Les systèmes s'efforcent de trouver si le nom de locuteur détecté est celui qui va parler après, celui qui va parler avant, celui qui parle actuellement ou une autre

*Projet Miles financé par la Région des Pays de la Loire

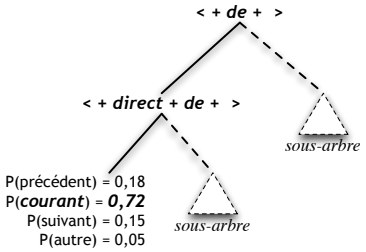

Fig. 2: Arbre de classification sémantique

personne étrangère au document, ou encore qui parle dans un tour de parole plus éloigné. Les méthodes actuelles se cantonnent donc à essayer d'attribuer le nom d'un locuteur aux tours de parole contigus au tour de parole courant (figure 1).

2.2. Approche retenue

La méthode proposée utilise un arbre de décision binaire reposant sur le principe des arbres de classification sémantique (SCT — [5]) qui apprend automatiquement des règles lexicales à partir des mots présents autour des noms des locuteurs détectés dans le corpus d'apprentissage. Il permet, lors de la phase de test, d'attribuer l'étiquette *"courant"*, *"prochain"*, *"précédent"* ou *"autre"* à chaque locuteur cité.

Les arbres de classification sémantique intègrent dans chaque noeud une expression régulière. La suite d'expressions régulières activées depuis la racine jusqu'à une feuille de l'arbre permet de classifier les contextes lexicaux suivant les 4 étiquettes décrites précédemment. En complément des expressions régulières, l'arbre peut intégrer des questions globales. Le système proposé utilise la position du nom détecté dans le tour de parole comme question globale. La position correspond aux situations où le nom détecté apparaît au début, au milieu ou à la fin d'un tour de parole. La figure 2 illustre que les étiquettes sont associées à des scores. Lorsqu'un exemple atteint une feuille de l'arbre, les scores correspondent aux probabilités de chacune de ces étiquettes. Ces scores sont établis lors de l'apprentissage de l'arbre et reflètent les cas observés dans le corpus d'apprentissage. À partir de ces scores une décision locale est prise en attribuant l'étiquette la plus probable à l'exemple. Pour les décisions correspondant aux étiquettes *courant*, *suivant* ou *précédant*, un nom de locuteur potentiel et son score sont alors associés aux tours de parole correspondants.

Chaque classe de locuteur est renommée en sélectionnant un nom final parmi les noms potentiels attribués aux tours de parole de la classe. Le nom final correspond au nom dont la somme des scores des occurrences de ce nom est maximale. Cette formule permet de prendre en compte le nombre d'occurrences d'un nom de locuteur pour une classe donnée, en pondérant chaque occurrence par son score.

L'approche par arbre de décision, comparée à la méthode proposée par S. Tranter utilisant des trigrammes [4, 6], donne des résultats similaires sur des transcriptions manuelles, mais est beaucoup plus robuste sur des transcriptions réalisées par un système de reconnaissance de la parole (précision supérieure d'au moins 5%). Dans l'optique d'un travail sur des transcriptions automatiques, nous avons choisi de remplacer l'étiquetage manuel des entités nommées effectué dans les précédents travaux par un étiquetage réalisé de manière automatique sur le corpus d'apprentissage et de test. L'outil utilisé est Nemesis (cf 3) développé par l'équipe TALN (Traitement Automatique des Langues Naturelles) du LINA.

3. Détection d'entités nommées : Nemesis

Nemesis [7] est un système d'identification et de catégorisation d'entités nommées pour le français. Ses spécifications ont été élaborées à la suite d'une étude en corpus et s'appuient sur des critères graphiques et référentiels. Ces derniers ont permis de construire une typologie des entités la plus fine et la plus exhaustive possible fondée sur celle de Grass [8]. L'architecture logicielle de Nemesis se compose principalement de 4 modules (prétraitement lexical, première reconnaissance, apprentissage, seconde reconnaissance) qui effectuent un traitement immédiat des données à partir de textes bruts. L'identification des entités nommées est réalisée en analysant leur structure interne et leurs contextes gauches et droits immédiats à l'aide de lexiques de mots déclencheurs, ainsi que de règles de réécriture. Leur catégorisation, quant à elle, s'appuie sur la typologie construite précédemment. L'outil atteint environ 90% de précision et 80% de rappel sur des textes écrits en langage naturel.

4. Expérimentation

4.1. Corpus ESTER

Les méthodes proposées sont développées et évaluées en utilisant les données de la campagne ESTER 2005. ESTER est une campagne d'évaluation sur les systèmes de transcription d'émissions radiophoniques en français[1]. Les données comportent six radios différentes dont les émissions durent de 10 à 60 min et sont décomposées en 3 corpora (apprentissage, développement et test).

Le système de transcription du LIUM [9] a été utilisé pour la transcription en mots. Ce système a évolué depuis 2005, les changements majeurs portent sur la paramétrisation acoustique et la simplification du processus de segmentation. Ce système permet d'obtenir 20.5% de taux d'erreur mots sur le corpus de test et un taux d'erreur de segmentation et de classification acoustique en locuteurs de 11,5%.

4.2. Métrique d'évaluation

Les résultats sont évalués en comparant l'hypothèse générée à la référence distribuée avec le corpus. Cette comparaison met en évidence 5 cas d'erreur ou de succès possibles relatifs aux situations suivantes :

- L'identité proposée est correcte (C_1) : le système propose une identité correspondant à celle indiquée dans la référence.
- Erreur de substitution (S) : le système propose une identité différente de l'identité présente dans la référence.
- Erreur de suppression (D) : le système ne propose pas d'identité alors que le locuteur est identifié dans la référence.

[1] `http://www.afcp-parole.org/ester/index.html`

Trans.	Seg. Class.	R	P	F
manu	**manu**	65,4%	91,2%	76,1
manu	**auto**	37,8%	69,9%	49,0
auto	**manu**	20,8%	73,7%	32,5
auto	**auto**	17,4%	68,0%	27,7

Fig. 3: Résultats sur le corpus de test
Trans. : transcription manuelle ou automatique.
Seg. Class. : segmentation et classification en locuteurs manuelle ou automatique.
R, P, F : respectivement mesure de rappel et de précision, F-mesure.

• Erreur d'insertion (I) : le système propose une identité alors que le locuteur n'est pas identifié dans la référence.
• Il n'y a pas d'identité (C_2) : le système ne propose pas d'identité et la référence ne contient pas d'identité.

A partir de ces cas, il est possible de définir une mesure de précision et de rappel :

$$P = \frac{C_1}{C_1 + S + I} \quad ; \quad R = \frac{C_1}{C_1 + S + D} \qquad (1)$$

La précision et le rappel peuvent être synthétisés en calculant la F-mesure : $F = (2 \times P \times R)/(P + R)$. Les résultats sont donnés en terme de durée, tous les instants de l'hypothèse et de la référence sont évalués.

4.3. Résultats

Le SCT a été appris à partir du corpus d'apprentissage, tandis que les paramètres du système ont été fixés en utilisant le corpus de développement. Le tableau 3 illustre les résultats obtenus sur le corpus de test. Les résultats sont uniquement donnés pour le seuil d'acceptation des décisions maximisant la F-mesure.

L'utilisation de la segmentation et de la classification automatique en locuteur (avec transcription manuelle) dégrade considérablement les performances du système, le taux de rappel est presque divisé par 2 alors que la précision chute de 21%. L'utilisation de transcriptions automatiques en laissant la segmentation et la classification manuelles donne des résultats encore plus mauvais. En effet, la précision ne chute que de 18% mais le rappel est quant à lui divisé par 3. L'utilisation de transcriptions automatiques a donc un impact plus négatif sur le système que l'utilisation de segmentations et de classifications automatiques. Les deux combinés (transcription et segmentation / classification automatiques) dégradent encore plus les performances.

4.4. Étude des erreurs

Afin de comprendre le comportement du système, nous avons choisi de l'étudier avec les données permettant de maximiser la F-mesure. Cette première étude utilise des transcriptions et des segmentations/classifications manuelles sur le corpus de test. Nous chercherons donc à comprendre pourquoi ce système a un rappel faible (65,4%) bien que les conditions soient optimales (transcriptions et segmentation / classification manuelles).

Méthode
Toutes les mauvaises décisions ont été étiquetées manuellement et comparées avec les différents scores générés par l'arbre. Cette étude a permis de mettre en évidence deux grandes catégories d'erreurs : les erreurs de détection d'entités nommées et les erreurs de décision de l'arbre. D'autres types d'erreurs ont aussi été constatés mais elles sont moins significatives.

Erreurs de détection d'entités nommées
La première catégorie d'erreur est due à des problèmes rencontrés lors de la détection d'entités nommées. En effet, les entités nommées, et plus précisément les noms de locuteurs non étiquetés uniquement par leur prénom / nom, ont une influence directe sur les performances du système en terme de rappel :

(N1) Un locuteur n'est pas correctement étiqueté pour le système d'identification nommée : quand il n'est pas détecté, quand seul son prénom est étiqueté ou quand ses nom et prénom sont étiquetés avec un ou plusieurs autres mots en plus (comme la fonction de la personne). Lorsque le nom du locuteur détecté ne correspond pas exactement au nom du locuteur de la référence, il est alors comptabilisé comme une erreur de suppression et par conséquent le rappel chute. Par exemple, l'entité nommée "[Joël Collado de Météo France]" est étiquetée comme une personne et ne correspond pas exactement au couple prénom / nom attendu lors de l'identification.

(N2) Un ensemble de mots ne contenant pas de nom et prénom est détecté comme une personne. Cette détection fait chuter le rappel en attribuant ce *faux locuteur* à un tour de parole. Par exemple, Nemesis étiquette l'entité nommée "[président du Fetia Api]" comme personne qui est ensuite attribuée à un tour de parole. "[président du Fetia Api]" est bien une personne, mais cette entité nommée n'est pas un locuteur au sens de l'identification. Il est identifié par sa fonction au lieu de son nom et prénom.

Erreurs de décision de l'arbre
La seconde catégorie d'erreurs provient des erreurs d'étiquetage commises par l'arbre. Elles ont une incidence à la fois sur le rappel et sur la précision du système :

(A1) Les erreurs affectant le rappel proviennent majoritairement des locuteurs qui ont été étiquetés comme *autres* alors qu'ils correspondent à une des 3 autres étiquettes.
"*Valérie Crova* `dit : (...) qui a toujours entretenu selon [Jean Christophe Bouisson] des rapports tendus avec le gouvernement.` *Jean Christophe Bouisson* `dit : On ne peut pas penser (...)`"
Dans cet exemple "[Jean Christophe Bouisson]" est étiqueté comme *autre* alors qu'il est le *suivant*.

(A2) Les erreurs affectant principalement la précision sont dues à des locuteurs qui sont mal étiquetés avec les étiquettes *suivant* et *précédent*. Dans la majorité des cas, l'étiquette *suivant* a été attribuée au lieu de *autre*. Par exemple, l'annonce d'interview correspond à une erreur typique :
"`[Jean Michel Hibon] au micro de [Jérôme Susini].`"
On s'attend à ce que "[Jean Michel Hibon]" parle ensuite (ce qui est le cas), mais le système détecte

Erreurs	Fréquence	Pourcentage Total
N1	12	14%
N2	4	4,6%
A1	32	37,2%
A2	30	34,9%
E	8	9,3%
Total	86	100%

Fig. 4: Répartition des erreurs sur le corpus de test.
N1, N2 : erreurs issues de la détection d'entités nommées.
A1, A2 : erreurs issues de l'arbre de classification.
E : autres erreurs.

"[Jérôme Susini]" comme le locuteur du prochain tour de parole car son score est plus élevé.

Autres erreurs

D'autres erreurs plus indépendantes du système ont été relevées :
(E) Certaines personnes ne sont citées que partiellement (seul leur prénom est cité dans la transcription), d'autres ont été mal orthographiées par le transcripteur ou encore certaines ne sont pas du tout citées ou annoncées dans la transcription. Sur les 11 heures du corpus, ce dernier cas ne concerne que 3 personnes. Cette constatation permet de valider l'hypothèse de départ : les noms des locuteurs sont présents dans la transcription.

Répartition des erreurs

Le tableau 4 montre la répartition et l'importance des différents types d'erreurs sur le corpus de test. Les erreurs provenant de l'arbre de classification (A1 et A2) sont clairement dominantes avec plus de 72% des erreurs totales alors que les erreurs dues à Nemesis (N1 et N2) ne représentent qu'un peu moins de 19% des erreurs. Pour l'arbre de classification, les erreurs A1 et A2 ont le même ordre de grandeur (environ 36% en moyenne). En revanche, en ce qui concerne Nemesis les erreurs de type N1 sont beaucoup plus fréquentes que celles de type N2 (3 fois plus).

Cas de l'anaphore pronominale

Avant cette étude, nous pensions que la résolution d'anaphore pronominale à l'intérieur d'un tour de parole pouvait conduire à améliorer les prises de décisions. Lors de l'analyse des erreurs, nous n'avons relevé que deux cas où cette prise en compte aurait été bénéfique. L'un des deux cas correspond à la situation suivante :
" *Fabrice Drouelle* : `écoutez ce qu'en pense [Jérôme Savary] (...) il le dit à [Christine Siméone].`"
Comme le pronom personnel *il* est en rapport avec Jérôme Savary, l'arbre aurait pu l'étiqueter comme *suivant*.

5. Conclusion

Nous avons présenté notre système d'identification de locuteur ayant comme principales caractéristiques de s'appuyer sur les données de transcription, d'effectuer la reconnaissance des entités nommées grâce à un outil issu du traitement automatique des langues et de réaliser l'apprentissage et la prise de décision à l'aide d'un arbre de classification sémantique. Nous avons d'abord analysé les résultats obtenus en terme de précision et de rappel et nous avons obtenu des résultats prometteurs. Nous avons ensuite mené une étude exhaustive concernant les erreurs commises par le système, erreurs qui se situent principalement au niveau du système de reconnaissance d'entités nommées Nemesis et de l'arbre de classification.

Au niveau de Nemesis, il faudrait premièrement calibrer ses sorties pour les adapter aux entrées attendues de notre système d'identification. En effet, Nemesis a été conçu pour détecter des entités nommées simples et complexes (par exemple "président de la république Jacques Chirac" est une entité nommée étiquetée comme anthroponyme par Nemesis) et le système d'identification exploite des entités nommées sous une forme simple de type "nom prénom", ce qui génère des erreurs. Deuxièmement, l'utilisation de Nemesis peut aussi être optimisée par un ajout de ponctuation au niveau de la transcription qui lui permettrait de mieux détecter les entités nommées (notamment quand une entité se trouve à la fin d'un segment et l'autre au début du suivant, il étiquette les deux entités nommées ensemble).

Au niveau de l'arbre de classification, il serait intéressant de prendre en compte le numéro d'ordre du nom détecté dans le tour de parole lors de l'apprentissage de l'arbre et du test. Ceci permettrait dans certains cas de faire abstraction des mots pouvant suivre le dernier nom de locuteur du tour de parole.

Références

[1] W. Antoni, C. Fredouille, and J.-F. Bonastre. On the use of linguistics information for broadcast news. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.

[2] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain. A comparative study using manual and automatic transcriptions for diarization. In *Proc. of ASRU, Automatic Speech Recognition and Understanding*, San Juan, Porto Rico, USA, November 2005.

[3] J. Mauclair, S. Meignier, and Y. Estève. Speaker diarization : about whom the speaker is talking ? In *IEEE Odyssey 2006*, San Juan, Puerto Rico, USA, 2006.

[4] S. E. Tranter. Who really spoke when ? Finding speaker turns and identities in broadcast news audio. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 1013–1016, Toulouse, France, May 2006.

[5] R. Kuhn and R. De Mori. The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5) :449–460, 1995.

[6] Yannick Estève, Syvlain Meignier, Paul Deléglise, and Julie Mauclair. Extracting true speaker identities from transcriptions. In *Proc. of Interspeech, European Conference on Speech Communication and Technology*, Antwerp, Belgium, 2007.

[7] N Fourour. Identification et catégorisation automatiques des entités nommées dans les textes français. In *Thèse en informatique de l'université de Nantes*, 2004.

[8] T Grass. Typologie et traductibilité des noms propres de l'allemand vers le français. In *Traitement automatique des langues*, volume 41(3), pages 643–670, 2000.

[9] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.