« Hé Manu, tu descends ? » : identification nommée du locuteur dans les dialogues

Léo Galmant[1] — Hervé Bredin[2] — Camille Guinaudeau[1] — Anne-Laure Ligozat[3]

[1] *LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay, France*
[2] *LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay*
[3] *LIMSI, CNRS, ENSIIE, Université Paris-Saclay, F-91405 Orsay, France*

prenom.nom@limsi.fr

RÉSUMÉ. *L'identification du locuteur est la tâche qui consiste à associer un locuteur à chaque tour de parole d'un dialogue, utilisée notamment pour enrichir les corpus de transcriptions automatiques.Le traitement de la tâche peut totalement différer selon le média : vidéo (films, séries, etc.), audio (séries, radio, etc.) ou textuel (scripts, transcriptions, etc.). Dans cet article, nous proposons une méthode d'identification du locuteur à partir des scripts et transcriptions de séries. Dans un dialogue de série, il est courant que les personnages mentionnent le nom des autres personnages à travers les discussions. Ces mentions peuvent servir d'indices afin d'identifier les locuteurs des tours de parole par leur nom. Nous proposons donc une approche d'identification du locuteur neuronale fondée sur la propagation des mentions à travers le dialogue. Cette approche parvient à trouver correctement le locuteur de 34% tours de parole pour la série* The Big Bang Theory, *et 34% pour la série* Friends.

ABSTRACT. *Speaker identification consists in matching a speaker for each speech turn of a given dialog. Its main use is to enhance automatic transcripts corpus.The task can differ totally from a media to another: video (movies, series, etc.), audio (series, radio, etc.) or textual (scripts, transcripts, etc.). In this paper, we propose a speaker identification method for scripts and transcripts of series. In series dialogs, characters usually mention other characters names. These mentions can be used to identify speakers by their names. We present a neuronal speaker identification approach based on mentions propagation along dialogs. We successfully identify speakers for 34% of speech turns for* The Big Bang Theory*'s script, and 34% for* Friends.

MOTS-CLÉS : *Identification du locuteur, Entité nommées, Réseau de neurones.*

KEYWORDS: *Speaker identification, Named entities, Neural networks.*

1. Introduction

Un dialogue consiste en un échange d'informations entre plusieurs individus, portant généralement sur un sujet précis. Un tour de parole correspond à un ensemble de phrases qu'un interlocuteur prononce consécutivement. La tâche d'identification du locuteur consiste à retrouver le locuteur de chaque tour de parole d'un dialogue donné.

Le traitement de la tâche peut totalement différer selon le média : vidéo (films, séries, etc.), audio (séries, radio, etc.) ou textuel (scripts, transcriptions, etc.). En fonction du cadre dans lequel on souhaite résoudre cette tâche, le locuteur peut être représenté sous différentes formes. On peut ainsi identifier un locuteur par son nom, sa voix, son visage, etc.

La plupart des approches textuelles d'identification du locuteur tentent de résoudre cette tâche en modélisant le style propre à chaque interlocuteur du dialogue (Kundu *et al.*, 2012 ; Ma *et al.*, 2017). Lors de la phase d'entraînement, un modèle stylistique est créé pour chaque interlocuteur que l'on souhaite identifier. Lors de la phase de test, le modèle d'expression du tour de parole dont on souhaite identifier le locuteur est comparé aux modèles des interlocuteurs appris afin de sélectionner le plus susceptible d'être locuteur du tour de parole. Ces méthodes nécessitent donc des connaissances stylistiques a priori sur les personnages, et restreignent notamment l'efficacité de la reconnaissance de personnages peu présents, car leurs exemples de tours de parole d'entraînement sont restreints. Les caractéristiques stylistiques des personnages ne peuvent être obtenues qu'en disposant de suffisamment d'exemples de tours de parole où les personnages s'expriment. Cependant, les transcriptions automatiques et de nombreuses transcriptions manuelles possèdent l'information de ce qui est dit mais pas par qui. Ces corpus sont donc inutilisables pour de telles approches.

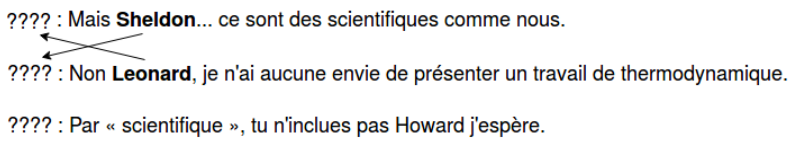

Figure 1. *Exemple de résolution de la tâche d'identification du locuteur dans un dialogue en utilisant les mentions. La mention* Howard *est ignorée car le personnage correspondant n'est pas locuteur dans le dialogue.*

Dans un dialogue de série télévisée, il est courant que les personnages mentionnent le nom des autres personnages à travers les discussions. Ces mentions peuvent servir d'indice afin d'identifier les locuteurs des tours de parole par leur nom. La figure 1 est un exemple de résolution d'identification du locuteur fondée sur la propagation des mentions faites dans le dialogue pour retrouver le nom des locuteurs. Cette méthode présente l'avantage de ne pas nécessiter de connaissances a priori sur les personnages, et permet par exemple un apprentissage sur une série et une évaluation sur une autre avec des résultats encourageants.

Dans cet article, nous explorons une approche d'identification du locuteur fondée sur la propagation des mentions dans le dialogue, en implémentant un réseau de neurones qui apprend à propager les noms dans les dialogues, indépendamment de qui ils mentionnent : par exemple, en remplaçant les mentions de la figure 1 par d'autres noms, la propagation reste similaire. La mention faite dans le premier tour de parole est propagée au second tour de parole et inversement. Les résultats sont notamment comparés aux résultats obtenus lors d'une expérience menée auprès de 22 participants à qui nous avons demandé de résoudre la tâche d'identification du locuteur dans les mêmes conditions que notre modèle.

2. État de l'art

De nombreux travaux proposent des méthodes d'identification du locuteur fondées sur des médias audiovisuels. Certains effectuent une tâche de *segmentation et regroupement en locuteur* (Anguera *et al.*, 2012) à savoir regrouper les tours de parole qui possèdent des similarités vocales entre eux, permettant d'identifier les locuteurs par leur voix. D'autres travaux fusionnent cette approche de segmentation et regroupement en locuteur avec une tâche de détection de visages (Gay *et al.*, 2014 ; Bredin et Gelly, 2016), permettant d'identifier les locuteurs par leur voix et leur visage.

Certains travaux proposent l'utilisation des transcriptions pour étiqueter les regroupements en locuteur et ainsi identifier les locuteurs par leur nom. (Bredin *et al.*, 2014) proposent d'unifier la tâche de segmentation et regroupement en locuteur avec la propagation des mentions présentes dans la transcription automatique d'émissions télévisées à travers un problème d'optimisation global, modélisé par une architecture basée sur les graphes.
(Everingham *et al.*, 2006) utilisent l'alignement du script (qui fournit l'information « qui parle ? » mais pas « quand ? ») avec les sous-titres (qui fournissent l'information « quand ? » mais pas « qui parle ? ») de séries afin d'étiqueter les visages reconnus. Ici les sous-titres et transcriptions ne sont utilisés que pour obtenir les noms des locuteurs tandis que leurs propos sont ignorés. De plus, il est rare que le script et les sous-titres soient disponibles pour un même média. (Haurilet *et al.*, 2016) proposent la propagation des mentions faites dans les sous-titres de séries associée avec une reconnaissance des visages pour étiqueter les visages des personnages. (Azab *et al.*, 2018) proposent une structure unifiée mêlant des informations acoustiques, visuelles et linguistiques (issues des sous-titres de films), ainsi qu'une propagation des mentions inspirée de (Haurilet *et al.*, 2016) pour l'identification du locuteur.

Dans ces approches, les informations textuelles apportées par les scripts et les transcriptions sont souvent utilisées comme indices pour attribuer un nom aux différents regroupements, mais très peu d'approches sont fondées sur ces informations pour la tâche d'identification du locuteur. Il existe donc très peu de références concernant la tâche d'identification du locuteur basée sur des données textuelles. (Canseco-Rodriguez *et al.*, 2004) sont parmi les premiers à utiliser des données linguistiques pour l'identification des locuteurs en exploitant des patrons de langage sur les trans-

criptions d'informations télévisées pour savoir si une mention correspond au locuteur précédent, au locuteur actuel ou au locuteur suivant, permettant la propagation des mentions et ainsi l'identification des locuteurs. (Kundu *et al.*, 2012 ; Ma *et al.*, 2017) proposent une approche de classification supervisée fondée sur les différents styles des personnages. Les auteurs extraient des descripteurs stylistiques (fréquence des mots courts, fréquence des ponctuations, etc.) pour chaque personnage afin d'assigner à chaque tour de parole un locuteur. (Kundu *et al.*, 2012) obtiennent 30% de locuteurs correctement identifiés en moyenne sur différents films avec une approche basée sur les k plus proches voisins. (Ma *et al.*, 2017) parviennent à identifier correctement le locuteur de 31% des tours de parole avec une approche basée sur un réseau de convolution pour la série *Friends.*

La tâche d'identification du locuteur est souvent utilisée pour l'enrichissement de corpus de sous-titres. Parfois, les données audiovisuelles viennent à manquer, notamment dans le cadre de transcriptions manuelles ou en cas d'alignements coûteux. Certaines situations nécessitent de recourir à des approches uniquement textuelles, sur les forums ou les *chats* par exemple. Les méthodes textuelles fondées sur les informations stylistiques obtenues pour chaque locuteur sont alors difficiles à mettre en place car elles nécessitent des connaissances a priori sur chaque locuteur potentiel.

C'est pourquoi nous proposons une approche fondée sur la propagation des mentions dans les dialogues qui ne nécessite que la liste des noms des participants du dialogue. Notre approche pourra ainsi être comparée avec celle de (Ma *et al.*, 2017) pour l'identification des locuteurs de la série *Friends.*

3. Description de l'approche

Dans cette section nous décrivons l'architecture globale de l'approche et les descripteurs utilisés pour résoudre la tâche d'identification du locuteur.

3.1. *Principe*

On appelle t_0 le tour de parole dont on souhaite connaître le locuteur. On note $T_{t_0} = \{t_{-n} \dots t_0 \dots t_n\}$ l'ensemble des $2n + 1$ tours de parole qui précèdent et succèdent t_0 en incluant t_0. n est un hyperparamètre qui correspond au nombre de tours de parole, avant et après t_0, utilisés pour retrouver le locuteur de t_0.

Pour chaque saison des séries que l'on traite, on dispose d'une liste de candidats C qui correspond à l'ensemble des personnages qui parlent dans la saison, avec $c \in C$ qui représente le nom normalisé d'un candidat sous la forme *prénom_nom.* La génération de cet ensemble C est décrite dans la section 4.1. Cette liste de candidats est la seule connaissance nécessaire a priori pour l'approche proposée.

Lors de la phase d'entraînement de notre modèle, pour un couple (t_0, c) donné, où t_0 est le tour de parole dont on souhaite trouver le locuteur et c un candidat, on

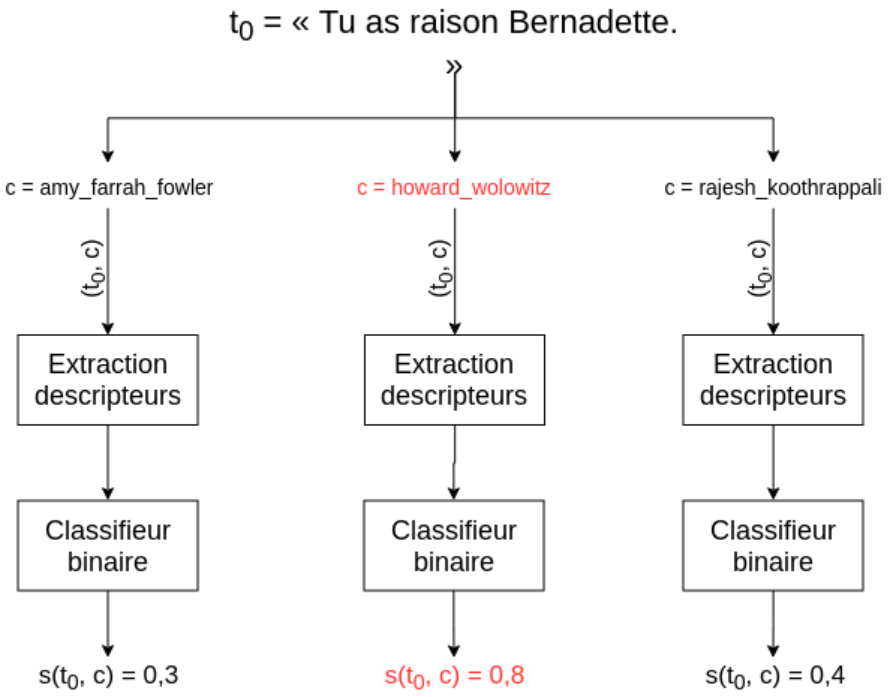

Figure 2. *Identification du locuteur d'un tour de parole t_0 parmi 3 candidats. Le candidat sélectionné par le modèle (en rouge) est celui dont le score est le plus élevé.*

extrait un ensemble de descripteurs que l'on fournit en entrée à un classifieur binaire. La sortie attendue idéale du classifieur est 1 si c est le locuteur de t_0, et 0 sinon.

Pour chaque paire (t_0, c), on calcule un score $s(t_0, c)$. Le candidat ayant le score le plus élevé est ensuite désigné locuteur de t_0 : $c_{t_0} = \text{argmax}_{c \in C}\ s(t_0, c)$. La figure 2 présente un exemple de cette approche pour un tour de parole et trois candidats.

3.2. *Extraction des descripteurs*

Pour déterminer le score $s(t_0, c)$ en se fondant sur les mentions faites dans le dialogue, il convient d'utiliser l'ensemble des tours de parole de T_{t_0}. Notamment, les tours de parole de T_{t_0} où le nom du candidat c est prononcé sont déterminants pour obtenir $s(t_0, c)$. Pour chaque couple (t_0, c), des descripteurs sont donc extraits pour chaque tour de parole t_i de T_{t_0} avec $i \in [-n, n]$, et sont ensuite concaténés.

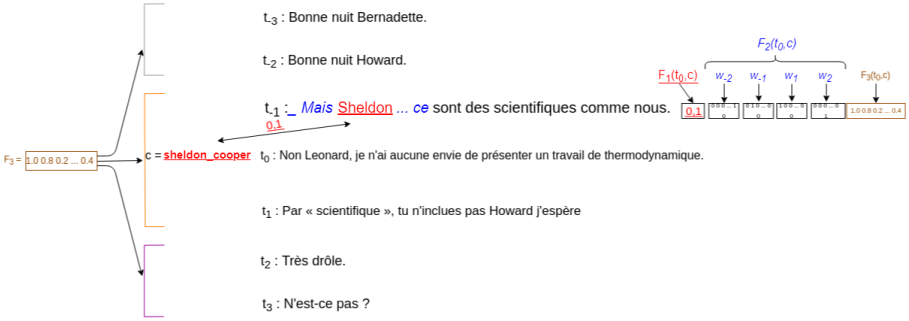

Figure 3. *Exemple d'extraction de descripteurs pour un couple (t_0, c) donné. La figure décrit l'extraction des descripteurs de t_{-1}. En rouge (souligné) est décrit le premier descripteur F_1, en bleu (italique) est décrit le deuxième descripteur F_2 et en jaune est décrit le troisième descripteur F_3.*

3.2.1. *Descripteur de mentions*

Pour chaque couple (t_0, c), on cherche les mentions qui font référence à c dans T_{t_0}. Cette approche est semblable aux tâches d'annotation sémantique mais avec une liste de candidats restreinte. Pour cela, on compare les mentions de noms de personnes faites dans les tours de parole de T_{t_0} avec c grâce à une distance d'édition. Une distance d'édition donne une indication sur la similarité entre deux chaînes de caractères, en l'occurrence, une mention et c.

On note M_{t_i} l'ensemble des mentions faites dans un tour de parole t_i. Cet ensemble est construit par une étape de reconnaissance des entités nommées. Pour un couple (t_0, c) donné, pour chaque tour de parole t_i de T_{t_0}, on note :

$$f_1(t_i, c) = \min_{m \in M_{t_i}} d(c, m)$$

$$m(t_i, c) = \operatorname*{argmin}_{m \in M_{t_i}} d(c, m)$$

$f_1(t_i, c)$ correspond au descripteur extrait par le modèle pour le couple (t_i, c). $d(a, b)$ renvoie la distance d'édition, normalisée entre 0 et 1, entre les chaînes a et b. Si M_{t_i} est vide, on pose $f_1(t_i, c)$ = 1. $m(t_i, c)$ correspond à la mention la plus proche de c pour le tour de parole t_i. Un exemple d'extraction de ce descripteur est fourni dans la figure 3 (en rouge).

Pour chaque couple (t_0, c), on obtient alors un vecteur de taille $2n+1$ qui constitue le premier descripteur de notre modèle :

$$F_1(t_0, c) = \begin{pmatrix} f_1(t_{-n}, c) \\ \dots \\ f_1(t_0, c) \\ \dots \\ f_1(t_n, c) \end{pmatrix}$$

Ce premier descripteur permet d'encoder la position des tours de parole où le nom du candidat est prononcé, mais ne fournit aucune information sur les conditions dans lesquelles les noms sont prononcés.

3.2.2. *Descripteur d'étiquetage morpho-syntaxique*

Si un tour de parole t_i commence par « Non Leonard. », il y a de fortes chances pour que le tour de parole précédent, t_{i-1}, soit prononcé par Leonard. À l'inverse, si on a t_i = « Où est parti Sheldon ? », il y a de fortes chances pour que les tours de parole autour de t_i ne soient pas prononcés par Sheldon.

Il convient donc de fournir au modèle un descripteur F_2, basé sur les mots prononcés autour d'une mention, qui va lui permettre de propager les mentions aux différents tours de parole plus finement. On note $W_{t_i,c} = \{w_{-m}, \dots, w_{-1}, w_1, \dots, w_m\}$ l'ensemble des m mots à gauche et m mots à droite de $m(t_i, c)$. Cependant, le nombre de combinaisons de m-grammes autour des mentions est très grand, et nous ne disposons pas de suffisamment d'exemples pour que notre modèle apprenne avec cette méthode. Cette méthode a été testée avec des m-grammes composés de plongements de mots et n'a pas obtenu de résultats. Pour réduire la dimension des m-grammes, il est possible, plutôt que d'extraire des m-grammes de mots, d'extraire des m-grammes de classes grammaticales. On encode les classes grammaticales $cg(w)$ des mots sous forme de vecteurs *one-hot.*

Pour un couple (t_i, c), on concatène les $2 * m$ vecteurs $cg(w_l), w_l \in W_{t_i,c}$ pour obtenir $f_2(t_i, c)$. Ainsi, les tours de parole « Où est parti Sheldon ? », « Où est allé Sheldon ? », « Où est passé Sheldon ? », etc. auront la même représentation f_2 car *parti*, *passé* et *allé* sont des participes passés. Lorsque t_i ne contient pas de mention, $f_2(t_i, c)$ est un vecteur de 0.

Pour un couple (t_0, c), on extrait $W_{t_i,c}$ pour chacun des tours de parole de T_{t_0}. On obtient alors une matrice :

$$F_2(t_0, c) = \begin{pmatrix} f_2(t_{-n}, c) \\ \dots \\ f_2(t_0, c) \\ \dots \\ f_2(t_n, c) \end{pmatrix}$$

Un exemple d'extraction de ce descripteur pour $m = 2$ est fourni dans la figure 3 (en bleu).

3.2.3. *Descripteur sémantique*

Un troisième descripteur est extrait, permettant de mieux rendre compte des similarités sémantiques des tours de parole.

Si les tours de parole t_{-3} et t_0 concernent le thème de la physique quantique et que le nom d'un candidat est prononcé dans t_{-3}, les chances que ce nom corresponde au locuteur de t_0 sont augmentées. De même, si les tours de parole t_{-10} à t_{-4} concernent un thème différent des tours de parole t_{-3} à t_{10}, il y a de grandes chances pour que les mentions faites dans les tours de parole t_{-10} à t_{-4} n'aident pas à identifier le locuteur de t_0.

Récemment, l'émergence des *plongements de phrases* (Cer *et al.*, 2018) a montré de bons résultats pour les tâches qui nécessitent des informations globales concernant les phrases. Ce descripteur consiste donc à enrichir les descripteurs précédents en ajoutant une matrice de similarité des plongements de phrases.

Pour un ensemble T_{t_0}, on extrait pour chaque tour de parole t_i le plongement de phrase p_i correspondant. On calcule ensuite la similarité cosinus entre chaque p_i : $S_{i,j} = sc(p_i, p_j)$.

$$F_3(t_0, c) = \begin{bmatrix} 0 & \dots & S_{i,j} \\ \vdots & \ddots & \vdots \\ S_{j,i} & \dots & 0 \end{bmatrix}$$

Ce descripteur est représenté dans la figure 3 (en jaune). Il est attendu de ce descripteur qu'il permette de diviser T en sous-parties, permettant d'isoler les tours de parole dont les locuteurs sont susceptibles d'interagir entre eux. Notons qu'en pratique il ne dépend pas du candidat c.

4. Expériences

Dans cette section, nous décrivons le corpus que nous avons utilisé pour nos expériences, les détails de l'implémentation ainsi que les résultats.

4.1. *Corpus*

Deux jeux de données ont été utilisés pour mener cette expérience :

– *The Big Bang Theory* : Le script des 10 premières saisons de la série *The Big Bang Theory*[1], avec le locuteur de chaque tour de parole. Ce jeu de données représente un total de 231 épisodes et plus de 48 000 tours de parole. En plus du script, un ensemble de 355 noms de personnages normalisés sous un format « prénom_nom »

1. https ://github.com/lgalmant/TBBT-corpus

ont été extraits de *l'Internet Movie Database*[2], ainsi que leurs saisons d'apparition, parmi lesquels on retrouve les 7 personnages principaux de la série. Ce jeu de données a fait l'objet d'un effort d'annotation pour la reconnaissance des entités nommées de type *personne*. Une première étape de reconnaissance des entités nommées a ainsi été effectuée puis une vérification manuelle a été réalisée pour augmenter sa précision. Pour chaque entité nommée reconnue, il a été aussi annoté si elle est prononcée à la première (« Je m'appelle Sheldon »), la deuxième (« Comment tu vas, Sheldon ? ») ou la troisième (« Sheldon n'est plus là. ») personne. De plus, pour un épisode donné, le corpus est segmenté en scènes. Nous n'utilisons cette annotation et cette segmentation que pour l'analyse d'erreurs de notre modèle.

– *Friends* : Une transcription manuelle des 10 saisons de la série *Friends*[3], avec le locuteur de chaque tour de parole, pour un total de 194 épisodes et plus de 49 000 tours de parole. Les entités nommées correspondant à des personnes ont été annotées avec le *Stanford Named Entity Recognizer (v3.9.1)* à travers l'interface *NLTK*. 414 personnages, ainsi que leurs saisons d'apparition, sont extraits pour cette série depuis *l'Internet Movie Database.*

4.2. *Implémentation*

La distance d'édition utilisée pour l'extraction du descripteur F_1 est la *constant gap penalty* (Vingron et Waterman, 1994), souvent utilisée dans l'alignement de séquences ADN, qui pénalise moins les insertions en fin de chaîne et permet donc de rendre compte du fait que la plupart de nos comparaisons se font entre « prénom » (mention dans le dialogue) et « prénom_nom » (candidat normalisé). Par exemple, la distance renvoyée en comparant *sheldon* et *sheldon_cooper* est de 1, là où une distance de Levenshtein renvoie 7. Les distances ont été normalisées sur $[0, 1]$, 0 étant la distance de deux chaînes identiques. Lorsqu'un tour de parole ne contient aucun nom prononcé, on associe à ce dernier une distance systématique de 1.

L'étiquetage morpho-syntaxique a été réalisé par le *PerceptronTagger*[4] de la bibliothèque *Natural Language Toolkit (NLTK) 3.3* avec le jeu d'étiquettes *Penn Treebank Tagset.* La trentaine de classes grammaticales sont encodées sous forme d'encodage *one-hot*, où chaque vecteur *one-hot* correspond à une classe grammaticale traitée par NLTK.

Afin d'obtenir les plongements de phrase du descripteur F_3, nous avons utilisé la bibliothèque *InferSent* (Conneau *et al.*, 2017)[5]. Les plongements de mots sont générés par *fastText* sur le corpus *Common Crawl.* Les plongements de phrases ont été entraînés sur le corpus *Stanford Natural Language Inference*.

2. https ://www.imdb.com/
3. https ://fangj.github.io/friends/
4. http ://www.nltk.org/_modules/nltk/tag/perceptron.html
5. https ://github.com/facebookresearch/InferSent

Le modèle d'apprentissage est un perceptron multicouche. Après la phase d'extraction des descripteurs, on obtient pour un couple (t_0, c) une entrée de taille L. Le modèle est constitué de deux couches linéaires de dimensions respectives $L/2$ et $L/4$. Les couches linéaires sont respectivement suivies par une fonction d'activation de type *ReLU*. Une dernière couche linéaire de dimension 1 est ensuite appliquée, suivie par une fonction *sigmoïde*, permettant d'obtenir le score $s(t_0, c)$.

L'entraînement du modèle est effectué par l'optimiseur *Adam* avec la fonction objectif *Binary Cross Entropy*.

4.3. *Points de comparaison*

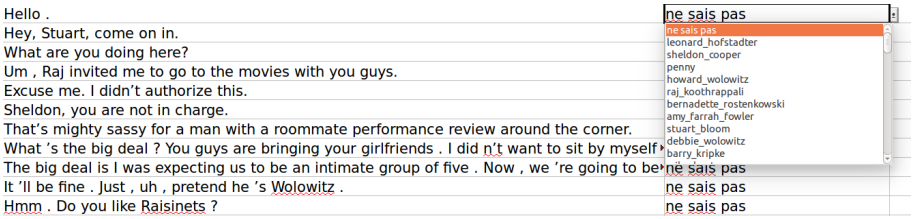

Figure 4. *Exemple de fichier Excel fourni à chaque participant de l'expérience.*

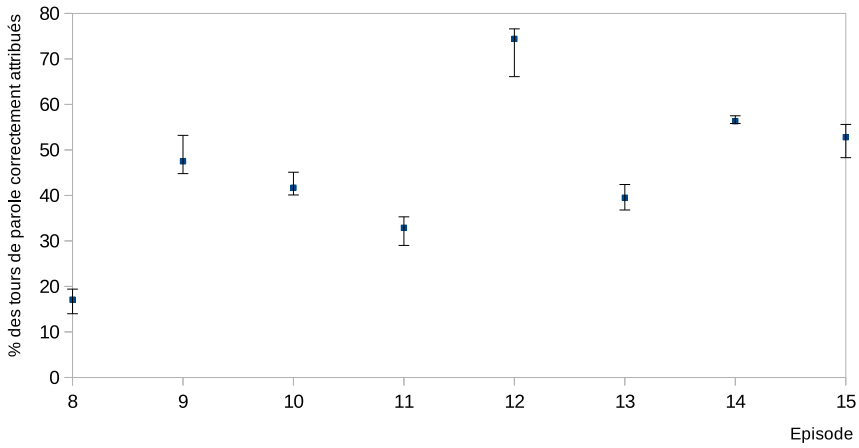

Figure 5. *Scores moyens, score minimum et score maximum obtenus par les participants de l'expérience pour les épisodes 8 à 15 de la saison 6 de* The Big Bang Theory.

Pour pouvoir comparer nos résultats avec des performances humaines, nous avons placé les participants de notre expérience dans les mêmes conditions que notre modèle.

Il est donc fourni à chaque volontaire un ou plusieurs épisodes de la série *The Big Bang Theory,* chacun sous forme d'un fichier Excel dont chaque ligne correspond à un tour de parole. Pour chaque tour de parole, une liste déroulante avec la liste des candidats est fournie. Les volontaires doivent alors associer à chaque tour de parole un locuteur. La figure 4 donne un aperçu du fichier Excel fourni aux participants de l'expérience.

Au total, 22 participants ont passé l'expérience sur un total de 16 épisodes issus de 3 saisons différentes. Nous avons volontairement fourni le même épisode à plusieurs participants pour constater le désaccord entre ces derniers.

La figure 5 présente la moyenne des scores obtenus par les participants pour les épisodes 8 à 15 de la saison 6 de *The Big Bang Theory,* ainsi que les scores minimum et maximum. La moyenne des tours de parole dont le locuteur est correctement identifié par les participants pour l'ensemble des épisodes est de 50%. On remarque une forte disparité entre les épisodes, avec une moyenne d'environ 17% pour l'épisode 8 et d'environ 74% pour l'épisode 12. Pour un épisode donné, les scores des participants sont en revanche proches les uns des autres, avec un score k de Cohen moyen de 0,91.

Les erreurs commises par les participants sont souvent dues à une mauvaise propagation de l'alternance des tours de parole dans le dialogue. Par exemple, pour une séquence où Penny et Bernadette se répondent, le participant attribue les tours de parole de Penny à Bernadette et inversement. Cela révèle à quel point l'humain a tendance à attribuer les différents tours de parole en prenant en compte la structure du dialogue.

Notre système étant fondé sur la propagation des mentions dans les dialogues, nous avons implémenté un algorithme oracle, qui identifie correctement le locuteur d'un tour de parole dès que son nom est prononcé dans les n tours de parole qui l'entourent. La figure 6 présente la courbe du taux de locuteurs correctement identifiés par l'oracle en fonction de n (en bleu). Pour un n donné, l'oracle fournit donc le score maximum que peut atteindre notre modèle.

4.4. *Résultats*

La figure 6 présente le taux de tours de parole dont le locuteur est correctement identifié par les différents modèles selon le nombre de tours de parole considérés n. Lorsque n est petit (environ < 8), les courbes des modèles ont la même allure que la courbe de l'oracle. Lorsque n augmente, la courbe de l'oracle se détache totalement des courbes des modèles. Pour comprendre cet écart, il est important de se souvenir que l'oracle parvient à identifier le locuteur de t_0 si son nom n'est par exemple prononcé qu'au tour de parole t_{18}, même lorsque les tours de parole t_0 et t_{18} ne font pas partie de la même scène. Pour notre modèle il est beaucoup plus difficile d'interpréter des mentions prononcées dans des tours de parole éloignés de t_0 et de les propager correctement.
Avec n qui augmente, le score des modèles augmente de façon sous-linéaire jusqu'à atteindre un seuil à $n = 12$. Il s'agit d'une limite empirique au-delà de laquelle le

Entraînement	Évaluation	F_1	F_1 + F_2	F_1 + F_3	F_1 + F_2 + F_3
TBBT	*TBBT*	26% ±0,03%	33% ±3,20%	28% ±0,08%	34% ±3,41%
Friends	*Friends*	17% ±0,04%	23% ±3,62%	19% ±0,10%	25% ±3,54%
Friends	*Friends**	25% ±0,04%	32% ±3,60%	27% ±0,09%	34% ±3,47%
TBBT	*Friends**	26% ±0,03%	32% ±3,43%	28% ±0,08%	34% ±3,39%

Tableau 1. *Taux moyen de tours de parole dont le locuteur est correctement identifié par notre modèle pour chaque descripteurs selon le corpus d'entraînement et d'évaluation. Pour chaque corpus on fournit la liste des personnages pour chaque saison comme candidats, sauf pour* Friends* *auquel on ne fournit que la liste des personnages principaux. Le taux moyen est calculé sur 20 échantillons pour chaque système.*

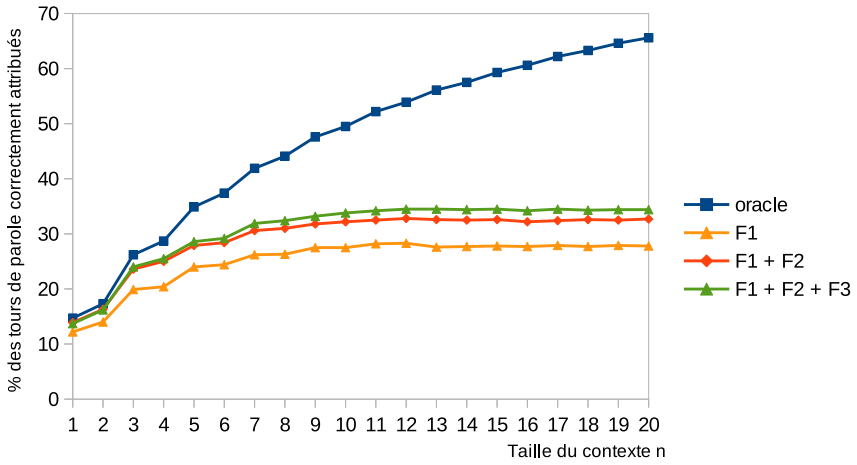

Figure 6. *Taux de tours de parole dont le locuteur est correctement identifié par nos différents modèles, ainsi que l'oracle, selon n.*

modèle ne semble plus propager les mentions correctement.
L'ajout du descripteur F_2 à F_1 permet une amélioration des résultats pour tout n. L'ajout du descripteur F_3 à F_1 et F_2 augmente légèrement les résultats du modèle à partir de $n = 7$.

Le tableau 1 représente le taux moyen de tours de parole dont le locuteur a été correctement identifié pour les différents descripteurs, selon le corpus d'entraînement

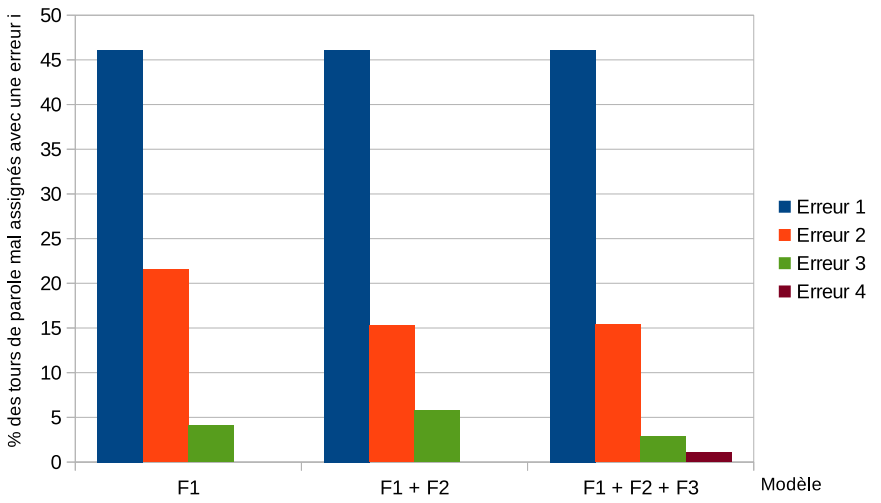

Figure 7. *Taux de chacune des erreurs sur le total d'erreurs selon le modèle pour* $n = 12$*. L'erreur 1 correspond à une mauvaise assignation du candidat locuteur par absence de mention, l'erreur 2 correspond à une mauvaise assignation d'une mention à la troisième personne, l'erreur 3 correspond à une mauvaise assignation de mention éloignée de* t_0 *et l'erreur 4 correspond à la mauvaise assignation d'une mention issue d'un tour de parole sémantiquement proche de* t_0*.*

et le corpus d'évaluation. Le corpus *Friends* est aussi testé avec une liste de candidats restreints aux personnages principaux de la série, pour nous comparer aux résultats de (Ma *et al.*, 2017). On remarque que, indépendamment des corpus, le descripteur F_2 améliore le taux de tours de parole dont le locuteur a été correctement identifié par F_1. Le descripteur F_3 apporte une légère amélioration du score lorsqu'il est assemblé avec F_1. L'écart type est trop important pour pouvoir interpréter l'apport de F_3 sur $F_1 + F_2$.

La dernière ligne du tableau représente le modèle entraîné sur le corpus *The Big Bang Theory* mais évalué sur le corpus *Friends,* avec comme candidats uniquement les noms des six personnages principaux. Les résultats prouvent que l'on peut entraîner notre modèle sur une série et l'évaluer sur une autre sans impacter négativement le score. Les deux évaluations sur le corpus *Friends* restreint à la liste des personnages principaux comme candidats obtiennent 34% de tours de parole dont le locuteur est correctement identifié. C'est un meilleur score que l'approche stylistique employée par (Ma *et al.*, 2017), qui parvient à correctement identifier le locuteur de 31% des tours de parole.

4.5. *Analyse d'erreurs*

Le modèle effectue plusieurs types d'erreurs lors de la sélection du candidat locuteur :

– Erreur 1 : Mauvaise assignation du candidat locuteur par absence de mention. Comme le montre l'oracle, même avec une propagation parfaite des mentions dans le dialogue, certains locuteurs sont impossibles à identifier du fait de l'absence de mention. Par exemple, pour $n = 20$, 35% des tours de parole restent impossibles à identifier. Pour identifier les 35% de tours de parole restants, il est nécessaire d'augmenter n ou d'associer la méthode d'identification par propagations des mentions avec une autre méthode.

– Erreur 2 : Mauvaise assignation du candidat locuteur par mention à la troisième personne. Lorsqu'une mention est prononcée à la troisième personne comme « Sheldon est parti », nous savons qu'il est peu probable que Sheldon soit locuteur du dialogue en cours. Il arrive que le modèle propage tout de même ces mentions pour assigner le candidat locuteur. Cette erreur est déterminée en mesurant le taux de mentions à la troisième personne faites lorsque le modèle se trompe.

– Erreur 3 : Mauvaise assignation du candidat locuteur par mention éloignée. Avec $t_{-20} =$ « Salut Sheldon ! », est-il vraiment possible de propager efficacement la mention *Sheldon* pour identifier t_0 ? Les scènes de la série *The Big Bang Theory* font en moyenne 22 tours de parole. Les mentions faites dans une scène différente de celle de t_0 ne sont pas pertinentes pour identifier son locuteur. Il est difficile de mesurer cette erreur en globalité. Il est cependant possible de mesurer le taux de mentions propagées issues de tours de parole d'une autre scène que t_0 lorsque le modèle se trompe.

– Erreur 4 : Mauvaise assignation du candidat locuteur par similarité sémantique. Lorsque t_i comporte une mention et est sémantiquement proche de t_0, il semble pertinent dans de nombreux cas de propager la mention. Cependant, il arrive que le modèle propage des mentions dans ces conditions sans qu'elles ne soient pertinentes. Par exemple, $t_{-1} =$ « Tu sais bien que Howard n'est pas docteur. », $t_0 =$ « Il n'a jamais eu le niveau pour être docteur. », $t_1 =$ « Je suis ingénieur au MIT, ça n'est pas suffisant ? ». Ici, le modèle propage la mention *Howard* au tour de parole t_0, sémantiquement proche de t_{-1}, au lieu de t_1. Cette erreur est déterminée en mesurant le taux de mentions propagées depuis t_i lorsque le modèle se trompe quand t_i et t_0 sont sémantiquement proches.

La figure 7 présente le taux de chaque erreur décrite sur le total d'erreurs, selon le modèle pour $n = 12$.
L'erreur 1 représente la limite de la méthode par propagation de mentions. Avec un système parfait (comme l'oracle), l'ensemble des erreurs est une erreur de type 1.
Avec l'ajout du descripteur F_2, le taux d'erreur 2 diminue. L'ajout du descripteur F_3 n'influe pas sur l'erreur 2. Le descripteur F_2 fournit au modèle des informations morpho-syntaxiques qui peuvent indiquer qu'une mention est prononcée à la troisième personne. Avec F_2, le modèle apprend à ignorer ces mentions.
Le taux d'erreur 3 augmente avec l'ajout de F_2. En effet, lorsque le modèle écarte les

mentions à la troisième personne, il doit parfois aller chercher des mentions dans des tours de parole éloignés de t_0. L'ajout du descripteur F_3 remédie à ce problème.
Le descripteur F_3 permet de diminuer le taux d'erreur 3 lorsque l'on utilise F_2, mais crée un nouveau type d'erreur (Erreur 4). Avec ce descripteur, le modèle choisit parfois de propager la mention d'un tour de parole t_i lorsqu'il est sémantiquement proche de t_0.

5. Conclusion

Cet article présente une approche d'identification du locuteur dans les scripts et transcriptions de films et séries fondée sur la propagation des mentions. Pour fonctionner sur un corpus, le modèle ne nécessite que la liste des personnages du corpus. Nous obtenons, avec cette méthode, des résultats comparables aux méthodes textuelles fondées sur la modélisation stylistique des locuteurs, qui nécessitent des connaissances a priori sur les personnages du corpus, notamment des exemples de tours de parole pour chaque personnage. Pour améliorer le modèle, il serait possible d'implémenter une étape préliminaire de recherche de candidats dans les corpus pour se défaire totalement de connaissances a priori.

Un futur travail pourrait permettre au modèle de prendre en compte l'alternance des tours de parole dans les dialogues. Une méthode serait d'appliquer un réseau neuronal récurrent qui englobe notre modèle, permettant de propager certaines mentions sur des tours de parole consécutifs deux à deux, plus proche de ce que font les humains pour traiter la tâche. Une autre méthode à explorer serait de fusionner notre modèle avec un modèle d'approche audiovisuelle.

6. Remerciements

Ce travail a été partiellement financé par l'Agence Nationale de la Recherche au travers du projet PLUMCOT (ANR-16-CE92-0025).

7. Bibliographie

Anguera X., Bozonnet S., Evans N., Fredouille C., Friedland G., Vinyals O., « Speaker Diarization : A Review of Recent Research », *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, n° 2, p. 356-370, Feb, 2012.

Azab M., Wang M., Smith M., Kojima N., Deng J., Mihalcea R., « Speaker Naming in Movies », *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, p. 2206-2216, 2018.

Bredin H., Gelly G., « Improving Speaker Diarization of TV Series using Talking-Face Detection and Clustering », *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, ACM Press, Amsterdam, The Netherlands, p. 157-161, 2016.

Bredin H., Laurent A., Sarkar A., Le V.-B., Rosset S., Barras C., « Person Instance Graphs for Named Speaker Identification in TV Broadcast », *Odyssey 2014*, vol. 2014 of *Proceedings of Odyssey : The Speaker and Language Recognition Workshop*, Joensuu, Finland, June, 2014.

Canseco-Rodriguez L., Lamel L., Gauvain J.-L., « Speaker Diarization from Speech Transcripts », p. 4, 2004.

Cer D., Yang Y., Kong S., Hua N., Limtiaco N., John R. S., Constant N., Guajardo-Cespedes M., Yuan S., Tar C., Sung Y., Strope B., Kurzweil R., « Universal Sentence Encoder », *CoRR*, 2018.

Conneau A., Kiela D., Schwenk H., Barrault L., Bordes A., « Supervised Learning of Universal Sentence Representations from Natural Language Inference Data », *CoRR*, 2017.

Everingham M. R., Sivic J., Zisserman A., « Hello ! My name is... Buffy" – Automatic Naming of Characters in TV Video », British Machine Vision Association, p. 92.1-92.10, 2006.

Gay P., Khoury E., Meignier S., Odobez J.-M., Deleglise P., « A conditional random field approach for audio-visual people diarization », *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Florence, Italy, p. 116-120, May, 2014.

Haurilet M.-L., Tapaswi M., Al-Halah Z., Stiefelhagen R., « Naming TV characters by watching and analyzing dialogs », *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Lake Placid, NY, USA, p. 1-9, March, 2016.

Kundu A., Das D., Bandyopadhyay S., « Speaker identification from film dialogues », *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, p. 1-4, Dec, 2012.

Ma K., Xiao C., Choi J. D., « Text-based Speaker Identification on Multiparty Dialogues Using Multi-document Convolutional Neural Networks », *Proceedings of ACL 2017, Student Research Workshop*, Association for Computational Linguistics, p. 49-55, 2017.

Vingron M., Waterman M. S., « Sequence alignment and penalty choice : Review of concepts, case studies and implications », *Journal of Molecular Biology*, vol. 235, n° 1, p. 1 - 12, 1994.