Détection automatique de l'âge à partir de transcriptions de l'oral

Iris Eshkol-Taravella[1] Angèle Barbedette[2] Vanessa Gaudray Bouju[1] Menel Mahamdi[1]

(1) MoDyCo, 200 Av. de la République, 92001 Nanterre, France
(2) ERTIM, 2 rue de Lille, 75007 Paris, France

ieshkolt@parisnanterre.fr, angele.barbedette@gmail.com,
v.gaudraybouju@parisnanterre.fr, mmahamdi@live.fr

MOTS-CLES : détection de l'âge, transcriptions de l'oral, âge, apprentissage supervisé.
KEYWORDS: age detection, oral transcriptions, age, supervised machine learning.

1 Introduction

L'âge constitue un des enjeux principaux de la sociolinguistique contemporaine (Wagner, 2012) en tant que facteur de variation du langage. Selon (Labov, 1966), les usages individuels se stabilisent au début de l'âge adulte, soit après l'adolescence. Deux types de langages sont souvent étudiés dans la littérature : le langage des jeunes et celui des personnes âgées. Age et normativité ont souvent été des concepts associés pour expliquer les variations entre adolescents et adultes, ces derniers ayant plus tendance à utiliser le langage sous sa forme standard (Gerstenberg & Voeste, 2015). Les études concernant les adultes plus âgés notent des marqueurs lies à des capacités cognitives et physiologiques amoindries (Kemper et al., 1992; Stine-Morrow & Payne, 2016). Dans le domaine du TAL, un certain nombre de travaux traitant l'âge dans les données orales s'intéressent à l'identification du locuteur (Bonastre et al., 2000; Przybocki & Martin, 1999), à la reconnaissance automatique de la parole de personnes âgées (Aman, 2014) ou à la classification de locuteurs par l'âge (Naini & Homayounpour, 2006). Tous ces travaux sont fondés sur des critères acoustiques et phonétiques de la parole.

Les travaux sur la détection de l'âge menés à partir de corpus de l'écrit se concentrent principalement sur la communication en ligne : les chats et forums (Tam & Martell, 2009) ou les posts sur divers réseaux sociaux (Simaki et al., 2016; Pentel, 2015a,b; Nguyen et al.,

2011; Demmelmaier & Westerberg, 2021; Van de Loo et al., 2016) afin de détecter les prédateurs dans la lutte pour la protection des mineurs (Van de Loo et al., 2016), ou d'étudier des préférences de certaines communautés (Alroobaea et al., 2020). Parmi ces travaux, le modèle de reconnaissance de l'âge développé par (Tam & Martell, 2009) obtient une F-mesure de 0.996 pour distinguer les adolescents des adultes, en utilisant un classifieur de type SVM, des traits linguistiques tels que les n-grammes et le nombre de mots et d'émoticônes, et six tranches d'âge de départ : 13-19, 20-29, 30-39, 40-49, 50-59 et 60+. Par ailleurs, ces différents travaux montrent qu'il est possible d'aborder la question de la détection de l'âge de deux façons différentes : en la considérant comme une tache de classification, mettant en jeu un nombre plus ou moins grand de classes d'âge, ou en la traitant comme une tache de régression, cherchant à approximer au mieux l'âge d'un locuteur donné.

2 Données

Le premier corpus utilisé dans cette étude est ESLO (Baude & Dugua, 2011; Eshkol-Taravella et al., 2011). Le jeu de données est constitué de 90 transcriptions des entretiens guidés, menés par des chercheurs en sociolinguistique avec 92 locuteurs. Un total de 136267 tours de parole a été extrait de ces fichiers. Les fiches des locuteurs et les transcriptions d'enregistrements sont complétées par des métadonnées comportant des informations telles que la catégorie socioprofessionnelle, le métier, le niveau d'études ou encore l'âge. Pour augmenter le nombre de jeunes locuteurs, nous avons choisi le corpus MPF (Gadet & Guerin, 2016), composé d'entretiens menés avec des jeunes de banlieue parisienne. Le corpus LangAge (Ismail et al., 2022) a été pris en compte pour augmenter les données de la tranche d'âge des plus âgés. Le nombre de tours de parole attribués à une tranche d'âge dans le jeu de données extraits à partir des trois corpus ESLO, MPF et LangAge est de 187720 au total, avec 32383 énoncés pour la classe -30, 65450 énoncés pour les 30-60 et 89887 énoncés pour les 60+. Les observations manuelles du corpus, à savoir l'essai de classification manuelle des tours de paroles dans trois classes prédéfinies : -30, 30-60, 60+, montrent que les interjections telles que *ouais*, *bah*, *bon*, les termes familiers comme *truc*, *vachement*, *machin*, *tu vois*, etc. sont utilisés par les plus jeunes alors que l'imparfait ainsi que certaines thématiques « la retraite », « les petits-enfants » sont attestés chez les personnes âgées.

3 Apprentissage supervisé

En premier lieu, nous avons testé la méthode de l'apprentissage de surface avec les SVM utilisés déjà dans (Tam & Martell, 2009). Le jeu de données utilisé a été constitué de 168

items à classer (41 pour la classe -30, 47 pour classe 30-60 et 80 pour la classe 60+) et a été représenté à l'aide d'une normalisation avec des poids TF-IDF prenant en compte les unigrammes et les bigrammes. Une série de traits linguistiques a été définie, se basant sur les travaux antérieurs, les premières observations du corpus et nos intuitions : la longueur moyenne des mots de l'énoncé et le nombre de caractères contenus dans l'énonce, le nombre d'amorces présentes au sein d'un énoncé et la répétition de mots, la présence de néologismes ou de mots à la graphie incertaine, du mot *quoi* en fin d'énoncé, d'adverbes en -ment et de connecteurs logiques. Nous avons testé la classification sans intégration de traits linguistiques, avec tous les traits linguistiques et en ne sélectionnant que des traits linguistiques significatifs (le nombre de connecteurs logiques par énoncé, la présence de quoi en fin d'énoncé et le nombre de répétitions par énoncé) dont la valeur de significativité p était inférieure au seuil standard de 0.05 pour la classification.

Nous nous sommes tournés ensuite vers l'apprentissage profond et avons utilisé le modèle CamemBERT (Martin et al., 2019) avec une fenêtre de 30 énoncés (un item). Le jeu de données est constitué de 3284 items découpés en 2305 items pour l'entrainement et 989 items pour le test. L'ajustement des paramètres a été fait selon la méthode d'optimisation Adam et les principaux hyperparametres utilisés sont le nombre d'épochs (8), la longueur maximum de la phrase (200), la taille du batch (12) et le taux d'apprentissage (2e-5), une couche de sortie (3 neurones).

La détection de l'âge d'un locuteur est traitée traditionnellement comme un problème de régression, l'âge étant une valeur continue et linéaire. C'est la raison pour laquelle le modèle de classification de CamemBERT a été adapté afin de correspondre à la tâche de régression. La fonction d'activation de la dernière couche du modèle a été supprimée pour que les sorties correspondent directement aux âges prédits. Nous avons choisi la répartition des données suivante : 70 % l'entrainement, 15 % la validation et 15 % l'évaluation. Différents tests ont été effectués en jouant sur trois paramètres principaux : la fenêtre de tours de parole par item (20 ou 30), le nombre minimal de tours de parole par item, la répartition des locuteurs au sein du jeu d'entrainement.

Ces trois méthodes ont été testées également sur un nouveau jeu de données issu du CFPP2000 (Branca-Rosoff et al., 2000). Le corpus utilisé comprend 37 transcriptions de 61 locuteurs (14 appartenant à la classe -30, 27 ayant 30-60 ans et 20 faisant partie des 60+), soit un total de 45811 tours de parole.

Les résultats montrent que l'utilisation des techniques d'apprentissage supervisé à l'aide des SVM obtient environ 87 % de bonnes classifications sur un jeu de test issu des trois corpus de départ et 62 % sur un échantillon test de données issues d'un nouveau corpus. Le modèle pré-entrainé CamemBERT atteint 82 % de bonnes classifications mais ne permet pas de généraliser sur un autre corpus. Les résultats de l'utilisation d'un modèle de régression sont globalement inferieurs avec 65.6 % de bonnes prédictions sur un jeu de test issu des trois

corpus utilisés pour l'entrainement, avec une marge d'erreur de 10 ans entre âge réel et prédit et 53.7 % de bonnes prédictions sur un échantillon test de données issues d'un nouveau corpus. Néanmoins, il demeure que la régression permet de mieux approximer l'âge d'un locuteur, palliant les limites qu'entraine le fait de se restreindre à trois classes assez vastes qui comprennent des âges avec parfois jusqu'à 30 ans de différence. La marge d'erreur fixée à 10 ans permet une approximation plus précise de l'âge des locuteurs.

Références

ALROOBAEA R., ALMULIHI A. H., ALHARITHI F. S., MECHTI S., KRICHEN M. & BELGUITH L. H. (2020). A deep learning model to predict gender, age and occupation of the celebrities based on tweets followers. In *CLEF* (Working Notes).

AMAN F. (2014). *Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile*. Thèse de doctorat, Université de Grenoble. HAL : tel-01347155

BAUDE O. & DUGUA C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus*, (10), p. 99–118. HAL : hal-01162479

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BONASTRE J.-F., DELACOURT P., FREDOUILLE C., MEIGNIER S., MERLIN T. & WELLEKENS C. (2000). Différentes stratégies pour le suivi du locuteur, reconnaissances des formes et intelligence artificielle. In *RFIA 2000*.

BRANCA-ROSOFF S., FLEURY S., LEFEUVRE F. & PIRES M. (2000). *Discours sur la ville. Corpus de français parlé parisien des années 2000*. HAL : halshs-00550127

DEMMELMAIER G. & WESTERBERG C. (2021). *Data segmentation using NLP : Gender and age*.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

ECKERT P. (2017). Age as a sociolinguistic variable. *The handbook of sociolinguistics*, p. 151–167. DOI : 10.1002/9781405166256.ch9

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral ≪disponible≫ : le corpus d'Orléans 1 1968-2012, *Revue TAL : traitement automatique des langues, 2011, Ressources Linguistiques Libres,* 53 (2), pp.17-46. HAL : halshs-01163053

GADET F. & GUERIN E. (2016). Construire un corpus pour des façons de parler non standard :≪Multicultural Paris French≫. *Corpus*, (15), p. 285-307. HAL : halshs-01658279

GERSTENBERG A. (2015). 14 langues et générations : enjeux linguistiques du vieillissement. In *Manuel de linguistique française*, p. 314–333. De Gruyter. DOI : 10.1515/9783110302219-016

GERSTENBERG A. & VOESTE A. Éds. (2015). *Language development: The lifespan perspective*, volume 37. John Benjamins Publishing Company.

ISMAIL E. E. S., GERSTENBERG A., SPAGNOLO M. L., SCHULZ F. & VANDENBROUCKE A. (2022). L'âge avance en perspective longitudinale et ses outils : Langage, un corpus au pluriel. *Actes du 8e Congrès Mondial de Linguistique Française (CMLF),* SHS Web of Conferences, volume 138, p. 10003 : EDP Sciences. DOI : 10.1051/shsconf/202213810003

KEMPER S., KYNETTE D. NORMAN S. (1992). Age differences in spoken language. In West, R.L., Sinnott, J.D. (eds) *Everyday Memory and Aging.* Springer, New York, NY. DOI : 10.1007/978-1-4613-9151-7_9

LABOV W. (1966). *The linguistic variable as a structural unit.* Report resumes.

MARTIN L., MULLER B., SUAREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE E. V., SEDDAH D. & SAGOT B. (2019). *Camembert : a tasty French language model*. arXiv preprint arXiv :1911.03894.

NAINI A. S. & HOMAYOUNPOUR M. (2006). Speaker age interval and sex identification based on jitters, shimmers and mean mfcc using supervised and unsupervised discriminative classification methods. In *2006 8th international Conference on Signal Processing*, volume 1 : IEEE. DOI : 10.1109/ICOSP.2006.345516.

NGUYEN D., SMITH N. A. & ROSE C. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, p. 115–123.

PENTEL A. (2015a). Automatic age detection using text readability features. In *Educational Data Mining* (Workshops).

PENTEL A. (2015b). Effect of different feature types on age based classification of short texts. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Corfu, Greece, 1–7 : IEEE. DOI: 10.1109/IISA.2015.7388069.

PRZYBOCKI M. A. & MARTIN A. F. (1999). The 1999 nist speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking. In *Sixth European Conference on Speech Communication and Technology*, Budapest.

SIMAKI V., MPORAS I. & MEGALOOIKONOMOU V. (2016). Evaluation and sociolinguistic analysis of text features for gender and age identification. *American Journal of Engineering and Applied Sciences*, 9(4), 868–876.

STINE-MORROW E. A. & PAYNE B. R. (2016). Age differences in language segmentation. *Experimental Aging Research*, 42(1), 83–96. DOI: 10.1080/0361073X.2016.1108751

Tam J. & Martell C. H. (2009). Age detection in chat. In *2009 IEEE International Conference on Semantic Computing*, p. 33–39 : IEEE. doi: 10.1109/ICSC.2009.37.

Van De Loo J., De Pauw G. & Daelemans W. (2016). Text-based age and gender prediction for online safety monitoring. *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, 5(1), 46–60.

Wagner S. E. (2012). Age grading in sociolinguistic theory. *Language and Linguistics Compass*, 6(6), 371–382. doi: 10.1002/lnc3.343