Université Paris Nanterre
200 av. de la République
92001 Nanterre Cedex
www.parisnanterre.fr

Membre de l'université Paris

Connaissance, Langage, Modélisation (CLM) – ED 139
MODAL'X – UMR 9023

El Mehdi Issouani

Modèles et algorithmes de simplification automatique de textes

Thèse présentée et soutenue publiquement le 23/06/2023
en vue de l'obtention du doctorat de Mathématiques appliquées et applications des mathématiques de l'Université Paris Nanterre
sous la direction de M. Patrice Bertail (Université Paris Nanterre)

Jury :

Rapporteur :	M. Amor Keziou	MCF (HDR), Université de Reims Champagne-Ardenne
Rapporteure :	Mme Estelle Kuhn	DR (HDR), INRAE (Jouy-en-Josas)
Membre du jury :	M. Antoine Chambaz	PR, MAP5, Université Paris Cité
Membre du jury :	Mme Delphine Battistelli	PR, MODYCO, Université Paris Nanterre
Membre du jury :	M. Jean-François Pradat-Peyre	PR, Université Paris Nanterre
Membre du jury :	Mme Marianne Clausel	PR, IECL - Université de Lorraine

Résumé

Le phénomène de la surdité à la naissance ou en bas âge pose des questions fondamentales sur l'accès à la langue, notamment via les expériences d'interaction précoce et le développement des compétences en lecture et écriture. Selon une étude menée par Swanwick et al. (2005) [29], la plupart des enfants sourds ont des difficultés de lecture en raison de leur déficit expérientiel et linguistique au début de leur vie, avec un développement cognitif et des compétences linguistiques insuffisantes (voir Quigley et Paul (1984) [26], Quigley (1984) [24], Marc Marschark and Patricia Elizabeth Spencer (2010) [18])). Kelly et al. (1996) [12], et Alegria (2004) [1] ont signalé que les lecteurs sourds ont du mal à tirer pleinement parti de leur vocabulaire jusqu'à ce qu'ils atteignent un niveau de compétence syntaxique adéquat à des âges plus avancés. De plus, il a été constaté que ce déficit entre enfants sourds et entendants se creuse au fil du temps (Harris (1994) [10] et Mahapatra (2016) [17]). Des études ont montré que les constructions passives, les propositions relatives, les conjonctions et les pronoms sont des éléments qui affectent la compréhension (voir Robbins & Hatcher (1981)).

Conrad (1979) [8] a montré que le niveau de lecture médian pour l'ensemble d'une population de malentendants (spécifiquement, tous les élèves quittant une école spéciale pour enfants malentendants en Angleterre et au Pays de Galles entre 1974 et 1976) était celui d'un enfant de 9 ans. Pour les personnes présentant une perte auditive supérieure à 86 dB, environ 50% des élèves étaient totalement illettrés. Enfin, même si l'on suppose qu'un niveau fonctionnel de compréhension n'est pas atteint avant un âge de lecture[1] de 11 à 12 ans, moins de 15% de cette population atteignait ce niveau. De plus, un niveau limite de lecture est atteint vers la troisième année d'apprentissage de la lecture (voir également Paul et Jackson (1994) [22], Jesus Alegria (2004) [1] et Quigley et al. (1977) [25]). Au final, à tout âge, les personnes ayant une surdité sévère depuis la naissance ou dès leur jeune âge ont souvent des problèmes de lecture et de compréhension des textes.

Dans de nombreux cas, la syntaxe des sites Web, y compris les sites administratifs pourtant indispensables aux citoyens est inadaptée au public souffrant de déficience auditive précoce. Par ailleurs, les services d'assistance vocaux ne peuvent suppléer à cette inadaptation des sites web car ils sont, eux aussi, par nature, inadaptés à ce public. Pour s'en convaincre on regardera la

[1] L'âge de lecture désigne le niveau de lecture ou de compréhension de la lecture d'une personne sans déficience (surdité, aphasie, dyslexie, etc.). Un âge de lecture de 8 ans représente le niveau de lecture moyen d'enfants de 8 ans n'ayant aucune déficience.

vidéo YouTube[2] où la responsable de l'entreprise DALiNK, Lynda Robillard, simule un appel à un opérateur téléphonique (ou dans d'autres conférences aux impôts, à des services administratifs ou à Enedis). Selon une enquête OMS[3] de 2021, près d'un million d'enfants naissent chaque année avec une surdité invalidante. En France, 6% des 15-24 ans sont concernés par le déficit auditif invalidant (perte d'audition supérieure à 40 dB pour un adulte et 30 dB pour un enfant). Contrairement à la déficience visuelle, ce problème est mal diagnostiqué, mal corrigé. Pourtant, la déficience auditive a des répercussions importantes sur la vie quotidienne.

Les salons de "chat" en langue des signes [4] soit n'existent pas, soit ne sont pas suffisamment nombreux pour la population concernée ou encore limités à certaines plages horaires. À ce jour, les acteurs du web ne sont pas encore tous conscients de ce type de problème.

Initialement, le but de cette thèse était de contribuer en partenariat avec la startup DALiNK[5] à la simplification de sites web à l'usage des malentendants, afin de garantir un accès équitable à l'information pour les personnes sourdes et malentendantes. Toutefois, il est apparu rapidement qu'il n'existait aucun corpus adapté construit avec des personnes malentendantes (ou trop insuffisant pour être utilisé). Aussi nous n'avons pas traité directement du processus d'adaptation à la surdité, ni d'un point de vue théorique, ni d'un point de vue appliqué. Nous nous sommes plutôt penchés sur le problème de la simplification automatique de texte dans un cadre général incluant les personnes entendantes et malentendantes. Cela inclut donc celles ayant d'autres problèmes impactant leur niveau de lecture ou de compréhension de la langue (dus à divers retard, aphasie, dyslexie etc.) et les locuteurs non natifs d'une langue.

Ainsi, le but de la thèse est de contribuer aux méthodes de simplification automatique du texte. Plus précisément, il s'agit de construire des mesures de complexité (classifieur binaire de

[2] https://www.youtube.com/watch?v=2qLYbVn_ehU

[3] https://www.who.int/fr/news-room/fact-sheets/detail/deafness-and-hearing-loss

[4] L'État a développé un nouvel outil, ANAE, qui est un assistant numérique d'accessibilité. Il est désormais disponible sur les pages "Pass vaccinal" et "Vaccin" de l'espace Covid-19. ANAE est un chatbot qui utilise l'intelligence artificielle pour générer une animation en langue des signes française (LSF) et des sous-titres pour les personnes sourdes ou malentendantes. Cette initiative s'inscrit dans l'effort du gouvernement français dans le cadre de la directive européenne du 17 avril 2019 sur l'inclusion des personnes handicapées et notamment de l'accessibilité aux services.

[5] DALiNK (https://bootcamp.dalink.fr/) construit des stratégies pour développer une communauté, en proposant des services opportuns, en identifiant les centres d'intérêts et en étudiant les nouvelles combinaisons gagnantes pour mieux servir les intérêts de la communauté et réagir efficacement aux nouvelles contraintes ou opportunités de marché.

textes simples vs textes complexes) et de contribuer au développement de modèles de langage prédictif (comme ceux utilisés dans le désormais célèbre chatbot ChatGPT). Pour ce faire, nous nous sommes intéressés aux méthodes d'entropie utilisées en NLP, que nous réinterprétons en terme de vraisemblance empirique généralisée et au comportement de la statistique de Hotelling en grande dimension. Cette statistique apparaît naturellement dans ce type de problème et permet d'effectuer des tests de moyenne en grande dimension (c'est-à-dire lorsque la dimension q des entrées $(X_i)_{i\in\{1,\dots,n\}}$ dépasse le nombre d'observations n).

L'objectif de ce projet transversal de mathématiques appliquées et de linguistique est d'apporter un éclairage aux problèmes de simplification automatique de texte, ce qui potentiellement pourrait à termes aider les personnes ayant une déficience auditive à faire face aux difficultés rencontrées. Les travaux sont situés dans des champs variés (statistique, linguistique, informatique) et revêt un caractère fortement pluridisciplinaire. Cette thèse comporte quatre chapitres que nous allons maintenant présenter.

Plan et contributions de la thèse

Chapitre 1

L'objectif du premier chapitre est d'une part d'introduire les principaux concepts du NLP (traitement automatique du langage naturel) telles que l'analyse syntaxique ou la classification de textes et d'autre part de présenter les formalismes utilisés pour modéliser les problèmes liés à l'analyse de texte (lexicale, syntaxique, sémantique, etc.). Ceci réduit de fait fortement notre champ d'investigation, puisque nous n'aborderons donc que marginalement les questions liées au traitement automatique du langage naturel, et pratiquement pas les question liées à l'adaptation de textes pour les déficients auditifs.

Ce chapitre s'adressant principalement aux lecteurs non familiers avec ce domaine, nous rappellerons les principales définitions et le vocabulaire utilisés en linguistique et en linguistique informatique (appelée aussi linguistique computationnelle).

Nous passerons succinctement en revue les méthodes existantes de constitution de corpus. Il s'agit d'un passage obligé pour qui veut développer des modèles d'analyse de texte ou de production et de génération de texte par des modèles prédictifs (de type chatGPT). Nous présenterons d'abord les théories linguistiques disponibles et les outils informatiques qui permettent leur mise en œuvre, ainsi que leur évolution dans le temps. Ensuite, afin de se familiariser avec les concepts du traitement automatique du langage naturel (TALN), nous poursuivrons cet exposé par la présentation de quelques méthodes largement utilisées pour l'exploration, l'analyse syntaxique de textes, ainsi que leur automatisation rendue possible grâce aux modèles statistiques et aux méthodes d'apprentissage automatique.

Ce premier chapitre sert également de guide pour l'étude, l'analyse et la synthèse selon des processus de découpage et de transformations classiques. Nous détaillerons notamment les processus de tokenisation (qui permet de découper le texte en petits morceaux porteurs de sens appelés tokens), puis du pos-tagging (qui permet d'attribuer une classe grammaticale à chaque token) et enfin de chunking et de parsing définis plus tard. Quelques aspects techniques de ces méthodes ainsi que leurs limites sont abordés ici.

Chapitre 2

Nous débutons ce chapitre par l'exposé des principaux modèles prédictifs utilisés pour réaliser automatiquement l'étiquetage morpho-syntaxique, sans entrer de manière approfondie dans l'écriture mathématique des objets associés. L'idée est d'estimer une probabilité conditionnelle d'un token ou d'un tag conditionnellement à un contexte construit à partir de mots ou d'une séquence textuelle (de lettres, de stems, de mots ou de phrases). En particulier, la méthode d'entropie maximale employée par Ratnaparkhi (1996) [27] permet de construire des modèles de classifications simples qui reviennent à estimer des modèles log-linéaires ou des modèles logit.

Nous montrons en quoi ces méthodes peuvent s'interpréter comme des méthodes de vraisemblances empiriques généralisées (construites pour l'entropie relative) sous l'existence d'un grand nombre de contraintes s'interprétant comme des moments. Un des premiers résultats de cette thèse est de proposer des extensions de ces méthodes de vraisemblance empirique généralisée capables de s'appliquer à la majorité des tâches utilisées en TALN et de donner des formules explicites (asymptotiques) pour les probabilités conditionnelles (de tokens ou de tags), y compris lorsque le nombres de contraintes est plus grand que le nombre d'observations.

Pour cela, nous combinerons des méthodes de vraisemblance empirique généralisée pénalisée [21, 20] avec des techniques d'extraction de "features" (ou caractéristiques), qui permettent de projeter des données textuelles dans des espaces numériques. Nous donnerons ensuite deux applications de ces approches :
1) le POS-tagging (l'étiquetage de parties du discours), qui a été largement étudié ces 20 dernières années (voir Ratnaparkhi et al. (1996) [27], Borthwick et al. (1998) [3], Thorsten Brants (2000) [4], Collins (2002) [7], Guyon et al. (2008) [9], Yogatama (2015) [31])
2) la classification de texte dans le cadre de la simplification automatique de texte qui fera l'objet de développements dans le chapitre 4.

Étant donné que les contraintes (ici des moyennes de caractéristiques) appartiennent à un espace de grande dimension, nous proposons l'utilisation de méthode de pénalisation basée sur la représentation duale du problème d'origine (selon les principes utilisés par Otsu (2007) [19], Chang et al. (2018) [5] et Shi (2016) [28]). Enfin, la même approche sera utilisée pour construire un classificateur qui prend une phrase en entrée et renvoie une sortie binaire indiquant si la phrase est complexe "0" ou simple "1", ce qui permet de construire un indicateur préalable de

complexité.

Chapitre 3

Nous obtenons des inégalités avec des bornes exponentielles pour la statistique de Hotelling T_n^2, qui prennent en compte le phénomène de grande dimension du problème. Nous explorons les propriétés de la fonction de survie de ces statistiques pour des échantillons finis/à horizon fini en dérivant des bornes exponentielles pour les distributions symétriques ainsi que pour les distributions générales sous des hypothèses de moments faibles (nous ne supposons jamais l'existence de moments exponentiels). Pour cela, nous utilisons un estimateur pénalisé de la matrice de covariance et proposons un choix optimal pour la pénalité.

Dans de nombreuses applications telles que le traitement automatique du langage naturel, la dimension du paramètre d'intérêt q est plus grande que la taille de l'échantillon n et croît avec n. Considérons par exemple le problème de l'estimation ou du test d'une moyenne de variables dans $\mathbb{R}^q$, avec $q > n$; dans ce cas, la matrice de covariance empirique n'est pas de plein rang et ne converge pas vers la vraie matrice lorsque n tend vers l'infini (voir Johnstone (2001) [11]). Par conséquent, les tests habituels de Hotelling T_n^2 dans un cadre de grande dimension ne sont plus valables. Il est donc important de construire des estimateurs et des procédures de test qui prennent en compte les aspects de grande dimension du problème (comme cela a été fait par exemple par Ledoit et Wolf (2000, 2022) [14, 15], voir également les références de ce travail). Une proposition pertinente est d'utiliser un estimateur pénalisé, non singulier, de la matrice de covariance au lieu de la matrice variance-covariance empirique dans les tests. Dans cet esprit, Chen et al. (2011) [6] ont obtenu des tests de Hotelling T_n^2 régularisées asymptotiquement valides pour la moyenne dans le cas gaussien, dans un cadre de grande dimension, lorsque n et $q \equiv q(n)$ tendent vers l'infini à partir d'un certain rang. Li et al. (2020) [16] ont étendu ces résultats à certaines distributions sous-gaussiennes spécifiques. Le but de ce chapitre (qui fait l'objet d'un article soumis à Journal of Multivariate Analysis) est d'approfondir ces propriétés de ces tests pour des échantillons à n fixe. Ainsi nous obtenons des bornes exponentielles pour la statistique de Hotelling T_n^2 correctement régularisée, dans le cas de distributions générales, en supposant l'existence de très peu de moments.

Pour cela, nous dérivons des bornes exponentielles pour la statistique de Hotelling T_n^2 régularisée dans l'esprit de Bertail et al (2008) [2], qui ont obtenu des bornes pour des formes quadratiques auto-normalisées (ou la statistique de Hotelling T_n^2) lorsque $q < n$. Nous montrons que pour les distributions symétriques, seuls l'existence des moments d'ordre 2 est requise et

nous ne supposons l'existence de moments d'ordre 8 que pour les distributions générales. Dans la mesure où ces résultats sont originaux et forment le coeur théorique de la thèse, nous en détaillons quelques résultats ci-dessous.

Principaux résultats théoriques du chapitre 3

Soient $Z, Z_1, ..., Z_n$ des vecteurs aléatoires i.i.d. centrés avec une distribution de probabilité P, définis sur un espace de probabilité $(\Omega, A, \mathbb{P})$ avec des valeurs dans $\left(\mathbb{R}^{q(n)}, \mathcal{B}, P\right)$ doté de $\|.\|_2$ la norme L_2. Nous notons $\mathbb{E}$ l'espérance sous P. On pose $Z^{(n)} = (Z_i)_{1 \leq i \leq n}$. Lorsque n et $q(n)$ tendent vers l'infini, on remarque que $(Z^{(n)})_n$ définit en fait un tableau triangulaire de variables aléatoires de dimensions variables. Cependant, comme nous nous intéressons aux propriétés des échantillons à distance finie, nous laisserons tomber la dépendance en n. En particulier, nous utilisons q au lieu de $q(n)$. La matrice de covariance de l'observation est donnée par $S^2 = \mathbb{E}(ZZ')$, où Z' est la transposée de Z et S la racine carrée de S^2. La moyenne empirique de l'échantillon est donnée par $\bar{Z}_n = n^{-1}\sum_{i=1}^n Z_i$ et la matrice de covariance empirique est définie ici par

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^{'}.$$

Nous rappelons que le T_n^2 de Hotelling, qui peut être considéré comme une forme quadratique de sommes auto-normalisées, est donné par

$$T_n^2 = n\bar{Z}_n' S_n^{-2} \bar{Z}_n,$$

avec, lorsque $q < n$, $S_n^{-2} = \left(S_n^2\right)^{-1}$. Soient ρ_1 et ρ_2 deux nombres réels strictement positifs. Nous considérons des estimateur de S^2 défini par $\Sigma_n^2(\rho_1, \rho_2)$, combinaison linéaire de la matrice identité avec la matrice de covariance de l'échantillon

$$\Sigma_n^2(\rho_1, \rho_2) = \rho_1 I_q + \rho_2 S_n^2,$$

avec I_q la matrice identité de taille q.

Dans ce qui suit, nous nous intéressons à la statistique de Hotelling T_n^2 qui utilise une combinaison linéaire de la matrice de covariance empirique de l'échantillon et de l'identité, que

nous appelons maintenant la statistique de Hotelling T_n^2 régularisée, définie par

$$T_n^2(\rho_1,\rho_2) = n\bar{Z}_n'\Sigma_n^{-2}(\rho_1,\rho_2)\,\bar{Z}_n$$

généralisant la proposition de Chen et al (2011)[6].

Bornes pour la fonction de survie d'une Hotelling T_n^2 régularisée

Nous obtenons dans un premier temps des inégalités Oracles dans le cas d'une distribution symétrique.

Théorème 1 *Supposons que Z a une distribution symétrique, avec une matrice de variance-covariance finie, nous avons alors, sans aucune hypothèse de moment supplémentaire, pour n'importe quel $n > 1$, pour $t > n$, quelque soit $\rho_1, \rho_2 > 0$,*

$$\begin{aligned}\mathbb{P}\left(T_n^2\left(\frac{\rho_1}{\rho_2},1\right) \geq t\right) &= \mathbb{P}\left(n\bar{Z}_n'\Sigma_n^{-2}(\rho_1,\rho_2)\,\bar{Z}_n \geq \frac{t}{\rho_2}\right)\\ &\leq \frac{2e^3}{9}\bar{F}_n(t)\\ &\leq \frac{2e^3}{9}\exp\left(-\frac{(t-n)^2}{4t}\right),\end{aligned}$$

où F_n est la fonction de répartition d'une distribution de $\chi^2(n)$.
De plus, pour tout $\rho > 0$, on a

$$\begin{aligned}\mathbb{P}\left(\frac{T_n^2(\rho,1)-n}{\sqrt{2n}} \geq t\right) &= \mathbb{P}\left(\frac{n\bar{Z}_n'\Sigma_n^{-2}(\rho,1)\,\bar{Z}_n - n}{\sqrt{2n}} \geq t\right)\\ &\leq \frac{2e^3}{9}\exp\left(\frac{-t^2}{2\left(1+\sqrt{2}\frac{t}{\sqrt{n}}\right)}\right).\end{aligned}$$

Nous pouvons obtenir une meilleure borne pour les statistiques de Hotelling T_n^2 régularisées et pénalisées en nous appuyant sur les résultats de Pinelis (1994) [23] et Laurent et Massart (2000) [13] (voir p.24 de leur article) qui contrôlent la queue de distribution d'une somme pondérée de variables aléatoires indépendantes suivant une loi de $\chi^2(1)$.

Soit $\lambda = (\lambda_j)_{j=1,\ldots,q} \in \mathbb{R}_+^q$ les valeurs propres de S_n^2 (classées par ordre décroissant). Nous définissons pour tout $\rho_1, \rho_2 > 0$, les dimensions effectives suivantes (voir [6] pour d'autres

expressions de ces quantités) :

$$\Theta_1(\lambda,\rho_1,\rho_2) = \sum_{j=1}^{\inf(n,q)} \frac{\lambda_j}{\rho_1+\rho_2\lambda_j}$$
$$\Theta_2(\lambda,\rho_1,\rho_2) = \sqrt{\sum_{j=1}^{\inf(n,q)} \frac{\lambda_j^2}{(\rho_1+\rho_2\lambda_j)^2}}$$
$$\Theta_\infty(\lambda,\rho_1,\rho_2) = \sup_{1\le j\le \inf(n,q)} \left(\frac{\lambda_j}{\rho_1+\rho_2\lambda_j}\right).$$

Théorème 2 *Supposons que la distribution de Z soit symétrique. Nous avons alors, sans aucune hypothèse de moment, pour tout $n>1$ et $q>0$, pour tout $t>0$ et quelque soit $\rho_1,\rho_2>0$,*

$$\mathbb{P}\left(\frac{T_n^2(\rho_1,\rho_2)-\Theta_1(\lambda,\rho_1,\rho_2)}{\sqrt{2\Theta_2(\lambda,\rho_1,\rho_2)^2}} \ge \sqrt{2}\left(\sqrt{t}+\frac{\Theta_\infty(\lambda,\rho_1,\rho_2)}{\Theta_2(\lambda,\rho_1,\rho_2)}t\right)\right) \le C\exp(-t).$$

avec $C=3824$. De manière équivalente, nous avons pour la statistique de Hotelling pénalisée, pour $n>1$ et $q>0$, pour tout $t>0$ et quelque soit $\rho>0$,

$$\mathbb{P}\left(\frac{T_n^2(\rho,1)-\Theta_1(\lambda,\rho,1)}{\Theta_2(\lambda,\rho,1)} \ge \sqrt{2t}+\frac{\Theta_\infty(\lambda,\rho,1)}{\Theta_2(\lambda,\rho,1)}t\right) \le C\exp\left(-\frac{t}{2}\right).$$

Les bornes du théorème ci-dessus peuvent être utilisées en pratique pour effectuer des tests, en particulier pour la détection d'anomalies dans le cadre de l'apprentissage statistique. Voir par exemple la littérature sur les systèmes de détection d'intrusion utilisant des cartes de contrôle multivariées basées sur T_n^2 de Hotelling (par exemple Tracy et al. (1992) [30] et d'autres travaux de ces auteurs).

Dans un second temps, nous obtenons des inégalités avec paramètres estimés dans le cas de distributions générales sous quelques hypothèses de régularités standards (voir Ledoit et Wolf (2000) [14]). Considérons Λ la matrice diagonale des valeurs propres de S^2 et O la matrice des vecteurs propres associés. Les valeurs propres sont notées $\mu_1,...,\mu_q$ avec $\mu_1\le\mu_2\le....\le\mu_q$. Nous avons $S^2=O'\Lambda^2 O$. Maintenant, pour $i=1,\dots,n$, nous définissons

$$Y_i = OZ_i \quad \text{avec } Y_i=(Y_{i,1},...,Y_{i,q})'.$$

Afin d'obtenir un estimateur adapté aux matrices de variance-covariance en grande dimension,

Ledoit et Wolf (2000) [14] ont étudié le minimum de

$$\mathbb{E}\left(\left\|\Sigma_n^2(\rho_1,\rho_2)-S^2\right\|^2\right).$$

Cette minimisation peut être considérée comme un problème de projection dans un espace de Hilbert pour les matrices aléatoires, équipé du produit scalaire $\langle A,B\rangle_{\mathcal{H}}=\mathbb{E}\left[\langle A,B\rangle\right]$ avec comme norme associée $\|.\|_{\mathcal{H}}^2=\mathbb{E}\|.\|^2$, où $\langle A,B\rangle=Tr\left(AB'\right)/q$ représente un produit scalaire de Frobenius modifié.

Ledoit et Wolf (2000) [14] ont montré que cette minimisation conduit à considérer des valeurs optimales de ρ_1^* et ρ_2^* qui nous permettent de définir une pénalité optimale

$$\rho^*=\rho_1^*/\rho_2^*,$$

que nous proposons d'estimer par un estimateur de type plug-in $\hat{\rho}^*$.

Considérons les hypothèses suivantes :

(A_1) $\exists K_0,K_1>0$ tel que pour tout n and pour tout $q\geq n$, $K_0\leq\frac{q}{n}\leq K_1$.

(A_2) $\exists K_2>0$ tel que pour tout n et pour tout $q\geq n$, $\frac{1}{q}\sum_{j=1}^{q}\mathbb{E}\left[Y_{1,j}^8\right]\leq K_2$.

(A_3) $\exists K_3>0$ tel que pour tout n et pour tout $q\geq n$, $\frac{1}{K_3}<\mu_1\leq\mu_q<K_3$.

(A_4) $\exists K_4>0$ tel que pour tout n et pour tout $q\geq n$,

$$\nu=\frac{q^2}{n^2}\times\frac{\sum_{(i,j,k,l)\in\mathbf{Q}}\left(Cov\left(Y_{1,i}Y_{1,j},Y_{1,k}Y_{1,l}\right)\right)^2}{Card\left(\mathbf{Q}\right)}\leq\frac{K_4}{n},$$

où $\mathbf{Q}$ désigne l'ensemble de tous les quadruples composés de quatre entiers distincts compris entre 1 et q.

Théorème 3 *Supposons que les hypothèses* (A_1) *à* (A_4) *sont vérifiées, nous avons alors*

pour tout $n > 1$, *pour tout* $q > n$, *pour tout* $t > 2n$ *et pour tout* $\epsilon > 0$ *petit,*

$$\begin{aligned}\mathbb{P}\left(T_n^2(\widehat{\rho}_n^*, 1) \geq t\left(1 + \hat{a}_n^* + 2\epsilon\right)\right) &= \mathbb{P}\left(n\bar{Z}_n'\hat{\Sigma}_n^{*-2}\bar{Z}_n \geq t\left(1 + \hat{a}_n^* + 2\epsilon\right)\right) \\ &\leq \frac{2e^3}{9}\left(\frac{t-n}{2}\right)^{\frac{n}{2}}\frac{e^{-\frac{t-n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} + \frac{C\left(\epsilon\right)}{n\epsilon},\end{aligned}$$

où $\hat{a}_n^* = 1 + \frac{K_3}{\hat{\rho}_n^*}$, *et* $C\left(.\right)$ *est une fonction réelle strictement positive, indépendante de* n, *définie par*

$$\begin{aligned}C\left(\epsilon\right) &= 4K_1\sqrt{K_2}\left(2 + \frac{1}{q} + K_1\right) + 2K_1 G\left(\sqrt{\frac{\epsilon}{2K_1}}\right) \\ &\quad + \frac{4K_1^2\sigma^4}{\epsilon}G\left(\frac{\epsilon}{2\sigma^2 K_1}\right) + \frac{K_3^2}{\epsilon}G\left(\frac{\epsilon}{K_3}\right).\end{aligned}$$

L'expression explicite de G est donnée par une fonction qui contrôle la proximité entre $1/\rho^*$ et $1/\hat{\rho}_n^*$.

Chapitre 4

Ce chapitre 4 met en oeuvre les outils des chapitres 1, 2 et 3, ainsi que leurs extensions aux réseaux de neurones à des données textuelles extraites d'encyclopédie. Ainsi nous avons constitué une base de données et un corpus à partir d'extraction de textes ou d'articles de Wikipédia en deux versions : en anglais standard et en anglais simplifié. Nous décrirons cette étape fondamentale qui s'est avérée très coûteuse en temps.

Plus précisément, nous présentons d'abord brièvement les origines des réseaux de neurones et rappelons les principes des méthodes de Deep Learning (DL) ou apprentissage profond, et les principaux termes utilisés dans ce domaine. Nous décrivons les réseaux de neurones convolutifs et récurrents, ainsi que les LSTM et/ou les réseaux avec des couches encodeur-décodeur qui sont mis en oeuvre sur nos données. Les aspects pratiques (calibration des paramètres du réseau) seront abordés ensuite. Nous résumons enfin les performances d'une série d'architectures d'apprentissage profond, que nous proposons sur notre corpus extrait de Wikipedia. Ceci nous permet de construire une mesure de complexité et un simplificateur de texte automatique (certes beaucoup moins performant que chatGPT mais qui permet de comprendre les mécanismes en oeuvre). Lorsque les données sont contextualisées à un thème donné (par exemple le cinéma ou aux artistes), nous obtenons des taux de performances de nos classifieurs de l'ordre 85%. Par ailleurs, une combinaison de ces différentes architectures nous permet de construire un générateur de texte en version simplifié qui a été testé pour l'instant uniquement sur un petit corpus mais que nous souhaitons mettre en oeuvre sur un corpus beaucoup plus grand (avec les ressources informatiques adéquates). Enfin, les codes python sont fournis en annexe.

References

[1] J. Alegria. Deafness and reading. In *Handbook of children's literacy*, pages 459–489. Springer, 2004.

[2] P. Bertail, E. Gautherat, and H. Harari-Kermadec. Exponential bounds for multivariate self-normalized sums. *Electronic Communications in Probability*, 13:628–640, 2008.

[3] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Sixth Workshop on Very Large Corpora*, 1998.

[4] T. Brants. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[5] J. Chang, C. Y. Tang, and T. T. Wu. A new scope of penalized empirical likelihood with high-dimensional estimating equations. *The Annals of Statistics*, 46(6B):3185–3216, 2018.

[6] L. S. Chen, D. Paul, R. L. Prentice, and P. Wang. A regularized hotelling's t 2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association*, 106(496):1345–1360, 2011.

[7] M. Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics, 2002.

[8] R. Conrad. The deaf schoolchild london harper & row. 1979.

[9] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

[10] M. Harris. Reading comprehension difficulties in deaf children. In *Workshop on Comprehension Disabilities. Centro Diagnóstico Italiano, Milán, Italia*, 1994.

[11] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

[12] L. Kelly. The interaction of syntactic competence and vocabulary during reading by deaf students. *The Journal of Deaf Studies and Deaf Education*, 1(1):75–90, 1996.

[13] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[14] O. Ledoit and M. Wolf. A well conditioned estimator for large dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2000.

[15] O. Ledoit and M. Wolf. Quadratic shrinkage for large covariance matrices. *Bernoulli*, 28(3):1519–1547, 2022.

[16] H. Li, A. Aue, D. Paul, J. Peng, and P. Wang. An adaptable generalization of hotelling's t^2 test in high dimension. *The Annals of Statistics*, 48(3):1815–1847, 2020.

[17] S. Mahapatra and J. Sabat. Comprehension difficulties in reading disabled children. *IOSR Journal of Humanities and Social Science*, 21:16–22, 2016.

[18] M. Marschark and P. E. Spencer. *The Oxford handbook of deaf studies, language, and education, vol. 2*. Oxford University Press, 2010.

[19] T. Otsu. Penalized empirical likelihood estimation of semiparametric models. *Journal of Multivariate Analysis*, 98(10):1923–1954, 2007.

[20] A. Owen et al. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

[21] A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

[22] P. V. Paul and D. W. Jackson. *Toward a psychology of deafness: Theoretical and empirical perspectives*. Allyn & Bacon, 1993.

[23] I. Pinelis et al. Extremal probabilistic problems and hotelling's t2 test under a symmetry condition. *The Annals of Statistics*, 22(1):357–368, 1994.

[24] S. Quigley and P. Paul. Language and deafness. college, 1984.

[25] S. P. Quigley et al. The language structure of deaf children. *Volta Review*, 79(2):73–84, 1977.

[26] S. P. Quigley and P. V. Paul. *Language and deafness*. College Hill Books, 1984.

[27] A. Ratnaparkhi et al. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, PA, 1996.

[28] Z. Shi. Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics*, 195(1):104–119, 2016.

[29] R. Swanwick and L. Watson. Literacy in the homes of young deaf children: Common and distinct features of spoken language and sign bilingual environments. *Journal of Early Childhood Literacy*, 5(1):53–78, 2005.

[30] N. D. Tracy, J. C. Young, and R. L. Mason. Multivariate control charts for individual observations. *Journal of Quality Technology*, 24(2):88–95, 1992.

[31] D. Yogatama. *Sparse models of natural language text*. PhD thesis, Ph. D. thesis, Carnegie Mellon University, 2015.