APPRENTISSAGE DE REPRÉSENTATION: DE LA DÉCISION NON LINÉAIRE À LA GÉNÉRATION DE DONNÉES

Vincent Guigue

SCIENCES SORBONNE UNIVERSITÉ

HABILITATION À DIRIGER LES RECHERCHES DE SORBONNE UNIVERSITÉ

Spécialité

Informatique

Institut de Formation Doctorale de Sorbonne Université

Présentée par

Vincent Guigue

APPRENTISSAGE DE REPRÉSENTATION: DE LA DÉCISION NON LINÉAIRE À LA GÉNÉRATION DE DONNÉES

soutenue le 17 février 2021

devant le jury composé de :

M. Massih-Reza Amini	PR Univiversité Grenoble-Alpes	Rapporteur
M. Patrice Bellot	PR Univversité Aix-Marseille	Rapporteur
Mme. Gabriella Pasi	PR Université de Milan	Rapporteur
M. Mohamed Chetouani	PR Sorbonne Université	Examinateur
M. Martin Trépanier	PR Polytechnique Montréal	Examinateur
M. Emmanuel Viennet	PR Université Sorbonne Paris Nord	Examinateur

> *Le travail d'équipe est essentiel.*
> *En cas d'erreur, ça permet d'accuser quelqu'un d'autre.*
>
> **— Bernard Menez**

Remerciements

Mes remerciements vont d'abord l'ensemble de l'équipe MLIA pour cette ambiance quotidienne à la fois détendue et productive qui permet se lever le matin de bonne humeur, de travailler des bonnes conditions tout en prenant du plaisir à manger ensemble le midi. Si ces quelques lignes semblent d'une banalité plate, je veux insister sur les valeurs humanistes d'empathie, de confiance, de partage et d'entraide que ce soit en recherche, sur les cours ou les financements qui donnent du relief à ces concepts. Un tel environnement est assurément rare et fragile : merci à Patrick d'avoir su le bâtir, merci à Nico, Laure, Sylvain, Ludo, Benjamin, Matthieu, Olivier, Edouard de l'entretenir au quotidien.

Je tiens également à remercier l'ensemble des doctorants et stagiaires avec qui j'ai collaboré : si je présente les travaux suivant à la première personne, c'est bien la première personne du pluriel ! Je suis conscient de l'aspect collaboratif de ces contributions et je le souligne dans chacune de mes présentations. Merci à Abdel, Mickael, Elie, Emeric, Ludmilla, Charles Emmanuel, Perrine, Clara, Bruno et aux autres : vous reconnaitrez vos travaux dans mes phrases. Je suis fier de ces collaborations et je vous remercie pour votre implication.

Je n'oublie pas non plus le soutien logistique de Christophe, Vincent, Nadine, Jaqueline, Ghislaine et tous les autres : n'étant doué ni avec un calendrier, ni avec un ordinateur, je ne suis pas un client facile ! Mais vous avez toujours su trouver des solutions efficaces et je vous en remercie chaleureusement.

Si l'HDR est l'aboutissement d'années de travail, c'est aussi un investissement qui déborde largement sur la vie de famille et souhaite remercier ici Sarah, Luna et Romane pour leur soutien dans la construction de ce projet.

Table des matières

> *Le raisonnement est aussi naturel à l'homme que le vol aux oiseaux.*
>
> **— Quintilien**

1 Introduction

Ce document vise à mettre en perspective nos travaux de recherche, effectués au Laboratoire d'Informatique de Paris 6, sur la dernière décennie. Nous avons opté pour une introduction en deux parties en commençant par un bref historique de l'apprentissage statistique afin d'expliquer le contexte de notre travail. La seconde partie de l'introduction revient sur l'émergence des grandes idées qui ont façonné nos contributions.

1.1 Contexte historique de l'apprentissage statistique

Cinq grandes périodes permettent de structurer les évolutions en apprentissage statistique depuis l'apparition de la discipline dans la seconde moitié du XX^{e} siècle. Notre équipe se positionnant sur la recherche appliquée, il est particulièrement important de rappeler le context général associé pour saisir la motivation de certaines contributions.

Fondations L'intelligence artificielle, comme l'informatique, est généralement considérée comme l'héritage des travaux d'Alan Turing. Le terme est inventé lors de la conférence de Dartmouth en 1956. Du coté de l'apprentissage statistique, qui ne constitue alors qu'une petite sous-communauté de l'IA, les contributions sont principalement théoriques avant 1985. Parmi les propositions célèbres, nous mentionnerons l'algorithme du perceptron de Rosenblatt [Ros58], le travail de Tikhonov sur les formulations régularisées [Tik+77] et les premiers modèles non paramétriques d'estimation de densité [Par61]. Ces travaux ont posé les bases des algorithmes qui sont encore utilisés en traitement de données.

Proof of Concept La décennie (1985-1995) porte la marque des premiers succès industriels. L'algorithme de la rétro-propagation du gradient [Rum+86] et les architectures à convolution [LeC+89b ; LeC+89a] ont permis l'industrialisation des systèmes de reconnaissance de chiffres manuscrits (au niveau de la poste, ou de l'industrie bancaire). Du coté de la transcription et de la traduction automatique, les chaines de Markov cachées permettent de franchir un palier de performances et ouvrent de nouvelles perspectives [RJ86]. De nouvelles formulations convexes [Bos+92 ; CV95] rendent l'accès à l'apprentissage beaucoup plus simple ; cette simplification permet la multiplication des expériences et l'augmentation des performances.

Industrialisation, essor de la communauté Une fois démontré l'intérêt de l'apprentissage automatique, la décennie 1995-2005 correspond à un essor et une structuration, à la fois de la communauté et des outils ; de nombreuses toolboxes voient le jour [Joa98a ; CL01 ; Can+05]. Les propositions de modèles continuent à façonner la communauté scientifique (modèles graphiques [Jor98], boosting [FS+96], arbres de décision & forêts aléatoires [Bre01], ...) ; ces nouveaux modèles ouvrent des perspectives sur la manipulations de données complexes et/ou structurées [Kas+03 ; Des+05], la classification et l'alignement de séquences [Laf+01]. Les premières propositions correspondant à des prédictions structurées datent de cette période [Tso+04].

Les applications se multiplient rapidement dans tous les domaines, de la maintenance prédictive aux filtres spam ; du traitement du signal à la classification de documents sur le web en passant par la médecine et la bioinformatique. Parmi les nouvelles problématiques applicatives, il faut mentionner la recommandation et l'introduction de la NMF (non negative matrix factorization) [LS99] permet de simplifier l'apprentissage de représentations et d'introduire de la personnalisation dans de nombreux algorithmes.

Réseaux sociaux & big data : explosion de l'apprentissage Les années qui suivent voient l'explosion des réseaux sociaux et l'entrée dans l'ère du big-data (2005-2015). Du coté algorithmique, cela se traduit notamment par une remise en cause de l'hypothèse classique i.i.d. (données indépendantes et identiquement distribuées) et de nouvelles problématiques autour des graphes comme la modélisation de la diffusion d'information [Gru+04] ou la découverte des influenceurs dans un réseau [Kem+03].

D'un point de vue applicatif, le web participatif (souvent étiqueté 2.0) permet de récolter des avis en temps réel à travers des blogs, des revues, des tweets, des pages Facebook [PL+08]... Cela ouvre de nouvelles possibilités pour effectuer des sondages, gérer la e-réputation des entreprises, prévenir les buzz négatifs... L'analyse des données textuelles est révolutionnée successivement par l'introduction de nouveaux espaces vectoriels sémantiques [Mik+13] et la modélisation des phrases dans les textes [Cho+14]. Les techniques de *profiling* se généralisent [BK07], la recherche d'information prend en compte des informations contextuelles voire personnelles et les systèmes de recommandation fleurissent des sites de e-commerce jusqu'aux pages d'accueil des journaux. Le domaine de la vision est révolutionné par les nouvelles architectures profondes de réseaux de neurones [Kri+12], qui promettent de remettre en cause à court terme le modèle de l'industrie automobile en basculant vers des véhicules (de plus en plus) autonomes.

Sur le plan industriel, cette décennie correspond à une prise de conscience de la valeur des données manipulées au quotidien. Le stockage et la valorisation sont rendus possibles grace à de nouvelles plateformes de base de données [Cat11] et de traitements distribués comme Hadoop [Whi12] puis Spark [Zah+10]. L'explosion du domaine est ressentie en terme de ressources humaines, les GAFAM (Google, Apple, Facebook, Amazon, Microsoft) investissent massivement et recrutent énormément ; l'audience de la conférence de référence, NIPS, passe de 800 à 4000 personnes.

Re-fondation et convergences Le domaine de l'apprentissage automatique est actuellement en cours de re-fondation autour de l'apprentissage de représentation [Ben+13], avec une attention particulière autour des nouvelles architectures de

réseaux de neurones [LeC+15 ; Sch15]. L'enjeux général est de décrire l'ensemble des éléments du problèmes (signaux, item, individu, ...) sous forme de représentations continues (souvent abstraites) et d'utiliser ces profils dans différentes tâches. De manière générale, multiplier les problématiques autour de concepts invariants revient à placer tous les problèmes rencontrés dans un cadre de raisonnement général [Wes+14 ; Tai+15]. Plusieurs problématiques signifie aussi plusieurs manière d'exploiter un concept ; par extension, la représentation du concept va être plus riche et progressivement, plus universelle [CW08]. Quasiment tous les problèmes d'apprentissage automatique reviennent à chercher la réponse à une question particulière *Y a-t-il un avion sur la photo ?*, *Cette revue est-elle positive ?*, *Ce signal présente-t-il une anomalie ?*... Il devient ainsi possible de répondre à des questions qui n'ont pas été envisagée lors de l'apprentissage des modèles (*0-shot learning*), que ce soit dans des espaces latents textuels [Mik+13], de la reconnaissance des formes [Soc+13b] ou du décryptage des signaux du cerveau [Pal+09].

Plusieurs exemples applicatifs illustrent aussi cette volonté de convergence. Au niveau architectural, les systèmes de traduction ont changé de paradigme récemment [Cho+14] : les représentations apprises à l'aide de réseaux récurrents, sur les phrases du langage d'origine, deviennent universelles ; elles peuvent être *décodées* vers n'importe quelle langue. Ces systèmes s'appuient sur des représentations de mots riches apprises sur des corpus gigantesques à l'aide d'algorithmes plus efficaces [Mik+13]. Dans le domaine de l'image et de la vision, il devient maintenant courant d'utiliser des systèmes appris sur des tâches spécifiques par certains [SZ14], pour traiter des tâches de plus haut niveau par d'autres [Dur+16] ou sur des petits jeux de données qu'il n'aurait pas été possible de traiter efficacement sans transfert de connaissances. Cette exploitation des représentations, non seulement dans plusieurs tâches, mais aussi dans plusieurs applications impliquant différents auteurs est un élément nouveau et structurant pour la communauté de l'apprentissage automatique.

La convergence concerne les implémentations et le matériel utilisé. Quelques bibliothèques de référence *open-source*, notamment SciKit-Learn [Ped+11], TensorFlow [Aba+16] et (py)Torch [Col+11b] (et sa récente version python), servent de base à la plupart des implémentations, l'exécution reposant sur des GPU toujours plus puissants. Cette concentration de plateformes est supportée par les GAFAM, elle aboutit a des cadres logiciels de très grande qualité qui favorisent la productivité, la reproductibilité des expériences, les échanges et, plus généralement, l'accès à des architectures très complexes.

Enfin, la convergence s'illustre au niveau des plateformes de données et des groupes industriels. La communauté informatique partage de longue date des jeux de données permettant une comparaison équitable des modèles [Lic13], cependant un nouveau palier a été franchi récemment avec la multiplication des compétitions ouvertes de grande envergure [Eve+15 ; Rus+15], l'initiative CLEF [Cre+10] ou le site Kaggle [Kag]. Sur le plan industriel, les rachats se sont multipliés (e.g. DeepMind par Google, KXEN par SAP) pour aboutir à la création de quelques laboratoires mondiaux privés dans le domaine de l'apprentissage automatique (DeepMind, Microsoft Research, OpenAI, Google Brain, FAIR).

1.2 Contributions

Nos contributions sont réparties dans les domaines applicatifs du traitement du signal et des séries temporelles, de l'analyse de documents (en particulier la classification de sentiments) et enfin du *profiling* et des applications de recommandation. Le dénominateur commun des contributions suivantes est l'apprentissage de représentation. Dans un premier temps, nous abordons la problématique de la classification robuste en comparant des méthodes linéaires et des approches non-linéaires de type réseaux de neurones reposant sur des espaces de représentation intermédiaires. Nous avons ensuite dressé un parallèle entre représentations d'utilisateurs et représentations de mots. Ainsi, la notion de sémantique est une clé dans plusieurs domaines : le texte, évidemment, mais aussi les profils utilisateurs dans le cadre de la recommandation ou de l'analyse de séries temporelles. Nos dernières contributions se concentrent autour de la fusion de données hétérogènes et du démêlage de facteurs latents pour la construction de représentations interprétables. Ces problématiques innovantes dans le domaine du machine learning nous conduisent à travailler sur des modèles génératifs appris dans des architectures de bout en bout. Nous montrons que la génération de données et notamment de texte est une manière élégante d'intégrer la mouvance xAI qui prône une Intelligence Artificielle plus explicable.

Le parallèle avec la section précédente sur l'historique des approches en IA est intéressant : notre premier chapitre est directement inspiré de la période (1995-2005). Ayant à disposition des approches linéaires et non linéaires, le problème est de construire le classifieur le plus efficace possible en comparant les architectures et en sélectionnant les caractéristiques les plus pertinentes. Si le second chapitre prend sa source dans les années 2000 en croisant sémantique et profiling, il est ancré dans les problématiques réseaux sociaux & big-data correspondant à la période (2005-2015). Nos travaux récents et perspectives de recherche correspondent quant à elles aux problématiques listées dans le paragraphe *r*e-fondation et convergences : nos propositions reposent sur des architectures complexes requérant à la fois des supports logiciels avancés (pyTorch) et des modèles de langues pré-entrainés sur de larges corpus comme ELMO ou BERT. Ces architectures sont, pour partie, entrainées de manière end-to-end et contiennent donc un module génératif.

Robustesse, réduction du bruit et sélection de caractéristiques Suite aux travaux menés en thèse sur la classification de signaux [Gui+05 ; Gui+06], nous avons travaillé sur les interfaces cerveau-machine (BCI –Brain Computer Interfaces–) [RG08]. Cette application requiert un traitement particulièrement robuste, les EEG (électroencéphalogrammes) étant particulièrement bruités. Les EEG ont aussi une composante spatiale : chaque signal est composé de 64 canaux correspondant à autant d'électrodes sur le scalp. De plus, un des enjeux consiste à traiter les signaux de patients qui n'ont pas été vus en apprentissage : il s'agit donc d'effectuer un transfert vers des nouveaux échantillons dans un contexte différent ; une situation très défavorable.

Dans le domaine de la classification de sentiments, l'espace de description (textuel) est encore une fois très bruité (nombre de mots élevé, complexité des phrases, fautes d'orthographe...) et la problématique du transfert très présente. En effet, si les revues du web participatif constituent une ressource étiquetée quasi-infinie, l'enjeu est bien d'utiliser les modèles sur d'autres données (blogs, tweets, ...). Nous avons effectué des tests au niveau des architectures [Raf+11 ; Raf+12d]

en mesurant l'apport des réseaux de neurones sur les architectures linéaires. Nous avons également proposer une nouvelle technique de régularisation pour augmenter la robustesse [Raf+12c ; Raf+12a]. Nous avons étudié l'impact des différents jeux d'apprentissage sur le modèle appris [Raf+13] et les possibilité de passer d'un media à l'autre [GS+13]. Enfin, nous avons étudié l'impact des liens sociaux sur la classification de sentiments [Noz+14]. D'une manière amusante, ces travaux sur des étiquettes subjectives font echo à notre début de doctorat sur la perception des émotions dans les signaux physiologiques [GUI+03 ; Loo+06].

Récemment, nous avons aussi abordé la problématique de la classification de trajectoires à partir de données radar, intrinsèquement fortement bruitées. Nous avons mis au point une stratégie de discrétisation très simple [Del+16] mais particulièrement efficace pour ces signaux.

Si la robustesse au bruit est un facteur important, les applications mentionnées ci-dessus font aussi apparaitre d'autres aspects : le transfert d'un patient à l'autre, d'un domaine thématique à l'autre ; le traitement de séries spatio-temporelles. Nous avons envisagé plusieurs angles d'attaque : celui de la sélection de variables, de la multiplication des classifieurs, de la volumétrie des données d'apprentissage (associée notamment aux architectures de réseaux de neurones) et, enfin, celui des formulations régularisées. Par rapport à la frise chronologique esquissée précédemment, ces travaux s'inscrivent dans la logique d'industrialisation de l'apprentissage automatique. Nous avons abordé plusieurs domaines applicatifs et proposé des modèles ré-utilisables, notamment intégrés à la toolbox [Can+05].

Ces contributions sont détaillées dans le chapitre 2 de ce document.

Apprentissage de profils L'explosion du web puis celle des réseaux sociaux, au milieu des années 2000 ont redéfini le problème de l'accès à l'information pour chacun. Les moteurs de recherche sont de plus en plus affutés sur l'indexation, mais le potentiel d'amélioration se situe plutôt au niveau de la modélisation du contexte : les informations *deviennent* pertinentes (ou non), en fonction du contexte géographique, temporel, situationnel –travail *vs* loisir– mais surtout en fonction de la personne à qui elles sont destinées. L'apprentissage de représentation des usagers d'un service permet d'extraire des profils qui sont ensuite exploités pour filtrer les résultats des moteurs de recherche ou, de manière proactive, pour ré-organiser des pages d'accueil, personnaliser des *newsletters* ou suggérer des films et des produits à acheter.

Nos travaux autour de l'apprentissage de profils concernent d'abord l'intégration de différentes sources hétérogènes pour l'enrichissement des profils issus des données d'interaction. Nous avons travaillé sur l'exploitation des données textuelles pour améliorer les résultats du filtrage collaboratif [Pou+14d ; Pou+14a ; Pou+15b ; Dia+16 ; Dia+17b]. Nous avons démontré dans ces travaux que le texte permet non seulement d'extraire les centres d'intérêt thématiques de l'utilisateur mais nous avons aussi pointé l'importance du style et du choix des mots pour affiner le profil d'une personne. Nous avons également travaillé sur des approches multi-tâches pour mieux structurer l'espace latent en recommandation [SG16].

Nous avons également travaillé sur la modélisation du temps et de l'enchainement des évènements : certaines recommandation sont plus pertinentes à un instant donné, avant ou après un évènement donné. Sur une échelle temporelle réduite, l'ordre des œuvres dans une exposition n'est certainement pas fortuit : il faut re-

commander et expliquer *dans l'ordre*. Sur un intervalle de temps plus vaste, c'est l'utilisateur du système qui évolue, son niveau d'expertise qui s'affine... Et ses goûts qui changent. Nous avons étudié ces aspects dans plusieurs publications [GS+14 ; GS+15 ; GS+16]. Nous avons aussi proposé des travaux mêlant analyse du texte et de la dynamique dans l'espace de représentation pour améliorer la modélisation de l'utilisateur [Dia+17a ; Gab+20].

L'apprentissage de profil a pris récemment beaucoup d'importance dans le monde des transports intelligents, pour mieux cerner les usagers et leur proposer des alternatives personnalisées en cas d'incident sur le réseau etc... Afin de mieux comprendre des usagers dans un contexte de données très bruitées, nous avons proposé des modèles à la fois fins et robustes, capables de saisir les habitudes quotidiennes (trajets domicile-travail) mais aussi les comportements de plus faibles énergies. Ces travaux, décrits dans [Pou+14c ; Pou+16 ; Ton+16 ; Ton+18c], ont ensuite été étendus à l'analyse des anomalies [Ton+17 ; Ton+18a ; Ton+18b]. En effet, en définissant une anomalie comme un écart au comportement habituel, nous voyons le lien direct entre les deux problématiques.

Il est intéressant de constater que les algorithmes de factorisation matricielle peuvent s'appliquer indifféremment sur des personnes ou sur d'autres concepts comme les mots. Les premiers algorithmes décrivent des profils tandis que les seconds sont motivés par l'introduction d'une sémantique, c'est à dire d'une meilleure compréhension des concepts. De ce point de vue, l'essor de l'apprentissage de représentation en deep learning correspond à une volonté d'analyse très fine des concepts manipulés. La question centrale du chapitre 3 est donc celle de la sémantique (littéralement, de l'ajout de sens) ; l'étape initiale est de dépasser la métrique binaire entre deux objets distincts –je suis identique à moi-même & arbitrairement loin du reste du monde– pour aboutir à une quantification pertinente de la ressemblance –deux mots très différents, deux mots appartenant au même champ lexical, deux mots synonymes– voire une qualification de cette ressemblance –ces deux utilisateurs partagent une affinité pour les films de science-fiction–. La transition vers le chapitre suivant est immédiate : il s'agira d'approfondir la compréhension pour tenter de bâtir des raisonnements et des explications autour des concepts manipulés.

De la compréhension à la génération de données En imaginant que nous disposons de profils pertinents pour les utilisateurs et les concepts associés à une application (articles d'un journal, objets sur un site marchand, films chez un diffuseur, ...), il serait dommage de se limiter à estimer des affinités entre *usagers* et *items*. L'enjeu consiste donc à dépasser la fonction de régression exploitant les profils pour prédire un score de matching afin de répondre à un ensemble de questions, éventuellement en langage naturel et d'expliquer la décision, de la motiver en se fondant sur des aspects plus fin qu'une simple note.

D'un point de vue plus général, il s'agit d'au moins quatre problématiques distinctes :

1. l'affinage et la compréhension de l'espace de représentation des données. Nous avons abordé ce problème sous deux angles différents : d'abord via l'introduction de données textuelles dans des algorithmes qui n'en exploitaient habituellement pas –e.g. le filtrage collaboratif pour la recommandation– ensuite en exploitant le paradigme du démêlage – *disentanglement*– qui vise à rendre chaque dimension de l'espace de représentation interprétable. Nous passons alors d'un espace latent à un espace explicite.

2. Motiver une décision, l'expliquer revient dans un premier temps à analyser la pondération des facteurs dans un espace explicite. Etre capable de re-générer la donnée d'entrée à partir de sa représentation permet de valider le fait que toutes les informations correspondant au signal d'entrée sont bien encodées dans l'espace de représentation. Cette génération est également un moyen d'affiner la compréhension des facteurs importants dans l'espace de représentation. Ainsi, les deux premières problématiques exposées ici sont généralement abordées simultanément comme dans les architectures GAN.
3. Pour aller plus loin, notre idée consiste à générer une explication associée à la décision. Cette explication est, jusqu'ici, de nature textuelle. L'enjeu est d'exploiter les modèles de langues les plus récents pour dépasser les approches par extraction de phrases. L'explication doit à la fois être une synthèse d'un concept –à la manière du résumé automatique– et correspondre à une personne cible –à la manière du filtrage collaboratif–.
4. La dernière étape consiste à introduire du raisonnement causal dans les explications. Cet aspect rappelle les fondements de l'intelligence artificielle mais ce retour aux sources est particulièrement ambitieux, les systèmes logiques fondés sur la causalité, ayant été largement dépassés dans l'aide à la décision par les systèmes fondés sur les corrélations.

Du coté des systèmes de recommandation, nous avons d'abord envisagé des extensions légères pour accompagner les recommandations de textes [Pou+14a ; Dia+16]. Nous avons ensuite envisagé des architectures où l'ensemble du filtrage collaboratif est effectué dans un espace textuel intrinsèquement explicatif [Dia+16 ; Dia+17b]. Cette stratégie apporte une solution élégante et novatrice au problème du démarrage à froid. L'attention sur une séquence de texte constitue aussi une approche intéressante pour expliquer les éléments sur lesquels s'appuie une décision [Dia+18b ; Dia+18a]. Nos contributions les plus récentes tentent de générer un texte personnalisé dépendant non seulement d'un item et d'une opinion, mais aussi d'une personne [Dia+19a]. Le domaine applicatif du matching CV/offre d'emploi est très proche de celui de la recommandation, il s'agit toujours d'évaluer une affinité entre deux objets hétérogènes. La différence réside dans la nature *u*nique d'une offre d'emploi : il est impossible d'apprendre un profil en misant sur la répétition des apparitions d'un item unique ; de même, il est impossible d'évaluer notre système en exploitant un historique de ceux qui ont *a*imé ou pas une offre. Les systèmes génératifs offrent une issue à ce problème qui peut alors être reformulé de la manière suivante : étant donné un historique de CV, suis-je capable de générer la description du prochain poste occupé par une personne en particulier ? [Dia+17a ; Gab+20].

L'extraction et l'exploitation de connaissances requièrent une forme avancée de compréhension des textes reposant conjointement sur des modèles de langues –génératifs– appris sur de très grands corpus et des bases étiquetées permettant d'affiner l'apprentissage. La tâche elle-même constitue une forme d'aboutissement par rapport à la compréhension d'un message : il s'agit d'encoder un texte brut en triplets exploitables par des algorithmes de raisonnement logique. Nous nous sommes intéressés à la catégorisation des liens entre entité d'une part [Sim+19b ; Sim+19a] et à l'extraction conjointe des entités et des relations [Tai+19a ; Tai+19b ; Tai+20a ; Tai+20b].

Nos contributions sur la compréhension et le raisonnement, ne se limitent pas aux données textuelles. Dans le domaine du traitement du signal, nous avons proposé un algorithme de *brain-reading* basé sur du *0-shot learning* [Pip+14 ; Pip+15]. Dans

ce cadre, nous devons retrouver des classes non vues lors de l'apprentissage en nous basant sur une sémantique de concepts exogène. Les systèmes de transports intelligents requièrent également une compréhension fine des contextes ayant un impact sur les affluences et nous avons montré que l'apprentissage de représentations associées à différents éléments de contexte permet d'envisager de nouvelles approches de prédiction fortement inspirées des algorithmes génératifs [Gui+19 ; CD+20b ; CD+20a].

Projet de recherche L'ensemble de ces contributions nous a menés à la construction d'un projet de recherche autour des techniques d'apprentissage de représentation pour l'extraction et la gestion des connaissances.

L'extraction de connaissances est un domaine historique de l'IA qui a profondément évolué ces dernières années [Tai+20b]. A l'inverse, les stratégies d'inférence de nouvelles connaissances sont récentes dans la communauté de l'apprentissage de représentation. Les premières propositions de systèmes capables de raisonner de cette manière datent du début de la décennie 2010 [Bor+11 ; Soc+13a]. L'intégration de la dynamique pour le suivi de dialogue a ensuite conduit aux architectures remarquables décrites dans [Wes+14 ; Suk+15]. L'avènement des architectures génératives, la multiplication des problèmes basés sur des sources hétérogènes et la prise en compte de supervisions de plus en plus distantes nous amènent à la problématique récente du *data-to-text* [Yan+17 ; Reb+19]. Nous sommes convaincus que ces évolutions doivent être mises en perspective des stratégies de transfert qui permettent des gains en performances significatifs en ré-utilisant des modèles pré-appris sur de larges corpus textuels [Pet+18 ; Dev+18] ou image [Kri+12].

Nous proposerons ainsi des pistes pour redéfinir la représentation des connaissances à la fin de ce document.

1.3 Collaborations

Les contributions listées précédemment et détaillées dans la suite de ce document sont le fruit de nombreuses collaborations, notamment avec les stagiaires de M2 et les doctorants que nous avons encadrés, seul ou avec des collègues, ces dernières années.

Stages de M2

2009 **Anastasio Bellas**. Modèles francophones en classification d'opinion. Il a continué en thèse à l'Université de Grenoble.

2011 **François Rousseau**. Apprentissage de représentations de mots pour la classification de sentiments. Il a continué en thèse à Polytechnique.

2012 **Elie Guardia-Sebaoun**. Diffusion des opinions dans les réseaux sociaux. Il a continué ses recherches en thèse sous ma direction.

2014 **Debora Nozza**. A latent representation model for sentiment analysis in heterogeneous social networks.

2014 **Ludmilla Tajtelbom**. Classification de signaux EEG en transfert inter-patient par exploitation de la géométrie Riemanienne. Elle a continué en thèse sous la direction de T. Artières et la mienne.

2016 **Damien Siléo**. Système de recommandation par filtrage collaboratif, interaction de facteurs latents & régularisation. Il a continué en thèse à Toulouse (IRIT).

2017 **Matthieu Crilout**. Système de recommandation et e-learning. Il a été embauché en ingénieur de recherche sur les systèmes d'e-learning.

2017 **Cynthia Delauney**. Analyse des traces radars d'avions au sol pour la détection d'anomalies. Elle a ensuite entamé un carrière dans le privé.

2019 **Valentin Guiguet**. Prise en compte du contexte dans les architectures de deep learning pour la prédiction de séries temporelles. Il a continué en thèse en se ré-orientant vers le NLP dans notre équipe.

Thèses

2010-13 **Abdelhalim Rafrafi**. *Classification de sentiments sur le Web 2.0*. Co-direction avec **P. Gallinari**.

2011-15 **Mickael Poussevin**. *Representation Learning of User-generated Data*. Analyses de traces utilisateurs, transports intelligents et systèmes de recommandation. Co-direction avec **P. Gallinari**.

2012-16 **Elie Guardia-Sebaoun**.*Accès Personnalisé à l'Information : Prise en Compte de la Dynamique Utilisateur*. Analyse de sentiments sur les réseaux sociaux et systèmes de recommandation.

2014-18 **Ludmilla Tajtelbom**. Brain-reading et analyse des signaux du cerveau. Co-direction avec **T. Artières**, thèse abandonnée au bout de 4 ans.

2015-19 **Emeric Tonnelier**. *Apprentissage de représentations pour les traces de mobilité*. Analyses de traces utilisateurs et transports intelligents. Co-direction avec **N. Baskiotis**.

2016-19 **Charles-Emmanuel Dias**. *Expliciter le filtrage collaboratif par le traitement automatique des langues*. Systèmes de recommandation et données textuelles.

2017- **Perrine Cribier-Delande**. Modélisation des utilisateurs et du contexte dans l'analyse et la prédiction de séries temporelles. Co-direction avec **L. Denoyer**, thèse CIFRE (Renault).

2018- **Clara Gainon de Forsan de Gabriac**. Modélisation d'utilisateurs et d'entreprises à partir de bases de CV.

2018- **Bruno Taillé**. Systèmes robustes pour l'extraction de connaissances. Co-direction avec **P. Gallinari**, thèse CIFRE (BNP-Paribas).

2020- **Darius Afchar**. Attribution des caractéristiques pour des systèmes de recommandation plus transparents, thèse CIFRE (Deezer).

2020- **Maya Sahraoui**. Extraction d'entités multi-modales. Analyse des corpus images et textuels du Museum National d'Histoire Naturelle et supervision distante par des bases de connaissances. Co-direction avec **R. Vignes Lebbe**.

2021- **Tristan Luigi**. Extraction d'informations et modèles génératifs de textes. Exploration du domaine naissant du *data-to-text*. Co-direction avec **L. Soulier**, thèse CIFRE (Upskills).

2021- **Etienne Le Naour**. Analyse contextuelle de séries temporelles et apprentissage de profils utilisateur. Co-direction avec **N. Baskiotis**, thèse CIFRE (EDF).

1.4 Organisation

Le plan de cette habilitation à diriger les recherches découle directement de l'organisation des contributions proposée ci-dessus. Le chapitre 2 présente nos contributions pour l'apprentissage de classifieurs robustes en jouant sur les architectures

–linéaires ou non–, la sélection de caractéristiques et la volumétrie des données. Ce chapitre permet aussi de clarifier les principales notations et de rappeler le formalisme de l'apprentissage statistique.

Dans un deuxième temps, le chapitre 3 détaille les différentes techniques d'apprentissage de profil que nous avons développées et les applications associées. Le formalisme des problèmes d'apprentissage de représentations vise à étendre celui introduit dans le chapitre précédent en soulignant les opportunités sur la gestion de données hétérogènes et les bénéfices des architectures multi-tâches.

Enfin, nous présentons nos travaux les plus récents autour de l'explicitation des représentations et de l'exploitation avancée de ces espaces, notamment à travers les approches génératives dans le chapitre 5. Ce chapitre aboutit logiquement sur la conclusion de ce manuscrit qui intègre notamment un projet de recherche pour les années à venir autour des nouvelles approches pour la gestion des connaissances.

> *Le fait d'être chauve et jeune*
> *n'est pas un élément de classification politique.*
>
> **— Laurent Fabius**

2

Classification robuste : du transfert à la sémantique avec des approches simples

Nous abordons dans ce chapitre les formulations usuelles de classification supervisée. A partir de données associées des étiquettes, nous apprenons une fonction capable de prédire cet étiquetage ; dans un second temps, nos fonctions sont utilisées en inférence pour traiter de nouvelles données non supervisées. Théoriquement, ces nouvelles données doivent suivre la même distribution que les données d'apprentissage. Dans la pratique, le contexte thématique, temporel ou humain peut changer et nous n'avons que peu de garantie sur la distribution des données de test. La robustesse est donc une notion clé et nous montrons que plusieurs options sont envisageables en jouant sur la sélection de caractéristiques, la formulation des problèmes d'apprentissage ou simplement la taille des jeux de données.

Application par application, l'enjeu principal consiste à analyser les spécificités d'un problème, anticiper ou simuler les difficultés de passage vers les données de test puis à trouver des solutions pour y faire face efficacement. Les approches se classent en trois grandes familles :

— la construction de **descripteurs avancés** dédiés à une tâche. Ces descripteurs peuvent être calculés suite à une discussion avec un expert du domaine ou appris sur un jeu de données étiquetées –toujours par des experts du domaine–. Bien que plébiscité dans le milieu industriel, ce type d'approche intéresse moins la communauté scientifique pour des raisons évidentes de limitation d'accès aux spécialistes.

— La construction de **classifieurs avancés** où l'idée est d'attaquer la difficulté de la tâche via des architectures avancées. Modéliser des interactions complexes entre les descripteurs de base peut par exemple se faire en exploitant des réseaux bayesiens [Jor98] ou des architectures multi-couches dans les réseaux de neurones [Bis95].

— La dernière approche, détaillée dans ce chapitre, consiste à exploiter des **approches linéaires naïves** mais très robustes pour traiter plus de descripteurs et/ou plus de données. Nous montrerons que ces approches fonctionnent très bien dans un certain nombre de domaines applicatifs.

Le fil directeur de ce manuscrit est la sémantique des données : la définition de la sémantique est plus complexe qu'il n'y parait au premier abord mais elle est

centrale dans la plupart des domaines applicatifs. Les stratégies d'extraction de cette sémantique sont multiples. La ré-émergence récente de la problématique xAI [Goe+18] –qui insiste sur le besoin de compréhension de la décision en opposition à certaines approches de deep-learning de type boite noire– est aussi une manière de recentrer les débats autour de la sémantique.

Nous nous intéressons dans ce chapitre aux stratégies les plus simples, les modèles linéaires, afin de montrer qu'ils sont à la fois efficaces et porteurs de sens dans de nombreux cas de figure.

Organisation & contributions Après avoir décrit les notations et formulations utilisées, section 2.1, nous détaillons deux cas d'usage illustrant des problématiques spécifiques. La classification de signaux du cerveau est abordée en section 2.2, du point de vue de la robustesse et du transfert entre patient pour les applications d'interfaces cerveau-machine –*Brain Computer Interface*, BCI– [Rak+05 ; RG08]. Nous aborderons ensuite les données textuelles et nous nous focaliserons sur la classification de sentiments et la fouille d'opinion en section 2.3 [Raf+11 ; Raf+12d ; Raf+12b ; Raf+12c ; Raf+12a ; Raf+13 ; GS+13]. La difficulté réside alors dans le transfert entre domaines : les données étiquetées étant chères, il est intéressant d'utiliser des données de revues utilisateurs, disponibles en masse, pour construire des outils de sondage automatique dédiés à d'autres thématiques ou d'autres média tel que Twitter. Nous évoquerons en conclusion des travaux menés dans d'autres domaines comme l'analyse de trajectoires [Del+16].

2.1 Enjeux, notations et formulation générale

Le cadre le plus classique de l'apprentissage est la classification supervisée. Nous considérons un ensemble de N données brutes $X^b = \{\mathbf{x}_0^b, \ldots, \mathbf{x}_N^b\}$ décrites dans un espace d'origine quelconque $\mathbf{x}^b \in \mathcal{X}$. Classiquement, ces données sont considérées comme des variables i.i.d, c'est à dire indépendantes et identiquement distribuées ; elles sont donc traitables une par une à l'aide d'un modèle unique adapté à la distribution sous-jacente des données. Dans le cadre supervisé, chaque donnée $\mathbf{x}^b$ est associée à une étiquette $y \in \mathcal{Y}$. Dans le cas de la classification, $\mathcal{Y} = \{C_1, \ldots, C_{ncl}\}$ est un ensemble fini (généralement petit, souvent binaire) ; dans le cas de la régression $\mathcal{Y} = \mathbb{R}$. Cette présentation, très classique, est directement inspirée des références du domaine [Dud+73 ; Fri+01].

Représentation La première étape consiste à décrire les données, les segmenter, éliminer une partie du bruit à l'aide d'informations issues du domaine applicatif. En bref, il s'agit d'appliquer une fonction de représentation f_{repr} pour passer des données brutes à des données formatées dans un espace vectoriel de dimension d :

$$\begin{array}{rccc} & \mathcal{X} & & \mathbb{R}^d \\ f_{repr} : & \mathbf{x}^b & \rightarrow & \mathbf{x} \end{array} \tag{2.1}$$

Nous distinguerons trois approches pour la construction de fonction de représentation : le calcul de caractéristiques métier (*feature engineering*), l'extraction de nombreux descripteurs (à sélectionner ultérieurement) –*shapelets*, N-grammes de mots, motifs complexes, bancs de filtres, ...–, et enfin, l'apprentissage de descripteurs optimisés pour une ou plusieurs tâches.

Apprentissage et évaluation L'étape suivante consiste à faire le lien entre la représentation des données $\mathbf{x} \in \mathbb{R}^d$ et les étiquettes $y \in \mathcal{Y}$. Nous cherchons à apprendre la fonction f :

$$\begin{array}{rccc} & \mathbb{R}^d & & \mathcal{Y} \\ f: & \mathbf{x} & \rightarrow & y \end{array} \tag{2.2}$$

telle que $f(\mathbf{x})$ soit une bonne approximation de y. Commençons par les fonctions les plus simples, modèles linéaires : $f(\mathbf{x}) = \mathbf{x} \cdot \mathbf{w}$ de paramètres $\mathbf{w} \in \mathbb{R}^d$.

L'évaluation de ces systèmes est critique, l'enjeu n'étant pas d'optimiser les performances d'estimation des y vus en apprentissage (minimisation de l'erreur empirique), mais de viser une performance optimale sur les données de test, non vues lors de l'apprentissage. Les données sont donc divisées en ensembles d'apprentissage, de validation et de test $X = \{X_{train}, X_{val}, X_{test}\}$ de tailles respectives $N_{train}, N_{val}, N_{test}$. Dans le cadre supervisé, ces notations sont étendues aux étiquettes Y. L'enjeu est donc d'optimiser l'erreur en généralisation. Les formulations régularisées permettent d'aller vers cet objectif [Tik+77] en pénalisant à la fois l'erreur empirique mais aussi la complexité de f afin d'empêcher la fonction d'apprendre par cœur les données d'apprentissage :

$$f^\star = \arg\min_f \underbrace{\underbrace{\sum_{i=1}^{N_{train}} \Delta(f(\mathbf{x}_i), y_i)}_{\text{Erreur empirique}} + \quad \lambda \quad \underbrace{\Omega(f)}_{\substack{\text{Pénalisation des classifieurs complexes} \\ \Rightarrow \text{ limiter le sur-apprentissage}}}}_{\text{Estimation de l'erreur en généralisation}} \tag{2.3}$$

où Δ est une fonction de coût mesurant les écarts entre la vérité terrain y_i et la prédiction $f(\mathbf{x}_i)$ (typiquement l'erreur au carré), Ω est la fonction de régularisation (par exemple $\|\mathbf{w}\|^2$) et λ est le compromis entre ces deux objectifs généralement antagonistes. Les données d'apprentissage permettent d'entrainer le classifieur, i.e. trouver les $\mathbf{w}^\star$ optimaux dans le cas linéaire. Les données de validation sont utiles pour le réglage des hyper-paramètres (λ dans la formulation ci-dessus). Les données de test fournissent une évaluation non biaisée des performances du système.

Il est courant d'optimiser les hyper-paramètres en validation croisée, c'est à dire sur l'ensemble des données disponibles : le protocole est plus simple et surtout mieux adapté aux petits jeux de données, mais les performances du système sont alors évidemment légèrement sur-estimées.

Techniques d'optimisation L'optimisation des hyper-paramètres est réalisé en *grid-search*, c'est à dire en testant un ensemble fini de valeurs. L'optimisation des paramètres peut être réalisée par une résolution analytique de (2.3) lorsque cela est possible (moindres carrés régularisés L_2 par exemple) ou par descente de gradient. Nous passerons rapidement sur cette question qui n'est pas centrale dans le travail exposé ici.

Robustesse et transfert : des enjeux majeurs La problématique générale de la robustesse consiste à faire face efficacement à des données $\mathbf{x}^b$ avec un fort niveau de bruit ou à des entorses à l'hypothèse i.i.d.

Dans notre première application –classification de signaux du cerveau–, le bruit est particulièrement présent dans les données EEG (électro-encéphalogramme) à cause de la captation des signaux à l'extérieur de la boite crânienne ; ces signaux sont

par exemple indéchiffrables à l'œil nu. Dans ce cas, il est recommandé de filtrer les hautes fréquences et en particulier le 50Hz (en Europe) ou le 60Hz (en Amérique du Nord). Les signaux ne sont pas non plus complètement indépendants, l'état d'esprit (ou d'énervement) des personnes créant des corrélations sur une session d'acquisition. Mais la principale difficulté réside dans le transfert d'une personne à l'autre : les données ne sont plus identiquement distribuées et les pertes de performances en test sur une personne inconnue sont très importantes.

Nous abordons ce problème à l'aide de la technique du bootstrap (multiplication des classifieurs sur des sous-ensembles de données) et de la sélection de variables dans le but d'obtenir des performances élevées et stables. Nous mélangeons donc déjà la phase de description des signaux et la phase d'apprentissage, la sélection des variables explicatives revenant de fait à pénaliser en amont la complexité $\Omega(f)$ du classifieur.

Dans la seconde application, il s'agit de classer des revues textuelles, écrites par différentes personnes en fonction de leur polarité. Le bruit est plus léger (fautes d'orthographes, formulations ambiguës) mais la difficulté non moindre étant donné la taille de l'espace des mots (voire des N-grammes –groupes de mots–) et la diversité des thématiques abordées. La grande dimensionnalité de l'espace correspond à des variables très redondantes ; ces variables apparaissent en majorité dans très peu de documents entrainant un fort risque de sur-apprentissage. Le fait d'avoir des documents appartenant à plusieurs thématiques est une entorse à l'hypothèse i.i.d. d'autant plus dommageable que certains mots changent de polarité en fonction du contexte (*léger* est positif pour un ordinateur, ambivalent pour un livre et négatif pour tous les ustensiles demandant de la robustesse ou pour l'autonomie des téléphones portables).

Nous cherchons simplement à construire le modèle le plus général possible, c'est à dire le plus robuste aux changements. Nous nous appuyons sur différentes techniques de régularisation et sur la masse de données étiquetées disponibles (l'ensemble des commentaires récupérables sur Internet !).

2.2 Classification de signaux multi-canaux & transfert entre patients

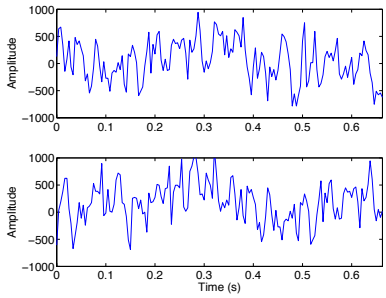


Figure 2.1: Exemples de signaux à classer : les deux classes sont presque impossible à distinguer à l'oeil nu. Exemple positif en haut, négatif en bas.

Dans la continuité de nos travaux de doctorat, nous avons participé à une compétition de classification de signaux EEG (électro-encéphalogrammes) P300 et nous l'avons remporté. Les détails de notre stratégie sont donnés dans [RG08]. Les enjeux du concours concernent principalement la robustesse au bruit et le transfert d'un patient à l'autre. Mais nous montrons également qu'une stratégie linéaire permet une analyse qualitative très intéressante sur des données complexes telles que les séries temporelles multi-variées.

2.2.1 Cadre expérimental classique

mise en forme des données A l'état brut, un EEG est une série spatio-temporelle :

$$\mathbf{x}^b \in \mathbb{R}^{C\times T}, C = 64 \text{ capteurs}, T = 158400 \ (660ms \text{ à } 240Hz) \tag{2.4}$$

Les 64 capteurs étant répartis sur le scalp (Figure 2.2, droite). Après filtrage passe-bande $1-10Hz$ et décimation, il reste 14 points de mesure par capteur. Les différents

canaux sont concaténés de manière à obtenir une représentation vectorielle compacte :

$$f_{repr}(\mathbf{x}^b) = \mathbf{x} \in \mathbb{R}^{896}, \qquad 896 = 64 \times 14 \tag{2.5}$$

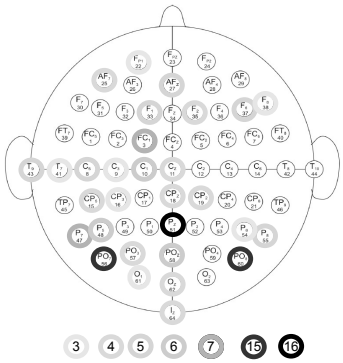
Chaque EEG est une réponse à un stimulus visuel, l'enjeu de la classification est de différencier les stimuli *intéressants* pour l'utilisateur des autres. Dans le système P300-*speller* (Interface cerveau machine de dictée) [1], ce sont des lettres qui sont illuminées (par ligne/colonne dans une matrice 6×6) et notre classifieur doit déterminer si cette illumination correspond à la lettre souhaitée ou pas (Figure 2.2, gauche). Chaque signal $\mathbf{x}$ est donc associé à une étiquette binaire $y \in \{-1, 1\}$ mais aussi à une personne $p \in \{p_1, p_2\}$. Etant donné la structure de la matrice de lettre, il y a 1 signal pertinent pour 5 non pertinents. Le problème est très difficile et la détection d'une seule lettre repose sur la multiplication des expériences : chaque ligne et chaque colonne sont illuminées 15 fois.

Figure 2.2: Interface cerveau-machine de dictée de lettres. EEG 64 canaux, problème de classification binaire.

Apprentissage robuste Les signaux sont très bruités et très variables d'une personne à l'autre. Pour faire face à ce phénomène, nous avons mis en place une stratégie inspirée du *bootstrap* en multipliant les classifieurs sur des sous-ensembles de données homogènes (17 partitions de 900 signaux par utilisateur). Nous avons aussi mis en place une stratégie de sélection de capteurs pour chaque classifieur. Il s'agit donc d'éliminer les blocs de variables non pertinentes. L'algorithme proposé est un algorithme itératif glouton d'élimination, basé sur un critère robuste aux données mal équilibrées :

$$C_{cs} = \frac{tp}{tp + fp + fn} \tag{2.6}$$

où tp, fp, fn désignent respectivement les taux de vrais-positifs, faux-positifs et faux-négatifs [2]. Le nombre de canaux à retenir varie considérablement d'un classifieur à l'autre (entre 4 et 40 sur 64) mais les régions correspondantes demeurent stables, comme le montre la Figure 2.3.



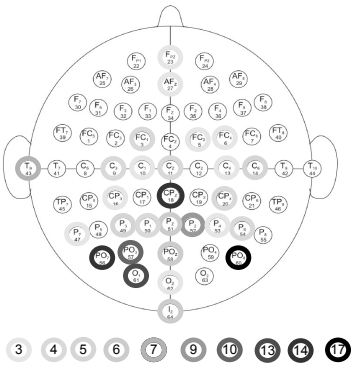


Figure 2.3: Répartition des capteurs sélectionnés par les 17 classifieurs associés aux patients p_1 et p_2. L'échelle de gris indique le nombre de sélection.

Les 34 classifieurs sont des SVM linéaires f_k, appris indépendamment après la phase de sélection des capteurs. En notant ϕ_k la fonction de sélection des variables, de $\mathbb{R}^d$ dans $\mathbb{R}^d$ mettant à 0 les dimensions non pertinentes, nous obtenons $k = 34$ problèmes d'optimisation définis sur 34 sous-ensembles de données distincts :

$$\phi_k(\mathbf{x}) \in \mathbb{R}^d,\ \phi_k(x_j) = \begin{cases} 0 & \text{si } j \text{ n'est pas pertinente} \\ x_j & \text{sinon} \end{cases},\ \phi_k^\star = \arg\min_{\phi_k} C_{cs}(\phi_k) \tag{2.7}$$

$$\mathbf{w}_k^\star, b_k^\star = \arg\min_{\mathbf{w}_k, b_k} \sum_i [1 - \phi_k(\mathbf{x}_i) \cdot \mathbf{w}_k y_i - b_k]_+ + \lambda \|\mathbf{w}_k\|^2 \tag{2.8}$$

Le compromis de régularisation λ est optimisé par validation croisée. $[]_+$ désigne la partie positive d'un réel ; cet opérateur est utile pour écrire simplement la fonction coût charnière des SVM. Les décisions sont ensuite agrégées sur les lignes (respectivement les colonnes) durant la séquence d'illuminations [3]. Chaque signal est toujours associé à une ligne ou une colonne, que nous notons en exposant j : $\mathbf{x}_i^j$. L'agrégation

1. Le nom du système vient d'un pic qui apparait dans l'EEG 300 ms après un stimulus pertinent pour l'utilisateur. Ce phénomène s'observe sur des signaux agrégés mais est indétectable à l'œil nu sur des données brutes, c'est à dire pour une réaction correspondant à un stimulus unique.

2. Le calcul du critère requiert donc l'apprentissage d'un classifieur sur chaque sous-ensemble de variables. La combinatoire pose évidemment des problèmes de passage à l'échelle.

3. Une séquence = 15 illuminations de chaque ligne et chaque colonne dans un ordre aléatoire.

des scores sur les $L = 36$ lignes/colonnes est effectuée par sommation à la fois sur les différents classifieurs et sur les redondances des signaux.

$$score(\ell) = \sum_k \sum_{i \ t.q. \ \mathbf{x}_i^{j=\ell}} \phi_k(\mathbf{x}_i^j) \cdot \mathbf{w}_k + b_k \tag{2.9}$$

Pour obtenir les meilleures performances, il est nécessaire, lorsque c'est possible, d'isoler les classifieurs correspondant à la personne p source du signal à classer. Nous nous restreignons alors aux indices $k \in p$ dans la première somme. Dans le cadre du challenge, les signaux de tests appartenaient aux deux patients

Performances & analyses Le Tableau 2.1 montre les performances des différents systèmes sur les données du challenge Berlin BCI III (dataset II). Au niveau des lettres, après agrégation des décisions sur les lignes et les colonnes pour une série d'illumination (15 au maximum), nous obtenons 96.5% de reconnaissance. Parmi les différentes propositions, nos ensembles de classifieurs fonctionnent très bien et permettent d'acquérir une robustesse certaine : non seulement la performance sur 15 séquences est la meilleure, mais même en se réduisant à l'analyse de 5 séquences d'illuminations les performances sont moins impactées que la concurrence. La manière de sélectionner les capteurs pertinents est très sensible : passer d'une sélection adaptée à chaque classifieur à une sélection globale (bien moins chère à calculer) provoque une chute dans la reconnaissance.

Algo.	Sel. Var.	Nb de séquences 5	15
Contribution	Adapt.	73.5	96.5
Multi-SVM	-	74.5	95.5
Multi-SVM	8 capt.	40.0	80.0
Mono-SVM	8 capt.	31.0	70.0
$2^{ème}$ pos.		55.0	90.5
$3^{ème}$ pos.		59.5	90

Table 2.1: Performances obtenues sur les données de test de la compétition Berlin BCI III (dataset II). Taux de reconnaissance sur les lettres (36 classes)

Il est intéressant de noter que la combinaison linéaire de classifieurs linéaires eq. (2.9) est un nouveau classifieur linéaire. Le gain de performances vient donc bien d'une amélioration de la robustesse et non d'un passage à des méthodes non-linéaires.

Sur le challenge considéré, les données de test appartiennent aux patients vus en apprentissage. Parmi les différents paramétrages de *bootstrap* considérés, le meilleur consiste à n'utiliser que les classifieurs appris sur les données d'une personne p pour classer les signaux de test de p. Cela montre que les signaux sont très variables d'une personne à l'autre et que le transfert d'un modèle appris sur un patient vers un autre patient requiert encore nettement plus de robustesse que la solution exposée ci-dessus.

2.2.2 Evolution du domaine et problématique du transfert

Ce challenge nous a permis d'aborder la problématique de la robustesse et de la sémantique dans un contexte de données très fortement bruitées. Nous avons montré l'intérêt des méthodes ensemblistes inspirées du *bootstrap* et de la sélection de variables pour améliorer à la fois les performances et l'interprétation des résultats. Ces conclusions ne sont pas très originales et viennent simplement conforter les avantages connus de longue date des méthodes ensemblistes [ET94] et de la réduction de la dimensionnalité [Tib96 ; GE03].

Le réel enjeu du domaine se situe plus loin, au niveau du transfert de modèle entre patients. En effet, les signaux sont très variables d'une personne à l'autre. Les compétitions DecMeg 2014 (magneto-encéphalogrammes), BCI NER 2015 (EEG) proposent un tel challenge et Alexandre Barachant a démontré l'intérêt de la géométrie Riemannienne pour y faire face [Bar+13a ; Con+17]. L'un des points forts de la méthode proposée consiste à résoudre un problème d'adaptation durant la phase

d'inférence : le coût de cette étape est largement compensé par les performances de tout premier plan obtenues sur cette tâche difficile.

Le domaine du traitement des signaux du cerveau n'est pas encore passé au *deep learning* (DL) : la plupart des descripteurs sont extraits manuellement à l'aide de différents filtres passe-bande et de combinaison de capteurs (agrégation spatiale) [Dor+06]. Les principales tentatives d'optimisation des caractéristiques restent actuellement cantonnées aux outils traditionnels du traitement du signal [Bla+08]. L'apprentissage de filtres spatio-temporels dans des architectures de réseaux de neurones à convolution (CNN) pour l'analyse d'EEG a donné quelques résultats préliminaires sur la détection de l'épilepsie [Mir+08]. Cependant ces analyses impliquent une temporalité beaucoup plus longue que celle des interfaces cerveau-machine. Les tentatives d'utilisation des CNN en BCI ne permettent pas encore d'égaler l'état de l'art [CG11]. Deux facteurs pouvaient expliquer ce blocage : le fort niveau de bruit combiné avec la taille limitée des jeux de données. Les données publiquement disponibles sont beaucoup plus massives depuis 2014 (challenge Kaggle DecMeg), cependant les méthodes de *deep learning* n'ont toujours pas percé dans ce domaine.

De même, alors que l'intérêt grandissant pour les problématiques d'analyse de réseaux, de villes intelligentes ou de maintenance prédictive ont engendré beaucoup d'expérimentations autour des réseaux de capteurs et des séries spatio-temporelles avec des techniques de *deep-learning* [Zia+17 ; Bou+16] ; cependant, ces travaux ne semblent pas encore assez robustes pour faire face au niveau de bruit des électroencéphalogrammes.

2.3 Analyse de textes & classification de sentiments

La classification automatique de textes a reçu une attention particulière de la communauté de l'apprentissage à la fin des année 90 dans le sillage des travaux d'A. McCallum [MN+98], T. Joachims [Joa98b] et bien d'autres [SS00 ; Seb02]. Au niveau des méthodes d'apprentissage, les *Support Vector Machines* (SVM) ont démontré l'intérêt d'une formulation régularisée pour traiter des données en très grande dimension [Joa98b].

La question de la sémantique est critique en texte puisque la communauté cherche une manière de dépasser le fossé sémantique introduit par la représentation en sac de mots en calculant des distances entre les mots. Des approches à base de thésaurus ont été proposées [Mil95] et des algorithmes non supervisés ont permis d'extraire des champs lexicaux thématiques ; les modèles graphiques ont démontré leur efficacité dans cette tâche à travers les algorithmes PLSA (*Probabilistic Latent Semantic Analysis*) puis LDA (*Latent Dirichlet Allocation*) [Hof99 ; Ble+03].

Néanmoins, la compréhension du sens des mots est une tâche difficile avec des mots polysémiques, elle est parfois subjective et souvent liée à un domaine applicatif ou un contexte de phrase. Ainsi, aborder cette problématique avec un simple classifieur linéaire sur un domaine particulier est loin d'être absurde : c'est la thèse défendue par [Pan+02] pour la classification de sentiments. [4]. L'enjeu est de

4. Nous utilisons indifféremment les mots *sentiment*, *opinion* ou *polarité* dans l'ensemble de la section. Cela peut sembler réducteur, notamment pour les sentiments, mais correspond au vocabulaire de la littérature du domaine, i.e. à la croisée de chemins entre apprentissage statistique et traitement automatique de la langue naturelle. C'est aussi une traduction directe de l'anglais *Sentiment Analysis*.

caractériser la polarité de l'ensemble des textes issus des sources ouvertes comme des blogs, tweets, ou pages Facebook pour sonder l'opinion. Classer un document en polarité, même efficacement, ne permet pas d'identifier le sens individuel des mots mais permet néanmoins d'identifier les groupes de mots associés à l'expression des sentiments.

Dans la communauté de l'apprentissage statistique –hors Traitement Automatique de Langue Naturelle–, l'article de B. Pang agit comme un déclencheur [Pan+02] et engendre de nombreux travaux ultérieurs. La revue [PL+08] donne un bon panorama des approches utilisées, avant le passage au *deep learning*, sur cette tâche. Elle permet notamment de comprendre les enjeux du transfert, entre un domaine source étiqueté et un domaine cible potentiellement vierge. La première référence traitant ce cas difficile propose une stratégie d'alignement basée sur des mots pivots [Bli+07] mais d'autres approches ont été proposées, reposant sur une extension de l'espace de représentation [DI07], sur une modélisation thématique (bayesienne) des corpus [Mei+07] ou sur une factorisation matricielle [Li+10]. Mais des approches très simples, comme l'assimilation progressive des données de test par ordre de confiance avec ré-entrainement du modèle [Che+11], peuvent donner de très bonnes performances. Les principales contributions en transfert sont comparées dans la revue [PY10]. Le cadre évaluatif usuel est décrit dans [Bli+07], il consiste à appliquer les modèles appris sur une source s sur des données étiquetées d'un autre domaine (cible t). Ce cadre permet une évaluation quantitative fiable. Passer à d'autres media (non étiquetés) est plus cher, étant donné le manque de labélisation et l'écart entre les distributions de mots entre la source et la cible [MS12].

2.3.1 Formalisations linéaires & régularisations

Dans une première série d'expériences, nous nous focalisons sur les classifieurs linéaires avec une idée en tête : démontrer que des solutions simples peuvent être efficaces pour la classification de sentiments dans le cadre multi-domaines.

Descripteurs Pour la représentation des documents textuels, nous sommes repartis des modèles vectoriels classiques, comme dans [Pan+02] : sacs de mots sur des unigrammes et bigrammes. Nous n'avons pas poussé les analyses jusqu'aux descripteurs de type skip-gram ou sous-arbres syntaxiques, différentes références montrant un gain très limité par rapport à l'investissement nécessaire pour extraire les décompositions syntaxiques [Mat+05 ; PP11].

Dans tous les cas (N-grammes ou sous-séquences), nos documents sont vectorisés sous la forme $\mathbf{x} \in \mathbb{R}^d$, où d est grand (typiquement entre $5k$ et $350k$). Les caractéristiques suivent une loi de Zipf [SJ72] et les documents ont des représentations très parcimonieuses (entre 100 et 700 termes en moyenne par document). Comme la plupart des articles de la littérature [PL+08], nous avons opté pour un codage binaire des caractéristiques :

$$x_j = \begin{cases} 0 & \text{si la caractéristique (mot, N-gramme, ...) } j \text{ n'est pas dans le document} \\ 1 & \text{sinon (quelque soit le nombre d'occurences)} \end{cases} \tag{2.10}$$

Toutes les données sont étiquetées de manière binaire. Sur les données Amazon présentées ci-dessous, les revues étaient étoilées entre 1 et 5. Le consensus dans la littérature consiste –ou consistait à ce moment là– à considérer que les notes 1 et 2 correspondent à un avis négatif tandis que les notes 4 et 5 sont positives. Les notes

intermédiaires (ici 3 étoiles) sont éliminées car considérées comme ambiguës. Ainsi, pour reprendre les notations précédentes :

$$f_{repr}(\mathbf{x}^b) = \mathbf{x} \in \{0,1\}^d,\ 5k < d < 350k \tag{2.11}$$

Classification Comme la plupart du temps dans le domaine du texte, nous nous limitons dans cette première série d'experiences à la classe des modèles linéaires (sans biais) $f(\mathbf{x}) = \mathbf{x} \cdot \mathbf{w}$. Nous repartons de l'equation (2.3) et nous cherchons une combinaison de fonction Δ (adéquation aux données d'apprentissage) et Ω (capacité de généralisation) apte à répondre à cette problématique de transfert. Nous distinguons les données d'apprentissage, issues d'une source s et vectorisées $\mathbf{x}^s \in \mathbb{R}^d$ des données de test, issues de la source t, $\mathbf{x}^t \in \mathbb{R}^d$. L'enjeu était d'étudier l'impact de différentes fonctions de coût et différentes régularisations dans un cadre d'apprentissage efficace [Raf+12a ; Raf+12b ; Raf+12c]. Quatre formulations classiques sont issues de ce cadre générique. Elles correspondent aux combinaisons de deux fonctions de coût (charnière et moindres carrés) et de deux fonctions de régularisation (basées respectivement sur les normes $\mathcal{L}_1$ et $\mathcal{L}_2$ du vecteur de paramètres $\mathbf{w}$). Le tableau 2.2 synthétise ces formulations.

		Régularisation	
		$\mathcal{L}_1$ $\Omega^{(\mathcal{L}_1)}(f) = \sum_{j=1}^{d} \lvert w_j \rvert$	$\mathcal{L}_2$ $\Omega^{(\mathcal{L}_2)}(f) = \sum_{j=1}^{d} w_j^2$
Coût	Charnière (*hinge*) : $\Delta_h(\mathbf{X}, \mathbf{Y}) = \sum_{i=0}^{N} (1 - y_i f(\mathbf{x}_i))_+$	$\mathcal{L}_1$ SVM [BM98]	SVM [Bos+92]
	Moindres carrés : $\Delta_{ls}(\mathbf{X}, \mathbf{Y}) = \sum_{i=0}^{N} (y_i - f(\mathbf{x}_i))^2$	LASSO [Tib96]	Spline [Tik+77]

Table 2.2: Les quatre formulations classiques (et leurs références) obtenues en combinant deux fonctions coûts avec deux fonctions de régularisation.

Ces modèles de base permettent d'illustrer des comportements différents en terme de sélection d'informations discriminantes. Ils couvrent une partie importante de la bibliographie sur l'apprentissage supervisé. La fonction coût charnière (*hinge loss*) est une approximation convexe assez fine de l'erreur de classification. Elle se focalise sur les documents ambigus proches de la frontière de décision. A l'inverse, les moindres carrés sont plus souvent utilisés en régression. Ils optimisent un critère de corrélation et se focalisent sur les moyennes des documents d'une classe pour construire la frontière de décision. La régularisation $\mathcal{L}_2$ permet généralement de lutter contre le sur-apprentissage en conservant de bonnes performances. Lorsque nous utilisons la descente de gradient pour optimiser le vecteur $\mathbf{w}$, la mise à jour est la suivante : $\mathbf{w} \leftarrow \mathbf{w} - 2\varepsilon\mathbf{w}$ (en tenant seulement compte de la régularisation) : un poids w_j non nul n'est donc jamais mis à zéro. A l'inverse, la régularisation $\mathcal{L}_1$ est parcimonieuse : durant la descente de gradient, la mise à jour est effectuée selon $\mathbf{w} \leftarrow \mathbf{w} - \varepsilon\ \text{sign}(\mathbf{w})$. Si le signe de w_j change, le poids est mis à zéro (cf. [Fri+07]). Pour résumer, à chaque pas les w_j se déplacent vers zéro et les poids suffisamment faibles sont éliminés.

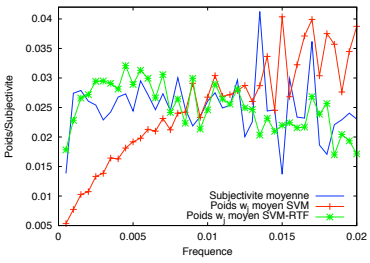


Figure 2.4: (Bleu, ligne pleine) subjectivité moyenne des termes par rapport à leur fréquence en se basant sur SentiWordNet. (Rouge+/Vert∗) poids moyen des termes après l'apprentissage en fonction de leur fréquence (courbe rouge : SVM, $\lambda = 0$ et courbe verte : SVM-RFT, $\lambda = 0.001$)

Expériences préliminaires & biais fréquentiel Les premières expériences ont montré des performances très proches d'un modèle à l'autre, elles nous ont permis de nous

Régularisation freq.			
Caract.	U	UB	UBS
Books	82.35	**85.5**	85.1
Dvd	84.25	**87.1**	85.6
Electr.	84.4	**88.2**	87.3
Kitchen	85.4	**88.3**	88.0
Mov. rev.	88.4	88.8	**92.2**
baseline = SVM			
Caract.	U	UB	UBS
Books	80.8	84.1	83.5
Dvd	82.7	84.0	84.3
Electr.	82.2	85.6	85.5
Kitchen	83.5	86.2	86.2
Kitchen	87.1	88	91.4

Ref.	[1]	[2]	[3]	[4]
Caract.	U	UBS+	UB	UBS
Books	80.4	81.4	-	-
Dvd	82.4	82.55	-	-
Electr.	84.4	84.6	-	-
Kitchen	87.7	87.1	-	-
Mov. rev.	-	-	87.1	88.9

Table 2.3: Performances obtenues en validation croisée (5 ensembles pour Amazon, 10 pour M.R. en accord avec la bibliographie).
Références : [1] (Blitzer *et al.*, 2007) [2] (Pan *et al.*, 2010) [3] (Pang *et al.*, 2004) [4] (Matsumoto *et al.*, 2005)
UBS+ correspond à des caractéristiques avancées issues de la langue naturelle.

situer par rapport à l'état de l'art (table 2.3). Vis à vis de la dimensionnalité des données, nous visions clairement à mettre en évidence l'intérêt de la régularisation $\mathcal{L}_1$ pour la sélection de caractéristiques pertinentes ; face à cet échec, nous avons donc cherché à caractériser les variables retenues par nos modèles. Afin d'associer une polarité aux mots de notre dictionnaire, nous sommes partis d'une des ressources ouvertes les plus populaires : SentiWordNet (3.0) [Bac+10]. Une fois la polarité des mots agrégée par rapport à la fréquence d'apparition des mots dans le corpus [5], nous avons obtenu la Figure 2.4. D'après SentiWordNet, la polarisation moyenne des mots est approximativement constante par rapport à leurs fréquences d'apparition. En étudiant en parallèle les poids $\mathbf{w}$ du classifieur (SVM), nous voyons à l'inverse que les mots les plus fréquents sont sur-pondérés par rapport aux autres. En conclusion, les différents modes de régularisation tendent à mettre à zéro des coefficients faibles, qui ne sont pourtant pas dénués d'opinion : il y a une marge d'amélioration.

Régularisation fréquentielle Afin de ré-équilibrer le comportement de la régularisation vis à vis des écarts pointés en Figure 2.4, nous proposons simplement de plus pénaliser les termes fréquents en introduisant un terme basé sur la fréquence documentaire ν :

$$\Omega(f) = \sum_{j=1}^{d} \nu_j \Omega_j(f), \quad \nu_j = \text{fréquence documentaire}(j) \in [0,1] \tag{2.12}$$

où $\Omega_j(f)$ décrit la composante de Ω relative au terme j. Le passage à cette formulation est aisé à implémenter et conduit à des améliorations significatives en reconnaissance d'opinion, Table 2.3, colonnes de gauche. Les différentes formes de régularisation fonctionnent toujours de manière assez semblables. Etonnamment, formulation basées sur la fonction de coût MSE (moindres carrés $-$*mean squared error*$-$) dépassent légèrement les SVM et $\mathcal{L}_1$-SVM. La Table 2.4 permet de mieux comprendre l'effet de la régularisation fréquentielle : avant la pondération fréquentielle, les 100 coefficients w_j les plus forts correspondaient à des mots apparaissant en moyenne 25 fois (coefficients les plus positifs) ou 250 fois (coefficients les plus négatifs). Après la pondération, les mots mis en avant apparaissent beaucoup moins (respectivement 12 et 15 fois en moyenne dans le corpus).

SVM-Régularisation fréq.	
Nb. occ. (+)	Nb. occ. (−)
11.85	15.04
SVM classique	
Nb. occ. (+)	Nb. occ. (−)
24.46	249.17

Table 2.4: Nombre d'occurrences moyen des termes du top 100 (100 termes les plus influents en positif et négatif) pour les classifieurs SVM à régularisation fréquentielle et SVM classique sur *Books*

En conclusion, nous montrons que jouer avec les formulations est essentiel face à des problèmes mal posés : même si le gain n'est pas à chercher entre les modèles connus, le fait de garder la main sur la formulation générale nous a permis d'introduire une pondération qui fait la différence.

Transfert de domaine Comme nous l'avons mentionné précédemment, le principal enjeu de la robustesse réside dans la capacité de transfert des modèles d'un domaine à l'autre. Pour quantifier cette robustesse, nous avons repris le protocole de [Bli+07] : nous apprenons les modèles (à régularisation fréquentielle) dans un domaine source et nous évaluons les performances sur un autre domaine (de test). Nous avons donc repris les bases classiques du domaine et confronté notre système avec les

5. Nous travaillons évidemment sur le sous-ensemble du vocabulaire correspondant à l'intersection entre SentiWordNet et notre corpus. De plus, SentiWordNet décrit la polysémie de nombreux termes que nous n'utilisons pas : nous moyennons simplement les polarités des différentes formes du terme cible.

approches de l'état de l'art [Bli+07 ; Pan+10]. Nous cherchons à montrer l'intérêt d'une formulation robuste et très efficace, mais sans modélisation de la problématique du transfert, sur la tâche de classification de sentiment dans une de ses versions les plus ardues.

La Table 2.5 montre les performances de nos modèles en transfert. Globalement, les résultats sont très intéressants, en particulier pour la formulation LASSO. Cependant, nous n'avons pas réussi à égaler l'état de l'art. Notre modèle garde pour lui sa simplicité à la fois en apprentissage et en inférence par rapport à des approches nettement plus coûteuses, mais il rend un point de performance en moyenne sur les 12 expériences de transfert.

	SVM			LASSO			L1 SVM			Spline		
Descr.	U	UB	UBS	U	UB	UBS	U	UB	UBS	U	UB	UBS
B → D	81.35	81.75	82.1	79.6	82.65	**83.5**	81.4	83.25	83.2	80.6	82.8	82.45
E → D	73.95	74.9	75.65	*68.3*	**77**	76.35	*72.95*	75.95	76.45	*72.1*	74.25	76.1
K → D	*73.15*	77.2	**77.5**	*70.3*	76.2	75.65	74.65	75.7	76	*71.45*	76.6	76.25
D → B	80.2	83.35	82.5	*78.35*	82.5	81.7	*78.95*	82.45	82.6	80.8	**83.6**	83
E → B	*68.95*	71.8	71.65	70.95	72.2	72.25	69.55	71.4	**72.9**	*68.55*	71.9	72.05
K → B	*69.6*	73.9	**74.1**	*67.6*	72.8	72.7	71.5	73.65	73.6	*68.35*	73.45	72.85
B → E	69.45	70.1	70.95	*68.1*	72	72.35	70.28	**72.4**	71.65	*67.85*	71.3	71.95
D → E	*69.9*	72.45	73.15	*68.3*	73.6	**74.8**	*70.8*	73.75	74.3	*70.7*	73.65	73.85
K → E	*81.5*	85.9	85.75	*79.05*	85.45	86.1	82.2	**86.15**	85.9	*80.65*	85.05	84.95
B → K	73.25	75.35	*71.1*	*70.5*	75.35	74.8	72.55	**75.8**	74.85	72.2	75.3	73.85
D → K	*72.1*	76.1	*73.55*	*71.75*	*77*	73.4	*72.85*	75.4	*73.05*	73.9	**77.65**	*73.4*
E → K	*81.9*	85.6	*78.9*	*78.55*	86.2	82.15	*81.5*	**86.3**	*81.85*	*81.25*	86.15	*79.65*
Moy.	74.61	77.37	76.41	72.61	**77.75**	77.15	74.93	77.68	77.20	74.03	77.64	76.70

Table 2.5: Taux de reconnaissance sur les données Amazon dans le cadre multi-domaines. La première colonne décrit les sous-corpus utilisés (par exemple, B → D signifie que *Books* a été utilisée pour apprendre le modèle tandis que *DVD*, la cible, a été utilisée pour évaluer les performances). Les meilleures performances de chaque ligne sont mises en gras et les résultats médiocres ($\leq 95\%$ du meilleur) sont en italique. Les moyennes de performances sur les 12 expériences doivent être comparées à 78.65% (Pan *et al.*, 2010) et 77.95% (Blitzer *et al.*, 2007)

2.3.2 Masse de données

En conservant la même philosophie, consistant à chercher des solutions simples face à des problèmes complexes, nous avons exploré la piste de l'augmentation de la taille des ensembles d'apprentissage. Une telle approche serait absurde dans pratiquement n'importe quelle autre application, les étiquettes étant habituellement la ressource la plus chère et donc le facteur limitant des expériences. Mais les revues utilisées en classification de sentiments sont justement étiquetées par leurs auteurs sur le web participatif et disponibles à l'infini. Tous les détails associés à ces expériences sont donnés dans [Raf+13].

Jeux de données Nous sommes donc parti sur un éventail de corpus plus large, chacun étant plus étoffé (Table 2.6). Les documents sont représentés en sacs de mots et nous limitons le dictionnaire aux 5000 uni-grammes et bi-grammes les plus fréquents comme cela est généralement fait dans la littérature. Comme précédemment, nous utilisons un codage binaire (présentiel). Les données comptent 25 domaines

Amazon (25 domaines) [Bli+07]			
Domaine	# app.	# test	% ex. neg.
Toys	6318	2527	19.63%
Software	1032	413	37.77%
Apparel	4470	1788	14.49%
Video	8694	3478	13.63%
Automotive	362	145	20.69%
Jewelry	982	393	15.01%
Grocery	1238	495	13.54%
Camera	2652	1061	16.31%
Baby	2046	818	21.39%
Magazines	1195	478	22.59%
Cell	464	186	37.10%
Outdoor	729	292	20.55%
Health	3254	1301	21.21%
Music	10625	24872	8.33%
Videogame	720	288	17.01%
Beauty	1314	526	15.78%
Sports	2679	1072	18.75%
Food	691	277	13.36%
Office	195	78	0.16%
Instruments	164	65	0.15%
Tools	32	11	0.03%
Ensembles de test [Bli+07]			
Domaine	# app.	# test	% ex. neg.
Books	10625	10857	12.08%
DVDs	10625	9218	14.16%
Electronics	10196	4079	21.94%
Kitchen	9233	3693	20.96%
Amazon fusionné [Bli+07]			
Tous domaines	90535	68411	15.04%
Movie Reviews (large) [Maa+11]			
Movies	50000	-	50%
TripAdvisor [Wan+10]			
Hotel & cars	50000	-	50%

Table 2.6: Description des données.

mais nous n'envisageons d'abord que 4 jeux de test. Cet ensemble de test est fixé une fois pour toutes : il s'agit des parties de test des sous-corpus d'Amazon *Books*, *Dvd*, *Electronics* et *Kitchen*. Cela représente un total de 27847 documents. Les expériences montrent deux comportements types : d'une part *Books* et *Dvd* qui sont assez proches et d'autre part, *Electronics* et *Kitchen*, qui réagissent de la même manière aux différents tests.

En fonction des expériences, nous utilisons les ensembles d'apprentissage suivants :

— pour le cas i.i.d (intra-domaine), les parties d'apprentissage des domaines cibles sont utilisés (*Books*, *Dvd*, *Electronics* et *Kitchen*).

— Dans les expériences suivantes, nous utilisons les sous-domaines Amazon externes (différents de la cible). Amazon compte 25 domaines, il y a donc 24 domaines externes pour chaque cible. En adaptation mono et multi-sources, aucune donnée cible n'est utilisée lors de l'apprentissage.

— Pour le transfert hétérogène, l'apprentissage est réalisé sur les corpus 50k *movie reviews* [Maa+11] et 50k *TripAdvisor* [Wan+10] dont les critiques ne concernent pas des produits physiques mais respectivement des films, des hôtels et des chauffeurs. Les revues ont également une forme différente : les critiques de films font 740 mots en moyenne contre 200 pour les revues Amazon.

Expériences et performances Les performances de référence sur les 4 sous corpus cibles d'Amazon sont données en Figure 2.5. Il ne faut pas confondre ces chiffres et ceux de la section précédente, obtenus sur d'autres corpus de taille et d'équilibrage différents.

La première question est de quantifier la perte lorsque nous apprenons nos modèles à partir d'autres sources. Tous les résultats sont présentés en Figure **??**. Nous voyons immédiatement que la perte est importante et que la nature de la source impacte fortement les résultats. La seconde question, que nous approfondirons ensuite, est de savoir si la prise en compte de plus d'informations augmente effectivement les performances. La Figure 2.5 (droite) démontre que c'est bien le cas [6].

Domaine	Tx de reco.	♯ ens. test
Books	91.1%	10857
DVD	90.6%	9218
Electronics	90.6%	4079
Kitchen	91.7%	3693
Total	91%	27847

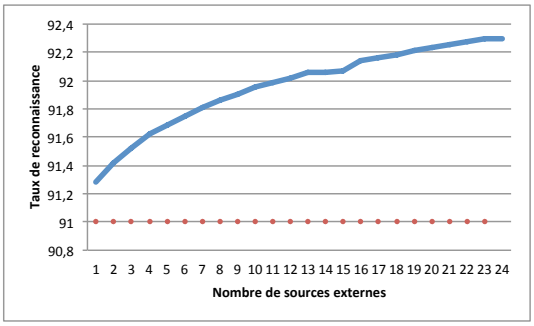
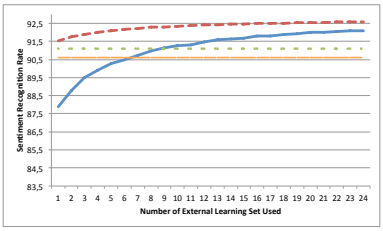


Figure 2.5: Tableau de gauche : taux de reconnaissance intra-domaine sur *Books*, *DVD*, *Electronics* et *Kitchen*. Courbe de droite : évolution du taux (moyenné sur les 4 expériences) lorsque l'ensemble d'apprentissage est enrichi avec des sources externes.

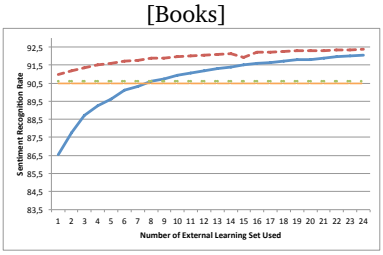
Nous mettons maintenant en perspective les performances intra-domaine avec les performances en transfert, lorsque le nombre de sources utilisées grandit. Les

6. les chiffres correspondent à des moyennes d'expériences avec un tirage aléatoire des sources. Tous les détails sont dans [Raf+13].

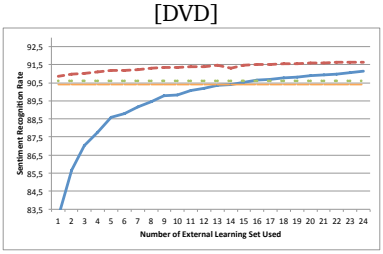
quatre ensembles de test ne changent pas mais *Books*, *Dvd*, *Electronics* et *Kitchen* ne sont plus utilisés en apprentissage. La Figure 2.6 montre l'évolution des performances en fonction du nombre de sources externes utilisées, les taux de reconnaissance sont moyennés sur plusieurs expériences. La figure compare quatre performances pour chaque cible : (1) ligne verte pointillée : expérience intra-domaine, chiffres issus du tableau 2.5, (2) ligne fine orange horizontale : transfert mono-source avec un oracle donnant la meilleure source pour chaque cible, (3) ligne bleue : transfert multi-sources, taux de reconnaissance en fonction du nombre de sources externes utilisées (le premier point de la courbe correspond au transfert mono-source), (4) ligne rouge pointillée : expériences intra-domaine + enrichissement de $n = 1$ à 24 sources externes (la cible est utilisée en apprentissage).



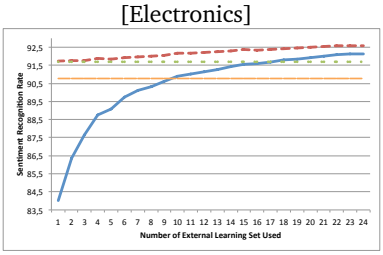
[Books]



[DVD]



[Electronics]



[Kitchen]

Figure 2.6: Taux de reconnaissance sur les 4 cibles en fonction du nombre de sources externes utilisées. Performance intra-domaine (pointillés verts), transfert mono-source+oracle (ligne fine orange), transfert multi-sources (bleu), intra-domaine+enrichissement (pointillés rouges).

Le transfert multi-sources permet de dépasser systématiquement la performance mono-source+oracle (pour choisir la meilleure source). Nous reproduisons ici les résultats de [Dre+10] mais avec un système nettement plus simple. Les gains en taux de reconnaissance vont de 0,7 à 1,6% comparé à l'oracle. Quelque soit la cible, la courbe multi-sources dépasse toujours la courbe intra-domaine (de 0,4 à 1,4%) : cette observation est nouvelle et à notre connaissance aucun article publié ne propose un tel résultat. Ces performances doivent être comparées notamment à [Glo+11] qui utilise le même jeu de données [7] sans aboutir à un gain systématique. Par rapport aux performances intra-domaine+enrichissement, l'écart reste toujours inférieur à 0,5% : la complexité et le coût des modèles de transfert deviennent des freins importants face aux maigres perspectives de gain.

Topologie & sélection des sources de données La question que nous nous sommes posée vis à vis des premiers résultats obtenus est de savoir s'il est utile de sélectionner les sources en fonction de la cible visée, à la manière de [Dre+10]. Afin de répondre très simplement, nous avons utilisé la divergence de Kullback-Leibler (symétrique) pour mesurer la distance entre corpus (modélisés comme des distributions sur les mots). Cette première étape nous donne une topologie des domaines et permet d'implémenter une stratégie de sélection basée sur les corpus les plus proches.

La Figure 2.7 illustre deux blocs principaux dans les données *Amazon*, autour de divers appareillages d'une part et de biens culturels (livre, dvd, ...) d'autre part. Quelques bases sont peu reliées aux autres, mais c'est principalement du à leur taille, infiniment plus petite que pour les autres domaines. Coté performances, les résultats sont mitigés : les bases *kitchen* et *electronics* profitent d'un apprentissage excluant les biens culturels (+2%). A l'inverse, les bases *books* et *dvd* tirent profit de l'ensemble des ressources disponibles.

Adaptation en contexte hétérogène La robustesse des classifieurs devient critique lorsque les distributions entre la source et la cible s'écartent. C'est le cas lorsque nous considérons deux média différents, comme par exemple des revues étiquetées pour apprendre des modèles et des tweets en inférence [MS12].

Comme dans la série d'expériences précédente, nous avons utilisé plusieurs jeux de données et comparé les performances sur un *Golden Standard* de tweets étiqueté [Che+12a]. Nous nous sommes également intéressés à la mise en concurrence entre des systèmes basiques appris sur des corpus larges et des systèmes plus complexes [Dre+10 ; DI07]. L'article [GS+13] rassemble les résultats obtenus. La première

7. sans les étiquettes des domaines externes

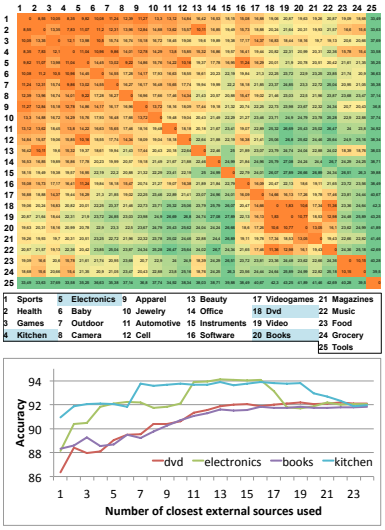


Figure 2.7: Matrice des divergences KL (symétrisées) entre les 25 sources *Amazon*. Evolution des performances sur les 4 cibles en fonction du nombre de sources (les plus proches) utilisées.

série d'expériences met en évidence l'intérêt d'utiliser une source centrée sur une thématique proche de la cible (les films dans le cas présent). La seconde série démontre encore une fois empiriquement le gain lié à l'utilisation de plus de données en apprentissage, même si la thématique est plus éloignée. Une fois passée un certain seuil sur la masse de données utilisées en apprentissage, le gain des algorithmes de transfert devient négligeable : suffisamment de cas de figures ont été modélisés en apprentissage pour qu'un modèle simple réussisse à interpréter correctement les données de test.

Les expériences que nous avons menées sur l'adaptation entre différents media sont plus larges que la synthèse ci-dessus. Tous les détails sont donnés dans la publication [GS+13].

2.3.3 Vers des architectures neuronales

Dans le domaine de l'analyse de textes en général et de la classification de sentiments en particulier, trois familles d'architectures se distinguent clairement à la fin des année 2000 : les modèles génératifs, dans le sillage de [Mei+07] ; les approches linéaires discriminantes qui occupent une place importante dans la revue [PL+08] ; les architectures neuronales qui émergent en texte autour de [CW08] et vont donner lieu à des applications en classification de sentiments [Bes+11].

Nous avons mis de coté les modèles génératifs dont les performances semblaient en retrait pour comparer les approches linéaires et neuronales dans [Raf+11 ; Raf+12d ; Raf+12c]. La conclusion générale est que les performances sont proches, même si les erreurs ne sont pas aux mêmes endroits. La capacité des réseaux de neurones à exploiter de grands corpus de données est intéressante mais, malgré une architecture faisant la part belle à l'apprentissage de représentation [CW08], la sémantique extraite reste très aléatoire. Ces architectures permettent déjà de limiter les pré-traitements et d'attaquer des données brutes...Néanmoins, le surcoût de réglage et d'apprentissage de l'architecture est encore supérieur aux économies évoquées précédemment.

A l'époque de ces travaux, les réseaux de neurones offrent donc déjà des perspectives très prometteuses. Malgré les difficultés d'implémentation et le temps de calcul nécessaire pour apprendre ces architectures complexes, ils permettent de tirer parti de grandes masses d'informations textuelles, ce qui est une clé en analyse de sentiments. De plus, [Col+11a] propose déjà de repartir de modèles de langue pré-entrainés et de multiplier les tâches à partir de ces représentations, ce qui deviendra la clé du succès dans de nombreuses applications de la fin des années 2010. Cette possibilité est aussi liée à l'utilisation d'un cadre logiciel ouvert et performant [Col+11b] qui illustre bien l'aspect précurseur des travaux de R. Collobert.

Malgré l'essor des performances des approches neuronales sur les données textuelles, nous avons néanmoins tenu à conserver cette section sur les sacs de mots pour deux raisons. D'une part, dans un milieu scientifique qui ne travaille plus que sur les architectures profondes, il est très important de garder à l'esprit les bonnes performances générales de ces approches très simples et légères. D'autre part, il ne faut pas négliger l'évolution des ressources –corpus et matériel– qui permettent directement de gagner en performance : si les modèles de langues pré-entrainés

représentent évidemment une évolution majeure du domaine, les anciens systèmes bénéficient également de l'évolution des ressources.

2.4 Conclusion et perspectives

Le parallèle entre les deux tâches considérées dans ce chapitre est intéressant : la nature des données et la problématique du transfert posent des questions qui nous amènent à réfléchir autour de la construction des caractéristiques descriptives, de leur sélection à l'entrée des modules d'apprentissage statistique puis aux formulations utilisées pour entrainer nos modèles.

Nous avons montré que la régularisation d'une part ou les procédures de sélection de caractéristiques d'autre part permettent d'améliorer les performances lorsque les données sont décrites en très grande dimension. Nous avons également montré que ces procédures de réduction de la dimensionnalité et de pondération des caractéristiques permettent de proposer une première interprétation des décisions en isolant les mots discriminants ou les parties du scalp participant activement à la décision. Néanmoins, l'interprétabilité reste limitée et les architectures étudiées ne proposent pas –ou très peu– d'analyse sémantique. Cette limite est critique par rapport au transfert entre les tâches lorsque nous dressons un parallèle avec le raisonnement humain : il semble clair que le pouvoir de généralisation de nos décisions provient essentiellement de notre compréhension globale des concepts, notre capacité à prendre du recul sur différentes situations et dresser des parallèles à haut niveau –ce que les anglophones désignent par *common sense reasoning*–. La sémantique sera effectivement l'enjeu majeur des travaux présentés dans la suite de ce manuscrit.

Ces deux études de cas permettent également de bien cerner la chronologie autour de la popularité des différentes approches d'apprentissage statistique. Autour de 2010, nous voyons clairement que l'usage de réseaux de neurones devient une possibilité mais n'est pas encore une évidence. L'opportunité de traiter des données avec moins de réglages experts en entrée n'est pas encore rentable au regard du coût de développement, réglage et apprentissage de ces architectures. Les travaux qui modifieront radicalement le rapport des communautés scientifiques et industrielles à la vision par ordinateur [Kri+12 ; SZ14] et au traitement du langage naturel [Mik+13 ; Cho+14] réduiront justement ces coûts en apportant des cadres efficaces et flexibles exploitant massivement le GPU pour réduire le temps de calcul.

Cette introduction à l'apprentissage de représentation constitue une transition vers le prochain chapitre, où la représentation de données devient une fin en soi. Nous chercherons alors à apprendre des profils de films, de personnes ou d'objets, la tâche de classification supervisée étant un moyen d'obtenir une représentation significative. L'idée de base du chapitre suivant est que l'apprentissage de représentations doit nous permettre de mieux classer... dans différents contextes, pour différents problèmes ou différentes tâches.

> *On a beaucoup parlé de la face de Dieu.*
> *Jamais de son profil.*
>
> **— Jean-Claude Brisville**

3

Apprentissage de représentation : du profiling à la sémantique

La problématique de la recommandation est devenue incontournable dans la communauté de l'apprentissage automatique au milieu des années 2000, portée, entre autres, par le challenge Netflix sur la prédiction des goûts cinématographiques des clients de la firme éponyme [BK07]. Nous nous rendons aujourd'hui compte que la notion de profil est intégrée à de nombreuses applications d'accès à l'information. Les algorithmes initialement dévolus aux systèmes de recommandation sont maintenant utilisés partout où il est possible d'introduire de la personnalisation. Le champ de mise en œuvre ne se limite pas à Amazon ou Netflix mais englobe la publicité, les newsletters voire, la contextualisation et la personnalisation dans l'analyse de données en général. Les moteurs de recherche se personnalisent ; les compteurs d'eau ou d'électricité deviennent des capteurs pour construire des villes plus intelligentes ; les systèmes d'information des transports en commun pourraient demain intégrer des prédiction personnalisées.

Nous abordons ainsi la problématique générale de l'apprentissage de représentation via une application à la fois emblématique et précurseure : la recommandation. Mais l'ambition de ce chapitre est de montrer comment ces mêmes techniques peuvent être exploitées pour améliorer la modélisation des usagers dans les transports en commun ou l'analyse sémantique des textes.

Recommandation Une part du succès d'Amazon autour des années 2000 est souvent attribuée –à tort ?– aux suggestions intégrées dans l'interface. Dès lors, les systèmes de recommandation sont perçus comme une source de valeur et vont se développer très rapidement entrainant des avancées significatives sur le plan algorithmique. Les premières propositions de recommandation étaient fondées sur la proximité entre les descriptions d'item [MR00] : elles manquaient de pertinence, notamment dans les grands catalogues où le manque de score d'autorité était critique. Le passage aux algorithmes de filtrage collaboratif, où les suggestions proviennent des goût partagés des utilisateurs a amélioré la qualité des propositions tout en ouvrant la voie de l'apprentissage de profil [LS99 ; DK11]. Du point de vue de l'apprentissage statistique, la problématique de la recommandation par filtrage collaboratif pose de nombreux challenges :

— encoder efficacement les goûts d'une personne à partir de ses interactions avec le système et exploiter ce profil pour faire des propositions pertinentes [Kor+09],

— trouver des solutions pour amorcer le système pour les nouveaux utilisateurs ou items –le démarrage à froid, ou *cold start*– [Bur07],

— expliquer les propositions faites à l'utilisateur pour mieux le convaincre [TM07],

— évaluer la pertinence des suggestions, ce qui est très difficile sur un système pro-actif et pose des questions sur la formulation même de la tâche de recommandation [Ste10 ; Ren+12],

— mesurer l'originalité des propositions par rapport à ce que l'utilisateur aurait découvert sans –la sérendipité– [Ge+10].

Profiling & apprentissage de représentation N'importe quelle architecture neuronale standard contient des couches cachées qui reflètent l'encodage des données d'entrée par rapport à la –ou les– tâches affectées à l'architecture. De ce point de vue, de nombreuses approches peuvent être considérées comme de l'apprentissage de représentation [Ben+13]. Néanmoins, les opportunités ne sont pas les mêmes selon les problématiques abordées : quand les données d'entrée sont des images et la tâche de la reconnaissance d'objets, la représentation représente un espace très compact encodant le contenu voire le style d'une image [Gat+15] ; lorsque les entrées sont des mots, il est intéressant d'exploiter les représentations apprises dans plusieurs tâches [Col+11a ; Mik+13 ; Dev+18]. Les personnes sont alors un cas particulier : chercher à prédire les items d'intérêt pour quelqu'un revient à encoder ses préférences. Rien n'empêche d'enrichir la représentation en cherchant également à prédire la classe d'age de la personne ou n'importe quel autre attribut disponible en apprentissage [SG16].
De ce point de vue, le profiling est un cas particulier de l'apprentissage de représentation.

Apprentissage de représentation & données hétérogènes Si la recommandation est une instance d'apprentissage de représentation parmi d'autres, la problématique introduit malgré tout une spécificité : la gestion de données hétérogènes. En effet, la nature même du problème impose de projeter dans le même espace les profils des personnes et des items.
Les possibilités sont vastes pour différentes applications. Par exemple, partant de deux architectures apprises séparément pour encoder respectivement du texte et de l'image il est possible d'aligner les représentations correspondant aux mêmes concepts dans les deux modalités. Dans ce nouvel espace multimodal, nous pouvons générer du texte correspondant à des images –*c*aptioning– [Mao+14], une image correspondant à du texte [Ree+16] ou répondre à des questions formulées textuellement et portant sur une image –*V*isual Question Answering, VQA– [BY+17] [1].
Pour revenir à la recommandation, il est possible de modéliser les aspects d'une image qui interpellent un utilisateur [HM16] ou sur les spécificités d'une bande audio [DS14].

Vers plus de sémantique Comme nous l'avons souligné en conclusion du chapitre précédent, l'amélioration des performances en transfert ou sur des tâches différentes requière une meilleure compréhension des données, une forme de prise de recul par rapport à la problématique pour faire un parallèle simpliste avec le raisonnement humain. De ce point de vue, l'apprentissage de modèles de langue ré-utilisables est une avancée majeure pour la gestion des données textuelles [Col+11a ; Mik+13 ; Pet+18 ; Dev+18]. Ces contributions ont permis une bien meilleure compréhension

1. Les aspects génératifs sont étudiés dans le chapitre suivant.

des mots puis des phrases. La prise en compte de données hétérogènes est une opportunité pour encore mieux décrire le sens des mots en faisant le lien entre différentes modalités d'expression [Soc+14].
Les avancées des dix dernières années autour de la sémantique sont donc intrinsèquement liées aux possibilités offertes par la plasticité des nouvelles architectures neuronales. Pour bien saisir ces opportunités, il faut considérer ces architectures dans leur contexte en incluant les plateformes logicielles avancées permettant des développements rapides [Col+11b ; Aba+16] et l'évolution des bonnes pratiques en recherche vers plus de reproductibilité des résultats à travers des mises à disposition systématiques de code.

Contributions et organisation L'organisation de ce chapitre reprend celle de cette introduction : nous commençons par rappeler la formulation classique d'un problème de recommandation en faisant le lien avec l'apprentissage de représentation. Nous étudions ensuite les possibilités offertes par ces algorithmes dans le cadre de l'extraction de profils à partir de bases de revues en exploitant une composante textuelle [Pou+14d ; Noz+14 ; Pou+15a]. Ces travaux ont débouché sur une proposition originale pour aborder la problématique du démarrage à froid [Dia+17b].

Partant du constat que les utilisateurs laissent de plus en plus de traces de navigation nécessitant une modélisation de la dynamique, nous avons travaillé dans ce sens et proposé plusieurs contributions [GS+14 ; GS+15 ; GS+16]. Ces travaux permettent de prédire les comportements des utilisateurs dans le temps ; ils reposent sur des algorithmes efficaces permettant de passer à l'échelle.

La section suivante s'éloigne de la recommandation de produits pour présenter nos travaux dans le cadre des *smart-cities* ; en particulier, nous avons développé des solutions innovantes pour la modélisation des usagers des transports en commun. Nos publications concernent la modélisation des usagers, sur les comportements fréquents et occasionnels [Pou+14c ; Pou+16 ; Ton+16] ; ces modèles nous permettent ensuite de différencier des situations normales et les anomalies sur le réseau [Ton+17 ; Ton+18a].

La conclusion de ce chapitre sera axée sur les autres possibilités offertes par les espaces de représentation afin d'attaquer des problématiques plus avancées. Ainsi, elle constituera une transition logique vers le dernier chapitre de manuscrit.

3.1 Formulation classique du filtrage collaboratif

Les entrées du problème classique de recommandation sont particulières : il s'agit de deux entités discrètes, un utilisateur $u \in \mathcal{U}$ d'une part et un item $i \in \mathcal{I}$ d'autre part. Ce couple est associé à une note $r_{ui} \in \mathbb{R}$. Ces deux entités sont tirés dans des ensembles finis et les prédictions ne seront disponibles que dans ce cadre. La formulation du problème est donc transductive. L'idée originale consiste à apprendre la fonction de représentation :

$$\begin{array}{rccc} & U & & \mathbb{R}^z \\ f_{repr} : & u & \rightarrow & \mathbf{u} \end{array} \qquad \begin{array}{rccc} & I & & \mathbb{R}^z \\ f_{repr} : & i & \rightarrow & \mathbf{i} \end{array} \tag{3.1}$$

La formulation est d'autant plus originale que la fonction de prédiction ne comporte pas de paramètres. Ainsi, les seuls paramètres sont ceux des représentations.

$$f : \begin{array}{ccc} \mathbb{R}^z \times \mathbb{R}^z & & \mathbb{R} \\ \mathbf{u}, \mathbf{i} & \rightarrow & \mathbf{u} \cdot \mathbf{i} \end{array}, \qquad \text{soit : } \hat{r_{ui}} = \mathbf{u} \cdot \mathbf{i} \tag{3.2}$$

Lorsque nous considérons les r_{ui} comme des éléments d'une matrice incomplète R, le problème peut être formulé comme une factorisation dans un espace réduit à z dimensions. [BK07] ont montré l'intérêt de travailler sur des matrices de termes positifs, aboutissant à l'algorithme de la NMF *–Non-negative Matrix Factorization–*. Pour rendre la formulation plus efficace et mieux diriger la phase d'apprentissage, la plupart des approches utilisent en fait une fonction de prédiction de la forme $f(\mathbf{u}, \mathbf{i}) = b + b_u + b_i + \mathbf{u} \cdot \mathbf{i}$ où les $b_.$ sont des scalaires encodant respectivement le biais général, celui lié à l'utilisateur u et celui de l'item i. L'apprentissage de ces représentations optimise l'erreur quadratique moyenne, sur l'ensemble des critiques de la base d'entraînement (3.3) avec une régularisation $\mathcal{L}_2$ sur les paramètres pour contrer un possible sur-apprentissage [2].

$$\{\mathbf{u}, \mathbf{i}\}^\star = \underset{\mathbf{u},\mathbf{i}}{\operatorname{argmin}} \sum_{(u,i)} (r_{u,i} - f(\mathbf{u}, \mathbf{i}))^2 + \alpha_u \|U\|_F^2 + \alpha_i \|I\|_F^2 \tag{3.3}$$

en notant $U \in \mathbb{R}^{|\mathcal{U}| \times z}$ et $I \in \mathbb{R}^{|\mathcal{I}| \times z}$ les matrices contenant tous les profils $\mathbf{u}$ et $\mathbf{i}$. Les compromis de régularisation α sont optimisés en validation croisée sur les données d'apprentissage. Les profils sont appris dans un espace de dimension z assez faible : la seule manière de faire monter une note est donc de créer un point de correspondance entre $\mathbf{u}$ et $\mathbf{i}$: une dimension commune non nulle. A l'issue du processus d'apprentissage, l'espace est donc structuré en différentes communautés (groupes de coordonnées élevées sur les mêmes dimensions).

Ces approches et leurs variantes représentent l'état de l'art pour estimer la note que mettrait un individu à un produit. Les algorithmes de recommandation ont franchi un premier palier d'efficacité (et de popularité) avec le challenge Netflix [BL+07] qui a consacré le filtrage collaboratif par factorisation matricielle [Kor+09]. Ces dernières années, ces algorithmes sont devenus de plus en plus performants grâce à la prise en compte d'un nombre croissant de facteurs comme le temps [Kor10 ; ML13b], les liens sociaux [Guy15] ou encore le texte [ML13a]. Cependant, même si les notes sont obtenues sur des correspondances d'aspects latents, ceux-ci sont difficilement interprétables et les méthodes sont encore régulièrement qualifiées de boites noires [BEK14]. Pourtant, la précision des suggestions est tout aussi importante que d'expliquer en quoi celles-ci sont appropriées [TM07].

Nous allons maintenant étudier différentes propositions que nous avons faites pour exploiter les données textuelles dans le double but d'améliorer la recommandation et de l'expliquer.

3.2 Profiling & données textuelles en recommandation

La place du texte dans les systèmes de recommandation a beaucoup fluctué depuis le début des années 2000 : elle est centrale dans les systèmes basés sur le

2. Il y a en général plus de paramètres que de notes à estimer : la régularisation est impérative.

contenu [Lop+11], marginale dans les approches de filtrage collaboratif plus centrées sur les données d'interactions utilisateurs [BK07]. Pour tirer parti des avantages des deux approches, différents systèmes hybrides ont été envisagés [Bur07].

[Gan+09] propose un système multi-aspects où le texte des revues de restaurants permet d'extraire les points (présentation, cuisine,...) qui ont été appréciés ou pas pour améliorer la recommandation en se focalisant sur les aspects importants pour l'utilisateur. Deux autres études récentes se positionnent à l'intersection de la recommandation et de la fouille d'opinion. Dans la référence [Poi+10], les auteurs ont eu recours à la classification de sentiments avant la phase de recommandation pour annoter des textes. Nous nous positionnons plus près des travaux [ML13a] où le texte des revues d'un utilisateur est utilisé pour affiner son profil : l'idée est de combiner les informations de notes et des traces écrites du passé pour améliorer la recommandation. Nos propositions reprennent cette philosophie qui vise non seulement à améliorer les profils mais aussi à expliquer les propositions du système.

3.2.1 Intégration du texte brut *vs* modélisation thématique

Nous avons proposé plusieurs options pour l'intégration des données textuelles dans les systèmes de recommandation. Historiquement, il y a une proximité évidente entre les techniques d'extraction thématique en analyse de textes [Dee+90] et la modélisation des utilisateurs en recommandation ; dans les deux cas, il s'agit de factoriser une matrice parcimonieuse –matrice de sacs de mots ou matrice des notes d'utilisateurs– pour faire émerger des motifs caractéristiques, des co-occurences significatives. Coté recommandation, l'idée est de construire des facteurs latents modélisant des affinités communes fréquentes pour des groupes de produits ; coté texte, les facteurs latents correspondent à des groupes de termes souvent présents ensemble dans les documents (i.e. des champs lexicaux associés à des thématiques). L'idée de construire un espace latent commun pour représenter –et enrichir– à la fois les profils utilisateurs et thématiques est récente [ML13a].

Modélisation. Nous avons cherché dans [Pou+14d] à mettre en perspective la modélisation thématique de [ML13a] par rapport à une idée encore plus simple : intégrer directement les mots sur lesquels un utilisateur se basent dans ses revues comme un élément de son profil. Nous construisons un modèle pour approximer $r_{u,i}$ en utilisant la formulation suivante :

$$\begin{array}{rcll} \hat{r}_{u,i} & = & \lambda_0\phi_0+ & \text{historique moyen} \\ & & \lambda_1\phi_1(u)+ & \text{historique de l'utilisateur} \\ & & \lambda_2\phi_2(i)+ & \text{historique du produit} \\ & & \lambda_3\phi_3(u,i)+ & \text{factorisation matricielle} \\ & & \lambda_4\phi_4(d_{u,i}) & \text{documents associés à l'utilisateur et au produit} \end{array} \tag{3.4}$$

Le dernier terme pourra modéliser du texte brut, associé à du sentiment, ou une distribution de thématiques pour estimer un matching entre i et u.

Dans [ML13a], les auteurs utilisent une variante de *LDA*, (*Latent Dirichlet Allocation*) pour projeter les textes dans l'espace latent. Ils proposent d'intégrer directement la représentation latente des documents dans l'algorithme de factorisation matricielle présenté précédemment, en utilisant une technique d'optimisation alternée. Nous proposons d'utiliser *LDA* comme une fonction de projection dans l'espace latent ψ. En notant $d_{u,*}$ et $d_{*,i}$ les concaténations respectives de l'ensemble des textes de

l'utilisateur u ou sur l'objet i, nous définissons le modèle de concordance thématique suivant :

$$\phi_{L4}(u,i) = \psi(d_{u,*}) \cdot \psi(d_{*,i}) \tag{3.5}$$

Dans un deuxième temps, nous proposons un calcul de matching basé sur du texte brut, en séparant les contributions associées aux différentes polarités pour mieux estimer le profil de l'utilisateur. Après avoir passé l'ensemble des données d'apprentissage en sacs de mots (bow), nous utilisons un modèle Bayesien naïf pour représenter respectivement : les revues de l'utilisateur (bow_u), les revues positives/négatives de u ($bow_u^{(+)}$ et $bow_u^{(-)}$) et les revues associées à l'item i (bow_i, $bow_i^{(+)}$ et $bow_i^{(-)}$). La prédiction basée sur le texte brut est calculée comme une combinaison linéaire des comparaisons entre les modèles de sentiments de l'utilisateur u et de l'item i :

$$\begin{aligned}\phi_{T4}(u,i) \quad = \quad & \lambda_{t1}\cos(bow_u, bow_i) \quad + \lambda_{t2}\cos(bow_u, bow_i^{(+)}) \quad + \lambda_{t3}\cos(bow_u, bow_i^{(-)}) + \\ & \lambda_{t4}\cos(bow_u^{(+)}, bow_i) \quad + \lambda_{t5}\cos(bow_u^{(+)}, bow_i^{(+)}) \quad + \lambda_{t6}\cos(bow_u^{(+)}, bow_i^{(-)}) + \\ & \lambda_{t7}\cos(bow_u^{(-)}, bow_i) \quad + \lambda_{t8}\cos(bow_u^{(-)}, bow_i^{(+)}) \quad + \lambda_{t9}\cos(bow_u^{(-)}, bow_i^{(-)}))\end{aligned} \tag{3.6}$$

Les coefficients λ sont optimisés sur les données de validation.

Données. Les données utilisées sont des revues anglophones annotées extraites des sites *ratebeer.com* [3] et *amazon.com* [4]. Nous avons fait varier la taille des bases de données pour étudier l'impact sur l'intérêt des données textuelles (cf suffixe des noms de BD & détails dans la publication [Pou+14d]).

Résultats. Nous présentons dans cette section les résultats obtenus sur les différents corpus *Amazon* et *RateBeer*. Dans les tableaux de performances 3.1, nous présentons d'abord les 3 références ϕ_0, $\phi_1(u)$, $\phi_2(i)$ correspondant aux différentes notes moyennes. Nous nommons les modèles $\phi_3(u,i)$, ϕ_{L4} et ϕ_{T4} mais ce sont en fait des modèles composites qui intègrent les biais.

Base	ϕ_0	$\phi_1(u)$	$\phi_2(i)$	$\phi_3(u,i)$	ϕ_{L4}	ϕ_{T4}
RB_U50_I200	0,67575	0,65325	0,20913	0,19776	**0,19208**	0,19508
RB_U500_I2k	0,56850	0,52563	0,25089	0,22377	0,22182	**0,22087**
RB_U5k_I20k	0,67744	0,58782	0,30791	0,28466	0,27193	**0,27155**
RB_U30k_I110k	0,70296	0,60644	0,34876	0,33157	0,31070	**0,30889**
A_U200_I120	1,53480	1,56583	1,49159	1,97755	1,37034	**1,34089**
A_U2k_I1k	1,53155	1,30432	1,27850	1,21357	**1,05542**	1,06147
A_U20k_I12k	1,47107	1,28584	1,23608	1,21267	1,04996	**1,04524**
A_U210k_I120k	1,50721	1,44538	1,32229	1,29709	1,15504	**1,14716**
A_U2M_I1M	1,60510	1,63127	1,49281	1,48153	1,33138	**1,32666**

Table 3.1: Résultats des modèles sur les différentes bases en erreur quadratique moyenne sur les critiques de test.

Le tableau précédent nous permet de tirer quelques conclusions importantes : parmi les modèles de référence, $\phi_2(i)$ est nettement plus performant que les autres. Le modèle ϕ_0 semble trop pauvre. La comparaison entre $\phi_1(u)$ et $\phi_2(i)$ montre que l'avis des utilisateurs est assez uniforme sur un produit donné alors qu'un utilisateur a un avis changeant d'un item à l'autre (ce qui semble assez intuitif) [5].

3. Revues sur des bières, collectées par [ML13b]
4. Revues sur différents types de produits vendus sur le site Amazon, collectées par [Jin+10]
5. Le biais item est tellement performant qu'il est souvent omis dans la littérature –e.g.[ML13a]– pour mettre l'accent sur la factorisation matricielle.

Si le filtrage collaboratif apporte un gain significatif par rapport aux références (comme cela a été montré plusieurs fois dans la littérature), il est intéressant de constater que la prise en compte du texte permet systématiquement de réduire l'erreur par rapport à ces techniques. Sur *RateBeer* , le gain est significatif (entre 2,5 et 7% d'amélioration) mais sur *Amazon* il devient très important (entre 10 et 25% de gain).

Ce travail démontre l'intérêt de la prise en compte du texte dans la construction des profils : le gain est aussi important (voire plus) que lors du passage des systèmes à base de biais par rapport au filtrage collaboratif. La comparaison entre les deux approches textuelles est plus serrée : même si ϕ_{T4} est souvent plus performant, les écarts sont minimes. Par contre, les traitements sont beaucoup moins lourds (notamment sur les grandes bases) car il n'y a plus besoin de calculer le modèle LDA (ni de l'appliquer). Philosophiquement, les deux modèles ne s'attaquent pas exactement à la même tâche : compréhension des aspects ou thématiques d'une part (ϕ_{L4}), analyse des mots utilisés d'autre part (ϕ_{T4}). Néanmoins, les résultats sont très proches sur les données réelles : il apparait donc que les approches envisagées (et/ou la masse des données traitées) font converger la modélisation des mots et la modélisation du sens des mots.

L'analyse générale des performances par rapport à la taille des bases considérées est étonnante : sur *Amazon*, les performances s'améliorent avant de repartir à la baisse sur les grandes bases tandis que sur *RateBeer*, la tendance est totalement à la hausse. Plus il y a de données pour apprendre, plus les modèles sont mauvais ! En réalité, l'explication est simple : nous avons réalisé les plus petites expériences sur les données les plus favorables (items largement commentés et utilisateurs les plus actifs) : plus nous avançons dans les données, plus les prédictions sont délicates, ce qui explique la tendance générale à la baisse des performances.

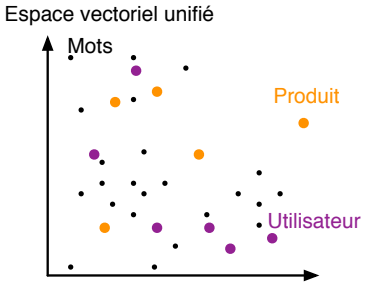


Figure 3.1: Représentation de l'espace latent d'avis en ligne. Une fois les concepts hétérogènes projetés dans le même espace, différentes tâches deviennent traitables.

3.2.2 Recommandation dans un espace latent textuel

Le domaine de l'analyse des données textuelles a été largement impacté par l'algorithme *w*ord2vec [Mik+13] : la similarité entre les représentations de mots a beaucoup gagné en sémantique par rapport aux approches précédentes basées sur les modèles graphiques. La définition de translations constantes dans l'espace latent correspondant à des concepts ou des propriétés grammaticales esquisse même un pont entre sémantique et connaissances.

Notre modèle de recommandation dérive de *paragraph2vec*, une variante de *w*ord2vec [LM14]. Plus exactement, nous utilisons un modèle appelé *Distributed Bag-of-Word* (DBOW), extension du modèle *Skip-Gram* (SG). En effet, en considérant les utilisateurs, les produits et les notes comme des éléments conceptuels caractérisés par un ensemble de mots les représentant il est possible de les introduire dans l'algorithme précédemment cité comme des paragraphes : le but est de trouver la meilleure représentation pour chaque concept, celle qui permette au mieux de prédire ses mots (Figure 3.2).

Une fois les différents éléments projetés dans l'espace latent, nous utilisons une similarité faire de la recommandation au sens des plus proches voisins, mais nous pouvons également utiliser cette similarité pour trouver les mots les plus proches d'un utilisateur, d'un item ou d'un couple utilisateur/item.

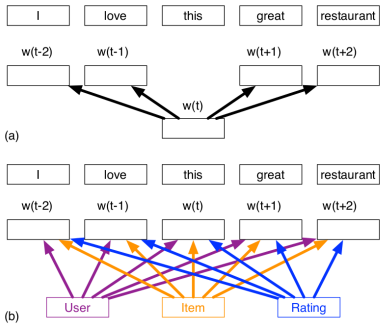


Figure 3.2: (a) Skip-Gram classique sur du texte. (b) Conceptual Skip-Gram qui projette les concepts (utilisateur, produit, note) dans l'espace des mots.

3.2.2.1 Modélisation

Les représentations latentes de chaque mot et de chaque "concept" des avis (utilisateur, produit, note) sont apprises ensemble avec deux objectifs séparés, mais similaires : les mots co-occurrents localement dans les avis (à l'intérieur d'une fenêtre) doivent être proches les uns des autres (et les autres doivent être plus loin) ; chaque concept doit être proche de ses mots (et éloigné de ceux des autres). Dans la pratique, les positions des mots sont apprises en premier, les concepts sont ensuite positionnés par rapport aux mots. Autrement dit, avec $\sigma(x) = \frac{1}{1+e^{-x}}$, les fonctions objectifs sont les suivantes :

$$\arg\max_{v} \left(\log \sigma(v'^{\intercal}_a . v_w) + \sum_{i=1}^{k} \mathbb{E}_{v_b \sim P_n(\mathbf{w})} \log \sigma(-v'^{\intercal}_b . v_w) \right) \tag{3.7}$$

$$\arg\max_{v} \left(\log \sigma(v'^{\intercal}_a . v_c) + \sum_{i=1}^{k} \mathbb{E}_{v_b \sim P_n(\mathbf{w})} \log \sigma(-v'^{\intercal}_b . v_w) \right) \tag{3.8}$$

Où v_w et v_c sont les mots (resp. concepts) et leurs mots associés v_a (cooccurrences ou contextes). Les v_b correspondent aux exemples négatifs, tirés selon une loi unigramme, élevée à la puissance 3/4[6]. Contrairement à [LM14] qui utilise un softmax hiérarchique pour l'apprentissage, nous utilisons seulement le negative sampling comme décrit ci-dessus. Après entraînement, on obtient un espace vectoriel en n dimension où les concepts et les mots similaires sont proches. Le nombre de représentations obtenues est de $|mots| + |concepts|$. À l'instar de word2vec, le cosinus s'utilise comme mesure de similarité, afin de trouver les éléments proches. Par convenance, nous basculons cette mesure sur l'intervalle $[0, 1]$ afin de l'utiliser comme pondération dans (3.10).

$$\alpha_{uv} = \frac{\frac{<u,v>}{\|u\| \times \|v\|} + 1}{2} \in [0, 1] \tag{3.9}$$

En considérant chaque utilisateur, produit et note comme concepts grâce aux mots qui les représentent, nous obtenons un espace où chaque composante des avis en ligne est représentée (Figure 3.1). Il est également possible de projeter individuellement chaque avis et chaque phrase, mais le nombre de représentations deviendrait trop important. Les phrases sont donc projetées dans l'espace latent en moyennant les représentations de leurs mots, comme suggéré par [Kag+14].

L'implémentation repose sur la technique dite du *negative sampling* de [Mik+13]. Cette méthode est une simplification de la *Noise-Contrasive Estimation* [GH10] qui postule qu'un bon modèle doit être capable de différencier les données réelles du bruit grâce à la régression logistique.

Prédiction de notes La mesure de similarité (4.5) rend possible le tri et la sélection des produits, des utilisateurs ou même des phrases et mots voisins à un autre produit, utilisateur ou même mot. C'est à l'aide de cette mesure que nous effectuons nos recommandations. Plus précisément, pour prédire la note $\hat{r_{ui}}$ d'un utilisateur u sur un nouveau produit i on utilise les notes déjà données sur ce même produit i par les autres utilisateurs $v \in \mathcal{N}, r_{uv}$, pondérées par leur similarité à l'utilisateur u, α_{uv}. En pratique, on calcule seulement cette moyenne sur le k plus proche voisinage $\mathcal{N}$. Par

6. La puissance 3/4 permet de lisser la distribution unigramme et de faire ressortir les mots rares [Lev+15]

ailleurs, les notes données par chaque utilisateur sont normalisées par la moyenne de chaque utilisateur μ_u pour prendre en compte leur biais de notation. Symétriquement, la même opération est possible avec les notes déjà données par l'utilisateur u sur des produits similaires au produit i. De ce fait, on obtient deux modèles de prédiction (3.10) avec les α issus de (4.5) :

$$\hat{r_{ui}} = \mu_u + \frac{\sum_{v \in \mathcal{N}} \alpha_{uv}(r_{vi} - \mu_v)}{\sum_{v \in \mathcal{N}} \alpha_{uv}}, \ \hat{r_{ui}} = \mu_i + \frac{\sum_{j \in \mathcal{N}} \alpha_{ij}(r_{uj} - \mu_j)}{\sum_{j \in \mathcal{N}} \alpha_{ij}} \tag{3.10}$$

La prédiction est donc obtenue en moyennant soit :

— les avis du même utilisateur sur des produits proches,

— les avis des autres utilisateurs sur le produit cible.

Prédiction de revues La similarité (4.5) étant basée sur le texte écrit par les utilisateurs, il est également légitime de l'utiliser pour extraire du texte des avis existant afin de le présenter en accompagnement de la recommandation. Deux possibilités seront étudiées :

— présenter l'avis complet le plus proche,

— créer un résumé personnalisé par extraction de phrases selon la méthode de [Kag+14].

3.2.2.2 Expériences et résultats

Nous utilisons notre espace latent pour différentes tâches. D'abord sur l'épreuve classique de prédiction de notes. Puis sur la prédiction des remarques textuelles, tâche originale introduite par [Pou+14a]. Enfin, nous montrons les capacités d'extraction de mots-clés et d'analyse de sentiment de notre espace latent textuel.

Nous sélectionnons différents corpus d'avis en ligne afin d'évaluer nos systèmes de recommandation (tableau 3.2). Pour évaluer la prédiction de notes, nous utilisons l'erreur moyenne au carré (MSE), métrique standard de cette application. Nos résultats sont comparés avec ceux publiés dans [ML13a]. La prédiction de remarques étant une application sans méthode d'évaluation habituelle, nous gardons celle utilisée par [Pou+14a], afin de comparer nos résultats.

Dataset	#Users	#Items	#Reviews
Ratebeer	29073	110283	2844778
BeerAdv.	33368	66032	1585241
Movies	1223167	164446	2985563
Music	1133669	420466	3229487
Yelp	366402	60784	1566139

Table 3.2: Détails des corpus d'évaluation

Prétraitement des avis et paramètres d'apprentissage Afin d'obtenir l'ensemble d'apprentissage, il convient d'effectuer différents prétraitements sur les avis extraits. Les mots apparaissant moins de 10 000 fois sont supprimés, de ce fait chaque base d'avis contient entre 1500 et 3000 mots distincts. Les mots restants sont de véritables pivots du langage qui vont jouer le rôle des aspects latents des techniques de factorisation matricielle. Les avis sont enfin subdivisés en phrases reliées à leurs concepts à l'aide des points restants. Un extrait de phrases après prétraitement est présenté figure 3.3.

r_1 drinkable but not great
r_2 Not much else going on here
r_3 Spicy lactic dry taste
r_4 Taste quite bitter super clove esters big yeast
r_5 The best beer in the world

Figure 3.3: Phrases après prétraitement (associées à leurs notes) extraites du corpus Ratebeer.

Le nombre k de voisins est choisi grâce à une phase de validation. Selon les jeux de données, le nombre de voisins permettant de minimiser la MSE varie de 9 à 33.

Prédiction de Notes Pour prédire une note $\hat{r_{ui}}$ associée à un couple (utilisateur u, produit i), nous utilisons la méthode de filtrage collaboratif par voisinage, décrite par [DK11]. Deux techniques classiques ont été évaluées : La prédiction grâce au voisinage utilisateur et celle grâce au voisinage produit (3.10). Il s'agit d'agréger les

notes des utilisateurs similaires ou les notes précédemment mises sur des produits proches.

Dataset	Moyenne	Fact. Mat	HFT*	CSG-kNN util.	CSG-kNN prod.	k*
Ratebeer	0.701	0.306	0.301	0.336	**0.286**	23
Beeradv.	0.521	0.371	0.366	0.382	**0.357**	29
Movies	1.678	**1.118**	1.119	1.39	1.304	33
Music	1.261	**0.957**	0.969	-	1.201	26
Yelp	1.890	1.49	-	1.591	**1.407**	27

Table 3.3: Précision de la prédiction de note en MSE. La référence HFT* (meilleur modèle "Hidden Factor & Topic model") est reporté de [ML13a]. CSG-$k$nn : Meilleur k-voisins dans l'espace d'avis utilisant k* voisins

.

Les résultats obtenus en prédiction des notes sont présentés dans le tableau 3.3. Les meilleurs résultats obtenus à l'aide de nos modèles de plus proches voisins (CSG-$k$nn utilisateur ou produit) sont comparés à un modèle qui prédit systématiquement la note moyenne du corpus d'apprentissage, un modèle de factorisation matriciel classique et aux modèles HFT présentés par [ML13a]. Les résultats de HFT sur le corpus Yelp ne sont pas reportés, car le corpus utilisé est différent.

En règle générale nos performances sont proches de celles de [ML13a] et même parfois meilleures. Notre modèle est particulièrement performant sur les corpus des sites d'amateurs de bières. Le vocabulaire utilisé y est spécifique et les revues particulièrement riches comparées aux revues type Amazon. Il est important de comprendre que contrairement aux autres modèles, l'espace latent n'est pas optimisé sur un critère de prédiction de note. Obtenir des résultats proches de l'état de l'art, voire meilleurs, souligne la pertinence de ces profils textuels pour la recommandation.

Sur Movies et Music, à l'inverse, le texte agit comme un bruit par rapport à la factorisation matricielle simple : les résultats du modèle HFT étaient légèrement impactés, mais notre plus grande dépendance au texte nous expose plus durement au phénomène.

3.2.3 Démarrage à froid – *cold start*

Cette section décrit un travail dans la continuité du précédent pour souligner l'intérêt de la recommandation dans un espace latent textuel. Mais le démarrage à froid est l'une des problématiques les plus difficiles en filtrage collaboratif et nous séparons donc la contribution qui lui est propre. Historiquement, les approches basées sur le contenu (construction d'une topologie d'item basée sur les descriptions proches) permettent de traiter ce cas de figure. Dans le cas où les données sont manquantes, nous utiliserons les biais pour estimer les notes (moyenne des notes de l'utilisateur face à un nouvel item, moyenne des notes de l'item

Nous avons voulu vérifier si les données textuelles, potentiellement préexistantes –description de l'item, page web personnelle, profil de réseau social, requête sur le nom de la personne– pouvaient suffire pour construire des profils pertinents [Dia+17b]. Nous avons simulé ces données (non présentes dans les jeux publics) en utilisant les textes des revues sans les notes. Afin de ne pas trop biaiser les résultats, nous sommes partis sur une nouvelle modélisation du texte, sans classification de sentiments.

Modélisation La prédiction de notes est effectuée en utilisant les plus proches voisins dans l'espace latent. Comme le montre la figure 3.4, les meilleurs résultats sont obtenus en considérant 9 ou 80 voisins (k-NN) selon le cas de figure qui nous intéresse –respectivement l'introduction d'un nouvel item ou d'un nouvel utilisateur–. Tous les détails concernant la formulation et l'optimisation sont disponibles dans [Dia+17b].

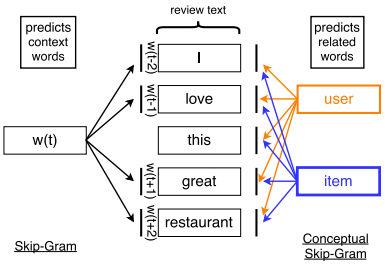


Figure 3.4: Apprentissage des représentations des mots et des profils par rapport à leurs contextes : les co-occurrences rapprochent les représentations.

Données	μ	Nouvel Utilisateur		Nouvel Item	
		μ_i	CSG-kNN	μ_u	CSG-kNN
Ratebeer	0.701	0.341	**0.333**	0.599	**0.371**
Beeradvocate	0.518	0.397	**0.386**	0.490	**0.419**

Table 3.4: Erreur en prédiction de notes (moindres carrés). μ, μ_i, μ_u : moyenne générale, moyenne item,utilisateur : prédicteurs à base de biais.

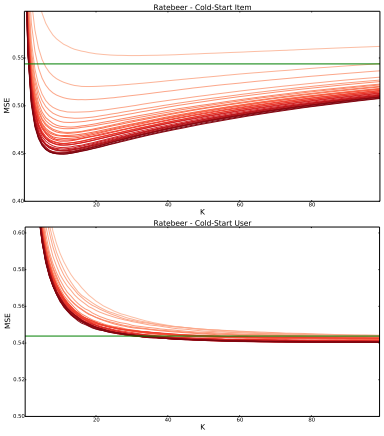


Figure 3.5: Item (haut) et Utilisateur (bas) intégrés dans un scénario de démarrage à froid sur le jeu de données *ratebeer*. Le niveau de coloration donne le nombre de revues (sans note) utilisées pour construire les profils initiaux (de une à 20 revue, plus de revues donnant toujours de meilleurs résutlats). La ligne verte correspond respectivement aux biais utilisateurs et items.
Abscisse : taille du voisinage $\mathcal{K}$,
Ordonnée : erreur au sens des moindres carrés en moyenne.

Données, expériences et résultats Les résultats présentés en figure 3.5 et tableau 3.4 sont éloquents à plusieurs titres : les courbes montrent l'intérêt de prendre en compte le plus de textes possibles. Les résultats s'améliorent significativement en prenant jusqu'à 20 revues pour établir les profils ; ensuite, les performances sont assez stables. L'ampleur du gain est variable d'un scénario à l'autre. L'insertion d'un nouvel item bénéficie largement de notre stratégie (entre 14.5% et 38%). Le gain est moins important pour l'insertion d'un nouvel utilisateur (autour de 2.5%).

Sur les items, notre système est une manière élégante d'hybrider un algorithme de filtrage collaboratif avec des gains que nous pouvions anticiper. Sur les utilisateurs, l'enjeux est plus fort : il s'agit de commencer à évaluer les possibilité de transfert de profils (qui seraient appris sur des sources externes comme les pages de réseaux sociaux ou les traces web de l'utilisateur). Concernant l'ampleur du gain, nous pouvons miser sur la richesse de ces sources externes pour améliorer la prédiction. De plus, il faut bien voir que notre modèle est mis en concurrence avec le biais item, qui constitue une référence particulièrement performante (colonne ϕ_2 dans le tableau 3.1).

3.3 Explications sur les recommandations : propositions & évaluation

Ce paragraphe présente une approche basée sur le modèle précédent. D'autres propositions seront faites dans ce chapitre et le suivant avec des approches de plus en plus ambitieuses. Nous avons choisi de présenter les contributions sur la construction d'explication chronologiquement et de séparer les explications extractives –sélection de phrases existantes–, présentées dans ce chapitre, des explications génératives, présentées dans le prochain chapitre.

L'exploitation des données textuelles représente donc un enjeu pour l'amélioration des suggestions ; il s'agit aussi d'un moyen de faire face à la problématique du démarrage à froid. Nous explorons maintenant une troisième valorisation des données textuelles dans le cadre de la recommandation : l'explication des suggestions. La définition d'une explication n'est pas arrêtée dans le domaine de la recommandation :

— il peut s'agir d'un modèle explicable, au sens Bayesien, en applicant une pondération sur des descripteurs explicites [Paz99] ou au sens des k-plus proches voisins en insistant sur l'origine de la suggestion ;

— il peut s'agir d'aspects, donnés [Gan+09] ou latent (et appris) [ML13a] qui sont pondérés dans les profils. Dans le premier cas, le système est interprétable, dans le second, il reste très *boite noire* ;

— il peut s'agir d'expliquer a posteriori un profil à travers les données textuelles, ce qui nous intéresse à présent. La tâche reste floue puisqu'il peut s'agir de trouver des mots clés, des phrases ou des textes à extraire ou à générer. En suivant cette philosophie, nous avons exploré plusieurs pistes : (1) greffer du texte sur la factorisation matricielle et retrouver les mots associés à des profils [Pou+14d], (2) basculer la recommandation dans un espace textuel [Dia+16], (3) mêler représentations des mots, des phrases et des utilisateurs dans une architecture de deep-learning [Dia+18b].

Concernant spécifiquement la tâche d'explication des suggestions, sa définition n'est pas précise et son évaluation ardue. Nous avons opté pour le cadre évaluatif du résumé automatique, en partant de l'hypothèse que la meilleure explication possible correspondait à la revue écrite (a posteriori) par l'utilisateur. Dans les contributions [Pou+15a ; Dia+16], nous avons utilisée la métrique ROUGE, calculant grossièrement le rappel sur les mots utilisés dans la revue à la manière de [Li+17b]. Les autres articles récents sur le sujet se basent soit exclusivement sur du qualitatif [Lip+15], sur la perplexité [Bar+16] ou sur la mesure BLEU [Don+17].

Modélisations Nous avons proposé plusieurs solutions pour expliquer les recommandations. Dans un premier temps [Pou+15a], nous avons chercher à extraire les mots, les phrases et les revues les plus proches de l'utilisateur parmi les revues d'un item cible. Le matching était effectué sur à la fois sur la correspondance entre la note prédite et la note de la revue et sur la correspondance entre les mots utilisés.
Qualitativement, les meilleurs résultats étaient obtenus en fournissant la revue la plus pertinente de l'historique. Quantitativement (au sens de la mesure ROUGE-N), il était plus intéressant de construire une revue par extraction de phrase en visant la longueur moyenne des revues de l'utilisateur cible.

Dans [Dia+16], Fig. 3.4, tous les éléments du problème sont projetés dans un espace latent textuel : il suffit alors d'exploiter la métrique *cosine* de word2vec pour récupérer les éléments (mots, phrases, revues) liés à un item cible et proches d'un utilisateur donné.

L'avantage des propositions précédentes est de coller de très près à un cadre algorithmique existant (respectivement la factorisation matricielle et word2vec). Notre proposition la plus ambitieuse sur l'explication des suggestions propose un cadre plus novateur mêlant modélisation multi-échelles des documents, intégration de profils item-utilisateur et modèle d'attention [Dia+18a ; Dia+19c]. L'attention permet de sélectionner les mots et les phrases qui expliquent la décision ; comme le montre la figure 3.6, cette attention correspond dans notre architecture à des profils utilisateurs et produits. La plasticité des architectures profondes nous a permis de mêler différentes fonctions de coût pour gagner en performances sur les tâches de recommandation et de classification de sentiments : notre modèle devient une sorte de lecteur personnalisé des avis produits, il extrait les passages d'intérêt pour l'utilisateur produisant ainsi une explication.

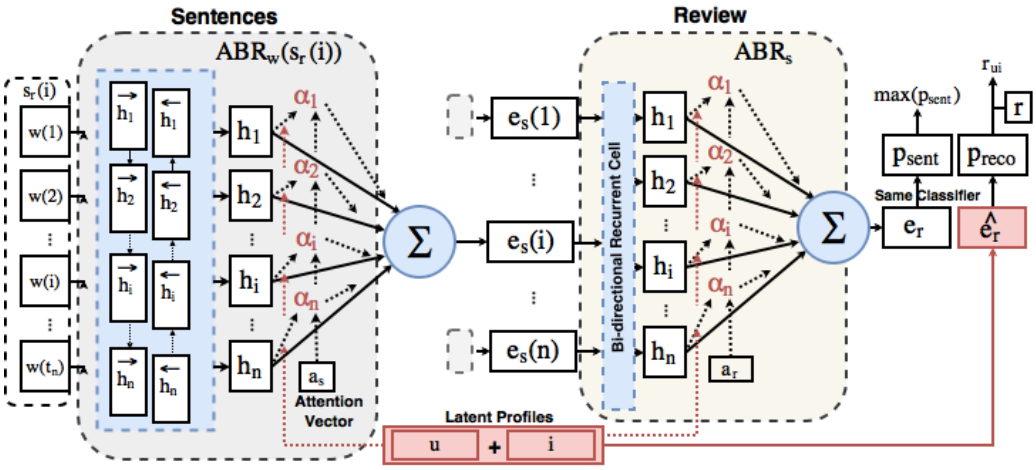


Figure 3.6: Architecture hiérarchique d'explication des suggestions à partir des revues : l'attention est liée aux utilisateurs et produits, elle permet de sélectionner les mots et les phrases qui intéresseront l'utilisateur.

Dataset	Rouge-1				Rouge-2			
	Rand	Prod.	Util.	Oracle	Rand	Prod.	Util.	Oracle
avis complet								
Ratebeer	0.243	0.271	0.351	0.568	0.037	0.052	0.117	0.278
Beeradv.	0.280	0.304	0.366	0.507	0.045	0.063	0.114	0.207
Music	0.255	0.407	0.403	0.545	0.043	0.212	0.215	0.255
Yelp	0.249	0.290	0.284	0.512	0.037	0.067	0.072	0.154
n phrases								
Ratebeer	0.157	0.239	0.217	0.591	0.011	0.043	0.046	0.288
Beeradv	0.122	0.225	0.244	0.606	0.010	0.046	0.057	0.249
Music	0.167	0.291	0.220	0.551	0.017	0.026	0.041	0.282
Yelp	0.150	0.246	0.279	0.550	0.011	0.041	0.050	0.208

Table 3.5: Évaluation de la prédiction de remarques au sens de la métrique ROUGE. *Prod.* : sélection des avis/phrases dans les données produit (les avis des autres utilisateurs). *Util.* : sélection dans les données de l'utilisateur (ses propres avis sur des produits similaires)

Performances Les performances sont très difficiles à évaluer : ni la supervision, ni les métriques existantes ne sont vraiment satisfaisantes. Le fait que la vérité terrain soit un avis a posteriori de l'utilisateur est biaisé : de nombreux avis font simplement état de sentiments sans description de l'item. La note de l'utilisateur ayant déjà fait l'objet d'une évaluation, le texte ne présente alors aucun intérêt... Néanmoins, en l'absence d'alternative, nous nous évaluons par rapport à ces étiquettes. Le choix de la métrique a déjà été abordé en introduction de la section. Aucune métrique n'est pleinement satisfaisante, les résultats doivent être considérés avec du recul, comme une manière de comparer différentes approches et sans signification absolue. De ce point de vue, certaines bases de données –Yelp sur les restaurants, BeerAdvocate sur les bières– se prêtes beaucoup mieux à l'analyse que les bases classiques de produits électroniques Amazon, où les revues sont plus souvent vides de description.

Predicted rating : 4.70
Extracted personalized summary : The staff is extremely friendly. On top of being extremely large portions it was incredibly affordable. Most of girls are good, one is very slow, one is amazing. The fish was very good but the Reuben was to die for. Both dishes were massive and could very easily be shared between two people.

Figure 3.7: Exemple de suggestion enrichie : prédiction personnalisée de la note, des mots clés voire de la revue associée à un item –ici, un restaurant de la base Yelp–.

Le tableau 3.5 illustre les performances obtenues dans [Dia+16] [7]. Les conclusions sont difficiles à tirer : la sélection de revue fonctionne mieux que l'extraction de phrases, au sens de la métrique ROUGE, mais pas du point de vue qualitatif. La faiblesse de la métrique est aussi mise en évidence lors de la comparaison des colonnes *Prod.* et *Util.* : il est plus efficace, quantitativement, d'extraire des phrases ou des revues parmi les contributions passées de l'utilisateur... Pourtant, cela n'a aucun sens dans le cadre de l'explication d'une suggestion sur un item cible différent. Quantitativement toujours, nous voyons que la marge de progression par rapport au modèle aléatoire est plus importante sur l'extraction de phrases que lors de la sélection de revue –l'oracle est également supérieur dans cette configuration–.

Si les performances numériques sont relatives, les revues générées présentent un réel intérêt qualitatif, notamment avec notre dernière proposition [Dia+18a] : le fait d'extraire des phrases parmi les revues présentant des notes proches de la note prédite apporte très souvent des informations pertinentes sur les aspects du produits recherché par l'utilisateur.

3.4 Modélisation de la dynamique en recommandation

En recommandation, un premier palier de performance a été franchi avec les algorithmes de filtrage collaboratif –notamment l'algorithme de factorisation matricielle positive–. Une grande partie de la communauté scientifique est persuadée que le prochain palier sera lié à la modélisation efficace des données de contexte de l'utilisateur : les données textuelles (cf section précédente) et les données temporelles, que nous allons aborder maintenant, sont les plus prometteuses pour apporter de nouvelles informations.

Définitions de la dynamique La modélisation du temps peut prendre différentes formes :

— en ajoutant des paramètres de temps sur les items, après la factorisation des profils, il est possible de prendre en compte les effets de mode et les tendances [Kor10].

— Des études ont mis en évidence l'aspect périodique du comportement des utilisateurs –y compris sur des tâches de notations de produits–. Dans [Cam+11], ils proposent une modélisation Bayesienne de l'individu pour prendre en compte ces biais temporels de notations.

— Plusieurs approches se focalisent sur la notion de séquence plus que de temps en modélisant les transitions entre items visités. [Che+12b ; Bar+13b] se focalisent respectivement sur les enchainements de morceaux musicaux et les déplacements de touristes, montrant dans les deux cas l'intérêt de cette modélisation souple, légère et robuste. Les travaux plus récents ont adopté le paradigme des enchainements locaux de word2vec et ont montré qu'il s'adaptait très bien aux navigations des utilisateurs dans des catalogues d'items [Grb+15]. Dans le cadre plus précis de la prédiction du prochain item visité (en recommandation), un modèle reconnu est décrit dans [TT12].

7. Nous avons arbitrairement pris les résultats de cet article : ces tableaux de résultats sont assez proches d'une architecture à l'autre et toujours très durs à exploiter.

— La logique de mémoire voudrait que nous accordions plus de poids aux actions récentes qu'aux actions plus lointaines. Dans cette idée, [DL05] introduisent une fonction d'oubli dans les algorithmes de filtrage collaboratif (pondération des voisinages).

— Plutôt qu'une mémoire, [ML13b] proposent de modéliser des niveaux d'expériences pour les utilisateurs, ce qui leur permet de revenir dans le cadre plus efficace de la factorisation matricielle. La ligne de temps de chaque individu comporte alors des ruptures, à déterminer, lorsqu'il passe d'un niveau à l'autre.

Contribution Nous avons cherché à construire des profils évoluant avec le temps. Tout d'abord, en repartant des modèles latents *entrée/sortie*, où chaque item est lié à deux profils permettant de définir une trajectoire. Les utilisateurs entrent par le premier profil et sorte par le second : il devient donc possible de prédire où un utilisateur risque de se retrouver dans le futur, mais indistinctement pour l'ensemble de la population [GS+14]. Pour personnaliser cette approche, nous avons opté pour un cadre très léger [GS+15] : les items ont des représentations uniques et fixes – l'espace étant appris par un algorithme de type word2vec– et les parcours deviennent personnels à partir du moment où l'utilisateur est représenté comme une translation permettant de prédire le prochain item visité.

Nos contributions s'apparentent à des variantes de l'algorithme prod2vec [Grb+15], qui a justement été proposé la même année et à la même conférence. Là où prod2vec se concentre sur les enchainements de produits dans une trace –à la manière de l'enchainement des mots dans une phrase– indépendamment des utilisateurs, nous avons chercher à introduire de la personnalisation.

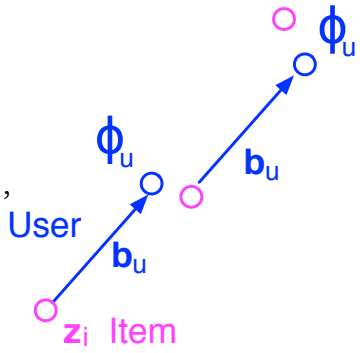


Figure 3.8: Modélisation d'une trace utilisateur dans les représentations de produits.

Modèles Soit les représentations d'items $\mathbf{z} \in \mathbb{R}^d$ et une trace d'un utilisateur u $\{\mathbf{z}^0, \ldots, \mathbf{z}^t\}$, nous avons d'abord envisagé d'estimer le prochain item visité comme :

$$\tilde{\mathbf{z}}^{t+1} = \mathbf{z}^t + \mathbf{b}_u, \qquad \mathbf{b}_u \in \mathbb{R}^d \tag{3.11}$$

en optimisant b_u pour minimiser $\|\tilde{\mathbf{z}}^{t+1} - \mathbf{z}^{t+1}\|$ sur l'ensemble des traces d'apprentissage. Dans une variante plus récente [GS+16], nous avons ajouté une déformation linéaire de l'espace A pour aboutir au modèle :

$$\tilde{\mathbf{z}}^{t+1} = A\mathbf{z}^t + \mathbf{b}_u, \qquad \mathbf{A} \in \mathbb{R}^{d \times d} \tag{3.12}$$

Etant donné le nombre de paramètres en jeu, il est impossible d'apprendre A au niveau individuel ; nous avons donc opté pour une version communautaire : les transformations A_c sont apprises pour des groupes d'individus homogènes.

Pour la partie recommandation, les profils utilisateurs et items précédants sont constants mais concaténés avec de nouveaux vecteurs $\bar{\gamma_u}, \bar{\gamma_i}$ pour offrir des degrés de libertés à la NMF :

$$\gamma_u = [\bar{\gamma_u}, \mathbf{b}_u], \forall u \in U, \qquad \gamma_i = [\mathbf{z}_i, \bar{\gamma_i}], \forall i \in I \tag{3.13}$$

Performances Nous avons évalué nos systèmes sur deux critères : le rappel@k pour la prédiction du prochain item visité et la MSE pour la tâche de recommandation. Dans la première tâche, nous nous sommes comparés à différentes méthodes de l'état de l'art en faisant varier k. Les résultats sont présentés en Figure 3.9 : ils montrent une nette avance de nos stratégies par rapport aux approches existantes.

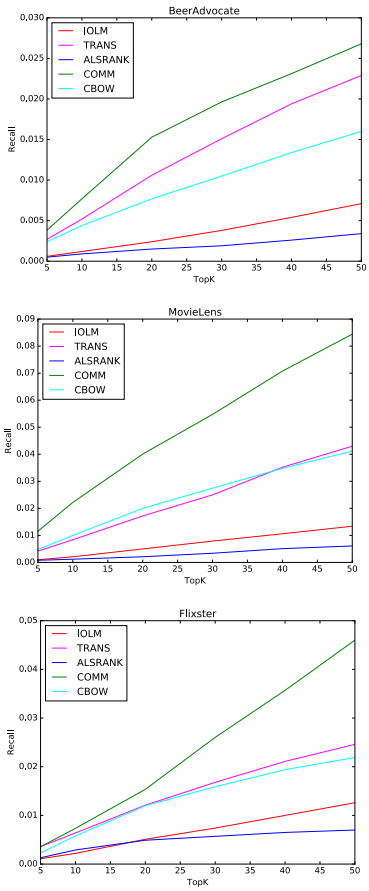


Figure 3.9: Performance dans la prédiction du prochain item visité, au sens du rappel@k, en fonction de k. Les différents modèles testés correspondent à l'état de l'art : IOLM [Che+12b], ALSRank [TT12], CBOW (=word2vec) [Mik+13]. Trans et Comm correspondent à nos deux modèles, respectivement en translation simple ou avec un endomorphisme appris par communauté.

Données \ Modèles	MF	TSVD	EXP	TRANS	COMM
BeerAdvocate	0.4	0.381	0.367	**0.361**	**0.360**
RateBeer	0.331	0.301	0.297	**0.279**	**0.279**
MovieLens	0.691	0.681	0.684	0.663	**0.660**
Flixster	0.912	0.867	0.827	0.816	**0.811**
Movies	1.377	1.211	1.05	**0.913**	**0.913**

Table 3.6: Erreur au sens des moindres carrés (MSE) sur différentes bases de données classiques de recommandation pour différents algorithmes de l'état de l'art. MF [Kor+09], TSVD [Kor10], EXP [ML13b]

Comme le montre les résultats, l'algorithme word2vec fonctionne particulièrement bien sur cette tâche. Le fait d'ajouter un modèle explicite de prédiction sur cet espace intrinsèquement pertinent nous permet d'obtenir de très bonnes performances. L'enjeu est ensuite de trouver le bon compromis entre simplicité –le modèle à translation– et expressivité –le modèle à endomorphisme, mais appris sur les communautés–.

La question suivante implique des approches différentes : il s'agit de vérifier si les paramètres appris constituent une base intéressante pour la problématique de la prédiction de notes. L'hypothèse sous-jacente est la suivante : si la prédiction séquentielle est bonne, c'est que les paramètres utilisateurs et items (leurs vecteurs de translation et les positions) offrent une certaine régularité. Cette régularité doit apporter une information a priori pour établir des profils pertinents dans le cadre de la prédiction de notes, même si les tâches sont distinctes.

Le Tableau 3.6 permet de comparer les performances en prédiction de notes de différentes techniques de l'état de l'art. La modélisation du temps apporte systématiquement des gains importants par rapport à la factorisation matricielle (MF). Nous montrons que notre approche est plus efficace que la modélisation des tendances –TSVD, [Kor10]– et que la prise en compte de niveaux d'expériences chez les utilisateurs –EXP, [ML13b]–. Par contre, comme nous nous y attendions, le modèle plus complexe –COMM– est équivalent au modèle TRANS sur cette seconde tâche : la modélisation des communautés n'apporte pas d'information pertinente pour la modélisation des individus.

3.5 Apprentissage de profils textuels intégrant de la dynamique

Après nos travaux portant sur l'exploitation du texte pour l'amélioration des profils et nos travaux sur la dynamique dans la recommandation, nous nous sommes naturellement intéressés au mélange de ces facteurs dans le cadre d'un projet centré sur l'analyse d'un large corpus de CV. Nous souhaitons comprendre les évolutions de carrières et proposer un outil de recommandation pertinent. Opter pour une modélisation de carrière en forme de séquence d'événements ouvre plus de finesse dans l'apprentissage des profils [Sav05].

L'enjeu est donc double : apprendre des représentations pertinentes des formations et des métiers sur une population, à la manière du filtrage collaboratif ; puis proposer une modélisation individualisée des évolutions de carrière pour dépasser les simples statistiques d'enchaînements de postes présents dans la base de données. En plus de ces aspects, nous travaillons ici sur des données très bruitées au niveau syntaxique et sémantique : les mots comportent beaucoup de fautes et les intitulés de postes sont quasi-uniques, chaque personne utilisant une description spécifique.

Cette section reprend les travaux publiés dans l'article [Gab+20].

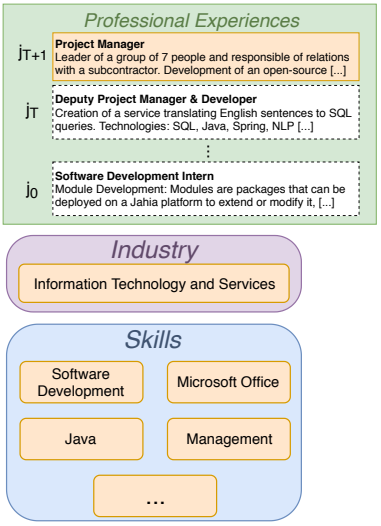


Figure 3.10: : Profil Linked-in composé d'expériences professionnelles, de secteur d'activité et de compétences.

Problématiques & modélisation Les données, issues de Linked-in, sont assez riches comme le montre la figure 3.10. A partir des intitulés de poste occupés précédemment, nous cherchons à prédire le secteur d'activité et les compétences mais aussi le poste actuel de l'utilisateur : c'est cette dernière tâche qui mêle analyse du texte et de la dynamique. Les questions qui se posent sont multiples : (1) pour la modélisation du texte, est-il possible de bénéficier de modèles de langue pré-entrainés ou la nature des données est-elle trop spécifiques – abréviation, phrases incomplètes, ...–? (2) comme dans la section sur l'explication des suggestions, l'enjeu de l'évaluation des performances est loi d'être trivial. A partir du moment où il n'est pas possible de catégoriser les intitulés de postes ou de se focaliser sur les intitulés les plus fréquents, il faut générer une prédiction textuelle puis l'évaluer. Nous avons retenu la métrique BLEU qui évalue la précision sur les mots prédits (BLEU-1) puis sur les n-grammes (BLEU-N) : l'enjeu est ainsi de prédire un maximum de n-grammes pertinents par rapport à l'intitulé.

Nous avons proposé une modélisation simple où chaque intitulé est encodé par :

— un modèle de langue de l'état de l'art pré-entrainé sur de vastes corpus textuels, ELMO [Pet+18].

— FT_{pt}, un modèle FastText pré-entrainé, reconnu pour sa robustesse notamment aux fautes d'orthographe [Boj+17],

— FT_{CV}, un modèle FastText entrainé sur les données de CV.

Les intitulés sont agrégés par une simple moyenne [8] et décodés par un LSTM pour la tâche de prédiction d'intitulé.

Les entrées sont donc des textes ordonnés chronologiquement pour chaque utilisateur : $\mathcal{J}_u = \{j_0, \ldots, j_T\}$. Chaque intitulé est encodé dans un vecteur $\mathbf{z}_t$; les intitulés sont agrégés en $\mathbf{z}_u$. Malgré un grand nombre de test au niveau de la fonction d'agrégation, les meilleurs résultats ont toujours été obtenus avec une simple moyenne sur les $\mathbf{z}_t$: c'est donc cette stratégie qui a été retenue pour ce travail. Le décodeur, de type LSTM au niveau des mots, produit un j_{T+1} à partir de $\mathbf{z}_u$. Nous prédisons également les compétences de l'utilisateur encodées en *one-hot* dans un vecteur $\mathbf{s}_u \in \{0,1\}^S$ et le secteur d'activité $b_u \in \{1, \ldots, B\}$ parmi une liste établie.

Résultats et analyses Les résultats sont donc de deux natures : multi-classes pour les compétences et les secteurs industriels –respectivement multi et mono valeur– et comparaison de phrases pour les intitulés. Le tableau 3.7 montre les résultats obtenus en classification, ils sont comparés aux compétences et secteurs les plus fréquents dans la base. Nous procédons de la même manière sur la prédiction d'intitulés en nous comparant à une référence absurde du point de vue qualitatif mais efficace sur

8. Aucune approche plus fine n'a donné de meilleurs résultats (Moyenne pondérée, RNN, etc...).

Model	BLEU score (Last Job)				
	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Most Common	0.00	**33.3**	0.3	0.0	0.0
FT_{pt} (pre-trained FastText)	1.91	20.6	3.5	0.8	0.2
FT_{CV} (CV-oriented FastText)	**2.15**	22.2	**3.8**	**0.9**	**0.3**
ELMo	1.74	22.5	**3.8**	0.6	0.2

Table 3.8: Expériences en génération de texte évaluées en BLEU.

la métrique retenue : la prédiction systématique des mots les plus fréquents (e.g. *consultant*).

Dans les deux séries d'expériences, le modèle le plus robuste est un FastText qui n'exploite aucune donnée externe ni aucun pré-entrainement. Ce résultat est contre-intuitif vis à vis du nombre d'applications qui bénéficient justement du pré-entrainement massif des modèles de langues ; néanmoins, il met en relief la spécificité de l'écriture de CV qui ne relève justement pas de la langue courante. Pour bien mesurer le coté surprenant de ce résultat, il faut aussi prendre en compte le fait que ELMO est un modèle naturellement génératif alors que la représentation FastText requière en parallèle l'apprentissage d'un décodeur de type LSTM au niveau des mots pour la prédiction du prochain poste occupé.

Dans l'absolu, la génération de texte est une tâche difficile et notamment difficile à évaluer. Nous avons opté pour la métrique BLEU qui met en avant la précision, c'est à dire l'existence des phrases prédites dans la vérité terrain (cf Tab. 3.8). La métrique n'est pas parfaite, comme le montre la très haute performance du modèles donnant toujours les mots les plus fréquents en BLEU-1. Nos approches permettent toutefois de conserver plus de pertinences dans la prédiction concrète des intitulés qui sont majoritairement composés de plusieurs mots.

La dynamique était présente à deux niveau dans le problème : au niveau de la phrase et des enchaînements de mots d'une part et au niveau de la succession des postes occupés. Bien que cette problématique soit centrale, les meilleurs architectures par rapport aux métriques proposées ne tiennent compte d'aucune forme de dynamique. Cette situation s'explique par le fort niveau de bruit dans les données : les mots comportent de nombreuses fautes d'orthographes, les phrases sont souvent remplies d'omission, dans un style sténographique... Et les intitulés de postes eux-mêmes sont assez bruités, certaines expériences fusionnant différents postes occupés. La question qui se pose est très classique en machine learning : est-il possible d'apprendre un modèle complexe traduisant des interactions très fines dans un contexte très bruité ?

Pour passer un palier de performances, il est maintenant nécessaire de mettre en place des architectures génératives de bout en bout afin de modéliser à la fois la langue des CV et leur dynamique : c'est une piste que nous envisageons dans un travail présenté dans le chapitre suivant, en section 4.1.2.

Modèle	Comp.	Secteur
Most Com.	24.0%	6.3%
FT_{pt}	40.9%	35.6%
FT_{CV}	**42.4%**	**38.4%**
ELMo	39.0%	30.7%

Table 3.7: Résultats en classification multi-classes. Les compétences sont évaluées en f1, les secteurs d'activité en taux de bonne classification.

3.6 Autres domaines applicatifs & évolutions algorithmiques

Les algorithmes d'apprentissage de représentation, qui deviennent naturellement des algorithmes de profiling dans le cadre de la modélisation d'individus ont permis de passer des paliers de performances dans divers domaines applicatifs. Nous avons mis en évidence dans les sections précédentes l'intérêt de ces approches pour la représentation de données hétérogènes permettant de modéliser conjointement les individus et leur contexte textuel ou temporel.

D'autres domaines applicatifs bénéficient des progrès en apprentissage de représentations. Par exemple, la principale problématique associée aux *smart homes*, *smart cities* ou *smart factories* consiste à dresser des profils riches d'individus ou de situations à partir de multiples capteurs et sources d'informations, souvent parcellaires ou bruitées. De ce point de vue, les techniques explorées précédemment sont directement pertinentes.

3.6.1 Smart-city & transports en commun

La thématique *smart-city* regroupe différents domaines de recherches très éloignés, notamment : internet des objets, amélioration des systèmes de surveillance & sécurité, profiling des usagers des services public. Nous nous sommes intéressés en particulier au domaine des transports en commun, qui présente quelques spécificités. Dans le cadre d'un partenariat avec le STIF (Syndicat des Transports en Ile de France), nous avons travaillé sur la modélisation des usagers sur la région francilienne.

Les signaux de logs des usagers présentent plusieurs caractéristiques rendant leur analyse ardue :

— ils sont globalement très déséquilibrés en puissance : la plupart des enregistrements correspondent à des trajets domicile/travail. Même en filtrant ces signaux majoritaires, les enregistrements restent dominés par des variations toujours attachées aux trajets domicile/travail.

— Les signaux faibles sont donc particulièrement difficiles à extraire, et a fortiori, à caractériser.

— Les données présentent différents bruits (calendaire –jours fériés, ...–, problème réseau ou panne, absence de certains logs (défaut de la machine de validation ou problème de collecte des informations). Le niveau total de bruit est très élevé et nécessite des approches robustes.

— La privacy et le profiling sont souvent opposés et rendent certaines analyses analyses inacceptables quand bien même l'application serait intéressante pour l'usager. Par exemple, il est impossible (et non souhaitable) d'agréger plusieurs sources de données personnalisées autour des transports en commun, même l'ajout des transactions bancaires de l'usager dans son profil permettaient d'affiner son profil, de découvrir et d'anticiper ses motivations de transport et de lui fournir un service d'information de plus grande qualité [Lou+17].

Problématiques successives D'un point de vue général, nous avons cherché à apprendre des profils d'usagers et de stations en introduisant le minimum de connaissances métiers. En tenant compte de ces contraintes, nos travaux se sont articulés successivement autour de plusieurs axes.

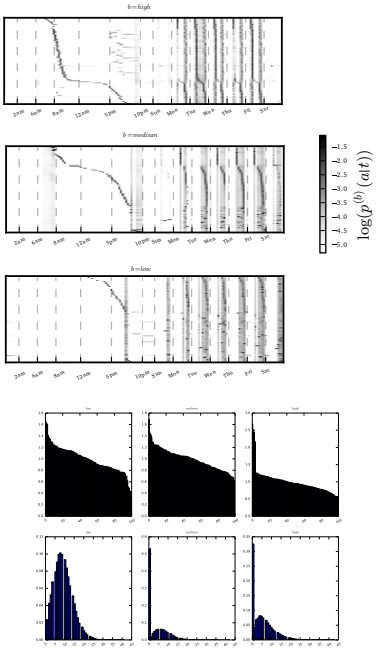


Figure 3.11: [haut] Atomes du dictionnaire localisés dans la journée et la semaine. [bas] Usage moyen des 100 atomes sur la base d'usager en haut, distribution du nombre d'atomes nécessaire pour représenter un usager en bas.

Tout d'abord, la caractérisation des usagers, en relevant un triple défi : (1) identifier automatiquement une sémantique dans les signaux, à la manière de [Can+08] mais en utilisant des approches agnostiques de machine learning ; (2) comprendre les différences entre usagers dans la réalisation d'une action identifiée ; (3) modéliser les signaux faibles, dans des traces utilisateurs particulièrement bruitées. Ces travaux ont fait l'objet de plusieurs publications [Pou+14c ; Pou+14b ; Pou+16 ; Ton+16 ; Ton+18c]

Nous avons ensuite travaillé sur une analyse réseau des logs. L'idée était non seulement de projeter les comportements sur une carte –ce qui était déjà fait dans les études précédentes–, mais aussi de caractériser en détail les stations pour attaquer la détection d'anomalie [Ton+17 ; Ton+18a].

La question centrale et récurrente autour des données de mobilité en général et des logs STIF en particulier est celle de l'évaluation quantitative des résultats obtenus. Nous avons proposé deux pistes dans nos contributions : la mesure de la capacité à prédire des logs futurs et la supervision via une source externe de données –les fils Twitter de la RATP–.

Modélisations des usagers Nous avons travaillé sur les modèles NMF –Non negative Matrix Factorization–, qui permettent d'exploiter les données de l'ensemble des usagers pour extraire de manière robuste un dictionnaire de comportements. Les profils individuels correspondent ensuite à une pondération de ces comportements types. L'enjeu principal est d'associer une sémantique aux atomes du dictionnaire.

Par rapport à la formulation proposée en section 3.1, seule les notations changent : la matrice à décomposer devient $X \in \mathbb{R}^{N \times T}$ où N est le nombre d'usagers et T le nombre de pas de temps (suite à une discrétisation). Le profil de l'usager u au cours du temps, $x_{u,\cdot} \in \mathbb{R}^T$, est estimé par $u \cdot D, u \in \mathbb{R}^Z, D \in \mathbb{R}^{Z \times T}$, avec Z désignant le nombre d'atomes du dictionnaire (hyper-paramètre). Pour obtenir un résultat satisfaisant, il faut que les Z atomes du dictionnaire soient interprétables, c'est à dire correspondent à une sémantique. Dans cette optique, nous avons exploité la flexibilité d'une optimisation en descente de gradient stochastique pour introduire différentes contraintes comme des atomes mono-modaux dans [Ton+16].

Notre plus grande contribution algorithmique est également décrite dans [Ton+16]. Il s'agit de distinguer la forme caractéristique d'un comportement (e.g. la dispersion temporelle des logs dans le départ au travail d'un individu) et le positionnement moyen du comportement dans la journée. Les atomes composant le dictionnaire D_r voient leur supports temporels réduits à $T_r << T$ et la reconstruction de $\hat{x}_{u,\cdot}$ introduit un facteur de décalage $\phi_{u,z}$ défini pour chaque usager et chaque atome :

$$\hat{x}_{u,\cdot} = \sum_{z \in 1,\dots,Z} \phi_{u,z}(u[z]D_r[z]) \tag{3.14}$$

La dernière étape concernant ces travaux consiste à intégrer la dimension spatiale dans la modélisation. Les logs sont alors stockés dans un cube usager/espace/temps : $x_{u,s,t}$. Pour la reconstruction, nous avons fait le choix de garder un dictionnaire d'atomes unique –pour plus de robustesse– et nous avons simplement localisé les activités pour chaque usager : la matrice de pondération $u[z]$ devient $u[z,s]$. Tous ces travaux sont décrits dans [Ton+18c].

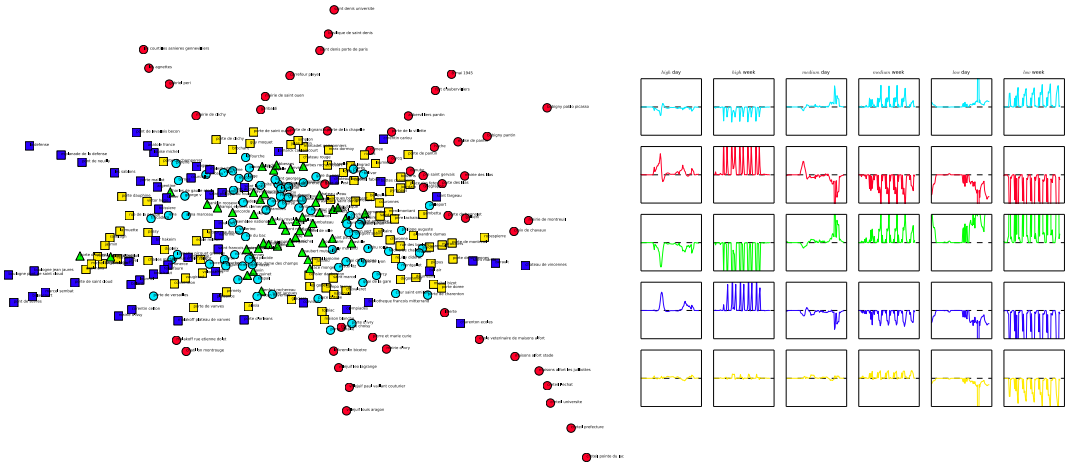


Figure 3.12: Clustering des comportements usagers projetés sur les stations qu'ils utilisent.

Synthèse des résultats obtenus Les premiers résultats obtenus dans [Pou+14c] sont de nature qualitative : nous avons extraits des atomes significatifs pour la reconstruction des usagers en Fig 3.11, gauche. Afin de capter autre chose que les aller-retours au travail, nous avons distingué les activités fréquentes, intermédiaires et rares des usagers. Les activités fréquentes sont décrites par des atomes concentrés sur le matin et le soir, 5 jours par semaine. Les autres activités montrent des atomes moins concentrés et présents également le week-end.

Nos usagers sont maintenant décrits dans un repère plus explicite. Nous sommes en mesure de les regrouper en fonction de leur comportement. L'évaluation quantitative est intrinsèquement problématique ; même l'évaluation qualitative de ces résultats est délicate. Nous avons opté pour une première vérification sur l'usage des différents atomes et de la distribution des atomes dans les profils : le résultat de l'optimisation semble cohérent. Pour juger les profils, nous avons projeté les comportements des usagers sur les stations qu'ils utilisent. Nous avons ensuite effectué un clustering spatial sur cette représentation des stations pour illustrer l'intérêt de notre approche. Le résultat est illustré en Fig 3.12.

Nous avons choisi de ne pas développer ici les résultats obtenus par les modèles suivants [Pou+16 ; Ton+16 ; Ton+18c]. Ces études ont permis d'intégrer de nouveaux éléments descriptifs (temps, espace) et l'exploration des profils montre que les profils sont effectivement plus riches. Cependant, il est très difficile de synthétiser visuellement des résultats.

Détection et catégorisation des anomalies Etudier les profils utilisateurs donne des clés pour mieux comprendre les usagers et anticiper les mouvements. C'est aussi une technique robuste pour caractériser le fonctionnement normal du réseau de transport. Nous avons ainsi mis en concurrence des systèmes de détection d'anomalies basés sur les signaux bruts –écarts entre une journée moyenne et une journée cible– et des systèmes basés sur des représentations latentes.

Nous avons montré l'intérêt de cette dernière approche dans [Ton+17 ; Ton+18a]. Nous avons aussi profité de cette tâche pour proposer une évaluation quantitative de nos résultats : en effet, nous avons collecté une base d'évènements sur Twitter –à partir des comptes de la RATP– qui nous donne une vérité terrain des anomalies sur le réseau métro.

Parmi les approches basées sur l'apprentissage de représentions, nous avons mis en perspective des systèmes basés sur la modélisation des stations et des systèmes modélisant les comportements individuels d'usagers. Nous avons montré que ces différentes approches ne captent pas les mêmes anomalies.

3.6.2 Discussion autour des données médicales

Les données médicales présentent un certain nombre de caractéristiques qui les rapprochent des données de transport. Bien que cette habilitation n'aborde pas directement le traitement de ces données, il est intéressant de dresser un parallèle avec certaines problématiques. Nous n'abordons pas ici les techniques d'imagerie ni les signaux complexes comme les électroencéphalogrammes ou les électrocardiogrammes, nous nous intéressons plutôt aux applications induites par nouveaux capteurs d'activités mesurant par exemple le rythme cardiaque.

Ces nouveaux appareils vont probablement déboucher sur des applications médicales, de diagnostic de certaines maladies ou de détection de certaines phases d'incubation : des applications existent déjà pour diagnostiquer l'apnée du sommeil, à quand un bracelet vous prévenant 48h avant l'apparition des symptômes de la grippe ?

Bruit & incomplétude Les capteurs sont assez loin des spécifications requises dans le milieu médical et présentent un plus fort niveau de bruit. De plus, les utilisateurs peuvent facilement enlever/remettre ces capteurs, engendrant des jeux de données incomplets. Les techniques d'analyse se devront donc d'être particulièrement robustes.

Etude de population Le principal moyen de faire face au bruit et à l'incomplétude des données réside dans l'étude d'une population plutôt que d'individus isolés. Il faut trouver un référentiel commun de description des données –avec une certaine invariance entre personnes– puis identifier les différents motifs caractéristiques des signaux.

Privacy Le passage à une population est incontournable pour les raisons citées précédemment ; mais il est critique du point de vue de la protection des données personnelles. Les enjeux doivent être analysés à deux niveaux : au moment de l'apprentissage, une anonymisation solide des données implique une transformation qui risque d'hypothéquer les résultats. Au moment de l'inférence, il est par contre assez aisé de construire un système personnel embarqué qui ne partage pas les informations personnelles.

Anomalie Deux grandes familles de problème se distinguent : les problèmes supervisés, où nous cherchons à classer une situation par rapport à une base de données étiquetées –un cas très rare dans les données de transport–; les problèmes non supervisés, où il s'agit de lever une alarme lorsque nous nous éloignons du régime stationnaire. Ce dernier cas est complètement partagé entre les applications de santé et de transport.

Cette étude comparée permet de bien résumer les différentes étapes clés dans le traitement de ce type de données de santé et de celles de transport.

3.7 Conclusion & discussion

Le débat autour de l'apprentissage de profil se développe principalement sur deux axes : celui de la *privacy* ou protection de la vie privée et celui de la pertinence et de l'intelligibilité des propositions. Ces deux aspects sont liés : amélioration des systèmes rime la plupart du temps avec affinage des profils, et mécaniquement, modélisation de l'intimité d'une personne. En contrepartie, les systèmes exploitant ces profils apportent un certains nombres de services : meilleure réponse aux requêtes, suggestions pertinentes, etc... Le problème éthique vient principalement de la manière de financer ces services : beaucoup d'entreprises monétisent directement les profils dont la part d'intimité devient alors critique. Ce chapitre traitant de profils au sens large, il convient d'étendre la notion de *privacy* à d'autres applications comme l'analyse des transports en commun ou le diagnostic médical.

L'apprentissage de représentation va plus loin que les tâches usuelles en apprentissage statistique. Alors qu'il s'agissait au départ d'étendre l'impact des algorithmes automatiques vers l'extraction de caractéristiques, nous voyons dans ce chapitre, que l'enjeu est en réalité de comprendre la donnée indépendamment de la problématique visée. A long terme, le but est simplement d'introduire une nouvelle modélisation des connaissances au sens large, quelle que soit la modalité des données considérées. De ce point de vue, chaque tâche devient un moyen d'améliorer les représentations des données d'entrée du problème pour tendre vers une représentation optimale. Cette vision des choses ne correspond pas encore à l'état de l'art, où l'impact des nouveaux algorithmes d'apprentissage de représentation est essentiellement apprécié par la performance sur une tâche donnée, comme la reconnaissance d'objets dans les images [Kri+12 ; SZ14] ou la classification de sentiments dans les textes [Che+16].

Néanmoins, la multiplication des articles abordant plusieurs tâches à la fois dans les domaines du texte, de la recommandation ou du traitement du signal [CW08 ; Liu+09 ; ML13a] illustre une tendance forte. De la même manière, la mise à disposition d'extracteurs de caractéristiques très puissants et réutilisés dans des applications divers, en image [Kri+12 ; SZ14] ou en texte [Mik+13 ; Pen+14 ; Boj+17] montre qu'il n'est pas utopique d'envisager des représentations universelles de concepts de haut niveau, transférable d'une tâche à l'autre.

Le dernier point que nous souhaitons aborder dans cette conclusion est relatif à la protection des données personnelles lorsque nous apprenons des profils de personnes. L'apprentissage des profils d'utilisateurs est motivé par la création ou l'amélioration de certains services –recherche d'informations, systèmes de recommandation, etc...– et souvent financé par la monétisation des profils eux-mêmes –publicités ciblées, vente de repertoires enrichis de préférences personnalisées–. Dès lors que nous sommes ciblés sur la base de critères ethnique ou politique –par exemple–, le problème de protection des données personnelles devient critique [Goe+12]. A l'heure des réseaux sociaux, se protéger est de plus devenu impossible : même en prenant un maximum de précautions, le fait d'être relié à nos proches –explicitement ou implicitement– dévoile la plupart des informations sensibles [Con+11].

Plusieurs approches, parfois complémentaire, sont possibles pour préserver ces données extrêmement sensibles, chacune présentant des lacunes évidentes :

— une des réponses possibles est règlementaire. L'entrée en vigueur du RGPD –Règlement général sur la protection des données– en 2018 marque une évolution significative dans cette direction. Au niveau national, la CNIL –Commission

nationale de l'informatique et des libertés– régule, entre autres, les fichiers recensant les personnes depuis 1978. Bien qu'essentielle, cette approche est pénalisée par l'aspect international d'Internet qui pose des problèmes évidents pour faire respecter des règles locales.

— il faut aussi des solutions techniques pour préserver à la fois l'anonymat des utilisateurs et la qualité de service. Dans cette optique, les travaux précurseurs sur la *privacy* visent à agréger des profils cohérents pour conserver les qualités descriptives sans cibler personnellement les utilisateurs [VC03]. Il est aussi possible de chercher à encoder un profil de manière robuste et non inversible, en utilisant par exemple des techniques d'apprentissage statistique [Maa+14]. Une troisième voie consiste à ne pas partager les données en distribuant plutôt les algorithmes [Van+17].

Les travaux que nous avons menés et étudiés jusqu'ici nous suggèrent une voie alternative. Les profils appris dans des espaces latents de mots [Dia+16 ; Dia+17a ; Dia+18b] nous permettent d'interpréter, de décoder des profils de personnes. En trouvant une technique pour plus facilement encoder ces profils, nous serions en mesure de donner les moyens aux utilisateurs de prendre la main sur le visage qu'ils montrent lors de leur navigation à travers une sorte de cookie universel. Une personne pourrait gérer une bibliothèque de profils correspondant à différents usage, il serait dès lors possible non seulement de conserver une qualité de service, mais même de l'améliorer à travers une forme d'apprentissage actif. En parallèle, toute collecte d'information personnelle par les acteurs industriels serait forcément faite dans le seul but de la monétisation : cette collecte pourrait plus facilement être régulée voire interdite.

En dé-corrélant la problématique de qualité de service et celle de collecte des données personnelles, nous simplifierions la réflexion sur le coût des services qui sont actuellement financés de manière obscure (mail, réseaux sociaux, moteurs de recherche) tout en donnant plus de marge au législateur.

> *À partir de prémisses fausses, tout peut se démontrer. Ce n'est pas parce qu'un raisonnement est implacable qu'il n'est pas délirant.*
>
> **— François Lelord**

4

De la compréhension à la génération de données : les architectures de bout en bout

Le deep-learning a explosé sur la dernière décennie : la communauté scientifique s'est agrandie de manière exponentielle tandis que le palier de performances franchi a ouvert très largement l'univers industriel à des techniques qui étaient encore peu répandues jusque là.

Dans un premier temps, cet essor a reposé sur les applications en détection d'objets dans les images qui requièrent des bases étiquetées de très grandes tailles [Kri+12]. Une seconde vague d'avancées plus récentes repose sur des modèles génératifs, entrainés à reconstruire les données d'entrées. La combinaison du coût de reconstruction avec d'autres objectifs a ouvert la voie aux modèles d'apprentissage de bout en bout qui bénéficient de toutes sortes de supervisions, plus ou moins explicites, plus ou moins distantes.

Il a été démontré très vite qu'une des forces des approches neuronales réside dans leur capacité à exploiter de très grandes masses de données [LeC+89b ; Bes+11 ; Kri+12]. Cependant, cette propriété a longtemps été contrebalancée par l'absence de cadre efficace pour l'apprentissage non supervisé : les stratégies les plus efficaces sur de grandes masses de données étaient paradoxalement cantonnées aux données supervisées qui sont les plus rares et les plus chères.

Les choses ont évolué avec l'introduction de modèles entrainés à prédire une partie manquante des données [KB05 ; Col+11a] ou à supprimer le bruit dans les données [Vin+08]. Ces approches correspondent à des modèles d'apprentissage de représentations génériques qui peuvent ensuite être adaptés à des problèmes spécifiques via une procédure de *fine-tuning*. Cette mécanique permet élégamment de construire un pont entre de nombreuses applications concrètes et le domaine du *big-data*.

Atouts des techniques génératives d'apprentissage de représentation Les avancées induites par les stratégies génératives en deep learning sont très vastes :

— la prise en compte de masses de données gigantesques pour pré-entrainer des modèles ;

— la généralisation du transfert des modèles d'un problème à l'autre –sur la modalité textuelle puis image– ;

— la multiplication des fonctions de coût autour d'une architecture présentant une grande plasticité (ajout / retrait de couches dédiées à une tâche spécifique) ;
— la mise en avant des capacités génératives des approches neuronales qui étaient jusqu'ici plutôt l'apanage des modélisations bayesiennes [Jor98].

Les opportunités sont très nombreuses autour des principes énoncés ci-dessus. En débutant par les plus ambitieuses, il est légitime de se demander si le fait d'aborder une nouvelle étude de cas avec la connaissance amassée précédemment sur de si vastes corpus ne correspond pas à une avancée majeure sur une des problématiques historiques de l'IA, le *common-sense reasoning* [McC86]. Comprendre la manière dont ont été générées des données est indéniablement une forme de connaissance ; exploiter des capacités génératives pour produire un texte ou une image est aussi une manière d'expliquer le processus de décision, il s'agit d'une issue au paradigme de la boite noire [Goe+18].

D'un point de vue opérationnel, la capacité à multiplier les fonctions de coût sur une architecture flexible constitue une double avancée. D'une manière générale, il s'agit de tirer profit de toutes les contraintes existantes autour d'un problème que ce soit au niveau d'une observation ou d'un échantillon, en supervision directe ou distante [Zen+15]. Mais il s'agit aussi d'une manière remarquable de fusionner différentes modalités de données. En effet, les approches d'apprentissage de représentations projettent toutes sortes de données dans des espaces vectoriels de grande dimension : des images [Kri+12], des mots [Mik+13], des profils d'utilisateurs ou de produits [LS99],... Mixer des images ou des sons et des représentations de produits dans le cadre de la construction d'un système de recommandation devient trivial [HM16 ; DS14] ; mêler textes et images est même devenu rapidement un domaine de recherche à part entière à travers le *captioning* [Mao+14] puis le Visual Question Answering (VQA) [Ant+15]. C'est toute la fusion de données qui s'est retrouvée impactée : il ne s'agit pas de choisir entre fusion précoce ou tardive mais d'inventer une fusion sémantique dans un espace vectoriel dont la structuration bénéficie des deux modalités [Zab+17].

Limites actuelles Les architectures récentes de machine learning présentent malgré tout un certain nombre de limites :

— l'espace de représentation des concepts est un espace vectoriel latent dans lequel la construction d'un raisonnement n'est pas simple et souvent le fruit de calculs lourds et abstraits [Mik+13 ; Bor+11 ; Suk+15].
— Si l'idée de construire de manière non supervisée une vision générale du monde pour ensuite aborder d'autres tâches est séduisante, le phénomène du *catastrophic forgetting*, lors du passage d'une tâche à l'autre, interroge sur la nature des informations qui ont réellement été encodées [Goo+14a].
— De la même manière, l'extrême sensibilité des approches actuelles par rapport à des bruits adverses minimes remet sérieusement en question l'hypothèse du *common sense reasoning* qui serait présent dans ces espaces de représentations [Goo+14b].

Contributions Nous avons exploré ces stratégies génératives dans différents contextes applicatifs ces dernières années. Chaque application nous a permis de mettre en évidence certains aspects liés à ces approches.

En matière de système de recommandation, la question est de savoir si nous sommes capables de générer des revues puis de faire le lien entre la représentation

de notre utilisateur et le caractère des revues générées [Dia+19a ; Dia+19b]. En restant dans l'apprentissage de profils à partir de données textuelles, nous avons aussi travaillé sur la dynamique des utilisateurs dans les bases de CV et montré qu'une approche générative robuste –basée sur les caractères et non les mots– permet de donner des prédictions exploitables dans un contexte particulièrement bruité [Dia+17a].

A l'interface entre la modalité textuelle et les bases de connaissances, nous avons également travaillé sur les approches *end-to-end* en tentant de démêler les réelles avancées des biais de métriques d'évaluation [Tai+19b ; Tai+20a]. Dans une extension de ce travail, nous avons montré que le foisonnement autour de ces techniques conduit malheureusement régulièrement à la publication de résultats approximatifs voire erronés [Tai+20b].

Si la construction automatique de bases de connaissances est l'un des objectifs historiques de l'Intelligence Artificielle, le décodage des signaux du cerveau constitue également un objectif prisé. L'amélioration des techniques d'imagerie médicale (functional Magnetic Resonance Imaging) combinée aux nouvelles représentations sémantiques des mots ont ouvert des options réalistes pour attaquer cette problématique ambitieuse [Pip+14 ; Pip+15].

Concernant la modalité signal, les architectures récurrentes ont permis de représenter efficacement des données de tailles variables présentant régulièrement un fort niveau de bruit. Dans le cadre des signaux issus de la ville intelligent, nous avons travaillé sur le démêlage de ces espaces de représentations afin de mettre en évidence les spécificités des différents facteurs expliquant ces signaux [Gui+19 ; CD+20b ; CD+20a]. Afin de boucler avec le premier paragraphe de cette section, notons que nous travaillons actuellement sur l'extraction de profils à partir de signaux de conduite.

Organisation du chapitre L'organisation du chapitre suit les applications évoquées ci-dessus : nous étudierons d'abord l'intérêt des modèles génératifs sur l'apprentissage de profils et les données textuelles avant de décrire nos expériences sur les bases de connaissances. Nous verrons ensuite comment est formulée la tâche du *brain-reading* et comment elle bénéficie naturellement des progrès sur la représentation des mots. Enfin, nous reviendrons aux profils à travers la modalité signal.

4.1 Extraction de profils textuels par approche générative

Comme dans le chapitre précédent, la plus grande partie de nos contributions concerne la modalité textuelle et, notamment, la construction de profils à partir de données textuelles.

4.1.1 Profils ancrés dans un espace textuel en recommandation

Cette partie correspond à la suite des travaux présentés dans le chapitre précédent. Nos contributions précédentes étaient axées sur des espaces de mots puis sur l'apprentissage d'attention personnalisée sur les mots et les phrases : dans tous les cas, il s'agissait de modèles de type lecteur.

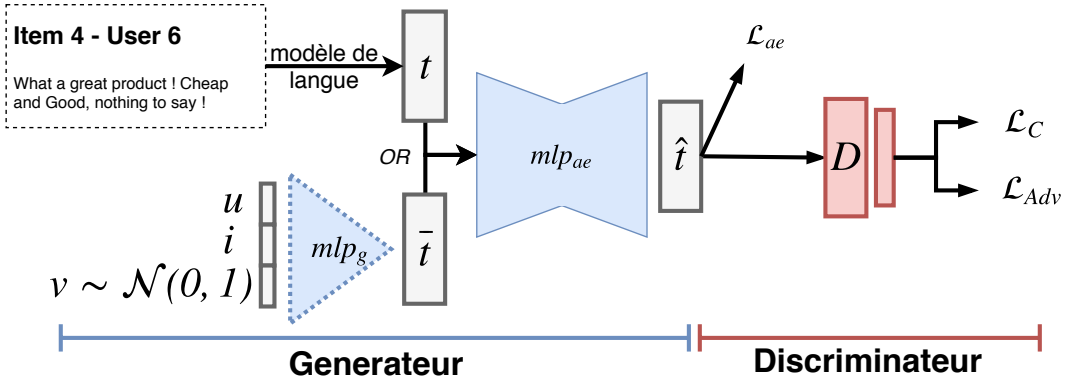


Figure 4.1: (à gauche) Generateur : (a) Génère une représentation clone $\bar{t}$ à partir d'une paire de représentations (utilisateur,produit). (b) Auto-encode t ou $\bar{t}$ en $\hat{t}$. (à droite) Discriminateur : Predit le sentiment et la source de $\hat{t}$

Nous présentons ici un modèle génératif reposant sur une architecture adverse pour effectuer du filtrage collaboratif supervisé par la génération de textes plutôt que via la prédiction directe de notes. L'idée revient à faire une sorte d'analyse de sentiments en aveugle : le profil de l'utilisateur et de l'item permet de générer un texte, lui-même analysé en opinion. Les détails sont fournis dans [Dia+19a].

Modélisation Formellement soit des tuples utilisateur, produit, note, texte (u, i, r, t) comme données, nous avons deux tâches : l'analyse de sentiments et le filtrage collaboratif. L'objectif de la première tâche est de prédire r, à l'aide du texte t mais aussi avec les profils u, i. Pour la seconde, l'objectif final est le même ; prédire la note mais seulement avec les profils (u, i). Dans cette contribution, nous proposons de mettre les deux tâches en compétitions : la prédiction d'une note r_{ui} se fera donc de la manière suivante :

$$r_{ui} = \underbrace{f_s(\overbrace{f_r(u,i)}^{\text{Recommandation générative}})}_{\text{Analyse de Sentiments}}, \qquad f_r(u,i) = \bar{t} \tag{4.1}$$

Afin de nous abstraire des contraintes liées aux données textuelles, nous proposons d'encoder le texte à l'aide de *word2vec* [Mik+13] : chaque avis t est représenté comme le barycentre de ses représentations de mots $w_i \in \mathbb{R}^d$, $t = \frac{1}{n}\sum_{i=1}^{n} w_i$

Notre modèle adverse (fig. 4.1) est composé de deux sous-réseaux – un générateur et un discriminateur – optimisés à tour de rôle dans un schéma d'apprentissage adverse. Nous détaillons d'abord séparément les fonctions de chaque réseau avant d'expliciter les coûts optimisés.

Générateur : Le premier sous-réseau – f_r – fait office de générateur. Il est lui même composé de deux modules : un auto-encodeur mlp_{ae} et un perceptron multicouches mlp_g qui génère un texte $\bar{t}$ à partir d'un triplet (u, i, v) ; où v est un vecteur de "bruits" tirés selon une loi normale centrée réduite. L'auto-encodeur peut sembler superflu dans cette architecture, mais il apprend la distribution des données en encodant le texte réel t dans une représentation $\hat{t}_r$. De ce fait, mlp_g se focalise sur l'extraction des informations critiques à partir des profils sans s'occuper de la nature générale d'une

revue. Les expériences ont montré que que seule la combinaison de ces éléments permettait d'apprendre à générer un texte consistant.

$$u, i, v \in \mathbb{R}^d, \qquad \bar{t} = mlp_g(u, i, v), \qquad \hat{t} = mlp_{ae}(t \vee \bar{t}) \tag{4.2}$$

L'architecture repose sur trois fonctions de coût séparées :

($\mathcal{L}_{\mathbf{ae}}$) Le générateur doit être capable d'auto-encoder un texte original t. Pour cela, il doit minimiser l'erreur de reconstruction au sens des moindres carrés : $MSE(t, \hat{t}_r)$ [1]. En apprenant à auto-encoder les données réelles, le générateur apprend la distribution des données [Gal+87]. En parallèle, les clones générés doivent également être "reconstructibles". Nous minimisons donc également l'erreur de leur reconstruction $MSE(\bar{t}, \hat{t}_f)$. De cette façon, nous garantissons que tous les $\hat{t}$ proviennent de la distribution des données apprises.

($\mathcal{L}^{\mathbf{g}}_{\mathbf{Adv}}$) Cependant, rien ne garantit que les reconstructions de textes générés $\hat{t}_f$, et celles de textes réels $\hat{t}_r$ soient similaires. Pour contraindre le générateur à s'aligner sur les données réelles, les reconstructions doivent être indissociables. Cette propriété est apprise grâce à une optimisation min-max où le générateur et le discriminateur ont des objectifs opposés $\mathcal{L}^g_{adv}$ et $\mathcal{L}^d_{adv}$. Concrètement, le générateur, via l'auto-encodeur, va essayer de faire passer la reconstruction du texte réel $\hat{t}_r$ pour la reconstruction du clone $\hat{t}_f$ et vice-versa. Formellement, le coup minimisé est l'entropie croisée binaire : $BCE(mlp_d(\hat{t}_r), 0)$ et $BCE(mlp_d(\hat{t}_f), 1)$ [2]

($\mathcal{L}_{\mathbf{C}}$) Le sentiment r lié aux reconstructions $\hat{t}_r$ doit être correctement prédit. Cela se fait dans le cadre d'une minimisation simultanée de la log-vraisemblance négative par le générateur.

Discriminateur : Le second sous-réseau – f_s – est composé d'un perceptron multi-couches mlp_d à deux têtes. Il fait à la fois office de discriminateur et de classifieur : Il doit parvenir à prédire correctement les polarités associées aux reconstructions $\hat{t}$ tout en arrivant à faire la différence entre celles provenant d'un texte légitime t ou d'un clone $\bar{t}$.

Formellement, le discriminateur minimise donc deux fonctions de coût distinctes :

($\mathcal{L}_{\mathbf{C}}$) le discriminateur, qui fait de l'analyse de sentiments, doit être capable de prédire correctement la polarité d'une reconstruction $\hat{t}$.

($\mathcal{L}^{\mathbf{d}}_{\mathbf{adv}}$) Le discriminateur doit aussi réussir à distinguer une reconstruction issue d'un texte généré $\bar{t}$ de celle issue d'un texte original t. Cette compétition entre générateur et discriminateur doit forcer le générateur à mêler $\bar{t}$ et t au sein d'une même reconstruction $\hat{t}$ et, donc, aligner t et $\bar{t}$.

Finalement, avec $\mathcal{L}_C$ l'entropie croisée et $\mathcal{L}^d_{adv}$ la fonction de coût adverse, le coût minimisé est le suivant :

$$\begin{aligned} \mathcal{L}_d &= \mathcal{L}_C + \mathcal{L}^d_{adv}. \\ &= CCE(c, r) + BCE(\hat{t}_f, 0) + BCE(\hat{t}_r, 1) \end{aligned} \tag{4.3}$$

1. $MSE(x, \hat{x}) = (x - \hat{x})^2$
2. $BCE(x, y) = -\left[y \cdot \log x + (1 - y) \cdot \log(1 - x)\right]$

Dataset (#reviews)	Mean (μ)	w/offset	MF	HFT	BSA
Instant Video (37.126)	1.25	1.137	1.024	0.933	**0.924**
Digital Music (64,706)	1.19	0.965	0.903	0.844	**0.832**
Video Games (231,780)	1.45	1.281	1.267	1.097	**1.082**
CSJ (278,677)	1.215	1.323	1.365	1.107	**1.101**
Movie (1,697,533)	1.436	1.148	1.118	1.020	**1.009**

Table 4.1: Erreur quadratique moyenne en prédiction de notes. Les modèles de références sont : La moyenne globale de la base de donnée μ, le biais de notation utilisateur-item global et une factorisation matricielle de référence [Kor+09]. HFT utilise le texte, il est issu de [ML13a]. Les valeurs présentées sont des moyennes obtenues par validation croisée sur 5 parties

Expériences les expériences ont été conduites sur 5 sous domaines de la base Amazon [McA+15]. Cette architecture, dénommée BSA –Blind Sentiment Analysis– dans le tableau 4.1, nous permet d'obtenir de très bonnes performances en prédiction de notes.

Nous sommes en mesure d'expliciter les suggestions faites en recherchant les phrases et les avis proches du texte généré $\hat{t}$. Le mécanisme est sommaire et pourrait être avantageusement remplacé par un décodeur récurrent, mais l'enjeu de cet article est surtout de démontrer la faisabilité d'une architecture GAN multimodale reposant largement sur des données textuelles.

L'étape suivante consiste à démêler cet espace latent de représentations pour faire émerger des dimensions significatives. Une recherche préliminaire montre qu'au moins une dimension intègre des opinions, comme le montre la figure 4.2.

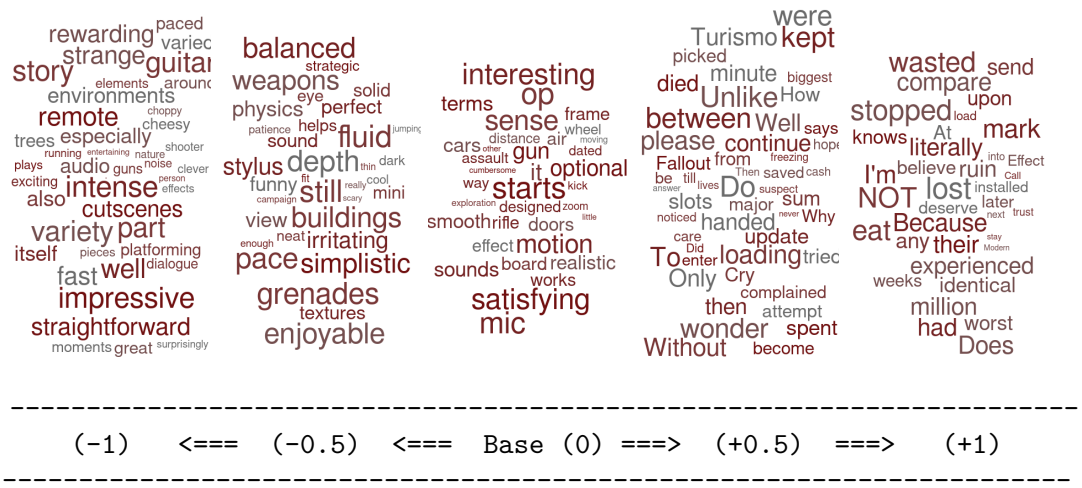


Figure 4.2: Evolution du nuage de mots associé à $\hat{t}$ en changeant une dimension du vecteur de bruit v.

4.1.2 Modélisation textuelle & temporelle

Dans cette section, nous nous intéressons à l'apprentissage d'un espace latent textuel pour la modélisation de la dynamique. Le cas d'usage concerne la représen-

tation –et la manipulation– des parcours professionnels extraits de *curriculum vitae*. Il s'agit d'une vision générative et intégrée de ce qui a été présenté dans le chapitre précédent 3.5.

Nous proposons d'adapter les méthodes d'encodeur-décodeur issues de la traduction automatique [Cho+14], pour construire un espace latent textuel intégrant de bonnes propriétés sémantiques et syntaxiques. Ces techniques à base de réseaux de neurones récurrents permettent notamment d'encoder caractère par caractère un champ textuel et de décoder la représentation obtenue en une autre chaîne de caractères. Le décodeur donne une interprétation claire de l'espace de représentation en langage naturel. La dynamique est modélisée par un second réseau récurrent qui est défini sur le premier espace latent.

4.1.2.1 Modélisation

Les réseaux de neurones récurrents (RNN) présentent l'avantage de traiter naturellement des séquences de longueurs variables, ce qui est bien adapté à la modélisation des données textuelles et aux carrières décrites dans les CV. En effet, contrairement aux réseaux de neurones classiques, le calcul d'une sortie y_t à l'instant t prend en compte à la fois l'entrée actuelle x_t et une représentation des entrée précédentes h_{t-1} [Elm90].

$$\begin{cases} h_t = \sigma_h(W_h x_t + U_h h_{t-1} + b_h) \\ y_t = \sigma_y(W_y h_t + b_y) \end{cases} \tag{4.4}$$

C'est cet état caché h_t qui est une composition non linéaire de l'ensemble des entrées $x_1, .., x_t$ modulée par une fonction d'activation σ_h qui permet de conserver une mémoire de la séquence, il s'agit donc d'une représentation latente de l'ensemble du passé. Néanmoins, cette version classique du réseau de neurone récurrent souffre du problème de disparition du gradient ; en effet, celui-ci décroit de manière exponentielle du fait des nombreuses multiplications lors de la rétro-propagation temporelle, ce qui rend difficile l'apprentissage sur de longues séquences. Face à ce problème, [Le+15] montrent qu'une initialisation spécifique, combinée à des fonctions d'activations *ReLU*[3] rend possible l'apprentissage de dépendances plus longues. Cependant, les meilleurs résultats sont obtenus en modifiant l'architecture afin de diminuer l'enchaînement des multiplications [Mik+14]. C'est dans ce contexte qu'ont été développées les cellules "à portes". Les deux cellules à portes les plus connues sont le LSTM (Long Short Term Memory) [HS97] et sa variante simplifiée, la GRU (Gated Recurrent Unit) [Cho+14]. [Col15] expose en détail l'intuition et le fonctionnement de ces cellules. En règle générale et en particulier sur les données textuelles, les cellules GRU et LSTM offrent des performances similaires [Gre+15 ; Joz+15 ; Chu+15]. Nous privilégions donc la GRU pour nos travaux car elle est plus simple. Tous les détails sont donnés dans [Dia+17a].

Encodeur - Décodeur : Modèle de langue syntaxique et sémantique Nous proposons d'adapter les méthodes d'encodeur-décodeur issues de la traduction automatique [Cho+14] pour construire un espace latent de carrières professionnelles. La force des techniques à base de réseaux de neurones récurrents est de pouvoir encoder directement caractère par caractère un champ textuel. Cette faculté est particuliè-

3. $f(x) = \max(0, x)$

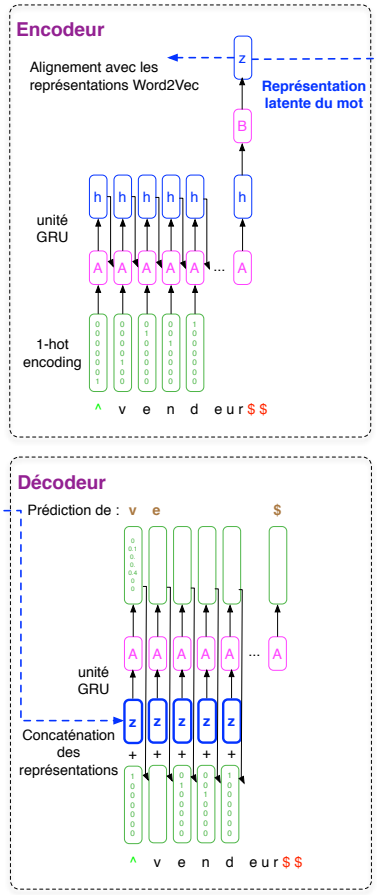


Figure 4.3: Encodage robuste des intitulés de postes au niveau des caractères par RNN.

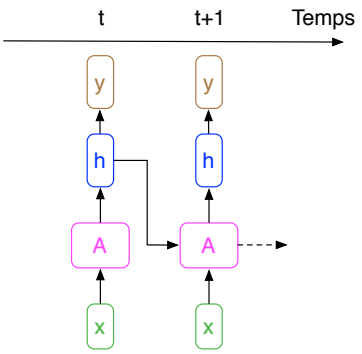


Figure 4.4: Modélisation de l'évolution des postes occupés par un individu dans sa carrière

rement importante ici puisque nous devons unifier des représentations syntaxiques faites de différents termes, abréviations voire fautes d'orthographe. Dans un second temps, le décodeur offre la possibilité de ré-interpréter un concept de l'espace latent en générant une chaîne de caractères ; il est aussi basé sur un réseau de neurones récurrents.

Encodeur. Nous avons opté pour une architecture issue de la traduction car la problématique est la même : dans un texte libre $\mathbf{txt}$, non-normalisé, plusieurs graphies $\{\mathbf{txt}_1, \mathbf{txt}_2, ...\}$ peuvent correspondre à la même signification. Le but est donc d'apprendre une représentation unifiée $\mathbf{z} \in \mathbb{R}^d$ captant le sens général de la phrase, ainsi qu'une fonction pour passer d'un texte à sa représentation, avec la propriété suivante : $\forall n,\ f(\mathbf{txt}_n) \approx \mathbf{z}$. Cette fonction correspond à l'encodeur. Comme le montre la figure 4.3, l'unité de base du système est la lettre, passée à l'encodeur dans un format *1-hot* : un vecteur de la taille de l'alphabet $\mathcal{A}$ rempli de 0, sauf à la position du caractère visé. Deux caractères spéciaux ^ et $ encodent respectivement le début et la fin de séquence. Un texte $\mathbf{txt}$ correspond donc à un ensemble de lettres $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ avec $\mathbf{x}_{lettre} \in \{0,1\}^{|\mathcal{A}|}$. Notre architecture permet d'encoder l'information lue dans un vecteur de $\mathbf{z} \in \mathbb{R}^d$ ($d = 256$ dans le cadre de cet article).

Décodeur. Le décodeur est également un réseau de neurones récurrents, dans lequel nous jouons sur l'entrée pour garantir la signification du résultat. A chaque pas de temps, nous fournissons le code $\mathbf{z}$ du concept, concaténé avec la représentation d'une lettre (au format *1-hot*). Nous sommes ainsi en mesure de prédire la prochaine lettre la plus vraisemblable étant donné le contexte $\mathbf{z}$.

Pour générer du texte, nous partons du caractère *neutre* ^ (codant le début de séquence) ; le début de la séquence dépend donc principalement du contexte ; ensuite, le modèle de langue appris garantit l'intelligibilité du texte généré. Nous nous arrêtons lors de la prédiction du caractère $.

Apprentissage. Dans le détail, les CV sont anonymisés et pré-traités ; chaque ligne (professionnelle ou cursus scolaire) est divisée en un intitulé et une description (ensemble de mots de contexte). Pour l'apprentissage, chaque intitulé est transformé en une suite de vecteurs 1-hot $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}, \mathbf{x}_i \in \{0,1\}^{|\mathcal{A}|}$ représentant l'enchaînement des caractères. Les séquences sont limitées à 32 caractères pour des raisons computationnelles. L'encodeur et le décodeur sont appris simultanément par descente de gradient stochastique temporelle modifiée selon la technique de RMSProp [Hin+12].

Afin que ces représentations $\mathbf{z}$ soient pertinentes, nous combinons différents critères d'apprentissage. Nous utilisons d'abord des représentations externes pour stabiliser et accélérer l'apprentissage : nous rapprochons les $\mathbf{z}$ d'un modèle pré-entrainé sur un large corpus avec l'algorithme Word2Vec [4]. Nous cherchons ensuite à reconstruire les intitulés de postes à l'identique, à la manière d'un auto-encodeur : les corrections ont alors lieu à la fois sur l'encodeur et le décodeur. Enfin, nous cherchons à reconstruire le contexte de l'intitulé de poste à partir de l'intitulé lui-même afin de rapprocher les représentations des synonymes et donner une véritable dimension sémantique à la représentation. La philosophie est la même que pour le *negative sampling* de l'approche Word2Vec.

4. Corpus `frWac` disponible : `http://fauconnier.github.io/#data`

Prédiction de parcours professionnels La prédiction du prochain item (travail ou cursus scolaire) dans un CV est une tâche très difficile pour deux raisons principales. La barrière syntaxique, à travers la personnalisation des intitulés, empêche de capitaliser l'expérience entre les utilisateurs. La nature même des carrières pose ensuite problème, la majorité des transitions renvoyant vers un poste équivalent ou un changement de contexte mais plus rarement vers une promotion. Nous avons abordé la variabilité de la syntaxe dans la section précédente, nous nous intéressons ici à la modélisation des hiérarchies de postes et à la prédiction des évolutions de carrières des utilisateurs.

Le prédicteur de prochain item est de nouveau un réseau de neurones récurrent avec une couche GRU. Pour étudier la topologie de l'espace des items, nous pouvons lui fournir une seule entrée $\mathbf{z}_t$; nous obtenons alors le comportement $\mathbf{z}_{t+1}$ dominant à partir de ce point de départ. Le même réseau est capable d'assimiler une séquence d'items passés $S_u = [\mathbf{z}_1, \ldots, \mathbf{z}_t]$ pour prédire $\mathbf{z}_{t+1}$. La sortie est alors personnalisée par rapport à l'ensemble du parcours d'un utilisateur u. Ce système conserve et apprend l'ordonnancement des items mais pas les dates ; le fait de rester longtemps sur un poste n'est pas encore pris en compte.

Nous obtenons donc un système global composé de deux RNN disposés orthogonalement : l'encodeur/décodeur et le modèle de prédiction du prochain item. L'ensemble des CV est utilisé pour apprendre le premier RNN, c'est à dire le modèle de langue. Le prédicteur est lui classiquement évalué en validation croisée (5 *folds*) : une partie des CV servant à apprendre le modèle et l'autre à l'évaluer.

4.1.2.2 Expériences

Données extraites des *curriculum vitae* Nous avons à notre dispositions un ensemble de 915 925 *curriculum vitae* anonymisés et pré-traités au format XML. Chaque CV devient ainsi une paire de séquences : le parcours éducatif et le parcours professionnel. Après filtrage des données abbérantes, il reste 656 134 paires de parcours éducatifs et professionnels. Le parcours éducatif est composé d'une séquence de noms de diplômes, triée par date d'obtention (du plus vieux au plus récent). Le parcours professionnel est quant à lui composé d'une séquence de titres de postes associés à leurs descriptions, triée par date de début de poste (du plus vieux au plus récent). Les noms de diplômes, les intitulés de postes et leurs descriptions associées sont extraits directement du texte des CV (sans traitement linguistique).

Après avoir enlevé les espaces redondants, les accents et les majuscules, il reste 266 622 intitulés de postes et 552 439 diplômes distincts. A titre de comparaison, le référentiel métier le plus complet actuellement (ESCO) compte 4800 entrées. En apprentissage statistique, l'approche la plus classique consiste à conserver uniquement les N intitulés/noms les plus fréquents et de remplacer les autres par un même identifiant *inconnu*. Malheureusement, le tableau 4.2 montre que la distribution des intitulés de postes et surtout de diplômes est trop uniforme : la proportion d'*inconnus* est trop élevée avec ce système. Les variations de formulation, les abréviations et les flexions sont trop nombreuses : c'est ce qui nous a poussé à développer le modèle de normalisation linguistique.

# Int. conservés :	5000
# Int. inc.	443k (15%)
# Diplomes inc.	737k (55%)
# Int. conservés :	3000
# Int. inc.	536k (18%)
# Diplomes inc.	782k (58%)
# Int. conservés :	2000
# Int. inc.	631k (21%)
# Diplomes inc.	817k (61%)

Table 4.2: Nombre et proportion d'intitulés inconnus dans l'ensemble des parcours en fonction du nombre d'intitulés uniques fréquents conservés.

Propriétés syntaxiques et sémantiques Nous souhaitons apprendre un encodeur robuste. De ce fait, l'encodeur doit être insensible aux différentes graphies possibles d'un même mot ; en d'autres termes, un mot et ses variantes mal orthographiées

doivent avoir des représentations proches. Pour évaluer cette propriété nous procédons à une expérience simple : nous suivons l'évolution d'une représentation (par rapport à son original) en fonction de modifications orthographiques additives, soustractives et commutatives. Les représentations étant en grande dimension (256), nous utilisons la similarité cosinus normalisée :

$$\text{sim}(\mathbf{z}, \mathbf{z}_{mod}) = \frac{1}{2}\left(\frac{\mathbf{z}\cdot\mathbf{z}_{mod}}{\|\mathbf{z}\|\|\mathbf{z}_{mod}\|} + 1\right) \in [0,1] \tag{4.5}$$

La figure 4.5 montre la robustesse de l'encodeur : jusqu'à environ 4 modifications, la similarité reste supérieure à 80%. Ce résultat est important : une seconde illustration (figure 4.7) montre que des représentations proches mènent à la même reconstruction.

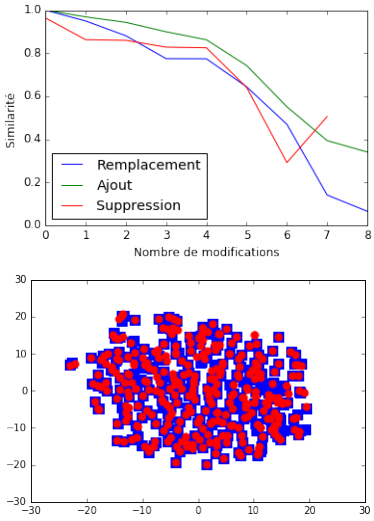


Figure 4.5: (en haut) Évolution de la similarité cosinus normalisée (eq. 4.5) du vecteur bruité par rapport à l'original en fonction de différentes modifications syntaxiques. (en bas) Projection T-SNE des représentations de métiers masculins (carrés bleus) et de leurs équivalents féminins (ronds rouges)

```
^manutention$      => ^manutentionnaire$
^manutencion$      => ^manutentionnaire$
^manutentionaire$ => ^manutentionnaire$
```

Figure 4.7: Processus d'encodage/décodage de différentes flexion d'un intitulé : la reconstruction est identique.

Nous nous sommes également intéressés au problème du genre : en effet, la plupart des métiers ont une version féminine de leur intitulé. Celle-ci se limite souvent à des modifications mineures d'orthographe, mais pas toujours. Comme pour l'expérience précédent, l'enjeu consiste à avoir des représentations proches. Pour vérifier expérimentalement cela, nous disposons d'une liste de 312 métiers féminisés [5]. En projetant les représentations des métiers masculins (carrés bleus) et féminins (ronds rouges) dans un espace bi-dimensionnel grâce à l'algorithme T-SNE [MH08], nous montrons à quel point les deux déclinaisons d'un même métier ont des représentations proches.

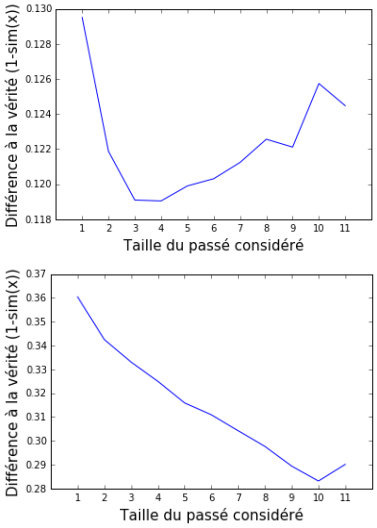


Figure 4.6: Courbe de l'erreur de prédiction du prochain emploi d'un opérateur hiérarchique en fonction de la taille du passé considéré, pour des représentations issues de Word2Vec (haut) et de notre encodeur (bas).

Evaluation de la prédiction de carrière Comme dans la section précédente, nous évaluons les performances directement dans l'espace latent en mesurant l'écart entre la position prédite de l'utilisateur à l'instant $t+1$ et la représentation encodée de l'item de son CV à $t+1$ en utilisant la métrique de l'équation 4.5. Une telle mesure est certes relativement abstraite, cependant, il est impossible de faire autrement à cause de la variabilité dans les intitulés employés par les utilisateurs. Après avoir appris la structure encodeur/décodeur sur l'intégralité de nos données, nous divisons les 656 134 parcours en cinq parts égales ($\approx$ 131 226 parcours chacune) pour faire de la validation croisée. Nous apprenons un RNN sur 4/5ème des données et évaluons son comportement sur la base de test restante. Nos représentations étant partiellement alignées sur des représentations Word2Vec et apprises avec une philosophie proche, nous comparons les comportements de deux modèles prédictifs dans les deux univers. Plus important : nous avons fait varier la taille de l'historique considéré dans chaque CV. Avec un historique de taille unitaire, nous avons simplement un modèle général d'enchaînement des intitulés ; mais lorsque l'historique augmente, nous obtenons un prédicteur personnalisé compilant les informations sur l'utilisateur pour affiner la proposition.

5. Liste disponible sur : `http://ameliorersonfrancais.com/grammaire/genres/feminisation-des-metiers-et-des-titres/`

```
 t    | vérité:                | prédiction:
-------------------------------------------------------------
t(0)  | ^assistant$            |
t(1)  | ^assistant commercial$ | ^assistante commercial$
t(2)  | ^conseiller clientele$ | ^conseiller commercial$
t(3)  | ^conseiller clientele$ | ^conseiller commercial$
```

Figure 4.8: Exemple de prédiction d'un parcours professionnels

Les résultats sont présentés en figure 4.6, la notion de similarité est simplement renversée pour mesurer une erreur : $\text{err}(\mathbf{z}, \mathbf{z}') = 1 - \text{sim}(\mathbf{z}, \mathbf{z}')$. La première conclusion concerne la capacité d'exploitation de l'historique : les représentations E/D (basées sur l'architecture encodeur/décodeur) permettent de tirer parti de l'ensemble de l'historique pour améliorer la prédiction alors que celles W2V (issues de Word2Vec) plafonnent à un horizon 3. L'axe des ordonnées n'est pas directement interprétable, les représentations évoluant dans 2 espaces distincts non normalisés. Les expériences complémentaires réalisées dans l'article [Dia+17a] montrent que notre système est bien supérieur au meilleur word2vec même avec un historique faible.

La figure 4.8 présente un exemple de résultat, en partant d'un exemple issu de la base de CV et en confrontant les prédictions avec l'évolution effective de l'utilisateur.

4.1.3 Perspectives sur le traitement des données textuelles

La capacité à générer des données textuelles vraisemblables a révolutionné la traduction automatique [Cho+14]. Mais ce n'était qu'un début et ces approches, couplées à des stratégies prédictives –des mots suivants, des phrases suivantes, des mots qui ne devraient pas figurer dans la phrase– ont ouvert la voie aux modèles de langues et à la représentation contextuelle des mots qui ont fait évoluer significativement l'état de l'art dans toutes les tâches liées au traitement de la langue naturelle [Pet+18 ; Dev+18].

Les exemples détaillés précédemment montrent bien la force actuelle des architectures de deep-learning :

— la possibilité de partir de modèles de langue pré-appris sur de large corpus pour intégrer une connaissance a priori qui guide l'apprentissage dans la bonne direction ;

— un espace latent unificateur de modalités, bien plus efficace qu'une fusion tardive pour réunir textes, traces utilisateurs voire modéliser la dynamique d'un système ;

— une architecture dont la plasticité permet la prise en compte de multiples contraintes afin de coller au plus près du cadre défini par les experts du domaine tout en gardant une robustesse par rapport au bruit présent dans les données ;

— un espace paramétrique vaste qui tire parti des très grandes masses de données disponibles à l'heure actuelle et un cadre logiciel de grande qualité permettant la réalisation rapide d'expériences très lourdes.

La double question que nous devrons nous poser dans les années qui viennent est la suivante : sommes-nous capables d'interpréter l'espace de représentation et

de manipuler le sens des phrases en jouant sur ces dimensions ? Et, à plus long terme, sommes-nous capables de générer un texte consistant, plus long, sur un sujet complexe intégrant par exemple des causalités ? La première problématique est difficile dans un cadre non-supervisé mais des expériences prometteuses existent déjà quand la supervision est disponible [Lam+19]. La seconde problématique est abordée par la communauté du data-to-text : l'idée est de structurer un résumé automatique à partir d'un ensemble de données en construisant un raisonnement logique [Reb+20].

L'ensemble de ces avancées très récentes pose la question du passage de la représentation des textes bruts à celle des connaissances. C'est cette problématique que nous abordons frontalement dans la section suivante.

4.2 Modèles end-to-end pour la construction de bases de connaissances

La construction et l'exploitation des bases de connaissances est une tâche historique en IA comme en témoigne les campagnes annuelles TREC –*Text REtrieval Conference*– sur la recherche d'informations organisées par l'institut NIST à partir de 1992 et les conférence MUC –*Message Understanding Conference*– organisées par la DARPA à partir de 1987.

L'amélioration récente des modèles de langues et la structuration logique des espaces de représentations de mots et de phrases [Mik+13] nous poussent à répondre à de nouvelles questions :

1. à la base du système, la représentation historique des connaissances sous forme de triplet RDF (entité$_1$, Relation, entité$_2$) [Dec+00] est-elle encore légitime ?
2. Dans l'alternative, l'espace de représentation des phrases, en franchissant des paliers sur le plan sémantique, devient-il un espace de représentation des connaissances ?
3. Les modèles de langue peuvent-ils nous aider à améliorer l'extraction de connaissances à partir de textes bruts ?
4. L'apprentissage de représentation permet-il de développer des outils plus performant que la logique formelle pour la déduction de nouvelles connaissances à partir d'une base existante ?

Dans cette section, nous conservons la logique de triplet, d'identification des entités nommées –Named Entity Recognition, NER– et de relation entre ces entités ; ainsi, nous ne considérons pas –encore– les représentations latentes de phrases comme des connaissances. Il est par contre évident que les avancées récentes en traitement automatique du langage permettent d'extraire des connaissances de manière beaucoup plus fiable : c'est le thème principal abordé ici. Malgré le nombre croissant de travaux pertinents sur le sujet [Bor+11 ; Soc+13a ; Lin+17], nous n'avons pas encore étudié la question de l'inférence de nouvelles connaissances à l'aide des techniques d'apprentissage de représentation.

4.2.1 Avancées récentes en reconnaissance d'entités nommées (NER)

Les tâches d'extraction d'informations remontent à longtemps et fédèrent une large communauté qui produit une littérature foisonnante attestant de progrès

réguliers par rapport à l'état de l'art. Dans un premier temps, nous avons cherché à faire la part des choses au niveau de l'extraction des entités nommées (NER en anglais ou REN en traduction littérale). Nos expériences préliminaires ont montré que le sur-apprentissage était un problème critique sur cette tâche : en effet, les réseaux de neurones exploitent leurs millions de paramètres pour extraire des entités, mais aussi pour apprendre par coeur celles qui ont été vues en apprentissage. Notre première contribution concerne ainsi les protocoles d'évaluation et la distinction entre la reconnaissance d'entités déjà vues en apprentissage de la détection des nouvelles entités. Cette contribution est détaillée dans [Tai+19a ; Tai+19b ; Tai+20a]

Recoupement lexical et données utilisées La partie anglaise de **CoNLL03** [SDM03] est le benchmark standard en REN et est composé d'articles Reuters datés de 1996 et annotés pour quatre types : Organisation (ORG), Personne (PER), Localité (LOC) et Divers (MISC).

OntoNotes 5.0 [Wei+13] est composé de documents de six domaines annotés pour la REN et la Résolution de Coréférence. Il est annoté manuellement pour onze types d'entités et sept types de valeurs qui sont généralement traités sans distinction. La partition entraînement/test classique pour la REN [Str+17] est la même que celle de la tâche de Résolution de Coréférence de CoNLL-2012. Nous avons aligné les types de OntoNotes sur ceux de CoNLL afin d'unifier le cadre expérimental.

Nous le quantifions en séparant les occurrences des mentions dans les sets d'évaluation en trois catégories : recoupement exact, recoupement partiel et recoupement nul, de manière similaire à [Aug+17]. Une mention d'un jeu d'évaluation est un recoupement exact si elle apparaît sous l'exacte même forme sensible à la capitalisation dans le jeu d'entraînement et annotée avec le même type. Le recoupement est partiel s'il n'est pas exact mais qu'au moins un des mots non vides de la mention apparaît dans une mention de même type. Toutes les autres mentions ont un recoupement nul : leurs mots non vides ne sont jamais rencontrés pendant l'entraînement. Ainsi, la proportion de recoupements partiels et nuls reflète la capacité d'un jeu de test à évaluer la capacité de généralisation d'un algorithme aux entités non rencontrées, ce qui est un premier pas nécessaire à l'adaptation de domaine.

	CoNLL03				
	LOC	MISC	ORG	PER	Tous
Exact	82%	67%	54%	14%	52%
Partiel	4%	11%	17%	43%	20%
Nul	14%	22%	29%	43%	28%
	OntoNotes réaligné				
	LOC	MISC	ORG	PER	Tous
Exact	87%	93%	54%	49%	69%
Partiel	6%	2%	32%	36%	20%
Nul	7%	5%	14%	15%	11%

Table 4.3: Recoupement lexical des occurrences de mentions des jeux de test avec les jeux d'entraînement respectifs pour CoNLL03 et OntoNotes original et réaligné. La dernière colonne montre le recoupement entre le test de OntoNotes réaligné et l'entraînement de CoNLL03 dans l'évaluation extra domaine.

Performances *vs* taux de recoupement Comme reporté dans la Table 4.3, les deux jeux de données montrent un important recoupement lexical de mentions. Dans CoNLL03, plus de la moitié des occurrences de mentions du jeu de test est présente dans le jeu d'entraînement alors que seulement 28% sont totalement nouvelles. Dans OntoNotes, le recoupement est encore pire avec 67% de recoupement exact contre 9% de nouvelles mentions. De plus, nous remarquons une influence significative du type d'entité puisque LOC et MISC présentent le recoupement le plus important alors que PER et ORG ont un vocabulaire plus varié.

Cela montre que les deux principaux jeux de données étalons en REN en anglais évaluent surtout la performance d'extraction des mentions déjà rencontrées lors de l'entraînement, bien qu'apparaissant dans des phrases différentes. De telles proportions de recoupement lexical ne sont pas réalistes dans des applications réelles où un modèle doit traiter quelques ordres de grandeurs de documents de plus en inférence qu'en entraînement pour rentabiliser le coût de l'annotation. L'amélioration spécifique des performances sur les nouvelles mentions revêt donc une importance cruciale en usage réel ; cet aspect est sous-estimé dans les benchmarks actuels.

				CoNLL03				OntoNotes			
Entraînement	Modèle	Représentation	Dim	Exact	Partiel	Nul	Tous	Exact	Partiel	Nul	Tous
CoNLL03	**BiLSTM-CRF**	BERT	4096	95.7	88.8	82.2	90.5	95.1	82.9	73.5	**85.0**
		ELMo	1024	95.9	89.2	85.8	**91.8**	94.3	79.2	72.4	83.4
		Flair	4096	95.4	88.1	83.5	90.6	94.0	76.1	62.1	79.0
		GloVe + char	350	95.3	85.5	83.1	89.9	93.9	73.9	60.4	77.9
		GloVe	300	95.1	85.3	81.1	89.3	93.7	73.0	57.4	76.9
	Map-CRF	BERT	4096	93.2	85.8	73.7	86.2	93.5	77.8	67.8	80.9
		ELMo	1024	93.7	87.2	80.1	**88.7**	93.6	79.1	69.5	**82.2**
		Flair	4096	94.3	85.1	78.6	88.1	93.2	74.0	59.6	77.5
		GloVe + char	350	93.1	80.7	69.8	84.4	91.8	69.3	55.6	74.8
		GloVe	300	92.2	77.0	61.7	81.5	89.6	62.8	38.5	68.1
OntoNotes	**BiLSTM-CRF**	BERT	4096					96.9	88.6	81.1	**93.5**
		ELMo	1024					97.1	88.0	79.9	93.4
		Flair	4096					96.7	85.8	75.0	92.1
		GloVe + char	350					96.3	83.3	69.9	91.0
		GloVe	300					96.2	82.9	63.8	90.4

Table 4.4: Scores micro-F1 séparés par degré de recoupement en évaluation intra et extra domaine. Nos résultats sont obtenus en moyennant cinq entraînements.

Nous proposons donc une évaluation extra domaine en entraînant les modèles sur CoNLL03 et en les testant sur OntoNotes, plus grand et plus diversifié, ce qui correspond mieux au cas concret. Nous gardons les types de CoNLL03 et y alignons ceux d'OntoNotes : ORG et PER correspondent déjà et nous alignons LOC + GPE dans OntoNotes à LOC dans CoNLL et NORP + LANGUAGE à MISC. Cela réduit le recoupement exact à 42%, ce qui nous semble encore une surestimation du recoupement en utilisation réelle.

Performances Intra Domaine Tout d'abord, dans toutes les configurations le score F1 est le plus haut pour les recoupements exacts, puis partiels et nuls ce qui confirme le biais dans les jeux de données avec un recoupement lexical important. Ensuite, bien que ELMo apparaît comme la solution la plus stable intra domaine, il est difficile de dégager une hiérarchie claire entre plongements contextuels puisque les données de pré-entraînement ainsi que la dimension des représentations diffèrent. Pour BERT et Flair, le BiLSTM-CRF performe relativement moins bien sur CoNLL03 que sur OntoNotes, probablement par surapprentissage sur CoNLL03. De plus, le gain maximal de la contextualisation sur CoNLL03 est de +0.6 F1 en recoupement exact contre +3.7 en partiel et +2.7 en nul. D'autre part, Map-CRF avec ELMo ou Flair arrive presque au même niveau que BiLSTM-CRF et Glove + char, ce qui montre que les modèles de langues capturent intrinsèquement des représentations utiles à la REN. Enfin, quelle que soit la représentation le BiLSTM réduit l'écart de performance entre mentions vues et non vues.

Généralisation Extra Domaine En évaluation extra domaine, les performances se dégradent et l'écart se creuse entre les mentions vues et non vues. De plus, la contextualisation est encore plus bénéfique aux mentions non vues avec +1.2 F1 en recoupement exact, +9.0 en partiel et +13.1 en nul avec le BiLSTM-CRF et BERT. Cette amélioration provient clairement du pré-entraînement du modèle de langue puisque même avec Map-CRF, les plongements contextuels atteignent au moins 77.5 F1 contre 77.9 pour BiLSTM-CRF et GloVe + char. Nous distinguons néanmoins une séparation entre plongements contextuels puisque Flair, issu d'un modèle de langue à l'échelle des caractères, généralise moins bien que ELMo ou BERT en extra domaine

pour les deux modèles. Il ressort ainsi que contextualiser des mots ou sous-mots conduit à une meilleure généralisation en REN.

Conclusion Les benchmarks actuels de REN sont donc biaisés en faveur des mentions déjà rencontrées, à l'exact opposé des applications concrètes. D'où la nécessité de séparer les performances par degré de recoupement des mentions pour mieux évaluer les capacités de généralisation. Dans ce cadre, les plongements contextuels bénéficient plus significativement aux mentions non rencontrées, d'autant plus en extra domaine.

4.2.2 Modèles d'extraction de relation de bout en bout

Si l'extraction d'entités nommées est une tâche à la fois difficile et très importante pour l'indexation des connaissances, le but est aussi d'identifier les relations entre les entités. L'analyse conjointe des entités et des relations s'appelle l'extraction de relation de bout en bout.

Dans ce domaine, toujours très concurrentiel, notre état de l'art a montré qu'un certain nombre de comparaisons dans la littérature étaient inexactes [Tai+20b]. Il y a notamment des confusions au niveau des métriques, de l'usage ou non des ensembles de validation dans l'apprentissage, de la publication des statistiques de résultats ou simplement du meilleur modèle. Nous avons également constaté que certaines expériences avaient été conduites sur des jeux de données altérés.

L'état de l'art réalisé dans le tableau 4.5 pointe donc des erreurs à éviter mais il met surtout en évidence la difficulté de reproduire les expériences dans un domaine scientifique foisonnant. Dans le même temps, l'évolution récente des performances souligne des améliorations significatives. La question qui se pose avant d'aller vers des contributions plus importantes est de quantifier l'apport de différentes propositions récentes dans un cadre unifié. Nous avons réalisé ces expériences d'ablation sur les modèles de langues et l'approche span, qui permet de prendre en compte des entités se superposant. Nous avons également quantifié les avantages indus liés à la confusion dans les métriques et/ou à l'usage des données de validation pour apprendre nos modèles. Les résultats sont rassemblés dans le tableau 4.6.

Les expériences montrent l'impact positif fort des modèles de langues pré-entrainés et notamment de BERT. A l'inverse les modélisations span, si elles sont utiles pour traiter de nouvelles bases de données spécifiques, sont inutiles sur les données ne présentant pas de superpositions. Les différentes erreurs de la littérature sont bien de nature à fausser les comparaisons avec l'état de l'art.

Conclusion Ces expériences très riches constituent une base solide pour construire de nouvelles approches ; elles permettent également de mettre en avant des hypothèses de travail pour améliorer les performances en extraction de connaissances. Nous citerons par exemple la mise en place de supervision distante et de transfert à l'intérieur de la tâche d'extraction pour mieux valoriser les jeux de données hétérogènes existants. Le fait de mieux prendre en compte les relations prédites dans la phase d'extraction des entités est également une option intéressante alors qu'actuellement, la plupart des approches traitent les entités en premier.

En parallèle, nous travaillons aussi avec Benjamin Piwowarski et Etienne Simon sur l'extraction de relations non supervisée [Sim+19b ; Sim+19a]. Dans l'optique d'un apprentissage à partir de sources hétérogènes, il est essentiel de comprendre la topologie de l'espace de représentation des relations indépendamment de la

Reference	ACE 05 Ent	ACE 05 Rel	ACE 04 Ent	ACE 04 Rel	CoNLL04 Ent	CoNLL04 Rel	ADE Ent	ADE Rel	SciERC Ent	SciERC Rel	
Strict Evaluation	μF1		μF1		μF1		MF1		μF1		
[Gio+19]	**87.2**	58.6†	**87.6**	**54.0**†	89.5	66.8†	**89.6**	**85.8**			+
[EU20]	-	-	-	-	**88.9**	**71.5**†	89.3	79.2†			-
[DAO19]	86.0	**62.8**	-	-	-	-	-	-			-
[Li+19]	84.8	60.2	83.6	49.4	87.8	68.9*	-	-			+
[Sun+18]	83.6	59.6	-	-	-	-	-	-			+
[Bek+18a]	-	-	81.6	47.5†	83.6	62.0†	86.7	75.5			+
[Bek+18b]	-	-	81.2	47.1†	83.9	62.0†	86.4	74.6			+
[Zha+17]	83.6	57.5	-	-	85.6	67.8*	-	-			-
[Li+17a]	-	-	-	-	-	-	84.6	71.4			+
[KC17]	82.6	53.6	79.6	45.7	-	-	-	-			-
[Li+16]	-	-	-	-	-	-	79.5	63.4			-
[MB16]	83.4	55.6	81.8	48.4	-	-	-	-			-
[MS14]	-	-	-	-	80.7	61.0*	-	-			-
[LJ14]	80.8	49.5	79.7	45.3	-	-	-	-			-
Boundaries Evaluation											
[EU20]	-	-	-	-					**70.3**	**50.8**†	-
[Wad+19] ✗	**88.6**	**63.4**	-	-					67.5	48.4	+
[Lua+19] ✗	88.4	63.2	**87.4**	**59.7**					65.2	41.6	+
[Lua+18]	-	-	-	-					64.2	39.3	+
[Zhe+17] ✗	-	52.1	-	-					-	-	-
[LJ14]	80.8	52.1	79.7	48.3					-	-	-
Relaxed Evaluation					MF1						
[NV19] ✗					**93.8**	**69.6**					-
[Bek+18a]					93.0	68.0†					+
[Bek+18b]					93.3	67.0†					+
[AS17]					82.1	62.5					-
[Gup+16]					92.4	**69.9**†					-
Not Comparable											
[San+19] ✗	85.5	60.5									-

Table 4.5: Résumé des contributions récentes en extraction de relation de bout en bout sur 5 jeux de données.
* = partitions issues de [MS14]. † = Usage explicite train+dev. + = Expériences sur d'autres données.
✗= résultats erronés (confusion sur les métriques). Les approches au-dessus des lignes en pointillés exploitent des modèles de langues pré-entrainés.

			μF1											
			CoNLL04						ACE05					
			NER		RE (S)		RE (B)		NER		RE (S)		RE (B)	
			Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
BERT	Span	train	$85.2_{1.9}$	$86.5_{1.4}$	$69.5_{1.9}$	$\mathbf{67.8}_{.6}$	$69.6_{2.0}$	$\mathbf{68.0}_{.5}$	$84.6_{.6}$	$86.2_{.4}$	$\mathbf{60.1}_{1.9}$	$\mathbf{59.6}_{1.0}$	$\mathbf{63.2}_{.9}$	$\mathbf{62.9}_{1.2}$
		+dev	-	$87.5_{.8}$	-	$\mathbf{70.1}_{1.2}$	-	$\mathbf{70.4}_{1.2}$	-	$86.5_{.4}$	-	$\mathbf{61.2}_{1.3}$	-	$\mathbf{64.2}_{1.3}$
	Seq	train	$\mathbf{86.4}_{1.8}$	$\mathbf{87.4}_{.8}$	$\mathbf{71.0}_{1.9}$	$\mathbf{68.3}_{1.9}$	$\mathbf{71.1}_{1.7}$	$\mathbf{68.5}_{1.8}$	$\mathbf{85.7}_{.2}$	$\mathbf{87.0}_{.3}$	$\mathbf{60.1}_{.8}$	$\mathbf{59.7}_{1.1}$	$62.6_{1.9}$	$\mathbf{62.9}_{1.2}$
		+dev	-	$\mathbf{88.9}_{0.6}$	-	$\mathbf{70.0}_{1.2}$	-	$\mathbf{70.2}_{1.2}$	-	$\mathbf{87.4}_{.3}$	-	$\mathbf{61.2}_{1.1}$	-	$\mathbf{64.4}_{1.6}$
BiLSTM	Span	train	$79.8_{1.8}$	$80.3_{1.2}$	$61.0_{1.5}$	$56.1_{1.4}$	$61.2_{1.5}$	$56.4_{1.4}$	$80.0_{.2}$	$81.3_{.4}$	$46.5_{.8}$	$49.4_{1.3}$	$49.3_{.9}$	$51.9_{1.3}$
		+dev	-	$82.7_{1.2}$	-	$58.2_{1.5}$	-	$58.5_{1.6}$	-	$82.2_{.3}$	-	$49.3_{.2}$	-	$51.9_{.6}$
	Seq	train	$80.5_{.7}$	$82.0_{.3}$	$62.8_{.6}$	$60.6_{1.9}$	$63.3_{.9}$	$60.7_{1.8}$	$80.8_{.5}$	$82.5_{.4}$	$47.2_{.5}$	$50.3_{1.4}$	$49.3_{.5}$	$52.8_{1.4}$
		+dev	-	$82.6_{.9}$	-	$61.6_{1.8}$	-	$61.7_{1.6}$	-	$82.8_{.2}$	-	$50.1_{1.4}$	-	$52.9_{1.6}$

Table 4.6: Double ablation study of BERT and Span-level NER. We report the average of five runs and their standard deviation in subscript. For RE we consider both the Strict and Boundaries settings, RE Strict score is used as the criterion for early stopping.

supervision : ces travaux sont donc parfaitement complémentaires vis à vis de nos perspectives à court et moyen termes.

4.3 *Brain reading & 0-shot learning*

Les classifieurs automatiques sont des outils puissants très intéressants à étudier. Ils ouvrent des perspectives dans de nombreux domaines, notamment en classification de sentiments –pour la fouille d'opinion, les sondages, la détection de buzz– ou dans l'analyse des signaux EEG –pour la compréhension du fonctionnement du cerveau–; nous avons choisi ces domaines applicatifs pour étudier les propriétés des classifieurs dans un cadre particulièrement exigeant.

Les signaux EEG sont très exigeants au niveau du bruit, avec un rapport signal sur bruit (SNR) très faible, tandis que les données textuelles, codées en sacs de mots, sont en très grandes dimensions –5k à 100k dimensions–, avec beaucoup de variables redondantes et une parcimonie difficile à gérer. Dans les deux cas, nous nous sommes intéressés au cadre du transfert, c'est à dire au traitement d'individus jamais vus auparavant.

Le formalisme du *0-shot learning* va plus loin : il s'agit de d'intégrer en test de nouvelles classes, des concepts qui n'ont pas été modélisés dans la phase d'apprentissage. Cette opération est rendue possible par l'existence d'une sémantique au niveau des classes. La difficulté est donc double : d'un part, apprendre une sémantique efficace entre les classes ; d'autre part, construire des classifieurs suffisamment robustes pour mettre en évidence des concepts non encore croisés.

4.3.1 Brain-reading

Nous avons travaillé à prédire la classe du concept visualisé par un patient à partir d'une image fMRI de son cerveau –*functional Magnetic Resonance Imaging*–; cette tâche est ambitieusement dénommée *brain-reading*. Les images fMRI contiennent environ $20\,000$ voxels (volumantic pixels) qui décrivent l'activité cérébrale locale [Pol08]. Cette technique d'imagerie a permis d'associer des concepts avec des activités cérébrales et d'identifier des motifs caractéristiques [Kay+08 ; Har+07]. Les travaux précurseurs de [Mit+08] ont démontré l'intérêt des techniques d'apprentissage automatique pour le décodage de ces signaux.

Dans la foulée, [Pal+09] ont démontré la possibilité d'identifier un concept nouveau dans le cadre *0-shot* en décrivant tous les concepts dans un espace explicite de 218 dimensions binaires explicites correspondant à des questions telles que *is it manmade ?* ou *can you hold it ?*. La Figure 4.9 présentant notre contribution permet de clarifier la chaine de traitement utilisée. Notre contribution a consisté à tenter de se passer de cet espace de médiation explicite couteux et intrinsèquement lié au (petit) nombre de concepts du jeu de données [6].

4.3.2 Positionnement et modèle

Les expériences préliminaires montrent la qualité de la sémantique explicite utilisée et la difficulté à la substituer par des sémantiques apprises ou tirées de vastes ressources linguistiques telles que *WordNet* [LC98]. Notre idée consiste à exploiter les avancées récentes en traitement du langage comme word2vec [Mik+13] [7] et à combiner différentes propositions pour dépasser ce verrou. Notre modèle s'apparente au modèle d'origine de [Pal+09], mais en combinant différents espaces de représentation intermédiaires, comme le montre la Figure 4.9.

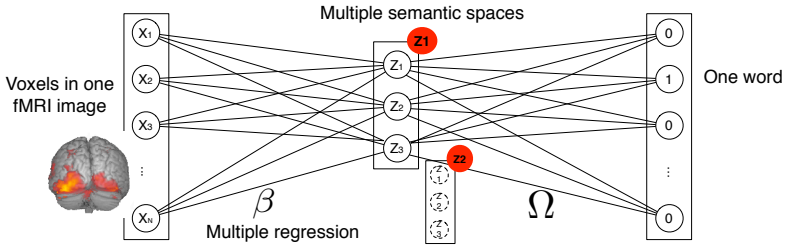


Figure 4.9: Chaine de traitement du *brain-reading* avec un espace sémantique intermédiaire pour pouvoir travailler en *0-shot*.

En terme de formulation, nous nous trouvons donc dans un système à double fonction de projection : coté étiquette –concept–, la fonction Ω permet de passer d'un mot w à un vecteur continu multi-dimensionnel $\Omega(w) = \mathbf{z} \in \mathbb{R}^p$; coté fMRI, la fonction β permet de passer d'un ensemble de voxels $\mathbf{x} \in \mathbb{R}^d$ à la représentation intermédiaire $\mathbf{z} \in \mathbb{R}^p$. Si la sémantique Ω est fixée, le problème d'apprentissage devient :

$$\underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2} \sum_{i=1}^{N} \|\mathbf{z}_i - \mathbf{x}_i\beta\|^2 + \lambda \sum_{j=1}^{p} \|\beta_j\|_\infty \right) \text{ with } \|\beta_j\|_\infty = \max_\ell |\beta_{\ell j}| \tag{4.6}$$

Pour β, nous avons choisi d'explorer l'espace des transformations linéaires en exploitant une régularisation robuste –$\mathcal{L}1$– par blocs –correspondant aux voxels–. Cette variante du LASSO est décrite dans [Liu+09] et résolue par un algorithme standard de descente de coordonnées. Dans la littérature, cette formulation de regression multi-variée est généralement associée à des problématiques multi-tâches ; ici, chaque tâche est simplement une dimension de la sémantique. La notion de bloc permet d'éliminer les voxels qui ne contribuent à aucune des dimensions. Cette formulation est nommée MTL dans la suite. Comme référence, nous avons aussi utilisé une simple régression Ridge régularisée L2.

6. Les données rendues publiques avec l'article [Mit+08] comptent 60 concepts.
7. Nous détaillons notre travail sur les espaces sémantiques dans le chapitre suivant. Les espaces sont utilisés sans modification par rapport aux implémentations de référence.

Après avoir appris K modèles, correspondant aux K sémantiques, nous calculons la similarité entre la sortie associée à une activité cérébrale et la représentation des 60 concepts de la base :

$$sim(\mathbf{x}, w) = \sum_{k=1}^{K} \lambda_k \frac{\langle \mathbf{x}\beta^{(k)}, \Omega^{(k)}(w) \rangle}{\|\mathbf{x}\beta^{(k)}\| \, \|\Omega^{(k)}(w)\|} \text{ s.t. } \sum_k \lambda_k = 1 \tag{4.7}$$

Nous voyons ici que la fusion des contributions des différentes sémantiques est un simple opérateur linéaire. La classe choisie est la classe présentant la plus grande similarité avec la projection de l'activité cérébrale courante.

En *0-shot* learning, 58 concepts (classes) sont utilisés en apprentissage, pour optimiser Ω au sens de la formulation (4.6). Les résultats sont calculés, en taux de bonne classification, sur les 2 concepts non vus en apprentissage, en exploitant la similarité (4.7). Les classes étant équilibrées, 50% de bonne classification correspond à un résultat aléatoire. Les concepts appartiennent à 12 grandes catégories : il est généralement plus facile de distinguer des concepts appartenant à des catégories différentes.

4.3.3 Expériences, résultats

Nous avons étudié la combinaison de trois espaces sémantiques, basés sur des constructions très différentes.

WordNet (WN) WordNet est un vaste lexique dans lequel les mots sont reliés (synonymie, inclusion, etc...). Ce graphe de mots permet de définir des distances [Res95 ; LC98] ; nous avons exploité ces distances pour construire un espace de dimension p de la manière suivante : chaque dimension correspond à la distance d'un mot aux p mots les plus commun de wikipedia. Il s'agit de reprendre l'idée de mots pivots issus de [Bli+07] pour construire un espace robuste. Plusieurs métriques ont été envisagées à l'intérieur de WordNet : la distance au plus proche ancêtre commun (anc), la comparaison des sous-arbres des concepts (mi) ou le plus court chemin dans le graphe (path).

Word2Vec (W2V) Comme nous le verrons plus en détail dans le chapitre suivant, word2vec est le premier algorithme d'une nouvelle famille basée sur une analyse locale des textes [Mik+13]. Ces algorithmes donnent des résultats novateurs en matière de sémantique et nous espérons pouvoir en tirer profit dans cette application.

Human218 (H_{218}) Ce dernier espace est celui, très efficace et construit manuellement par crowdsourcing, utilisé dans l'article de référence [Pal+09]. Pour chaque concept, un ensemble de volontaires a répondu à 218 questions telles que *is it manmade ?* ou *can you hold it ?*.

Nous avons tout d'abord travaillé sur la dimension de l'espace latent en utilisant les 2000 voxels les plus stables ; comme le montre la Figure 4.10, un espace de dimension 150 correspond à un compromis intéressant, quelque soit la sémantique considérée. Nous remarquons également que le LASSO en bloc dépasse la régression Ridge dans tous les cas de figure.

Nous avons donc sélectionné le LASSO et la dimension de l'espace sémantique et fait varier le nombre de voxels en entrée (sélectionnés selon le critère de stabilité de [Mit+08]) et combiné les espaces sémantiques pour étudier les performances par

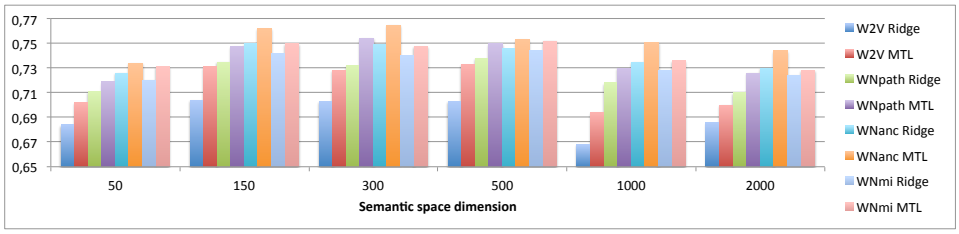


Figure 4.10: Taux de bonne classification (zero-shot learning) pour les trois espaces sémantiques considérés en fonction de leurs dimensions et des stratégies d'apprentissage (MTL = Multitask LASSO/ Ridge = RégressionRidge)

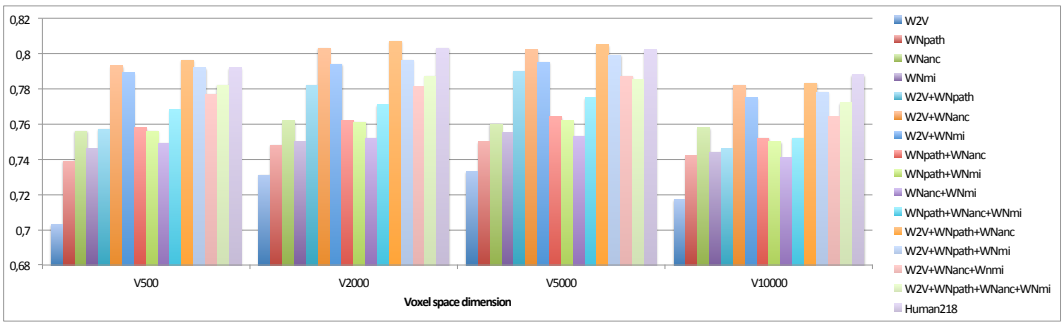


Figure 4.11: Taux de bonne classification (zero-shot learning) avec le LASSO par rapport aux nombres de voxels retenus et aux combinaisons d'espaces sémantiques considérées.

rapport au système de référence basé sur Human218. Tous les résultats sont décrits en Figure 4.11.

Combiner les espaces sémantiques permet de dépasser l'état de l'art qui reposait jusqu'ici sur Human218 (80.3% de bonne classification). Le meilleur espace sémantique appris ($WNanc$) permet d'atteindre 76.2% ; combiner deux espaces sémantiques (très différents) provoque une nette hausse de performance (80.3% de bonne classification). Passer à trois espaces sémantiques nous permet d'atteindre 80.7%.

Les nouveaux algorithmes d'analyse sémantique –combinés à d'autres ressources– permettent bien d'égaler et de dépasser les performances d'un système basé sur un espace ad hoc, très performant mais couteux et non transférable vers de nouvelles classes de concepts.

4.3.4 Discussion

Ce cadre de classification de données non textuelles fait intervenir beaucoup de concepts qui sont centraux dans notre vision de l'apprentissage statistique :

— la gestion de données très bruitées en entrée et la nécessité de réduire la dimension du problème d'apprentissage,

— la projection des données dans un espace de représentation intermédiaire permettant une meilleure compréhension de ces données,

— l'apprentissage de la sémantique contenue dans cet espace de représentation.

En plus de ces aspects, le *0-shot* est une manière très originale d'aborder le problème de classification en repoussant encore les limites de ce que nous pouvons envisager en apprentissage automatique. Une telle approche requiert l'exploitation conjointe de plusieurs modèles.

Réussir à rendre cette chaîne de traitement plus autonome, en minimisant les ressources ad'hoc est une grande satisfaction. C'est aussi la démonstration que les données textuelles peuvent être utiles dans presque tous les domaines applicatifs.

Ces travaux datent de la première révolution du deep-learning en texte, concernant la représentation des mots. Les espaces de représentations contextualisées des nouvelles approches pourraient nous aider à comprendre les associations d'idées du patient. Les espaces vectoriels étant mieux structurés, ils nous permettraient sans doute enfin de rattraper les performances obtenues avec Human218. Nous comptons réaliser ces expériences dans les années qui viennent.

4.4 Analyse et prédiction des séries temporelles

La prédiction des séries temporelles et leurs analyses, classifications, catégorisations est un domaine scientifique à part entière, à la frontière entre les communautés du traitement du signal et du machine learning. Les applications sont très nombreuses : des signaux médicaux à la maintenance prédictive, de l'analyse du comportement d'un conducteur à la gestion d'une flotte dans un système de transport intelligent. Le développement d'une ville –plus– intelligente passe aussi par la multiplication des capteurs sur les compteurs d'eau ou d'électricité, sur les usagers des transports en commun, sur le mobilier urbain ou les véhicules publics pour mesurer des flux ou de la pollution.

Nous avons séparé cette section du *brain-reading* où l'aspect modélisation des connaissances et données bi-modales éloignait la problématique de notre centre d'intérêt présent : l'analyse et la compréhension des signaux mono ou multi-variés et de leur dynamique temporelle.

Catégories de problèmes Les problèmes sont de natures diverses : le plus classique est sans doute la prédiction des valeurs futures à partir du passé [Box68]. La classification de signaux est importante, par exemple dans le domaine médical, pour reconnaitre les symptômes associés à des pathologies [RG08].

Les approches non-supervisées sont largement sollicitées dans les séries temporelles, d'une part pour catégoriser des ensembles de signaux [Pet+11] mais aussi dans le cadre de la détection d'anomalies [Ton+18b]. L'enjeu est alors de caractériser la normalité d'un phénomène puis de définir une métrique robuste pour détecter les écarts significatifs : la médecine et surtout la maintenance prédictive exploite largement ce paradigme.

Dernière catégorie mêlant d'une certaine manière les deux premières : la segmentation. L'idée est alors de classer –ou au moins distinguer– des portions de signaux de tailles variables, ce qui est souvent fait en détectant des changements dans la dynamique, c'est à dire une forme d'anomalie locale.

Spécificité des séries temporelles Nous avons opté pour une présentation séparée de ces applications car les séries temporelles constituent un matériau très particulier. Tout d'abord, les aspects temporels et la dynamique sont très importants et la supervision est rare –un petit peu comme dans les données textuelles–. La comparaison avec les données textuelles s'arrête cependant rapidement : alors que le texte peut être discretisé en lettres ou en mots, les séries sont continues, parfois très bruitées et souvent multi-variées.

En terme de vocabulaire, il est essentiel d'intégrer la notion de stationnarité (ou au contraire de non-stationnarité) qui est à la base de nombreuses hypothèses en traitement du signal. Un signal stationnaire voit ses statistiques demeurer constantes sur une période de temps : moyenne, écart-type, fréquence ou plage de fréquences... Ainsi, la plupart des problèmes précédents peuvent se ramener à une estimation locale des statistiques du signal ou une recherche de modification de ces statistiques.

Avancées récentes et perspectives à court terme L'amélioration des architectures récurrentes type RNN, GRU ou LSTM bénéficie directement à l'analyse des séries temporelles. La flexibilité des approches permet de prendre en compte une supervision locale et/ou une supervision globale de la même manière que pour les données textuelles –supervision au niveau des mots en NER, supervision au niveau de la phrase ou du paragraphe en traduction ou classification d'opinion–. Les architectures à portes (GRU ou LSTM) permettent de conserver une mémoire pertinente sur 10 à 50 pas de temps, ce qui représente une avancée considérable par rapport aux chaines de Markov [HS97].

Cependant, comme pour les données textuelles, la principale contribution des architectures récurrentes est leur capacité à apprendre à re-générer des entrées. Ainsi, le système va modéliser la dynamique des observations avant d'optimiser plus finement la tâche visée. Encore une fois, l'apprentissage de bout en bout, combiné à la multiplication des fonctions de coût sur ces approches flexibles, sont la clé pour franchir un palier à la fois sur les performances et l'interprétabilité des décisions.

La limite actuelle des systèmes *end-to-end* réside dans l'absence de modèle de langue pour les signaux. Les modèles TAL pré-entrainés sur wikipedia améliorent les performances d'analyse des réseaux sociaux qui correspondent pourtant à une grammaire et une syntaxe très différentes [Xu+19] ; les systèmes de reconnaissances d'objets dans les images sont plus performants pour détecter des défauts de fabrication sur des plaquettes de frein s'ils ont appris à distinguer les éléphants des pingouins [Kor+19]. Pour les séries temporelles, ces expériences n'ont pas encore été menées de manière concluantes.

Contribution Nous avons déjà abordé la classification, la catégorisation et la détection d'anomalies dans les signaux dans le chapitre précédent [Pou+14c ; Ton+18b] mais avec des modélisations plus classiques.

Nous nous focalisons ici sur l'apprentissage de modèles génératifs de bout en bout avec le but d'identifier les facteurs de contexte qui expliquent la forme du signal observé : le profil du conducteur correspondant à une trace GPS, le jour de la semaine associé à des logs dans les transports en commun, la météo à l'origine des pics de pollution observés sur des zones urbaines [Gui+19 ; CD+20b ; CD+20a].

4.4.1 Modélisation générative et facteurs de contexte explicite

Les modélisations possibles pour le contexte explicite d'une série temporelle sont multiples. Soit un ensemble de séries $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, chaque signal $\mathbf{x}_i \in \mathbb{R}^T$ pouvant être de longueur différente. Ces séries sont chacune associées à des facteurs de contexte $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ pouvant correspondre à des localisations, des types de jours, des conditions météorologiques etc... Afin de simplifier, nous nous limitons dans un premier temps à d facteurs discrets tous observés : $\mathbf{c}_i \in \mathbb{N}^d, c_{ij} \in \{0, \dots, K_i\}$.

Ce formalisme permet d'intégrer élégamment les séries multi-variées qui correspondent simplement à l'ensemble des signaux associés à un sous-ensemble de facteurs de contexte constant.

Famille de modèles et expériences préliminaires Les trois grandes familles de modèles que nous avons envisagées sont les suivantes (Figure 4.12) :

Uni. un modèle par contexte, afin de saisir les spécificités de chaque cas de figure : l'approche est simple mais la base de données est divisée, ce qui peut poser des problèmes notamment si les signaux sont peu nombreux ou si les contextes sont très variés.

Multi. Un modèle par groupe de contextes : le modèle travaille alors sur des séries multi-variées et prédit des valeurs multiples.

Context. Un modèle unique recevant en entrée une représentation du contexte, à apprendre sur les exemples fournis, en plus de la série elle-même.

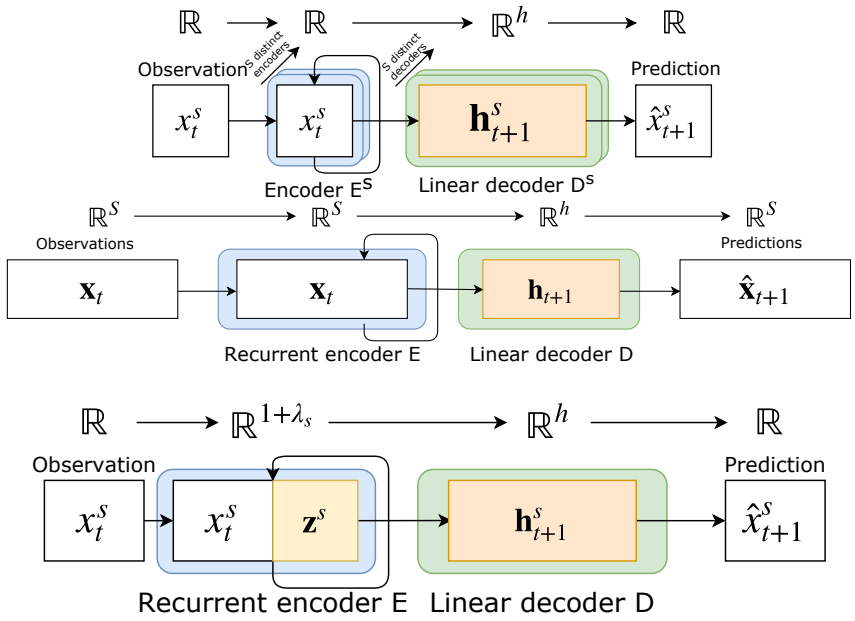


Figure 4.12: Trois familles de modèles pour gérer le contexte associés à des séries temporelles.

Les expériences conduites sur la prédiction d'affluence dans les transports en commun avec la prise en compte du facteur *station* (i.e. une localisation discrète) sont présentées dans le tableau 4.7. Le modèle de référence est celui des experts métiers qui consiste à agréger toutes les données de l'historique concernant une station et un jour donné : il s'agit d'une approche très efficace. En effet, les signaux sont quasi-stationnaires à contexte constant : une simple agrégation constitue donc une très bonne approximation. L'agrégation permet aussi de supprimer le bruit quotidien des données.

Les résultats obtenus sont très prometteurs puisque notre modélisation basée sur l'apprentissage de représentation permet à la fois de réduire l'erreur de prédiction et la variance du modèle.

En étudiant les prédictions à plus long terme, où la référence sus-mentionnée est très performante, les résultats sont encore plus intéressants (Figure 4.13). Alors que

	RNN	GRU	Réf.
Multi	$28.31_{0.09}$	$27.83_{0.12}$	
Uni	$26.15_{0.08}$	$26.73_{0.18}$	31.98
Context.	$\mathbf{24.98}_{0.05}$	$\mathbf{24.96}_{0.05}$	

Table 4.7: RMSE de la prédiction des fréquentations sur des fenêtres de 15 minutes pour chacune de nos architectures et comparaison avec la référence.

les modélisations récurrentes classiques sont bonnes à court terme mais voient leurs erreurs s'envoler avec l'horizon de prédiction, nous voyons que la prise en compte des facteurs de contexte –ici non seulement la station, mais aussi le jour de la semaine et l'heure– permet d'améliorer les choses.

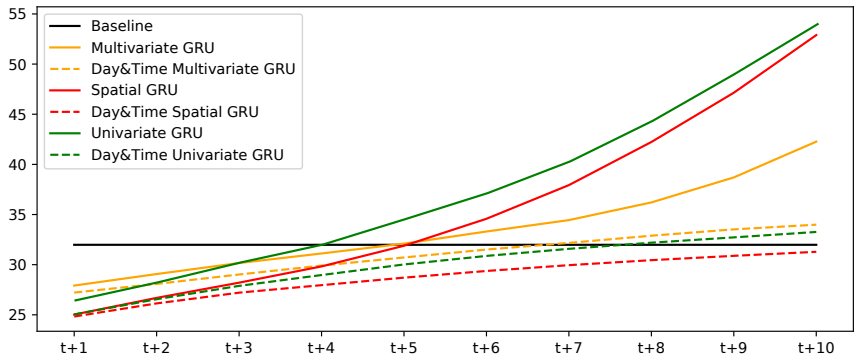


Figure 4.13: Evolution de l'erreur des différents modèles en fonction de l'horizon de prédiction.

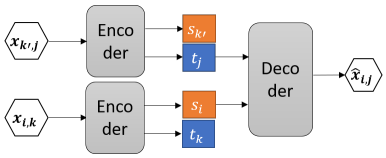


Figure 4.14: Procédure d'apprentissage de représentation démêlée : la reconstruction du signal associé aux contextes i, j est effectuée à partir de deux autres signaux i, k et k', j afin de capturer les aspects spécifiques aux contextes i et j.

4.4.2 Représentation du contexte et démêlage

L'étape suivante consiste à structurer l'espace de représentation des contextes en garantissant que les représentations n'intègrent que les aspects spécifiques à un contexte sans modéliser des phénomènes annexes [CD+20b ; CD+20a].

Nous nous reposons sur une architecture encodeur-décodeur (de bout en bout) mais nous l'apprenons de manière originale, comme le montre la figure 4.14. L'idée est de reconstruire une série cible à partir d'autres séries ne partageant qu'une partie des contextes avec la cible.

Complétion de données en filtrage collaboratif A delà de la prédiction, cette modélisation permet de prédire des séries entières pour compléter des bases de données ou estimer ce qui se passerait dans de nouveaux contextes, Figure 4.15. En effet, si l'observation de plusieurs signaux associés à un contexte permet de modéliser finement celui-ci, nous avons montré qu'un seul signal permet déjà d'avoir une estimation très intéressante de ce qui peut se passer dans toutes les combinatoires de contextes impliquant ce facteur unique.

Nous avons représenté graphiquement le problème de complétion comme du filtrage collaboratif dans le cas où deux facteurs sont appris. La gestion d'un nouveau facteur s'apparente ainsi au problème du démarrage à froid qui est connu pour être très difficile en recommandation : construire un modèle robuste à ce cas de figure est une démonstration de la robustesse de l'architecture mise en place.

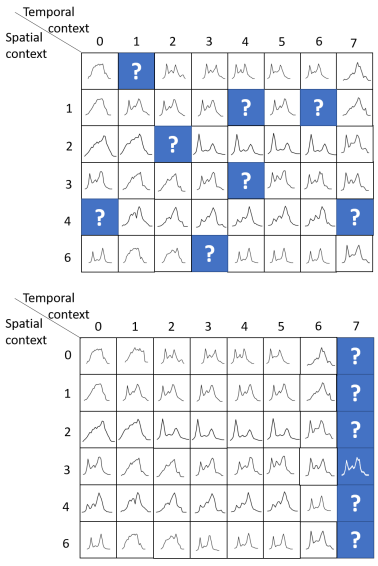


Figure 4.15: Problèmes généraux de complétion de base de données et de prédiction dans de nouvelles situations.

Performances Nous avons appliqué nos modèles sur des données d'affluence dans les transports en commun, sur la consommation d'électricité dans le nord-est des Etats Unis et sur la pollution madrilène. L'article [CD+20a] détaille les résultats en prédiction classique, en complétion de données et en estimation face à un nouveau contexte. Les tableaux ne sont pas reproduits ici mais montrent des performances de l'état de l'art dans la quasi-totalité des configurations ; cette architecture est donc efficace sur les tâches historiques du traitement du signal tout en étant capable d'aborder de nouvelles problématiques.

4.4.3 Perspectives autour des nouvelles approches de modélisation des séries temporelles

Les perspectives à explorer sont nombreuses autour de ce modèle étant donné les domaines applicatifs concernés. Au niveau de la prédiction de vente, l'enjeu est de pouvoir travailler à différentes échelles en regroupant les produits par catégories et les magasins par secteurs ; les cours de bourses constituent également un cas d'usage très difficile mais qui bénéficierait du fait que chaque action devienne associée à un contexte temporel, un secteur économique, une actualité spécifique, etc... Au niveau médical, l'idée serait de caractériser explicitement des groupes de patients voire les individus eux-mêmes, les antécédents ou le contexte saisonnier pour mieux identifier les symptômes de différentes pathologies.

Une modélisation plus fine apporte dans le même temps des opportunités en détection d'anomalies. Il sera alors possible de détecter des situations anormales mais aussi de caractériser quel ou quels facteur(s) s'écarte(nt) de la normale. La compréhension en profondeur des objets manipulés est une clé indéniable pour mieux cerner la normalité ou l'anormalité d'une situation.

En simulation, ces approches génératives sont aussi en mesure de déboucher sur des outils en rupture par rapport à l'existant. En imaginant un tableau de bord où chaque facteur prend la forme d'un potentiomètre, l'expert métier peut immédiatement visualiser les séries temporelles associées ; en imaginant maintenant un espace bi-dimensionnel où les représentations de facteur sont projetées, il devient possible de rapprocher le nouveau contexte global de tel ou tel facteur pour identifier l'impact associé.

Si la plupart des applications précédentes requièrent des bases étiquetées ou partiellement étiquetées en contexte, la question qui se pose est maintenant de découvrir de nouveaux facteurs de manière non supervisée. Des approches intermédiaires sont envisageables : si un facteur est étiqueté (par exemple le jour de la semaine dans des mesures d'affluences dans les transports en commun), nous pouvons retrouver les stations en cherchant les motifs stables dans la topologie des différentes journées à la manière de [Lam+17].

A long terme, l'enjeu est assurément d'apprendre un modèle générique associé aux signaux –ou plus vraisemblablement associé à une classe de signaux– afin de construire une base d'analyse transférable d'un problème à l'autre à la manière de ce qui se fait en texte ou en image.

4.5 Conclusion

Ce chapitre pose plusieurs questions qui feront l'objet d'un développement dans mon projet de recherche pour les années futures.

Approches génératives Les modèles génératifs ont toujours été à la base des techniques de machine learning :

— les premiers modèles d'analyse de données se sont naturellement appuyés sur des modélisations probabilistes paramétriques, souvent associées à des hypothèses naïves telles que l'indépendance des variables explicatives [MN+98]. Ces modèles étaient naturellement génératifs même si ces propriétés n'ont que peu été exploitées.

— Les années 2000 ont été marquées par l'essor des modèles graphiques qui exploitent des dépendances entre variables pour estimer la classe ou la valeur associée à une entrée [Jor98]. Les premiers modèles efficaces pour la gestion de séquences complexes en reconnaissance du texte ou de la parole sont également des modèles graphiques génératifs [Gal98].

— L'avènement du deep learning est étroitement lié à l'architecture des *denoising auto-encoders* [Vin+08], qui a rapidement donné lieu à des variantes génératives plus performantes [KW13].

L'explosion actuelle des approches *end-to-end* à base de modélisation récurrente [Pet+18] ou transformer [Dev+18] s'inscrit donc dans la continuité par rapport à l'évolution du machine learning.

L'intérêt de ces nouvelles approches génératives par rapport aux propositions plus anciennes est double :

d'une part, ces architectures sont puissantes, robustes et flexibles. Elles sont capables d'apprendre à partir de bases de données bien plus larges que dans le passé, sans saturer l'architecture tout en limitant l'impact des exemples mal étiquetés. La combinaison des fonctions de coût permet aux experts d'insérer élégamment l'ensemble des contraintes d'un domaine et de tirer parti de toutes les supervisions.

D'autre part, ces architectures sont transférables d'une problématique à l'autre ce qui n'avait jamais été envisagé –de manière systématique– jusqu'ici.

Common sense reasoning La question fondamentale qui est posée est celle de la représentation des connaissances et du bon sens ou *common sense reasoning*. L'intelligence artificielle forte requière des meta-connaissances [Pit95]. La robustesse dans les systèmes critiques tels que le véhicule autonome ou plus généralement la robotique passe par des mécanismes de bon sens permettant de prendre les bonnes décisions, d'avoir une attitude sécurisante face à des situations complètement inconnues.

Est ce que ces connaissances doivent prendre une forme explicite comme un *knowledge graph* à très grande échelle ? C'est une voie qui permet la construction de systèmes déductifs transparents et prouvables mais la phase de construction et d'apprentissage continu pose problème : ces systèmes risquent de ne pas être robustes aux bruits présents dans les données réelles et encore moins aux perturbations adverses [Mit+15].

S'il n'est pas possible de construire une base de connaissances discrètes suffisamment robuste et étendue, ne faudrait-il pas miser sur des modélisations continues ? Dans cette optique, les espaces sémantiques induits par les modèles de langues possèdent des propriétés très intéressantes : ils contiennent des informations syntaxiques, sémantiques et grammaticales, encodent de manière robuste le sens des phrases et sont même capables de désambigüiser le sens des mots polysémiques en fonction de leurs contextes [Mik+13 ; Dev+18]. Les performances ont significativement augmenté dans tous les domaines du TAL et notamment dans le *question answering*. A défaut d'être transparent ou explicite, le processus d'analyse du texte peut donc générer des réponses à des questions et être utilisé dans une logique de dialogue.

Evidemment, cette vision du monde est imparfaite, incomplète et biaisée. La combinaison de différentes modalités peut néanmoins nous permettre d'éviter certains

écueils [Zab+17]. De plus, la plasticité de ces approches rend possible l'intégration de ces *lookup tables* dans de nombreuses architectures.

Les deux grandes difficultés techniques actuellement au centre des réflexions de la communauté scientifique sont les exemples adverses, qui faussent la vision des algorithmes sans être perceptibles par les humains [Goo+14b] et les phénomènes d'oublis catastrophiques lors du passage d'une tâche à l'autre [Fre99 ; Goo+14a]. Cependant, les algorithmes évoluent vite et nous pouvons estimer que ces problèmes seront bientôt dépassés, notamment via les approches ensemblistes [Pan+19].

Abstraction faites de ces problèmes, la question sociétale est de savoir si nous pouvons nous reposer, pour des applications critiques, sur une vision du monde opaque et non prouvable.

Multi-modalité Généralement et historiquement, les approches ensemblistes fonctionnent très bien [FS97 ; CG16]. Le fait de mélanger différentes modalités de données dans les problèmes de machine learning a toujours été une source de gain de performances, ne serait-ce que par la combinaison de deux décisions basées sur des critères très différents. Les architectures de deep-learning ont changé la donne en intégrant les différentes modalités au cœur du système.

Les propositions les plus ambitieuses ont naturellement émergé à l'intersection des deux modalités les plus traitées en machine learning : le texte et l'image. La construction d'espaces latents multi-modaux a permis de traiter de nouvelles tâches comme le *captioning* [Mao+14], le *visual question answering* [Ant+15] et l'ancrage sémantique [Zab+17].

L'opportunité de fusionner différentes sources de données a permis de faire avancer les systèmes de recommandation avec des profils appris à partir de textes [Dia+19a], de signaux musicaux [DS14], d'images [Ned+16] ou de vidéo [Cov+16].

> *Le raisonnement n'est bon que dans les matières où nous n'y voyons goutte.*
> *C'est le vrai bâton de l'aveugle.*
>
> **— Joseph Joubert**

5

Conclusions scientifiques, éthiques et perspectives

Ce dernier chapitre nous permet de tirer des conclusions générales et de discuter les questions éthiques soulevées par l'évolution des technologies d'une part et des usages d'autre part. Nous présentons ensuite notre projet scientifique autour de l'apprentissage de bout en bout et de la multi-modalités.

5.1 Conclusions scientifiques

Ce manuscrit nous a permis, à travers quelques contributions, de retracer l'évolution, ces dernières années, de quelques domaines d'applications en machine learning. Partir des applications nous donne un objectif clair : optimiser les performances opérationnelles si elles sont directement intégrables dans une fonction coût ou sinon maximiser des indicateurs via la construction de fonctions de coût approchées. Cet objectif quantitatif est régulièrement associé à des aspects qualitatifs pour expliquer d'où vient la décision ou motiver une suggestion dans un système de recommandation.

Cet objectif clair correspond néanmoins à un chemin semé d'embuches. Nous énumérons ainsi les différentes étapes nécessaires pour obtenir un système performant, les écueils rencontrés et les propositions émanant de la communauté scientifique pour y faire face.

Evaluation des performances L'évaluation des performances n'est pas si simple dans l'apprentissage à partir d'exemples : au niveau théorique, les scientifiques travaillent quasi-systématiquement sur l'hypothèse d'échantillons i.i.d. –indépendants et identiquement distribués– en séparant des jeux d'apprentissage et de test. Une telle procédure est imparfaite dans de nombreux cas pratiques ; par exemple, dans les problèmes de reconnaissance d'entités nommées, les entités mentionnées à la fois en apprentissage et en test, même dans des contextes différents, posent un problème de sur-apprentissage. En classification d'opinions, les textes étiquetés –majoritairement les revues du web participatif– ne sont pas distribués selon les mêmes lois que la plupart des données de test correspondant à des applications cibles. Dans le formalisme du 0-shot learning, nous faisons en sorte que les classes des données de test soient spécifiques...

D'une part, il faut faire attention à ne pas introduire de biais dans l'évaluation ; d'autre part il faut vraiment distinguer la procédure d'apprentissage de la procédure d'évaluation. Dans cette optique, nous pouvons nous demander s'il ne faut pas consi-

dérer des tâches d'évaluation plus distantes à mesure que nos systèmes s'améliorent. Un *bon* système deviendrait alors un système quasi-parfait sur la tâche cible mais aussi très bon sur des tâches connexes.

Aujourd'hui, la majorité des problèmes sont encore spécifiques car la valeur ajoutée se trouve la plupart du temps dans les performances sur des données iid par rapport aux données d'apprentissage. Cependant, l'évaluation systématique en transfert des algorithmes d'apprentissage est clairement un moyen d'analyser le niveau de compréhension des données par le système, c'est à dire leurs capacités à généraliser à d'autres tâches.

Evolution dans l'optimisation de la robustesse La robustesse est intimement liée à l'évaluation des performances puisqu'il s'agit de ne pas perdre –ou peu– perdre en reconnaissance dans les cas défavorables (fort niveau de bruit en entrée, données de test éloignées de celles d'apprentissage, etc...). Durant la décennie des SVM ($\approx$ 1995-2005), la solution a consisté à limiter l'espace des hypothèses en régularisant et sélectionnant les caractéristiques. Ce faisant, les SVM ont en réalité éliminé une partie des informations en entrée du système (que ce soit des variables descriptives dans la version linéaire ou des échantillons dans l'espace des caractéristiques). Cette approche a démontré sa robustesse face aux données aberrantes et son pouvoir de généralisation mais aussi ses limites : la réduction de l'espace des hypothèses empêche de prendre en compte les informations présentes dans de très grandes masses de données.

A l'inverse, l'optimisation de la robustesse dans les réseaux de neurones prend des formes différentes. Il s'agit d'apprendre à éliminer le bruit dans les *de-noising auto-encoders*, d'apprendre des invariances en translation ou rotation via des stratégies de *pooling*, de combiner les fonctions de coût en génération et en discrimination pour mieux guider l'apprentissage, de s'appuyer sur des stratégies adverses pour contraindre l'espace de représentation, etc... Même la notion de régularisation prend une signification différente dans ces architectures : il ne s'agit plus d'éliminer les informations non pertinentes en entrée du système mais plutôt d'éviter des combinations de caractéristiques peu significatives au centre du système. Le rasoir d'Occam ne coupe plus la même chose.

Dans le même temps, il est intéressant de constater que les réseaux de neurones reviennent à des caractéristiques plus basiques et des espaces de représentation bruts plus concis : les démonstrations de génération de textes de A. Karpathy fonctionnent au niveau caractère où la centaine d'entrées possibles doit être mise en perspective des centaines de milliers de mots exploités en SVM. Dans les images, chaque pixel à l'entrée du réseau est composé de 3 entiers là où les chaines SVM basées sur les mots visuels requéraient jusqu'à des centaines de millers de caractéristiques pré-établies.

Les réseaux de neurones misent donc sur un espace d'entrée moins grand mais moins filtré puis sur des traitements plus complexes que les architectures précédentes. Ces conclusions sont classiques, la plupart des explications sur les performances en deep-learning étant centrées sur la capacité des réseaux de neurones à apprendre automatiquement des caractéristiques de haut niveau pertinentes à partir des données brutes. La question plus large qui se pose néanmoins est la suivante : le fait de travailler sur des données brutes élémentaires est-il une condition nécessaire pour appréhender la sémantique des données ? Les capacités des modèles de langues récents en transfert sont-elles liées au fait que ces systèmes travaillent

au niveau caractère ou sous-mot ? (en plus de la masse des données assimilées en apprentissage).

Robustesse et bio-mimétisme : une histoire d'échelle ? Historiquement, la dénomination *réseaux de neurones* a toujours correspondu à deux classes de modèles complètement distincts : les travaux issus des architectures d'Hinton et LeCun, où le réseau est un simple opérateur mathématique d'une part et les approches bio-mimétiques comme le modèle d'Hopfield d'autre part.

La première classe de modèle propose pourtant aujourd'hui d'assimiler de grandes masses de données dans différentes modalités à l'aide d'un réseau complexe reposant sur un espace latent abstrait, puis de transférer les connaissances encodées d'une application à l'autre... le processus d'apprentissage étant soumis à un ensemble de contraintes pour garantir la cohérence des informations extraites. Cette dernière définition correspond également à un système bio-mimétique, mais à une échelle différente : alors que les neurones d'Hopfield collent aux propriétés électro-chimiques des neurones biologiques au niveau microsopique, le deep learning semble plus implémenter la matrice des Robots humanoïdes d'I. Asimov à une échelle macroscopique.

Cette boucle historique est amusante : alors que le terme *réseaux de neurones* dans le cadre de l'apprentissage statistique avait une vocation plus marketing que bio-mimétique, 35 ans plus tard avec l'avènement des modèles pré-entrainés, nous entrevoyons des bribes de comportement humain dans la manière dont ces réseaux apprennent une grammaire sans règles, une sémantique sans dictionnaire, etc...

Multiplication des sources de données et modélisation des connaissances Ce manuscrit a pu donner l'impression de travaux très ou trop épars sur différents types de données d'entrées. L'évolution actuelle du machine learning montre cependant que les différentes modalités tendent à augmenter la richesse sémantique de l'espace latent appris, comme avec l'ancrage visuel des représentations de mots.

Nous avons montré que le texte et les traces utilisateurs étaient complémentaires pour une application, que les représentations de mots aidaient le *brain-reading*, que les données des réseaux sociaux permettraient bientôt de superviser l'analyse des logs dans les transports collectifs. Les étapes suivantes naturelles seraient l'extraction de connaissances multi-modales ou la création de modèles de langues dans le domaine du traitement du signal.

Ainsi, le fait de combiner des sources hétérogènes de données n'est pas un défi scientifique en soit mais un moyen d'imaginer des architectures pour améliorer la compréhension des données brutes. Les contributions à la fois diverses et remarquables de Tom Mitchell plaident pour cette hypothèse.

Limites Si les progrès en analyse de données sur les trente dernières années sont impressionnant, que dire de la dernière décennie ? Il faut néanmoins garder à l'esprit que les conclusions générales précédentes correspondent à l'état de l'art actuel en recherche et que nous risquons de rencontrer des limites opérationnelles lors de l'industrialisation de ces algorithmes.

Nous avons insisté dans les chapitres précédents sur les problèmes d'oublis catastrophiques de nos modèles lors du passage d'une tâche à l'autre ainsi que sur le risque de hacking, en particulier sur les applications critiques. Nous sommes confiants

sur le fait que ces problèmes techniques reçoivent des solutions techniques dans un avenir proche.

Les plus grandes limites sont peut-être sociétales : si les algorithmes développés dans notre communauté scientifique sont plus utilisés à des fin de limitation ou de privation des libertés individuelles qu'au développement du bien commun alors leur acceptation sera remise en cause et le financement de leur développement sera réduit. Cela nous conduit à discuter de l'éthique autour de ces algorithmes.

5.2 Questionnement éthique

Détecter une tumeur dans un scanner des poumons requière la même technologie que la surveillance des citoyens dans les rues. Cette affirmation, encore renforcée par l'exploitation grandissante du transfert de modèles entre différentes tâches, montre que nos algorithmes se définissent aussi par l'usage que nous en faisons.

Il est donc nécessaire de positionner chaque application en terme de faisabilité d'une part et d'acceptabilité d'autre part. Si le premier axe évolue chaque jour au gré des avancées technologiques, le second doit donner au citoyen le contrôle sur l'usage de ces technologies.

L'acceptation de l'IA, un problème à différents niveaux L'acceptabilité des applications et indirectement des algorithmes doit être considérée à différentes échelles. Il faut aussi distinguer deux types de régulations : les régulations comportementales (individuelles ou collectives) et les régulations législatives (nationales ou internationales).

Individu Sur certaines applications, comme les outils *Google* ou le choix des boutiques en ligne, c'est l'utilisateur individuel qui fait le choix, en fonction de ses convictions personnelles, d'utiliser (ou d'être utilisé) par les algorithmes d'IA.

Société Au niveau collectif, la population est impactée par l'analyse des comportements et les *fakenews*, elle prend conscience des risques de manipulation de type *Cambridge Analytica* et il est possible d'infléchir des politiques en limitant l'usage de *Google* dans les services publics ou en appelant au boycott du réseau *Facebook*.

Législations nationales Nous passons rapidement du niveau collectif au niveau législatif national. En effet, la révélation des dérives avérées ou possibles pousse les autorités à légiférer sur la gestion et l'exploitation des données individuelles ou le stockage des données sensibles. Les politiques varient fortement d'un pays à l'autre.

Echanges internationaux Ce niveau est encore balbutiant mais appelé à se développer dans les années à venir. Le RGPD –Règlement Général sur la Protection des Données– a été conçu au niveau européen et de plus en plus de traités diplomatiques détaillent les modalités de protection ou d'échange de données critiques entre services de renseignements ou de justices.

Vers une régulation différentiée par type d'acteur ? Il est évident que le secteur requière une régulation à la manière de la biologie ou de la physique nucléaire. La place des chercheurs est cependant à part : étudier les possibilités et les limites de nos algorithmes est nécessaire pour mieux cerner les risques ; ces risques ne sont pas les mêmes pour un démonstrateur technologique restant dans un laboratoire et sur une application à destination du public ou encore des forces de l'ordre.

A ce titre, la question de l'anonymat est révélatrice : du coté des chercheurs, le problème est souvent annexe, le but étant de cerner ce que nous pouvons savoir d'une personne à partir de ses traces. Pour le public, la question est critique afin de ne pas exposer l'intimité des uns ou des autres de manière irrémédiable. Pour les forces de l'ordre, il existe déjà beaucoup de droit d'accès à l'intimité à des fins d'enquête : la réponse doit donc encore être spécifique.

Le législateur va donc jouer un role fondamental dans les années à venir pour définir un cadre où chacun doit pouvoir bénéficier de la valeur associée à ces algorithmes sans remettre en cause les libertés individuelles et collectives dont nous jouissons. Les différents acteurs –y compris les commanditaires– doivent jouer un rôle pro-actif en faisant remonter vers le législateur les possibilités réelles et les risques associés de manière transparente. La dépendance des chercheurs aux financements sur projets –publics, privés ou militaires– peut alors être un frein important dans la boucle de retour d'expériences où les intérêts des uns et des autres vont inévitablement diverger, au moins à court terme.

5.2.1 Apprentissage des profils de personnes

De la même manière que pour les mots, les produits ou les images, nous apprenons efficacement des profils de personnes. Les implications sociétales sont évidemment très différentes.

Profiling et personnalisation de l'accès à l'information Le profiling est directement associé au modèle économique des entreprises liées à Internet, notamment autour du précepte *Si c'est gratuit, c'est vous le produit*. En effet, les outils de *Google*, *Facebook* et de bien d'autres entreprises dans leurs sillages donnent l'illusion de la gratuité. Ils sont financés par de la publicité ciblée et reposent donc sur la connaissance individuelle que l'entreprise a de l'utilisateur, elle-même valorisée par l'exploitation et la vente de ces connaissances.

Nous recevons donc des produits, régulièrement d'excellente qualité, en échange d'informations très intimes sur nos localisations, nos catégories professionnelles, nos centres d'intérêt, nos relations professionnelles ou amicales et directement ou indirectement, nos opinions. Il ne faut pas non plus négliger le fait que certaines applications modernes impose une forme de profiling : le catalogue des sites de e-commerce est bien trop grand pour permettre une navigation aléatoire dans les rayons virtuels, Internet est bien trop vaste pour se passer d'un moteur de recherche efficace... Les recherches –nombreuses– sur des algorithmes plus équitables et sur la *fairness* ont montré qu'il est possible de maintenir une qualité de service avec des profils moins intrusifs mais le chemin reste étroit.

Il faut aussi prendre du recul et ne pas s'arrêter aux traces que l'utilisateur laisse en ligne : les logs de transports en commun, de péage ou les données bancaires fournissent des profils tout aussi intrusifs et le modèle économique d'entreprises comme Alipay (ANT) montre que ces profils sont également revendus.

Toutes les études sociologiques montrent que : (1) il est illusoire de croire que nous pouvons séparer un profil *de consommateur* d'un profil *d'opinion* ; (2) nous sommes perméables aux stratégies d'orientation de l'opinion, en particulier le marketing ; (3) s'il existe des outils de manipulation de l'opinion, ils seront utilisés.

Philosophiquement, l'opinion tient une place à part (par rapport à notre catégorie d'âge ou même notre adresse personnelle). Pour les algorithmes de profiling, cette

distinction n'existe pas et c'est au régulateur d'indiquer d'une part quelle information peut être utilisée ou pas (en entrée du système) et quelle application peut être envisagée ou pas (à la sortie du système cette fois).

Profiling d'état Les forces de l'ordre ont toujours maintenu des fichiers individuels. Les sources de données et les outils actuels dévoilent une intimité jusqu'ici inédite. En croisant les empreintes digitales voire ADN, les caméras de surveillance, les transactions bancaires, les logs de transport en commun, les requêtes que font un utilisateur, le contenu de ses communications, son historique médical, etc... il est possible de dresser un profil particulièrement détaillé de celui-ci.

Les dérives, en particulier chinoises, observé ces dernières années doivent pousser les sociétés démocratiques à réguler très fortement le stockage de ces sources de données et leur croisement. Dans le même temps, la vente des outils de surveillance de masse doit également faire l'objet d'un contrôle, notamment à l'exportation où les conséquences dépendent directement des systèmes politiques et législations locales.

Néanmoins, en prenant du recul, il ne faut pas perdre de vue que la prochaine évolution significative de la médecine reposera probablement la constitution de larges bases de données de santé très détaillées –et donc intrusives– au niveau des états voire au niveau international. Le législateur devra donc faire la part des choses pour bénéficier des aspects positifs de ces technologies.

5.2.2 Sécurité, explicativité et preuve

La plupart de la valeur marchande des algorithmes actuels est localisée sur la publicité en ligne. Le coût d'une erreur est ainsi très limité. D'une manière générale, nos algorithmes sont capables de prendre des décisions plus fiables que des humains dans des contextes particulièrement difficiles... Mais ils restent des assistants sous la supervision d'un médecin ou d'un opérateur. Ainsi, nos algorithmes ne prennent actuellement pas ou peu de décisions critiques.

Passage vers des applications critiques Les choses évoluent rapidement avec le développement de véhicules plus intelligents voire autonomes : chaque décision met alors en jeu la sécurité des personnes. Contrairement aux systèmes experts, les réseaux de neurones sont des systèmes intrinsèquement non prouvables. La question devient donc : sommes-nous prêt à confier notre sécurité à des algorithmes dont le fonctionnement n'est pas borné.

Notons dans un premier temps que la question est très éloignée du dilemme du tramway, où il faut décider de tuer plutôt une personne –le chauffeur– ou un groupe qui se trouverait sur la voie de l'autre coté de l'aiguillage. Ce dilemme est fondamentalement impossible à trancher et la question de l'automatisation de cette décision met simplement en lumière des intérêts divergents inconciliables relevant plutôt de la théorie des jeux. Cet énoncé est de plus un faux problème pour le deep learning : d'une part, un algorithme fait ce pour quoi il a été supervisé et d'autre part la représentation des connaissances du réseau de neurones est probablement trop abstraite pour formaliser ce dilemme en inférence.

La vraie question est donc celle des limites et de la fiabilité de notre algorithme : (1) il doit savoir estimer la confiance dans sa décision de manière fiable, (2) il doit probablement être encadré par un système prouvable capable de reprendre le contrôle en cas de problème, (3) il doit être validé sur une masse de données suffisantes et avoir démontré sa robustesse à des tentatives de détournement. (4) Enfin, il doit être

évolutif : à la manière de tout système informatique connecté, il doit être en mesure de recevoir des correctifs pour s'améliorer au fil de l'évolution des usages. Ces quatre conditions sont évidemment plus simples à énoncer qu'à implémenter.

Pour le point (2), il est possible d'imaginer des systèmes hybrides en profondeurs où les règles sont imprimées à l'intérieur du réseau. C'est par exemple le cas des modèles hybrides en océanographie où des EDP sont au cœur du réseau et garantissent une modélisation dynamique crédible.

Ayant fait le choix d'un positionnement technique, je n'aborde pas ici la question de la responsabilité. Il est cependant évident qu'une acceptation des algorithmes d'IA dans les applications critiques passe par cette question ainsi que celle des assurances financières associées. Ces algorithmes ne seront déployés que si leur modèle économique, incluant risques et assurances, est viable.

Le débat trompeur sur l'explicativité Beaucoup d'articles, notamment à destination du grand public, créent de la confusion en se focalisant sur la notion d'explicativité sans distinguer les notions de preuves a priori des explications post hoc. Les réseaux ne sont pas prouvables car il n'est pas possible de valider les sorties associées à toutes les entrées possibles, ce qui pose les problèmes que nous avons listés précédemment.

Par contre, les décisions prises par les réseaux sont tout à fait explicables a posteriori : nous savons retrouver les éléments qui poussent une image dans une classe, les mots qui font qu'une partie de la phrase est détectée comme une entité nommée, les points qui entrainent la classification d'un signal comme anormal. Nous avons su très vite –mais malheureusement a posteriori– pourquoi le véhicule automatique Uber avait renversé une piétonne en 2018.

Le grand public a actuellement du mal à faire la part des choses et il est du ressort de la communauté scientifique de communiquer pour éduquer la société sur ces algorithmes qui prennent une place de plus en plus importante dans le quotidien des citoyens.

5.2.3 Régulation, éducation et société

Ces nouvelles technologies engendrent des modifications en profondeur dans la société. Ces mutations entrainent des risques que nous discutons ici.

Une partie importante des problèmes liés aux algorithmes d'analyse de données s'explique par le manque de recul du public et des pouvoirs politiques par rapport à ces nouvelles technologies qui ont connu une expansion particulièrement rapide.

Législation *vs* éducation A ce titre, la comparaison avec l'essor de Wikipedia est intéressante. A l'aube des années 2000, certains acteurs se sont sentis dépassés par cette source d'informations à la fois gigantesque mais souvent non vérifiées et par essence pas complètement vérifiables. Les plus alarmistes ont crié à la mort du système éducatif puisque l'ensemble du savoir était maintenant accessible à tous en un clic.

La suite des évènements a montré que c'était exactement l'inverse : Wikipedia est un formidable outil pédagogique à condition de savoir s'en servir. Il vient donc parfaitement compléter les enseignements de l'école si le système éducatif forme les enfants aux avantages et limites de l'outil.

L'exemple des *Fakenews* est édifiant : il de plus en plus facile de générer un texte ou une image trompeuse. Le législateur est en parti coincé car l'intention de celui qui génère le texte ou l'image n'est pas forcément mauvaise tandis que la détection automatique du phénomène est difficile et imparfaite.

La solution est claire mais malheureusement très longue à mettre en œuvre par rapport à un phénomène qui se répand très vite : il faut éduquer la population à lire une information dans un nouveau contexte où elle peut s'avérer fausse. C'est le moyen le plus fiable pour éliminer ce type de risque tout en bénéficiant pleinement des nouvelles technologies.

Statistiques interdites Le développement de nos outils d'analyse statistique permet ou permettra à court terme de nouvelles études, notamment au niveau des ressources humaines. Ces outils peuvent permettre de donner des indicateurs quantitatifs fiables sur les discriminations homme-femme ou celles liées aux patronymes.

Les débats sur l'usage de ces statistiques vont se multiplier dans différents domaines mais la question est cependant sensible, au même titre que celle des statistiques ethniques. Faut-il utiliser cet outil pour tenter de juguler des inégalités ou faut-il craindre des détournements qui fausseraient les conclusions et, au contraire, interdire ces études statistiques.

Comme c'est le cas actuellement, les réponses seront différentes d'un pays à l'autre. Elles doivent probablement être discutées au cas par cas localement et, la population étant de plus en plus sensibilisée à ces technologies, ces réponses ne doivent pas être figées mais au contraire ré-étudiées à intervalle régulier. Nous pourrons conclure sur une citation célèbre : *Il y a trois sortes de mensonges : les petits mensonges, les gros mensonges et les statistiques. Graham McNeill*.

Militarisation La question de l'usage militaire de ces algorithmes... n'est pas vraiment une question. Les militaires financent depuis la veille de la seconde guerre mondiale une grande partie des technologies informatiques et IA : ils en sont donc les premiers utilisateurs. Dès lors, une seule question reste ouverte : qu'est ce qui est autorisé ou pas, au niveau national ou international ?

Le scénario *Terminator* est cependant très éloigné des algorithmes actuels : les plus avancées de nos applications relèvent toujours de l'IA faible, nos machines ne raisonnent pas et l'embryon de vision du monde que nous leur intégrons doit leur permettre de prendre le volant d'une voiture mais pas la tête d'une armée. Quand bien même, nos algorithmes ne font que répéter ou interpoler des scénarios fournis en apprentissage, ils ne jugent pas et n'inventent rien. *Terminator* relève en réalité d'un paradigme très différent de celui développé en ce moment.

Notons enfin un autre parallèle malheureux dressé par de nombreux journalistes entre l'arme atomique et les outils de l'IA. Cette comparaison est infondée car il existe une barrière technologique pour accéder à l'arme atomique : le processus est long, difficile, requière un matériel qu'il est difficile de cacher. A l'inverse, les outils de l'IA sont accessibles sur github et reposent sur un cadre logiciel ouvert et robuste qui permet un transfert immédiat d'une application médicale à une application de surveillance.

Transition énergétique Dernier volet de cette section, la question énergétique. Les algorithmes d'analyse de données sont à la fois fort consommateurs et potentiellement source d'économie d'échelle en énergie. En effet, les approches prédictives

permettent un meilleur usage de l'énergie produite et moins de gâchis. De nombreuses applications ou processus de fabrication peuvent bénéficier des algorithmes d'optimisation pour réduire leur consommation. La maintenance prédictive, à l'opposée de l'obsolescence programmée est un outil de plus pour limiter le renouvellement des machines.

Reste la question de la consommation des algorithmes eux-mêmes. A ce titre, il faut distinguer trois étapes distinctes : (1) la mise au point des algorithmes voire des processeurs capable de les faire tourner ; (2) l'utilisation des algorithmes pour apprendre sur des données cibles ; (3) l'exploitation des modèles appris, en inférence, sur de nouvelles données.

Le point (1) représente actuellement la principale source de dépense à l'instar du développement du gigantesque GPT-3 par openAI. Il faut cependant noter le caractère dispendieux du développement de n'importe quelle nouvelle technologie et le besoin scientifique de comprendre le fonctionnement de ces réseaux de neurones géants. En parallèle, un nombre important de recherches sont dirigées vers des modèles moins chers (*budget learning*), notamment des modèles légers capables d'imiter des architectures plus complexes (*teacher-student*). Si le point (2) est relativement incompressible, notons que le point (3) a fait des progrès spectaculaires ces dernières années sur le plan de la consommation. Les systèmes embarqués de Mobileye sont par exemple remarquables : ils tournent en temps réel sur une batterie 12V de voiture, sans surchauffe alors qu'ils implémentent des algorithmes qui nécessitaient un super calculateur quelques années plus tot.

5.3 Projet de recherche

Des pistes des recherches ont été détaillées dans les conclusions intermédiaires et parfois même à la fin des expériences décrites dans les chapitres précédents. Cette dernière section du manuscrit vise à donner une image générale et structurée des pistes que nous allons explorer dans les 3 à 5 années à venir.

Nous verrons, de manière générale, que le positionnement très appliqué de nos recherches ne change pas : il s'agit toujours de poser de nouvelles hypothèses autour d'un problème, de proposer une architecture instanciant ces hypothèses puis de démontrer l'efficacité de la proposition à travers des tests sur des données réelles.

Sources hétérogènes et modèles génératifs Dans tous les cas, ces travaux impliquent d'une manière ou d'une autre des données hétérogènes : soit différentes modalités de données, soit, a minima, différentes sources potentiellement non iid. L'apprentissage de modèles génératifs ou en partie génératifs est un autre dénominateur commun. La multiplication des fonctions de coût est aujourd'hui reconnue comme le moyen le plus efficace d'empiler des hypothèses et/ou des contraintes métiers dans le processus d'apprentissage.

Des systèmes génératifs basés sur l'apprentissage de représentation qui plus est dotés de mécanismes d'attention sont capables d'expliquer les décisions prises. Nous veillerons à développer ces propriétés pour inscrire nos recherches dans la dynamique de l'IA explicable.

Vers de nouvelles approches pour la gestion des connaissances Les modèles génératifs et même la prise en compte de données hétérogènes structurent donc notre

projet de recherche mais la problématique sous-jacente qui nous motive est celle de la gestion de connaissances.

Nous cherchons à apprendre des représentations de concepts explicites –ou explicables– et à multiplier les relations entre ces concepts. Si la problématique est très ancienne en IA symbolique, dans la communauté de l'apprentissage statistique, les premières propositions de systèmes capables de raisonner de cette manière datent du début de la décennie 2010 [Bor+11 ; Soc+13a]. Ces contributions doivent aussi être mises en perspective avec les systèmes de dialogue suivi qui modélisent la dynamique pour générer une réponse tenant compte des échanges récents [Wes+14 ; Suk+15]. L'enjeu est ensuite de multiplier les sources de données tant à l'entrée du système qu'en supervision, afin, par exemple, de générer des explications textuelles même lorsque cette modalité n'est pas présente dans les données d'entrée. Ce dernier domaine de recherche est souvent appelé *d*ata-to-text [Yan+17 ; Reb+19].

La dernière étape, la plus critique, est de transférer des connaissances d'une application à l'autre à la manière des modèles de langue. A court terme l'enjeu est l'extraction et la compréhension des espaces latents encodant ces connaissances. A moyen terme, l'enjeu est le transfert de ces connaissances au delà des applications textes/images. A long terme, il s'agit de construire une base universelle exploitable par un large ensemble d'applications.

5.3.1 Sémantique utilisateur et modélisation de la dynamique

Au niveau de la dynamique dans les systèmes de recommandation, nos travaux ont montré comment les enchainements d'items dans une trace utilisateur pouvait permettre une meilleure compréhension de ceux-ci [GS+15]. Nous abordons une nouvelle phase de ces travaux avec le début de la thèse de **Darius Afchar** fin 2020 où l'enjeu est de mêler des données de type signal –de la musique–, des successions d'interactions utilisateurs –écoutes, skips, replays, ...– et différents contextes d'écoute pour mieux comprendre les intentions de l'utilisateur et améliorer les systèmes de recommandation.

A l'interface entre les modalités textuelles et traces d'utilisateurs, nous travaillons encore sur les évènements dans les bases de CV [Dia+17a ; Gab+20]. L'idée est de modéliser à la fois la dynamique et les échelles présentes dans les données multi-instances –poste occupé, individu, service, entreprise–. Au delà du matching d'experts ou de l'identification des compatibilités CV/offres d'emploi, l'enjeu est double : (1) mieux comprendre les individus par rapport à leurs compétences, leur dynamique de carrière, leur scolarité et expliquer ces éléments les uns par rapport aux autres. (2) Obtenir une sémantique et une modélisation dynamique au niveau des agrégats pour comprendre et prédire des évènements au niveau des entreprises. A plus long terme, nous pourrions imaginer des supervisions aux différentes échelles dans un système tirant parti à la fois des CV d'individu et du cours de bourse de l'entreprise. Ce travail s'appuie sur la thèse de **Clara Gainon de Forsan**. Ces travaux nous semble d'autant plus importants qu'ils permettront aux utilisateurs de mieux comprendre comment leur CV est perçu ; les services des ressources humaines bénéficieraient également d'une meilleure vision des domaines techniques, des mots clés à bien choisir pour éviter les confusions entre domaines de compétences proches.

Toujours à la même interface, nous nous intéressons à la problématique du *review spam* dans le cadre du postdoc de **Lynda Said**. Cette problématique mêle

séries temporelles (évolution de la note d'un produit) et interactions d'utilisateurs sous la forme de notes et de textes. Ce projet nous permet d'ajouter la notion de cotation d'informations mais exploite fondamentalement les mêmes ingrédients que les problématiques précédentes : enchainements d'évènements, détection et caractérisation de ruptures dans la dynamique de ces évènements et profiling utilisateur à partir de différentes modalités de données.

5.3.2 Entités multi-modales et représentation des connaissances

La représentation des connaissances est au centre de la plupart de nos travaux. Nous revenons cependant ici à la notion de connaissances formelles type RDF, c'est à dire composées de deux entités et d'une relation.

Au niveau du texte, les modèles de langue introduits récemment ont fait évoluer l'état de l'art de manière significative. Nous avons quantifié ce phénomène pour les tâches d'extraction d'entités nommées et de relations [Tai+20a ; Tai+20b] ; nous travaillons maintenant sur la construction de nouvelles architectures plus efficaces pour la construction de bases de connaissances. A court terme, l'enjeu est de tirer parti des différentes sources étiquetées de données même si elles sont très différentes (étiquetage complet, étiquetage distant partiel, étiquetage des entités et/ou des relations...). Ce travail est mené en collaboration avec **Patrick Gallinari** et **Benjamin Piwowarski**, il s'appuie sur les thèses de **Bruno Taillé** et **Etienne Simon**.

La thèse de **Maya Sahraoui**, co-encadrée avec **Régine VigneLebbe** qui débute fin 2020 va nous permettre d'explorer la notion d'entité multimodale en s'appuyant sur les bases de données du Muséum National d'Histoire Naturelle. L'enjeu de ce projet est d'extraire les entités dans des descriptions d'espèces animales ou végétales regroupant des données textuelles, images ainsi que des bases de connaissances. La richesse de ces données doit nous permettre de faire un lien explicite entre espace de représentation multi-modales et base de connaissances. Il s'agit d'une grande opportunité pour mieux comprendre la structuration des espaces latents et consolider l'ensemble des approches sur lesquelles nous travaillons.

L'amélioration des modèles de langue génératifs a ouvert la possibilité de générer des textes crédibles. L'enjeu est maintenant de diriger et encadrer cette génération pour transmettre des informations ; cette tâche est le *data-to-text*. Dans le cadre de cette tâche, nous souhaitons également travailler sur l'extraction d'informations en utilisant le texte comme une supervision pour comprendre comment les journalistes, par exemple, ordonnent les informations qu'ils transmettent. Ce travail est effectué en collaboration avec **Laure Soulier** et soutenu par la thèse de **Ismael Bonneau** qui a débuté en octobre 2020.

5.3.3 Vers des modèles de langues en traitement du signal

En traitement du signal, les modèles génératifs permettent de mieux cerner les facteurs qui expliquent la forme d'une série temporelle [CD+20b]. La compréhension de ces facteurs est une clé pour construire de meilleurs modèles prédictifs mais aussi pour mieux caractériser la normalité et ainsi, détecter les anomalies.

Nous abordons cette problématique sous deux angles : d'une part celui du transfert. A l'instar du texte ou de l'image, la question est de savoir s'il est possible de construire une chaine universelle dédiée à tous les signaux (ou plus vraisemblable-

ment à tous les échantillons d'une classe de signaux) puis à transférer cette chaine d'application en application pour l'enrichir à la manière des modèles de langue. Le second angle d'attaque réside dans l'analyse du contexte des signaux : tout signal est capté dans un environnement (localisation, météo, utilisateur, ...) qu'il est essentiel de modéliser pour comprendre le signal.

Cette combinaison de caractères intrinsèques universels et de facteurs exogènes ouvrent de nombreuses perspectives applicatives dans le domaine des véhicules intelligents, de la santé ou de la maintenance prédictive. Nous travaillons sur ces problématiques avec **Nicolas Baskiotis** et nous sommes en train de monter plusieurs projets pour prendre la suite des travaux de thèse de **Perrine Cribier-Delande**.

5.3.4 Le mot de la fin

Nous avons la chance de travailler dans un domaine de recherche en mutation rapide où beaucoup de choses ont changé ces dernières années. Nous sommes convaincus que la gestion des connaissances, qui est centrale en IA, va encore largement évoluer dans les années à venir et notre projet de recherche doit nous permettre de jouer un role dans cette redéfinition fondamentale.

Notre domaine est particulièrement attractif ; il nous permet de bénéficier de conditions privilégiées dans le domaine de recherche, au niveau des financements, de la qualité et de la richesse des flux étudiants. Mais la mutation de la société autour de ces algorithmes nous donne aussi une grande responsabilité. A ce titre, nous nous devons de conserver une recherche libre et indépendante, constructive mais critique. Nous devons aussi mesurer l'importance de l'enseignement de l'IA à tout les niveaux pour former des futurs concepteurs réfléchis et raisonnables mais aussi des citoyens éclairés face à ces nouvelles technologies.

Bibliographie

[Aba+16] Martín ABADI, Paul BARHAM, Jianmin CHEN et al. « TensorFlow : A System for Large-Scale Machine Learning. » In : *OSDI*. T. 16. 2016, p. 265–283 (cf. p. 3, 29).

[Ant+15] Stanislaw ANTOL, Aishwarya AGRAWAL, Jiasen LU et al. « Vqa : Visual question answering ». In : *Proceedings of the IEEE international conference on computer vision*. 2015, p. 2425–2433 (cf. p. 52, 77).

[AS17] Heike ADEL et Hinrich SCHÜTZE. « Global Normalization of Convolutional Neural Networks for Joint Entity and Relation Classification ». In : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark : Association for Computational Linguistics, sept. 2017, p. 1723–1729 (cf. p. 66).

[Aug+17] Isabelle AUGENSTEIN, Leon DERCZYNSKI et Kalina BONTCHEVA. « Generalisation in named entity recognition : A quantitative analysis ». In : *Computer Speech & Language* 44 (juil. 2017), p. 61–83 (cf. p. 63).

[Bac+10] Stefano BACCIANELLA, Andrea ESULI et Fabrizio SEBASTIANI. « SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. » In : *LREC*. 2010, p. 2200–2204 (cf. p. 20).

[Bar+13a] Alexandre BARACHANT, Stéphane BONNET, Marco CONGEDO et Christian JUTTEN. « Classification of covariance matrices using a Riemannian-based kernel for BCI applications ». In : *Neurocomputing* 112.Supplement C (2013), p. 172 –178 (cf. p. 16).

[Bar+13b] Ranieri BARAGLIA, Cristina Ioana MUNTEAN, Franco Maria NARDINI et Fabrizio SILVESTRI. « LearNext : learning to predict tourists movements ». In : *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM. 2013, p. 751–756 (cf. p. 40).

[Bar+16] Alberto BARTOLI, Andrea DE LORENZO, Eric MEDVET et Fabiano TARLAO. « Your paper has been accepted, rejected, or whatever : Automatic generation of scientific paper reviews ». In : *International Conference on Availability, Reliability, and Security*. Springer. 2016, p. 19–28 (cf. p. 38).

[Bek+18a] Giannis BEKOULIS, Johannes DELEU, Thomas DEMEESTER et Chris DEVELDER. « Adversarial training for multi-context joint entity and relation extraction ». In : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium : Association for Computational Linguistics, 2018, p. 2830–2836 (cf. p. 66).

[Bek+18b] Giannis BEKOULIS, Johannes DELEU, Thomas DEMEESTER et Chris DEVELDER. « An attentive neural architecture for joint segmentation and parsing and its application to real estate ads ». In : (2018), p. 1–26 (cf. p. 66).

[BEK14] Shay BEN-ELAZAR et Noam KOENIGSTEIN. « A Hybrid Explanations Framework for Collaborative Filtering Recommender Systems. » In : *RecSys Posters*. Citeseer. 2014 (cf. p. 30).

[Ben+13] Yoshua BENGIO, Aaron COURVILLE et Pascal VINCENT. « Representation learning : A review and new perspectives ». In : *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), p. 1798–1828 (cf. p. 2, 28).

[Bes+11] Dmitriy BESPALOV, Bing BAI, Yanjun QI et Ali SHOKOUFANDEH. « Sentiment classification based on supervised latent n-gram analysis ». In : *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2011, p. 375–382 (cf. p. 24, 51).

[Bis95] C.M. BISHOP. *Neural networks for pattern recognition*. Oxford University Press, USA, 1995 (cf. p. 11).

[BK07] Robert M BELL et Yehuda KOREN. « Lessons from the Netflix prize challenge ». In : *Acm Sigkdd Explorations Newsletter* 9.2 (2007), p. 75–79 (cf. p. 2, 27, 30, 31).

[BL+07] James BENNETT, Stan LANNING et al. « The netflix prize ». In : *Proceedings of KDD cup and workshop*. T. 2007. New York, NY, USA. 2007, p. 35 (cf. p. 30).

[Bla+08] Benjamin BLANKERTZ, Ryota TOMIOKA, Steven LEMM, Motoaki KAWANABE et K-R MULLER. « Optimizing spatial filters for robust EEG single-trial analysis ». In : *IEEE Signal processing magazine* 25.1 (2008), p. 41–56 (cf. p. 17).

[Ble+03] David M BLEI, Andrew Y NG et Michael I JORDAN. « Latent dirichlet allocation ». In : *Journal of Machine Learning Research* 3.Jan (2003), p. 993–1022 (cf. p. 17).

[Bli+07] John BLITZER, Mark DREDZE, Fernando PEREIRA et al. « Biographies, bollywood, boom-boxes and blenders : Domain adaptation for sentiment classification ». In : *ACL*. T. 7. 2007, p. 440–447 (cf. p. 18, 20–22, 69).

[BM98] Paul S BRADLEY et Olvi L MANGASARIAN. « Feature selection via concave minimization and support vector machines. » In : *ICML*. T. 98. 1998, p. 82–90 (cf. p. 19).

[Boj+17] Piotr BOJANOWSKI, Edouard GRAVE, Armand JOULIN et Tomas MIKOLOV. « Enriching word vectors with subword information ». In : *Transactions of the Association for Computational Linguistics* 5 (2017), p. 135–146 (cf. p. 43, 49).

[Bor+11] Antoine BORDES, Jason WESTON, Ronan COLLOBERT, Yoshua BENGIO et al. « Learning Structured Embeddings of Knowledge Bases. » In : *AAAI*. T. 6. 1. 2011, p. 6 (cf. p. 8, 52, 62, 88).

[Bos+92] Bernhard E BOSER, Isabelle M GUYON et Vladimir N VAPNIK. « A training algorithm for optimal margin classifiers ». In : *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, p. 144–152 (cf. p. 1, 19).

[Bou+16] Simon BOURIGAULT, Sylvain LAMPRIER et Patrick GALLINARI. « Representation learning for information diffusion through social networks : an embedded cascade model ». In : *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM. 2016, p. 573–582 (cf. p. 17).

[Box68] George EP BOX et Gwilym M. « Some recent advances in forecasting and control ». In : *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 17.2 (1968), p. 91–109 (cf. p. 71).

[Bre01] Leo BREIMAN. « Random forests ». In : *Machine learning* 45.1 (2001), p. 5–32 (cf. p. 2).

[Bur07] Robin BURKE. « Hybrid web recommender systems ». In : *The adaptive web*. Springer, 2007, p. 377–408 (cf. p. 28, 31).

[BY+17] Hedi BEN-YOUNES, Rémi CADENE, Matthieu CORD et Nicolas THOME. « Mutan : Multimodal tucker fusion for visual question answering ». In : *Proc. IEEE Int. Conf. Comp. Vis*. T. 3. 2017 (cf. p. 28).

[Cam+11] P. G. CAMPOS, F. DIEZ et A. BELLOGIN. « Temporal rating habits : A valuable tool for rating discrimina- tion ». In : *CAMRa*. 2011 (cf. p. 40).

[Can+05] Stéphane CANU, Yves GRANDVALET, Vincent GUIGUE et Alain RAKOTOMAMONJY. « Svm and kernel methods matlab toolbox ». In : *Perception Systmes et Information, INSA de Rouen, Rouen, France* (2005) (cf. p. 2, 5).

[Can+08] Julián CANDIA, Marta C GONZÁLEZ, Pu WANG et al. « Uncovering individual and collective human dynamics from mobile phone records ». In : *Journal of physics A : mathematical and theoretical* 41.22 (2008), p. 224015 (cf. p. 46).

[Cat11] Rick CATTELL. « Scalable SQL and NoSQL data stores ». In : *Acm Sigmod Record* 39.4 (2011), p. 12–27 (cf. p. 2).

[CD+20a] Perrine CRIBIER-DELANDE, Raphael PUGET, Vincent GUIGUE et Ludovic DENOYER. « Time series prediction generation from disentangled latent factors : new opportunities for smart cities ». In : *IEEE-ITSC*. 2020 (cf. p. 8, 53, 72, 74).

[CD+20b] Perrine CRIBIER-DELANDE, Raphael PUGET, Vincent GUIGUE et Ludovic DENOYER. « Time Series Prediction using Disentangled Latent Factors ». In : *European Symposium on Artificial Neural Networks*. 2020 (cf. p. 8, 53, 72, 74, 89).

[CG11] Hubert CECOTTI et Axel GRASER. « Convolutional neural networks for P300 detection with application to brain-computer interfaces ». In : *IEEE transactions on pattern analysis and machine intelligence* 33.3 (2011), p. 433–445 (cf. p. 17).

[CG16] Tianqi CHEN et Carlos GUESTRIN. « Xgboost : A scalable tree boosting system ». In : *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, p. 785–794 (cf. p. 77).

[Che+11] Minmin CHEN, Kilian Q WEINBERGER et John BLITZER. « Co-training for domain adaptation ». In : *Advances in neural information processing systems*. 2011, p. 2456–2464 (cf. p. 18).

[Che+12a] Lu CHEN, Wang WENBO, Meenakshi NAGARAJAN, Shaojun WANG et Amit P. SHETH. « Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter. » In : *ICWSM*. Sous la dir. de John G. BRESLIN, Nicole B. ELLISON, James G. SHANAHAN et Zeynep TUFEKCI. The AAAI Press, 2012 (cf. p. 23).

[Che+12b] Shuo CHEN, Josh L MOORE, Douglas TURNBULL et Thorsten JOACHIMS. « Playlist prediction via metric embedding ». In : *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2012, p. 714–722 (cf. p. 40, 42).

[Che+16] Huimin CHEN, Maosong SUN, Cunchao TU, Yankai LIN et Zhiyuan LIU. « Neural sentiment classification with user and product attention ». In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, p. 1650–1659 (cf. p. 49).

[Cho+14] Kyunghyun CHO, Bart van MERRIENBOER, Caglar GULCEHRE et al. « Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation ». In : *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. 2014 (cf. p. 2, 3, 25, 57, 61).

[Chu+15] Junyoung CHUNG, Caglar GÜLÇEHRE, Kyunghyun CHO et Yoshua BENGIO. « Gated feedback recurrent neural networks ». In : *ICML*. 2015 (cf. p. 57).

[CL01] C Chung CHANG et Chih-Jen LIN. « LIBSVM : a library for support vector machines ». In : *WWW* (2001) (cf. p. 2).

[Col+11a] Ronan COLLOBERT, Jason WESTON, Léon BOTTOU et al. « Natural language processing (almost) from scratch ». In : *Journal of machine learning research* 12.Aug (2011), p. 2493–2537 (cf. p. 24, 28, 51).

[Col+11b] Ronan COLLOBERT, Koray KAVUKCUOGLU et Clément FARABET. « Torch7 : A matlab-like environment for machine learning ». In : *BigLearn, NIPS Workshop*. EPFL-CONF-192376. 2011 (cf. p. 3, 24, 29).

[Col15] Christopher COLAH. *Understanding LSTM Networks*. `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`. 2015 (cf. p. 57).

[Con+11] Michael CONOVER, Jacob RATKIEWICZ, Matthew R FRANCISCO et al. « Political polarization on twitter. » In : *Icwsm* 133 (2011), p. 89–96 (cf. p. 49).

[Con+17] Marco CONGEDO, Alexandre BARACHANT et Rajendra BHATIA. « Riemannian geometry for EEG-based brain-computer interfaces ; a primer and a review ». In : *Brain-Computer Interfaces* 4.3 (2017), p. 155–174 (cf. p. 16).

[Cov+16] Paul COVINGTON, Jay ADAMS et Emre SARGIN. « Deep neural networks for youtube recommendations ». In : *Proceedings of the 10th ACM conference on recommender systems*. 2016, p. 191–198 (cf. p. 77).

[Cre+10] Fabio CRESTANI, Martin BRASCHLER, Jacques SAVOY et al. « Conference and Labs of the Evaluation Forum ». In : *Lecture Notes in Computer Science*. 2010 (cf. p. 3).

[CV95] Corinna CORTES et Vladimir VAPNIK. « Support vector machine ». In : *Machine learning* 20.3 (1995), p. 273–297 (cf. p. 1).

[CW08] Ronan COLLOBERT et Jason WESTON. « A unified architecture for natural language processing : Deep neural networks with multitask learning ». In : *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, p. 160–167 (cf. p. 3, 24, 49).

[DAO19] Kalpit DIXIT et Yaser AL-ONAIZAN. « Span-Level Model for Relation Extraction ». In : *Proceedings ofthe 57th Annual Meeting ofthe Association for Computational Linguistics*. 2019, p. 5308–5314 (cf. p. 66).

[Dec+00] Stefan DECKER, Sergey MELNIK, Frank VAN HARMELEN et al. « The semantic web : The roles of XML and RDF ». In : *IEEE Internet computing* 4.5 (2000), p. 63–73 (cf. p. 62).

[Dee+90] Scott DEERWESTER, Susan T DUMAIS, George W FURNAS, Thomas K LANDAUER et Richard HARSHMAN. « Indexing by latent semantic analysis ». In : *Journal of the American society for information science* 41.6 (1990), p. 391 (cf. p. 31).

[Del+16] Cynthia DELAUNEY, Nicolas BASKIOTIS et Vincent GUIGUE. « Trajectory Bayesian indexing : The airport ground traffic case ». In : *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE. 2016, p. 1047–1052 (cf. p. 5, 12).

[Des+05] Frédéric DESOBRY, Manuel DAVY et William J FITZGERALD. « A Class of Kernels For Sets of Vectors. » In : *ESANN*. 2005, p. 461–466 (cf. p. 2).

[Dev+18] Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE et Kristina TOUTANOVA. « Bert : Pre-training of deep bidirectional transformers for language understanding ». In : *arXiv preprint arXiv :1810.04805* (2018) (cf. p. 8, 28, 61, 76).

[DI07] Hal DAUMÉ III. « Frustratingly Easy Domain Adaptation ». In : *ACL 2007* (2007), p. 256 (cf. p. 18, 23).

[Dia+16] Charles-Emmanuel DIAS, Vincent GUIGUE et Patrick GALLINARI. « Recommandation et analyse de sentiments dans un espace latent textuel ». In : *CORIA-CIFED*. 2016, p. 73–88 (cf. p. 5, 7, 38, 40, 50).

[Dia+17a] Charles-Emmanuel DIAS, Vincent GUIGUE et Patrick GALLINARI. « Passé, présent, futurs : induction de carrières professionnelles à partir de CV. » In : *CORIA*. 2017, p. 281–296 (cf. p. 6, 7, 50, 53, 57, 61, 88).

[Dia+17b] Charles-Emmanuel DIAS, Vincent GUIGUE et Patrick GALLINARI. « Text-based collaborative filtering for cold-start soothing and recommendation enrichment ». In : *AISR*. 2017 (cf. p. 5, 7, 29, 36, 37).

[Dia+18a] Charles-Emmanuel DIAS, Vincent GUIGUE et Patrick GALLINARI. « Regularize and Explicit Collaborative Filtering With Textual Attention ». In : *European Symposium on Artificial Neural Networks*. 2018 (cf. p. 7, 38, 40).

[Dia+18b] Charles-Emmanuel DIAS, Clara Gainon de Forsan de GABRIAC, Vincent GUIGUE et Patrick GALLINARI. « RNN & modèle d'attention pour l'apprentissage de profils textuels personnalisés. » In : *CORIA*. 2018 (cf. p. 7, 38, 50).

[Dia+19a] Charles-Emmanuel DIAS, Vincent GUIGUE et Patrick GALLINARI. « Filtrage collaboratif explicite par analyse de sentiments à l'aveugle ». In : *CAp*. 2019 (cf. p. 7, 53, 54, 77).

[Dia+19b] Charles-Emmanuel DIAS, Vincent GUIGUE et Patrick GALLINARI. « Personalized attention for textual profiling and recommendation ». In : *Worshop EARS, ACM SIGIR*. 2019 (cf. p. 53).

[Dia+19c] Charles-Emmanuel DIAS, Clara Gainon de Forsan de GABRIAC, Vincent GUIGUE et Patrick GALLINARI. « RNN & modèle d'attention pour l'apprentissage de profils textuels personnalisés ». In : *Les Cahiers du Numérique* (2019) (cf. p. 38).

[DK11] Christian DESROSIERS et George KARYPIS. « A comprehensive survey of neighborhood-based recommendation methods ». In : *Recommender Systems Handbook* 69.11 (2011), p. 107–144 (cf. p. 27, 35).

[DL05] Y. DING et X. LI. « Time weight collaborative filtering ». In : *CIKM*. 2005 (cf. p. 41).

[Don+17] Li DONG, Shaohan HUANG, Furu WEI et al. « Learning to generate product reviews from attributes ». In : *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*. T. 1. 2017, p. 623–632 (cf. p. 38).

[Dor+06] Guido DORNHEGE, Benjamin BLANKERTZ, Matthias KRAULEDAT et al. « Combined optimization of spatial and temporal filters for improving brain-computer interfacing ». In : *IEEE transactions on biomedical engineering* 53.11 (2006), p. 2274–2281 (cf. p. 17).

[Dre+10] Mark DREDZE, Alex KULESZA et Koby CRAMMER. « Multi-domain learning by confidence-weighted parameter combination ». English. In : *Machine Learning* 79 (1-2 2010), p. 123–149 (cf. p. 23).

[DS14] Sander DIELEMAN et Benjamin SCHRAUWEN. « End-to-end learning for music audio ». In : *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, p. 6964–6968 (cf. p. 28, 52, 77).

[Dud+73] Richard O DUDA, Peter E HART et David G STORK. *Pattern classification*. Wiley, New York, 1973 (cf. p. 12).

[Dur+16] Thibaut DURAND, Nicolas THOME et Matthieu CORD. « Weldon : Weakly supervised learning of deep convolutional neural networks ». In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, p. 4743–4752 (cf. p. 3).

[Elm90] Jeffrey L ELMAN. « Finding structure in time ». In : *Cognitive science* 14.2 (1990), p. 179–211 (cf. p. 57).

[ET94] Bradley EFRON et Robert J TIBSHIRANI. *An introduction to the bootstrap*. CRC press, 1994 (cf. p. 16).

[EU20] Markus EBERTS et Adrian ULGES. « Span-based Joint Entity and Relation Extraction with Transformer Pre-training ». In : *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI)*. 2020 (cf. p. 66).

[Eve+15] M. EVERINGHAM, S. M. A. ESLAMI, L. VAN GOOL et al. « The Pascal Visual Object Classes Challenge : A Retrospective ». In : *International Journal of Computer Vision* 111.1 (jan. 2015), p. 98–136 (cf. p. 3).

[Fre99] Robert M FRENCH. « Catastrophic forgetting in connectionist networks ». In : *Trends in cognitive sciences* 3.4 (1999), p. 128–135 (cf. p. 77).

[Fri+01] Jerome FRIEDMAN, Trevor HASTIE et Robert TIBSHIRANI. *The elements of statistical learning*. T. 1. Springer series in statistics New York, 2001 (cf. p. 12).

[Fri+07] Jerome FRIEDMAN, Trevor HASTIE, Holger HÖFLING, Robert TIBSHIRANI et al. « Pathwise coordinate optimization ». In : *The Annals of Applied Statistics* 1.2 (2007), p. 302–332 (cf. p. 19).

[FS+96] Yoav FREUND, Robert E SCHAPIRE et al. « Experiments with a new boosting algorithm ». In : *Icml*. T. 96. 1996, p. 148–156 (cf. p. 2).

[FS97] Yoav FREUND et Robert E SCHAPIRE. « A decision-theoretic generalization of on-line learning and an application to boosting ». In : *Journal of computer and system sciences* 55.1 (1997), p. 119–139 (cf. p. 77).

[Gab+20] Clara Gainon de Forsan de GABRIAC, Vincent GUIGUE et Patrick GALLINARI. « Resume : A Robust Framework for Professional Profile Learning & Evaluation ». In : *European Symposium on Artificial Neural Networks*. 2020 (cf. p. 6, 7, 43, 88).

[Gal+87] Patrick GALLINARI, Yann LECUN, Sylvie THIRIA et Francoise FOGELMAN-SOULIE. « Memoires associatives distribuees ». In : *Proceedings of COGNITIVA* 87 (1987), p. 93 (cf. p. 55).

[Gal98] Mark JF GALES. « Maximum likelihood linear transformations for HMM-based speech recognition ». In : *Computer speech & language* 12.2 (1998), p. 75–98 (cf. p. 76).

[Gan+09] Gayatree GANU, Noemie ELHADAD et Amélie MARIAN. « Beyond the Stars : Improving Rating Predictions using Review Text Content. » In : *WebDB*. 2009 (cf. p. 31, 38).

[Gat+15] Leon A. GATYS, Alexander S. ECKER et Matthias BETHGE. « A Neural Algorithm of Artistic Style ». In : *CoRR* abs/1508.06576 (2015). arXiv : 1508.06576 (cf. p. 28).

[Ge+10] Mouzhi GE, Carla DELGADO-BATTENFELD et Dietmar JANNACH. « Beyond accuracy : evaluating recommender systems by coverage and serendipity ». In : *Proceedings of the fourth ACM conference on Recommender systems*. 2010, p. 257–260 (cf. p. 28).

[GE03] Isabelle GUYON et André ELISSEEFF. « An introduction to variable and feature selection ». In : *Journal of machine learning research* 3.Mar (2003), p. 1157–1182 (cf. p. 16).

[GH10] M GUTMANN et A HYVÄRINEN. « Noise-contrastive estimation : A new estimation principle for unnormalized statistical models ». In : *International Conference on Artificial Intelligence and Statistics*. 2010, p. 1–8 (cf. p. 34).

[Gio+19] John M GIORGI, Xindi WANG, Nicola SAHAR et al. « End-to-end Named Entity Recognition and Relation Extraction using Pre-trained Language Models ». In : *arXiv preprint arXiv :1912.13415* (2019) (cf. p. 66).

[Glo+11] Xavier GLOROT, Antoine BORDES et Yoshua BENGIO. « Domain Adaptation for Large-Scale Sentiment Classification : A Deep Learning Approach ». In : *ICML*. 2011 (cf. p. 23).

[Goe+12] Sharad GOEL, Jake M HOFMAN et M Irmak SIRER. « Who Does What on the Web : A Large-Scale Study of Browsing Behavior. » In : *ICWSM*. 2012 (cf. p. 49).

[Goe+18] Randy GOEBEL, Ajay CHANDER, Katharina HOLZINGER et al. « Explainable AI : the new 42 ? » In : *International cross-domain conference for machine learning and knowledge extraction*. Springer. 2018, p. 295–303 (cf. p. 12, 52).

[Goo+14a] Ian J GOODFELLOW, Mehdi MIRZA, Aaron Courville DA XIAO et Yoshua BENGIO. « An empirical investigation of catastrophic forgeting in gradientbased neural networks ». In : *In Proceedings of International Conference on Learning Representations (ICLR*. Citeseer. 2014 (cf. p. 52, 77).

[Goo+14b] Ian J GOODFELLOW, Jonathon SHLENS et Christian SZEGEDY. « Explaining and harnessing adversarial examples ». In : *arXiv preprint arXiv :1412.6572* (2014) (cf. p. 52, 77).

[Grb+15] Mihajlo GRBOVIC, Vladan RADOSAVLJEVIC, Nemanja DJURIC et al. « E-commerce in your inbox : Product recommendations at scale ». In : *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, p. 1809–1818 (cf. p. 40, 41).

[Gre+15] Klaus GREFF, Rupesh Kumar SRIVASTAVA, Jan KOUTNÍK, Bas R STEUNEBRINK et Jürgen SCHMIDHUBER. « LSTM : A search space odyssey ». In : *arXiv preprint arXiv :1503.04069* (2015) (cf. p. 57).

[Gru+04] Daniel GRUHL, Ramanathan GUHA, David LIBEN-NOWELL et Andrew TOMKINS. « Information diffusion through blogspace ». In : *Proceedings of the 13th international conference on World Wide Web*. ACM. 2004, p. 491–501 (cf. p. 2).

[GS+13] Élie GUÀRDIA-SEBAOUN, Abdelhalim RAFRAFI, Vincent GUIGUE et Patrick GALLINARI. « Cross-media sentiment classification and application to box-office forecasting ». In : *Proceedings of the 10th conference on open research areas in information retrieval*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. 2013, p. 201–208 (cf. p. 5, 12, 23, 24).

[GS+14] Elie GUÀRDIA-SEBAOUN, Vincent GUIGUE et Patrick GALLINARI. « Recommandation Dynamique dans les Graphes Géographiques ». In : *MARAMI*. 2014 (cf. p. 6, 29, 41).

[GS+15] Elie GUÀRDIA-SEBAOUN, Vincent GUIGUE et Patrick GALLINARI. « Latent trajectory modeling : A light and efficient way to introduce time in recommender systems ». In : *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM. 2015, p. 281–284 (cf. p. 6, 29, 41, 88).

[GS+16] Elie GUARDIA-SEBAOUN, Vincent GUIGUE et Patrick GALLINARI. « Apprentissage de trajectoires temporelles pour la recommandation dans les communautés d'utilisateurs ». In : *CAp*. 2016 (cf. p. 6, 29, 41).

[GUI+03] Vincent GUIGUE, Alain RAKOTOMAMONJY et Stéphane CANU. « SVM et k-ppv pour la reconnaissance d'émotions ». In : *19° Colloque sur le traitement du signal et des images, FRA, 2003*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images. 2003 (cf. p. 5).

[Gui+05] Vincent GUIGUE, Alain RAKOTOMAMONJY et Stéphane CANU. « Kernel basis pursuit ». In : *ECML*. 2005, p. 146–157 (cf. p. 4).

[Gui+06] Vincent GUIGUE, Alain RAKOTOMAMONJY et Stéphane CANU. « Translation-invariant classification of non-stationary signals ». In : *Neurocomputing* 69.7 (2006), p. 743–753 (cf. p. 4).

[Gui+19] Valentin GUIGUET, Perrine CRIBIER-DELANDE, Nicolas BASKIOTIS et Vincent GUIGUE. « Prédiction de séries temporelles multi-variées stationnaires : modélisation du contexte pour l'analyse des données de transports ». In : *GRETSI*. 2019 (cf. p. 8, 53, 72).

[Gup+16] Pankaj GUPTA, Hinrich SCHÜTZE et Bernt ANDRASSY. « Table Filling Multi-Task Recurrent Neural Network for Joint Entity and Relation Extraction ». In : *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*. Osaka, Japan : The COLING 2016 Organizing Committee, déc. 2016, p. 2537–2547 (cf. p. 66).

[Guy15] Ido GUY. « Social recommender systems ». In : *Recommender Systems Handbook*. Springer, 2015, p. 511–543 (cf. p. 30).

[Har+07] David R HARDOON, Janaina MOURAO-MIRANDA, Michael BRAMMER et John SHAWE-TAYLOR. « Unsupervised analysis of fMRI data using kernel canonical correlation ». In : *NeuroImage* 37.4 (2007), p. 1250–1259 (cf. p. 67).

[Hin+12] Geoffrey HINTON, N SRIVASTAVA et Kevin SWERSKY. *Lecture 6a Overview of mini–batch gradient descent*. Coursera Lecture slides https ://class. coursera. org/neuralnets-2012-001/lecture,[Online. 2012 (cf. p. 58).

[HM16] Ruining HE et Julian MCAULEY. « VBPR : visual bayesian personalized ranking from implicit feedback ». In : *Thirtieth AAAI Conference on Artificial Intelligence*. 2016 (cf. p. 28, 52).

[Hof99] Thomas HOFMANN. « Probabilistic latent semantic analysis ». In : *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1999, p. 289–296 (cf. p. 17).

[HS97] Sepp HOCHREITER et Jürgen SCHMIDHUBER. « Long short-term memory ». In : *Neural computation* 9.8 (1997), p. 1735–1780 (cf. p. 57, 72).

[Jin+10] Nitin JINDAL, Bing LIU et Ee-Peng LIM. « Finding unusual review patterns using unexpected rules ». In : *CIKM*. 2010, p. 1549–1552 (cf. p. 32).

[Joa98a] Thorsten JOACHIMS. *Making large-scale SVM learning practical*. Rapp. tech. Technical Report, SFB 475 : Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998 (cf. p. 2).

[Joa98b] Thorsten JOACHIMS. « Text categorization with support vector machines : Learning with many relevant features ». In : *Machine learning : ECML-98* (1998), p. 137–142 (cf. p. 17).

[Jor98] Michael Irwin JORDAN. *Learning in graphical models*. T. 89. Springer Science & Business Media, 1998 (cf. p. 2, 11, 52, 76).

[Joz+15] Rafal JOZEFOWICZ, Wojciech ZAREMBA et Ilya SUTSKEVER. « An empirical exploration of recurrent network architectures ». In : *ICML*. 2015 (cf. p. 57).

[Kag] *Kaggle : site d'hébergement de compétitions en machine learning*. 2010 (cf. p. 3).

[Kag+14] Mikael KAGEBACK, Olof MOGREN, Nina TAHMASEBI et Devdatt DUBHASHI. « Extractive Summarization using Continuous Vector Space Models ». In : *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@EACL 2014* (2014), p. 31–39 (cf. p. 34, 35).

[Kas+03] Hisashi KASHIMA, Koji TSUDA et Akihiro INOKUCHI. « Marginalized kernels between labeled graphs ». In : *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003, p. 321–328 (cf. p. 2).

[Kay+08] Kendrick N KAY, Thomas NASELARIS, Ryan J PRENGER et Jack L GALLANT. « Identifying natural images from human brain activity ». In : *Nature* 452.7185 (2008), p. 352–355 (cf. p. 67).

[KB05] Mikaela KELLER et Samy BENGIO. « A neural network for text representation ». In : *International Conference on Artificial Neural Networks*. Springer. 2005, p. 667–672 (cf. p. 51).

[KC17] Arzoo KATIYAR et Claire CARDIE. « Going out on a limb : Joint Extraction of Entity Mentions and Relations without Dependency Trees ». In : *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Vancouver, Canada : Association for Computational Linguistics, juil. 2017, p. 917–928 (cf. p. 66).

[Kem+03] David KEMPE, Jon KLEINBERG et Éva TARDOS. « Maximizing the spread of influence through a social network ». In : *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2003, p. 137–146 (cf. p. 2).

[Kor+09] Yehuda KOREN, Robert BELL et Chris VOLINSKY. « Matrix factorization techniques for recommender systems ». In : *Computer* 42.8 (2009) (cf. p. 27, 30, 42, 56).

[Kor+19] Simon KORNBLITH, Jonathon SHLENS et Quoc V LE. « Do better imagenet models transfer better ? » In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019, p. 2661–2671 (cf. p. 72).

[Kor10] Yehuda KOREN. « Collaborative filtering with temporal dynamics ». In : *Communications of the ACM* 53.4 (2010), p. 89–97 (cf. p. 30, 40, 42).

[Kri+12] Alex KRIZHEVSKY, Ilya SUTSKEVER et Geoffrey E HINTON. « Imagenet classification with deep convolutional neural networks ». In : *Advances in neural information processing systems*. 2012, p. 1097–1105 (cf. p. 2, 8, 25, 49, 51, 52).

[KW13] Diederik P KINGMA et Max WELLING. « Auto-encoding variational bayes ». In : *arXiv preprint arXiv :1312.6114* (2013) (cf. p. 76).

[Laf+01] John LAFFERTY, Andrew MCCALLUM et Fernando CN PEREIRA. « Conditional random fields : Probabilistic models for segmenting and labeling sequence data ». In : *international conference on Machine learning*. ACM. 2001 (cf. p. 2).

[Lam+17] Guillaume LAMPLE, Alexis CONNEAU, Ludovic DENOYER et Marc'Aurelio RANZATO. « Unsupervised machine translation using monolingual corpora only ». In : *arXiv preprint arXiv :1711.00043* (2017) (cf. p. 75).

[Lam+19] Guillaume LAMPLE, Sandeep SUBRAMANIAN, Eric SMITH et al. « Multiple-Attribute Text Rewriting ». In : *International Conference on Learning Representations*. 2019 (cf. p. 62).

[LC98] Claudia LEACOCK et Martin CHODOROW. « Combining local context and WordNet similarity for word sense identification ». In : *WordNet : An electronic lexical database* 49.2 (1998), p. 265–283 (cf. p. 68, 69).

[Le+15] Quoc V LE, Navdeep JAITLY et Geoffrey E HINTON. « A simple way to initialize recurrent networks of rectified linear units ». In : *arXiv preprint arXiv :1504.00941* (2015) (cf. p. 57).

[LeC+15] Yann LECUN, Yoshua BENGIO et Geoffrey HINTON. « Deep learning ». In : *Nature* 521.7553 (2015), p. 436–444 (cf. p. 3).

[LeC+89a] Yann LECUN, Bernhard BOSER, John S DENKER et al. « Backpropagation applied to handwritten zip code recognition ». In : *Neural computation* 1.4 (1989), p. 541–551 (cf. p. 1).

[LeC+89b] Yann LECUN et al. « Generalization and network design strategies ». In : *Connectionism in perspective* (1989), p. 143–155 (cf. p. 1, 51).

[Lev+15] Omer LEVY, Yoav GOLDBERG et Ido DAGAN. « Improving Distributional Similarity with Lessons Learned from Word Embeddings ». In : *Transactions of the Association for Computational Linguistics* 3 (2015), p. 211–225 (cf. p. 34).

[Li+10] Tao LI, Vikas SINDHWANI, Chris DING et Yi ZHANG. « Bridging domains with words : Opinion analysis with matrix tri-factorizations ». In : *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM. 2010, p. 293–302 (cf. p. 18).

[Li+16] Fei LI, Yue ZHANG, Meishan ZHANG et Donghong JI. « Joint Models for Extracting Adverse Drug Events from Biomedical Text ». In : *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI'16. New York, New York, USA : AAAI Press, 2016, 2838–2844 (cf. p. 66).

[Li+17a] Fei LI, Meishan ZHANG, Guohong FU et Donghong JI. « A neural joint model for entity and relation extraction from biomedical text ». In : *BMC bioinformatics* 18.1 (2017), p. 1–11 (cf. p. 66).

[Li+17b] Piji LI, Zihao WANG, Zhaochun REN, Lidong BING et Wai LAM. « Neural Rating Regression with Abstractive Tips Generation for Recommendation ». In : *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '17. Shinjuku, Tokyo, Japan : ACM, 2017, p. 345–354 (cf. p. 38).

[Li+19] Xiaoya LI, Fan YIN, Zijun SUN et al. « Entity-Relation Extraction as Multi-Turn Question Answering ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy : Association for Computational Linguistics, juil. 2019, p. 1340–1350 (cf. p. 66).

[Lic13] M. LICHMAN. *UCI Machine Learning Repository*. 2013 (cf. p. 3).

[Lin+17] Hailun LIN, Yong LIU, Weiping WANG, Yinliang YUE et Zheng LIN. « Learning entity and relation embeddings for knowledge resolution ». In : *Procedia Computer Science* 108 (2017), p. 345–354 (cf. p. 62).

[Lip+15] Zachary C LIPTON, Sharad VIKRAM et Julian MCAULEY. « Generative Concatenative Nets Jointly Learn to Write and Classify Reviews ». In : *arXiv preprint arXiv :1511.03683* (2015) (cf. p. 38).

[Liu+09] Han LIU, Mark PALATUCCI et Jian ZHANG. « Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery ». In : *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM. 2009, p. 649–656 (cf. p. 49, 68).

[LJ14] Qi LI et Heng JI. « Incremental Joint Extraction of Entity Mentions and Relations ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Baltimore, Maryland : Association for Computational Linguistics, juin 2014, p. 402–412 (cf. p. 66).

[LM14] QV LE et T MIKOLOV. « Distributed Representations of Sentences and Documents ». In : *ICML*. T. 32. 2014, 1188–1196 (cf. p. 33, 34).

[Loo+06] Gaëlle LOOSLI, Sang-Goog LEE, Vincent GUIGUE, Alain RAKOTOMAMONJY et Stéphane CANU. « Perception d'états affectifs et apprentissage ». In : *Revue d'Intelligence Artificielle* (2006) (cf. p. 5).

[Lop+11] Pasquale LOPS, Marco DE GEMMIS et Giovanni SEMERARO. « Content-based recommender systems : State of the art and trends ». In : *Recommender systems handbook*. Springer, 2011, p. 73–105 (cf. p. 31).

[Lou+17] Thomas LOUAIL, Maxime LENORMAND, Juan Murillo ARIAS et José J RAMASCO. « Crowdsourcing the Robin Hood effect in cities ». In : *Applied Network Science* 2.1 (2017), p. 11 (cf. p. 45).

[LS99] Daniel D LEE et H Sebastian SEUNG. « Learning the parts of objects by non-negative matrix factorization ». In : *Nature* 401.6755 (1999), p. 788 (cf. p. 2, 27, 52).

[Lua+18] Yi LUAN, Luheng HE, Mari OSTENDORF et Hannaneh HAJISHIRZI. « Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction ». In : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium : Association for Computational Linguistics, 2018, p. 3219–3232 (cf. p. 66).

[Lua+19] Yi LUAN, Dave WADDEN, Luheng HE et al. « A general framework for information extraction using dynamic span graphs ». In : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota : Association for Computational Linguistics, juin 2019, p. 3036–3046 (cf. p. 66).

[Maa+11] Andrew L MAAS, Raymond E DALY, Peter T PHAM et al. « Learning word vectors for sentiment analysis ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*. Association for Computational Linguistics. 2011, p. 142–150 (cf. p. 22).

[Maa+14] Maria Laura MAAG, Ludovic DENOYER et Patrick GALLINARI. « Graph anonymization using machine learning ». In : *Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on*. IEEE. 2014, p. 1111–1118 (cf. p. 50).

[Mao+14] Junhua MAO, Wei XU, Yi YANG et al. « Deep captioning with multimodal recurrent neural networks (m-rnn) ». In : *arXiv preprint arXiv :1412.6632* (2014) (cf. p. 28, 52, 77).

[Mat+05] Shotaro MATSUMOTO, Hiroya TAKAMURA et Manabu OKUMURA. « Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees. » In : *PAKDD*. T. 5. Springer. 2005, p. 301–311 (cf. p. 18).

[MB16] Makoto MIWA et Mohit BANSAL. « End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures ». In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Berlin, Germany : Association for Computational Linguistics, août 2016, p. 1105–1116 (cf. p. 66).

[McA+15] Julian MCAULEY, Rahul PANDEY et Jure LESKOVEC. « Inferring networks of substitutable and complementary products ». In : *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM. 2015, p. 785–794 (cf. p. 56).

[McC86] John MCCARTHY. « Applications of circumscription to formalizing common-sense knowledge ». In : *Artificial intelligence* 28.1 (1986), p. 89–116 (cf. p. 52).

[Mei+07] Qiaozhu MEI, Xu LING, Matthew WONDRA, Hang SU et ChengXiang ZHAI. « Topic sentiment mixture : modeling facets and opinions in weblogs ». In : *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, p. 171–180 (cf. p. 18, 24).

[MH08] Laurens van der MAATEN et Geoffrey HINTON. « Visualizing data using t-SNE ». In : *Journal of Machine Learning Research* 9 (2008), p. 2579–2605 (cf. p. 60).

[Mik+13] Tomas MIKOLOV, Ilya SUTSKEVER, Kai CHEN, Greg S CORRADO et Jeff DEAN. « Distributed representations of words and phrases and their compositionality ». In : *Advances in neural information processing systems*. 2013, p. 3111–3119 (cf. p. 2, 3, 25, 28, 33, 34, 42, 49, 52, 54, 62, 68, 69, 76).

[Mik+14] Tomas MIKOLOV, Armand JOULIN, Sumit CHOPRA, Michael MATHIEU et Marc'Aurelio RANZATO. « Learning longer memory in recurrent neural networks ». In : *arXiv preprint arXiv :1412.7753* (2014) (cf. p. 57).

[Mil95] George A MILLER. « WordNet : a lexical database for English ». In : *Communications of the ACM* 38.11 (1995), p. 39–41 (cf. p. 17).

[Mir+08] Piotr W MIROWSKI, Yann LECUN, Deepak MADHAVAN et Ruben KUZNIECKY. « Comparing SVM and convolutional networks for epileptic seizure prediction from intracranial EEG ». In : *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*. IEEE. 2008, p. 244–249 (cf. p. 17).

[Mit+08] Tom M MITCHELL, Svetlana V SHINKAREVA, Andrew CARLSON et al. « Predicting human brain activity associated with the meanings of nouns ». In : *science* 320.5880 (2008), p. 1191–1195 (cf. p. 67–69).

[Mit+15] T. MITCHELL, W. COHEN, E. HRUSCHKA et al. « Never-Ending Learning ». In : *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*. 2015 (cf. p. 76).

[ML13a] Julian MCAULEY et Jure LESKOVEC. « Hidden Factors and Hidden Topics : Understanding Rating Dimensions with Review Text ». In : *ACM Conference on Recommender Systems*. 2013, p. 165–172 (cf. p. 30–32, 35, 36, 38, 49, 56).

[ML13b] Julian John MCAULEY et Jure LESKOVEC. « From amateurs to connoisseurs : modeling the evolution of user expertise through online reviews ». In : *WWW*. 2013, p. 897–908 (cf. p. 30, 32, 41, 42).

[MN+98] Andrew McCallum, Kamal Nigam et al. « A comparison of event models for naive bayes text classification ». In : *AAAI-98 workshop on learning for text categorization*. T. 752. Madison, WI. 1998, p. 41–48 (cf. p. 17, 75).

[MR00] Raymond J Mooney et Loriene Roy. « Content-based book recommending using learning for text categorization ». In : *Proceedings of the fifth ACM conference on Digital libraries*. 2000, p. 195–204 (cf. p. 27).

[MS12] Yelena Mejova et Padmini Srinivasan. « Crossing Media Streams with Sentiment : Domain Adaptation in Blogs, Reviews and Twitter. » In : *ICWSM*. 2012 (cf. p. 18, 23).

[MS14] Makoto Miwa et Yutaka Sasaki. « Modeling Joint Entity and Relation Extraction with Table Representation ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar : Association for Computational Linguistics, oct. 2014, p. 1858–1869 (cf. p. 66).

[Ned+16] Thomas Nedelec, Elena Smirnova et Flavian Vasile. « Content2vec : Specializing joint representations of product images and text for the task of product recommendation ». In : (2016) (cf. p. 77).

[Noz+14] Debora Nozza, Daniele Maccagnola, Vincent Guigue, Enza Messina et Patrick Gallinari. « A latent representation model for sentiment analysis in heterogeneous social networks ». In : *International Conference on Software Engineering and Formal Methods*. Springer, Cham. 2014, p. 201–213 (cf. p. 5, 29).

[NV19] Dat Quoc Nguyen et Karin Verspoor. « End-to-end neural relation extraction using deep biaffine attention ». In : *European Conference on Information Retrieval*. Springer. 2019, p. 729–738 (cf. p. 66).

[Pal+09] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton et Tom M Mitchell. « Zero-shot learning with semantic output codes ». In : *Advances in neural information processing systems*. 2009, p. 1410–1418 (cf. p. 3, 68, 69).

[Pan+02] Bo Pang, Lillian Lee et Shivakumar Vaithyanathan. « Thumbs up ? : sentiment classification using machine learning techniques ». In : *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics. 2002, p. 79–86 (cf. p. 17, 18).

[Pan+10] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang et Zheng Chen. « Cross-domain sentiment classification via spectral feature alignment ». In : *Proceedings of the 19th international conference on World wide web*. ACM. 2010, p. 751–760 (cf. p. 21).

[Pan+19] Tianyu Pang, Kun Xu, Chao Du, Ning Chen et Jun Zhu. « Improving adversarial robustness via promoting ensemble diversity ». In : *arXiv preprint arXiv :1901.08846* (2019) (cf. p. 77).

[Par61] Emanuel Parzen. « Mathematical considerations in the estimation of spectra ». In : *Technometrics* 3.2 (1961), p. 167–190 (cf. p. 1).

[Paz99] M. J. Pazzani. « A Framework for Collaborative, Content-Based and Demographic Filtering ». In : *AI Review* 13 (1999), p. 393–408 (cf. p. 38).

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort et al. « Scikit-learn : Machine Learning in Python ». In : *Journal of Machine Learning Research* 12 (2011), p. 2825–2830 (cf. p. 3).

[Pen+14] Jeffrey PENNINGTON, Richard SOCHER et Christopher D MANNING. « GloVe : Global Vectors for Word Representation ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 2014 (cf. p. 49).

[Pet+11] François PETITJEAN, Alain KETTERLIN et Pierre GANÇARSKI. « A global averaging method for dynamic time warping, with applications to clustering ». In : *Pattern Recognition* 44.3 (2011), p. 678–693 (cf. p. 71).

[Pet+18] Matthew PETERS, Mark NEUMANN, Mohit IYYER et al. « Deep Contextualized Word Representations ». In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*. T. 1. 2018, p. 2227–2237 (cf. p. 8, 28, 43, 61, 76).

[Pip+14] Luepol PIPANMAEKAPORN, Thierry ARTIERES et Vincent GUIGUE. « Learning to combine Semantic Features for Neurolinglistic Decoding ». In : *Workshop MLSB*. 2014 (cf. p. 7, 53).

[Pip+15] Luepol PIPANMAEKAPORN, Ludmilla TAJTELBOM, Vincent GUIGUE et Thierry ARTIERES. « Designing Semantic Feature Spaces for Brain-Reading ». In : *European Symposium on Artificial Neural Networks*. 2015, p. 433–438 (cf. p. 7, 53).

[Pit95] Jacques PITRAT. « Des métaconnaissances pour des systèmes intelligents ». In : *Quaderni* (1995) (cf. p. 76).

[PL+08] Bo PANG, Lillian LEE et al. « Opinion mining and sentiment analysis ». In : *Foundations and Trends® in Information Retrieval* 2.1–2 (2008), p. 1–135 (cf. p. 2, 18, 24).

[Poi+10] D. POIRIER, F. FESSANT et I. TELLIER. « De la classification d'opinion à la recommandation : l'apport des textes communautaires ». In : *TAL* 51 (2010) (cf. p. 31).

[Pol08] Russell A POLDRACK. « The role of fMRI in cognitive neuroscience : where do we stand ? » In : *Current opinion in neurobiology* 18.2 (2008), p. 223–227 (cf. p. 67).

[Pou+14a] Mickaël POUSSEVIN, Vincent GUIGUE et Patrick GALLINARI. « Extended recommendation framework : Generating the text of a user review as a personalized summary ». In : *arXiv preprint arXiv :1412.5448* (2014) (cf. p. 5, 7, 35).

[Pou+14b] Mickaël POUSSEVIN, Nicolas BASKIOTIS, Vincent GUIGUE et Patrick GALLINARI. « Factorisation matricielle sous contraintes pour l'analyse des usages du métro parisien ». In : *CAp'2014 : Conférence d'Apprentissage Automatique*. 2014 (cf. p. 46).

[Pou+14c] Mickaël POUSSEVIN, Emeric TONNELIER, Nicolas BASKIOTIS, Vincent GUIGUE et Patrick GALLINARI. « Mining ticketing logs for usage characterization with nonnegative matrix factorization ». In : *International Workshop on Modeling Social Media*. Springer International Publishing. 2014, p. 147–164 (cf. p. 6, 29, 46, 47, 72).

[Pou+14d] Mickaël POUSSEVIN, Elie GUARDIA-SEBAOUN, Vincent GUIGUE et Patrick GALLINARI. « Recommandation par combinaison de filtrage collaboratif et d'analyse de sentiments ». In : *CORIA 2014*. 2014, pp–27 (cf. p. 5, 29, 31, 32, 38).

[Pou+15a] Mickaël POUSSEVIN, Vincent GUIGUE et Patrick GALLINARI. « Extended recommendation framework : Generating the text of a user review as a personalized summary ». In : *Workshop on New Trends in Content-Based Recommender Systems, RecSys 2015*. 2015 (cf. p. 29, 38).

[Pou+15b] Mickaël POUSSEVIN, Vincent GUIGUE et Patrick GALLINARI. « Extraction d'un vocabulaire de surprise par mélange de filtrage collaboratif et d'analyse de sentiments. » In : *CORIA*. 2015, p. 123–138 (cf. p. 5).

[Pou+16] M. POUSSEVIN, E. TONNELIER, N. BASKIOTIS, V. GUIGUE et P. GALLINARI. « Mining ticketing logs for usage characterization with nonnegative matrix factorization ». In : *LNCS Big Data Analytics in the Social and Ubiquitous Context* (2016) (cf. p. 6, 29, 46, 47).

[PP11] Alexander PAK et Patrick PAROUBEK. « Text representation using dependency tree subgraphs for sentiment analysis ». In : *Database systems for adanced applications* (2011), p. 323–332 (cf. p. 18).

[PY10] Sinno Jialin PAN et Qiang YANG. « A survey on transfer learning ». In : *IEEE Transactions on knowledge and data engineering* 22.10 (2010), p. 1345–1359 (cf. p. 18).

[Raf+11] Abdelhalim RAFRAFI, Vincent GUIGUE et Patrick GALLINARI. « Réseau de neurones profond et svm pour la classification de sentiments ». In : *CORIA : COnférence en Recherche d'Information et Applications*. Éditions Universitaires d'Avignon. 2011, p. 121–133 (cf. p. 4, 12, 24).

[Raf+12a] Abdelhalim RAFRAFI, Vincent GUIGUE et Patrick GALLINARI. « Coping with the Document Frequency Bias in Sentiment Classification. » In : *ICWSM*. 2012 (cf. p. 5, 12, 19).

[Raf+12b] Abdelhalim RAFRAFI, Vincent GUIGUE et Patrick GALLINARI. « Pénalisation des mots fréquents pour la classification de sentiments ». In : *Les Cahiers du numérique* 7.2 (2012), p. 63–84 (cf. p. 12, 19).

[Raf+12c] Abdelhalim RAFRAFI, Vincent GUIGUE et Patrick GALLINARI. « Représentations et régularisations pour la classification de sentiments ». In : *CORIA*. 2012, p. 285–300 (cf. p. 5, 12, 19, 24).

[Raf+12d] Abdelhalim RAFRAFI, Vincent GUIGUE et Patrick GALLINARI. « Réseau de neurones à double convolution pour la classification de sentiments multi-domaines ». In : *Conférence Francophone sur l'Apprentissage Automatique-CAp 2012*. 2012, 16–p (cf. p. 4, 12, 24).

[Raf+13] Abdelhalim RAFRAFI, Vincent GUIGUE et Patrick GALLINARI. « Classification de Sentiments Multi-Domaines & Passage à l'Echelle ». In : *CORIA*. 2013 (cf. p. 5, 12, 21, 22).

[Rak+05] Alain RAKOTOMAMONJY, Vincent GUIGUE, Gregory MALLET et Victor ALVARADO. « Ensemble of SVMs for improving brain computer interface P300 speller performances ». In : *Artificial Neural Networks : Biological Inspirations–ICANN 2005* (2005), p. 45–50 (cf. p. 12).

[Reb+19] Clément REBUFFEL, Laure SOULIER, Geoffrey SCOUTHEETEN et Patrick GALLINARI. « A Hierarchical Model for Data-to-Text Generation ». In : *arXiv preprint arXiv :1912.10011* (2019) (cf. p. 8, 88).

[Reb+20] Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten et Patrick Gallinari. « A Hierarchical Model for Data-to-Text Generation ». In : *European Conference on Information Retrieval*. Springer. 2020, p. 65–80 (cf. p. 62).

[Ree+16] Scott Reed, Zeynep Akata, Xinchen Yan et al. « Generative adversarial text to image synthesis ». In : *JMLR* (2016) (cf. p. 28).

[Ren+12] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner et Lars Schmidt-Thieme. « BPR : Bayesian personalized ranking from implicit feedback ». In : *arXiv preprint arXiv :1205.2618* (2012) (cf. p. 28).

[Res95] Philip Resnik. « Using information content to evaluate semantic similarity in a taxonomy ». In : *arXiv preprint cmp-lg/9511007* (1995) (cf. p. 69).

[RG08] Alain Rakotomamonjy et Vincent Guigue. « BCI competition III : dataset II-ensemble of SVMs for BCI P300 speller ». In : *IEEE transactions on biomedical engineering* 55.3 (2008), p. 1147–1154 (cf. p. 4, 12, 14, 71).

[RJ86] Lawrence Rabiner et B Juang. « An introduction to hidden Markov models ». In : *ieee assp magazine* 3.1 (1986), p. 4–16 (cf. p. 1).

[Ros58] Frank Rosenblatt. « The perceptron : A probabilistic model for information storage and organization in the brain. » In : *Psychological review* 65.6 (1958), p. 386 (cf. p. 1).

[Rum+86] David E Rumelhart, Geoffrey E Hinton et Ronald J Williams. « Learning representations by back-propagating errors ». In : *Nature* 323 (1986), p. 533–536 (cf. p. 1).

[Rus+15] Olga Russakovsky, Jia Deng, Hao Su et al. « ImageNet Large Scale Visual Recognition Challenge ». In : *International Journal of Computer Vision (IJCV)* 115.3 (2015), p. 211–252 (cf. p. 3).

[San+19] Victor Sanh, Thomas Wolf et Sebastian Ruder. « A hierarchical multi-task approach for learning embeddings from semantic tasks ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 33. 2019, p. 6949–6956 (cf. p. 66).

[Sav05] Mark L Savickas. « The theory and practice of career construction ». In : *Career development and counseling : Putting theory and research to work* 1 (2005), p. 42–70 (cf. p. 42).

[Sch15] Jürgen Schmidhuber. « Deep learning in neural networks : An overview ». In : *Neural networks* 61 (2015), p. 85–117 (cf. p. 3).

[SDM03] Erik F. Tjong Kim Sang et Fien De Meulder. « Introduction to the CoNLL-2003 shared task ». In : *Proceedings of the seventh Conference on Natural Language Learning at NAACL-HLT 2003*. T. 4. 2003, p. 142–147 (cf. p. 63).

[Seb02] Fabrizio Sebastiani. « Machine learning in automated text categorization ». In : *ACM computing surveys (CSUR)* 34.1 (2002), p. 1–47 (cf. p. 17).

[SG16] Damien Sileo et Vincent Guigue. « Apprentissage relationnel pour la recommandation et la prédiction de données manquantes ». In : *CAp*. 2016 (cf. p. 5, 28).

[Sim+19a] Etienne Simon, Vincent Guigue et Benjamin Piwowarski. « Extraction d'information non supervisée avec des modèles discriminants ». In : *CAp*. 2019 (cf. p. 7, 65).

[Sim+19b] Etienne SIMON, Vincent GUIGUE et Benjamin PIWOWARSKI. « Unsupervised Information Extraction : Regularizing Discriminative Approaches with Relation Distribution Losses ». In : *ACL*. 2019 (cf. p. 7, 65).

[SJ72] Karen SPARCK JONES. « A statistical interpretation of term specificity and its application in retrieval ». In : *Journal of documentation* 28.1 (1972), p. 11–21 (cf. p. 18).

[Soc+13a] Richard SOCHER, Danqi CHEN, Christopher D MANNING et Andrew NG. « Reasoning with neural tensor networks for knowledge base completion ». In : *Advances in neural information processing systems*. 2013, p. 926–934 (cf. p. 8, 62, 88).

[Soc+13b] Richard SOCHER, Milind GANJOO, Christopher D MANNING et Andrew NG. « Zero-shot learning through cross-modal transfer ». In : *Advances in neural information processing systems*. 2013, p. 935–943 (cf. p. 3).

[Soc+14] Richard SOCHER, Andrej KARPATHY, Quoc V LE, Christopher D MANNING et Andrew Y NG. « Grounded compositional semantics for finding and describing images with sentences ». In : *Transactions of the Association for Computational Linguistics* 2 (2014), p. 207–218 (cf. p. 29).

[SS00] Robert E SCHAPIRE et Yoram SINGER. « BoosTexter : A boosting-based system for text categorization ». In : *Machine learning* 39.2-3 (2000), p. 135–168 (cf. p. 17).

[Ste10] Harald STECK. « Training and testing of recommender systems on data missing not at random ». In : *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, p. 713–722 (cf. p. 28).

[Str+17] Emma STRUBELL, Patrick VERGA, David BELANGER et Andrew MCCALLUM. « Fast and Accurate Entity Recognition with Iterated Dilated Convolutions ». In : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, p. 2670–2680 (cf. p. 63).

[Suk+15] Sainbayar SUKHBAATAR, Jason WESTON, Rob FERGUS et al. « End-to-end memory networks ». In : *Advances in neural information processing systems*. 2015, p. 2440–2448 (cf. p. 8, 52, 88).

[Sun+18] Changzhi SUN, Yuanbin WU, Man LAN et al. « Extracting Entities and Relations with Joint Minimum Risk Training ». In : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium : Association for Computational Linguistics, 2018, p. 2256–2265 (cf. p. 66).

[SZ14] Karen SIMONYAN et Andrew ZISSERMAN. « Very deep convolutional networks for large-scale image recognition ». In : *arXiv preprint arXiv :1409.1556* (2014) (cf. p. 3, 25, 49).

[Tai+15] Kai Sheng TAI, Richard SOCHER et Christopher D MANNING. « Improved semantic representations from tree-structured long short-term memory networks ». In : *arXiv preprint arXiv :1503.00075* (2015) (cf. p. 3).

[Tai+19a] Bruno TAILLÉ, Vincent GUIGUE et Patrick GALLINARI. « Contextualized Embeddings in Named-Entity Recognition : An Empirical Study on Generalization ». In : *EurNLP*. 2019 (cf. p. 7, 63).

[Tai+19b] Bruno TAILLÉ, Vincent GUIGUE et Patrick GALLINARI. « Une Etude Empirique de la Capacité de Généralisation des Plongements de Mots Contextuels en Extraction d'Entités ». In : *CAp*. 2019 (cf. p. 7, 53, 63).

[Tai+20a] Bruno TAILLÉ, Vincent GUIGUE et Patrick GALLINARI. « Contextualized Embeddings in Named-Entity Recognition : An Empirical Study on Generalization ». In : *ECIR*. 2020 (cf. p. 7, 53, 63, 89).

[Tai+20b] Bruno TAILLÉ, Vincent GUIGUE, Geoffrey SCOUTHEETEN et Patrick GALLINARI. « Let's Stop Incorrect Comparisons in End-to-end Relation Extraction ! » In : *EMNLP*. 2020 (cf. p. 7, 8, 53, 65, 89).

[Tib96] Robert TIBSHIRANI. « Regression shrinkage and selection via the lasso ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), p. 267–288 (cf. p. 16, 19).

[Tik+77] Andrei Nikolaevich TIKHONOV, Vasilii Iakovlevich ARSENIN et Fritz JOHN. *Solutions of ill-posed problems*. V. H. Winston et Sons, 1977 (cf. p. 1, 13, 19).

[TM07] Nava TINTAREV et Judith MASTHOFF. « A survey of explanations in recommender systems ». In : *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*. IEEE. 2007, p. 801–810 (cf. p. 28, 30).

[Ton+16] Emeric TONNELIER, Nicolas BASKIOTIS, Vincent GUIGUE et Patrick GALLINARI. « Smart card in public transportation : Designing a analysis system at the human scale ». In : *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE. 2016, p. 1336–1341 (cf. p. 6, 29, 46, 47).

[Ton+17] Emeric TONNELIER, Nicolas BASKIOTIS, Vincent GUIGUE et Patrick GALLINARI. « Anomaly detection and characterization in smart card logs using NMF and Tweets ». In : *European Symposium on Artificial Neural Networks*. 2017 (cf. p. 6, 29, 46, 47).

[Ton+18a] Emeric TONNELIER, Nicolas BASKIOTIS, Vincent GUIGUE et Patrick GALLINARI. « Anomaly detection in smart card logs and distant evaluation with Twitter : a robust framework ». In : *Neurocomputing* (2018) (cf. p. 6, 29, 46, 47).

[Ton+18b] Emeric TONNELIER, Nicolas BASKIOTIS, Vincent GUIGUE et Patrick GALLINARI. « Anomaly detection in smart card logs and distant evaluation with Twitter : a robust framework ». In : *Neurocomputing* (2018) (cf. p. 6, 71, 72).

[Ton+18c] Emeric TONNELIER, Nicolas BASKIOTIS, Vincent GUIGUE et Patrick GALLINARI. « Factorisation de Tenseurs pour l'analyse de réseaux de mobilité ». In : *Rencontre Francophone Transport et Mobilité*. 2018 (cf. p. 6, 46, 47).

[Tso+04] Ioannis TSOCHANTARIDIS, Thomas HOFMANN, Thorsten JOACHIMS et Yasemin ALTUN. « Support vector machine learning for interdependent and structured output spaces ». In : *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004, p. 104 (cf. p. 2).

[TT12] G. TAKÁCS et D. TIKK. « Alternating least squares for personalized ranking ». In : *RecSys*. 2012 (cf. p. 40, 42).

[Van+17] Paul VANHAESEBROUCK, Aurélien BELLET et Marc TOMMASI. « Decentralized Collaborative Learning of Personalized Models over Networks ». In : *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Fort Lauderdale, Florida., United States, avr. 2017 (cf. p. 50).

[VC03] Jaideep VAIDYA et Chris CLIFTON. « Privacy-preserving k-means clustering over vertically partitioned data ». In : *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2003, p. 206–215 (cf. p. 50).

[Vin+08] Pascal VINCENT, Hugo LAROCHELLE, Yoshua BENGIO et Pierre-Antoine MANZAGOL. « Extracting and composing robust features with denoising autoencoders ». In : *Proceedings of the 25th international conference on Machine learning*. 2008, p. 1096–1103 (cf. p. 51, 76).

[Wad+19] David WADDEN, Ulme WENNBERG, Yi LUAN et Hannaneh HAJISHIRZI. « Entity, relation, and event extraction with contextualized span representations ». In : *arXiv preprint arXiv :1909.03546* (2019) (cf. p. 66).

[Wan+10] Hongning WANG, Yue LU et Chengxiang ZHAI. « Latent Aspect Rating Analysis on Review Text Data : A Rating Regression Approach ». In : *ACM SIGKDD*. 2010, p. 783–792 (cf. p. 22).

[Wei+13] Ralph WEISCHEDEL, Martha PALMER, Mitchell MARCUS et al. « OntoNotes Release 5.0 LDC2013T19 ». In : *Linguistic Data Consortium, Philadelphia, PA* (2013) (cf. p. 63).

[Wes+14] Jason WESTON, Sumit CHOPRA et Antoine BORDES. « Memory networks ». In : *arXiv preprint arXiv :1410.3916* (2014) (cf. p. 3, 8, 88).

[Whi12] Tom WHITE. *Hadoop : The definitive guide*. " O'Reilly Media, Inc.", 2012 (cf. p. 2).

[Xu+19] Hu XU, Bing LIU, Lei SHU et Philip S YU. « Bert post-training for review reading comprehension and aspect-based sentiment analysis ». In : *arXiv preprint arXiv :1904.02232* (2019) (cf. p. 72).

[Yan+17] Zichao YANG, Phil BLUNSOM, Chris DYER et Wang LING. « Reference-Aware Language Models ». In : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, p. 1850–1859 (cf. p. 8, 88).

[Zab+17] Eloi ZABLOCKI, Benjamin PIWOWARSKI, Laure SOULIER et Patrick GALLINARI. « Learning multi-modal word representation grounded in visual context ». In : *arXiv preprint arXiv :1711.03483* (2017) (cf. p. 52, 77).

[Zah+10] Matei ZAHARIA, Mosharaf CHOWDHURY, Michael J FRANKLIN, Scott SHENKER et Ion STOICA. « Spark : Cluster computing with working sets ». In : *HotCloud* 10.10-10 (2010), p. 95 (cf. p. 2).

[Zen+15] Daojian ZENG, Kang LIU, Yubo CHEN et Jun ZHAO. « Distant supervision for relation extraction via piecewise convolutional neural networks ». In : *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, p. 1753–1762 (cf. p. 52).

[Zha+17] Meishan ZHANG, Yue ZHANG et Guohong FU. « End-to-End Neural Relation Extraction with Global Optimization ». In : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark : Association for Computational Linguistics, sept. 2017, p. 1730–1740 (cf. p. 66).

[Zhe+17] Suncong ZHENG, Yuexing HAO, Dongyuan LU et al. « Joint entity and relation extraction based on a hybrid neural network ». In : *Neurocomputing* 257 (2017), p. 59–66 (cf. p. 66).

[Zia+17] Ali ZIAT, Edouard DELASALLES, Ludovic DENOYER et Patrick GALLINARI. « Spatio-Temporal Neural Networks for Space-Time Series Forecasting and Relations Discovery ». In : *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*. 2017, p. 705–714 (cf. p. 17).