Apprentissage de plongements de mots dynamiques avec régularisation de la dérive

Syrielle Montariol[1, 2] Alexandre Allauzen[1]
(1) LIMSI, CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, F-91405 Orsay, France
(2) Société Générale, 17 Cours Valmy 92043 Puteaux, France
syrielle.montariol@limsi.fr, alexandre.allauzen@limsi.fr

RÉSUMÉ

L'usage, le sens et la connotation des mots peuvent changer au cours du temps. Les plongements lexicaux diachroniques permettent de modéliser ces changements de manière non supervisée. Dans cet article nous étudions l'impact de plusieurs fonctions de coût sur l'apprentissage de plongements dynamiques, en comparant les comportements de variantes du modèle *Dynamic Bernoulli Embeddings*. Les plongements dynamiques sont estimés sur deux corpus couvrant les mêmes deux décennies, le *New York Times Annotated Corpus* en anglais et une sélection d'articles du journal *Le Monde* en français, ce qui nous permet de mettre en place un processus d'analyse bilingue de l'évolution de l'usage des mots.

ABSTRACT

Learning dynamic word embeddings with drift regularisation

Word usage, meaning and connotation change throughout time. Diachronic word embeddings are used to grasp these changes in an unsupervised way. In this paper, we use variants of the *Dynamic Bernoulli Embeddings* model to learn dynamic word embeddings, in order to identify notable properties of the model. The comparison is made on the *New York Times Annotated Corpus* in English and a set of articles from the French newspaper *Le Monde* covering the same period. This allows us to define a pipeline to analyse the evolution of words use across two languages.

MOTS-CLÉS : Diachronie, Plongements lexicaux, analyse bilingue.

KEYWORDS: Diachrony, word embeddings, cross-lingual analysis.

1 Introduction

Les langues peuvent être considérées comme des systèmes dynamiques : l'usage des mots évolue au cours du temps, reflétant les nombreux aspects des évolutions de la société, qu'ils soient culturels, technologiques ou dûs à d'autres facteurs (Aitchison, 2001).

La diachronie désigne l'étude de ces variations temporelles d'usage et de sens au sein d'une langue. Ici, nous étudions un corpus journalistique d'une plage temporelle de deux décennies : l'usage des mots évolue suite à des événements ayant un retentissement médiatique. Par exemple, l'usage du mot "Katrina" a connu un important changement au cours de ces deux décennies. Si par le passé, il fut exclusivement utilisé comme un prénom féminin, comme *Justine* et *Sonja*, dès 1999, son sens se rapproche de celui d'*ouragan*, avec l'arrivée du premier orage tropical éponyme. Puis à partir de

2005 où le cyclone Katrina eut lieu, ce qui fut un prénom féminin partage désormais le même champs lexical que les mots "désastre", "dévastation" et "inondation".

Détecter et comprendre ces changements avec le concours de méthodes d'apprentissage automatique est utile à la recherche linguistique, mais aussi à de nombreuses tâches de traitement automatique des langues. Ajouter une notion temporelle aux représentations de mots permet d'étudier des corpus qui s'étendent sur des plages temporelles longues avec une plus grande acuité. Le problème se pose particulièrement aujourd'hui, alors qu'un nombre croissant de documents historiques sont numérisés et rendus accessibles ; leur analyse conjointe à celle de corpus contemporains, pour des tâches allant de la classification de documents à la recherche d'information, nécessite de prendre en compte la diachronie.

Suivant les travaux de Bengio *et al.* (2003) puis Mikolov *et al.* (2013), de nombreuses méthodes de représentations vectorielles de mots ont été mises au point depuis deux décennies. Elles permettent de représenter les mots par des vecteurs continus de faible dimension : les plongements lexicaux ou *word embeddings*. Néanmoins, ces plongements lexicaux reposent sur l'hypothèse que le sens d'un mot est inchangé sur l'ensemble du corpus. Cette hypothèse d'une représentation statique peut s'avérer limitée. Ainsi en supposant qu'un changement dans le contexte usuel d'un mot reflète un changement dans la signification de ce mot, il est possible d'entraîner des plongements de mots diachroniques : qui évoluent au cours du temps en suivant les changements d'usage des mots.

Récemment Rudolph & Blei (2018) ont proposé un tel modèle, nommé *Dynamic Bernoulli Embedding* (DBE). Il apprend des représentations de mots qui évoluent au cours du temps selon les strate temporelles d'un corpus, en caractérisant la dérive de ces représentation d'une strate à l'autre au moyen d'un processus aléatoire gaussien. Des choix de modélisation différents ont été effectués par d'autres auteurs dans la littérature. Les approches de Han *et al.* (2018) et Hamilton *et al.* (2016) impliquent d'apprendre des plongements lexicaux pour chaque strate temporelle sans les relier chronologiquement ; tandis que Kim *et al.* (2014) apprend les plongements diachroniques de façon incrémentale, mais sans contrôler la dérive de ces plongements.

Dans cet article nous prenons pour base le modèle DBE, qui présente un bon compromis entre simplicité et modulabilité, pour questionner l'importance de ces différents choix de modélisation. Dans ce but, nous analysons le comportement des plongements de mots appris à partir de ce modèle (décrit à la section 3) sur deux tailles de strates temporelles – mensuelle et annuelle – et l'appliquons (dans la section 4) à des corpus dans deux langues différentes : français et anglais. Les données en anglais proviennent du *New York Times Annotated Corpus* [1] (Sandhaus, 2008), qui s'étend de 1987 à 2006 ; le corpus en français est constitué d'articles du journal *LeMonde* collectés de façon à couvrir la même période. L'étude de ces deux corpus nous permet, dans un deuxième temps, d'étudier de façon conjointe l'évolution d'un mot à travers les deux langages.

2 État de l'art

Les premières méthodes automatiques d'étude de la diachronie se basent sur la détection de changements dans les co-occurrences des mots, puis sur des approches basées sur la similarité distributionnelle (Gulordava & Baroni, 2011) en construisant des mesures d'information mutuelle à partir de matriças de co-occurrences.

1. https ://catalog.ldc.upenn.edu/LDC2008T19

L'usage de méthodes d'apprentissage automatique basées sur les plongements lexicaux est récent et a connu une forte hausse d'intérêt depuis deux ans, avec la publication consécutive de trois articles dédiés à l'état de l'art de ce domaine (Kutuzov *et al.*, 2018; Tahmasebi *et al.*, 2018; Tang, 2018).

Dans un des premiers articles employant ce type de méthode (Kim *et al.*, 2014), les auteurs estiment des plongements lexicaux pour la première strate temporelle t_0 puis mettent à jour ces plongements pour les strates temporelles suivantes, considérant les plongements au temps $t-1$ comme initialisation pour la strate t. D'autres travaux ont ensuite vu le jour, reposant sur l'apprentissage de façon indépendante des plongements lexicaux pour chaque strate temporelle. Néanmoins, les plongements ainsi obtenus ne sont pas directement comparables car appartiennent à des espaces vectoriels différents. Deux approches sont alors envisageables : d'une part, déterminer la meilleure transformation linéaire afin d'aligner les espaces de représentation à travers les périodes (Hamilton *et al.*, 2016; Dubossarsky *et al.*, 2017; Szymanski, 2017; Kulkarni *et al.*, 2015) ; d'autre part, calculer la similarité cosinus entre chaque paire de mot à l'intérieur d'une strate temporelle, les similarités étant alors comparables d'une strate sur l'autre sans nécessiter d'alignement (Kim *et al.*, 2014).

Les méthodes dites dynamiques constituent un second type d'approche. Le corpus d'étude est toujours divisé en strates temporelles, mais cette fois les plongements lexicaux diachroniques sont appris de façon conjointe sur l'ensemble des strates. Ils sont ainsi placés dans un même espace de représentation dès l'apprentissage. Pour cela, Bamler & Mandt (2017) utilisent des modèles bayésiens d'apprentissage de plongements lexicaux : les vecteurs sont liées à travers les périodes à l'aide d'un processus de diffusion temporel qui contrôle leur évolution. Poursuivant le même objectif, différentes méthodes ont été proposées (Yao *et al.*, 2018; Rudolph & Blei, 2018; Han *et al.*, 2018) afin de mettre en évidence de façon jointe l'évolution continue du sens des mots. Ces méthodes permettent de s'affranchir de la limite de volume de données par strate temporelle lors de l'apprentissage.

La majorité de ces modèles sont évalués sur des corpus en anglais. À notre connaissance, bien que plusieurs auteurs ont expérimenté sur d'autres langues que l'anglais (Hamilton *et al.*, 2016; Eger & Mehler, 2016), aucun travaux n'a tenté de comparer l'évolution de mots à travers plusieurs langues à l'aide de méthodes de plongements diachroniques.

3 Modèles de plongements de mots dynamiques

Nous partons du modèle *Dynamic Bernoulli Embeddings* (DBE) de Rudolph & Blei (2018). Il se base sur les plongements de mots de la famille exponentielle (Rudolph *et al.*, 2016), qui sont une généralisation probabiliste du modèle *Continuous Bag-of-Words* (CBOW) de Mikolov *et al.* (2013). Nous en fournissons une brève description avant de présenter les différentes variantes mises en place.

3.1 Le modèle DBE

L'objectif de ce modèle est de prédire un mot à partir de son contexte. Afin de désigner un mot v parmi un vocabulaire de taille V, on considère un vecteur de variables aléatoires binaires $\mathbf{x}_v \in \{0,1\}^V$, où seule la composante associée au mot v vaut 1. Le mot v à la position i dans le corpus est donc représenté par le vecteur binaire $\mathbf{x}_{iv}$ et son contexte $\boldsymbol{c}_i$ est constitué des C mots avant et des C mots après (C étant la taille de la fenêtre). Ainsi, $\mathbf{x}_{\boldsymbol{c}_i}$ regroupe l'ensemble des points constituant le contexte du mot i. Le modèle DBE prédit le vecteur binaire du mot $\mathbf{x}_{iv}$ à partir de son vecteur de contexte $\mathbf{x}_{\boldsymbol{c}_i}$

selon la loi de Bernoulli suivante : $\mathbf{x}_{iv}|\mathbf{x}_{\boldsymbol{c}_i} \sim Bern(p_{iv})$. Le paramètre de la loi de Bernoulli p_{iv} est calculé à partir des plongements lexicaux $\boldsymbol{\rho}_v$ du mot à prédire et $\boldsymbol{\alpha}_{v'}$ des mots du contexte :

$$p_{iv} = \sigma\Big(\boldsymbol{\rho}_v^T(\sum_{j\in\boldsymbol{c}_i}\sum_{v'\in V}\boldsymbol{\alpha}_{v'}\mathbf{x}_{jv'})\Big). \tag{1}$$

Ainsi la somme sur v' sélectionne les plongements $\boldsymbol{\alpha}_{v'}$ des mots du contexte, qui sont ensuite additionnés (somme sur j) afin de créer un vecteur représentant le contexte $\boldsymbol{c}_i$. Le paramètre de Bernoulli résulte de l'application de la fonction sigmoïde σ au produit scalaire de ce vecteur avec le plongement $\boldsymbol{\rho}_v$ du mot à prédire.

Vers un modèle dynamique : Pour rendre ce modèle dynamique, considérons un corpus composé de T strates temporelles indicées par t. Dans chaque strate, chaque mot v a deux types de représentions : celle en tant que mot de contexte $\boldsymbol{\alpha}_v$, et celle en tant que mot central $\boldsymbol{\rho}_v$. Le vecteur $\boldsymbol{\alpha}_v$ est considéré comme invariant : il est commun à toutes les strates temporelles. Seuls les plongements $\boldsymbol{\rho}_v$ évoluent au cours du temps selon la marche aléatoire gaussienne suivante :

$$\boldsymbol{\rho}_v^{(0)} \sim \mathcal{N}(0, \lambda_0^{-1}I), \text{ puis pour } \forall t \geq 1,\ \boldsymbol{\rho}_v^{(t)} \sim \mathcal{N}(\boldsymbol{\rho}_v^{(t-1)}, \lambda^{-1}I). \tag{2}$$

Le paramètre λ, nommé *dérive*, est le même pour l'ensemble des strates et contrôle l'évolution du vecteur $\boldsymbol{\rho}_v$ d'une strate temporelle sur l'autre.

Apprentissage : L'apprentissage de ce modèle, plus précisément décrit par Rudolph & Blei (2018), s'appuie sur une variante de la stratégie du *negative sampling* (Mikolov *et al.*, 2013). L'objectif est d'optimiser la fonction suivante :

$$\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\alpha}) = \mathcal{L}_{pos}(\boldsymbol{\rho}, \boldsymbol{\alpha}) + \mathcal{L}_{neg}(\boldsymbol{\rho}, \boldsymbol{\alpha}) + \mathcal{L}_{prior}(\boldsymbol{\rho}, \boldsymbol{\alpha}) \tag{3}$$

Le premier terme $\mathcal{L}_{pos}$ représente la log-probabilité associée aux exemples positifs, tandis que le second ($\mathcal{L}_{neg}$) correspond à celle associée à des exemples négatifs tirés aléatoirement. Le troisième terme agit comme un terme de régularisation sur $\boldsymbol{\alpha}$ et sur la dynamique des plongements $\boldsymbol{\rho}$, et consiste à pénaliser le vecteur $\boldsymbol{\rho}_v^{(t)}$ lorsqu'il s'éloigne trop fortement du vecteur $\boldsymbol{\rho}_v^{(t-1)}$, de la manière suivante :

$$\mathcal{L}_{prior}(\boldsymbol{\rho}, \boldsymbol{\alpha}) = -\frac{\lambda_0}{2}\sum_v \|\boldsymbol{\alpha}_v\|^2 - \frac{\lambda_0}{2}\sum_v \|\boldsymbol{\rho}_v^{(0)}\|^2 - \frac{\lambda}{2}\sum_{v,t}\|\boldsymbol{\rho}_v^{(t)} - \boldsymbol{\rho}_v^{(t-1)}\|^2. \tag{4}$$

3.2 Variantes de régularisation

La première variante du modèle se rapproche du principe d'apprentissage incrémental proposé par Kim *et al.* (2014). Elle consiste à supprimer la régularisation sur la dérive des plongements de mots. Dans ce cas, la fonction de coût ne prend en compte que les deux premiers termes de la log-priore ainsi que $\mathcal{L}_{pos}$ et $\mathcal{L}_{neg}$. Par la suite, nous intitulons cette variante DBE-I (Incrémental).

La seconde variante consiste à abolir l'obligation de chronologie dans les vecteurs temporels successifs. En remplaçant la troisième composante de $\mathcal{L}_{prior}$ par $\sum_{v,t}\|\boldsymbol{\rho}_v^{(t)} - \boldsymbol{\rho}_v^{(0)}\|^2$, on force le vecteur $\boldsymbol{\rho}_v^{(t)}$ à rester proche du plongement d'origine $\boldsymbol{\rho}_v^{(0)}$. Ce principe est similaire à celui de Han *et al.* (2018), où les plongements de mots diachroniques sont appris de façon indépendante sur chaque strate temporelle. Cette variante est désignée par DBE-NC (Non Chronologique).

Une autre version de cette dernière fonction est mise en place, afin de prendre en compte l'éloignement temporel. Dans ce but, la troisième composante de la log-priore $\sum_{v,t} \|\boldsymbol{\rho}_v^{(t)} - \boldsymbol{\rho}_v^{(0)}\|^2$ est multipliée par un facteur temporel : le coefficient devient $-\frac{\lambda}{2} * t$ et permet de contrôler l'éloignement à la priore sans ajouter de dépendance entre les strates temporelles successives. Cette dernière version est nommée DBE-SC (Semi Chronologique).

4 Expérimentation

Nous expérimentons à partir de notre propre implémentation du modèle DBE en PYTORCH. Dans un premier temps, nous analysons de façon quantitative le comportement des différentes variantes du modèle DBE définies précédemment. Puis nous observons de plus près la dérive des mots ; en particulier, nous mettons en place un processus pour comparer les évolutions des mots dans deux langues de façon conjointe.

4.1 Données et hyper-paramètres

Les données du *New York Times Annotated Corpus* sont composées de 1 855 000 articles s'étalant sur une période d'environs 20 ans, du 1^{er} janvier 1987 au 19 juin 2007. Le journal *Le Monde* est un des quotidiens les plus lus en France ; nous en collectons des articles entre le 1^{er} janvier 1987 et le 31 décembre 2006. Ces deux corpus sont divisés en $T = 20$ strates temporelles annuelles [2] et $T = 240$ strates temporelles mensuelles.

Pour construire le vocabulaire, nous sélectionnons pour les deux langues $V = 40\,000$ mots selon leurs fréquences après avoir retiré les mots-outils. De même que Mikolov *et al.* (2013), nous sous-échantillonnons les mots fréquents en retirant chaque mot i avec une probabilité $p = 1 - \sqrt{\frac{10^{-5}}{\text{fréquence}(i)}}$. Dans le corpus *LeMonde*, le nombre moyen de mots par strate temporelle est d'environ 3.5 millions pour les strates annuelles et 300k pour les strates mensuelles. Dans le corpus *NYT*, ce nombre est d'environ 9 millions pour les strates annuelles et 750k pour les strates mensuelles. Le corpus est ensuite divisé en échantillons d'apprentissage, de validation et de test. Ces derniers comprennent chacune 10 % des données tirées aléatoirement. Les embeddings sont entraînés avec 1000 mini-batchs par strates temporelles pour l'analyse annuelle et 100 mini-batchs pour l'analyse mensuelle.

Les hyper-paramètres sont sélectionnés à partir de l'étude de la log-probabilité sur les exemples positifs $\mathcal{L}_{pos}$ calculée sur l'échantillon de validation de chaque corpus, à partir du modèle DBE classique. Afin de permettre la comparaison, les valeurs de $\mathcal{L}_{pos}$ sont mises à l'échelle selon la règle suivante :

$$\text{Échelle} = \frac{\text{Nb de mots dans l'échantillon de validation}}{\text{Nb de mots dans chaque mini-batch}}.$$

Dans un premier temps, le modèle est entraîné sur l'ensemble du corpus sans composante temporelle (modèle statique). Ainsi, les plongements lexicaux $\boldsymbol{\rho}$ et $\boldsymbol{\alpha}$ peuvent servir par la suite d'initialisation pour les modèles temporels. Suite à l'évaluation sur l'échantillon de validation, la fenêtre de contexte choisie est $C = 4$ [3]. La dimension des plongements lexicaux est de 100 et le nombre d'exemples

2. La dernière année du corpus *NYT* étant incomplète, elle n'est pas prise en compte dans l'analyse.
3. Deux mots précédant et deux mots suivant le mot central.

négatifs tirés pour chaque exemple positif est fixé à 10. Pour finir, la dérive $\lambda = 1$ est celle qui offre les meilleurs résultats. La dérive initiale λ_0 est fixée, comme le font Rudolph & Blei (2018), à $\frac{\lambda}{1000}$.

4.2 Évaluation quantitative

Dans un premier temps, nous analysons l'effet des différentes variantes de la fonction de coût sur les performances du modèle mesurées en terme de log-vraisemblance et sur la distribution des dérives des mots.

4.2.1 Évolution de la log-vraisemblance

Dans cette partie, nous calculons la log-probabilité du modèle DBE sur les exemples positifs $\mathcal{L}_{pos}$ sur les données de test de chaque corpus, *NYT* et *LeMonde*, pour les deux tailles de strates temporelles (Figure 1). Nous appliquons le terme d'échelle décrit dans la partie 4.1.

Dans un premier temps, les plongements lexicaux sont appris sur l'ensemble du corpus de façon statique. Puis ils sont utilisés sur chaque strate annuelle du corpus pour calculer $\mathcal{L}_{pos}$. La courbe obtenue est plus basse que la courbe associée au modèle dynamique appris en initialisant les vecteurs à partir de ce modèle statique. On constate logiquement que l'apprentissage dynamique permet aux plongements d'être adaptés à chaque strate temporelle, et donc plus efficace pour prédire les données de test.

À l'inverse, pour les deux corpus, le modèle dynamique sans initialisation a la performance la plus faible. Cette tendance est confirmée par la log-probabilité moyenne sur l'ensemble des strates temporelles (Table 1). Une explication se trouve peut être dans le faible volume de données sur chaque strate. Quand à la performance du modèle statique, elle dépend de l'homogénéité temporelle du corpus étudié ; nos deux corpus couvrent une plage de temps relativement faible, justifiant la performance du modèle statique par rapport à celle du modèle dynamique sans initialisation.

Comme le montre le tableau 1, les variantes du modèle définies par les différentes fonctions de coût ont des performances très proches ; l'ajout du coefficient de dérive croissant (DBE-SC) au modèle non chronologique (DBE-NC) permet d'augmenter légèrement sa performance, mais dans l'ensemble, c'est le modèle sans régularisation sur la dérive (DBE-I) qui obtient le score le plus élevé quelle que soit la taille de la strate temporelle.

L'étude de la log-probabilité ne reflète qu'une vision globale des performances du modèle ; afin de mieux comprendre son comportement, nous observons ensuite la distribution des dérives pour chaque variation du modèle.

4.2.2 Caractérisation de la dérive des plongmements

Dans le but d'analyser plus en détail le comportement du modèle et l'effet des variations de la fonction de coût, nous représentons les histogrammes superposés des dérives successives observées sur le corpus *LeMonde* (Figure 2). Les histogrammes pour le corpus *NYT* présentent des tendances similaires, de même que le cas des strates temporelles mensuelles. La dérive de chaque mot est calculée à partir de la distance euclidienne entre le plongement du mot au début du corpus $\boldsymbol{\rho}_v^{(t_0)}$ et ses

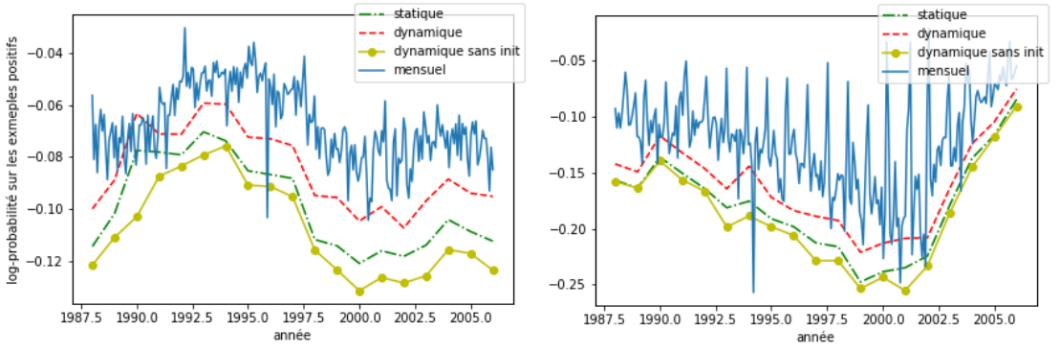

FIGURE 1 – Log-probabilités sur les exemples positifs sur l'échantillon de test du corpus *NYT* (à gauche) et *LeMonde* (à droite), à partir des modèles statique, dynamiques, et dynamiques sans pré-entraînement pour des strates annuelles, et dynamique pour des strates mensuelles.

	NYT		**Le Monde**	
	annuel	mensuel	annuel	mensuel
Statique	-0.09875	-0.06771	-0.1794	-0.1259
DBE	-0.08476	**-0.06720**	-0.1606	-0.1231
DBE sans initialisation	-0.10774	-0.07305	-0.1873	-0.1498
DBE-I	**-0.08448**	-0.06752	**-0.1593**	**-0.1227**
DBE-NC	-0.08517	-0.06817	-0.1607	-0.1236
DBE-SC	-0.08455	-0.06752	-0.1598	-0.1228

TABLE 1 – Log-probabilités moyennes sur l'ensemble des strates temporelles, sur l'échantillon de test des deux corpus, pour les différentes variantes d'apprentissage et de fonction de coût du modèle DBE.

plongements successifs $\boldsymbol{\rho}_v^{(t)}$ à chaque nouvelle strate temporelle t. Sur les histogramme, les couleurs plus claires correspondent aux distributions des dérives aux strates récentes : ainsi, la courbe la plus claire représente la distribution des dérives de mots calculées entre $t_0 = 1987$ et $t = 2006$ tandis que la plus sombre représente la distribution des dérives entre $t_0 = 1987$ et $t = 1988$.

Une première propriété intéressante est le caractère dirigé des dérives. Comme le montre le premier histogramme de la figure 2, les valeurs des dérives augmentent à travers le temps pour le modèle DBE classique. Cela signifie que le modèle capture principalement des dérives possédant une tendance, plutôt que de brefs changements de plongements suivis de retours à la normale. Ainsi le terme de régularisation décrit par l'équation 4 réalise bien le compromis attendu, en considérant comme partie de l'objectif la détection des grandes tendances d'évolution du sens des mots et en omettant leurs brèves variations.

Ces brèves variations sont dues à des évènements qui modifient temporairement le contexte dans lequel apparaît un mot sans avoir un impact à long terme sur son sens. Elles sont plutôt capturées par la version DBE-NC du modèle, dont l'histogramme ne présente pas d'évolution dirigée de la dérive en fonction de la distance à t_0, donc ne distingue pas ces "bruits" de la tendance générale d'évolution des mots. Pour finir, malgré l'absence de terme de régularisation sur la dérive, le modèle DBE-I capture naturellement une dérive relativement dirigée dans le temps bien que l'histogramme montre une plus grande sensibilité au bruit.

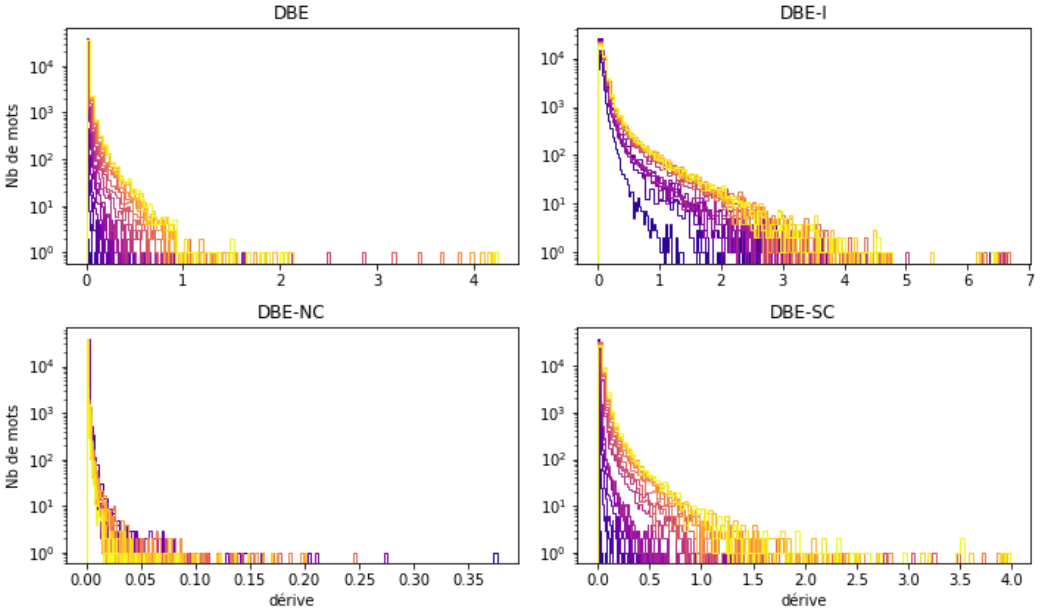

FIGURE 2 – Histogramme des dérives entre les plongements de mots à $t_0 = 1987$ et à chaque strate temporelle annuelle successive du corpus *LeMonde*, pour le modèles DBE et ses trois variantes. Plus la couleur est claire, plus la différence est calculée par rapport à une strate temporelle récente. Les nombres de mots (en ordonnée) sont en échelle logarithmique.

La seconde propriété mise en évidence est la capacité du modèle à distinguer les mots stables des mots dont l'usage évolue. En effet dans un intervalle de deux décennies, la majorité des mots est supposée peu évoluer. Le modèle DBE-NC, en introduisant une régularisation par rapport aux plongements de mots initiaux, permet de forcer le respect cette propriété : une grande part des mots sont presque invariants sur tout le corpus, et seule une sélection de dérives se démarque. Le modèle DBE classique permet aussi, dans une certaine mesure, de garder une faible dérive pour une grande partie des mots ; de même pour le modèle DBE-SC. Seul le modèle DBE-I ne distingue pas naturellement les mots qui dérivent peu.

4.3 Évaluation qualitative

La seconde étape de l'analyse est d'observer directement l'évolution des mots. À notre connaissance, il n'existe pas de corpus annotés permettant une évaluation directe des modèles diachroniques. Il est par contre possible d'observer l'évolution de certains mots choisis, permettant un premier diagnostic sous la forme d'une évaluation qualitative et subjective (Tahmasebi *et al.*, 2018). Ainsi pour chaque variante du modèle, nous observons les mots qui dérivent le plus pour mieux en comprendre le comportement. Puis, afin d'étudier de façon conjointe l'évolution des mots sur les deux corpus, nous mettons en place un processus d'analyse diachronique inter-langues.

NYT	Le Monde					
Annuel	Annuel			Mensuel		
DBE	DBE	DBE-I	DBE-NC	DBE	DBE-I	DBE-NC
google	euros	clearstream	royal	euros	rfa	seznec
skilling	ump	arcelor-mittal	sarkozy	sarkozy	euros	tramway
bloomberg	villepin	raimond	gdf	ump	ségolène	pinochet
email	rfa	shultz	euros	francs	sarkozy	hamas
katrina	sarkozy	ségolène	hezbollah	ségolène	ump	euros
cellphone	al-qaïda	outreau	liban	villepin	villepin	ahmadinejad
darfur	poutine	eads	thaksin	internet	monory	abbas
contras	gorbatchev	zapatero	ump	ue	climatique	révision
blog	katrina	villepin	islam	euro	contras	fibre
euros	internet	zidane	suez	bush	réévaluation	mahmoud

TABLE 2 – Listes des 10 mots ayant la plus grande dérive totale (distance entre la première et la dernière strate temporelle) pour les modèles DBE, DBE-I et DNE-NC sur le corpus *LeMonde* et DBE sur le corpus *NYT*. Les mots ayant un arrière-plan coloré sont communs à plus d'un modèle.

4.3.1 Analyse des fortes dérives

Nous nous concentrons ici sur le corpus *LeMonde*. La période étudiée ne couvrant que deux décennies, on observe principalement des évolutions de contexte liées aux événements ayant un impact médiatique ; les mots subissant de fortes dérives sont en majorité des entités nommées, et sont liés au contexte politique de la période, un thème récurrent dans ce journal.

Nous listons les 10 mots dont l'usage a le plus dérivé au cours des deux décennies selon chaque modèle dans le corpus, pour les deux tailles de strates temporelles. Du corpus *NYT*, nous ne reportons que les 10 mots ayant le plus varié d'après le modèle DBE classique sur des strates temporelles annuelles (Table 2).

Dans le cas du modèle DBE, pour les strates mensuelles et annuelles, les mots qui dérivent le plus sont associés à des concepts ayant subi un changement notable et continu au cours de la période (*euros*, *al-qaïda*, *internet*...). Cette observation est en accord avec la propriété observée précédemment au sujet du caractère dirigé des dérives. À l'inverse, les modèles DBE-I et DBE-NC annuels mettent principalement en valeur des dérives très fortes de mots liées à des événements uniques (*zidane*, *clearstream*, *royal*). C'est particulièrement le cas pour le modèle DBE-NC sur les strates mensuelles , où les dérives rapportées sont sujettes à un bruit important.

Afin de confirmer ces observations, le tableau 3 moyenne pour chaque modèle les rapports entre la dérive moyenne (sur toutes les strates temporelles) et la dérive totale (entre la première et la dernière strate) des 10 et 500 mots qui évoluent le plus. Notons que les valeurs des dérives moyennes normalisées ne sont pas directement comparables, selon que l'on considère les strates annuelles ou mensuelles (car le nombre de strates diffère). La valeur moyenne pour les 10 mots qui dérivent le plus est toujours plus faible que celle sur les 500 mots qui dérivent le plus. Les dérives importantes sont donc les plus dirigées quel que soit le modèle. Le modèle DBE-I présente des valeurs très proches de celles du modèle DBE-SC, montrant que l'absence de régularisation par rapport à la dérive permet, dans une certaine mesure, de conserver une certaine robustesse au bruit, en adéquation avec les observations de la figure 2.

Pour finir, remarquons que parmi les mots qui ont le plus dérivé au cours des deux décennies, certains

	Annuel		Mensuel	
	top 500	**top 10**	**top 500**	**top 10**
DBE	0.00642	0.00785	0.00356	6.280e-05
DBE-I	0.1152	0.0272	0.1149	0.0253
DBE-NC	0.5259	0.1054	0.0767	0.0530
DBE-SC	0.1096	0.0301	0.0715	0.0164

TABLE 3 – Valeurs moyennes pour les 10 mots et les 500 mots dérivant le plus, de leur dérive moyenne normalisée. Les dérives sont calculées pour les 4 variantes du modèle DBE sur le corpus *LeMonde*.

sont communs aux deux corpus (*euros*, *google*, *katrina*). Nous proposons donc par la suite une méthode pour observer l'évolution conjointe de ces mots dans les deux langues.

4.3.2 Analyse conjointe en anglais et français

Dans cette partie, nous étudions l'évolution d'un mots en français dans le corpus *LeMonde* et de sa traduction anglaise dans le corpus *NYT*. Les modèles de plongements de mots étant appris de façon indépendante sur ces deux corpus, les vecteurs ne sont pas directement comparables. Nous effectuons alors un alignement des deux espaces de représentation en utilisant un dictionnaire bilingue comme outil de supervision (Conneau *et al.*, 2018). Dans un premier temps, les plongements de mots des deux langues appris de façon statique sur l'ensemble du corpus sont normalisés ; puis, l'espace de représentation des plongements en français est aligné sur l'espace vectoriel des plongements en anglais. Nous choisissons ce sens car les données du *NYT* sur lesquelles les plongements anglais sont appris ont une volumétrie plus élevée, permettant des plongements lexicaux plus robustes.

Nous utilisons l'outil MUSE [4] pour l'alignement. La supervision est effectuée au moyen d'un dictionnaire bilingue construit à partir des vocabulaires des deux corpus. Nous sélectionnons tous les mots ayant un équivalent dans l'autre langue à partir du dictionnaire fournit par MUSE, puis ajoutons manuellement une sélection de mots spécifiques aux données (principalement des entités nommées). À partir des deux vocabulaires de 40 000 mots, nous obtenons un vocabulaire bilingue de 27 351 mots. Pour finir, les plongements sémantiques alignés $\boldsymbol{\rho}_{\text{align}}$ and $\boldsymbol{\alpha}_{\text{align}}$ sont utilisés pour initialiser les modèles dynamiques entrâinés sur chaque corpus.

Suite à l'apprentissage, pour chaque couple de mot dans le dictionnaire bilingue, nous calculons leurs dérives dans les deux corpus. Puis, nous calculons le cosinus comme une similarité entre les plongements des deux mots à la première et la dernière strate temporelle. Appelons cette valeur la similarité inter-langues. Nous calculons la dérive de cette similarité entre la première et la dernière strate temporelle, en mesurant la distance euclidienne entre ces deux valeurs.

En observant la distribution de ces grandeurs, nous mettons en évidence 4 types de comportement de mot à travers les deux langues :

1. Les mots qui dérivent dans la même direction dans les deux langues ;
2. Les mots qui dérivent dans les deux langages, mais dont le sens diverge (la similarité inter-langues décroît entre le première et la dernière strate temporelle) ;
3. Les mots qui dérivent dans une seule des deux langues, tandis que l'autre reste stable ;
4. Les mots qui sont stables dans les deux langues.

4. https ://github.com/facebookresearch/MUSE

Nous différencions les classes de mots en utilisant la moyenne des dérive des mots de chaque langue ainsi que la moyenne de la dérive de la similarité inter-langues, et reportons leur répartition au sein du vocabulaire bilingue dans le tableau 4. La majorité des mots appartiennent à la catégorie (4) (mots stables dans les deux langues), ce qui confirme la propriété du modèle DBE énoncée à partir de la figure 2 ; tandis que les mots qui évoluent dans les deux langues (catégories (1) et (2)) sont les plus rares.

Classe	(1)	(2)	(3-fr)	(3-en)	(4)
Pourcentage	5.4	5.5	16.1	15.2	57.8
Exemple	renouvelable	soviétique	francs	patrie	savon

TABLE 4 – Proportion des mots dans les différentes classes de comportement de dérive inter-langues, avec un exemple pour chaque classe.

Considérons à titre d'exemple le mot *Barbie*, qui appartient à la 3ème catégorie. L'espace aligné des plongements de mots est réduit à deux dimensions au moyen de la méthode t-SNE (van der Maaten & Hinton, 2008) pour représenter l'évolution de ce mot dans les deux langues (figure 3). Les mots les plus similaires au mot *Barbie* dans chaque langue, et à chaque strate temporelle, sont indiqués sur le graphique. En français (en rouge), le mot cible ne subit pas d'évolution notable. Il est majoritairement associé au criminel nazi *Klaus Barbie* et à son procès. Ces évènements ont eu une couverture médiatique moins importante et plus ponctuelle dans les journaux américains ; l'équivalent anglais du mot cible évolue rapidement en direction du champs lexical de la mode, en association avec la célèbre marque de poupée homonyme. Son plongement sémantique se stabilise dans ce voisinage à partir des années 2000.

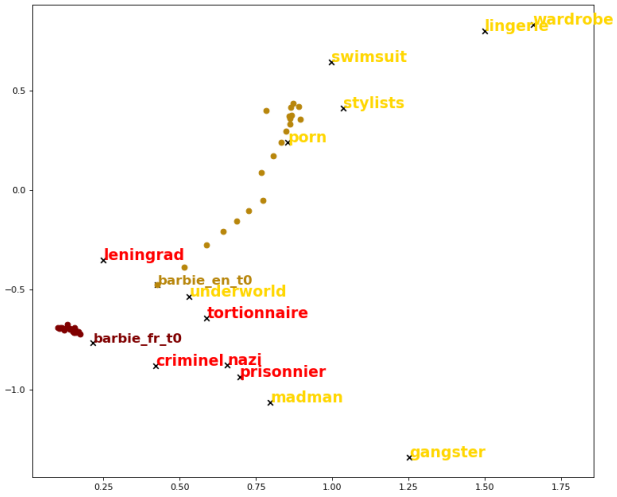

FIGURE 3 – T-SNE des espaces de plongements sémantiques alignés, représentant l'évolution temporelle du mot *Barbie* en français (rouge) et en anglais (jaune) ainsi que leurs plus proches voisins au cours du temps sur des strates annuelles.

5 Discussion

Dans cet article, nous étudions en détail le comportement d'un modèle d'apprentissage de plongements lexicaux dynamiques. Le point de départ de notre étude est le modèle *Dynamic Bernoulli Embeddings*, dont nous définissons plusieurs variantes. Nous répliquons ainsi le comportement d'autres modèles d'apprentissage de plongements de mots diachroniques de la littérature. Deux propriétés nous paraissent importantes à distinguer pour bien caractériser ces modèles : la capacité à mettre en évidence des évolutions dirigées des plongements de mots, et la capacité à garder une partie du vocabulaire stable au cours du temps. Nous montrons ensuite qu'il est possible d'analyser l'évolution d'un mot dans deux langues de façon conjointe. Un processus d'analyse préliminaire est mis en place en initialisant les modèles dynamiques à partir de plongements de mots alignés, et en analysant la dérive de la similarité inter-langues.

Le domaine en plein essor qu'est l'apprentissage de plongements de mots dynamiques manque encore de la cohésion que possèdent les tâches plus anciennes du traitement automatique des langues. Les publications sur ce sujet portent sur des corpus très diversifiés et les évaluations se font le plus souvent de façon qualitative, en l'absence de base d'évaluation robuste. Notons qu'un cadre d'évaluation est difficile à définir dans le cas qui nous intéresse, tant les attentes applicatives vis-à-vis d'un modèle diachronique peuvent varier. De plus, un cadre mathématique commun et rigoureux n'a pas encore été défini (Kutuzov *et al.*, 2018) et devrait s'appuyer sur les modèles d'apprentissage conjoint sur toutes les strates temporelles tel que celui décrit ici.

En effet, l'apprentissage conjoint à travers toutes les strates permet de s'affranchir dans une certaine mesure de la nécessité d'avoir un grand volume de données dans chacune d'elle. Néanmoins, le caractère discontinu des strates temporelles induit le modèle à détecter seulement les dérives d'une strate à l'autre ; plus les strates sont larges, plus les variations internes sont cachées, ou du moins moyennées. Ainsi, la question de la juste granularité temporelle se pose et dépend de l'application visée. Il est cependant important que les modèles étudiés puissent travailler à différents niveaux de finesse temporelle. Par exemple, lors de la recherche de dérives sémantiques brusques et de court terme, un type de modèle en temps continu (Rosenfeld & Erk, 2018) pourrait être plus adéquat, mais nécessite des informations temporelles très précises qui en pratique se retrouvent presque exclusivement dans les corpus issus de médias sociaux.

Une alternative est d'explorer l'usage de processus temporels de diffusion plus complexes, travaux initiés par exemple par (Bamler & Mandt, 2017) avec le processus d'Ornstein-Uhlenbeck. Enfin, l'emploi de strates temporelles non fixes dont les ruptures seraient apprises en même temps que les plongements lexicaux, assorti d'une régularisation sur la fonction de coût similaire à celle du modèle DBE-SC, serait une alternative à explorer. Néanmoins le cadre théorique approprié à cette sorte de quantification temporelle apprise par le modèle reste à définir.

Références

AITCHISON J. (2001). Language change : Progress or decay ? In *Cambridge Approaches to Linguistics*. Cambridge University Press.

BAMLER R. & MANDT S. (2017). Dynamic word embeddings. In D. PRECUP & Y. W. TEH, Eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings*

of Machine Learning Research, p. 380–389, International Convention Centre, Sydney, Australia : PMLR.

BENGIO Y., DUCHARME R., VINCENT P. & JAUVIN C. (2003). A neural probabilistic language model. In *Journal of Machine Learning Research*, p. 1137–1155.

CONNEAU A., LAMPLE G., RANZATO M., DENOYER L. & JÉGOU H. (2018). Word translation without parallel data. *CoRR*, **abs/1710.04087**.

DUBOSSARSKY H., WEINSHALL D. & GROSSMAN E. (2017). Outta control : Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1136–1145 : Association for Computational Linguistics.

EGER S. & MEHLER A. (2016). On the linearity of semantic change : Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 52–58, Berlin, Germany : Association for Computational Linguistics.

GULORDAVA K. & BARONI M. (2011). A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, p. 67–71 : Association for Computational Linguistics.

HAMILTON W. L., LESKOVEC J. & JURAFSKY D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1489–1501 : Association for Computational Linguistics.

HAN R., GILL M., SPIRLING A. & CHO K. (2018). Conditional word embedding and hypothesis testing via bayes-by-backprop. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 4890–4895 : Association for Computational Linguistics.

KIM Y., CHIU Y.-I., HANAKI K., HEGDE D. & PETROV S. (2014). Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, p. 61–65 : Association for Computational Linguistics.

KULKARNI V., AL-RFOU R., PEROZZI B. & SKIENA S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, p. 625–635, Republic and Canton of Geneva, Switzerland : International World Wide Web Conferences Steering Committee.

KUTUZOV A., ØVRELID L., SZYMANSKI T. & VELLDAL E. (2018). Diachronic word embeddings and semantic shifts : a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1384–1397 : Association for Computational Linguistics.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

ROSENFELD A. & ERK K. (2018). Deep neural models of semantic shift. In *NAACL 2018*.

RUDOLPH M. & BLEI D. (2018). Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, p. 1003–1011 : International World Wide Web Conferences Steering Committee.

RUDOLPH M., RUIZ F., MANDT S. & BLEI D. (2016). Exponential family embeddings. In *Advances in Neural Information Processing Systems*, p. 478–486.

SANDHAUS E. (2008). The new york times annotated corpus. In *Philadelphia : Linguistic Data Consortium*. Vol. 6, No. 12.

SZYMANSKI T. (2017). Temporal word analogies : Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 448–453 : Association for Computational Linguistics.

TAHMASEBI N., BORIN L. & JATOWT A. (2018). Survey of computational approaches to diachronic conceptual change. *CoRR*, **1811.06278**.

TANG X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering*, **24**(5), 649–676.

VAN DER MAATEN L. & HINTON G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

YAO Z., SUN Y., DING W., RAO N. & XIONG H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, p. 673–681 : ACM.