
MODÈLE ‘WORD2VEC’ POUR LA CONSTRUCTION DE VECTEURS DE MOTS

Projet final de master 1 linguistique-informatique Université de Paris-cité

Eléonora Khachaturova Armand Garrigou Léo Rongieras



Table of contents

1	Introduction	1
2	Approche théorique	2
3	Méthodologie et implémentation	5
3.1	Données d'entraînement	5
3.2	Prétraitement des données	6
3.3	Modèle CBOW	6
4	Resultats	8
4.1	Calculer la qualité des embeddings	8
5	Limitation et piste d'amélioration	12
	References	14

List of Figures

2.1	Architecture de modèle CBOW	4
4.1	visualisation des embeddings en 2 dimension	9
4.2	accuracy en fonction du nombre d'epoch sur une liste de tautologie	11

List of Tables

4.1	mots les plus proches de (King-men+women)	10
-----	---	----

1 Introduction

Le modèle Word2Vec a marqué une étape significative dans le domaine du traitement du langage naturel, en révolutionnant la manière d'obtenir des représentations du sens des mots. Depuis son introduction par Mikolov et al. (2013), Word2Vec est devenu l'un des modèles les plus influents et largement adoptés pour la représentation des mots dans les tâches de traitement automatique du langage.

L'objectif de ce rapport est de présenter en détail notre implémentation du modèle Word2Vec et de discuter des résultats obtenus lors de nos expérimentations. Nous aborderons en premier lieu les fondements théoriques du modèle.

Le modèle Word2Vec se base sur l'hypothèse distributionnelle, selon laquelle les mots ayant des contextes similaires ont tendance à partager des significations similaires. En exploitant de vastes corpus de textes, Word2Vec apprend des représentations vectorielles denses pour chaque mot, capturant ainsi les relations sémantiques et syntaxiques entre les mots. Ces représentations vectorielles, souvent appelées embeddings, ont été utilisées avec succès dans de nombreuses applications telles que la traduction automatique, la recherche d'informations et la classification de documents.

Dans notre projet, nous avons souhaité développer notre propre implémentation du modèle Word2Vec, afin de mieux comprendre son fonctionnement interne. Cette approche "from scratch" nous a permis d'explorer en profondeur les mécanismes de Word2Vec, depuis la création des fenêtres contextuelles jusqu'à l'entraînement du réseau neuronal.

Nous avons défini les objectifs suivants pour notre projet : (1) implémenter l'algorithme Word2Vec en utilisant le modèle CBOW, (2) entraîner notre modèle sur un corpus de texte de grande envergure, et (3) évaluer la qualité des embeddings appris, en utilisant différentes mesures de similarité sémantique et en effectuant des tâches d'analogie de mots.

2 Approche théorique

Le modèle Word2Vec propose deux architectures principales : Skip-gram et CBOW (Continuous Bag-of-Words). Dans notre implémentation, nous avons choisi de nous concentrer sur l'architecture CBOW. Contrairement au modèle Skip-gram qui prédit les mots environnants à partir d'un mot central, CBOW utilise le contexte environnant pour prédire le mot central. Cette approche nous permet d'apprendre des embeddings de mots en exploitant les relations contextuelles.

Par exemple en supposant une taille de fenêtre de deux : pour la phrase "le chat dort les souris chantent", si notre mot central est "dort" notre contexte sera défini par [le, chat, les, souris]. Comme le montre le schéma suivant :

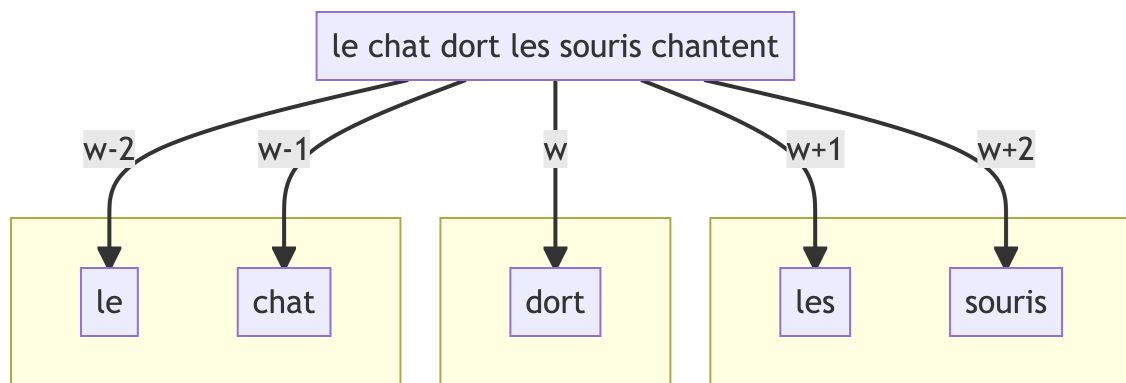


Schéma de la décomposition mot-cible/contexte

L'idée qu'un mot est déterminé par son contexte est parfaitement explicité dans les modèles de langue comme les modèles n -grams, on l'on assume que la probabilité de la présence d'un mot dans une phrase est déterminé par les mots qui précèdent, soit pour un modèle bi -gram :

$$P(w_1, w_2, \dots, w_n) = \prod_{i=2}^n P(w_i | w_{i-1})$$

.

L'intérêt de l'architecture CBOW est qu'elle permet d'exploiter aussi les mots qui apparaissent *après* notre mot cible. Comme le montre notre schéma, contrairement à un modèle n -grams traditionnel, le contexte *global* est pris en compte comme information. Un des apports majeur

2 Approche théorique

du modèle word2vec est donc le passage à la *bi-directionnalité*, prendre le contexte de gauche et de droite.

Nous allons donc voir comment CBOW peut apprendre ces probabilités.

Notre modèle prend en entrée un vecteur contexte x et retourne un vecteur mot y qui correspond au mot au centre de notre contexte. On définit deux matrices $\mathcal{A}^{(c)} \in \mathbb{R}^{|V| \times n}$ et $\mathcal{A}^{(w)} \in \mathbb{R}^{n \times |V|}$. On note :

- w_i le mot en position i du vocabulaire V
- $\mathcal{A}^{(c)} \in \mathbb{R}^{|V| \times n}$ la matrice des mots en entrée où a_i correspond à la i -ème ligne de $\mathcal{A}^{(c)}$ c'est à dire le vecteur qui représente le mot en entrée w_i
- $\mathcal{A}^{(w)} \in \mathbb{R}^{n \times |V|}$ la matrice en sortie où y_i correspond à la i -ème colonne de $\mathcal{A}^{(w)}$ c'est à dire le vecteur qui représente le mot en sortie w_i
- n correspond à la dimension arbitraire des embeddings

Les étapes du fonctionnement du modèle peuvent être décrites de telle sorte:

1. On crée un vecteur d'entrée composé des indices i des mots $\in V$ en contexte avec une fenêtre de taille N soit $v^c = (x^{c-N}, \dots, x^{c-1}, x^{c+1}, \dots, x^{c+N})$
2. On obtient nos embeddings pour ce contexte. Soit $\mathcal{A}^{(c)} v^c$, on obtient une matrice de taille $N \times n$ où chaque ligne i correspond à l'embedding du mot en contexte. Chaque vecteur dense est de dimension n
3. On veut récupérer la somme des vecteurs appartenant au contexte C , c'est à dire la somme des éléments **par colonne** de notre matrices $\mathcal{A}^{(c)} v^c$ soit

$$\hat{x} = \sum_{i=1}^{N \times 2} a_i$$

4. Notre vecteur score z est obtenu par

$$z = \mathcal{A}^{(w)} \sum_{i=1}^{N \times 2} a_i = \mathcal{A}^{(w)} \hat{x}$$

5. On retourne ce score transformé en log probabilité soit $\hat{y} = \log_softmax(z)$

On peut donc remarquer que après l'application du *softmax* on obtient un vecteur \hat{y} de la taille du vocabulaire V , où chaque \hat{y}_i correspond à la *log-probabilité* que \hat{y}_i soit le mot cible du contexte en entrée.

2 Approche théorique

On peut le vérifier algébriquement : en effet notre score est obtenu par le scalaire entre la matrice $\mathcal{A}^{(w)} \in \mathbb{R}^{n \times |V|}$ et le vecteur \hat{x} de taille n on a donc :

$$\begin{bmatrix} y_{1,1} & y_{2,1} & \cdots & y_{|V|,1} \\ y_{1,2} & \cdots & \cdots & \cdots \\ y_{1,3} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ y_{1,n} & \cdots & \cdots & y_{|V|,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = [\hat{y}_1 \quad \hat{y}_2 \quad \hat{y}_3 \quad \cdot \quad \cdot \quad \cdot \quad \hat{y}_{|V|}]$$

On obtient donc un vecteur \hat{y} de taille $|V|$

On peut résumer l'ensemble avec le schéma suivant¹ :

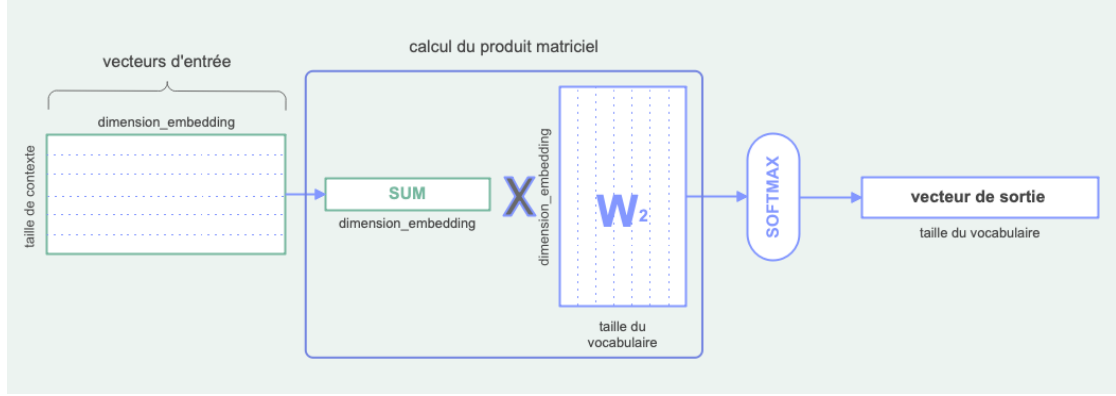


Figure 2.1: Architecture de modèle CBOW

en dernier lieu, le *softmax* prend en entrée le vecteur de score z de taille $|V|$ et renvoie un vecteur de même dimension, où chaque composant i est défini comme :

$$(\text{softmax}(z))_i = \frac{e^{z_i}}{\sum_{j=1}^{|V|} e^{z_j}}$$

¹On peut noter qu'il n'y pas dans l'architecture CBOW de fonction d'activation, la seule couche cachée du réseau correspond justement à la somme des embeddings des mots en contexte que nous avons décrit, le nombre de neurone correspond simplement au nombre de mot dans le vocabulaire.

3 Méthodologie et implémentation

Nous implémentons notre modèle avec les caractéristiques suivantes :

- utilisations de batch
- pas de negative sampling

En choisissant une implémentation en batch, on assure une optimisation de computation. Si l'on regarde la définition du *softmax* on peut se rendre compte que ce calcul est en fait computationnellement lourd, il parcourt l'ensemble du vocabulaire **pour tout mot**. Ce calcul est donc particulièrement lourd. En utilisant les batchs, on se permet de passer quelques exemples à chaque itération, réduisant la lourdeur de ce calcul. De fait, l'utilisation d'une telle technique va nous permettre d'entraîner notre modèle bien plus rapidement, et sur de plus gros corpus de texte.

Le negative sampling est une technique introduite par Mikolov et son équipe. Bien qu'il s'agisse aussi d'une technique d'optimisation computationnelle, nous ne l'implémenterons pas dans ce projet.

Notre implémentation repose sur l'utilisation du module [pytorch](#). Ce module propose un ensemble de classe et de fonction permettant d'optimiser nos tâches et nos modèles bien plus facilement. Nous verrons dans les parties suivantes comment nous avons utilisé ce module.

3.1 Données d'entraînement

Nous disposons de trois corpus d'entraînement. Deux sont de langue anglaise et directement issue de texte Wikipédia. Nous avons utilisés deux datasets Wikitext2 et Wikitext103¹ qui est une collection de plus de 100 millions de tokens.

Pour le corpus en langue française, nous disposons d'une partie du corpus *frcow*. Il s'agit d'extraction de texte issue du web. La particularité de ce corpus est qu'il a été pré-traité avec un travail de lemmatisation².

L'intérêt d'avoir sous la main deux corpus de bonnes tailles est de pouvoir comparer l'apprentissage de notre modèle sur deux langues différentes et de deux types différents, un corpus

¹Les datasets sont par exemple disponibles sur [Hugging Face](#)

²Nous remercions le professeur Olivier Bonami pour nous avoir transmis une partie de ce corpus comprenant son travail de lemmatisation

plain text (Wikitext103) et un corpus lemmatiser (frcow) pour les besoins d’une étude morphologique.

Comme toutes tâches de NLP, nous devons dans un premier temps préparer nos données à être passer dans le modèle, c’est que nous allons décrire dans la partie suivante.

3.2 Prétraitement des données

A partir des données nous avons besoins d’extraire :

- le vocabulaire : un set des mots présents dans notre corpus
- un map des mots par leurs indices dans le corpus : un dictionnaire `word2idx`
- une liste des mots trié par leurs indices : cette liste `idx2word` permet à partir d’un indice de récupérer le mot correspondant
- pour chaque mot parcouru sur notre corpus, on doit récupérer son contexte : un tuple `([context], target)`

Pour récupérer le vocabulaire, `word2idx` et `idx2word` nous utilisons un module particulier de pytorch : [torchtext](#). Nous pourrions récupérer ces informations avec des fonctions écrite “à la main”, ce que nous avons fait lors des premiers tests de notre projet, cependant, si l’utilisation et la documentation de `torchtext` est parfois peu explicite et intuitive, ce module permet d’optimiser grandement notre phase de prétraitement.

Ce module permet aussi de chargé un dataset, dont notre dataset Wikitext103, Ce qui est particulièrement pratique car il n’est donc pas nécessaire de télécharger en amont notre dataset, et notre programme est exécutable sur la plupart des machines disposant d’une connexion internet, sans se préoccuper de la présence ou de l’emplacement du dataset.

La fonction `build_vocab_from_iterator` de `torchtext` permet d’obtenir depuis un texte un objet `vocab`. `Torchtext` optimise cette opération et permet de récupérer tout ce dont nous avons besoin (`word2idx`, `idx2word`) à partir des attributs de cette objet `vocab`.

3.3 Modèle CBOW

Nous avons contruit notre modèle à partir de la classe de base `Module`³ proposée par pytorch. Nous pouvons visualiser l’ensemble de notre classe :

³voir [nn.Module](#) pour la documentation officielle

3 Méthodologie et implémentation

```
class CBOWModeler(nn.Module):

    def __init__(self, vocab_size, embedding_dim):
        super(CBOWModeler, self).__init__()
        self.embeddings = nn.Embedding(vocab_size, embedding_dim)
        self.linear1 = nn.Linear(embedding_dim, vocab_size)

        initrange = 0.5 / embedding_dim
        self.embeddings.weight.data.uniform_(-initrange, initrange)

    def forward(self, input):
        '''
        calcul de la somme des contextes à laquelle
        on applique la transformation linéaire (tensor : [1, len(vocab)])

        returns: log_softmax appliqué à la transformation

        '''
        embedding = self.embeddings(input)
        embedding = torch.sum(embedding, dim=1)

        Z_1 = self.linear1(embedding)
        out = F.log_softmax(Z_1, dim=1)

        return out
```

4 Resultats

4.1 Calculer la qualité des embeddings

Notre objectif dans cet exercice est d'obtenir des embeddings de bonne qualité, c'est à dire représentatifs du sens des mots qu'ils représentent. Pour déterminer si nos embeddings encodent des informations sémantiques, nous avons plusieurs moyen de procéder.

Nous allons nous baser sur la comparaisons de nos vecteurs entre eux, sur des taches particulières. Nous pouvons dans un premier temps nous contenter de visualiser la distribution de nos embeddings dans l'espace vectoriel. Nos embeddings étant de taille 200 nous devons d'abord passer par un algorithme de réduction de dimensions afin de ne garder que les 2 dimensions les plus importantes de notre espace vectoriel. Nous obtenons donc des embeddings de dimension 2 et nous pouvons les visualiser.

Afin de confirmer ces résultats, nous pouvons mettre de coté la réduction de dimension, qui implique forcément une perte d'information. Nous cherchons à utiliser une métrique fiable et constante pour calculer la proximité entre vecteurs. Traditionnellement on utilise la distance euclidienne ou la similarité cosinus.

La distance euclidienne est une mesure de la distance entre deux points dans un espace vectoriel à plusieurs dimensions. La distance euclidienne entre deux points A et B est calculée en prenant la racine carrée de la somme des carrés des différences entre les coordonnées correspondantes des points. Mathématiquement, la formule de la distance euclidienne est la suivante :

$$||AB|| = \sqrt{(A_1 - B_1)^2 + (A_1 - B_2)^2 + \dots + (A_n - B_n)^2}$$

Le calcul de similarité cosinus mesure l'angle entre deux vecteurs dans l'espace vectoriel. La similarité cosinus est une mesure de similarité normalisée qui varie entre -1 et 1. Une valeur de similarité cosinus proche de 1 indique une similarité élevée entre les vecteurs, tandis qu'une valeur proche de -1 indique une similarité inverse. Il est basé sur la formule mathématique suivante :

$$\cos(A, B) = \frac{(A \bullet B)}{(||A|| \times ||B||)}$$

Ces deux métriques sont pertinentes pour notre tache, elles présentent toutes deux des avantages et des inconvénients. La distance euclidienne permet de calculer les vecteurs les plus proches

4 Resultats

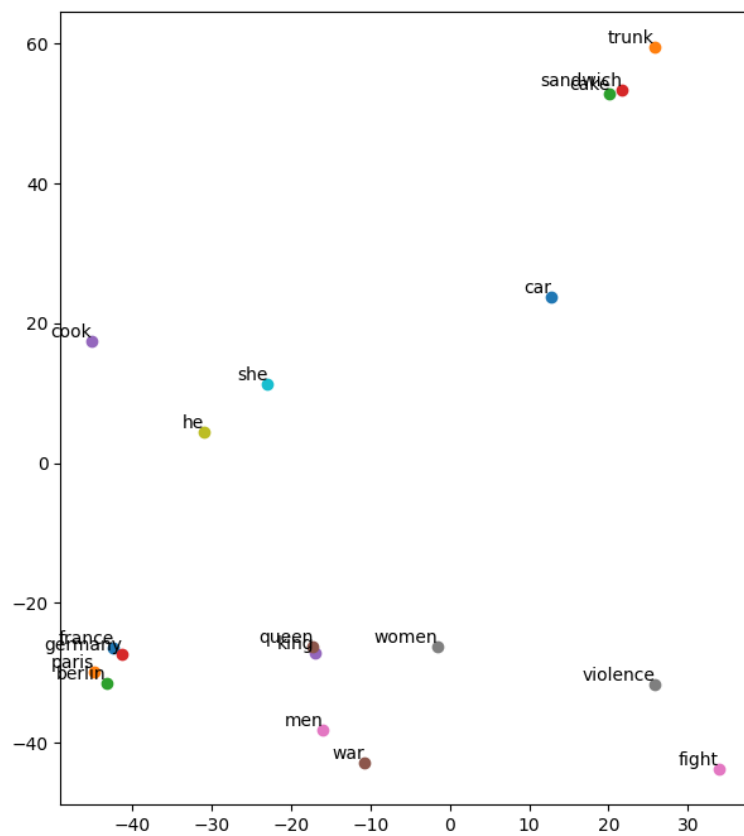


Figure 4.1: visualisation des embeddings en 2 dimension

(littéralement) d'un embedding en particulier. Toutefois, cette distance n'est pas entièrement bornée: elle peut être égale à 0 si deux vecteurs se superposent, mais n'a pas de limite naturelle supérieure. Se problème est résolu par la similarité cosinus qui est naturellement bornée entre -1 et 1. Le seul problème de cette métrique est que deux vecteurs peuvent avoir une similarité cosinus de 1 et pourtant ne pas avoir une distance euclidienne égale à 0. Traditionnellement on préfère utiliser la similarité cosinus.

FIGURE COMPARAISON DISTANCE/SIMILARITE

La tâche sur laquelle nous allons pouvoir vraiment évaluer la qualité de nos embeddings sont les tautologies. Elles sont utilisées dans l'article original qui introduit le modèle Word2Vec (Mikolov & al., 2013). Nos vecteurs sont représentables dans un espace euclidien et en respectent les règles:

- La distance entre deux points est toujours positive.
- La distance entre deux points est nulle si et seulement si les points sont identiques.
- La distance entre deux points est symétrique, c'est-à-dire que la distance entre A et B est la même que la distance entre B et A.
- La distance entre deux points obéit à l'inégalité triangulaire. Cela signifie que la distance entre deux points A et C ne peut jamais être plus courte que la somme des distances entre A et B, et entre B et C.

Ainsi on peut appliquer correctement les opérations mathématiques simples comme l'addition et la soustraction entre vecteurs. (Mikolov et al, 2013) montre que le modèle Word2Vec permet d'appliquer l'addition entre vecteurs pour combiner les sens de deux mots, et inversement avec la soustraction. Nous arrivons donc au fameux exemple de leur article:

$$ROI - HOMME + FEMME = REINE$$

Si nous vérifions cette égalité avec notre modèle:

Table 4.1: mots les plus proches de (King-men+women)

word	cosine-sim
monarch	0.49
kings	0.41
queen	0.41
nobility	0.4
prince	0.4

On peut voir que le mot "queen" apparaît en troisième position des mots les plus similaires au vecteur résultant du calcul algébrique.

4 Resultats

Nous allons à présent vérifier que notre modèle performe des résultats similaires sur une liste de tautologies créée par le groupe de travail de Mikolov. Nous considérons qu'une tautologie est validée si le vecteur attendu en sortie est dans les 5 embeddings les plus proches du vecteur résultant de la soustraction et de l'addition.

Les résultats sont bons, sachant que la random baseline de cet exercice serait de $5 \times \frac{1}{|V|}$, si nous considérons 5 exemples par tautologie. Nous pouvons essayer de comparer les performance de notre modèle selon le nombre d'itérations et aussi par rapport à FastText, qui fournit des embeddings entraînés sur la même architecture Word2Vec que la notre.

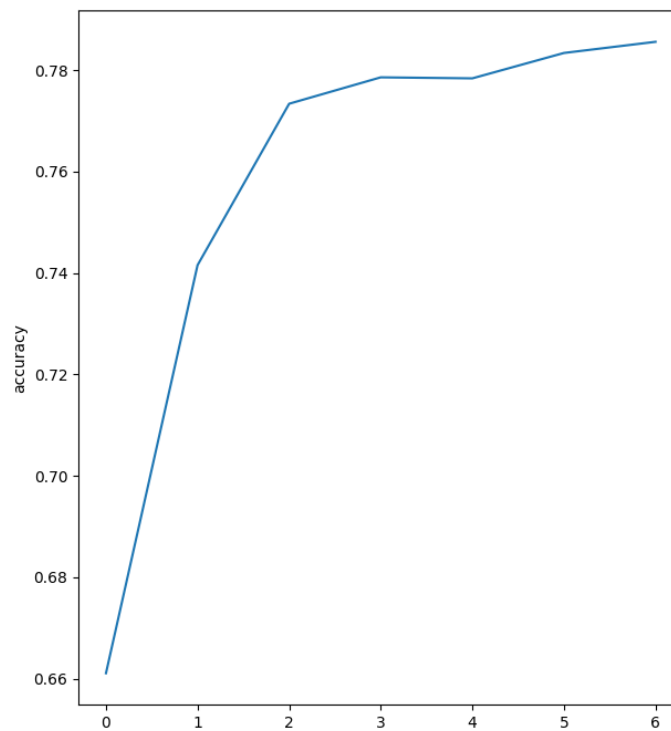


Figure 4.2: accuracy en fonction du nombre d'epoch sur une liste de tautologie

5 Limitation et piste d'amélioration

Comme vu précédemment, notre implémentation de word2vec atteint déjà de bonnes performances. Il existe toutefois de nombreuses pistes à explorer pour améliorer la qualité de nos embeddings.

Nous pourrions explorer plusieurs types de pré-traitement de notre corpus avant de lancer la vectorisation. Nous avons par exemple décidé de passer notre corpus en minuscule, pour réduire le bruit lié aux tokens en début de phrase, mais cela se fait au détriment d'une perte d'information sur les noms propres notamment: "Bordeaux" devient "bordeaux" et les sens se melangent. Toutefois nous avons pensé que cette opération nous coute moins qu'elle nous rapporte. D'autres techniques de nettoyage du corpus comme la lemmatisation, la suppression des caractères spéciaux ou des nombres pourraient être prises en compte. Le cout de pré-traitement est particulièrement long maintenant que nous somme capables de faire marcher notre modèle sur de grandes quantités de données.

Nous avons fait de nombreux tests sur les hyperparamètres afin de trouver ceux qui puissent allier bonne qualité d'embeddings et coût de l'entraînement (en termes de temps et de ressources). Nous avons notamment fait varier plusieurs hyperparamètres comme la taille de fenêtre, la dimensionnalité des vecteurs, le nombre d'itérations et le taux d'apprentissage. Toutefois nous n'avons pas gardé de trace de manière systématique des performance de notre modèle selon ces différents paramètres. Il faudrait à l'avenir procéder de façon plus stricte, en passant par exemple par une grid search pour trouver la combinaison de paramètres optimale, ou bien random search qui permettrait de trouver une combinaison suffisamment bonne en moins de temps.

Aussi, un axe d'amélioration serait d'augmenter la quantité de données, et de les diversifier. Notre modèle est déjà entraîné sur XXXXXXXX exemples, toutefois nous savons que plus les données sont conséquentes et variées et plus notre modèle pourra représenter fidèlement les mots de son vocabulaire. De manière générale, l'augmentation de la taille du corpus est souvent la première solution avancée. Toutefois, cela a un cout que nous ne jugions plus nécessaires vu l'état d'avancement du projet. Aussi, les données ajoutées au corpus d'entraînement doivent être de qualité: la diversité dans les exemples est primordiale car sinon les embeddings ne captureront pas totalement le sens des mots qu'ils encodent. Par exemple l'embedding de "robe" ne pourra pas encoder le sens de "couleur" d'un vin si il n'existe pas dans notre corpus d'entraînement de données parlant de vin. Cette problématique est présente dans le domaine de la désambiguïsation lexicale, que plusieurs de nos camarades ont expérimenté cette année en projet de TAL.

5 Limitation et piste d'amélioration

Une autre problématique est la gestion des déséquilibres lexicaux. Les mots fréquents ont mécaniquement une meilleure qualité de représentation puisqu'ils sont présents dans une quantité plus grande et plus de contextes diversifiés. A l'inverse, les mots rares auront une qualité de représentation amoindrie. Afin de contrer cela, nous pourrions ne considérer qu'une quantité bornée d'exemples par mot de vocabulaire, afin de permettre une couverture plus grande du corpus d'entraînement et ainsi augmenter le nombre d'exemples rencontrés pour les mots rares. Nous pourrions aussi implémenter le négative sampling, qui permet d'augmenter la quantité d'exemples par mot en associant un mot avec des mots aléatoires du vocabulaire et en soustrayant leurs sens (EXEMPLE)

Une des lacunes de notre implémentation vient aussi de la métrique d'évaluation. Elle se repose principalement sur la résolution de tautologies. Cette méthode est choisie car elle permet de juger de la qualité de plusieurs embeddings en même temps, tout en montrant qu'une addition/soustraction de vecteurs est similaire à une addition/soustraction de sens. Toutefois, cette métrique est biaisée, notamment car le corpus de ces tautologies comporte de nombreux cas de mots rares. Nous ne considérons pas les tautologies où l'embedding d'un mot n'est pas encodé par notre modèle, mais il subsiste de nombreux exemples de mots mal encodés qui ne valident pas les tautologies. C'est notamment le cas du sous corpus des tautologies sur les monnaies par exemple, où les mots sont assez fréquents pour être encodés mais pas assez pour avoir une représentation fidèle de leurs sens. Une autre proposition de métrique pourrait être de se baser sur un modèle dont on sait que les embeddings sont de bonne qualité. En effet, nous pourrions proposer une métrique qui calculerait l'embedding le plus proche (distance ou similarité) d'un mot x dans notre modèle et de déterminer si il est aussi présent dans les n embeddings les plus proches d'un modèle tiers. Nous pourrions par exemple nous baser sur les embeddings fournis par fastext, dont l'utilisation est simple et rapide.

Enfin, nous pourrions initialiser nos vecteurs non pas de manière aléatoire mais en récupérant directement ceux fournis par un modèle tiers. Nous pourrions ensuite fine-tuner les embeddings sur notre corpus d'entraînement. Cette technique nous permettrait de spécialiser nos vecteurs sur un corpus spécifique, ce qui pourrait être utile si nous avons besoin de vecteurs spécialisé sur un domaine en particulier. L'exemple du milieu viticole donné au dessus est une bonne représentation d'un cas où le fine-tuning donnerait de bon résultats. Pouvoir spécialiser des embeddings génériques tels que "robe", "mère", "cru", "sec", "moelleux".

References

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” <https://arxiv.org/abs/1301.3781>.