
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
2.1	Utilisation de Batch	4
3	Resultats	5
3.1	Calculer la qualité des embeddings	5
	References	9

List of Figures

2.1	Architecture de modèle CBOW	4
3.1	visualisation des embeddings en 2 dimension	6

List of Tables

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 :

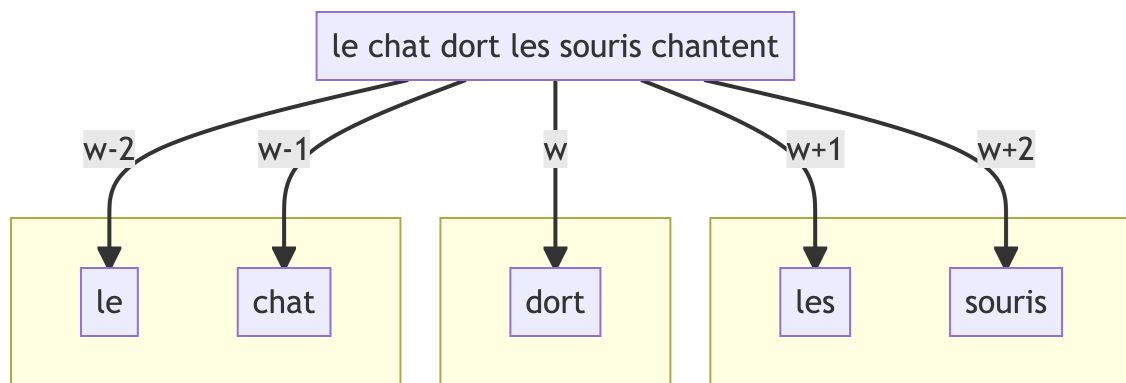


Schéma de la décomposition mot-cible/contexte

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 :

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

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

On peut résumer l'ensemble avec le schéma suivant¹ :

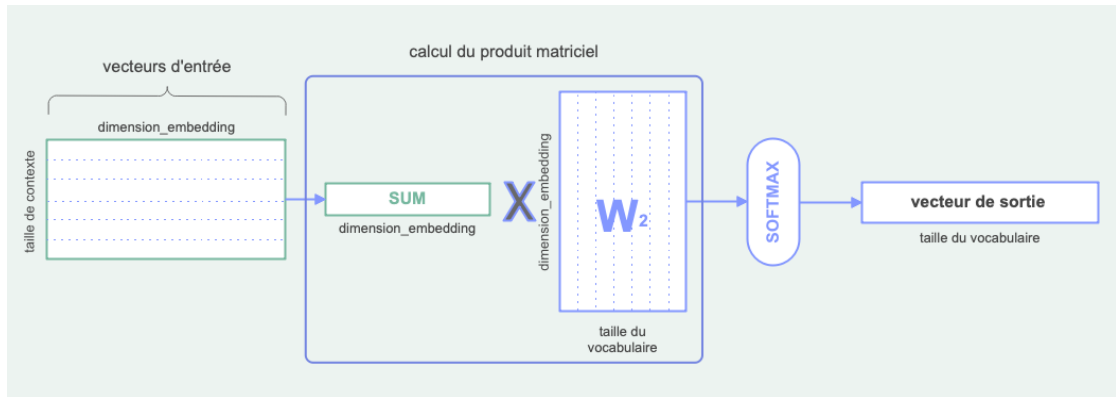


Figure 2.1: Architecture de modèle CBOW

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 :

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

2.1 Utilisation de Batch

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 Resultats

3.1 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

3 Resultats

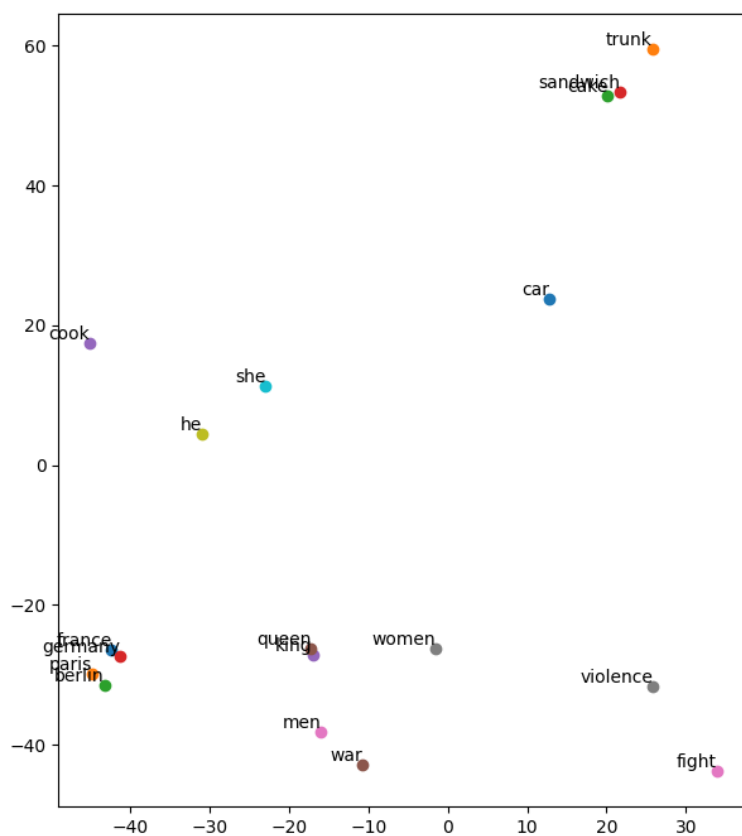


Figure 3.1: visualisation des embeddings en 2 dimension

3 Resultats

(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:

$$\text{ROI} - \text{HOMME} + \text{FEMME} = \text{REINE}$$

Si nous vérifions cette égalité avec notre modèle:

FIGURE TAUTOLOGY

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.

TAB Accuracy

Les résultats sont bons, sachant que la random baseline de cet exercice serait de $5 * 1/|\text{voc}|$, si nous considérons 5 exemples par tautologie. Nous pouvons essayer de comparer les performances 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.

3 Resultats

TAB/Graph accuracies

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>.