Word translations without parallel data

Résumé de l'article Conneau et al. 2017 par Rachel Keraron et Gabriel Melki

2018

Introduction

L'article Word translation without parallel corpora [1] cherche à résoudre la problématique de traduction de mots de manière non supervisée.
Pour cela, les auteurs procèdent aux étapes suivantes :

1. Pré-alignement de deux espaces d'embeddings monolingues via des **modèles adversaires**.
2. Alignement avec **Procrutes**
3. Evaluation selon un critère non supervisé utilisant **CSLS**.

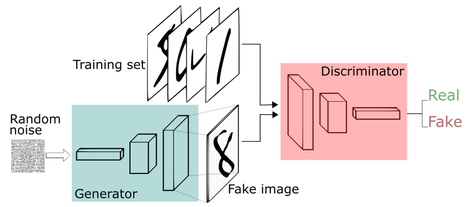
Les auteurs ont obtenu pour certaines langues (eng-it, it-en, en-es, es-en) des résultats supérieurs à la baseline, même supervisée. D'autres paires de langues (de-en, fr-en) se rapprochent de la baseline supervisée. Enfin certaines langues très différentes restent nettement en dessous (russe et anglais, par ex).

Choix des embeddings

Depuis l'apparition d'embeddings de mots (Word2vec, fastText), il a été remarqué que les espaces de représentations de mots d'une langue à l'autre partagent des structures similaires, même pour des langues très différentes. Mikolov et al. [3] ont par exemple tenté pour la première fois un alignement de ces embeddings supervisés.
Il y a deux paramètres importants lors du choix d'embedding :

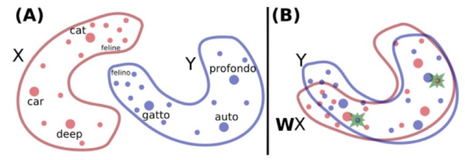
- la transformation elle même (fastText, Word2Vec, CBOW)
- les données d'entraînement de l'embedding

Modèles adversaires



Modèles adversaires génératifs

Les modèles adversaires génératifs ont été introduits par Goodfellow et al. (2014)[2]. On entraîne deux réseaux de neurones en parallèle, un discriminateur et un générateur (pour générer une image à partir de bruit, par exemple).
Le discriminateur doit déterminer si l'image est vraie ou fausse, le générateur a pour objectif de recevoir une classification positive du discriminateur sur les images générées.
L'entraînement de tels réseaux n'est pas simple en pratique : il faut éviter qu'un réseau devienne trop rapidemment beaucoup plus fort que l'autre. Il existe de nombreux paramètres (nombre d'itérations du discriminateur, nombre d'époques, taille du batch, learning rate) qui influent fortement les résultats.



Variante Mappeur et générateur

Les modèles adversaires ne sont pas nécessairement utilisés pour la génération.
Pour notre cas, le mappeur remplace le générateur et va transformer un embedding dans un espace à un autre. Le discriminateur doit reconnaître l'espace auquel appartient le mot (classification binaire source/target).
Cela conduit aux fonctions objectifs suivantes pour le discriminateur :

$$L_D(\theta_D/W) = -\frac{1}{n}\Sigma_{i=1}^{n} logP_{\theta_d}(source = 1/W_{x_i}) - \frac{1}{m}\Sigma_{i=1}^{m} logP_{\theta_d}(source = 0/y_i) \quad (1)$$

Et le mappeur :

$$L_W(W/\theta_D) = -\frac{1}{n}\Sigma_{i=1}^{n} logP_{\theta_d}(source = 0/W_{x_i}) - \frac{1}{m}\Sigma_{i=1}^{m} logP_{\theta_d}(source = 1/y_i) \quad (2)$$

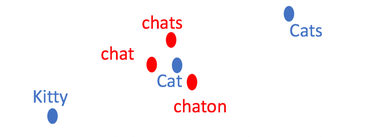
CSLS

Mesure de distance utilisée pour améliorer la construction de paires ($mot_{source}, mot_{target}$ (éviter les hubs des NN). Construction d'un graphe où chaque mot d'un dictionnaire est connecté à ses K mots les plus proches.
Il faut d'abord calculer une mesure de similarité entre un mot et son voisinage W_{x_s} :

$$r_T(W_{x_s}) = \frac{1}{K}\Sigma_{y_t \in N_t(W_{xs})} cos(W_{x_s}, y_t)$$

Le CSLS est alors égal à :

$$CSLS(W_{x_s,y_t}) = 2cos(W_{x_s}, y_t) - r_T(W_{x_s}) - r_s(y_t)$$



Le Mot Cat sera un hub et apparaitra dans 3 traductions alors que ni Kitty, ni Cats ne seront utilisés

Résultats

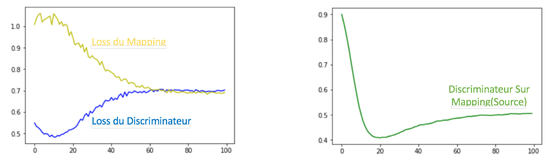


FIGURE 1: Evolution du loss du discriminator sur W.x sur une base très petite (500 mots)

Sélection de points d'ancrage et réalignement

Les auteurs sélectionnent les points d'ancrage à fournir à la méthode Procrustes.

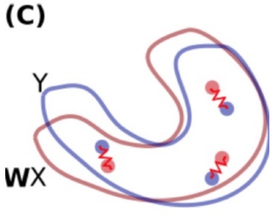
Points d'ancrage

La bonne sélection de points d'ancrages est nécessaire pour une bonne application de Procrustes. Le choix des plus proches voisins comme point d'ancrages ne suffit pas, car cette distance est asymétrique et pénalisée par les «hubs» : beaucoup de points seraient proches du même point. Il y a deux méthodes étudiés (la meilleure méthode étant la combinaison des deux) :

1. Les plus proches voisins mutuels.
2. Le critère CSLS.

Procrustes

On dispose donc maintenant de solides points d'ancrages et de deux espaces d'embedding rapprochés.



Références

[1] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou.
Word translation without parallel data.
arXiv preprint arXiv :1710.04087, 2017.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial nets.
In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig.
Linguistic regularities in continuous space word representations.
In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 746–751, 2013.