Apprentissage de transformation vectorielle pour la mise en correspondance de plongements lexicaux

Guillaume Wisniewski
guillaume.wisniewski@linguist.univ-paris-diderot.fr

décembre 2019

1 Représentation des plongements lexicaux

Dans la suite de TP, nous allons utiliser les bibliothèques suivantes :

— `sklearn` qui fournit une implémentation de TSNE, une méthode permettant de projeter des vecteurs dans un espace de plus petites dimensions en conservant, au mieux, les distances entre deux éléments. TSNE est la méthode généralement utilisée pour visualiser les plongements lexicaux.

— `spacy` une bibliothèque de TAL pour python qui fournit des modèles pré-entrainés pour de nombreuses langues (et, notamment, la possibilité d'obtenir les plongements lexicaux de mots).

Une fois ces bibliothèques installées (avec `pip`), il est possible de télécharger des modèles pré-entrainés pour `Spacy` à l'aide des commandes :

```
> ./local/bin/python -m spacy download fr_core_news_sm
> ./local/bin/python -m spacy download en_core_web_md
```

Il est alors possible d'obtenir la représentation d'un mot sous forme de vecteurs :

```
import spacy

fr = spacy.load("fr_core_news_sm")
f = fr("bonjour")

# f représente une phrase (liste de mots)
f[0].vector
```

① Représentez dans, un même espace, les représentations des mots suivants : cheval/horse/caballo, vache/cow/vaca, cochon/pig/cerdo, chien/dog/perro, chat/cat/gato. Qu'en concluez-vous ?

2 Transformation des plongements lexicaux

Les différentes méthodes de plongements lexicaux permettent de représenter les mots d'une langue dans un espace vectoriel dans lequel la distance entre éléments peut être interprétée comme une mesure de la similarité entre mots. Une idée naturelle, pour développer des méthodes multilingues est de définir une *transformation vectorielle* pour que la représentation d'un mot français soit proche de sa traduction en anglais.

Dans la suite de ce TP, nous proposons de considérer la projection linéaire définit par la matrice W qui permet de transformer un vecteur x de $\mathbb{R}^n$ représentant un mot français en sa représentation $W \cdot x \in \mathbb{R}^m$ dans l'espace de représentation des mots anglais.

② Quelle est la dimension de la matrice W ?

③ Quel est l'intérêt de définir une telle transformation ?

La matrice W peut être apprise en définissant un lexique initial $(e_i, f_i)_{i=1}^n$ contenant n mots anglais avec leur traductions en français et en cherchant la matrice W tel que :

$$\min_W \sum_{i=1}^{n} (W \cdot e_i - f_i)^2 \tag{1}$$

④ Comment pouvez-vous interpréter le critère d'apprentissage défini par l'Équation 1 ?

⑤ Trouvez la matrice W permettant de minimiser le critère d'erreur défini par l'Équation 1. Comment vous assurer que vous avez obtenu la « bonne » matrice ?

⑥ Comment peut-on déterminer des mots et leur traductions afin d' « apprendre » la matrice W. De combien de paires a-t-on besoin ?

⑦ Une fois la matrice W apprise, comment peut-on trouver la traduction de nouveaux mots ? comment évaluer l'approche ?

⑧ Tester cette méthode dans une situation « réelle ».