

Learning noun countability using minimally supervised (deep) learning

Cindy Aloui

LIF

Encadrant(s)

Carlos Ramisch et Alexis Nasr (co-encadrant)

Résumé

Résumé en français obligatoire.

Mots-clés : mot-clés obligatoires.

1 Introduction

A faire.

2 État de l'art

A faire

3 Description des données

Comme dis précédemment, nous souhaitons résoudre ce problème avec un minimum de supervision. Nous avons donc décidé d'utiliser pour l'entraînement seulement une liste de 200 mots typiquement massif et une liste de 200 mots typiquement comptable et du corpus frWaC qui a été annoté par le TreeTagger qui nous donne pour chaque mot sa forme lemmatisé et sa partie de discours.

Dans un premier temps nous avons séparé ces listes en un ensemble de mots pour l'entraînement et un autre pour le test (100 mots comptable et 100 mots massifs pour l'entraînement et la même chose pour le test).

Dans un second temps, nous avons choisi aléatoirement 100 noms qui apparaissent dans le frWaC et ensuite choisi à nouveau aléatoirement pour chacun de ces mots 100 phrases où il apparaît. Nous avons donné cet ensemble de phrases à une linguiste qui pour chaque mot a annoté 50 des phrases extraite et à indiqué si ce nom dans ce contexte précis est plutôt massif, comptable ou ni l'un ni l'autre.

Exemple :

C Le [[balai]] est tombé dans un trou de plusieurs mètres de profondeur .

Pour cette phrase le balai dans ce contexte a été annoté comptable.

Ces 5000 phrases vont nous servir lors de l'évaluation de notre classifieur.

4 Nom à trouver

Pour réaliser la tâche de classification massif/comptable nous avons penser à trois approches différentes :

- classifier le nom hors contexte en fonction des contextes dans lequel il apparaît (par exemple comme le nom table apparaît dans des contextes comptable, il serait classé en temps que nom comptable)
- classifier le contexte seul sans regarder le nom à classer
- classifier le contexte en utilisant aussi des informations lexicales du mot (éventuellement utilisé la classe du nom hors contexte)

Je vais par la suite entrer plus en détail sur les motivations de chacune de ces approches et expliquer les différentes implémentations.

4.1 Classifieur lexical

Nous avons tout d'abord pensé à classifié les mots hors contexte et à appliquer la classe prédite à tout les contexte où ce mot apparaît. Nous pensions que

ça permettrait d'avoir une première baseline assez rapidement, car la grande partie des mots sont soit très majoritairement massifs ou très majoritairement comptable, de plus c'est une information que nous pensons utile pour la suite en particulier pour la dernière approche.

Word Embedding

Nous avons tout d'abord pensé à utiliser des word embedding qui sont une représentation vectorielle des mots calculer à l'aide des contextes où les mots apparaissent. Cette représentation a la particularité que les mots apparaissant dans des contextes similaires possèdent des vecteurs proches, ce type de représentation semble donc utile pour notre tâche de classification. Nous avons donc calculé les word embedding des mots de tout le frWaC, à l'aide de la bibliothèque Python Gensim avec comme paramètres :

- une taille de 200, ce qui est une taille standard
- un fenêtre de 2, c'est à dire que pour l'on créer les word embeddings en prenant comme contexte deux mots avant et deux mots après
- skip-gram comme modèle pour apprendre les embeddings

Nous avons ensuite entraîné un classifieur KNN (Méthode des k plus proches voisins) implémenté dans la librairie Python Scikit-Learn sur les word embeddings des 100 mots massifs et les 100 mots comptables de nos données d'apprentissage. Le nombre k de voisins est déterminé à l'aide d'une validation croisée sur l'ensemble d'entraînement.

Ce système a été évalué de deux manières :

—

5 Consignes

Insérer ici le texte du rapport dans le format et la structure que vous voulez. N'oubliez pas de mettre à la fin du rapport la bibliographie.

Une fois que votre rapport est prêt, produisez-en une version pdf. Relisez-la avec attention. Vérifiez qu'il n'y a pas de problèmes d'accentuation.

Merci d'avoir bien voulu suivre ces instructions.