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

Dans de nombreuses langues dont le français un nom peut être comptable ou massif. Les noms comptables peuvent être "comptés" comme leur nom l'indique. Ils peuvent être au singulier ou au pluriel, par exemple : un stylo, deux stylos. Les noms massifs eux ne peuvent être comptés, ils ont donc seulement une forme singulier comme *eau* ou *sable*.

La comptabilité des noms est importante dans plusieurs tâches de traitement automatique du langage, notamment pour la traduction automatique. En effet, certains langages tels que le Japonais n'ont pas de système d'article ni de marque du pluriel, dans ces cas-là, il est utile de connaître la comptabilité des mots pour déterminer les articles à utiliser. Par exemple en japonais : "gohan o tabemasu" se traduira "je mange du riz" alors que "pasuta o tabemasu" se traduira "je mange des pâtes" alors que le seul mot changeant dans la phrase est *gohan*(riz) qui devient *pasuta*(pates).

La comptabilité peut aussi être utilisé pour lever l'ambiguïté de certain mots tel que gourmandise qui peut être comptable ou massif. Il peut donc être utile de pouvoir déterminer la comptabilité d'un contexte. Cela permet toujours en traduction automatique de connaître la traduction la plus pertinente, par exemple savoir si gourmandise sera plutôt traduit *greed* ou *delicacy* en anglais.

On peut aussi utiliser la comptabilité des mots pour détecter des erreurs dans l'utilisation d'article indéfini et de forme pluriel notamment chez les personnes n'ayant pas le français comme langue maternelle. Par exemple si un mot typiquement massif tel que sable se trouve dans un contexte comptable, il s'agit sûrement d'une erreur (Nagata et al., 2006).

Nous pensons que connaître seulement la classe d'un mot hors contexte que l'on appelera la **classe lexical** du mot n'est pas suffisant, en effet, il y a tout d'abord le problème des mots ambigus qui peuvent être massifs ou comptables en fonction du contexte comme gourmandise dont on a déjà parlé précédemment. On peut aussi rencontrer le cas de mots typiquement massif ou comptable qui sont employé dans des contextes ne correspondant pas à leurs classes, par exemple : "Il s'agit d'une **eau** contenant beaucoup de minéraux".

On peut aussi essayé de classifié le "contexte" c'est à dire les mots entourant le nom à classifier, par exemple si un nom est précédé de "un peu de" le contexte est plutôt massif, par contre si un mot est précédé de "plusieurs" le contexte sera comptable. Cependant cette **classe du contexte** ne suffit pas non plus, car un grand nombre de contextes ne sont pas discriminants, par exemple pour les phrase "Le tableau est beau" et "Le sable est beau" le seul mot changeant est le nom à classifié, dans le premier cas le nom est comptable et dans le deuxième cas le nom est massif.

On pense donc avoir besoin pour répondre à un plus grand nombre de problèmes de pouvoir classifier un mot hors contexte, un contexte seul et un mot dans un contexte.

Nous souhaitons aussi que notre système soit généralisable à d'autres grandes classes sémantique telle que concret/abstrait ou animé/inanimé, car cela permettrait de désambiguïser un plus grand nombre de mots, pour cela, il faudrait que notre système soit général et non pas spécialement adapté à la tâche de classification comptable/massif. De plus les annotations en contexte sont très onéreuse on souhaiterais donc limité ces annotations et donc que notre système n'ai besoin que d'un minimum de données supervisé pour que l'on puisse l'en-

traîner pour d'autres classes facilement.

Faire un paragraphe deep learning, word embedding. Et le plan.

2 État de l'art

Comme il existe un certain nombre d'applications, plusieurs chercheurs ont travaillé sur le problème de la comptabilité (pour l'anglais).

Baldwin and Bond (2003a ; 2003b) ont proposé une méthode pour apprendre la comptabilité des mots hors contexte à partir de données d'un corpus annoté. Ils ont distingués 4 classes dont massif et comptable qui sont fortement majoritaires, les noms peuvent avoir différentes classes, gourmandise par exemple serait classé comptable et massif. Il représentent les noms par un vecteurs de caractéristiques (tel que la proportion d'apparition au pluriel de ces noms), ces caractéristiques sont extraites à partir de données annotés. Ils utilisent 4 classifieurs (un pour chaque classe) entraînés grâce à l'algorithme KNN. Ils ont avec ce modèle 94,6% de réussite.

Lapatta et Keller (2005) ont proposé des modèles web pour certains problème du traitement de la langue dont le problème de la comptabilité des noms hors contexte. Ils ont tester un modèle qui prédit la comptabilité des nom en fonction de leur fréquence d'apparition au pluriel et leur fréquence d'apparition précédé de certain determinant. Ils obtiennent 88,62% de réussite sur les mots comptables et 91,53% de réussite sur les mots massifs avec cette méthode.

Pend et Araki (2005) ont proposé des modèles Web pour apprendre la comptabilité.

À notre connaissance il n'y a pas de travaux sur la comptabilité des mots français ni de modèle utilisant peu de données annotées et cherchant un modèle généralisable. Les autres travaux portant sur la comptabilité utilisent des informations spécialisées pour définir la comptabilité d'un certain nom comme par exemple la proportion d'apparition de ce nom au pluriel ainsi que sa proportion d'apparition précédé par un article indéfini qui sont des informations très adaptés à cette tache en particulier et ne permettent donc pas de généraliser cette approche à d'autres classes sémantique. De plus ces travaux tentent uniquement de définir seulement la classe lexical des mots. Les plus grandes différences entre notre travail et les travaux préexistant sont que nous allons utilisé des réseaux de neurones et des représentations vectoriels pour cette tâche, ce qui n'avait pas était fait à l'époque de plus nous allons travaillé sur très peu de données supervisés.

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 était 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 à 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 était 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 soit 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 utilisé des word embedding qui sont une représentation vectorielle des mots calculer à l'aide des contextes où les mots apparaissent. Cette représentation à la particularité que les mots apparaissant dans des contextes similaires possèdent des vecteurs proches, ce type de représentation semble donc utile pour notre tache 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 entraîner 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 croisé sur l'ensemble d'entraînement.

Ce système a été évalué de deux manières :
— Nous avons classé les 100 mots comptables et les 100 mots massifs de test, cela donne une précision de 90,3%, cependant il s'agit de mot que l'on considère "facile" à classer.
— Nous avons classé les 100 noms que nous avons choisis aléatoirement et appliqué leur classe prédite à toutes les phrases annoté et calculé la précision en fonction de l'annotation, ce qui donne 68,6% de réussite.

Nous avons testé une autre approche qui consiste pour chaque mot que l'on souhaite classer à faire la moyenne des cosinus pour tout les mots comptable et massifs de l'entraînement et de la classer dans la classe ayant la plus grande moyenne. L'évaluation est faite de la même manière que précédemment est les résultats sont :
— 85% sur les listes de mots massifs et comptables hors contexte.
— 66,9% sur les phrases annotées.

Cette approche ne s'avère pas concluante, cela est du au peu de données d'apprentissage, en effet les points les plus proches peuvent être en réalité très éloignés, de plus les word embeddings ne sont pas calculé pour cette tâche en particulier, ils ont donc des informations non utile et deux mots peuvent être proches car leur sens est proche (ils apparaissent donc dans des contextes semblables) mais leurs classes peuvent être différentes. C'est pour cela que nous avons décidé d'abandonner les word embeddings

Nom à trouver

Nous avons ensuite décidé d'utiliser le classifieur de contexte (nous reviendrons sur ce classifieur dans la section 4.2) pour donner un score de massivité aux lemmes. Ce classifieur nous donne un score de massivité associé à un certain contexte, pour un lemme donné nous faisons la moyenne des scores de massivité pour tout les contextes où il apparaît. Ce score de massivité est un score entre 0 et 1, on considère que si ce score est supérieur à 0.5, le mot est massif, s'il est inférieur à 0.5 le mot est comptable. Cette approche donne les résultats suivants :
#TODO