TP2 - Lemmatisation à l'aide d'un GRU

MASCO1 - apprentissage automatique

21 mars 2021

1 Objectif

L'objectif de ce projet est de programmer un réseau récurrent permettant de lemmatiser des mots. Etant donné les différentes formes que peut prendre un mot variable : *manger*,*manges*,*mangeront*,*mangeâmes* ..., le **lemme** est une de ces formes choisie de manière arbitraire. Dans le cas des verbes, il s'agit du verbe à l'infinitif, dans le cas d'un nom, d'un déterminant ou d'un adjectif, il s'agit du mot au masculin singulier.

Le réseau prend donc en entrée un mot (on dit en général une *forme*) pour produire le lemme associé à cette forme.

2 Données

Les données permettant d'entraîner le modèle sont composées de triplet *(forme, lemme, catégorie)*. La catégorie est une information importante pour déterminer le lemme d'une forme. En effet, dans certains cas, une forme se lemmatise différement selon sa catégorie. Par exemple : *(couvent, V)* $\rightarrow$ *couver*, tandis que *(couvent, N)* $\rightarrow$ *couvent*.

Les quatre fichiers suivants vous permettront d'entraîner et d'évaluer vos lemmatiseurs :

- `lemmatisation_train_N.txt`
- `lemmatisation_test_N.txt`
- `lemmatisation_train_V.txt`
- `lemmatisation_test_V.txt`
- `lemmatisation_train_A.txt`
- `lemmatisation_test_A.txt`
- `lemmatisation_train.txt`
- `lemmatisation_test.txt`

Les deux premiers fichiers ne comportent que des noms, les deux suivants que des verbes, les deux d'après que des adjectifs. Les deux derniers fichiers comportent des mots de toutes les catégories.

Les lignes de ces fichiers se présentent de la manière suivante :

```
traversent###### traverser####### V
```

comme on peut l'observer un certain nombre de dièses ont été ajoutés à la fin de la forme et du lemme. Le but de ces caractères suppélementaires est de s'assurer que la forme et le lemme possèdent le même nombre de caractères. La raison de cet ajout est que les réseaux que l'on va utiliser sont du type *transducteurs*, ils produisent un caractère en sortie pour tout caractère en entrée. Le principe d'ajout est le suivant. On ajoute une séquence de 6 dièses à la fin de chaque forme et un nombre variable de dièses à la fin du lemme, de manière à obtenir des chaînes de même longueur.

3 Lemmatiseur catégoriel

Le lemmatiseur catégoriel ne permet de lemmatiser que les formes appartenant à une catégorie. Pour l'entraîner, on ne lui fournit donc que des exemples relevant d'une catégorie donnée. La couche d'entrée du lemmatiseur catégoriel est donc composée de la concaténation des caractères qui constituent la forme à lemmatiser, ainsi que les dièses qui ont été ajoutés à la fin de la forme.

Vous trouverez dans le fichier `alphabet.txt` les différents caractères qui peuvent constituer les formes à lemmatiser (ainsi que les lemmes). Il y a en tout 66 caractères différent. Chaque lettre est donc représentée par un vecteur *one-hot* de dimension 66.

Afin d'encoder un caractère sous la forme d'un vecteur *one-hot*, vous pourrez utiliser la librairie `EncodeDecode` qui définit les deux méthodes `oneHotEncode` et `oneHotDecode`.

4 Lemmatiseur général

Le lemmatiseur général accepte en entrée des formes de toutes les catégories. La couche d'entrée du lemmatiseur général est composée, comme pour le lemmatiseur catégoriel, des représentations *one-hot* des caractères de la forme, ainsi que de la représentation *one-hot* de la catégorie.

5 Ce qu'il faut faire

1. Concevoir un réseau de type GRU pour le lemmatiseur catégoriel. Entraîner trois lemmatiseurs catégoriels, pour les trois catégories *Nom*, *Verbe* et *Adjectif*. Evaluer les performances du lemmatiseur sur les fichiers de test, faire une analyse des erreurs commises par ces lemmatiseurs.
2. Concevoir un réseau de type GRU pour le lemmatiseur général. Réflechir au meilleur moyen d'encoder la catégorie de la forme, afin qu'elle soit bien prise en compte par le réseau. Entraîner le réseau sur le fichier `lemmatisation_train.txt` et l'évaluer sur le fichier `lemmatisation_test.txt`. Donner les performances du lemmatiseur par catégorie et faire une analysee d'erreur.

3. L'étude est à rendre sous la forme d'un fichier pdf ou d'un notebook, à rendre pour le 11 avril 2021.