Certaines langues sont-elles plus difficiles à analyser que d'autres ?

Projet du cours M2 IAAA - Traitement du Langage Naturel

24 octobre 2019

1 Objectif

L'objectif de ce projet est d'analyser et de comprendre les raisons pour lesquelles des analyseurs syntaxiques partageant la même architecture et entraînés sur la même quantité de données obtiennent des performances très différentes sur différentes langues. On peut observer ce phénomène dans la Table 1 qui présente les performances calculées à l'aide des mesures LAS (Labeled Accuracy Score) et UAS (Unlabeled Accuracy Score) obtenues par un analyseur sur 36 langues différentes.

Comme on le voit, les performances (LAS/UAS) présentent une très grande variablilité, variant de 47.28%/55.20% pour le turc à 79.47%/86.80% pour le hindi.

2 Analyseur et données d'apprentissage

L'analyseur que vous utiliserez dans le cadre de ce projet est un analyseur en transition de type Arc-eager. Vous pouvez utiliser l'analyseur MACAON qui vous a été présenté en cours ou un autre analyseur. Dans un premier temps, vous utiliserez une version délexicalisée de l'anlayseur, ce qui veut dire que la forme des mots n'est pas utilisée pour prédire la structure syntaxique.

Les données sur lesquelles seront entraînés et évalués les analyseurs sont disponibles sur le site du cours. Il s'agit de données provenant du projet *Universal Dependencies*. Elles ont été constituées de manière à ce que les données d'apprentissage soient équilibrées : 20000 mots à peu près pour chaque langue.

L	LAS	UAS	L	LAS	UAS	L	LAS	UAS
hi	79.47	86.80	ru	69.70	73.85	sl	63.47	71.78
it	78.38	82.15	da	68.12	74.18	hr	63.58	72.10
ur	76.33	83.55	id	67.05	72.21	cs	63.84	72.45
pl	76.18	84.41	en	67.18	74.39	lv	62.30	69.83
ja	75.74	85.60	es	66.93	74.52	hu	62.73	68.86
no	73.25	78.91	uk	65.85	74.19	fi	62.77	70.83
bg	73.40	82.36	ro	65.13	72.53	zh	59.91	65.15
el	72.55	78.52	ga	65.13	74.02	vi	59.77	62.68
ca	72.06	79.70	fa	65.22	73.42	eu	58.80	68.78
sv	71.10	77.36	he	64.68	72.34	nl	57.44	68.43
fr	71.36	77.02	et	64.76	75.40	ko	53.12	63.21
pt	70.73	76.95	ar	64.28	71.65	tr	47.28	55.20

TABLE 1 – Labeled Accuracy Score (LAS) et Unlabeled Accuracy Score (UAS) pour 36 langues différentes dans des conditions d'apprentissage proches.

3 Analyse statistique

Comme nous l'avons mentionné dans l'introduction, l'objectif du projet est d'expliquer pourquoi observe-t-on les différences indiquées dans la Table 1. Pour cela, nous nous placerons dans le cadre de la regression multiple. Nous supposerons que la variable à expliquer, notée Y est le LAS calculé sur les corpus de test et que les variables explicatives X_i sont diverses observations effectuées sur les corpus d'apprentissage.

Votre objectif est de trouver des variables explicatives qui expliquent le plus possible la variabilité observée. Certaines variables explicatives sont observables directement sur les données annotées (longueur moyenne des phrases, des dépendances ...), tandis que d'autres sont plus difficiles à estimer, tel que la cohérence des annotations. On pourra s'inspirer pour cela de l'article suivant : *Divergences entre annotations dans le projet Universal Dependencies et leur impact sur l'évaluation de l'étiquetage morpho-syntaxique* que l'on trouvera sur le site du cours.

4 Ce qu'il faut faire

— Refaire les expériences qui ont mené aux résultats de la Table 1 en utilisant l'analyseur par transition de votre choix. Le fait de réaliser les expériences permet de modifier les conditions d'apprentissage de l'analyseur (en particulier les *features*) afin de tester certaines hypothèses.
— Extraire à partir des données d'apprentissage des variables explicatives et réaliser la regression multiple. C'est le cœur du travail. Vous pouvez commencer cette partie sans avoir encore un analyseur qui fonctionne, en utilisant les résultats de la Table 1.
— Si vous trouvez des variables explicatives qui ne sont pas couvertes par les features utilisées par l'analyseur, modifiez le jeu de features (fichier `fm` dans le cas de MACAON) et évaluez le nouvel analyseur.
— Rédiger un rapport de huit à dix pages décrivant le travail effectué ainsi que les réslutats obtenus.

Annexe : Le fichier `fm`

```
#les mots du buffer
b.-2#POS
b.-2#XPOS
b.-2#MORPHO
b.-2#GOV
b.-2#LABEL
b.-1#POS
b.-1#XPOS
b.-1#MORPHO
b.-1#GOV
b.-1#LABEL
b.0#POS
b.0#XPOS
b.0#MORPHO
b.1#POS
b.1#XPOS
b.1#MORPHO
b.2#POS
b.2#XPOS
b.2#MORPHO

#les mots de la pile
s.0#POS
s.0#XPOS
s.0#MORPHO
s.0#GOV
s.0#LABEL
s.1#POS
s.1#XPOS
s.1#MORPHO
s.1#GOV
s.1#LABEL
s.2#POS
s.2#XPOS
s.2#MORPHO
s.2#GOV
```

```
s.2#LABEL

#la distance entre les deux mots à relier (b0 et s0)
b.0#DIST.s.0

#l'historique des actions
tc.0
tc.1
tc.2
tc.3
tc.4

#on force la fin de phrase
s.0#EOS
```