# Analyse syntaxique multilingue

étudiant(e)1, étudiant(e)2, étudiant(e)3

12 novembre 2020

Ce document est un canevas de rapport pour le projet de TLNL. Il correspond à ce qui est attendu de vous.

Il n'est pas à suivre à la lettre, il est à considérer comme un guide.

Bonne courage !

## 1 Analyse monolingue

### 1.1 Les hyper-paramètres du classifieur

Vous décrirez ic les hyper-paramètres que vous avez choisi et comment vous avez fait pour les déterminer (on n'attend pas de vous de faire du grid search sur l'ensemble des hyper-paramètres)

Vous donnerez à la fin la structure du classifieur, par exemple son code en keras (si vous avez utilié keras) :

```
model = Sequential()
model.add(Dense(units=128, activation='relu', input_dim=inputSize))
model.add(Dropout(0.4))
model.add(Dense(units=outputSize, activation='softmax'))
model.compile(loss='categorical_crossentropy',
                  optimizer='adam',
                  metrics=['accuracy'])

model.fit(x_train, y_train, epochs=10, batch_size=32,
                  validation_data=(x_dev,y_dev))
```

### 1.2 La fonction de décomposition

Vous décrirez ici la fonction de décomposition que vous avez utilisé, vous décrirez les configurations que vous avez testé et quelles sont les features qui sont les plus porteuses d'informations.

Vous pourriez aussi dire se certaines fonctions sont plus adaptées à certaines langues (par exemple, parmi $n$ fonctions de décomposition, la meilleure pour le français n'est pas la meilleure pour l'arabe).

Vous donnerez à la fin la fonction de décomposition que vous avez choisi :

```
W B -2 POS
W B -1 POS
W B 0 POS
```

```
W B 1 POS
W B 2 POS
W S 0 POS
W S 1 POS
  ...
```

### 1.3 Les résultats

Vous mettrez dans le tableau 2 les résultats obtenus par les analyseurs monolingues. Je vous conseille de mettre dans ce tableau les résultats obtenus par tous les modèles, afin de pouvoir les comparer facilement.

Commentez les résultats obtenus par les analyseurs monolingues. Est ce que les performances sont très différentes selon les langues ? pourquoi à votre avis ?

## 2 Analyse multilingue

### 2.1 Le modèle $\Sigma$

Vous décrirez ici ce qui se passe lorsqu'on mélange tous les fichiers d'apprentissage sans dire à l'analyseur qu'il s'agit de langues différentes.

Décrivez ce qui se passe au niveau des résultats de la table 2 en comparant les modèles monolingues au modèle $\Sigma$ est ce que toutes les langues sont affectées de la même manière ? pourquoi ? essayez d'analyser le phénomène.

### 2.2 Le modèle $\Sigma ID$

La question qui nous intéresse ici est la suivante : est ce que le fait de dire à l'analyseur de quelle langue il s'agit permet de retrouver les performance des analyseurs monolingues ? pourquoi ? la situation est elle la même pour toutes les langues ?

### 2.3 Le modèle $\Sigma W$

L'analyseur n'a plus maintenant accès à la langue qu'il est en train d'analyser, mais il en possède une description partielle à partir d'un vecteur de 12 features issues du WALS.

Ces features sont à choisir avec soin.

#### 2.3.1 Hypothèses

Une manière de procéder consiste à regarder en détails ce qui se passe avec le modèle $\Sigma$ et repérer des dépendances qui sont mal prédites et qui ont des performances différentes selon les langues.

Vous pourrez présenter les valeurs de précision rappel et f-mesure pour ces dépendances sur certaines langues, pour étayer votre propos.

Vous pouvez alors choisir quels features vous semblent en rapport avec ces dépendances là.

Vous émettrez alors des hypothèses. Par exemple on s'attend que le trait $x$ du WALS ait un effet positif sur la dépendance $d$. Est ce le cas ? est ce le cas pour toutes les langues ? pourquoi ?

| | $L$ | | $\Sigma$ | | $\Sigma ID$ | | $\Sigma W$ | |
|---|---|---|---|---|---|---|---|---|
| L | LAS | UAS | LAS | UAS | LAS | UAS | LAS | UAS |
| ar | | | | | | | | |
| bg | | | | | | | | |
| ca | | | | | | | | |
| cs | | | | | | | | |
| da | | | | | | | | |
| de | | | | | | | | |
| el | | | | | | | | |
| en | | | | | | | | |
| es | | | | | | | | |
| et | | | | | | | | |
| eu | | | | | | | | |
| fa | | | | | | | | |
| fr | | | | | | | | |
| he | | | | | | | | |
| hi | | | | | | | | |
| hr | | | | | | | | |
| hu | | | | | | | | |
| id | | | | | | | | |
| it | | | | | | | | |
| ja | | | | | | | | |
| ko | | | | | | | | |
| lv | | | | | | | | |
| nl | | | | | | | | |
| nno | | | | | | | | |
| nob | | | | | | | | |
| pl | | | | | | | | |
| pt | | | | | | | | |
| ro | | | | | | | | |
| sl | | | | | | | | |
| sv | | | | | | | | |
| vi | | | | | | | | |
| zh | | | | | | | | |

TABLE 1 – Labeled Accuracy Score (LAS) et Unlabeled Accuracy Score (UAS) pour 32 langues différentes dans des conditions d'apprentissage proches.

#### 2.3.2 Matrice des features

Vous mettrez ici la liste des features que vous aurez choisi et leurs valeurs pour chacune des 32 langues.

Vous direz aussi commment vous avez géré les valeurs manquantes.

#### 2.3.3 Les résultats

Vous commenterez ici les résultats obtenus dans le tableau 2.

Il s'agit d'une première analyse générale.

### 2.4 Analyse

On entre ici dans les détails, en particulier des hypothèses que vous avez formulées dans la section 2.3.1 est ce que vos hypothèses se sont révélées correctes ou pas ? essayer d'aller le plus loin possible dans l'analyse.

## 3 Conclusion

Quelles conclusions tirez vous du travail que vous avez effectué et des performances obtenues.

| L | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F12 | F12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ar | | | | | | | | | | | | |
| bg | | | | | | | | | | | | |
| ca | | | | | | | | | | | | |
| cs | | | | | | | | | | | | |
| da | | | | | | | | | | | | |
| de | | | | | | | | | | | | |
| el | | | | | | | | | | | | |
| en | | | | | | | | | | | | |
| es | | | | | | | | | | | | |
| et | | | | | | | | | | | | |
| eu | | | | | | | | | | | | |
| fa | | | | | | | | | | | | |
| fr | | | | | | | | | | | | |
| he | | | | | | | | | | | | |
| hi | | | | | | | | | | | | |
| hr | | | | | | | | | | | | |
| hu | | | | | | | | | | | | |
| id | | | | | | | | | | | | |
| it | | | | | | | | | | | | |
| ja | | | | | | | | | | | | |
| ko | | | | | | | | | | | | |
| lv | | | | | | | | | | | | |
| nl | | | | | | | | | | | | |
| nno | | | | | | | | | | | | |
| nob | | | | | | | | | | | | |
| pl | | | | | | | | | | | | |
| pt | | | | | | | | | | | | |
| ro | | | | | | | | | | | | |
| sl | | | | | | | | | | | | |
| sv | | | | | | | | | | | | |
| vi | | | | | | | | | | | | |
| zh | | | | | | | | | | | | |

TABLE 2 – Labeled Accuracy Score (LAS) et Unlabeled Accuracy Score (UAS) pour 32 langues différentes dans des conditions d'apprentissage proches.