# Modèles proposés pour répondre à la FNC.

Enzo Poggio

1er mai 2018

## 1 Analyse des résultats des participants et création d'hypothèses.

Dans cette troisième partie nous tentons de faire une analyse d'erreurs comparative des participants de la FNC. Nous avons reproduit les résultats de tous les participants [1].

### 1.1 Un aperçu de l'ensemble de test.

| Stance | unrelated | agree | disagree | discuss | Somme |
|---|---|---|---|---|---|
| Nombre | 18349 | 1903 | 697 | 4464 | 25413 |
| Pourcentage | 72.20% | 7.48% | 2.74% | 17.56% | 100% |
| Score FNC | 4587.25 | 1903 | 697 | 4464 | 11651.25 |

TABLE 1 – Répartiotion des données dans l'ensembe de test.

Il y a 7064 points à marquer avec les *Related* et seulement 4587.25 points avec les *Unrelated*. Ainsi comme nous l'avions vu précédement trouvé un *Related* rapporte 4 fois plus de points que trouvé un *Unrelated*.

Afin de produire des hypothèses non-biaisées [2], nous allons faire des observations des résultats des différents modèles sur 80% des entrées de l'ensemble de test.

1. À l'exeption de Talos in the Swen qui ont laissé une copie de leur CSV de leurs résultats, pour leurs deux modèles sur leur répertoire Github.

2. Du moins des hypothèses qui ne soient pas ajustées totalement à l'ensemble de test final.

| Stance | unrelated | agree | disagree | discuss | Somme |
|---|---|---|---|---|---|
| Nombre | 14662 | 1568 | 550 | 3550 | 20330 |
| Pourcentage | 72.12% | 7.71% | 2.70% | 17.46% | 100% |
| Score FNC | 3665.5 | 1568 | 550 | 3550 | 9333.5 |

TABLE 2 – Répartition des données dans l'ensembe de test à 80%.

Le corpus de test étant partiellement ordonné nous avons retiré les 20% de manière ordonnée [3] aussi.

# 2 Les résultats et les scores des participants.

Ici nous présentons les différents résultats des modèles de chaque participant sur notre ensemble de test à 80%. Notez que les tables de confusions ont horizontalement les labels de vérité et verticalement les prédictions. Nous présentons les participants du premiers au derniers selon le score FNC.

## 2.1 Talos in the Swen.

Talos Complet

| | agree | disagree | discuss | unrelated | Somme |
|---|---|---|---|---|---|
| agree | 927 | 11 | 478 | 152 | 1568 |
| disagree | 217 | 8 | 235 | 90 | 550 |
| discuss | 661 | 5 | 2700 | 184 | 3550 |
| unrelated | 26 | 0 | 172 | 14464 | 14662 |
| Somme | 1831 | 24 | 3585 | 14890 | 20330 |

TABLE 3 – Table de confusion du modèle : Talos Complet

On remarque que la classe **disagree** est largement sous-représentée. Cette classe est aussi sous-représentée dans l'ensemble d'entrainement. Ce qui explique potentiellement pourquoi nous ne détectons pas ces traits distinctifs.

3. En effet, nous avons retiré toutes les entrées dont l'indice été divisible par 5 ou 10

Talos Complet

| Mesure | agree | disagree | discuss | unrelated |
|---|---|---|---|---|
| Précision | 0.59 | 0.01 | 0.76 | 0.99 |
| Rappel | 0.51 | 0.33 | 0.75 | 0.97 |
| F1score | 0.55 | 0.03 | 0.76 | 0.98 |
| Exactitude | 89.03 | | | |
| Score FNC | 7652.75 | | | |
| Pourcentage FNC | 81.99 | | | |

TABLE 4 – Mesures pour le modèle : Talos Complet

Bien que Talos in the Swen ait remporté la 1er place de la FNC, l'exactitudes et les scores ci-dessus ne sont pas les maximaux.

### 2.1.1 Les sous-modèles de Talos in the Swen.

#### 2.1.1.1 Sous-modèle arborescent de Talos in the Swen

Les sous-modèles de Talos n'ont pas eu de soumission à la FNC mais ils sont quand même intéressants à étudier car ils expliquent les résultats et les biais du modèles.

Talos Arborescent

| | agree | disagree | discuss | unrelated | Somme |
|---|---|---|---|---|---|
| agree | 807 | 0 | 690 | 71 | 1568 |
| disagree | 147 | 1 | 334 | 68 | 550 |
| discuss | 500 | 1 | 2902 | 147 | 3550 |
| unrelated | 16 | 0 | 160 | 14486 | 14662 |
| Somme | 1470 | 2 | 4086 | 14772 | 20330 |

TABLE 5 – Table de confusion du modèle : Talos Arborescent

On voit d'où le modèle principal tient son aversion du label **disagree**.

Talos Arborescent

| Mesure | agree | disagree | discuss | unrelated |
|---|---|---|---|---|
| Précision | 0.51 | 0.0 | 0.82 | 0.99 |
| Rappel | 0.55 | 0.5 | 0.71 | 0.98 |
| F1score | 0.53 | 0.0 | 0.76 | 0.98 |
| Exactitude | 89.5 | | | |
| Score FNC | 7749.5 | | | |
| Pourcentage FNC | 83.03 | | | |

TABLE 6 – Mesures pour le modèle : Talos Arborescent

Ce sous-modèles ne tient pas vraiment compte de la classe **disagree** mais il a paradoxalement tout de même les plus hauts scores du challenge. En effet si Talos avait proposé uniquement ce modèle il aurait pris encore plus d'avance par rapport aux autres participants.

#### 2.1.1.2 Sous-modèle de Deep Learning de Talos in the Swen

Talos DeepLearning

| | agree | disagree | discuss | unrelated | Somme |
|---|---|---|---|---|---|
| agree | 911 | 115 | 143 | 399 | 1568 |
| disagree | 271 | 62 | 41 | 176 | 550 |
| discuss | 1396 | 137 | 1397 | 620 | 3550 |
| unrelated | 2321 | 449 | 766 | 11126 | 14662 |
| Somme | 4899 | 763 | 2347 | 12321 | 20330 |

TABLE 7 – Table de confusion du modèle : Talos DeepLearning

Nous voyons clairement que ce sous-modèle vient équilibrer le sous-modèle arborescent. Sa table de confusion montre une tentative d'uniformisation de la classe **disagree**.

Talos deep

| Mesure | agree | disagree | discuss | unrelated |
|---|---|---|---|---|
| Précision | 0.58 | 0.11 | 0.39 | 0.76 |
| Rappel | 0.19 | 0.08 | 0.6 | 0.9 |
| F1score | 0.28 | 0.09 | 0.47 | 0.82 |
| Exactitude | 66.38 | | | |
| Score FNC | 5677.25 | | | |
| Pourcentage FNC | 60.83 | | | |

TABLE 8 – Mesures pour le modèle : Talos DeepLearning

Cette uniformisation a pour conséquence que ce sous-modèle est le pire de tous les modèles. Combiné avec le sous-modèle arborescent, il permet d'avoir une augmentation minime de la classe **disagree** au détriment des autres classes.

## 2.2 Le système Athene

| Athene | | | | | |
|---|---|---|---|---|---|
| | agree | disagree | discuss | unrelated | Somme |
| agree | 709 | 57 | 669 | 133 | 1568 |
| disagree | 193 | 56 | 193 | 108 | 550 |
| discuss | 388 | 30 | 2853 | 279 | 3550 |
| unrelated | 16 | 3 | 86 | 14557 | 14662 |
| Somme | 1306 | 146 | 3801 | 15077 | 20330 |

TABLE 9 – Table de confusion du modèle : Athene

La table de confusion d'Athene ressemble beaucoup à celle de Talos. Ce modèle tente beaucoup plus de labéliser des titres d'articles comme **diagree**. Mais il a beaucoup plus de mal à distinguer la classe **discuss** de la classe **agree**.

| Athene | | | | |
|---|---|---|---|---|
| Mesure | agree | disagree | discuss | unrelated |
| Précision | 0.45 | 0.1 | 0.8 | 0.99 |
| Rappel | 0.54 | 0.38 | 0.75 | 0.97 |
| F1score | 0.49 | 0.16 | 0.78 | 0.98 |
| Exactitude | 89.4 | | | |
| Score FNC | 7639.75 | | | |
| Pourcentage FNC | 81.85 | | | |

TABLE 10 – Mesures pour le modèle : Athene

Athene a la meilleure exactitude de toute la FNC car elle classe très bien les **unrelated** (qui ne valent pas beaucoup de points dans le score FNC).

## 2.3 UCL Machine Reader

| Uclmr | | | | | |
|---|---|---|---|---|---|
| | agree | disagree | discuss | unrelated | Somme |
| agree | 697 | 7 | 770 | 94 | 1568 |
| disagree | 144 | 37 | 277 | 92 | 550 |
| discuss | 424 | 35 | 2882 | 209 | 3550 |
| unrelated | 44 | 2 | 261 | 14355 | 14662 |
| Somme | 1309 | 81 | 4190 | 14750 | 20330 |

TABLE 11 – Table de confusion du modèle : Uclmr

Nous observons une tentative d'uniformisation proportionnelle de la table de confusion. UCLMR sait très bien classé les **unrelated**. Néanmoins, il manque apparemment de traits pour distinguer les **agree** des **discuss**.

| Uclmr | | | | |
|---|---|---|---|---|
| Mesure | agree | disagree | discuss | unrelated |
| Précision | 0.44 | 0.07 | 0.81 | 0.98 |
| Rappel | 0.53 | 0.46 | 0.69 | 0.97 |
| F1score | 0.48 | 0.12 | 0.74 | 0.98 |
| Exactitude | 88.4 | | | |
| Score FNC | 7619.0 | | | |
| Pourcentage FNC | 81.63 | | | |

TABLE 12 – Mesures pour le modèle : Uclmr

## 2.4 Analyses et Hypothèses

Voici un liste d'hypothéses basée sur nos constatations que nous testerons de vérifier par la suite avec les différents modèles. Nous voyons clairement une confusion entre la **agree** et la classe **discuss**. Mieux distinguer ces deux classe rapporterait gros Et pourrait faire un différence significative. La classe **disagree** est sous-représentée donc quantitativement elle ne rapporte que peu de points. Le modèls de Talos semble plus robuste que les autres en utilisant l'apprentissage par ensemble (*ensemble learning*). Utiliser ce type d'apprentissage entre les modèles des participants augmenterait certainement les performances.

# 3 Apprentissage par ensemble : Le vote de majorité

## 3.1 Explication du modèle.

Comme nous l'avons vu précédement les combinaisons entre les modèles déjà présentés à la FNC peuvent atteindre de hauts résultats. Notre premier modèle sera un jet naif de combinaisons de deux modèles. Le vote de majorité se base sur les classes prédites par les modèles des participants pour l'ensemble de test complet. Ce vote de majorité marche avec un modèle dominant départageant en cas de conflict entre les deux autres modèles assujettis. La classe la plus voté sera la classe prédite. Si jamais les trois classes sont toutes différentes alors est choisie la classe donné par le modèle dominant.

Vote de majorité entre 3 modèles

| Modèle Dominant | Modèle assujetti 1 | Modèle assujetti 2 | Classe prédite |
|---|---|---|---|
| **agree** | **agree** | unrelated | **agree** |
| disagree | **discuss** | **discuss** | **discuss** |
| **discuss** | disagree | agree | **discuss** |

TABLE 13 – Exemple du vote de majorité

Ainsi chaque modèle à tour de rôle peut devenir le modèle dominant. Ce qui donne les résultats dans la sous-sections suivantes.

## 3.2 Les résultats des votes de majorité

### 3.2.1 Dominant : Talos in the swen

Vote de majorité avec dominant : Talos

| Mesure | agree | disagree | discuss | unrelated |
|---|---|---|---|---|
| Précision | 0.49 | 0.03 | 0.84 | 0.99 |
| Rappel | 0.56 | 0.54 | 0.74 | 0.97 |
| F1score | 0.52 | 0.05 | 0.79 | 0.98 |
| Exactitude | 89.89 | | | |
| Score FNC | 9681.25 | | | |
| Pourcentage FNC | 83.09 | | | |

TABLE 14 – Mesures pour le modèle : Vote de majorité Talos

### 3.2.2 Dominant : Athene

Vote de majorité avec dominant : Athene

| Mesure | agree | disagree | discuss | unrelated |
|---|---|---|---|---|
| Précision | 0.48 | 0.07 | 0.84 | 0.99 |
| Rappel | 0.57 | 0.41 | 0.75 | 0.97 |
| F1score | 0.52 | 0.11 | 0.79 | 0.98 |
| Exactitude | 90.03 | | | |
| Score FNC | 9702.75 | | | |
| Pourcentage FNC | 83.28 | | | |

TABLE 15 – Mesures pour le modèle : Vote de majorité Athene

### 3.2.3 Dominant : UCL Machine Reader

Vote de majorité avec dominant : Uclmr

| Mesure | agree | disagree | discuss | unrelated |
|---|---|---|---|---|
| Précision | 0.48 | 0.05 | 0.84 | 0.99 |
| Rappel | 0.57 | 0.47 | 0.74 | 0.97 |
| F1score | 0.52 | 0.09 | 0.78 | 0.98 |
| Exactitude | 89.93 | | | |
| Score FNC | 9698.75 | | | |
| Pourcentage FNC | 83.24 | | | |

TABLE 16 – Mesures pour le modèle : Vote de majorité Uclmr

### 3.2.4 Commentaires sur les résultats du vote de majorité

Bien que ce modèle d'apprentissage par ensemble semble être le plus naïf ; ces résultats sont les plus haut que l'on obtiendra. On remarque que Athene est le meilleur des modèles dominants. C'est le modèle qui se trompe le moins en cas de conflit.

# 4 Apprentissage par ensemble : Par moyenne

## 4.1 Modèles utiliseant *50/50 weighted average*

Pour ce modèle, deux sous-modèles génèreront des probalités de label pour chaque entrée. Nous nous s'inspirerons de la méthode d'unification de Talos. Nous utiliserons donc aussi un *50/50 weighted average* pour combiner les résultats. Ainsi chaque sous-modèles produira un vecteurs de dimension 4 pour chaque labels. Chaque couple de même dimensions passera par un *50/50 weighted average*. Le maximum des 4 résultats des moyennes indiquera la classe prédite pour l'entrée.

Notre première architecture utilisera le sous-modèle arborescent de Talos et le modèle de UCL Machine Reader. Nous appelerons ce modèles « mixte UCLMR/Talos TF-Idf ».

Notre deuxième architecture utilisera aussi le sous-modèle arborescent de Talos et le modèle de UCL Machine Reader mais dans cette version le sous modèle arborescent de Talos n'utiliseras pas son de TF-Idf. La raison est que le système de UCL utilise déjà et uniquement ces traits-là. Ainsi pour augmenter la différence entre les résultats des deux sous-modèles, nous décidons de retirer ce module. De plus il sera entrainé seleument sur les données d'entraînement pour créer son espace sémantique [4]. Nous appelerons ce modèles « mixte UCLMR/Talos sans TF-Idf ».

4. En effet dans la version originale du modèle lors de la création de l'espace sémantique les données de test viennent augmentées les données d'entrainement. Cela a pour conséquence que le modèle ait déjà vu les données de test lors de son entrainement. Mais ne nous méprenons pas cela est autorisé dans le réglement de la FNC même si cela est assez questionnable.

## 4.2 Résultats

### 4.2.1 Modèles mixte UCLMR/Talos TF-Idf

Mixte UCLMR/Talos TF-Idf

| | agree | disagree | discuss | unrelated | Somme |
|---|---|---|---|---|---|
| agree | 963 | 0 | 819 | 121 | 1903 |
| disagree | 198 | 1 | 377 | 121 | 697 |
| discuss | 606 | 3 | 3612 | 243 | 4464 |
| unrelated | 21 | 0 | 166 | 18162 | 18349 |
| Somme | 1788 | 4 | 4974 | 18647 | 25413 |

TABLE 17 – Table de confusion du modèle : Mixte UCLMR/Talos TF-Idf

Cette table de confusion ressemble beaucoup à celle de Talos dans l'analyse sur 80% du corpus. Sans surprise, la classe disagree n'est pas améliorée. Par contre, les classes agree, discuss et unrelated sont légèrement augmentées.

Mixte UCLMR/Talos TF-Idf

| Mesure | agree | disagree | discuss | unrelated |
|---|---|---|---|---|
| Précision | 0.51 | 0.0 | 0.81 | 0.99 |
| Rappel | 0.54 | 0.25 | 0.73 | 0.97 |
| F1score | 0.52 | 0.0 | 0.77 | 0.98 |
| Exactitude | 89.47 | | | |
| Score FNC | 9617.25 | | | |
| Pourcentage FNC | 82.54 | | | |

TABLE 18 – Mesures pour le modèle : Mixte UCLMR/Talos TF-Idf

Nous constatons un gain de 0.52% par rapport au meilleur des systèmes de la FNC. Ceci est peu mais montre une amélioration minime de la précision de la classe agree et de la classe discuss. La classe disagree est encore négligée par le modèle arborescent de Talos. En conséquence, le modèle UCLMR biaisé par Talos ne permet pas d'améliorer la précision de la classe disagree.

### 4.2.2 Modéles mixte UCLMR/Talos sans TF-Idf

ixte UCLMR/Talos sans TF-Idf

| | agree | disagree | discuss | unrelated | Somme |
|---|---|---|---|---|---|
| agree | 1002 | 0 | 776 | 125 | 1903 |
| disagree | 216 | 1 | 357 | 123 | 697 |
| discuss | 602 | 2 | 3596 | 264 | 4464 |
| unrelated | 16 | 0 | 149 | 18184 | 18349 |
| Somme | 1836 | 3 | 4878 | 18696 | 25413 |

TABLE 19 – Table de confusion du modèle : mixte UCLMR/Talos sans TF-Idf

Par rapport au modèle mixte UCLMR/Talos TF-Idf, ce modèle améliore la classe des **agree** d'une quarantaine d'entrée. C'est certes un avantage minime, qui n'en reste pas moins surprenant. En effet le modèle avec TF-Idf a un espace sémantique plus exhaustif et plus proche de l'ensemble de test.

Mixte UCLMR/Talos sans TF-Idf

| Mesure | agree | disagree | discuss | unrelated |
|---|---|---|---|---|
| Précision | 0.53 | 0.0 | 0.81 | 0.99 |
| Rappel | 0.55 | 0.33 | 0.74 | 0.97 |
| F1score | 0.54 | 0.0 | 0.77 | 0.98 |
| Exactitude | 89.65 | | | |
| Score FNC | 9633.25 | | | |
| Pourcentage FNC | 82.68 | | | |

TABLE 20 – Mesures pour le modèle : mixte UCLMR/Talos sans TF-Idf

Certes l'avantagesur le modèle précédent n'est que de 0.16% pour les pourcentage FNC. Peut-être montrons-nous qu'un espace sémantique simplement formé avec l'ensemble d'entrainement est suffisant.

# 5 Apprentissage par ensemble : Avec *Single Layer Learner*

## 5.1 Explication du modèle

Une des meilleurs manière *a priori* de déterminer une classe à partir d'une liste de probabilité est certainement d'utiliser un couche neuronale. Ainsi

nous avons créer avec le framework Keras un *Single Layer Learner*. Celui-ci dispose d'un couche d'entrée de la longueur de nos vecteurs d'entrainement, d'un couche de 32 unités d'apprentissage et d'une couche de sortie de 4 unités, une par label. Afin de produire les vecteurs d'entrainement nous avons utiliser l'intégralité du corpus d'entrainement. Chaque dixième des vecteurs on était produits par les 90% restant du corpus. Les données qui ont servi à produire les vecteurs d'entraînement sont les vecteurs concaténé du modèle Mixte UCLMR/Talos sans TF-Idf. Nous nommerons ce modèle Mixte UCLMR/Talos sans TF-Idf SLL.

### 5.1.1 Code du modèle

```
import numpy as np
import keras
from keras.models import Sequential
from keras.layers import Dense, Activation
import tensorflow as tf


def create_trained_nn(vecs_train, labels_train, epochs=50):
    """Create a trained neural network model."""
    input_dim = vecs_train.shape[1]
    model = Sequential()
    model.add(Dense(units=32, input_dim=input_dim))
    model.add(Activation('relu'))
    model.add(Dense(units=4))
    model.add(Activation('softmax'))
    model.compile(loss='sparse_categorical_crossentropy',
                  optimizer='sgd',
                  metrics=['accuracy'])
    model.fit(vecs_train, labels_train, epochs=epochs,
              batch_size=32)

    return model
```

## 5.2 Résultats

Mixte UCLMR/Talos sans TF-Idf SLL

| | agree | disagree | discuss | unrelated | Somme |
|---|---|---|---|---|---|
| agree | 1059 | 3 | 729 | 112 | 1903 |
| disagree | 253 | 14 | 323 | 107 | 697 |
| discuss | 684 | 7 | 3530 | 243 | 4464 |
| unrelated | 45 | 0 | 288 | 18016 | 18349 |
| Somme | 2041 | 24 | 4870 | 18478 | 25413 |

TABLE 21 – Table de confusion du modèle : Mixte UCLMR/Talos sans TF-Idf SLL

Mixte UCLMR/Talos sans TF-Idf SLL

| Mesure | agree | disagree | discuss | unrelated |
|---|---|---|---|---|
| Précision | 0.56 | 0.02 | 0.79 | 0.98 |
| Rappel | 0.52 | 0.58 | 0.72 | 0.97 |
| F1score | 0.54 | 0.04 | 0.76 | 0.98 |
| Exactitude | 89.01 | | | |
| Score FNC | 9606.75 | | | |
| Pourcentage FNC | 82.45 | | | |

TABLE 22 – Mesures pour le modèle : Mixte UCLMR/Talos sans TF-Idf SLL

Etonnamment ce modèle a de moins bons résultats que le modèle précédent. Même en changeant les paramètres d'apprentissage (nombre d'unités d'apprentissage ou nombre de récusrion) le modèle n'excéde pas le précédents.