Analyse des résultats FNC

Enzo Poggio

11 avril 2018

Résumé

Dans cette première 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 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 qe 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 Solat 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. Notez que les tables de confusions ont horizontalement les labels de vérité et verticalement les prédictions.

2.1 Solat in the Swen.

solat					
	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

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. Nous présentons les participants du premiers au derniers selon le score FNC.

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

solat

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

Bien que Solat 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 Solat in the Swen.

2.1.1.1 Sous-modèle arborescent de Solat in the Swen

Les sous-modèles de Solat 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.

solat tree

	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

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

solat tree

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

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 Solat 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 Solat in the Swen

solat deep

	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

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**.

solat 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

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

La table de confusion d'Athene ressemble beaucoup à celle de Solat. 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

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

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