Stylo@DEFT2022

Notation automatique de copies d'étudiant·e·s par combinaisons de méthodes de similarité

Ibtihel Ben Ltaifa[1], Toufik Boubehziz[1], Andrea Briglia[1], Corina Chutaux[1], Yoann Dupont[2], Carlos-Emiliano González-Gallardo[3], Caroline Koudoro-Parfait[1,2,4] Gaël Lejeune[1,3]

July 5, 2022

[1] Sens Texte Informatique Histoire (STIH), Sorbonne Université
[2] Observatoire des Textes et des Connaissances (OBTIC), Sorbonne Université
[3] Laboratoire Informatique, Image et Interaction (L3i), La Rochelle Université, France
[4] Sorbonne Center for Artificial Intelligence (SCAI), Sorbonne Université, France

Sommaire

Problématique

Évaluation automatique de devoirs d'étudiant·e·s avec le corrigé de l'enseignant

Figure 1: Comment faire pour ne pas mettre que des 1 et des 0 ? Source : Français Langue Seconde [2]

[1] http://www.francaislangueseconde.fr
[2] http://www.francaislangueseconde.fr

Données

Deux ressources :

- questions (50 `train` et 21 `test`)+ éléments de réponse (et indication de notation)
- réponses (3.820 `train` et 1.644 `test`) produites par des étudiants (avec la note)

	Nb. Questions	Nb. Réponses	Nb. Etudiants
Données d'entraînement	50	3 820	118
Données d'évaluation	21	1 644	118

Table 1: Données fournies pour la tâche de notation automatique

Méthodes de similarité pour la correction automatique

Méthodes de similarité pour la correction automatique

$\rightarrow$ Approche basée sur l'usage de plusieurs méthodes de similarité pour la correction automatique

1. Extraction de caractéristiques ;
2. Mesures de similarité

Extraction de caractéristiques

→ Quatre types de caractéristiques pour la vectorisation des réponses d'étudiant·e·s

1. N-grammes sans frontières de mots (char) ;
2. N-grammes avec frontières de mots (char_wb);
3. Sous-chaînes de mots (WordPiece) ;
4. Vecteurs contextualisés SentenceBERT (sbert)

Mesures de similarité

$\rightarrow$ Deux mesures de similarité sont appliquées sur les différentes représentations vectorielles des textes générées à partir de quatre différents types de caractéristiques:

1. La similarité Cosinus ;
2. L'indice de dissimilarité de Bray-Curtis

Protocole expérimental

Protocole expérimental

Nous utilisons trois méthodes de notation différentes :

1. notation par régression;
2. notation par classification;
3. notation par réseau de neurones

Notation par régression (M1)

$\rightarrow$ tâche de notation définie comme une tâche de régression linéaire.
Nous avons utilisé les traits suivants :

1. char ;
2. char_wb ;
3. WordPiece ;
4. sbert

Notation par classification (M2)

$\rightarrow$ tâche de notation définie comme une tâche de classification avec 3 classes (0 , 0,5 , 1) en utilisant:

$\rightarrow$ un algorithme de régression logistique avec le solver Saga (meilleurs résultats).

$\rightarrow$ les mêmes traits que M1

Notation par réseau de neurones (M3)

$\rightarrow$ Appel à un modèle de réseau de neurones pour effectuer la notation.

$\rightarrow$ Le modèle est constitué de deux couches de neurones associées à une fonction d'activation de type sigmoïde.

$\rightarrow$ L'algorithme prend en entrée les mêmes traits de M1 et la note prédite en sortie.

Résultats et discussion

Résultats

	M1	M2	M3
Précision (P)	0,642	0,678	0,653
Corrélation de Spearman (rs)	0,633	0,543	0,644

Table 2: : Résultats de nos méthodes entrainées sur DEFT-TRAIN-2021 et évaluées DEFT-TEST2021

Résultats

Représentation	**Longueurs**	**Distances**	**# Dimensions**
char	[3 : 3], [3 : 4],[3 : 5] . . .[4 : 6]	Cosinus et Bray-Curtis	16
char_wb	[3 : 3], [3 : 4],[3 : 5] . . .[4 : 6]	Cosinus et Bray-Curtis	16
WordPiece			16

Table 3: : Les 48 dimensions utilisés pour la représentation de la similarité (nous y ajoutons deux représentations issues de Sentencebert pour aboutir à 50 dimensions)

Analyse de nos résultats

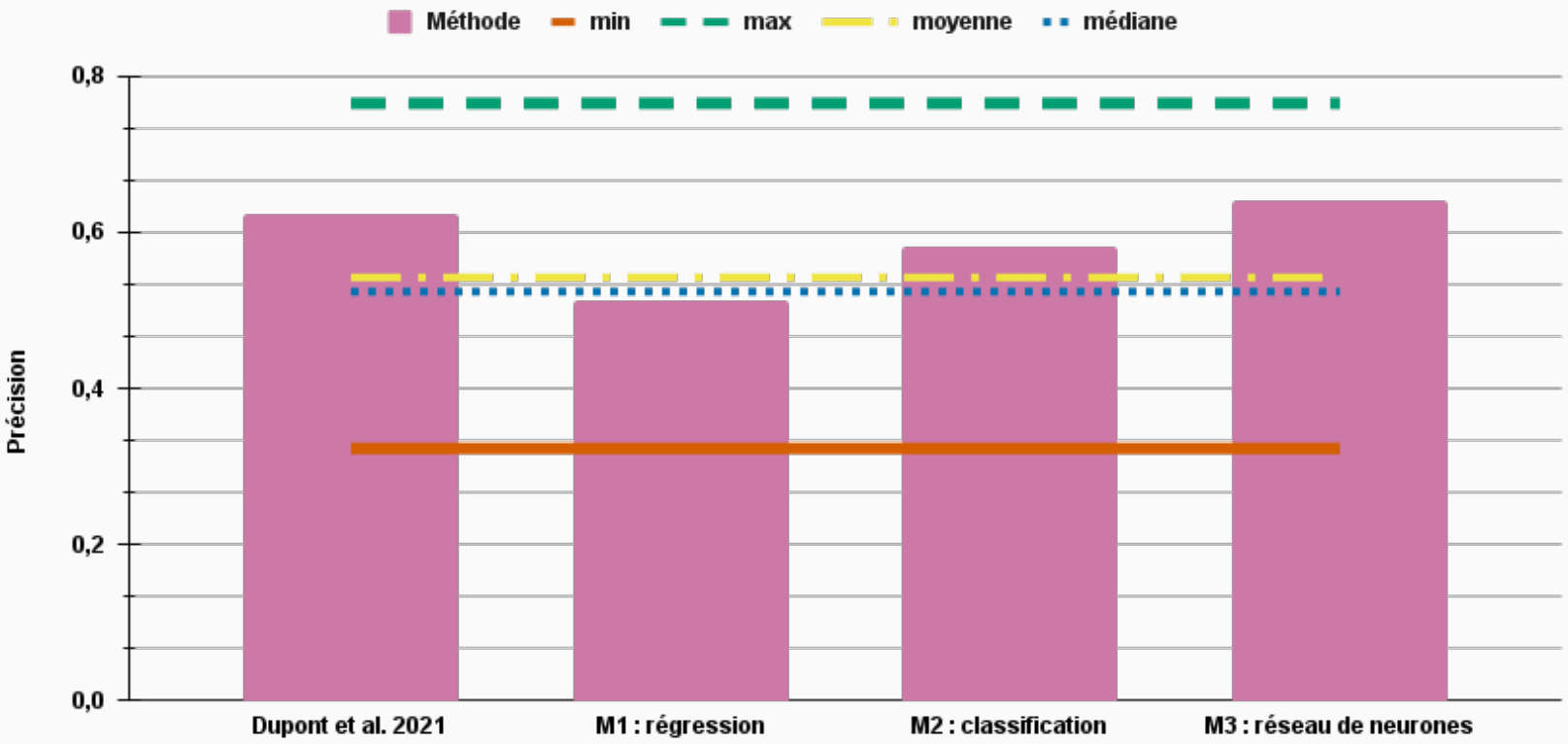

Figure 2: Résultats des nos méthodes et statistiques générales

Commentaire sur les résultats

- Les résultats obtenus sont supérieurs à la moyenne et à la médiane des résultats des autres équipes, méthode avec un certain potentiel
- améliorer la précision de **0,630** à une précision de **0,678** (M3), soit un gain de **4,8** points de pourcentage sur DEFT 2021.
- score de **0,654** pour DEFT 2022, soit un gain de **2,4** points de pourcentage par rapport à DEFT 2021.

Conclusion et perspectives

Conclusion

Proposition d'une approche simple pour l'évaluation de réponses d'étudiant·e·s étant donné une correction préalable :

- Nous avons pu améliorer les résultats depuis notre précédente contribution (DEFT2021)
- Notre score a augmenté de **4,8** points de pourcentage sur les données Deft 2021, **2,4** point de pourcentage sur les données Deft 2022.

Perspectives

- Tester d'autres similarités comme la distance de Jaccard ou l'indice de Dice pour estimer la proximité.
- Utiliser d'autres modèles de sentence embedding comme le modèle : Universal Sentence Encoder (Cer et al., 2018 ; Yang et al., 2019)