DEFT 2022

Notation automatique de réponses courtes d’étudiants : présentation de la campagne DEFT 2022

Cyril GROUIN – Gabriel ILLOUZ

LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

cnrs

université PARIS-SACLAY

Introduction

Présentation

Défi Fouille de Textes (https://deft.lisn.upsaclay.fr/), créé en 2005, sur le français

- presse ancienne [1800-1944] : origine géographique (2010), variation diachronique (2010, 2011)
- articles scientifiques : appariement articles / résumés (2011), indexation mots-clés (2012)
- recettes de cuisine (2013) : identification niveau de difficulté, type de plat, ingrédients, appariement titre / recette
- nouvelles littéraires courtes (2014) : classification, notation, degré de consensus
- tweets : polarité globale (2015, 2017, 2018), identification de l'ironie / l'humour / du sarcasme (2017) ; classification (2015, 2018), marqueurs de sentiment et cible (2015, 2018), portions de sentiments (2018)
- cas cliniques : indexation, similarité sémantique, extraction d'information démographique (2019) ; EI fine patients / pratique / traitements (2020) ; identification du profil clinique du patient (2021)
- réponses courtes d'étudiants (2021, 2022)

Calendrier

- janvier – février : réflexion sur les tâches à proposer, constitution des corpus (annotations, formatage)
- 1er mars – 30 avril : phase d'entraînement, distribution des corpus aux équipes inscrites
- 2 – 9 mai : phases de test, application des modèles, soumission des résultats
  - tâche de base (2 – 3 mai) puis tâche continue (4 – 9 mai)
- 3 juin : soumission des articles (EasyChair)
- 27 juin : atelier de clôture (conférence TALN 2022)

DEFT 2022 : données

Données

- Cours : base de données, programmation web (sur plusieurs années)
- Anonymisées (identifiant unique par étudiant, vérification manuelle)
- Pas de correction orthographique
- Notes normalisées entre 0 et 1 avec maximum deux décimales
  - 77,1 % des notes attribuées sont 0 ou 1 ; 88,9 % si on intègre la note 0,5
- Données issues de Moodle
  - conservation des balises de mise en forme (<p>
)
  - représentation des balises de code informatique par des entités HTML (< >)
  - mention « NO_ANS » en cas d’absence de réponse par l’étudiant

Questionnaires Moodle : questions

Question	Intitulé	Correction + commentaire
2032 Langue naturelle	Quel est l'intérêt d'utiliser du code AJAX ?	 Permet l'échange de données avec le serveur sans mise à jour complète de la page Ok pour permet de màj une partie de la page sans avoir à la recharger complètement.
2034 Code	Modifiez le code XML ci-dessous pour le rendre valide : <code> <ue id=PW2> </code> ? <code> <ue id=PW2> </code>	 <ue id=''PW2''> 1 si guillemets ajoutés 0 sinon 0.5 si transformé id en sous-élément (mais pas si supprimé la notion d'id)

Questionnaires Moodle : réponses

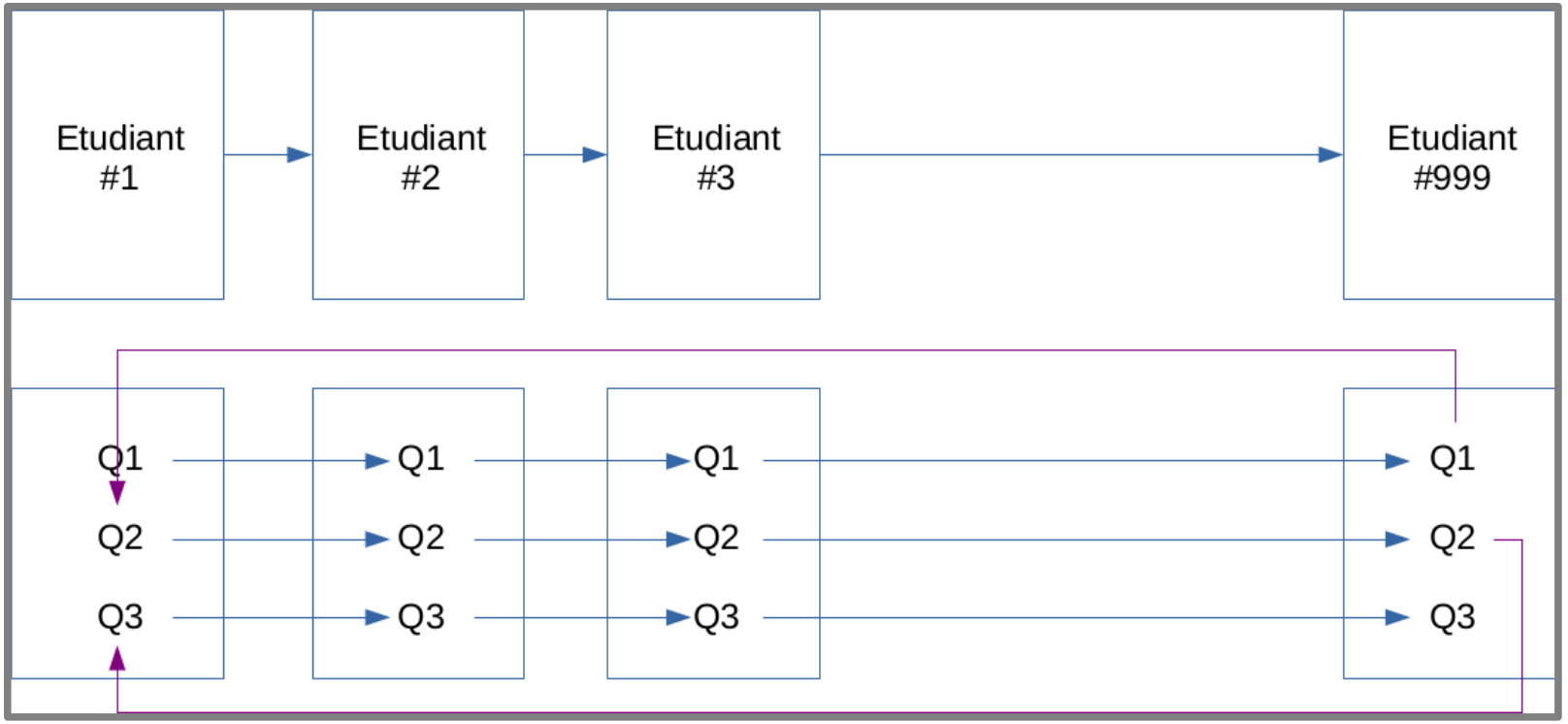
Q	Etud	Réponse	Note
2032	106	Il permet de mettre à jour dynamiquement la page sans avoir à recharger la page entière.	1
2032	107	NO_ANS	0
2032	108	AJAX permet de modifier en temps réel une page, sans avoir à faire appel au serveur. Par exemple, on peut changer le contour d'un bouton lorsque la souris passe dessus	0,5
2032	109	Cela permet d'appeler des scripts dans la page web	0,2
2032	12	Le code AJAX permet d'actualiser une partie d'une page web sans avoir à recharger toute la page.	1
2034	106	<code>\n <ue id=''PW2''>\n </code>	1
2035	108	<code>\n <ue id=PW2>\n </code>	0
2035	110	<code>\n <ue>\n <id>PW2</id>\n </ue>\n </code>	0,5

Code reproduit à l'identique

Problématique

Problématique

Quelle stratégie un enseignant humain doit-il adopter pour corriger efficacement un ensemble de réponses courtes d'étudiants ?



Problématique

- Une stratégie de correction par question permet :
  - d’avoir une vue globale de l’ensemble des réponses
  - de ne se concentrer que sur le même contenu évalué
  - de réévaluer son barème et d’ajuster l’attendu dans les réponses si nécessaire
- Cependant, le processus de correction peut/doit être optimisé
  - comment ? sur quelle base ?

Tâches

Tâches de notation automatique

- Tâche de base : prédire des notes à partir d'une référence
  - répartition classique entre données d'entraînement annotées et données de test non annotées
- Tâche continue : interrogation continue et raisonnée d'un serveur d'évaluation sécurisé
  - obtention des notes de référence pour chaque étudiant souhaité sur les différentes questions
  - sans demander toutes les notes (techniquement possible) : stratégie de correction inefficace

Tâche de base

- Objectif : attribuer automatiquement des notes
- Corpus d’entraînement (intégralité du corpus DEFT 2021)
  - 88 questions avec correction et indications de l’enseignant
  - 6620 réponses d’étudiants (nombre variable d’étudiants par question)
- Corpus de test :
  - 24 nouvelles questions
  - 2640 réponses d’étudiants à évaluer (110 étudiants)

Tâche continue

- Objectif : aider l’enseignant en organisant les corrections
  - pour minimiser le temps passé à corriger
  - pour proposer une correction automatique partielle jusqu’à atteindre un taux de réponses non corrigées acceptable
- Corpus d’entraînement identique à celui de DEFT 2021
  - 50 questions, 3820 réponses
- Corpus de test différent de celui de la tâche de base
  - 25 nouvelles questions, 2750 réponses (110 étudiants)

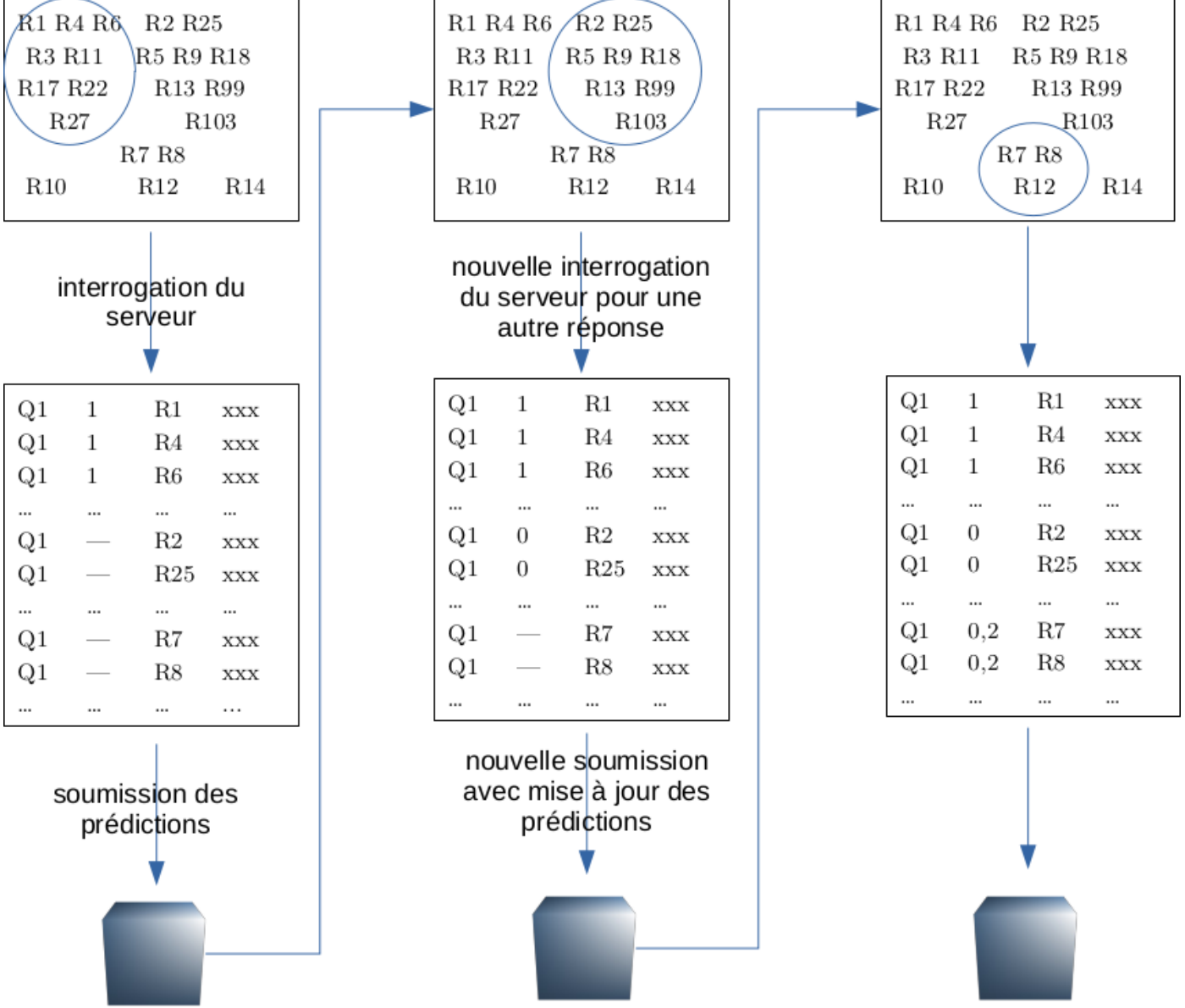
Tâche continue

- « *Le code PHP est-il exécuté sur la machine cliente ou sur le serveur web ?* »
- 116 réponses d'étudiants dont :
  - 43 « *le code PHP est exécuté sur le serveur web* »
  - 20 « *sur le/un serveur web* » (réponse courte)
  - 12 « *il est exécuté sur le serveur web* » (reprise pronominale)

  81 réponses proches (64,7 % de la promotion)
  - 6 « *le code PHP est exécuté sur la machine cliente* »

  6 réponses erronées (5,2 % de la promotion)
  - 3 : absence de réponse (= 0)

  3 absences de réponse (2,6 % de la promotion)
  - autres : variantes (« serveur » au lieu de « serveur web »), précisions (« … *c'est le javascript qui est exécuté sur la machine cliente* », « ... *et non sur la machine cliente* », « … *le client ne reçoit que le résultat du script* », etc.)


R1 R4 R6 R2 R25
R3 R11 R5 R9 R18
R17 R22 R13 R99
R27 R103
R7 R8
R10 R12 R14
interrogation du serveur
Q1 1 R1 xxx
Q1 1 R4 xxx
Q1 1 R6 xxx
... ...
Q1 — R2 xxx
Q1 — R25 xxx
... ...
Q1 — R7 xxx
Q1 — R8 xxx
... ...
soumission des prédictions
nouvelle interrogation du serveur pour une autre réponse
Q1 0 R2 xxx
Q1 0 R25 xxx
Q1 0,2 R7 xxx
Q1 0,2 R8 xxx
nouvelle soumission avec mise à jour des prédictions

Tâche continue

- Trois scripts Python, à intégrer dans vos outils, pour un processus itératif :
  - interrogation du serveur : demande la note d'un étudiant à une question
  - pour cette question, envoie les notes prédites (sur tout ou partie des étudiants)
  - dépôt de la soumission : pour cet envoi, horodatage avec commentaire éventuel
- Quatrième script (phase d'entraînement uniquement) :
  - vide la base de données : reprendre la procédure du début en testant une nouvelle configuration / stratégie
  - désactivé pour le test (éviter de récupérer la référence, d'effacer la soumission, puis de faire une soumission plus « avantageuse » du point de vue de la compétition)

Tâche continue

- Dans quel ordre les questions ont-elles été demandées ?
  - groupe par type de réponse attendu (code, langue naturelle, réponse courte) ?
  - priorité sur les grands ensembles de réponses ?
- Combien de notes demandées par question ?
  - y a t-il demande pour les réponses isolées ?
- Observe t-on une évolution rapide des performances des systèmes ?

Résultats

Résultats : tâche de base

Equipe	Prec. run 1	Prec. run 2	Prec. run 3
EDF R&D	0,752 (0,70)	0,756 (0,70) 1	0,323 (0,76)
LIA-LS2N	0,440 (0,00) 4	0,404 (0,00)	0,440 (0,00)
LIA-LS2N (hors compétition)	0,606 (0,39)	0,726 (0,69)	0,649 (0,48)
STyLO	0,512 (0,63)	0,580 (0,56)	0,641 (0,51) 2
TGV	0,491 (0,17)	0,536 (0,54)	0,624 (0,42) 3
Baseline		*0,522 (0,37)*	
Tirage aléatoire		*0,380 (0,47)*	

Précision moyenne = 0,542 ; médiane = 0,524 (calcul sur les soumissions officielles)
Corrélation de Pearson entre parenthèses. Classement officiel sur la précision

Résultats : tâche continue

- Un seul participant
- Pas de résultats significatifs pour l’objectif visé

Conclusion

Conclusion

- Tâche de base : 4 participants, 3 soumissions par équipe
  - meilleures précisions par équipe 0,440–0,756 (moyenne 0,542 ; médiane 0,524)
  - augmentation du nombre de questions dans l'entraînement : 50 questions, 3820 réponses en 2021, 88 questions, 6620 réponses en 2022
  - corpus de test différents et similaires entre DEFT 2021 et 2022, baisse du nombre de questions mais plus de réponses : 38 questions en 2021, 25 questions en 2022
  - amélioration globale des performances (EDF R&D 0,682→0,756, QUEER-StyLO 0,630→0,641) mais écarts plus importants sur les meilleures soumissions (0,630–0,682 en 2021, 0,624–0,756 en 2022)
- Tâche continue expérimentale mais complexe

Merci d’avoir participé
et essayé autant de méthodes

Programme DEFT 2022

- **14h00** *Notation automatique de réponses courtes d'étudiants : présentation de la campagne DEFT 2022* (Grouin et Illouz)
- **14h30** *Participation de l'équipe TGV à DEFT 2022 : Prédiction automatique de notes d'étudiants à des questionnaires en fonction du type de question* (Gaudray et al.)
- **15h00** *Stylo@DEFT2022 : Notation automatique de copies d'étudiant·e·s par combinaison de méthodes de similarité* (Ben Ltaifa et al.)

15h30 *Pause café*

- **16h00** *Correction automatique d'examens écrits par approche neuronale profonde et attention croisée bidirectionnelle* (Labrak et al.)
- **16h30** *Participation d'EDF R&D à DEFT 2022* (Suignard et al.)
- **17h00** Discussions et conclusion