Appel à participation à la compétition « *my taylor is rich* » de CAp 2018

Prédiction du niveau en anglais à partir de production écrite d'apprenants

1 Objectif

Le Cadre européen commun de référence pour les langues (CERL) découpe la compétence linguistique d'une langue étrangère en six niveaux de référence, décrits pour être partagés par les pays européens : A1, A2, B1, B2, C1 et C2.

Le but de cette compétition est de réaliser, par apprentissage, un système permettant de prédire le niveau de compétence d'un apprenant, à partir d'une de ces productions écrites comprenant entre 20 et 300 mots et d'un ensemble de caractéristiques calculées à partir de ce texte.

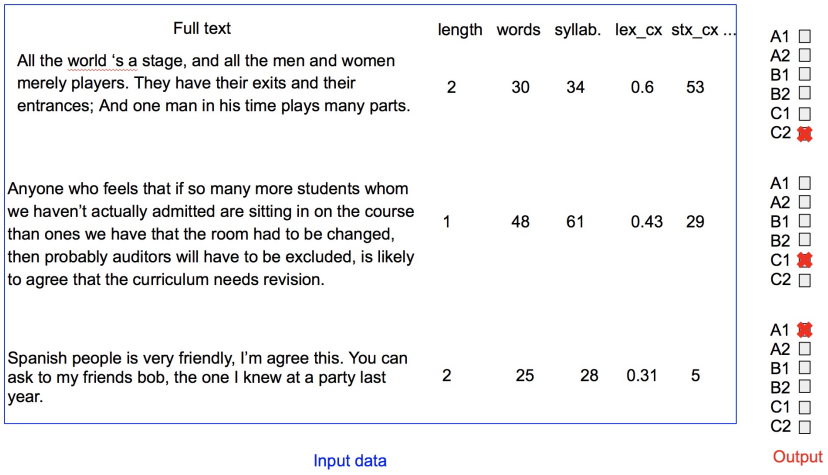

Illustration de la tâche à réaliser.

2 Jeux de données

Pour avoir accès aux données il faut :

1. vous inscrire sur le site
   `https://corpus.mml.cam.ac.uk/efcamdat2/public_html/explore/`

2. une fois inscrit, envoyer un email à `efcamdat.team@gmail.com` avec comme sujet :

   Request for CAp2018 Shared Task Data

Après confirmation de votre inscription, un e-mail avec les informations de connexion au dossier contenant les données vous sera envoyé dans les 24 heures.

Les données sont issues d'une extraction de la base de données publiée par Cambridge et Education First [Geertzen et al., 2013].

Avertissement : les données d'apprentissage et de test ont été sélectionnées et manipulées indépendamment de la participation des équipes de recherche de Cambridge et Education First.

Le jeu de données proposé comprend 27 310 exemples de textes écrits par des apprenants et un ensemble de caractéristiques associées comprenant :

- des métriques de complexité lexicale,
- des métriques de lisibilité,
- la classe à prédire

2.1 Les caractéristiques

Cinquante-neuf variables sont disponibles. La première (appelée *fulltext*) est un texte : le texte intégral produit par la personne à évaluer (en moyenne 70 mots).

Les 58 autres variables sont des métriques calculées à partir du texte. Elles décrivent le degré de sophistication de vocabulaire et la complexité du texte. Parmi ces métriques on trouve : le nombre de phrases, de mots, de lettres, de syllabes, les *type-to-token* ratio (et les mesures dérivées), des mesures de lisibilité calculées à partir de la corrélation entre le nombre de mots et la longueur des mots utilisés et la sophistication lexicale, qui mesure la richesse du lexique à l'aide d'inventaires de références. Une description précise des caractéristiques est disponible dans l'annexe.

Deux points importants concernant les caractéristiques :

1. Pour la compétition, on pourra utiliser tout ou partie des variables disponibles ; notamment, il n'est pas obligatoire d'utiliser les textes.

2. Un prix spécial du jury pourra être accordé à une solution performante ayant sélectionné les meilleures caractéristiques.

2.2 Classes à prédire

Les classes à prédire sont les 6 niveaux de référence du CERL (la dernière variable du fichier appelée *level1*). Les données proposées sur le site sont de 16 niveaux différents. Voici comment s'établit la conversion entre les niveaux estimés par EFCAMDAT et le CERL et les effectifs de chaque classe dans l'ensemble d'apprentissage :

EFCAMDAT	CERL	effectif par classe
1-3	A1	11361
4-6	A2	7688
7-9	B1	5383
10-12	B2	2337
13-15	C1	491
16	C2	50

Les données de test ont la même proportion par classe que les données d'apprentissage.

3 Evaluation

La mesure de performance utilisée sera :

$$E = \frac{1}{n} \sum_{i=1}^{6} \sum_{j=1}^{6} C_{ij} N_{ij}$$

où N est la matrice de confusion (N_{ij} compte le nombre de fois où un exemple de la classe i a été classé j), n le nombre d'exemples classés et C la matrice de coût ci-dessous.

Réel \ Estimée	A1	A2	B1	B2	C1	C2
A1	0	1	2	3	4	6
A2	1	0	1	4	5	8
B1	3	2	0	3	5	8
B2	10	7	5	0	2	7
C1	20	16	12	4	0	8
C2	44	38	32	19	13	0

Matrice C des coûts

Cette matrice de coût a été calculée comme une mesure d'entropie croisée pondérée entre les classes. Les probabilités prises en compte sont les probabilités d'apparition des classes telles qu'elles apparaissent dans les échantillons d'apprentissage et de test. Les poids ont été donnés par les experts du domaine pour prendre en compte l'importance de chacune des classes.

4 Dates importantes

La compétition se déroulera sur 2 mois de la manière suivante :

- Ouverture de la compétition : 28 mars 2018
- Disponibilité du jeu de test : 28 avril 2018
- Fin de la compétition : 28 mai 2018 minuit
- Annonce des résultats et remise des prix : 21 juin 2018

5 Prix

NVIDIA attribuera des cartes graphiques GPU (2 ou 3) aux 2 ou 3 meilleurs systèmes.

6 Comité d'organisation

- Nicolas Ballier (CLILLAC-ARP, Université Paris Diderot)
- Stéphane Canu (LITIS, INSA Rouen Normandie)
- Thomas Gaillat (Insight Centre for Data Analytics NUIG, Irlande)
- Gilles Gasso (LITIS, INSA Rouen Normandie)
- Caroline Petitjean (LITIS, Université Rouen Normandie)
- Alain Rakotomamonjy (LITIS, Université Rouen Normandie)

References

[Geertzen et al., 2013] Geertzen, J., Alexopoulou, T., and Korhonen, A. (2013). Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project.*