Shared task CMCL : Prédiction de variables oculométriques

Franck Dary

JTT

08/04/2021

Contexte

Dans le cadre du Workshop :
Cognitive Modeling and Computational Linguistics (CMCL 2021)

A été organisée une Shared Task portant sur la prédiction de variables comportementales (lecture de texte).

Participants TALEP : Franck Dary, Alexis Nasr, Abdellah Fourtassi.

Tâche

Étant donné du texte, il faut prédire 5 variables.

Entrées :

sentID	wordID	word
1	1	Carlucci
1	2	was
1	3	deputy
1	4	defense
1	5	secretary

Sorties :

sentID	wordID	word	nFix	FFD	GPT	TRT	fixProp
1	1	Carlucci	28.40	4.64	6.19	10.34	94.12
1	2	was	12.98	3.53	5.26	4.57	76.47
1	3	deputy	25.15	5.81	9.93	9.15	100.00
1	4	defense	20.28	5.26	8.09	8.36	88.24
1	5	secretary	17.85	4.14	4.51	5.69	88.24

Description des Variables

Variables à prédire :

- Valeurs moyennes sur 12 sujets humains.
- "Normalisées" entre 0 et 100.
- Toutes au niveau du mot.

nFix	Number of Fixations	Nombre de fixations.
FFD	First Fixation Duration	Durée de la première fixation.
GPT	Go Past Time	Somme de la durée des fixations ayant lieu avant que le mot ne soit franchi.
TRT	Total Reading Time	Somme des temps de fixation.
fixProp	Fixation Proportion	Proportion des sujets ayant fixé le mot.

Corpus

Les données viennent de ZuCo (Hollenstein et al., 2018) :

- en anglais (critiques de films, biographies Wikipedia)
- 12 lecteurs l'ayant pour langue maternelle
- une partie avec données d'électroencéphalographie

Pour la shared task :

- "Lecture Normale" de texte (pas de tâche)
- Train : 800 phrases (15.737 tokens)
- Test : 191 phrases (3.554 tokens)

Notre approche

Réseau de neurones qui combine 4 familles de features :

- **Extraites du texte brut :** mots, taille des mots, prefixes, position dans la phrase. . .
- **Fréquences :** fréquences unigrammes et bigrammes estimées à partir d'un autre corpus.
- **Prédictions Linguistiques :** PoS, morphologie, fonction syntaxique et distance au gouverneur.
- **Mesures de complexité :** obtenues en produisant les prédictions linguistiques, état de la pile et entropies.

Extraction des features

Corpus qu'on nous donne :

- pas de ponctuation
- segmentation = à l'espace
- comme montré aux sujets

Problème : Annotation linguistique apprise sur Universal Dependencies

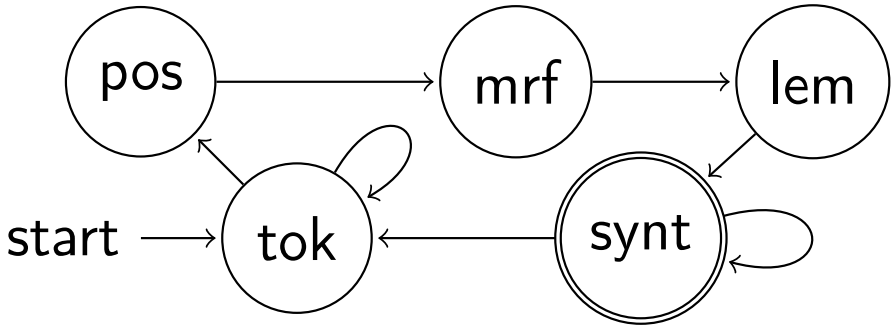
- ponctuation
- segmentation plus complexe (ex. $don't \rightarrow \{do, n't\}$)

Méthode :

1. Récupérer le texte brut qui correspond au corpus.
2. Annoter ce texte au format UD.
3. Aligner les annotations avec le corpus de la shared task.

Production des annotations linguistiques

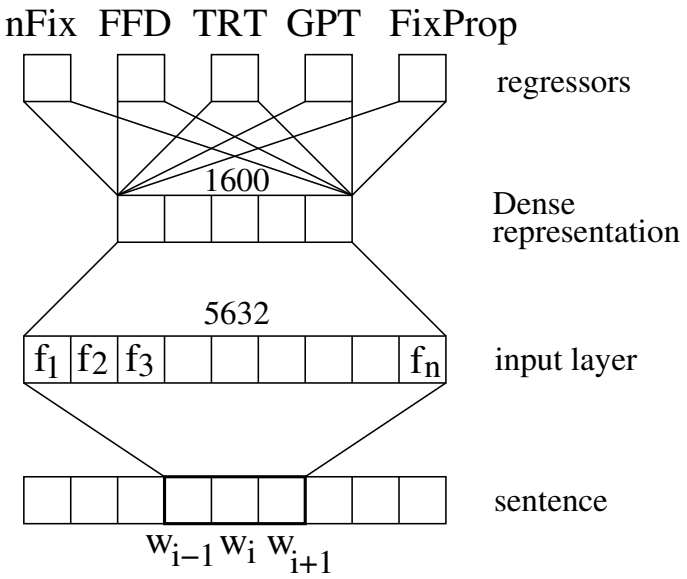
Analyseur en transitions incrémental :



- Transition = déplacer frontière de mot ou annoter mot courant
- Mémorise certains mots dans une pile (Transition Based Parsing)
- Analyse du texte = prédire une séquence de transitions
- A chaque étape, choix glouton d'une transition parmi n
- Mesures de complexité : pile, entropie de la distribution de probabilités

Entraîné sur UD : GUM+EWT+LinES+ParTUT = 515.228 mots

Prédiction des variables oculométriques



Prédiction des variables pour le mot w_i :

- Extraction des features dans un contexte centré sur w_i
- MLP 2 couches
- Une couche de régression différente pour chaque variable à prédire
- Loss = L1 (absolute error). Optimizer = Adagrad

Évaluation

Name	Description	Span	MAE (individual)	MAE (group)
Raw Textual Features				
length	Number of letters in the word.	111	$4.20_{\pm 0.00}$	$3.87_{\pm 0.01}$
prefix	First 3 letters of the word.	010	$4.15_{\pm 0.02}$	
suffix	Last 3 letters of the word.	010	$4.16_{\pm 0.01}$	
form	Contextualized word embedding.	110	$4.05_{\pm 0.01}$	
eos	Whether or not the word is the last of the sentence.	110	$4.15_{\pm 0.00}$	
word_id	Index of the word in the sentence.	110	$4.09_{\pm 0.01}$	
sent_id	Index of the sentence in the text.	110	$4.16_{\pm 0.01}$	
origin	Name of the Zuco file containing the raw text.	010	$4.18_{\pm 0.00}$	
Frequencies				
frequency	Logarithm of the frequency of the word.	111	$4.20_{\pm 0.00}$	$3.78_{\pm 0.02}$
cooc_p	Log frequency of the bigram with previous word.	111	$4.15_{\pm 0.00}$	
cooc_n	Log frequency of the bigram with next word.	111	$4.17_{\pm 0.00}$	
Linguistic Features				
pos	Part of speech.	110	$4.15_{\pm 0.00}$	$3.74_{\pm 0.01}$
morpho	Morphology.	110	$4.17_{\pm 0.00}$	
deprel	Syntactic function.	110	$4.14_{\pm 0.00}$	
dep_len	Distance to the syntactic governor.	110	$4.17_{\pm 0.01}$	
Complexity Metrics				
stack_size	Size of the stack when processing the word.	111	$4.12_{\pm 0.00}$	$3.73_{\pm 0.00}$
stack_dist	Distance between the two top elements of the stack.	111	$4.14_{\pm 0.01}$	
ent_tok	Entropy of the tokenizer.	110	$4.22_{\pm 0.00}$	
ent_tag	Entropy of the part of speech tagger.	110	$4.22_{\pm 0.01}$	
ent_morpho	Entropy of the morphological tagger.	110	$4.22_{\pm 0.00}$	
ent_parser	Entropy of the dependency parser.	110	$4.23_{\pm 0.01}$	
ent_mean	Mean of the entropies.	110	$4.22_{\pm 0.00}$	
ent_max	Highest entropy.	110	$4.23_{\pm 0.01}$	

Résultats

Team Name	User	MAE	nFix	FFD	GPT	TRT	fixProp	Rank
LAST	zz	3.8134	3.879	0.655	2.197	1.524	10.812	1
LAST	zz	3.8159	3.886	0.655	2.199	1.523	10.817	-
TALEP	franck.dary	3.8328	3.761	0.662	2.18	1.486	11.076	2
LAST	zz	3.8664	3.943	0.662	2.237	1.545	10.944	-
TorontoCL	bai	3.9287	3.944	0.671	2.227	1.516	11.286	3
TorontoCL	bai	3.9421	3.964	0.673	2.247	1.521	11.305	-
LRL_NC_IITD	raksha297	3.9488	4.039	0.674	2.248	1.568	11.216	4
CogNLP@Sheff	petervickers	3.9565	3.956	0.689	2.26	1.529	11.349	5
CogNLP@Sheff	petervickers	3.9567	3.957	0.689	2.259	1.529	11.35	-
TorontoCL	bai	3.974	4.016	0.676	2.243	1.537	11.398	-
OhioState	byungdoh	3.9767	3.987	0.682	2.364	1.54	11.311	6
LRL_NC_IITD	raksha297	3.9789	4.039	0.685	2.266	1.563	11.342	-
LRL_NC_IITD	raksha297	3.9794	4.0392	0.6849	2.2665	1.564	11.3424	-
OhioState	byungdoh	3.9816	4.079	0.668	2.407	1.544	11.21	-
OhioState	byungdoh	4.0035	4.079	0.682	2.407	1.54	11.311	-
MTL782_IITD	srz208250	4.0639	4.115	0.719	2.264	1.622	11.599	7
CogNLP@Sheff	rosaw-ai	4.0689	4.198	0.689	2.372	1.676	11.41	-
KonTra	kkalouli	4.2163	4.263	0.698	2.756	1.682	11.683	8
MTL782_IITD	srz208250	4.2247	4.242	0.737	2.493	1.65	12.001	-
Sabhay_Jain	sabhayj123	4.2565	4.264	0.848	2.476	1.721	11.974	9
ReadMe	balkoca	4.383	4.363	0.741	2.502	1.761	12.549	10
PIHKers	laviniasalicchi	4.3877	4.335	0.715	3.059	1.713	12.118	11
MTL782_IITD	LanguageRese	4.4383	4.492	0.728	2.632	1.749	12.59	-
ChiSquareX	AGP	4.6764	4.557	1.281	2.81	2.289	12.445	12
ReadMe	balkoca	4.7745	4.62	0.755	3.605	1.836	13.056	-
Sabhay_Jain	sabhayj123	4.8201	4.942	1.334	2.989	2.182	12.654	-
ChiSquareX	vmm	5.2606	4.975	0.809	3.065	2.31	15.144	-
MEAN BASELIN	MEAN BASELIN	7.3699	7.303	1.149	3.782	2.778	21.775	-
IIIT_DWD	Ankit011	9.7615	8.845	1.589	4.633	3.296	30.446	13
ChiSquareX	vmm	10.1822	9.244	1.654	4.374	3.346	32.292	-