Classifying learner level

Andrew Simpkin

Acknowledgement

With the financial support of the French Ministry for Europe and Foreign Affairs (Ministères de l’Europe et des affaires étrangères, MEAE) and the French Ministry of Higher Education, Research and Innovation (Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, MESRI).

PHC Hubert Currien Ulysse 2019 (ref 43121RJ)

Motivation

- Data from 49813 texts written by 8851 learners were available from the EFCAMDAT dataset.
- The aim of this study was to construct a classification model of learner level (A1, A2, B1, B2, C1, C2), based on a corpus submitted by the learner.
  - This is a multi-category or multinomial classification problem
- A simpler classifier task was to separate beginner (A1, A2, B1) and advanced (B2, C1, C2) learners
  - This is a binary classification problem

Features used

In order to test the efficacy of our novel microsystem variables, we built three classification models:

i. using 691 features from previous research (TAALES, TAASC)
ii. using the 49 microsystem variables introduced in this paper along with 29 ratios
iii. using the 49 microsystem variables introduced in this paper along with 12 interactions listed above

Models tested

- multinomial logistic regression
- random forests
- linear discriminant analysis
- k-nearest neighbours
- Gaussian naive Bayes
- support vector machine
- decision tree classifier
- neural network

Results: multinomial classifier

- We split the data into 75% training and 25% test data, resulting in 12454 texts in the testing data
- Among the eight model types tested here, the optimal classification performance in the testing dataset was found using **multinomial logistic regression.**

Confusion matrix (i) Old features

Predicted	A1	A2	B1	B2	C1	C2
Predicted A1	4481	356	73	27	7	**2**
Predicted A2	398	2696	349	42	9	**0**
Predicted B1	114	274	2004	323	41	**7**
Predicted B2	44	28	131	784	132	**20**
Predicted C1	7	4	10	28	59	**4**
Predicted C2	**0**	**0**	**0**	**0**	**0**	**0**

Results (i) Old features

level	precision	recall	f1	support
A1	0.91	0.89	0.90	**5044**
A2	0.77	0.80	0.79	**3358**
B1	0.73	0.78	0.75	**2567**
B2	0.69	0.65	0.67	**1204**
C1	0.53	0.24	0.33	**248**
C2	0.00	0.00	0.00	**33**
Mean	**0.80**	**0.80**	**0.80**	**12454**

Results (ii) Old + New + Ratios

Predicted	A1	A2	B1	B2	C1	C2
Predicted A1	4524	334	76	27	6	3
Predicted A2	359	2734	324	38	8	0
Predicted B1	112	258	2038	307	44	7
Predicted B2	42	29	116	796	117	19
Predicted C1	7	3	13	36	73	4
Predicted C2	0	0	0	0	0	0

Results (ii) Old + New + Ratios

level	precision	recall	f1	support
A1	0.91	0.90	0.90	**5044**
A2	0.79	0.81	0.80	**3358**
B1	0.74	0.79	0.76	**2567**
B2	0.71	0.66	0.69	**1204**
C1	0.54	0.29	0.38	**248**
C2	0.00	0.00	0.00	**33**
Mean	**0.81**	**0.82**	**0.81**	**12454**

Results (iii) Old + New + Interactions

Predicted	A1	A2	B1	B2	C1	C2
Predicted A1	4506	343	76	28	6	**3**
Predicted A2	376	2717	326	37	8	**0**
Predicted B1	115	267	2031	313	42	**7**
Predicted B2	40	27	122	793	120	**19**
Predicted C1	7	4	12	33	72	**4**
Predicted C2	**0**	**0**	**0**	**0**	**0**	**0**

Results (iii) Old + New + Interactions

level	precision	recall	f1	support
A1	0.91	0.89	0.90	**5044**
A2	0.78	0.81	0.79	**3358**
B1	0.73	0.79	0.76	**2567**
B2	0.71	0.66	0.68	**1204**
C1	0.55	0.29	0.38	**248**
C2	0.00	0.00	0.00	**33**
Mean	**0.81**	**0.81**	**0.81**	**12454**

Binary classification

When trying to separate out beginners and advanced learners, model performance was increased

level	precision	recall	f1	support
Advanced	0.83	0.74	0.78	**1485**
Beginner	0.96	0.98	0.97	**10969**
Mean	**0.95**	**0.95**	**0.95**	**12454**

Summary

- Using 769 features, multinomial logistic regression achieves over 80% accuracy in classifying learner level
- This is improved to 95% accuracy for determining whether a learner is a beginner or advanced
- Future work will investigate feature importance
- More advanced learner data (C1, and especially C2) is required to improve performance

Evaluation on external data set

SAG corpus (Tack et al. 2017)

99 texts

hort answers from 30 words at the A1 levels to 150 words at the C levels

Classification with model

	Precision	Recall	F1-score	Support
1	0.62	0.56	0.59	18
2	0.67	0.14	0.23	59
1	0.49	0.80	0.61	113
2	0.51	0.58	0.54	74
1	1	0.03	0.06	30
2	0	0	0	5

Confusion matrix

Predicted	A1	A2	B1	B2	C1	C2
redicted A1	10	2	6	0	0	0
redicted A2	6	8	45	0	0	0
redicted B1	0	2	90	21	0	0
redicted B2	0	0	31	43	0	0
redicted C1	0	0	12	16	1	1
redicted C2	0	0	0	5	0	0

Discussion

o approximately **55% accuracy in ASAG**, where we had 82% in the EFCAMDAT test data.
ome reasons why we've lost performance...

1. Performance in test data randomly taken from the training data is always optimistic, because the test and train sets are very similar
2. The ASAG data have few A1 learners (~16%), while the EFCAMDAT had approximately 40%
3. Overfitting to EFCAMDAT - this is the one we can work on! 768 features is a lot for this mode By reducing this to a smaller set of features, we'll lose performance on the EFCAMDAT test data, but possibly gain performance in external sets. This is something I'll work on next!

References

Tack, A., François, T., Roekhaut, S., & Fairon, C. (2017). Human and automated CEFR-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 169-179).

ALTERNATES : about the ASAG data

mean of ASAG texts

- [1] 157.6288

Standard deviation

- [1] 81.66867

Distribution of levels (for the majority vote)

```
A1  A2  B1  B2  C1  C2
18  59 113  74  30   5
```