Language Proficiency Scoring

Text Mining Project

Cristina Arhiliuc, Yuliia Barabash, Jean Pierre Char, Stepan Gagin

Universität Passau

July 23, 2019

1. Introduction
2. State Of The Art
3. Dataset Overview
4. Execution Environment
5. Original Paper Results Analysis
6. Cross-lingual extension
7. Experiments With Augmented Dataset
8. Reproducing Difficulties
9. Conclusion

Introduction

Introduction

- IELTS test takers: 3 millions in 2017[1]
- The Common European Framework of Reference (CEFR)
- Automated Essay Scoring (AES)
- REPROLANG 2020[2]

[1] IELTS news, https://www.ielts.org/news/2017/ielts-numbers-rise-to-three-million-a-year?fbclid=IwAR1Q7AyPRBH6XnTnx$_a m4xn0A8tKXKpVtsTefLYWv2EqGSnlgacFwgEvOQ,lastaccessedonJuly21,2019$

[2] REPROLANG 2020, https://www.clarin.eu/event/2020/reprolang-2020, last accessed on July 21, 2019

State Of The Art

State Of The Art

Common approaches to building AES systems are based on a monolingual evaluation[1, 3]. The researchers of 'Experiments with Universal CEFR Classification'[2] experiment with different approaches involving classification on multiple languages:

- Monolingual
- Multilingual
- Cross-lingual

State Of The Art

The feature space consists of:

- Word and POS n-grams[3].
- Embeddings of task-specific words and characters trained through a softmax layer[2].
- Dependency n-grams where each unigram consists of 3 elements: The dependency relation, the POS tag of the dependent and the POS tag of the head [2].

State Of The Art

Linguistic features such as:

- Document length: The number of words in a text.
- Lexical richness features: Lexical density, lexical variation and lexical diversity features.
- Error features: These are obtained by using LanguageTool[3] for spelling and grammar checking.

which are also called domain features.

[3] LanguageTool, https://languagetool.org/, last accessed on July 21, 2019

Dataset Overview

Dataset Overview

CEFR level	CZ	DE	IT
A1	0	57	29
A2	188	306	381
B1	165	331	394
B2	81	293	0
C1	0	42	0
Total	434	1029	804

Table: Distribution of labels in corpora

Dataset Overview

CEFR level	CZ	DE	IT
A1	-	32.23	39.86
A2	93.68	56.89	69.04
B1	169.81	112.48	145.61
B2	205.91	187.96	-
C1	-	220.95	-

Table: Average document length per level

Execution Environment

Execution Environment

- Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHZ 3.80GHZ
- 16 GB RAM
- 64-bit Windows 10 Home Edition, x64-based processor
- Python 3.7
- No information about the hardware / software used for the execution of the experiments whose results are reported in [2].

Original Paper Results Analysis

Monolingual

Features	DE	IT	CZ	Avg. Dev.
Baseline	0.477 (-0.020)^{RF}	0.573 (-0.005)^{LR}	0.613 (+0.026)^{LR}	0.017
Word n-grams(1)	0.589 **(-0.077)**^{RF}	0.799 (-0.028)^{RF}	0.727 (+0.006)^{RF}	0.037
POS n-grams(2)	**0.658** (-0.005)^{RF}	0.801 (-0.024)^{RF}	0.678(-0.021)RF	0.016
Dep. n-grams(3)	0.637 (-0.026)^{RF}	**0.800** (-0.006)^{RF}	0.706 (+0.002)^{RF}	0.011
Domain features	0.520 (-0.013) ^{LR}	0.654 (+0.001)^{LR}	0.629 **(-0.034)**^{RF}	0.016
(1)+Domain	0.644 (-0.042)^{RF}	0.793**(-0.044)**^{RF}	0.720 (-0.014)^{RF}	0.033
(2)+Domain	0.646 (-0.040)^{RF}	0.796 (-0.020)^{RF}	0.687 (-0.022)^{RF}	0.027
(3)+Domain	0.639 (-0.043)^{RF}	0.784 (-0.022)^{RF}	**0.730** (+0.018)^{RF}	0.027
Word embeddings	0.604 (-0.042)	0.777 (-0.017)	0.609 (-0.016)	0.025
Avg. Dev.	0.034	0.018	0.017	

Multilingual

Features	Lang (-)	Lang (+)	Avg. Dev.
Baseline	0.426 (-0.002)^{LR}	-	0.002
Word n-grams	0.605 **(-0.116)**^{RF}	0.607 **(-0.112)**^{RF}	0.114
POS n-grams	**0.680** (-0.046)^{RF}	**0.680** (-0.044)^{RF}	0.045
Dep. n-grams	0.650 (-0.053)^{RF}	0.652 (-0.041)^{RF}	0.047
Domain features	0.433 (-0.016)^{LR}	0.447 (-0.024)^{LR}	0.020
Word embeddings	0.652 (-0.041)	0.645 (-0.044)	0.042
Avg. Dev.	0.045	0.053	

Cross-lingual

Features	**Test: IT**	**Test: CZ**	**Avg. Dev.**
Baseline	0.553 (=)$^{\text{LR}}$	0.48 (=)$^{\text{LR}}$	0.000
POS n-grams	**0.752** (-0.006)$^{\text{RF}}$	**0.679 (+0.030)**$^{\text{RF}}$	0.018
Dep. n-grams	0.60 **(-0.023)**$^{\text{RF}}$	0.66 (-0.012)$^{\text{RF}}$	0.017
Domain features	0.62 (-0.001)$^{\text{LR}}$	0.46 (-0.009)$^{\text{RF}}$	0.005
Avg. Dev.	0.007	0.017	

→Pred	**A1**	**A2**	**B1**	**B2**	**C1**
A1	3 (-2)	26 (+2)	0	0	0
A2	9 (=)	330 (+19)	39 (-17)	3 (-2)	0
B1	2 (-1)	89 (+19)	260 (-19)	43 (-1)	0

→Pred	**A1**	**A2**	**B1**	**B2**	**C1**
A2	0	134 (+5)	54 (-3)	0 (-2)	0
B1	0	30 (+7)	98 (-3)	37 (-4)	0
B2	0	2 (-3)	24 (-1)	55 (+4)	0

Cross-lingual extension

With Italian As Training Language

Features	Test: DE	Test: CZ
Baseline	**0.711$^{\text{LinSVC}}$**	**0.770$^{\text{LR}}$**
POS n-grams	0.508$^{\text{RF}}$	0.657$^{\text{RF}}$
Dep. n-grams	0.549$^{\text{LinSVC}}$	0.602$^{\text{LinSVC}}$
Domain features	0.706$^{\text{LinSVC}}$	0.756$^{\text{RF}}$

→Pred	A1	A2	B1
A2	0	122	66
B1	0	14	151

→Pred	A1	A2	B1
A1	2	55	0
A2	0	227	79
B1	0	47	284

With Czech As Training Language

Features	Test: DE	Test: IT
Baseline	**0.528**$^{\text{RF}}$	0.697$^{\text{LR}}$
POS n-grams	0.444$^{\text{LR}}$	0.587$^{\text{RF}}$
Dep. n-grams	0.363$^{\text{LR}}$	0.531$^{\text{RF}}$
Domain features	0.478$^{\text{LR}}$	**0.796**$^{\text{LinSVC}}$

→Pred	A2	B1	B2
A2	283	12	3
B1	186	106	39
B2	23	146	124

→Pred	A2	B1	B2
A2	334	36	11
B1	98	164	132

Experiments With Augmented Dataset

Dataset Overview

CEFR level	CZ	DE	IT	EN
A1	0	57	29	0
A2	188	306	381	960
B1	165	331	394	3776
B2	81	293	0	464
C1	0	42	0	0
Total	434	1029	804	5200

Table: Distribution of labels in corpora

Dataset Overview

CEFR level	CZ	DE	IT	EN
A1	-	32.23	39.86	-
A2	93.68	56.89	69.04	214.28
B1	169.81	112.48	145.61	224.54
B2	205.91	187.96	-	232.92
C1	-	220.95	-	-

Table: Average document length per level

Monolingual

Features	English
Baseline	0.333^{LinSVC}
Word n-grams(1)	0.617^{RF}
POS n-grams (2)	0.615^{RF}
Dep. n-grams(3)	0.616^{RF}
Domain features	0.335^{LinSVC}
(1) + domain	$\mathbf{0.629}^{\textbf{RF}}$
(2) + domain	0.620^{RF}
(3) + domain	0.620^{RF}
Word embeddings	0.619

Multilingual

Features	Lang (-)	Lang (+)	Avg. Dev.
Baseline	0.308 **(-0.118)**^{LR}	-	0.118
Word n-grams	0.563 (-0.042)^{RF}	0.559 (-0.048)^{RF}	0.045
POS n-grams	**0.634** (-0.046)^{RF}	**0.634** (-0.046)^{RF}	0.046
Dep. n-grams	0.623 (-0.027)^{RF}	0.620 (-0.032)^{RF}	0.029
Domain features	0.318 (-0.115)^{LR}	0.365 **(-0.082)**^{LR}	0.098
Word embeddings	0.571 (-0.081)	0.572 (-0.073)	0.077
Avg. Dev.	0.071	0.056	

Cross-lingual With English As Training Language

Features	Test: DE	Test: IT	Test: CZ
Baseline	0.272$^{\text{LR}}$	**0.726$^{\text{LR}}$**	0.536$^{\text{LR}}$
POS n-grams	0.431$^{\text{RF}}$	**0.821$^{\text{RF}}$**	0.570$^{\text{RF}}$
Dep. n-grams	0.299$^{\text{LinSVC}}$	**0.580$^{\text{LinSVC}}$**	0.351$^{\text{RF}}$
Domain features	0.289$^{\text{LR}}$	**0.363$^{\text{LR}}$**	0.242$^{\text{LR}}$

Cross-lingual With English As Testing Language

Features	**Train: DE**	**Train: IT**	**Train: CZ**
Baseline	0.075^{RF}	$\mathbf{0.707}^{\textbf{LinSVC}}$	0.400^{RF}
POS n-grams	0.362^{RF}	$\mathbf{0.716}^{\textbf{RF}}$	0.567^{RF}
Dep. n-grams	0.449^{LR}	$\mathbf{0.718}^{\textbf{RF}}$	0.619^{RF}
Domain features	0.107^{RF}	$\mathbf{0.708}^{\textbf{RF}}$	0.614^{RF}

Reproducing Difficulties

Reproducing Difficulties

Execution Warnings:

- Default solver will be changed to 'lbfgs' in 0.22.
- Default multi_class will be changed to 'auto' in 0.22.
- The default value of n_estimators will change from 10 in version 0.20 to 100 in 0.22.

Missing Information:

- Unclear script arguments
- Multiple remaining TODOs

Coding Style

- Lack of comments
- Declaration of unused variables
- Declaration of unused functions
- Declaration of unused imports
- Duplicated functions
- Not following Python coding standards
  - Naming of variables
  - Naming of functions

Conclusion

Conclusion

- Remarkable difference with the published results, especially on the multilingual models.
- The presented approach doesn't scale well when English is added. Possible reasons: text length, lexical diversity and sentence structure.
- No better results in intra-family classification (DE-EN).
- Good correlation between English and Italian.
- Further work: Add features related to the semantic and syntactic analysis of the texts

References

[1] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[2] Sowmya Vajjala and Taraka Rama. Experiments with universal CEFR classification. *CoRR*, abs/1804.06636, 2018.

[3] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 180–189, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.