A Prompt-independent and Interpretable Automated Essay Scoring Method for Chinese Second Language Writing

Yupei Wang[2] Renfen Hu[1]

[1]Institute of Chinese Information Processing
Beijing Normal University

[2]School of Science
Beijing Jiaotong University

Dec 4th, 2021

Table of Contents

Table of Contents

Motivation

The motivation of this work can be summarized as follows:

- AES system for **Chinese L2** writing has received less attention;

Motivation

The motivation of this work can be summarized as follows:

- AES system for **Chinese L2** writing has received less attention;
- Existing models are mainly built in a **prompt-dependent** way;

Motivation

The motivation of this work can be summarized as follows:

- AES system for **Chinese L2** writing has received less attention;
- Existing models are mainly built in a **prompt-dependent** way;
- Neural models are weak in **interpretability** of the results.

Contribution

The contribution of this work is as follows:

- Presenting a model for both **narrative** and **argumentative** essays;

Contribution

The contribution of this work is as follows:

- Presenting a model for both **narrative** and **argumentative** essays;
- Integrating various dimensions of features emphasized in Chinese L2 acquisition, thus **interpretable**;

Contribution

The contribution of this work is as follows:

- Presenting a model for both **narrative** and **argumentative** essays;
- Integrating various dimensions of features emphasized in Chinese L2 acquisition, thus **interpretable**;
- The source code of our method is **publicly available**:) `https://github.com/iris2hu/L2C-rater`.

Table of Contents

Linguistic Complexity Features

We constructs a comprehensive set of linguistic complexity measures of Chinese L2 writing.

- Chinese characters and vocabulary
- Sentences and clauses
- Collocations and bigrams
- Dependency structures
- Constructions
- Writing error features

Chinese characters and vocabulary

We build **four** indices in this dimension:

- Number of Chinese characters **汉字数量**
- Number of Chinese words **词汇数量**
- Lexical diversity **词汇多样性**
- Lexical sophistication **词汇复杂度**

The lexical diversity index is computed as the root type token ratio (RTTR) of words.

The lexical sophistication is built as the ratio of sophisticated words.

Words of HSK-5 level, HSK-6 level and out of the HSK vocabulary are regarded as sophisticated.

Sentences and clauses

Seven indices are proposed to measure the sentence and clausal complexity (the first five):

- The mean length of sentences 平均大句长
- The mean length of clauses 平均小句长
- The mean length of T-units 平均 T 单位长
- Number of clauses per sentence 平均小句数
- Number of T-units per sentence 平均 T 单位数

T-units(T 单位)

A single clause that contains **one independent predicate** plus whatever other subordinate clauses or non-clauses are attached to, or embedded within, that one main clauses.

Sentences and clauses

The next **two**:

- The mean depth of the dependency trees 平均句法树深度
- The max depth of the the dependency trees 最大句法树深度

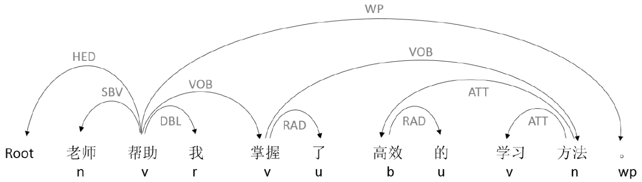


图 1: An example of dependency tree

Collocations and bigrams

First, **eight** types of collocations are considered. **Four** of which are **universal** collocation types existing in different languages, while the other **four** are **language-specific** types.

The **universal** four are as follows:

- Verb-Object(VO) **动宾** ← 喜欢看书；唱着歌
- Subject-Predicate(SP) **主谓** ← 歌曲流行；戒指找回来了
- Adjective-Noun(AN) **形名** ← 著名大学；专业书籍
- Adverb-Predicate(AP) **状中** ← 突然改变；有效地提高

Collocations and bigrams

The **language-specific** four are as follows:

- Classifier-Noun(CN) **量名**：条河；张纸
- Preposition-Postposition(PP) **框式介词**：在 X 上；像 X 似的
- Preposition-Verb(PV) **介动**：把 X 解决；被 X 吃完了
- Predicate-Complement(PC) **述补**：吃饱；玩得愉快

Collocations and bigrams

Besides, to measure the **collocation sophistication**, we introduce:

- Diversity of all the collocations 整体搭配多样性
- Diversity of Chinese unique collocations 特殊搭配多样性
- Diversity of language-independent collocations 一般搭配多样性
- Ratio of Chinese unique collocations 特殊搭配比例
- Ratio of sophisticated collocations [1] 低频 (复杂) 搭配比例

[1]基于某外部语料库定义
[2]同上

Collocations and bigrams

Besides, to measure the **collocation sophistication**, we introduce:

- Diversity of all the collocations 整体搭配多样性
- Diversity of Chinese unique collocations 特殊搭配多样性
- Diversity of language-independent collocations 一般搭配多样性
- Ratio of Chinese unique collocations 特殊搭配比例
- Ratio of sophisticated collocations [1] 低频 (复杂) 搭配比例

To cover *more* language usages, we implement the following two as well by considering the **bigrams** as a specific type of collocations.

- Bigram diversity 二元组多样性
- Bigram sophistication[2] 低频 (复杂) 二元组比例

[1] 基于某外部语料库定义

[2] 同上

Dependency structures

Drawbacks of collocations and bigrams features

- They only target at **part of** the syntactic relations, lacking a whole picture of the syntactic structures;
- They are **NOT** able to measure the fine-grained **phrasal complexity** underlying the structures(e.g. num and len of mod-s).

To address the above two problems, we proposes **41** dependency based indices that measure the **diversity**, **ratio** and **mean distance** (for num and len of mod-s), of all the **dependency triples**.

Dependency structures

Examples of **dependency triples**

- **主谓关系**: (SBV, 老师, 帮助)
- **动宾关系**: (VOB, 掌握, 方法)
- **定中关系**: (ATT, 高效, 方法)

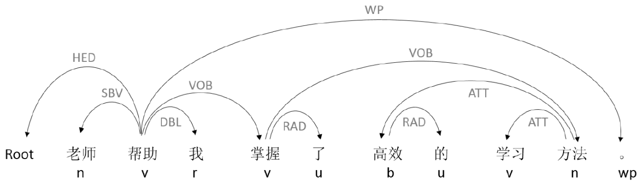


图 2: An example of dependency tree

For more detail of dependency triples, you could check `https://ltp.ai/docs/appendix.html#id5`

Constructions

- We measure the **density** and **ratio** of constructions with regarding to their levels[3].
- **15** indices are built to reflect the **density** and **ratio** of different levels of constructions after automatic recognition.

Example:

这 跟 一架 机器 一样 ， 搁在那里 不用 就要 生锈， 经常 运转才 能 保持 良好状态。

- 2 级： 常用量词, 意愿表达;
- 3 级： 能愿动词, 介词短语 _ 对象, 连动句;
- 4 级： 时间副词;
- 5 级： 地点补语;

[3] 目前系统所识别的语法点参考《国际汉语教学通用课程大纲》(2009 版) 中的 “常用汉语语法项目分级表”. 该表将 62 个常用语法项目由易到难分为五级

Writing Error Features

We adopt **five** indices of writting errors:

- Punctuation errors 标点错误数量
- Chinese character errors 汉字错误数量
- Word level errors 词汇错误数量
- Sentence level errors 句式错误数量
- Discourse level errors 篇章错误数量

be counting them with reference to the annotation in **HSK Dynamic Composition Corpus**.

Multi-granularity Text Features

It's still beneficial to retain the full picture of the textual features. We extract **character**, **word** and **part-of-speech unigrams**, **bigrams** and **trigrams** as features. We use the **tf–idf** weighted representations of these features, and each essay can be represented as a text vector:

$$TextVec = (tfidf_1, tfidf_2, \ldots, tfidf_N) \tag{1}$$

The Ordinal Logistic Regression Model

We proposes to use the Ordinal Logistic Regression (OLR) model in Chinese L2 AES since the it's effective for ordinal categories.
A practical loss of ordinal classification is **threshold-based**, which is divided into **Immediate-threshold loss** and **All-threshold loss**. We use All-threshold loss, which is represented as

$$\mathrm{Loss}_{\mathrm{AT}}(z) = \sum_{k=1}^{l-1} f\left(s(k;i)\left(\theta_k - z\right)\right) \quad s(k;i) = \begin{cases} -1 & k < i \\ +1 & k \geq i \end{cases} \tag{2}$$

where z is a specific predicted value, (θ_{i-1}, θ_i) refers to the **correct** segment, and $f(\cdot)$ could be any kind of loss function for multiclass classification.

The Ordinal Logistic Regression Model

Bringing $h(z) := \log(1 + \exp(z))$ into $\mathrm{Loss}_{AT}(\cdot)$ as $f(\cdot)$ gives the minimization objective for All-threshold Ordinal Logistic Regression:

$$\mathrm{Loss}_{\text{OLR-AT}} = \sum_{i=1}^{N} \left[\sum_{k=1}^{y_i - 1} h\left(\theta_k - \mathbf{x}_i^T \mathbf{w}\right) + \sum_{k=y_i}^{l-1} h\left(\mathbf{x}_i^T \mathbf{w} - \theta_k\right) \right] + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \tag{3}$$

where label $k \in \{1, \ldots, l\}$ corresponds to the segment (θ_{k-1}, θ_k). θ_0 and θ_l denotes $-\infty$ and $+\infty$ respectively. $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^{d^2}$ are training examples while $\{y_1, \ldots, y_n\}, y_i \in \{1, \ldots, l\}$ are their labels.

The Ordinal Logistic Regression Model

Bringing $h(z) := \log(1 + \exp(z))$ into $\mathrm{Loss}_{AT}(\cdot)$ as $f(\cdot)$ gives the minimization objective for All-threshold Ordinal Logistic Regression:

$$\mathrm{Loss}_{\text{OLR-AT}} = \sum_{i=1}^{N} \left[\sum_{k=1}^{y_i - 1} h\left(\theta_k - \mathbf{x}_i^T \mathbf{w}\right) + \sum_{k=y_i}^{l-1} h\left(\mathbf{x}_i^T \mathbf{w} - \theta_k\right) \right] + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \tag{3}$$

where label $k \in \{1, \ldots, l\}$ corresponds to the segment (θ_{k-1}, θ_k). θ_0 and θ_l denotes $-\infty$ and $+\infty$ respectively. $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^{d^2}$ are training examples while $\{y_1, \ldots, y_n\}, y_i \in \{1, \ldots, l\}$ are their labels.

◇ Contrast

The regularized logistic regression minimization objective:

$$\mathrm{Loss}_{\mathrm{RLR}} = \sum_{i=1}^{N} \log\left(1 + \exp\left(-y_i \cdot \mathbf{x}_i^T \mathbf{w}\right)\right) + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \tag{4}$$

Table of Contents

Dataset and Preprocessing

- Using essay data from **HSK Dynamic Composition Corpus**;
- The essays are rated from **40p to 95p** with an interval of **five**, yielding **12** different categories;
- 10277 **argumentative** and **narrative** essays are involved;
- **5-fold** cross validation with Train-7040/Dev-1760/Test-1477.

Feature Selection

We conduct **step-wise** linear regression in each dimension of linguistic complexity and writing error indices to examine their **predictive power**.

Dimension	R	R^2
Chinese characters and vocabulary (4, 3)	0.648	0.420
Sentences and clauses (7, 4)	0.197	0.039
Collocations and bigrams (23, 8)	0.587	0.345
Dependency structures (41, 16)	0.610	0.372
Constructions (15, 9)	0.248	0.061
Writing Error Features (5, 4)	0.254	0.065

表 1: Step-wise regression results in **each dim**. The numbers in brackets denote the number of indices **entered and remained** in the step-wise regression.

For the **90** linguistic complexity indices, **33** were selected by step-wise regression, and it yields **31** after **integrating the writing error features**.

Models

We build two types of baselines including **regression-based** and **tree-based** ML models that **share the same feature space** with OLR model:

- Linear Regression
- Logistic Regression
- Random Forest Regression
- XGBoost Regression

as well as two other effective **neural** models:

- `CNN+LSTM` by *Taghipour and Ng(2016)*
- `Att-BLSTM` by *Zhou et al.(2016)*

Evaluation Metrics

There are many metrics that can measure the **consistency** between AES systems and human experts. In this study we employ three of them:

- Quadratic Weighted Kappa(QWK) 二次加权 κ
- Root Mean Square Error(RMSE) 均方根误差
- Pearson coefficient(Pears.) 皮尔逊相关系数

Results

Method	Mode	QWK	RMSE	Pears.	Mode	QWK	RMSE	Pears.
LiR	L	0.640	**1.636**	**0.679**	L+T	0.269	3.576	0.299
	L+E	**0.668**	**1.585**	**0.702**	L+E+T	0.276	3.557	0.307
LoR	L	0.598	1.813	0.620	L+T	0.641	1.720	0.663
	L+E	0.640	1.715	0.661	L+E+T	0.663	1.667	0.681
RFR	L	0.625	1.657	0.668	L+T	0.652	1.603	0.694
	L+E	0.655	1.601	0.695	L+E+T	0.667	1.575	0.706
XGBR	L	0.576	1.690	0.652	L+T	0.587	1.676	0.659
	L+E	0.613	1.625	0.687	L+E+T	0.621	1.616	0.690
CNN+LSTM	Rand	0.496	1.845	0.551	Sogou	0.504	1.831	0.560
Att-BLSTM	Rand	0.520	1.825	0.568	Sogou	0.531	1.812	0.578
OLR-AT	L	**0.644**	1.650	0.674	L+T	**0.697**	**1.554**	**0.718**
	L+E	0.666	1.616	0.691	L+E+T	**0.714**	**1.516**	**0.734**

表 2: Results of Chinese L2 AES. The **bold** denotes the best result under the same feature setting.

Observations

- All models obtain the **best** results under L+E+T **except for** LiR;
- LiR achieves **almost the best** results under L and L+E;
- The effect of the neural AES model is **temporarily** weaker than methods based on feature engineering;
- After adding text features to L+E, the performance of OLR-AT improves by 7.2%, compared with 3.6% of LoR, 1.8% of RFR and 1.3% of XGBR.

Table of Contents

Analysis on Confusion Matrix

To illustrate the models' behaviors, Figure 3 shows the confusion matrix of the `OLR-AT` model under `L+E+T`.

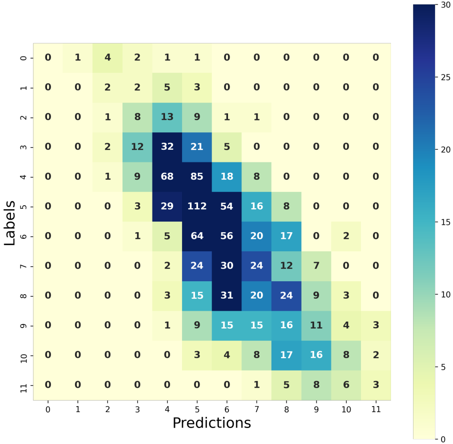


图 3: Confusion Matrix of OLR-AT Results

Bad Cases from Confusion Matrix

◇ For essays whose predicted scores too **high**:

- The contents **deviate from their prompts**;
- **Lacking of organization** when expressing opinions (for argumentative essays).

◇ For essays whose predicted scores too **low**:

- Rating exceptions by the human raters, e.g. giving high scores to **unfinished essays**.

Revisiting Linear Regression

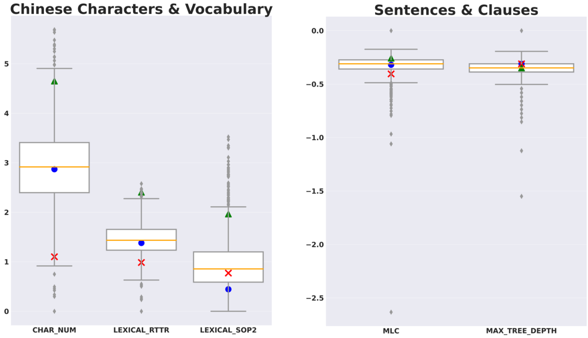
Mode	QWK	RMSE	Pears.
T	0.207	3.787	0.232
L+T	0.269	3.576	0.299
L+E+T	0.276	3.557	0.307

表 3: The results of Linear Regression with different feature sets.

Method	Mode	QWK	RMSE	Pears.	Mode	QWK	RMSE	Pears.
LiR	L	0.640	1.636	0.679	L+T	0.269	3.576	0.299
	L+E	0.668	1.585	0.702	L+E+T	0.276	3.557	0.307
Ridge	L	0.636	1.640	0.676	L+T	0.694	1.538	0.723
	L+E	0.667	1.585	0.702	L+E+T	0.709	1.510	0.735

表 4: The comparison of Linear Regression and Ridge Regression

Effect Plot

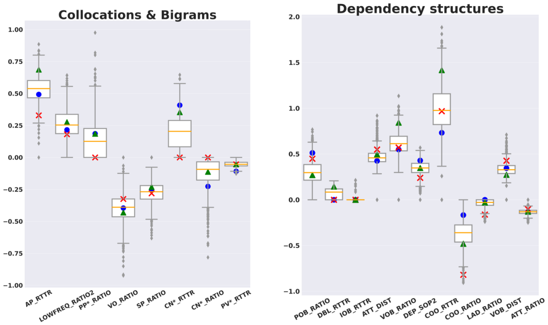


(a) CHN Char & Vocab (b) Sentences & Clauses

图 4: Effect plot[4] - Part 1

[4] The green triangle for a essay of *95p*; The blue circle for a essay of *65p*; The red cross for a essay of *45p*.

Effect Plot

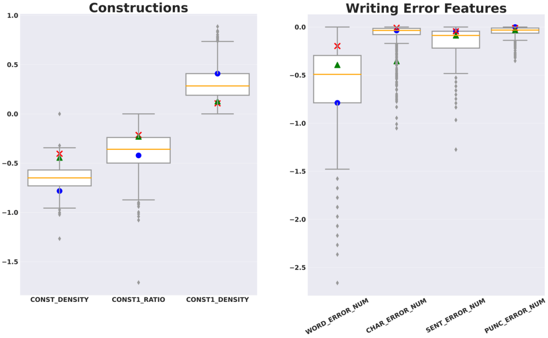


(a) Collocations&bigrams (b) Dependency structures

图 5: Effect plot[5] - Part 2

[5]The green triangle for a essay of *95p*; The blue circle for a essay of *65p*; The red cross for a essay of *45p*.

Effect Plot



(a) Constructions (b) Writing Error Features

图 6: Effect plot[6] - Part 3

[6] The green triangle for a essay of *95p*; The blue circle for a essay of *65p*; The red cross for a essay of *45p*.

Table of Contents

Conclusion and Future Work

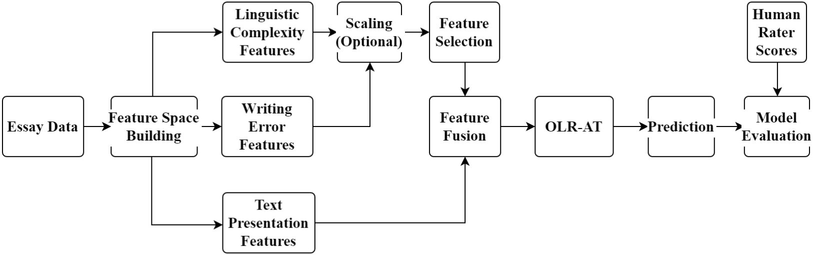


图 7: Pipeline of the model

Conclusion and Future Work

Summary:

- **Explainable representations** of both linguistic and text features are built;
- The most effective combination: `OLR-AT / L+E+T`;
- Potential to offer users writing **feedback**.

Next step:

- Modeling **more dimension** of essay quality such like **fluency** , **coherence**, **prompt-adherence** and so on;
- Trying to make **automatic feedback** more **accurate and helpful**;
- Further exploiting the potential of **neural models** on AES tasks.

Thank You!