Improving Automatic Grammatical Error Annotation for Chinese Through Linguistically-Informed Error Typology

Yang Gu[1] Zihao Huang[1] Min Zeng[2]
Mengyang Qiu[1,3] Jungyeul Park[1,4]

[1]Open Writing Evaluation, France [2]TEKsystems, Canada
[3]Department of Psychology, Trent University, Canada
[4]Department of Linguistics, The University of British Columbia, Canada
http://open-writing-evaluation.github.io

Introduction

- What is GEC?
  - Detects and corrects grammatical errors in text
  - Crucial for language learning and writing improvement
- Why focus on Chinese?
  - Growing demand for high-quality Chinese writing support tools
  - Chinese poses unique challenges due to its writing system
- Existing annotation tools:
  - `ERRANT` (English): Provides detailed annotations (Bryant et al., 2017)
  - `ChERRANT`: Adapted to Chinese but limited to basic character-level errors (Zhang et al., 2022)
- What this study does:
  - Refines the ChERRANT framework
  - Introduces a linguistically-informed typology for Chinese-specific errors for both L1 and L2 learners

Features of the Chinese Writing System

- No explicit word boundaries (e.g., continuous text without spaces)
  - Words can be $1-2+$ characters
  - Example: 被 (*bèi*) can be:
    - Standalone passive marker
    - Part of words like 被子 (*bèizi*, "blanket")
- Many homophones (e.g., 权力 vs. 权利) and visually similar characters (e.g., 四 vs. 西)
- Implications for GEC:
  - Tokenization errors
  - Misidentification of grammatical roles and structures

Previous CGEC Annotation System–ChEERRANT

- Examined:
  - Manual vs. automated annotations
  - L1 vs. L2 writing via automated annotations
- Validation sets of FCGEC (Xu et al., 2022) and MuCGEC (Zhang et al., 2022)

Dataset	# of sentences		Different numbers of errors per sentence					
	Ori.	Ref.	0	1	2	3	4	≥5
FCGEC (L1)	2000	2143	899	1138	101	3	2	0
MuCGEC (L2)	1138	1153	55	214	238	176	153	317

Previous CGEC Annotation System–ChEERRANT

- Manual vs. automated annotations
  - Automated annotations align with manual annotations in up to 95% of cases, with a match rate of 73% for more complex sentences containing multiple grammatical errors
  - Differences in word boundary may lead to different error type definitions over different spans of text
  - Problems of dealing with changing the order of large spans of text like phrases and clauses
- L1 vs. L2 writing
  - The ChERRANT analysis results are consistent with conventional assumptions about the differences between L1 and L2 writing
  - ChERRANT is less reliable when handling sentences with multiple complex errors (e.g., a potential underrepresentation of ordering errors)

Refined Annotation for CGEC

- A new error typology that better captures both L1 and L2 grammatical errors, along with a new implementation based on this scheme
  - Similar pronunciation
  - Similar shapes
  - Multifaceted similarity
  - The *de* particles–的, 地, and 得
  - Character order

Refined Annotation for CGEC

	Learner writing	Correction
(1)	每 个 抽烟 的 人 都 有 这样 的 权力 。 'Every smoker has this power.'	每 个 抽烟 的 人 都 有 这样 的 权利 。 'Every smoker has this right.'
(2)	交 朋友 的 时 后 ， 很 可能 会 碰到 矛盾 。 'Time after making friends, it's highly likely to run into conflicts.'	交 朋友 的 时候 ， 很 可能 会 碰到 矛盾 。 'When making friends, it's highly likely to run into conflicts.'
(3)	我 一 前 没 住 过 五星级 旅馆 ， 所以 我 很 惊讶 。 'I haven't stayed in five-star hotels one before, so I am very surprised.'	我 以前 没 住 过 五星级 旅馆 ， 所以 我 很 惊讶 。 'I haven't stayed in five-star hotels before, so I am very.'
(4)	她 有 两 个 姐姐 、一 个 妹妹 和 西 个 哥哥 。 'She has two older sisters, one younger sister, and west older brothers.'	她 有 两 个 姐姐 、一 个 妹妹 和 四 个 哥哥 。 'She has two older sisters, one younger sister, and four older brothers.'
(5)	从 十六 世纪 开始 ， 欧州人 就 抽烟 。 'Since the 16th century, Europe region people smoked.'	从 十六 世纪 开始 ， 欧洲人 就 抽烟 。 'Since the 16th century, Europeans smoked.'
(6)	经理 ， 新 得 计划 发 您 信箱 了 ， 您 看 了 吧 ？ 'Manager, the new obtain plan was sent to your mailbox, you've seen it, right?'	经理 ， 新 的 计划 发 您 信箱 了 ， 您 看 了 吗 ？ 'Manager, the new MOD plan was sent to your mailbox, have you seen it?'
(7)	简单 生活 ， 哪怕 对 身体 还是 精神 ， 还 大有 裨益 。 'A simple life, whether for the body or the mind, is greatly beneficial.'	简单 的 生活 ， 无论 对 身体 还是 精神 ， 都 大有 裨益 。 'A simple MOD life, whether for the body or the mind, is greatly beneficial.'
(8)	反而 那些 不 帅 ， 还 有 点 丑 但是 很 会 唱歌 就 被 淘汰 了 。 'Instead, those who are not handsome, a bit ugly, but can sing well were eliminated.'	反而 那些 不 帅 ， 还 有 点 丑 但是 很 会 唱歌 的 就 被 淘汰 了 。 'Instead, those who are not handsome, a bit ugly, but can sing well were eliminated.'
(9)	要 了解 一 个 人 ， 不妨 看 他 读 些 么 什 书 ， 观察 向 他 来往 得 朋友 一样 有效 。 'To understand a person, it's just as effective to see what books he reads and observe the friends he interacts with.'	要 了解 一 个 人 ， 不妨 看 他 读 些 什么 书 ， 这 跟 观察 与 他 来往 的 朋友 一样 有效 。 'To understand a person, it's just as effective to see what books he reads and observe the friends he interacts with.'

Refined Annotation for CGEC

- Stanza for segmentation and tagging

Algorithm 1 Pseudo-code for error classification

```
function ErrorClassification (S, T, {R|M|U}):
    if R ∧ (Sim (pinyin) > α1) ∧ (Sim (shape) > α2) then
        return R:MULTI
    else if R ∧ (Sim (pinyin) > α1) then
        if T == de then
            return R:DE
        else
            return R:PINYIN
        end if
    else if R ∧ (Sim (shape) > α2) then
        return R:SHAPE
    else if R ∧ (Set (S) == Set (T)) then
        if Len (T) == 1 then
            return R:CO
        else
            return R:WO
        end if
    else if M ∧ (T == de) then
        return M:DE
    end if
    return {R|M|U}
```

Refined Annotation for CGEC

- Identified 12-16% more errors than `ChERRANT`
- Greater detail in annotating spelling and structural errors

	L1		L2	
Error Type	Count	Ratio	Count	Ratio
M:DE	6	0.0051	8	0.0207
R:DE	5	0.0032	14	0.0031
R:MULTI	73	0.0466	197	0.0433
R:PINYIN	21	0.0134	146	0.0321
R:SHAPE	20	0.0128	164	0.0361
R:WO	63	0.0402	148	0.0326
Total	188	0.1212	677	0.1678

Conclusion and Future Work

- Key contributions:
  - Introduced a refined error typology tailored to Chinese
  - Highlighted differences in L1 and L2 errors
- Future directions:
  - Address more context-sensitive grammatical issues in addition to surface-level errors
  - Improve robustness of segmentation and annotation tools

Thank you!

References

Bryant, C., Felice, M., and Briscoe, T. (2017). Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Xu, L., Wu, J., Peng, J., Fu, J., and Cai, M. (2022). FCGEC: Fine-Grained Corpus for Chinese Grammatical Error Correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1900–1918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhang, Y., Li, Z., Bao, Z., Li, J., Zhang, B., Li, C., Huang, F., and Zhang, M. (2022). MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States. Association for Computational Linguistics.