CompLex-ZH: A New Dataset for Lexical Complexity Prediction in Mandarin and Cantonese

Le Qiu[1], Shanyue Guo[1], Tak-sum Wong[1], Emmanuele Chersoni[1], John S. Y. Lee[2], Chu-Ren Huang[1]
[1]The Hong Kong Polytechnic University, [2]City University of Hong Kong

TSAR 2024 @ EMNLP 2024

Introduction

Background

- Lexical complexity, being the first step of text simplification pipelines, has received increasing attention in the NLP community.
- However, new datasets and several shared tasks are available only for English and for a limited number of Western languages (Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021; Saggion et al., 2023; Shardlow et al., 2024)

- **The present study[1]**
  - Introduces CompLex-ZH, the first evaluation benchmark for lexical complexity prediction in Chinese.
  - Included two different Sinitic varieties: Mandarin, the standard Chinese, and Cantonese, a major variety of Chinese but having a low-resource status in terms of NLP research.
  - Provides a preliminary evaluation with a bassline regressor based on a combination of hand-crafted features and contextualized embeddings.

- Mandarin Chinese
  - The standard variety of Chinese
- Yue Chinese, or Cantonese
  - Colloquial
  - Different from Mandarin in vocabulary, grammar, and pronunciation
  - Hong Kong, Macao, Guangdong, Guangxi, South-East Asia, North America and Western Europe (Sachs and Li, 2007; Yu, 2013; Xiang et al., 2024)

Table 1. Mandarin vs. Cantonese

Related Work

Previous studies on text simplification:

❖ Mostly limited to Western languages, including English, Portuguese, Spanish, etc. (see Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021; Saggion et al., 2023; Shardlow et al., 2024)

❖ Notably, Qiang et al., (2021) only included *high-level* words in their research on Chinese lexical simplification.

Complexity prediction:

- Once seen as a binary problem (e.g., Paetzold and Specia, 2016; Yimam et al, 2018)
- First treated as a regression task at the Task 1 at SemEval-2021 (Shardlow et al., 2021)

CompLex (Shardlow et al., 2020, 2022)

- A gold standard dataset on English lexical complexity

CompLex-ZH:

- Benchmarking Chinese lexical complexity for the first time
- Includes varying degrees of complexity
- Carefully built from different sources and text genres
- Features complexity ratings provided by native speakers
- Incorporates both Mandarin Chinese and Cantonese

Dataset Creation

Target Selection

- **Data source**
  - Mandarin
    - Weibo
    - People's Daily,
    - BCC corpus (Xun et al., 2016),
    - Chinese Wikipedia
  - Cantonese
    - LIHKG
    - Cantonese Wikipedia
    - Counseling corpus (Lee et al., 2020)
    - PolyU Corpus of Spoken Chinese
- **The workflow of data processing**
  - Target word filtering based on frequency and part-of-speech
  - 1017 target words and 3240 samples collected for Mandarin;
  - 260 targets words and 2502 samples for Cantonese

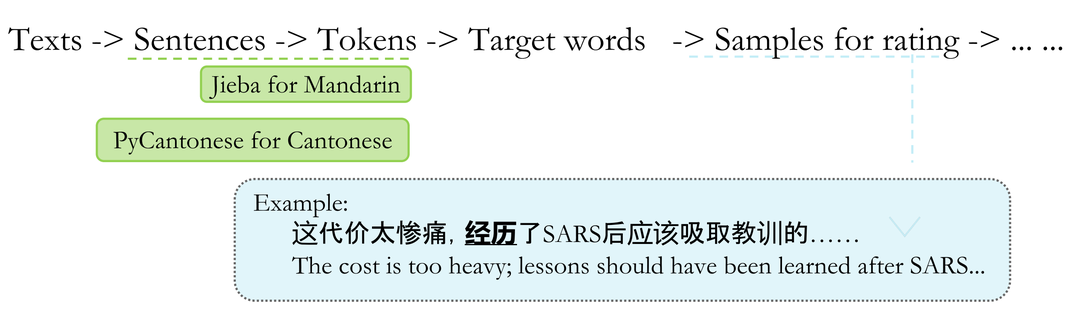

Figure 1. Workflow of data processing

Rating Collection

- Questionnaire (~300 for each variety)
  - Question (102)
    - 100 normal + 2 validation samples
  - Option (in a 5-point Likert scale)
    - 1 being very easy, 5 being very difficult
- Raters per sample (>= 5)
- Complexity score
  - Sample-wise complexity
    - Average of scores by all raters
  - Word-wise complexity
    - Average of all sample-wise scores

Context	Score
... 忽然变得澄清见底，翳障 全无。 ...it turns crystal, without obstacles in sight.	.213
此前有团队 已经在粪便里发现新冠病毒。 The team had found coronavirus in feces.	.893
... 感受到被失蹤、被跟蹤的實在... ...I truly felt disappeared and stalked...	.588
點解講GOOD JOB 佢反而又呆哂... Why he acts so dumb and ... when you said GOOD JOB?	.200

Table 2. Some examples with average high/ low complexity scores. The first 2 are in Mandarin and the last 2 in Cantonese. Target words are underlined.

Evaluation

Experimental Setting

Formulation: **Ridge regression**

- Given a sentence s with a target word t, the model tries to predict the complexity score c.

- Input:
  - Handcrafted features (**HC**)
    - Word length (**WLen**)
    - Word frequency (**LogF**)
    - Stroke
  - Contextualized word embeddings (**Emb**) from CINO[2], a PLM trained on Mandarin and several minority languages in China
- Metrics
  - Coefficient of determination ($\boldsymbol{R^2}$) -> [0, 1]
  - Mean absolute error (**MAE**) -> [0, $+\infty$), 0 means a perfect prediction
  - Spearman's rank correlation coefficient ($\boldsymbol{\rho}$) -> [-1, 1]

Train: val: test = 8: 1: 1

- Test set size:
  - 324 instances for Mandarin, 250 for Cantonese
  - 574 for a joint dataset of both

Results and Findings

- Contextualized embeddings outperform out-of-context HC features
- LogF is most predictive among HC features,
- Values in both languages are similar, but explained variance in Cantonese is much lower
  - confirming the Cantonese data pose a nontrivial challenge for Chinese NLP
- Scores in general are relatively low
  - suggesting the need for more sophisticated approaches to Chinese lexical complexity prediction

	Feat.	MAE	R^2	ρ
Mand.	HC	.065	.186	.091
	Stroke	.065	.083	.107
	WLen	.065	.055	.082
	LogF	.065	.201	.061
	Emb	**.059**	**.355**	**.338**
	Comb.	.060	.086	.322
Canto.	HC	**.060**	.051	.191
	Stroke	.063	-.001	.008
	WLen	.063	.0184	.158
	LogF	.061	.022	.149
	Emb	.061	**.056**	.353
	Comb.	.061	.045	**.354**
Joint	HC	.065	.047	.135
	Stroke	.066	-.002	-.015
	WLen	.066	-.002	-.109
	LogF	.066	.040	.116
	Emb	**.062**	.131	**.329**
	Comb.	**.062**	**.136**	.326

Table 3: Evaluation results. Comb. indicates the combination of the most influential LogF features and the embedding features.

Conclusion

- **Introduction of ComPlex-ZH**:
  - The first dataset for lexical complexity evaluation in Mandarin and Cantonese.
- **Preliminary Findings:**
  - Contextualized embeddings are more predictive of lexical complexity, compared to handcrafted, out-of-context features that were commonly used in the literature.
- **Challenges Noted**:
  - Limited accuracy, weak-to-moderate correlations, and low explained variance suggest improvement needed.

[1]Code and data are available at https://github.com/Laniqiu/CompLex-ZH.
[2]https://github.com/iflytek/cino?tab=readme-ov-file

Emmanuele Chersoni
emmanuele.chersoni@polyu.edu.hk