COMPILING: A Benchmark Dataset for Chinese Complexity Controllable Definition Generation

Jiaxin Yuan Cunliang Kong Chenhui Xie
Liner Yang Erhong Yang
Beijing Language and Culture University

Motivation and Background

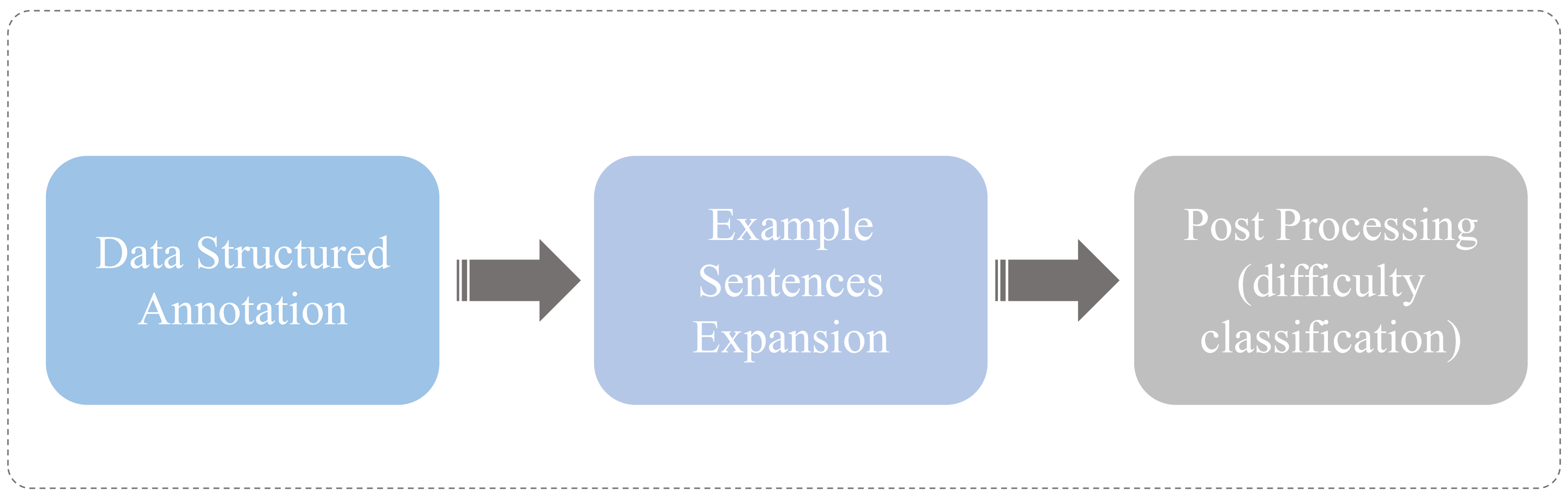
- High complexity problem in the studies of Definition Generation (DG) is prevalent: generated definitions contain words that **are more difficult** than the defined word, which are labored for language learners to understand.
- The existing Chinese learner dictionaries is not enough for Chinese as Foreign Language (CFL) learners:

Current issues of existing Chinese learner dicitionaries:

Complexity level	The number of words
The difficulty of definitions is not considered.	The existing dictionaries contain only a small number of words.

- We focus on the task of **generating defifinitions for CFL learners with appropriate complexities**.
- Considering there is no dataset providing the complexity of definitions, which is essential information in the controllable generation,we build a novel benchmark dataset named **COMPILING.**
- In order to quantitatively **measure the complexity of definitions**, we refer to the graded vocabularies formulated by HSK (Chinese Proficiency Test).

Dataset Construction


Data Structured Annotation
Example Sentences Expansion
Post Processing (difficulty classification)

- **Data Structured Annotation**:To turn disorganized data into structured ones,which is conducive for computers to extract this information automatically.
- **Example Sentences Expansion**:The original context attached to the targeted words given in dictionaries is too short to provide enough knowledge for the model to learn and generate descriptions.

```
Algorithm 1 Example Sentences Expansion
Input: phrase $p$, corpus $C$
Output: examples $E$
$D \leftarrow \{\}, E \leftarrow []$
for $sentence$ in $C$ do
    if $p$ in $sentence$ then
        $score \leftarrow pplScore(sentence)$        ▷ Compute the PPL score for each sentence.
        $D[sentence] \leftarrow score$
    end if
end for
$sortedExamples \leftarrow descSortByValue(D)$        ▷ Descendant sort by the scores.
for $i = 0 \rightarrow topN$ do        ▷ $topN$ is set to 5 in practice.
    $E.add(sortedExamples[i])$
end for
```

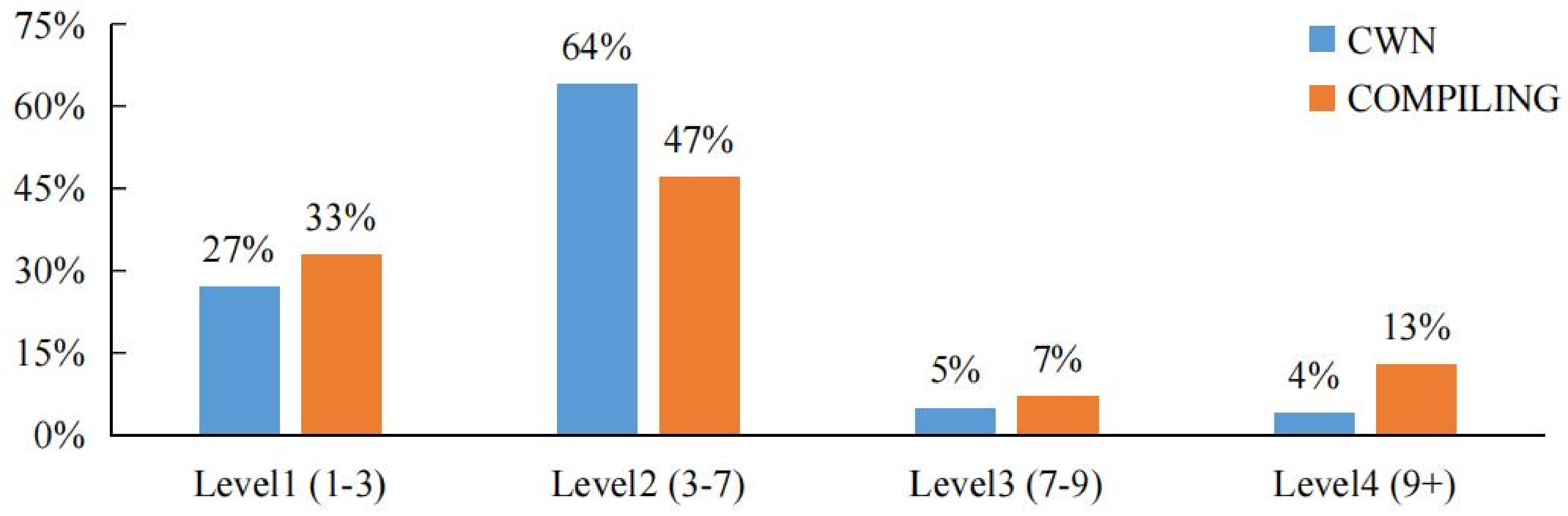
- **Post Processing**:We calculate the average and highest HSK level, and combine the HSK level into the dataset.
- Eventually, each entry of the dataset consists of a target word,its definition, the average and highest HSK level, and the contexts of the corresponding usage of this description.

Dataset Analysis

- Compared to another dataset of Chinese definition modeling, COMPILING dataset **includes more words and has longer definition and context**.

Datasets	Count		Average Length	
	Words	Entries	Definition	Context
CWN	8,221	84,542	9.07	21.57
COMPILING	74,303	127,757	13.60	23.95

- **The distribution of** definitions in the COMPILING dataset **in the three complexity levels is closer than** CWN, supporting model to generate definitions of any target complexity level.



Experiments and Results

Regardless of complexity levels

Models	Dev			Test		
	BLEU	NIST	HSK	BLEU	NIST	HSK
LOG-CaD	27.66	25.55	3.74	27.71	27.88	3.85
Transformer	28.61	25.85	3.92	28.58	31.00	3.96
BERT	**32.95**	29.66	4.05	**32.03**	30.56	4.08
BART	29.49	**36.90**	**4.76**	30.63	**42.79**	**4.80**

- The results show that PLMs outperforms the other two methods in terms of the BLEU and NIST scores apparently.

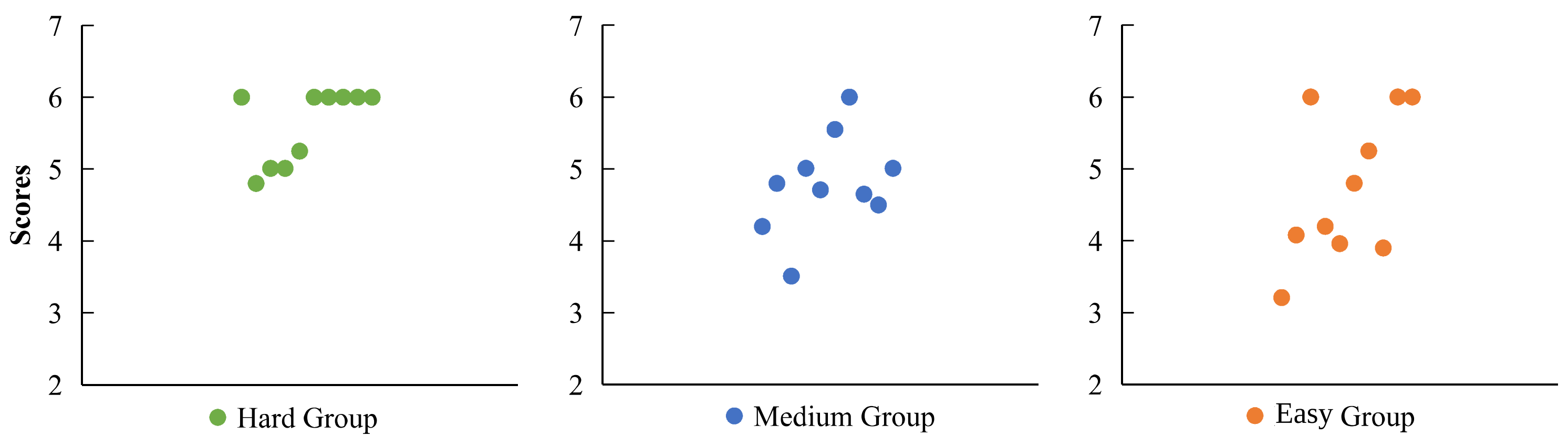
Complexity specific models

- Even on different test sets, definitions generated by the same model have similar complexity.

Models	Easy Set			Medium Set			Hard Set		
	BLEU	NIST	HSK	BLEU	NIST	HSK	BLEU	NIST	HSK
BART-Easy	**32.44**	**64.40**	2.40	21.56	27.61	2.73	25.89	7.95	2.74
BART-Medium	22.92	24.59	4.70	**27.69**	**40.68**	4.86	29.37	16.09	5.01
BART-Hard	22.49	3.55	**8.46**	23.70	7.04	**8.45**	**46.57**	**18.22**	**8.76**

Unified model based on prompt learning

- The definition in the Easy Group scored the lowest overall score.
- It means the difficulty level of the model-generated interpretations obtained by automatic evaluation is roughly in line with expectations.
- The result proves the effectiveness of prompt learning on complexity controllable task.



Conclusion

- We propose a novel task of generating definitions for a word with appropriate complexity.
- We propose the **COMPILING** dataset that is of large scale and high quality.
- We perform several experiments on the COMPILING dataset and the results demonstrate it could assist models to achieve effective complexity controllable generations.