NLPCC2023 Shared Task 8:

Chinese Spelling Check

1 Introduction

1.1 Background

In today's highly developed media landscape, news and other textual content are constantly being produced, but errors in spelling and word usage are not uncommon. Given the sheer volume of content, it is difficult for traditional manual review and regulation to effectively and comprehensively monitor for errors. Therefore, we hope to use AI technology to assist editorial staff in identifying errors in content, thereby improving content quality, increasing editing efficiency, and reducing the risk of errors. This task involves spelling errors such as homophonic errors, visually similar errors, and other types of errors in sentences from Chinese news articles.

1.2 Data Source

Data was obtained from publicly available news articles and books where errors were manually identified and labeled.

1.3 Task Requirements

Identify locations in the content where errors exist and provide corresponding correction suggestions.

1.4 Significance

By using AI technology to assist editorial staff in identifying errors in content, we can improve content quality, increase editing efficiency, and reduce the risk of errors.

2 Updates

All updates about this shared task will be posted on this page.

3 Important Dates

- 2023/03/15：registration open
- 2023/04/06：release of detailed task guidelines & training data
- 2023/05/05：registration deadline
- 2023/05/21：release of test data
- 2023/05/31：participants’ results submission deadline
- 2023/06/10：evaluation results release and call for system reports and conference paper

4 Dataset

The development dataset comprises of 1000 sentence pairs, consisting of both incorrect and correct sentences. The file format is such that each line contains a raw sentence followed by its corresponding correct answer, separated by a tab character: **Input** `\t` **Gold**. (For further details, please refer to the `data` folder.)

You are free to use the development dataset for local training and testing. You may also use any other open-source datasets such as SIGHAN13, SIGHAN14, SIGHAN15, Lang8, HSK, CGED, MuCGEC, YACLC, CTC2021 and so on.

The errors in the original sentences fall into three categories:

4.1 Homophonic Spelling Errors:

```
【例】公司在处理技术、产品设计、检验检测等方面有着坚实的基础和出色的造诣，形成了较强的技术优式。
【答案】公司在处理技术、产品设计、检验检测等方面有着坚实的基础和出色的造诣，形成了较强的技术优势。
【例】知觉告诉他，这是正确的选择。
【答案】直觉告诉他，这是正确的选择。
【例】碳成本激增或将危机油气行业。
【答案】碳成本激增或将危及油气行业。
【例】给予多特征融合的目标跟踪算法系统通过图像处理和分析技术、机器学习和模式识别来识别和分析人体的位置和运动。
【答案】基于多特征融合的目标跟踪算法系统通过图像处理和分析技术、机器学习和模式识别来识别和分析人体的位置和运动。
```

4.2 Visually Similar Errors:

```
【例】严格落实"开喷淋、常冲洗、勤酒水"等防治措施。
【答案】严格落实"开喷淋、常冲洗、勤洒水"等防治措施。
【例】书本是人类灵魂的桥梁,是人类思想选代升级的阶梯,是人类认知传承的纽带。
【答案】书本是人类灵魂的桥梁,是人类思想迭代升级的阶梯,是人类认知传承的纽带。
```

4.3 Other Types of Errors:

```
【例】然而几个以前，张大妈还深陷在窨井盖爆炸的阴影中。
【答案】然而几个月前，张大妈还深陷在窨井盖爆炸的阴影中。
【例】我县聚焦青年人才成长，以人才制度创新推进区域性青年人才高地建设，打造人擦荟萃、要素集聚、业态繁荣的创新创业热土。
【答案】我县聚焦青年人才成长，以人才制度创新推进区域性青年人才高地建设，打造人才荟萃、要素集聚、业态繁荣的创新创业热土。
【例】老一辈科学家身上充沛着为科学而献身的可贵精神。
【答案】老一辈科学家身上充满着为科学而献身的可贵精神。
```

5 Evaluation Metrics

The following metrics will be used to evaluate the performance of the system:

5.1 Detection Level

- The detection precision is calculated as the number of overlapping positions between the detected errors and the standard detection answers divided by the total number of detected errors, expressed as a percentage. The formula is as follows:

$$\text{Detection Precision} = \frac{(|\{\text{Detection Outputs}\} \cap \{\text{Detection Answers}\}|}{|\{\text{Detection Outputs}\}|)} * 100\%$$

- The detection recall is calculated as the number of overlapping positions between the detected errors and the standard detection answers divided by the total number of standard detection answers, expressed as a percentage. The formula is as follows:

$$\text{Detection Recall} = \frac{(|\{\text{Detected Outputs}\} \cap \{\text{Detection Answers}\}|)}{|\{\text{Detection Answers}\}|} * 100\%$$

- The detection F1-score is the harmonic mean of the detection precision and recall. The formula is as follows:

$$\text{Detection F1-score} = 2 * \frac{(\text{Detection Precision} * \text{Detection Recall})}{(\text{Detection Precision} + \text{Detection Recall})}$$

5.2 Correction Level

- The correction precision is calculated as the number of corrected characters that match the standard correction answers divided by the total number of edited characters, expressed as a percentage. The formula is as follows:

$$\text{Correction Precision} = \frac{(|\{\text{Corrected Characters}\} \cap \{\text{Correction Answers}\}|)}{|\{\text{Corrected Characters}\}|} * 100\%$$

- The correction recall is calculated as the number of corrected characters that match the standard correction answers divided by the total number of standard correction answers, expressed as a percentage. The formula is as follows:

$$\text{Correction Recall} = \frac{(|\{\text{Corrected Characters}\} \cap \{\text{Correction Answers}\}|)}{|\{\text{Correction Answers}\}|} * 100\%$$

- The correction F1-score is the harmonic mean of the correction precision and recall. The formula is as follows:

$$\text{Correction F1-score} = 2 * \frac{(\text{Correction Precision} * \text{Correction Recall})}{(\text{Correction Precision} + \text{Correction Recall})}$$

These evaluation metrics are used to assess the performance of the AI model in detecting and correcting spelling errors in textual content. The precision, recall, and F1 scores for both detection and correction provide a comprehensive evaluation of the model's ability to identify and rectify errors accurately and efficiently. The higher the scores, the better the performance of the model.

5.3 False Positive Rate

Given the practical considerations of real-world applications, we incorporate the **false positive rate** at the sentence level into our analysis. Specifically, this involves calculating the proportion of correctly formed sentences that are incorrectly altered by the model.

5.4 Evaluation Method

We provide an evaluation program in the `csc_evaluation` file. The usage is as follows:

```
python csc_evaluation.py --gold_file="gold.txt" --modelout_file="modelout.txt"
```

The Python version used is 3.*.

- --gold_file: The file path for the test set in the format "**input**`\t`**gold**`\n`". The two columns of text have equal lengths.
- --modelout_file: The file path for the model output in the format "**input**`\t`**output**`\n`". The two columns of text have equal lengths.

The data in the first column of the gold_file and modelout_file should be consistent.

Output:

- Sentence-level false positive rate (FPR),
- Character-level:
  - error detection precision, recall, F1
  - error correction precision, recall, F1.

6 Testing Instructions and Submission

The test dataset includes one original sentence per line. The distribution of the test dataset is similar to the development dataset. The test dataset will be released via email before May 21st, 2023. Kindly ensure to check the email address that you used during registration for the dataset.

For submission, the following materials should be packaged as one `zip` file and sent to xjyin@pku.edu.cn.

- Submission File: Please write the final results into a text file, encoded in utf-8 format. Each line of the file should consist of a sentence pair, with the original sentence and the model output separated by a tab character: **Input**`\t`**Output**. Specifically, the submission file must contain the **same number of lines** as the input file.

- Code: The code folder should contain all the codes of data augmentation, data processing, model training and model inference.

- Document:

  - Data Description: The document needs to contain a brief description of data used in the experiment, as well as the data augmentation methods.
  - Sharing Link of Additional Data: Additional data used in the experiment should be uploaded to a cloud storage, i.e., net disk, and the sharing link should be included in the document.
  - Code Reproduction Process: The submitted code may be checked and reproduced by us, so please briefly explain the process of reproducing the code in the document.

If you have any questions, please contact xjyin@pku.edu.cn via email.