

Error-Robust Retrieval for Chinese Spelling Check

Xunjian Yin, Xinyu Hu, Jin Jiang, Xiaojun Wan

Peking University

LREC-COLING 2024

Chinese Spelling Check

- Spelling errors in Chinese texts often occur between characters with similar pronunciations and morphologies.
- The purpose of Chinese Spelling Check (CSC) is to detect and correct spelling errors in Chinese texts.
- Challenge:
 - insufficient annotated data
 - underutilization of existing datasets
- How about using Retrieval-augmented methods?

Error Pair in	Input	因为设(set)是校长的工作。
	Output (correct)	因为这(this)是校长的工作。
	Output (model)	因为涉(related)是校长的工作。
	Translation	Because this is the principal's job.
Training Set	Samples in training set	我以为设(这)是她主演的电影。 我们设(这)个周末见面。 大家认为设(这)是正常的。...
	Correct Usage in	Input
Training Set	Output (correct)	旁边的人偷(steal)了我的手册。
	Output (model)	旁边的人投(cast)了我的手册。
	Translation	Someone nearby stole my manual.
Training Set	Samples in training set	有人偷了我的钱包。 有个女生偷了我的东西。 店里的珠宝都被人偷了。...

Table 1: Examples of Chinese spelling errors, including inputs, correct outputs, outputs from model REALISE, and related samples in the training set.

Retrieval-augmented Text Generation

- A new paradigm known as "open-book exam".
- Specially, algorithms based on KNN retrieval always predict tokens with a nearest neighbor classifier over a large datastore of cached examples.

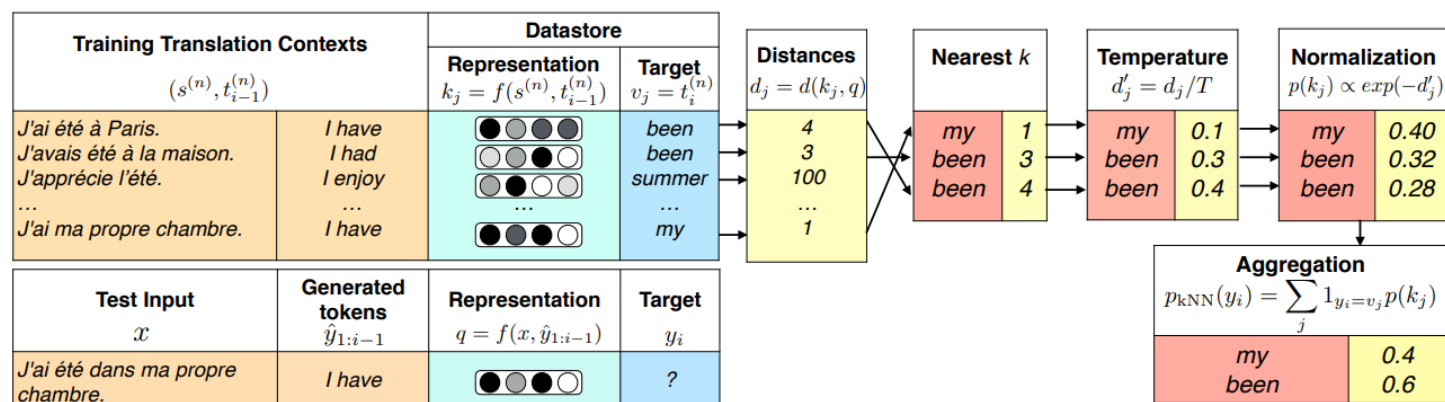


Figure 1: An illustration of how the k NN distribution is computed. The datastore, which is constructed offline, consists of representations of training set translation contexts and corresponding target tokens for every example in the parallel data. During generation, the query representation, conditioned on the test input as well as previously generated tokens, is used to retrieve the k nearest neighbors from the datastore, along with the corresponding target tokens. The distance from the query is used to compute a distribution over the retrieved targets after applying a softmax temperature. This distribution is the final k NN distribution.

Nearest Neighbor Machine Translation, Urvashi Khandelwal ICLR 2021

Challenge and Solution

- Challenge: Both correct and incorrect tokens exist in the input text, which makes it confusing and unreasonable to arbitrarily store the traditional semantic representations of each token for retrieval.
- As mentioned before, incorrect tokens, namely spelling errors, are often caused by phonetic and morphologic similarity. So we incorporate the phonetic and morphologic information of each token itself into the calculation of the query and key. (Error Robust Information)

RERIC: Retrieval method with Error-Robust Information for Chinese Spelling Check

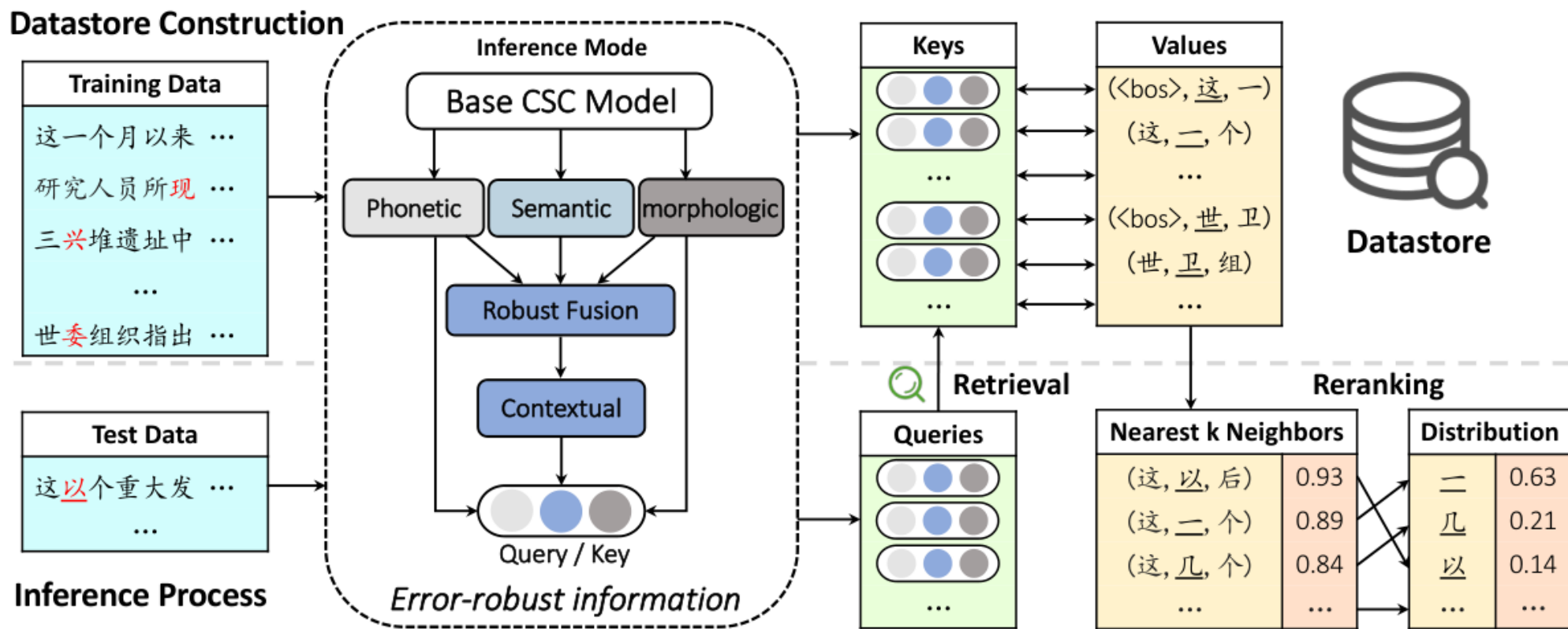
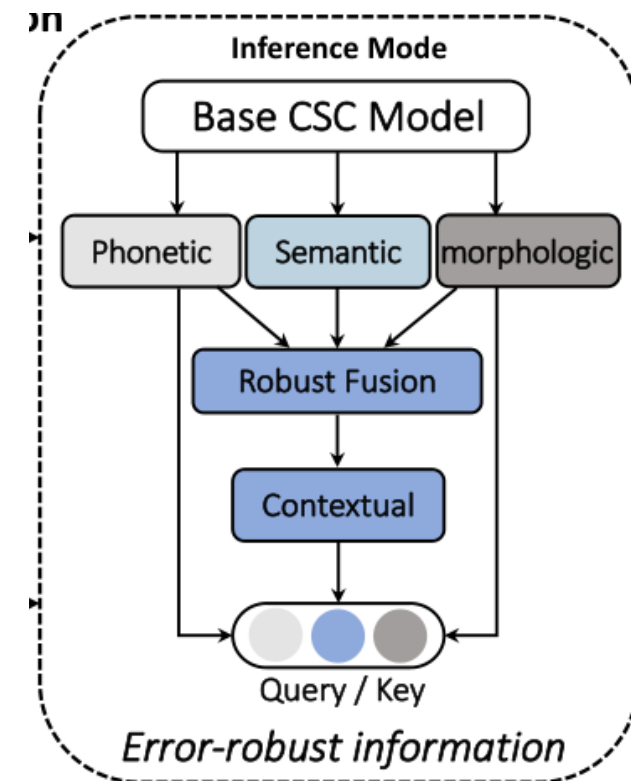


Figure 1: An illustration of our RERIC method with the datastore construction and the inference process including the KNN retrieval and reranking. The key contains the phonetic, morphologic, and contextual information of the token obtained from the base CSC model, and the value is in the form of 3-gram here. There are both correct (the majority) and incorrect tokens (marked in red) in the training and test data. Moreover, the target token and corresponding positions in n-gram values are underlined. And the test sample shows the correction process for the token "以(to)", which should be corrected to "一(one)".

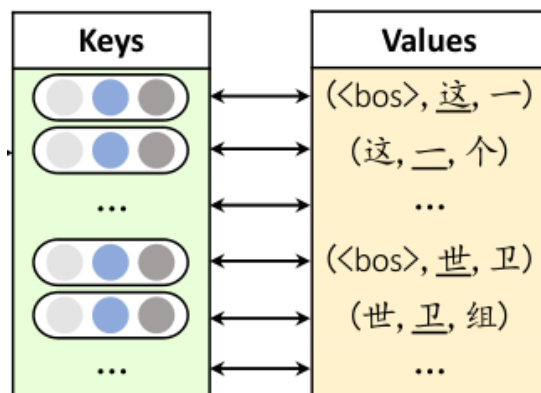
RERIC: Datastore Construction

- Key Design:
 - The goal of our key design is to alleviate the negative impact of mixing correct and incorrect tokens in the input and provide sufficient information for error correction. Each target token needs to be represented more rationally and robustly in the same high-dimensional space.
 - We use concatenation to combine and store three parts of information.



RERIC: Datastore Construction

- Value Design:
 - we propose to extend the single target token to the n-gram around it as the value for further matching and reranking.



RERIC: Retrieval and Reranking

- Retrieval:

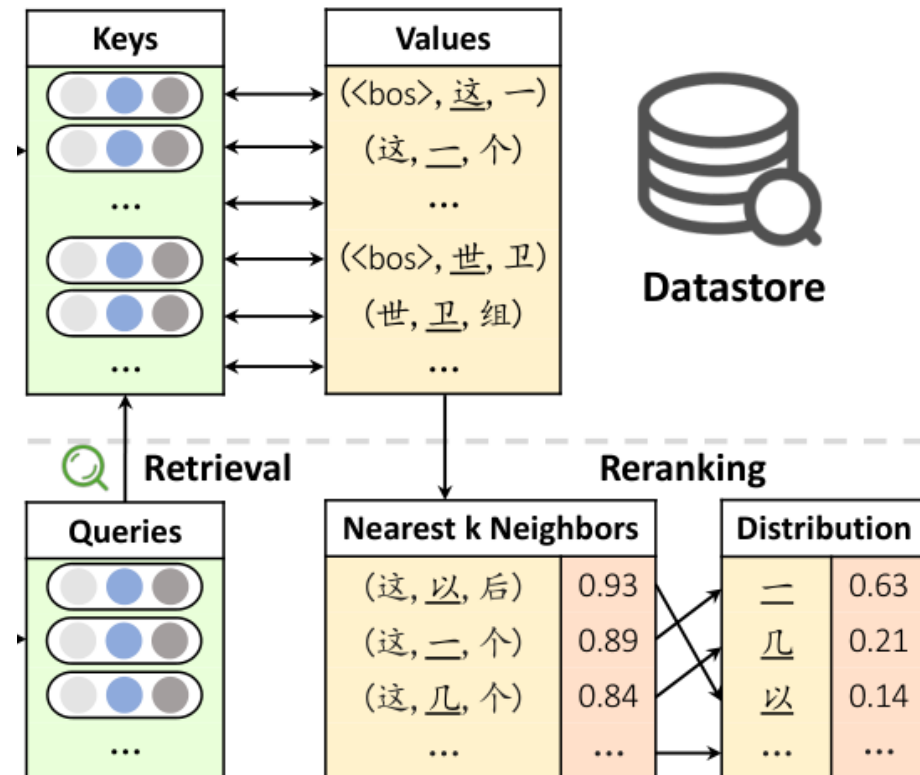
- the query is used to retrieve the k-nearest neighbors in the datastore we constructed before based on the measure of similarity.

- Reranking:

- We calculate the modified distance of the retrieved neighbor.

$$\alpha_t^j = \frac{\sum_{1 \leq i \leq n} \mathbb{I}(v^j(i), g_t(i)) w_i}{n},$$

$$d_t^j = (1 - \alpha_t^j) d(q_t, k^j),$$



Experimental Settings

- Training Data:
 - SIGHAN 13/14/15 training set
 - Wang271K
- Test Data:
 - SIGHAN 13/14/15 test set
- Evaluation Methods:
 - sentence-level metrics at both the detection level and the correction level

Results

Test Set	Model	Detection Level				Correction Level			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
SIGHAN13	FASpell (Hong et al., 2019)	63.1	76.2	63.2	69.1	60.5	73.1	60.5	66.2
	SpellGCN (Cheng et al., 2020)	-	80.1	74.4	77.2	-	78.3	72.7	75.4
	ECOPO [†] (Li et al., 2022b)	83.3	89.3	83.2	86.2	82.1	88.5	82.0	85.1
	REALISE [†] (Xu et al., 2021)	82.7	88.6	82.5	85.4	81.4	87.2	81.2	84.1
	RERIC [†] (Ours)	83.0	89.7	82.8	86.1	82.1	88.7	81.9	85.2
SIGHAN14	FASpell (Hong et al., 2019)	70.0	61.0	53.5	57.0	69.3	59.4	52.0	55.4
	SpellGCN (Cheng et al., 2020)	-	65.1	69.5	67.2	-	63.1	67.2	65.3
	ECOPO (Li et al., 2022b)	79.0	68.8	72.1	70.4	78.5	67.5	71.0	69.2
	REALISE (Xu et al., 2021)	78.4	67.8	71.5	69.6	77.7	66.3	70.0	68.1
	RERIC (Ours)	79.9	72.1	70.6	71.3	79.6	71.3	69.8	70.6
SIGHAN15	FASpell (Hong et al., 2019)	74.2	67.6	60.0	63.5	73.7	66.6	59.1	62.6
	SpellGCN (Cheng et al., 2020)	-	74.8	80.7	77.7	-	72.1	77.7	75.9
	PLOME (Liu et al., 2021)	-	77.4	81.5	79.4	-	75.3	79.3	77.2
	ECOPO (Li et al., 2022b)	85.0	77.5	82.6	80.0	84.2	76.1	81.2	78.5
	REALISE (Xu et al., 2021)	84.7	77.3	81.3	79.3	84.0	75.9	79.9	77.8
RERIC (Ours)	86.1	81.1	81.3	81.2	85.6	79.9	80.1	80.0	

Table 4: Sentence-level performance of our RERIC method and baseline models. REALISE is the backbone and base CSC model for RERIC to build the datastore. Results marked with "†" on SIGHAN 2013 are post-processed with removing all "的", "地", "得" from the model output, due to the low annotation quality about them, which is to follow the previous work (Xu et al., 2021) for convenient comparison.

Discussions with ChatGPT

- ChatGPT is not adept at performing tasks like CSC that strictly restrict the output format, and thus there are many over-correction problems.

Method	Detection Level			Correction Level		
	Pre	Rec	F1	Pre	Rec	F1
ChatGPT	36.8	79.4	50.3	26.5	57.2	36.2
RERIC	81.1	81.3	81.2	79.9	80.1	80.0

Table 6: Results of ChatGPT on the SIGHAN15 test set with the few-shot setting.

Ablation Experiments

- Ablation Experiments indicate that all three types of representations in ERI are critical, especially the phonetic information.
- When ERI or NVR are removed, the performance of the model drops significantly at both detection and correction levels.

Method	Detection Level			Correction Level		
	Pre	Rec	F1	Pre	Rec	F1
RERIC	81.1	81.3	81.2	79.9	80.1	80.0
w/o ERI-P	79.3	80.9	80.1	78.3	79.4	78.9
w/o ERI-M	79.5	81.0	80.2	78.7	79.5	79.1
w/o ERI-C	80.4	81.1	80.7	79.2	80.2	79.7
w/ ERI-S	74.6	78.4	76.5	72.0	75.7	73.8
w/o ERI	77.7	81.3	79.5	76.5	80.0	78.2
w/o NVR	79.7	81.2	80.4	78.4	79.9	79.1
w/o Retrieval	77.3	81.3	79.3	75.9	79.9	77.8

Table 5: Ablation results of our RERIC method on SIGHAN2015 test set. We apply the following changes: 1) removing each component of ERI (w/o ERI-P, w/o ERI-M, and w/o ERI-C denote the reduction of phonetic, morphologic, and contextual information, respectively); 2) using traditional semantic representation as the component of ERI (w/ ERI-S); 3) using the standard hidden representation of the token as the key (w/o ERI); 4) removing the reranking process and only using the single token as the value (w/o NVR).

Case Study

- One example is "我带上运动鞋出门(I take my sneakers and go out.)", which is correct, but REALISE incorrectly changed the "带(take)" in it to "戴(wear)" which is usually used in Chinese to refer to putting on a hat, glasses, etc. And RERIC does not make this mistake, because there are many similar uses of "带(take)" in the training set.

Input:	老师进教 师 来了。
Correct:	老师就进教 室 来了。
Translation:	The teacher came into the classroom。
CSC Output:	老师就 请 教 室 来了。
RERIC Output:	老师就进教 室 来了。
Traing Sample:	当老师的第一个脚步踏进教室时...

Input:	我 带 上运动鞋出门。
Correct:	我 带 上运动鞋出门。
Translation:	I take my sneakers and go out。
CSC Output:	我 戴 上运动鞋出门。
RERIC Output:	我 带 上运动鞋出门。
Traing Sample:	...带上半亿珠宝现身北京。 ...被老师带上街头。

Table 8: Some examples from SIGHAN 2015. The word in **red** means an error, and the word in **green** means correct. "CSC Output" means the prediction from standard REALISE model.

Conclusion

- In this paper, we propose RERIC to improve the current CSC model with our retrieval and reranking method.
- The key and value in the datastore for retrieval are designed according to the characteristics of CSC to effectively make use of the training data. More importantly, we employ multimodal representation that fuses phonetic, morphologic, and contextual information, together with n-gram matching and reranking, to improve error robustness during retrieval.
- The experimental results and relevant analyses prove the effectiveness of our method and its improvement over previous studies.
- Furthermore, our method can be simply applied in a plug-and-play manner without additional training, which shows superiority.

THANKS
