MCSCSet: A Specialist-annotated Dataset for Medical-domain Chinese Spelling Correction

Wangjie Jiang[†*], Zhihao Ye[‡*], Zijing Ou[♠], Ruihui Zhao[‡], Jianguang Zheng[‡], Yi Liu[‡],
Siheng Li[†], Bang Liu[♣], Yujiu Yang[†] and Yefeng Zheng[‡]

[†]Tsinghua University, [‡]Tencent Jarvis Lab, [♠]Sun Yat-sen University, [♣]Université de Montréal Mila & CIFAR
{jwj20,lisiheng21}@mails.tsinghua.edu.cn,{evanzhye,zacharyzhao,jaxzheng,yefengzheng}@tencent.com
{zijingou.mail,97liuyi}@gmail.com,bang.liu@umontreal.ca,yang.yujiu@sz.tsinghua.edu.cn

ABSTRACT

Chinese Spelling Correction (CSC) is gaining increasing attention due to its promise of automatically detecting and correcting spelling errors in Chinese texts. Despite its extensive use in many applications, like search engines and optical character recognition systems, little has been explored in medical scenarios in which complex and uncommon medical entities are easily misspelled. Correcting the misspellings of medical entities is arguably more difficult than those in the open domain due to its requirements of specific domain knowledge. In this work, we define the task of Medical-domain Chinese Spelling Correction and propose *MCSCSet*, a large-scale specialist-annotated dataset that contains about *200k* samples. In contrast to the existing open-domain CSC datasets, *MCSCSet* involves: *i)* extensive real-world medical queries collected from Tencent Yidian, *ii)* corresponding misspelled sentences manually annotated by medical specialists. To ensure automated dataset curation, *MCSCSet* further offers a medical confusion set consisting of the commonly misspelled characters of given Chinese medical terms. This enables one to create the medical misspelling dataset automatically. Extensive empirical studies have shown significant performance gaps between the open-domain and medical-domain spelling correction, highlighting the need to develop high-quality datasets that allow for Chinese spelling correction in specific domains. Moreover, our work benchmarks several representative Chinese spelling correction models, establishing baselines for future work.

1 INTRODUCTION

Misspelled characters frequently occur in hand-crafted Chinese sentences, easily leading to a wrong understanding of these sentences. To this end, we need a corrector to automatically detect and correct spelling mistakes in the text. The task of Chinese Spelling Correction (CSC) is to design such a corrector to correct spelling errors, which plays a vital role in various Natural Language Processing (NLP) applications such as search engine [19] and optical character recognition system [1]. To achieve the goal of efficient error correction, previous work has mainly focused on designing advanced error correction models [3, 11, 36, 40] and establishing canonical benchmark spelling correction corpora [17, 25, 30, 37]. For example, a well-known open-domain spelling correction corpus, *SIGHAN-15* [25], is a Chinese spelling correction corpus collected from a computer-based Test of Chinese as a Foreign Language (TOCFL). Although these models and benchmark datasets provide people with high-quality spelling error correction services in the open

domain, their effectivenss is reduced significantly in some specific domains, such as the medical domain. The reason is that open-domain corpora do not contain complex medical terms, and the spelling of medical terms requires specialized domain knowledge that ordinary people usually lack [21, 42].

Additionally, Chinese spelling correction for medical terms plays a crucial role in promoting the standardization and healthy development of the medical field [43]. Indeed, Chinese spelling correctors may improve the quality of medical application services, especially medical entity search systems, by automatically correcting medical terms with misspellings. Specifically, we get incorrect answers when using the medical entity query system to query medical terms with misspellings, leading to user misunderstandings and even severe medical malpractice. For example, when a doctor's hand-crafted electronic medical record contains misspellings for a malignant disease, a patient queries the term and may get results that misidentify themselves as having another illness, delaying patient care, and affecting a healthy doctor-patient relationship. This indicates that the spelling error in the medical field, especially in the medical entity query scenario, need to be corrected and resolved urgently [35]. Therefore, we need to find an effective way to correct spelling mistakes in the medical domain.

To achieve this goal, a straightforward method is to directly apply advanced methods [10, 12, 17, 33, 39] in open-domain CSC to medical-domain CSC. However, such a method is likely to fail on the medical CSC task due to the offset of the corresponding domain knowledge. To verify this, we choose an advanced BERT-based CSC model [15], which is first pre-trained on large-scale automatically-generated CSC data and then fine-tuned on *SIGHAN-15*. Then we validate the model on the test sets of *SIGHAN-15* and our proposed medical-domain dataset in this paper. The experimental results are shown in Table 1, and it can be seen that such a naive method shows a significant performance gap between in-domain and out-of-domain experiments. We conjecture that this is because the distribution of spelling errors differs significantly between an open domain and a specific domain. For instance, in Chinese medical texts, the vast majority of spelling errors occur in those complex and uncommon medical entities, which rarely occur in the open-domain Chinese texts, *e.g.*, *SIGHAN-15*, which is collected from TOCFL. In particular, we summarize the errors of medical terms into five categories of which the phonological errors and the visual errors belong to spelling errors, and show their corresponding examples in Table 2. We can observe from the table that the errors in the medical domain are not common in the open domain, which highlights the need to develop high-quality datasets that allow for medical-domain Chinese spelling correction.

---

[*]Equal contributions.

Table 1: Performance of a well-trained open-domain BERT-based CSC model on detection-level and correction-level tasks. Specifically, the model is first pre-trained on large-scale automatically-generated data [15] and then fine-tuned on *SIGHAN-15* [25] . We report the model's performances on test sets of *SIGHAN-15* and the proposed *MCSCSet*, respectively.

Test Set	Detection-level			Correction-level		
	Prec. (%)	Rec. (%)	F1 (%)	Prec. (%)	Rec. (%)	F1 (%)
SIGHAN-15	79.06	83.73	81.33	77.31	81.89	79.53
MCSCSet	43.83	38.94	41.24	28.58	25.38	26.89

Table 2: Examples of typical Chinese medical entity errors, which can be mainly divided into five categories: i) phonological errors, ii) visual errors, iii) order-confused errors; iv) repeated characters, and v) missing characters. Among the five categories, phonological and visual errors belong to spelling errors, which are the focus of our study. Erroneous characters are marked in red, and the corresponding phonics are given in brackets.

Type	Sentence	Correction
Phonological	如何闭(bi)孕 How to close pregnancy	避(bi) Contraception
Visual	拨智齿的过程 Wisdom tooth dial process	拔 Extraction
Order-confused	胃经常痛胀 Stomach is often bloated	胀痛 Pain
Repeated	头头疼的原因 Causes of head headaches	头疼 Headaches
Missing	血板高怎么降下来 How to reduce the high blood plate	血小板 Platelet

Here, we highlight the challenges of building a large-scale Chinese spelling correction benchmark dataset in the medical domain as follows:

(C1) Difficulty to Collect Real Data: To be able to provide the service of medical entity error correction in real application scenarios, annotated datasets must come from real medical scenarios and contain common error-prone medical entities among the hundreds of millions of queries generated by real-world applications.

(C2) High Demand of Medical Knowledge: To produce a high-quality medical term (or entity) spelling correction corpus, annotators are required to master specific medical knowledge and maintain high correction quality, which is a challenging and time-consuming task.

To address the above challenges, in this paper, we present *Medical Chinese Spelling Correction Dataset (MCSCSet)*, a large-scale and specialist-annotated dataset for Chinese spelling correction in the medical domain. Notably, we collect a large-scale query log dataset from a real-world medical application named Tencent Yidian[1] and construct a manually annotated dataset with about *200k* samples, in which each sample consists of a correct medical query and its corresponding wrong medical query with spelling errors. *MCSCSet* also provides a medical confusion set, consisting of a large number of error-prone characters from Chinese medical terminologies, each with its corresponding erroneous characters. This enables potential

[1]https://baike.qq.com/

researchers or practitioners to generate new medical-domain CSC datasets based on their specific needs by simply replacing the medical entities with misspelled characters defined in the confusion set. To distinguish from the open-domain CSC, we further provide a formal definition of the medical-domain Chinese spelling correction task, mainly focusing on the spelling error correction for medical entities. Moreover, our work benchmarks several Chinese spelling correction models for future comparisons. Overall, the following components summarize our major contributions:

- **Practical Task Definition of Medical-domain Chinese Spelling Correction:** We formally define the Chinese spelling correction task in the medical domain for the first time, which applies to all tasks involving user input such as search, question answering, and translation.
- **First CSC Dataset for Medical Domain:** We provide the first Chinese medical spelling correction dataset from the large-scale healthcare encyclopedia software Tencent Yidian, based on the annotation of medical specialists.
- **Rich Medical Confusion Set:** We present a corresponding medical confusion set, which consists of abundant error-prone medical entities. This allows great flexibility for future usage since one could exploit it to construct a new dataset.
- **Rigorous Medical-domain CSC Benchmarking:** We benchmark four representative Chinese spelling correction models, which verify the quality of the proposed *MCSCSet* dataset and provide reproducible comparisons for future studies.

Paper Organizations. Section 2 presents background and related work on Chinese spelling correction, including previous CSC algorithms, datasets, and benchmarks. In Section 3, we present the definition of the problem of medical-domain Chinese spelling correction. In Section 4 we provide details on the construction process of the *MCSCSet* dataset and present some statistical analysis. Section 5 provides specifics of benchmarking representative CSC algorithms, implementation details and experimental results. Lastly, Section 6 discusses and concludes the paper.

2 RELATED WORK

Chinese Spelling Correction. Chinese Spelling Correction (CSC) is a challenging task in Natural Language Processing (NLP) and plays an important part in various real-world applications, such as search engine [7, 8, 19], optical character recognition [6, 20, 27] and automatic speech recognition [2, 9, 34, 41]. CSC is similar to the task of Chinese Grammatical Error Correction (CGEC) [16]. The difference between them is that CSC only focuses on Chinese spelling errors that need replacement, while CGEC also includes errors that require deletion and insertion.

Table 3: The comparison of *MCSCSet* and existing CSC datasets.

Dataset	# Sentences	Avg. length	# Error types	Domain	Annotation
SIGHAN-13 [31]	1,700	60.9	2	Open	Human-annotation
SIGHAN-14 [37]	4,463	49.7	2	Open	Human-annotation
SIGHAN-15 [25]	3,438	31.1	2	Open	Human-annotation
OCRSet [11]	4,575	10.2	2	Open	Automatically-generated
HybridSet [27]	271,329	42.5	2	Open	Automatically-generated
MCSCSet	**199,877**	**10.9**	**5**	**Medical**	**Specialist-annotation**

Table 4: A training sample of Chinese medical query spelling correction task. Here, the misspelled medical entity "拨智尺" (dial the wisdom ruler) should be corrected to "拔智齿" (extract the wisdom tooth).

Wrong query: 拨智尺的过程。
Translation: The process of dial the wisdom ruler.

Output query: 拔智齿的过程。
Translation: The process of wisdom tooth extraction.

Existing Methods for CSC. CSC has been an active research topic for many years. Early works followed error detection, correction, candidate generation, and candidate selection. Most of the proposed correction methods [18, 32, 37] employed unsupervised n-gram language models and confusion set to detect errors and select candidate replacement characters to fix the errors. With the development of deep learning and especially pre-trained language models (PLM) [23] in recent years, great progress has been achieved in the CSC task. Since the pre-trained task (*i.e.*, masked language model) of BERT is similar to CSC, many approaches utilized BERT [5] in their models. Hong et al. [11] proposed the FASPell model based on BERT and exploited the phonological and visual similarity information to filter candidate characters. Cheng et al. [3] proposed to incorporate the character similarity knowledge into BERT via a Graph Convolution Network (GCN) [14]. Zhang et al. [40] proposed the soft-masked BERT model, in which Gated Recurrent Unit (GRU) [4] was used to detect the erroneous positions and BERT was used to predict correct characters. These methods achieved excellent results in *SIGHAN* [26, 31, 38], a series of well-known benchmark datasets in open-domain CSC. More recently, some studies [10, 12, 13, 15, 39] focused on treating CSC as a pre-trained task. Specifically, they utilized the open-domain confusion set to automatically generate a large number of pairs of spelling errors and the corresponding corrections. Models are first pre-trained on the large-scale automatically-generated data and then fine-tuned on the benchmark dataset. Benefiting from the CSC pretraining, better performances have been achieved. Although numerous excellent methods have been proposed to tackle open-domain CSC, few works studied CSC in a specific domain, limiting CSC's wide application.

Existing Datasets for CSC. There are several public datasets for CSC. Besides the above-mentioned *SIGHAN* series (including *SIGHAN-13* [31], *SIGHAN-14* [37] and *SIGHAN-15* [25]), *OCRSet* [11] and *HybridSet* [27] are also commonly used. Table 3 shows statistics of the existing CSC datasets and our proposed *MCSCSet*. We can find

that existing public CSC datasets suffer from either limited scale or low quality. The *SIGHAN* series has high quality due to human-annotation but poor in scale. As for *HybridSet*, although the dataset scale is large, the quality is relatively poor because automatically-generated data cannot exactly simulate the distribution of real-world spelling errors. Besides, these datasets are all collected from open-domain texts and cannot be used in a specific domain such as medicine.

Discussion. To tackle the above issues, we take the first step to study CSC in the medical domain. Specifically, we focus on a real-world medical scenario, *i.e.*, medical search engine, and propose the task of medical-domain Chinese spelling correction. To support this task, we construct a high-quality and large-scale dataset by specialist annotation.

3 TASK DEFINITION

In this section, we introduce our proposed medical-domain Chinese spelling correction, which is a new task on top of open-domain CSC. Specifically, the formal definition is provided in Definition 3.1 and a detailed discussion of the comparison between medical- and open-domain CSC follows immediately after.

Definition 3.1 (Medical-domain Chinese Spelling Correction). Given a Chinese medical query $X = \{x_1, x_2, \ldots, x_n\}$ with n characters, the goal of the task is to detect spelling errors on character level and output its corresponding correct sequence $Y = \{y_1, y_2, \ldots, y_m\}$. Following the general spelling correction task [24], the lengths of X and Y are assumed to be equal here, i.e., n = m. We model the task as a conditional generation problem by modeling and maximizing the log likelihood of probability $p_\theta(Y \mid X)$ with parameter θ. Thus, the training objective can be formulated as follows:

$$\underset{\theta}{\arg\max} \sum_{(X,Y) \sim \mathcal{D}} \log p_\theta(Y \mid X). \tag{1}$$

Where the data point (X, Y) is uniformly sampled from the dataset $\mathcal{D}$, X contains one or more medical entities that include misspelled characters, while the corresponding entities in Y are absolutely correct. Therefore, the key of medical-domain Chinese spelling correction is to detect and correct misspelled characters in medical entities.

To help better understand this task, we give an example in Table 4. It can be seen that the input query "拨智尺的过程" (*i.e.*, "the process of dialing the wisdom ruler" in English) contains a medical entity "拨智尺" that includes two misspelled characters "拨" (dial) and "尺" (ruler). Our task aims to predict the output query "拔智齿的过程" (the process of wisdom tooth extraction), in which the two misspelled characters of the medical entity are corrected to "拔"

Figure 1: Annotation interface.

(extract) and "齿" (tooth), respectively. We can see that "拨" (dial) is visually similar to "拔" (extract) while "尺" (ruler) is phonetically similar to "齿" (tooth), which are error-prone for one without both linguistic knowledge and medical knowledge.

Compared with open-domain CSC, our task is quite different and more difficult. First, Chinese medical queries usually contain medical terms that are generally complex and uncommonly used. Second, spelling errors are more likely to occur in medical entities than other characters. Third, correcting the spelling errors in medical entities requires medical knowledge in addition to linguistic knowledge. Therefore, as shown in the Table 1, directly transferring the model trained from the open domain would significantly degenerate the performance in the medical domain. To this end, it is necessary to construct a large-scale and high-quality dataset with specialist annotation to facilitate research on our task.

4 THE MEDICAL-DOMAIN CHINESE SPELLING CORRECTION DATASET

In this section, we present the medical-domain Chinese spelling correction dataset (*MCSCSet*). To help fully understand the property of *MCSCSet*, we first describe the data collection and annotation steps of medical query errors in Section 4.1 and Section 4.2, respectively. Then the detailed data format and statistics are given in Section 4.3. More details about *MCSCSet* can be found in the dataset document presented in A.

4.1 Medical Query Selection

Building a large-scale Chinese spelling correction benchmark dataset in the medical domain is challenging. As demonstrated in the Section 1, the first difficulty arises from the data collection. To ensure the practicality of medical-domain spelling corrector in real-world applications, the spelling errors should come from the real medical scenarios. To this end, we collect a large-scale query log set Q with more than *900k* samples from a real-world medical encyclopedia application named Tencent Yidian. Then, we remove queries that include personal information, such as name, ID number, and personal address. In addition, we further filter out queries with too long (more than 50 Chinese characters) or too short (less than 3), and queries without medical entities. Finally, we select about

---

Algorithm 1: A high level description of the annotation procedure.

Input: A large scale query log set Q with more than $900,000$ samples.
Output: A medical query dataset $\mathcal{D}$ with $|\mathcal{D}|$ sample pairs.

1: Clean the query set (e.g., removing queries with sensitive personal information and without medical entities, etc.)
2: **for** $t = 1 \rightarrow |\mathcal{D}|$ **do**
3:     Initialize the flag $F \leftarrow False$
4:     Expert annotators judge whether the raw sentence includes wrong entities and if so, assign flag $F \leftarrow True$, and correct the misspelled entities.
5:     **if** $F == False$ **then**
6:         Expert annotators annotate the corresponding erroneous entities.
7:     **end if**
8:     Expert annotators mark the error types and indexes of wrong medical entities, resulting the t-th annotated sample.
9: **end for**
10: **return**: The medical-domain Chinese spelling correction dataset (*MCSCSet*).

200k queries containing common error-prone medical entities as a dataset to be annotated.

4.2 Annotation Process

After going through medical query selection and getting a dataset to be annotated, we still have a second challenge to overcome. That is the annotation of misspelled medical entities requires the annotators to equip themselves with well-educated medical knowledge. In this regard, we hire annotators with medical backgrounds, such as medical students and hospital staff, to annotate medical entities of query and form a medical query error dataset $\mathcal{D}$ with $|\mathcal{D}| \approx 200,000$ samples. Figure 1 shows the user interface of our annotation process, and the annotation steps are as follows:

- **Step 1:** First, expert annotators find the medical entity in the query. Specifically, we first pre-label the medical entity with our medical entity recognition algorithm, and then the annotator needs to check the correctness of identified entities. After checking and correcting the entities identified by the algorithm, the annotator needs to mark the medical entities in the query with {} to annotate the location.
- **Step 2:** Expert annotators check whether the query includes a wrong medical entity and correct the error. Expert annotators also mark the medical entity's error type. It should be noted that common words other than medical entities are not considered.
- **Step 3:** Suppose the raw query does not include a wrong entity. Expert annotators replace the correct medical entity with an erroneous entity, and mark the corresponding error type.
- **Step 4:** Finally, The annotator puts the corresponding wrong query, correct query, and error type together to form an annotated sample.

Table 5: A sample from the *MCSCSet*.

Field	Content
Wrong query	拔智尺的过程。
Correct query	拔智齿的过程。
Error locations	(0, 2)
Error types	(visual, phonological)
Medical entity locations	([0, 2])

Table 6: Detailed statistics of *MCSCSet*.

# Queries	199,877
Avg. query length	10.90
Avg. # misspelled characters per query	1.86
Avg. # medical entities per query	1.46
# Unique medical entities	81,020
# Training samples	157,194
# Validation samples	19,652
# Test samples	19,650

The above annotation process is also formally summarized in algorithm 1.

4.3 Data Format and Statistics

Schema of *MCSCSet*: The schema of *MCSCSet* is shown in Table 5. Each sample includes five fields: wrong query, correct query, error location, error type, medical entity locations. Wrong query and correct query serve as input and output respectively in the medical-domain CSC task. In addition, exact error locations (*i.e.*, indexes of wrong characters in the sequence) and error types are attached for each sample. They can provide rich information about spelling errors. Medical entities in queries are also marked because spelling errors are more likely to occur.

Statistics of *MCSCSet*: In Table 6, we show the statistics of the *MCSCSet*. We find the average query length of 10.9 Chinese characters is short, and there are 1.86 wrong characters per query, which may reduce the difficulty of spelling error correction. We attribute this to the inherent property of search queries because people tend to input relatively short text. The specific length distribution is demonstrated in Figure 2 (a), and we can see that most medical queries contain less than twenty characters. Besides, there are an average of 1.46 medical entities in each query, and the total number of unique medical entities is up to 81, 020. Figure 2 (b) shows the distribution of medical entities' frequency, and we can see that most of the medical entities appear less than or equal to five times in our dataset. Figure 2 (c) shows the distribution of entity error types. From the figure, we can find that the error types of visual and phonological account for the highest proportion, about 96% in total. Besides, our dataset also contains error types of order-confused, repeated, and missing. In other words, our dataset can also be used for syntax error correction.

Medical Confusion Set: On top of the *MCSCSet*, we further construct a dictionary-like medical confusion set. Given a medical-domain error-prone character, we can easily find its corresponding

Table 7: Several examples in our medical confusion set.

Character	Corresponding misspelled characters
疣	犹, 由, 庞, 忧, 莜, 旁, 优, 疮, 尤
酮	醇, 侗, 痌, 桐, 同, 彤, 铜, 筒, 通, 桶,
椎	谁, 柱, 锥, 追, 锤, 坠, 垂
氨	铵, 安, 胺, 按
呤	令, 铃

Table 8: Top 5 high-frequency entities.

Entities	Frequency	Misspelled entities
怀孕 (pregnancy)	3,799	(坏孕,还孕,淮孕)
发烧 (fever)	2,483	(发稍,发梢,花烧)
咳嗽 (cough)	2,060	(核嗽,咳嗖,咳嗦)
糖尿病 (diabetes)	1,890	(唐尿病,糖尿崩,糖尿症)
感冒 (cold)	1,863	(感昌,感帽,赶冒)

common erroneous characters according to the medical confusion set. There are several public confusion sets in open-domain CSC, but they cannot be well applied in a specific domain like medicine. Therefore, we present the medical confusion set for the first time. Specifically, we collect all the spelling errors in the *MCSCSet* and obtain abundant correct-erroneous character pairs. We find 2, 623 different characters are misspelled in our dataset by statistical analysis. Most of the characters have one to twenty corresponding erroneous characters. Several examples are shown in Table 7. Additionally, we analyze high-frequency medical entities of all 81,020 that appear in our dataset, which means these entities are more prone to misspellings in real-world medical scenarios. As shown in Table 8, we can find the top 5 high-frequency entities and the corresponding wrong entity sets from our medical confusion set.

5 EXPERIMENTS

In this section, we conduct a series of studies on our *MCSCSet*. First, we benchmark some typical CSC models on the *MCSCSet* to establish baselines for future research. Then, we demonstrate the superiority of our dataset by comparing it with the automatically-generated dataset. Next, we show the effectiveness of our medical confusion set by comparing it with the open-domain confusion set. Finally, we investigate how the medical query's average number of misspelled characters affects the model performance.

5.1 Experiment Setup

Experiment Configurations: (i) Problem Revisiting. Chinese medical query spelling correction is to correct the misspelled characters in a medical query. In the experiment, the input is a medical query that possibly contains misspelled characters, and the output is a sentence with the same length as the input. For each character in the input sentence, if it is correct, the character in the same position in the output sentence should be unchanged; if it is misspelled, the character in the same position in the output sentence should be the corresponding correct character. **(ii) Dataset Filtering and Splitting.** There are five different types of errors in our dataset.
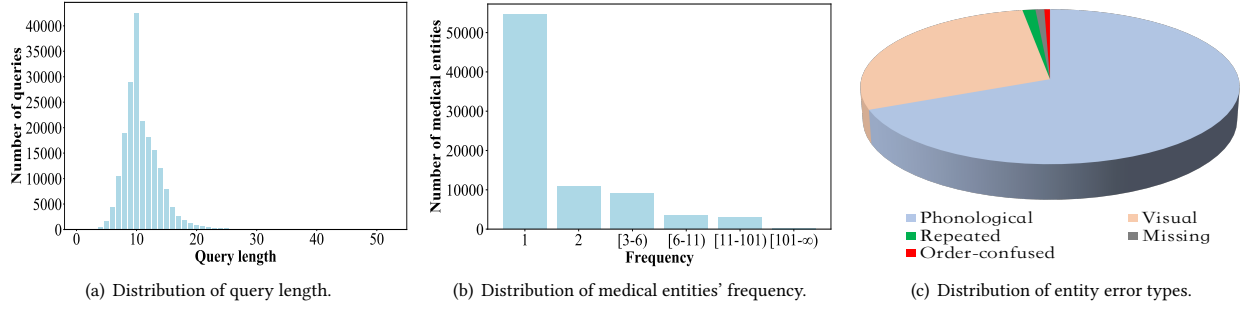
Wangjie Jiang and Zhihao Ye, et al.

(a) Distribution of query length.

(b) Distribution of medical entities' frequency.

(c) Distribution of entity error types.

Figure 2: Visualization of the statistics for *MCSCSet*.

Since we focus on the task of Chinese medical query spelling correction, we filtered out the medical queries containing errors other than spelling errors and obtained the remaining 196,496 samples. For simplicity, we still call the remaining data *MCSCSet*. Finally, we split *MCSCSet* into a training set (157,194), a validation set (19,652), and a test set (19,650) with a ratio of 8:1:1. **(iii) Samples that Need no Correction.** In real-world applications, only a small part of medical queries contains spelling errors, while all samples in *MCSCSet* include spelling errors and need correction. If we train the correction model with the original *MCSCSet*, the model would be prone to believe that every input query must have spelling errors simply. As a result, the model would likely transform initially correct medical queries into erroneous ones. Therefore, it is necessary to set a certain percentage of training samples to be in the form of correct-correct pair. Meanwhile, the percentage should not be too small if the correction model cannot be efficiently trained. In our actual practice, we set the percentage as 50%, which is in line with *SIGHAN-15*.

Evaluation Metrics: Results are reported at the detection level and the correction level. A medical query is processed correctly at the detection level if and only if all spelling errors in the query are recognized, and it can be further considered to be processed successfully at the correction level if and only if all misspelled characters are replaced with right ones. We report the precision (Prec.), recall (Rec.), and F1 scores on both levels.

5.2 Benchmark Models

The state-of-the-art methods in the *SIGHAN* benchmark are almost based on pre-training CSC tasks on a large-scale automatically-generated dataset. To help understand the existing correction methods, we show the general framework in Figure 3. As depicted in the figure, the general framework is structured as an encoder-decoder architecture, in which the encoder is usually a pre-trained language model (*e.g.*, BERT) and the decoder is typically a classifier (*e.g.*, linear layer). Some of them have not released the code, while the others have not provided a pre-training dataset or the pre-trained weights, which makes their work hard to reproduce or apply in our *MCSCSet*. Therefore, we experiment with some representative and widely used spelling error correction methods as follows:

N-gram Model [32]: This is a traditional rule-based method utilizing an n-gram language model and confusion set. To be specific, an unsupervised n-gram language model is first employed to detect

errors. Next, a confusion set, an external knowledge of the similarity between characters, is introduced to confine the candidates for correction. Finally, the best candidate with the highest n-gram language model probability is considered as the replacement character.

BERT-Correction [15]: This method treats CSC as a non-autoregressive generation task and employs BERT as the model backbone. First, the input sentence is encoded by BERT to obtain the hidden representations of each character. Then, a classifier (*e.g.*, linear layer) is used to pick the correction character from the whole vocabulary set for each character. According to our experiments, this is a simple but effective method that exceeds many methods with complicated designs.

Soft-Masked BERT [40]: This method introduces the soft-masking strategy in BERT to improve error detection performance. Concretely, *Soft-Masked BERT* first uses BERT as the encoder. Next, a Bi-GRU network is employed to detect the error probability for each character. The error probability is then used to weight the original and mask embedding to obtain the final character embedding. Last, a BERT-based network is applied to correct errors. With the error probability of character to incorporate the mask embedding, this method narrows the gap between CSC and the pretraining task of BERT (*i.e.*, masked language model), thus achieving excellent results.

MedBERT-Correction: We build this model to take advantage of medical-domain knowledge. This method is similar to the above *BERT-Correction*. The only difference is that the encoder is replaced by PCL-MedBERT, [2] which is a well-known pre-trained medical language model proposed by the Intelligent Medical Research Group at the Peng Cheng Laboratory, with excellent performance in medical question matching and named entity recognition.

Implementation Details: The *N-gram Model* is implemented with pycorrector, [3] which is a well-known toolkit for text error correction. The experiments of the other methods are conducted with Pytorch [22] and HuggingFace Transformers [29]. For all methods based on the pre-trained language model, we set different learning rates for the pre-trained language model and the subsequent classifier as this shows better performance in our experiments. We search the learning rate of the pre-trained language model in $[1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}]$, the best is 5×10^{-5}. Similarly, the

---

[2]https://code.ihub.org.cn/projects/1775

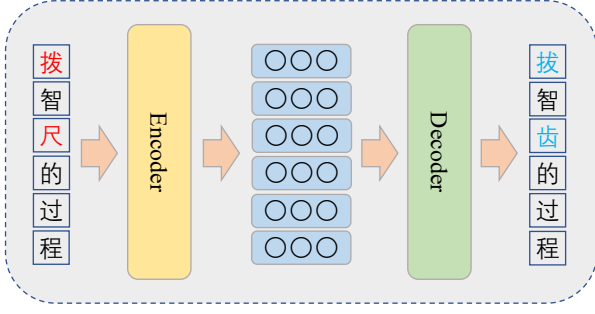[3]https://github.com/shibing624/pycorrector

Figure 3: The general framework of existing CSC models.

best learning rate of the classifier is 3×10^{-4} in $[1 \times 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}]$. The warmup and dropout ratio are both 0.1. We train the model for a total of 20 epochs with a batch size of 64, and early stopping is activated when the validation scores do not improve for 5 epochs. Especially for *Soft-Masked BERT*, we adopt the hyper-parameters proposed in [40] with the Bi-GRU, the Bi-GRU hidden unit of 256 and the balance weight λ in the objective function of 0.8. All the reported results are averaged over 5 runs with different random seeds. The experiments are conducted on NVIDIA Tesla V100 with 32GB memory.

5.3 Benchmark Experiments

Main Results: To bulid a valid benchmark and facilitate future research on our task, we experiment with the above representative CSC methods on our *MCSCSet*. Table 9 shows the main results, and we have the following observations: (1) *N-gram Model* performs unsatisfactorily on our *MCSCSet*, with the correction-level F1 score even under 25%. We attribute this phenomenon to the following aspects. First, the *N-gram Model* is trained on large-scale open-domain corpus in an unsupervised manner, which is unable to make full use of the supervision information offered by our dataset. Second, the open-domain confusion set is used to confine the candidates. However, the open-domain confusion set is not customized for the medical domain, which brings the noise to a certain extent. (2) Methods based on pre-trained language models achieve promising results on our *MCSCSet*. Specifically, they all have correction-level F1 score over 80%. The potential reasons are twofold: *i)* pre-trained language models can offer general linguistic knowledge, which is fundamental for spelling error correction. *ii) MCSCSet* provides the domain knowledge as well as correction supervision, which makes the correction models well generalizable in the medical domain. (3) *MedBERT-Correction* outperforms *BERT-Correction* by a bit of margin, which indicates the effectiveness of incorporating external medical knowledge to a certain extent. We should also notice that the effectiveness is limited for the pre-trained medical language model, which enlightens the need for other valid methods of introducing medical knowledge. Incorporating a medical knowledge graph during training probably works. (4) *Soft-Masked BERT* obtains the best results on our dataset. We attribute this to the specially-designed soft-masking strategy to narrow the gap between spelling error correction and masked language model. By

incorporating the mask embedding, the misspelled characters in medical entities are more accurately located.

In general, pre-trained language models have been proved to be highly effective in our task. They can improve the benchmark performance by effectively incorporating medical knowledge and considering more characteristics of spelling errors in the medical domain.

Comparison between *MCSCSet* and Automatically-generated Dataset: To demonstrate the necessity and great value of constructing the dataset with specialist annotation, we compare our *MCSCSet* with the automatically-generated dataset. The automatically-generated dataset is commonly used in open-domain CSC for the following reasons. First, existing open-domain CSC datasets are limited in scale incapable of training satisfactory CSC models. Second, the automatically-generated data is easy to acquire, and the dataset scale can be massive. Third, the quality of the automatically-generated data is acceptable, which can be used to pre-train the CSC models. Automatically generating a training sample (wrong-correct pair) is straightforward. According to the confusion set, the wrong query can be obtained by replacing some characters in the correct query with the corresponding erroneous characters. In this way, large-scale open-domain datasets can be generated automatically, and they are used to pre-train CSC models followed by fine-tuning on the target dataset. Although automatically-generated datasets have achieved open-domain CSC, we argue this paradigm is not suitable for CSC in the medical domain as the open-domain confusion set is inapplicable in the medical domain. To validate this, we first construct the automatically-generated dataset (*AGSet-Open*) based on the training set and validation set of *MCSCSet* with a commonly used open-domain confusion set [28]. For a fair comparison, we set the average number of misspelled characters to 1.8 to keep consistency to the original *MCSCSet*. Specifically, we train *BERT-Correction* on *AGSet-Open* and *MCSCSet* respectively and evaluate them on the same test set of *MCSCSet*. As the results shown in Table 10, we can observe that the correction-level F1 score of the model trained on *AGSet-Open* is under 40%, which is even lower than half of that on *MCSCSet*. We conjecture this is because the automatically-generated dataset has poor quality and fails to simulate the practical distribution of Chinese medical query spelling correction.

Effectiveness of the Medical Confusion Set: In this part, we compare our medical confusion set with the open-domain confusion set to illustrate the effectiveness and superiority of our medical confusion set. As mentioned in the last subsection, the confusion set can be used to generate training samples automatically. However, results show that the open-domain confusion set is not suitable for the medical domain. Compared with the open-domain confusion set, our medical confusion set is constructed on top of *MCSCSet*. Therefore, it is customized for spelling corrections in the medical domain, which is of high quality and can be used to further automatically generate training samples that are suitable for Chinese medical query spelling correction. To prove the above claims, we construct *AGSet-Med* based on *MCSCSet* and our medical confusion set in a way similar to the construction of *AGSet-Open* The only difference is we replace the original open-domain confusion set with our medical confusion set. For a fair comparison, we also set the

Table 9: Performances of benchmark models on *MCSCSet*. The best results are highlighted in bold.

Method	Detection-level			Correction-level		
	Prec. (%)	Rec. (%)	F1 (%)	Prec. (%)	Rec. (%)	F1 (%)
N-gram Model [32]	38.43	20.38	26.64	30.32	16.08	21.02
BERT-Correction [15]	**87.05**	86.08	86.55	80.93	80.05	80.49
MedBERT-Correction	87.01	86.25	86.63	80.98	80.24	80.61
Soft-Masked BERT [40]	87.03	**86.29**	**86.66**	**81.22**	**80.54**	**80.88**

Table 10: Performances of *BERT-Correction* trained on the automatically-generated datasets (*AGSet-Open, AGSet-Med*) and *MCSCSet*, respectively. Notably, *AGSet-Open* is constructed based on *MCSCSet* and open-domain confusion set [28], while *AGSet-Med* is built based on the *MCSCSet* and our medical confusion set.

Dataset	Detection-level			Correction-level		
	Prec. (%)	Rec. (%)	F1 (%)	Prec. (%)	Rec. (%)	F1 (%)
MCSCSet	87.05	86.08	86.55	80.93	80.05	80.49
AGSet-Open	62.77	52.69	57.29	43.61	36.62	39.81
AGSet-Med	72.25	61.32	66.34	62.24	55.16	60.19

average number of wrong characters as 1.86. Finally, we train *BERT-Correction* on *AGSet-Med* and report the evaluation results in Table 10. We can see that the model trained on *AGSet-Med* outperforms the one trained on *AGSet-Open*, which indicates that *AGSet-Med* has better quality than *AGSet-Open* and further demonstrates the effectiveness and superiority of our medical confusion set. The medical confusion set is of great value for medical-domain CSC. First, it is customized for medical-domain CSC and contains the misspelling feature of medical texts since it is constructed on top of *MCSCSet*. Second, the medical confusion set can be employed to help the candidate characters filtering. Specifically, once a character is detected misspelled, to determine the target corrected character from the whole vocabulary, we can set larger weights for the erroneous corresponding characters in the medical confusion set. Third, the medical confusion set can be used to automatically generate large-scale CSC data in the medical domain. Therefore, it is promising to obtain a superior correction model by first pre-training it on the generated large-scale medical-domain CSC data following fine-tuning it on the specific target dataset.

Impact of the Average Number of Misspelled Characters of the Medical Query: We study how the average number of misspelled characters affects the model performance. Intuitively, a medical query with more erroneous characters is more challenging to correct because the number of misspelled characters measures the correction difficulty to a certain extent. To obtain *MCSCSet* with the different average number of misspelled characters, we employ our medical confusion set to add erroneous characters on top of the original *MCSCSet*. Specifically, for each correct character in the medical entities of a medical query, with a certain probability, we change it to one of its corresponding erroneous characters according to the medical confusion set. We can obtain *MCSCSet* with a different average number of misspelled characters by controlling the probability. Then we train *BERT-Correction* on these datasets respectively and report their performances (F1 score) in Figure 4. As shown in the figure, the model performance deteriorates as the
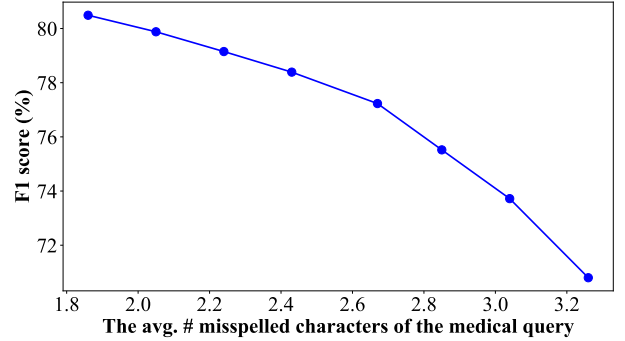
Figure 4: The influence of the average number of misspelled characters of the medical query on model performance.

average number of misspelled characters increases. Besides, when the average number of misspelled characters exceeds about 2.6, the model performance degrades relatively faster. We conjecture the reason is that after certain point, the model has less context to correct misspelled characters.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed the medical-domain Chinese spelling correction task and introduced *MCSCSet*, a large-scale dataset dedicated to this task, annotated by medical specialists. On top of *MCSCSet*, we also proposed the medical confusion set, which could facilitate the automatic data generation process. We conducted a series of experiments to demonstrate the necessity and effectiveness of *MCSCSet* and the medical confusion set. Moreover, we benchmarked several representative spelling correction methods to provide baselines for future research. As shown by our experiments, the current spelling correction algorithms still have much room for improvement in the medical domain, which means there still lies space to delve deeper for future work.

A APPENDIX

A.1 Dataset Introduction

MCSCSet is a large-scale specialist-annotated dataset for medical-domain CSC. It is constructed on top of large-scale real-world Chinese medical queries collected from Tencent Yidian, which is a well-known healthcare encyclopedia platform in China. Five different types of errors (including phonological, visual, order-confused, repeated and missing errors) are annotated in the dataset. Since spelling errors (phonological and visual errors) account for most of the errors in real application, we pick out samples that contain only spelling errors to make up the subset, which is dedicated for the task of medical-domain CSC. In addition, we provide the medical confusion set, which is constructed on top of *MCSCSet*. It consists of abundant medical-domain error-prone characters and their corresponding erroneous characters.

A.2 Dataset Statistics

The detailed statistics of *MCSCSet* shows in Table 11. Note that the training samples, validation samples and testing samples in the table consists of samples with only Chinese spelling errors. Besides, the number of unique misspelled characters is 2623, which also represents the size of the medical confusion set.

A.3 Data Content and Format

MCSCSet and the medical confusion set are stored in two JSON files ("MCSCSet.json" and "MedConfusionSet.json"). There are five fields in "MCSCSet.json", including wrong query, correct query, error locations (*i.e.*, indexes of erroneous characters), error types and medical entity locations (*i.e.*, start and end indexes of medical entities). Several examples are demonstrated in Table 12.

A.4 Data loading

The following code is provided for data loading:

```python
import json

data = []
with open("MCSCSet.json") as f:
    for line in f:
        sample = json.loads(line)
        data.append((sample["Wrong query"],
                     sample["Correct query"]))
confusionset = json.loads("MedConfusionSet.json")

return data, confusionset
```

The returned "data" is a list containing all wrong-correct pairs and "confusionset" is a dictionary containing all error-prone characters and their corresponding erroneous characters.

A.5 Application Scenarios

MCSCSet can be utilized to help construct text error correction systems in the medical domain. Further, these text error correction systems have potential to benefit a wide range of medical applications, such as medical search engine, medical dialogue system and medical named entity recognition.

Table 11: Detailed statistics of *MCSCSet*.

MCSCSet	
# Queries	199,877
Avg. query length	10.90
Avg. # error types	5
Avg. # misspelled characters per query	1.86
# Unique misspelled characters	2,623
Avg. # medical entities per query	1.4
# Unique medical entities	81,020
# Training samples	157,194
# Validation samples	19,652
# Test samples	19,650

Table 12: Several examples in "MCSCSet.json" and "MedConfusionSet.json".

Several examples in "MCSCSet.json":

```
{
    "Wrong query": "内风湿十年没复发正常吗？",
    "Correct query": "类风湿十年没复发正常吗？",
    "Error locations": (0),
    "Error types": (phonological),
    "Medical entity locations": ([0,2], [6,7])
},
{
    "Wrong query": "哪些过敏源会引起寻麻疹",
    "Correct query": "哪些过敏源会引起荨麻疹",
    "Error locations": (8),
    "Error types": (phonological & visual),
    "Medical entity locations": ([2,4], [8,10])
},
{
    "Wrong query": "一个月吃两次紧急闭孕药的危害",
    "Correct query": "一个月吃两次紧急避孕药的危害",
    "Error locations": (8),
    "Error types": (phonological),
    "Medical entity locations": ([6,10])
}
```

Several examples in "MedConfusionSet.json":

```
{
    "椎": ["谁","柱","锥","追","锤","坠","垂"],
    "氨": ["铵","安","胺","按"],
    "痕": ["行","印","恒","恨","狠","很","般","横"],
    "呤": ["令","铃"],
    "溃": ["匮","贵","亏","馈","愧","奎"]
}
```

REFERENCES

[1] Haithem Afli, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. Using SMT for OCR error correction of historical texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation.* 962–966.

[2] Recep Sinan Arslan, Necaattin Barisci, Nursal Arici, and Sabri Kocer. 2021. Detecting and correcting automatic speech recognition errors with a new model. *Turkish Journal of Electrical Engineering & Computer Sciences* 29, 5 (2021), 2298–2311.

[3] Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 871–881.

[4] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS Workshop on Deep Learning.*

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* 4171–4186.

[6] Senka Drobac and Krister Lindén. 2020. Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition* 23, 4 (2020), 279–295.

[7] Jianyong Duan, Tianxiao Ji, and Hao Wang. 2018. Error correction for search engine by mining bad case. *IEICE transactions on Information and Systems* 101, 7 (2018), 1938–1945.

[8] Jianfeng Gao, Chris Quirk, et al. 2010. A large scale ranker-based system for search query spelling correction. In *The 23rd International Conference on Computational Linguistics.*

[9] Jinxi Guo, Tara N Sainath, and Ron J Weiss. 2019. A spelling correction model for end-to-end speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 5651–5655.

[10] Zhao Guo, Yuan Ni, Keqiang Wang, Wei Zhu, and Guotong Xie. 2021. Global attention decoder for chinese spelling error correction. In *Findings of the Association for Computational Linguistics.* 1419–1428.

[11] Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 2019. FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm. In *Proceedings of the 5th Workshop on Noisy User-generated Text.* 160–169.

[12] Li Huang, Junjie Li, Weiwei Jiang, Zhiyu Zhang, Minchuan Chen, Shaojun Wang, and Jing Xiao. 2021. PHMOSpell: Phonological and morphological knowledge guided Chinese spelling check. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* 5958–5967.

[13] Tuo Ji, Hang Yan, and Xipeng Qiu. 2021. SpellBERT: A lightweight pretrained model for Chinese spelling check. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* 3544–3551.

[14] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations.*

[15] Chong Li, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2021. Exploration and Exploitation: Two Ways to Improve Chinese Spelling Correction Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers).* 441–446.

[16] Chuan-Jie Lin and Shao-Heng Chen. 2015. NTOU Chinese grammar checker for CGED shared task. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications.* 15–19.

[17] Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang. 2021. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* 2991–3000.

[18] Xiaodong Liu, Kevin Cheng, Yanyan Luo, Kevin Duh, and Yuji Matsumoto. 2013. A hybrid Chinese spelling correction using language model and statistical machine translation with reranking. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing.* 54–58.

[19] Bruno Martins and Mário J Silva. 2004. Spelling correction for search engine queries. In *International Conference on Natural Language Processing.*

[20] Kareem Mokhtar, Syed Saqib Bukhari, and Andreas Dengel. 2018. OCR error correction: State-of-the-art vs an NMT-based approach. In *13th IAPR International Workshop on Document Analysis Systems.* IEEE, 429–434.

[21] Jane Moon and Frada Burstein. 2007. Ontology-based spelling correction for searching medical information. *Semantic Web Technologies and E-business: Toward the Integrated Virtual Organization and Business Process Automation* (2007), 384–404.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* (2019).

[23] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* 63, 10 (2020), 1872–1897.

[24] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval.* Vol. 39. Cambridge University Press Cambridge.

[25] Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing.* 32–37.

[26] Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 Bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing.* Association for Computational Linguistics, Beijing, China, 32–37.

[27] Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for Chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.*

[28] Dingmin Wang, Yi Tay, and Li Zhong. 2019. Confusionset-guided pointer networks for chinese spelling check. In *Proceedings of the 2019 Conference on Association for Computational Linguistics.*

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[30] Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013.. In *SIGHAN@ IJCNLP.* Citeseer, 35–42.

[31] Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing.* Asian Federation of Natural Language Processing, Nagoya, Japan, 35–42.

[32] Weijian Xie, Peijie Huang, Xinrui Zhang, Kaiduo Hong, Qiang Huang, Bingzhou Chen, and Lei Huang. 2015. Chinese spelling check system based on n-gram model. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing.* 128–136.

[33] Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: leveraging multimodal information helps Chinese spell checking. In *Findings of the Association for Computational Linguistics.* Association for Computational Linguistics, Online, 716–728.

[34] Li Yang, Ying Li, Jin Wang, and Zhuo Tang. 2019. Post text processing of Chinese speech recognition based on bidirectional LSTM networks and CRF. *Electronics* 8, 11 (2019), 1248.

[35] Jui-Feng Yeh, Wen-Yi Chen, and Mao-Chuan Su. 2015. Chinese spelling checker based on an inverted index list with a rescoring mechanism. *ACM Transactions on Asian and Low-Resource Language Information Processing* 14, 4 (2015), 1–28.

[36] Jui-Feng Yeh, Yun-Yun Lu, Chen-Hsien Lee, Yu-Hsiang Yu, and Yong-Ting Chen. 2014. Chinese word spelling correction based on rule induction. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing.* 139–145.

[37] Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing.* 220–223.

[38] Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of SIGHAN 2014 Bake-off for Chinese Spelling Check. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing.* Association for Computational Linguistics, Wuhan, China, 126–132.

[39] Ruiqing Zhang, Chao Pang, Chuanqiang Zhang, Shuohuan Wang, Zhongjun He, Yu Sun, Hua Wu, and Haifeng Wang. 2021. Correcting Chinese spelling errors with phonetic pre-training. In *Findings of the Association for Computational Linguistics.* 2250–2261.

[40] Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 882–890.

[41] Shiliang Zhang, Ming Lei, and Zhijie Yan. 2019. Investigation of Transformer based spelling correction model for CTC-based end-to-end mandarin speech recognition.. In *Interspeech.* 2180–2184.

[42] Hai Zhao, Deng Cai, Yang Xin, Yuzhu Wang, and Zhongye Jia. 2017. A hybrid model for Chinese spelling check. *ACM Transactions on Asian and Low-Resource Language Information Processing* 16, 3 (2017), 1–22.

[43] Xiaofang Zhou, An Zheng, Jiaheng Yin, Rudan Chen, Xianyang Zhao, Wei Xu, Wenqing Cheng, Tian Xia, Simon Lin, et al. 2015. Context-sensitive spelling correction of consumer-generated content on health care. *JMIR Medical Informatics* 3, 3 (2015), e4211.