information

MDPI

Article

Spelling Correction of Non-Word Errors in Uyghur–Chinese Machine Translation

Rui Dong [1,2,3], Yating Yang [1,2,*] and Tonghai Jiang [1]

[1] Xinjiang Technical Institute of Physics and Chemistry Chinese Academy of Science, Urumqi 830011, China; dongrui@ms.xjb.ac.cn (R.D.); jth@ms.xjb.ac.cn (T.J.)
[2] Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China
[3] University of the Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: yangyt@ms.xjb.ac.cn; Tel.: +86-0991-3835330

Received: 22 April 2019; Accepted: 4 June 2019; Published: 6 June 2019

Abstract: This research was conducted to solve the out-of-vocabulary problem caused by Uyghur spelling errors in Uyghur–Chinese machine translation, so as to improve the quality of Uyghur–Chinese machine translation. This paper assesses three spelling correction methods based on machine translation: 1. Using a Bilingual Evaluation Understudy (BLEU) score; 2. Using a Chinese language model; 3. Using a bilingual language model. The best results were achieved in both the spelling correction task and the machine translation task by using the BLEU score for spelling correction. A maximum F1 score of 0.72 was reached for spelling correction, and the translation result increased the BLEU score by 1.97 points, relative to the baseline system. However, the method of using a BLEU score for spelling correction requires the support of a bilingual parallel corpus, which is a supervised method that can be used in corpus pre-processing. Unsupervised spelling correction can be performed by using either a Chinese language model or a bilingual language model. These two methods can be easily extended to other languages, such as Arabic.

Keywords: spelling correction; natural language processing; machine translation; language model; Uyghur

1. Introduction

Whether it is the traditional statistical machine translation (SMT), or the recent research focus on neural machine translation (NMT), out-of-vocabulary (OOV) has always been a problem affecting translation. SMT [1,2] and NMT [3,4] both have many problems with OOVs.

With the development of the internet and the increasingly wide use of social media, written sentences have become progressively more colloquial. In order to improve the accuracy of natural language processing tasks such as speech recognition [5] and machine translation, a great deal of in-depth research on spelling correction has been undertaken.

Damerau proposed a rule-based method which used dictionary matching to check and correct spelling errors [6]. Kernighan attempted to use a noisy channel model for spelling correction [7]. Church used probability scores to sort candidate words [8]. Brill proposed an improved noisy channel model for spelling correction, based on generic string-to-string edits [9]. WordNet was used by researchers to correct real-word spelling errors [10]. Monojit used a supervised HMM (Hidden Markov Model) method to convert non-standard words into standard words in text messages [11]. The model parameters had been estimated from a word-aligned SMS (Short Messaging Service) and standard English parallel corpus, through machine learning techniques. Bassam introduced two probabilistic measures for completing the correction task [12]. Mohammed believed that spelling correction require three components [13]: A dictionary, an error model and a language model. From these three aspects, he began to improve the accuracy of spelling correction in Arabic. Al-Jefri uses the n-gram language model to detect Arabic real-word errors and uses the n-gram language model for spelling correction [14]. Uyghur belongs to the Turkish language family of the Altaic language system. It is a complex morphological language with rich morphology. There are still few researches on the spelling correction for Uyghur. Azgurli used a Uyghur dictionary for spelling checks, and then used the Uyghur rules base for spelling correction [15]. Abdurexiti used Uyghur lexical analysis for spelling checking, and then used the shortest editing distance for spelling correction [16]. Maihefureti used a Uyghur dictionary and a Uyghur stem dictionary for spelling checking, and then used an N-gram model to judge whether the connection between stem and suffixes was correct [17]. Memet created a Bi-gram dictionary from the previous word and the next word of the current word, and then used the Bi-gram dictionary for spelling checking [18]. The shortest edit distance method was used to pick the best candidate.

The object of machine translation began to shift from the traditional official text to the colloquial text, and the OOVs caused by spelling mistakes was seen more often.

Due to the high degree of freedom in spelling of Uyghur languages, spelling errors are common in colloquial text spelling. Therefore, the OOV problem caused by misspelling is more serious than the translation of other language pairs in the Uyghur–Chinese machine translation of colloquial text. Consider the Latin–Uyghur colloquial sentence “dot, rasmv”. These two words are all misspelled words and cannot be translated into Chinese by Uygur–Chinese machine translation. The correct spelling of the sentence is “dǒt, rasmu”, which is translated into Chinese as “笨,真的 (Stupid, really)”. For the sentence “xape bolmay tilfunimgha 50 koy sep qoysingiz boptiken”, the result of translation is “不xape tilfunimgha 50元就好了.”, which cannot be translated completely. The words “xape” and “tilfunimgha” are misspelled words, the correct words are “xapa (please)” and “tilfunumgha (telephone)”. The correct sentence is translated into Chinese as “麻烦你把我电话冲上50元就好了. (Please put 50 yuan into my telephone)”. From the above examples we can see that OOVs have a great influence on Uygur–Chinese machine translation.

Most spelling correction methods use a labeled corpus training error model to generate candidate words, and then select the right words through the language model, or through linguistic rules. The purpose of this paper is to improve the accuracy of machine translation by solving non-word misspelling problems. The rest of the paper is organized as follows. In Section 2, we introduce the background. We describe the proposed method for spelling correction in Section 3. Section 4 presents the experimental corpus and results, and we conclude in Section 5.

2. Background

Spelling correction corrects the misspelled word in the text and returns the correct word. There are two main types of spelling errors: non-word and real-word errors.

Definition 1. ***Non-word error*** *is the result of a spelling error where the word itself is not in the dictionary and is not a known word. For example, mistakenly spelling “apple” into “appll” is a non-word error because “appll” is not in our dictionary.*

Definition 2. ***Real-word error*** *is due to misspelling a word to make another word that is in the dictionary. For example, mistakenly spelling “apple” in “I have an apple” as "apply", makes the sentence “I have an apply”. This is a real-word error because “apply” is in our dictionary, but is not the right word.*

Definition 3. ***Perplexity (PPL)*** *is an indicator used to measure the quality of a language model in the field of natural language processing. It mainly estimates the probability of occurrence of a sentence based on each word, and normalizes the length of the sentence.*

$$\text{Perplexity} = 2^{-l} \tag{1}$$

$$l = \frac{1}{M}\sum_{i=1}^{m} \log p(s_i) \tag{2}$$

Definition 4. ***N-gram*** *is a model based on the assumption that the nth word appears to be related to the first n-1 words and not to any other words. The probability that the entire sentence appears is equal to the probability product of the occurrence of each word. The probability of each word can be calculated by statistical calculations in the corpus.*

$$\begin{aligned} \mathrm{P(Sentence)} &= \prod_{i=1}^{n} P(w_i) \\ &= \prod_{i=1}^{n} P(w_i|w_1 \ldots w_{i-1}) \end{aligned} \tag{3}$$

Definition 5. ***The shortest edit distance*** *is the minimum number of operations required to change from one string to another. The operation here refers to inserting, deleting, and replacing characters at a certain position in a string. The formula for the shortest edit distance is:*

$$D_{\mathrm{ij}} = \min\begin{cases} D_{i,j-1} + 1(\mathrm{inser}t) \\ D_{i-1,j} + 1(delete) \\ D_{i-1,j-1} + 1(replace) \end{cases} \tag{4}$$

For machine translation, OOV can be defined as a non-word error, and the dictionary is defined as all of the Uyghur words in the parallel corpus.

The simplest non-word error for the unsupervised spelling error correction method is as follows:

Step 1. Use the shortest edit distance for misspelled words to find the most likely candidates.

Step 2. The lower the PPL, the more likely the word is the correct word, using the n-Gram language model to select candidate sentences.

3. Using Uyghur–Chinese Machine Translation for Uyghur Spelling Correction

The purpose of studying Uyghur spelling correction is to solve the problem of foreign words in machine translation, so as to improve the quality of Uyghur–Chinese machine translation. If we can solve a non-word misspelling, the quality of the machine translation must be improved. Conversely, we consider if Uyghur-Chinese machine translation can help improve Uyghur spelling correction? We know that the correct spelling of the words in a sentence gives a higher quality translation than the quality of translation with an error word in a sentence. Therefore, spelling correction can be performed using machine translation.

Using machine translation for spelling correction is an idea of error magnification. The errors are passed through multi-level natural language processing tasks, so the errors will be magnified. For example, in speech translation, speech recognition errors will be magnified by machine translation. For spelling correction, the influence of incorrect words on sentences will be magnified by machine translation, which makes it easier to select sentences containing correct words from candidate sentences.

Uyghur is a type of agglutinative language, and Uyghur words are composed of stems and additional suffixes. For example: "bEyjingda (in Beijing)" and "bEyjingdin (from Beijing)" have the same stem "bEyjing". Adding different suffixes to each stem can produce more than 1800 words. Word formation is very flexible. This flexible way of word formation will lead to many words with similar meanings when generating candidate word sets through the shortest editing distance. Using a Uyghur language model to sort candidate sentences set, there will be sentences with similar PPL values. It is difficult to ensure that the sentence with the lowest PPL value is the correct sentence. For example, the sentence "axsham uxlap qaptimenken sizge xet yazmaptim" contains two wrong words "qaptimenken" and "yazmaptim", the correct words are "qaptikenmen" and "yazmaptimen", the correct sentence is “axsham uxlap qaptikenmen sizge xet yazmaptimen”. Using a Uyghur language model to sort candidate sentences into a set, we then selected the five sentences with the lowest PPL value as shown in Table 1.

Table 1. Examples of Uygur language model for spelling correction.

No.	Candidate Sentences	Uyghur LM PPL Scores	Translation Result	BLEU(2-Gram)	Chinese LM PPL Scores
1	axsham uxlap qaptikenmen sizge xet yazmaptu	1247.37683387	昨晚睡觉了他没写信给你	30	39.4232909338
2	axsham uxlap qaptikenmen sizge xet yazmapsiz	1305.01848157	昨晚睡觉了你没写信给你	30	37.6604732108
3	axsham uxlap qaptikenmen sizge xet yazattim	1370.68035299	昨晚睡觉了我要写信给你	40	41.6196380647
4	axsham uxlap qaptikenmen sizge xet yazmamtim	1783.78385363	昨晚睡觉了yazmamtim信给你	37.5	270.052007821
5	axsham uxlap qaptikenmen sizge xet yazmaptimen	1796.53781059	昨晚睡觉了我没写信给你	70	35.522803018

From the table we can see that the PPL values of all Uyghur sentences are very high. This is because our Uyghur language model is trained by standard Uyghur texts. The fifth sentence in the table is the correct one, but it is not the best one after sorting with the Uyghur language model. Using the Uyghur language model to score can only guarantee correct sentences in the top-*n* sentences. This method cannot guarantee that the correct sentence score is the best choice. Therefore, we used machine translation to correct misspelt words.

3.1. Using BLEU Score for Spelling Correction

We analyzed the translation results and found that the OOVs that were caused by spelling errors have a great influence on the process of word alignment, sequencing and decoding in statistical machine translation. In other words, if an OOV word can become a known word, the BLEU score must be improved. So we can use the BLEU score as the basis for selection of candidate words. The flowchart of the algorithm is shown in Figure 1.

As shown in Table 1, using the BLEU score for spelling correction, we translated the candidate sentences into Chinese, and calculated the BLEU score. We found that the fifth sentence had the highest BLEU score of 70, so the fifth sentence is the correct sentence, and the candidate words in the sentence “qaptikenmen” and “yazmaptimen” are the correct words after error correction.

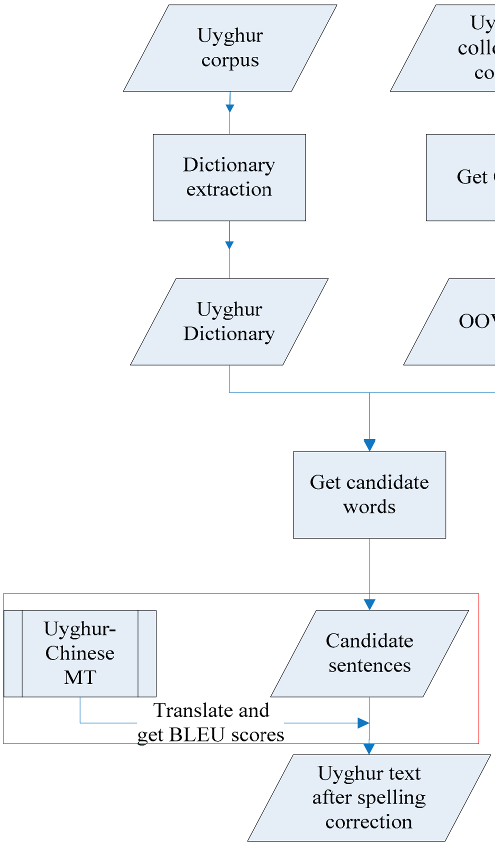

Figure 1. Using the BLEU score for spelling correction in an iteration. Generate Uyghur dictionary from Uyghur sentences in a bilingual corpus, the OOVs list from Uyghur colloquial, and then get all candidate words from the dictionary. Use candidate words to replace OOVs to generate candidate sentences. Translate candidate Uyghur sentences into Chinese sentences using Uyghur–Chinese machine translation, and then perform BLEU scores on sentences to return the highest scoring sentence. Candidates words contained in the sentence are the correct words.

3.2. Using the Chinese Language Model for Spelling Correction

The method of using the BLEU score for spelling correction needs the support of a bilingual parallel corpus, and cannot carry out unsupervised Uyghur spelling correction. Therefore, we propose an unsupervised method for Uyghur spelling correction based on the Chinese language model.

Because of the Uyghur–Chinese machine translation system, we not only have Uyghur monolingual resources, but also have bilingual resources. We can: 1. For each misspelled word in the Uyghur sentence, find the new candidate set by the shortest edit distance; 2. Translate the Uyghur sentences which contained all candidates into Chinese; 3. Score the Chinese sentences with a Chinese language model; and 4. Select the sentence that has the lowest PPL score for the candidate words in the selected sentence which are the spelling correction's result. The strategy of using the Chinese language model is based on the following considerations: one misspelled word in a sentence can have a huge impact on the overall machine translation. In short, when a misspelled word is translated, the error is magnified. If we can correct the word, then the quality of the final translation must be significantly improved. The flowchart of the algorithm is shown in Figure 2.

As shown in Table 1, using the Chinese LM for spelling correction, we translated the candidate sentences into Chinese, and calculated the PPL score. We found that the fifth sentence has the lowest PPL score of 35, so the fifth sentence is the correct sentence, and the candidate word in the sentence "qaptikenmen "and" yazmaptimen" is the correct word after error correction.

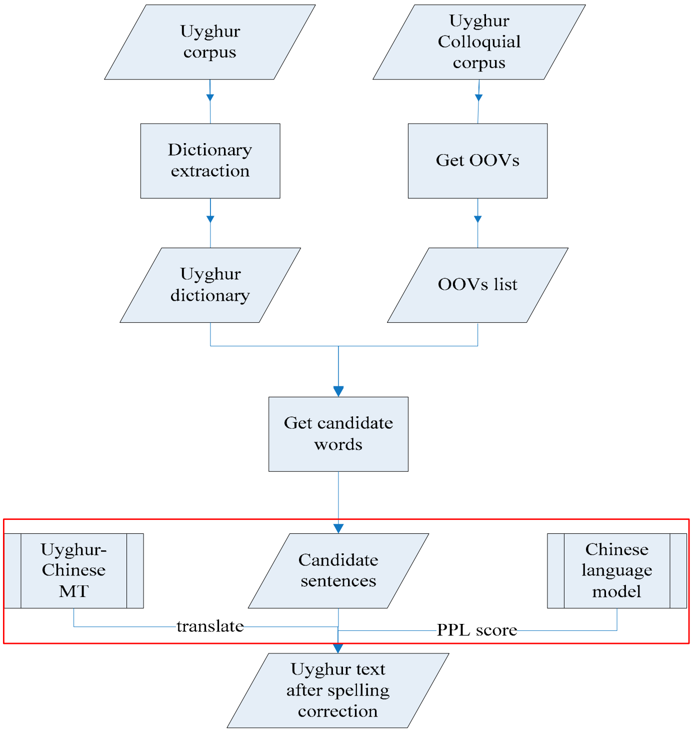

Figure 2. Using a Chinese language model (LM) for spelling correction in an iteration. Generate Uyghur dictionary from Uyghur sentences in bilingual corpus, the OOVs list from Uyghur colloquial, and then get all candidate words from the dictionary. Use candidate words to replace OOVs to generate candidate sentences. Translate candidate Uyghur sentences into Chinese sentences using Uyghur–Chinese machine translation, and then perform Chinese language model PPL scores on sentences, to return the highest score sentence. The candidate words contained in the sentence are the corrected words.

3.3. Using Bilingual Language Models for Spelling Correction

In Section 3.2, if a sentence contains multiple OOVs and there are multiple candidate words in the dictionary, the number of candidate sentences is very large and the translation process consumes a lot of resources. In order to improve the above problem, we propose a method of spelling error correction using the bilingual language model. The Uyghur monolingual model is used to prune the candidate sentences and select the optimal sentences. The flowchart of the algorithm is shown in Figure 3.

As shown in Table 1, by using the bilingual LM for spelling correction, pruning candidate sentences using the Uyghur language model, and then translating candidate sentences to Chinese, we calculated the PPL score as follows:

$$BiPPL = UYGHUR_{ppl} \div 1000 + CHINESE_{ppl} \tag{5}$$

We found that the fifth sentence had the lowest PPL score of 37, so the fifth sentence is the correct sentence, and the candidate words in the sentence “qaptikenmen” and “yazmaptimen” are the correct word after error correction.

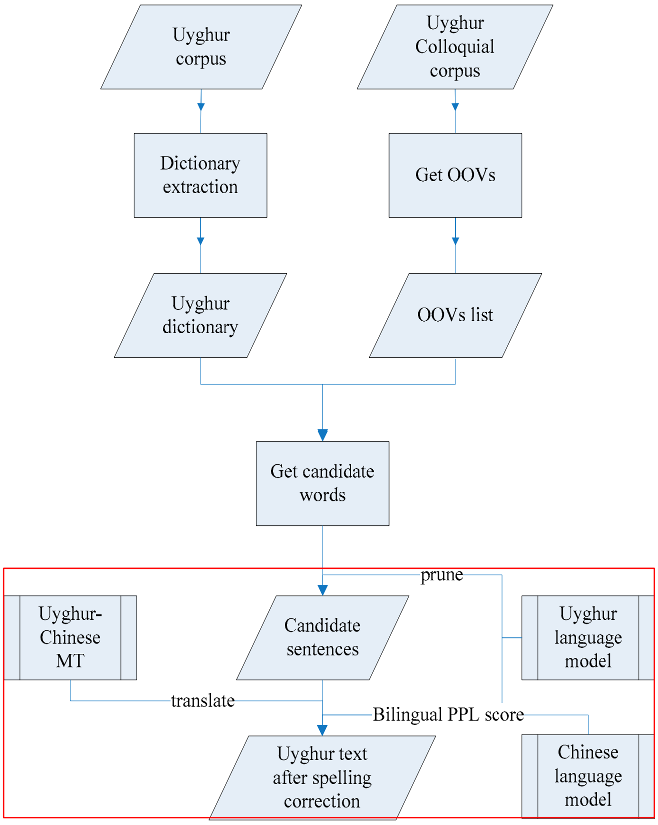

Figure 3. Using a bilingual LM for spelling correction in an iteration. Generate Uyghur dictionary from Uyghur sentences in bilingual corpus, get the OOVs list from Uyghur colloquial, and then get all candidate words from the dictionary. Use candidate words to replace OOVS to generate candidate sentences, simultaneously pruning using the Uyghur language model. Translate candidate Uyghur sentences into Chinese sentences using Uyghur–Chinese machine translation (MT), and then calculate Chinese language model and Uyghur language model PPL scores on sentences, returning the highest scoring sentence. The candidate words contained in the sentence are the corrected words.

4. Corpus and Result

We used 330,000 sentences of a Uyghur–Chinese bilingual parallel corpus (CWMT 2016) for training the Uyghur–Chinese machine translation system, and 692 colloquial sentences as the test corpus. As shown in Table 2, 68% of the sentences contain OOVs and 45% of the sentences contain non-word spelling errors, spelling errors are very common in colloquial Uyghur texts.

Table 2. Test corpus description.

Corpus	Numbers
Parallel corpus	692
Uyghur words	3993
Number of sentences containing OOVs	471
Number of non-word errors	312

From the experimental results in Tables 3 and 4, we can see that the BLEU score can be used for spelling correction, and the BLEU value of the translated result is significantly improved after the spelling error correction.

Table 3. The spell correction result.

Methods	Precision	Recall	F1
Using BLEU scores	**0.89**	0.60	**0.72**
Using Chinese LM PPL	0.70	0.51	0.59
Using Uyghur-Chinese LM PPL	0.74	0.65	0.69

Table 4. The translation result.

Methods	BLEU
baseline	11.67
Using BLEU scores correct	**13.64(+1.97**)
Using Chinese LM PPL	12.72(+1.05)
Using Uyghur-Chinese LM PPL	13.07(+1.4)

The accuracy and recall of spelling correction using the Chinese language model are lower than those when using BLEU scoring. An error analysis of the experimental results follows. The main reasons for the reduction in the accuracy and recall scores are: 1. Some of the candidate sentences selected by the Chinese monolingual model are very consistent with the grammatical rules of the Chinese language, and the reading is unobstructed, but it is not intended to be expressed in the original text; 2. This spelling correction method will be part of the correct spelling but not in the dictionary within the scope of the word spelling correction, so is over-corrected. The translation result is also a significant improvement over the baseline system.

The accuracy and recall of the spelling correction by using bilingual language models for spelling correction are higher than using the Chinese language model. The translation result BLEU score is also higher than using the Chinese language model for spelling correction, and lower than using the BLEU score for spelling correction.

5. Conclusions

In this paper, three methods of spelling correction using Uyghur–Chinese machine translation were assessed. Using the BLEU score for spelling correction achieved the best results in both spelling correction and machine translation. The F1 score of spelling correction reached 71%, and the BLEU score of machine translation reached 13.64, which was 1.97 points higher than the baseline system.

Using BLEU for spelling correction can be applied to the acquisition and processing of parallel corpus, such as the automatic acquisition of corpus and manual annotation of corpus, which can reduce sparsity caused by spelling errors, improve the quality of parallel corpus, and finally improve the quality of machine translation.

Unsupervised spelling correction can be performed using both the Chinese language model and the bilingual language model. The unsupervised spelling correction method is easily extended to machine translation of other language pairs.

Author Contributions: Conceptualization, R.D. and Y.Y.; methodology, R.D.; software, R.D.; validation, R.D., Y.Y. and T.J.; writing—original draft preparation, R.D.; writing—review and editing, Y.Y. and T.J.; funding acquisition, R.D. and Y.Y.

Funding: This research was funded by the Xinjiang Key Laboratory Fund under Grant No. 2018D04018, the Xinjiang Science and Technology Major Project under Grant No.2016A03007-3 and the Natural Science Foundation of Xinjiang under Grant No.2015211B034.c

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, J.; Zhai, F.; Zong, C. Handling unknown words in statistical machine translation from a new perspective. In *Natural Language Processing and Chinese Computing*; Springer: Berlin/Heidelberg, Germany, 2012.

2. Zhang, J.J.; Zhai, F.F.; Zong, C.Q. A substitution-translation-restoration framework for handling unknown words in statistical machine translation. *J. Comput. Sci. Technol.* **2013**, *28*, 907–918. [CrossRef]
3. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
4. Liu, A.; Kirchhoff, K. Context models for oov word translation in low-resource languages. *arXiv* **2018**, arXiv:1801.08660 2018.
5. Islam, N.; Ranga, K.K. A process to improve the accuracy of voice recognition system by using word correction system. *Compusoft* **2014**, *3*, 822.
6. Damerau, F.J. A technique for computer detection and correction of spelling errors. *Commun. ACM* **1964**, *7*, 171–176. [CrossRef]
7. Kernighan, M.D.; Church, K.W.; Gale, W.A. A spelling correction program based on a noisy channel model. In Proceedings of the 13th conference on Computational linguistics, Helsinki, Finland, 20–25 August 1990.
8. Church, K.W.; Gale, W.A. Probability scoring for spelling correction. *Stat. Comput.* **1991**, *1*, 93–103. [CrossRef]
9. Brill, E.; Moore, R.C. An improved error model for noisy channel spelling correction. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, China, 3–6 October 2000; pp. 286–293.
10. Pedler, J. Computer Correction of Real-Word Spelling Errors in Dyslexic Text. Ph.D. Thesis, University of London, London, UK, 2007.
11. Choudhury, M.; Saraf, R.; Jain, V.; Mukherjee, A.; Sarkar, S.; Basu, A. Investigation and modeling of the structure of texting language. *Int. J. Doc. Anal. Recognit.* **2007**, *10*, 157–174. [CrossRef]
12. Haddad, B.; Yaseen, M. Detection and correction of non-words in arabic: A hybrid approach. *Int. J. Comput. Process. Orient. Lang.* **2007**, *20*, 237–257. [CrossRef]
13. Attia, M.; Pecina, P.; Samih, Y.; Shaalan, K.; Van Genabith, J. Arabic spelling error detection and correction. *Nat. Lang. Eng.* **2016**, *22*, 751–773. [CrossRef]
14. Al-Jefri, M.M.; Mohammed, S.A. Arabic spell checking technique. U.S. Patent 9,037,967, 19 May 2015.
15. Azgurli, L.; Yusuf, A. Corpus-based Uyghur text proofreading system principle. *Inf. Comput.* **2012**, *12*, 156–157. (In Chinese)
16. Abdurexiti, R. Design and Implementation of Uyghur Word Auto-Proofreading System. Master's Thesis, University of Electronic Science and Technology, Chengdu, China, 2013. (In Chinese).
17. Maihefureti, A.W.; Aili, M.; Yibulayin, T.; Jian, Z.H.A.N.G. Spelling Check Method of Uyghur Languages Based on Dictionary and Statistics. *J. Chin. Inf. Process.* **2014**, *28*, 66–71.
18. Mutula, M.; Gurinigar, M.; Mauridan, N.; Escale, A. Uygur Spelling Error Detection and Automatic Correction Based on Context Relation. In Proceedings of the 14th National Academic Conference on Man-Machine Speech Communication, Lianyungang, China, 11–13 October 2017.