# Chinese Word Spelling Correction Based on N-gram Ranked Inverted Index List

**Jui-Feng Yeh*, Sheng-Feng Li, Mei-Rong Wu, Wen-Yi Chen, Mao-Chuan Su**
Department of Computer Science and Information Engineering,
National Chiayi University,
No.300 Syuefu Rd., Chiayi City 60004, Taiwan (R.O.C.).
{Ralph, s1010431, s1010432, s0992962,
s0992974}@mail.ncyu.edu.tw

## Abstract

Spelling correction can assist individuals to input text data with machine using written language to obtain relevant information efficiently and effectively in. By referring to relevant applications such as web search, writing systems, recommend systems, document mining, typos checking before printing is very close to spelling correction. Individuals can input text, keyword, sentence how to interact with an intelligent system according to recommendations of spelling correction. This work presents a novel spelling error detection and correction method based on N-gram ranked inverted index is proposed to achieve this aim, spelling correction. According to the pronunciation and the shape similarity pattern, a dictionary is developed to help detect the possible spelling error detection. The inverted index is used to map the potential spelling error character to the possible corresponding characters either in character or word level. According to the N-gram score, the ranking in the list corresponding to possible character is illustrated. Herein, E-How net is used to be the knowledge representation of tradition Chinese words. The data sets provided by SigHan 7 bakeoff are used to evaluate the proposed method. Experimental results show the proposed methods can achieve accepted performance in subtask one, and outperform other approaches in subtask two.

## 1 Introduction

Language is one of the most important capabilities of human for communication. Natural language cannot be absent in human communication either spoken communication or written text. As we known, word is the fundamental semantic unit in the most languages; it plays an essential role in natural language processing. Since the word is the building block for natural language processing, the spelling error or typos usually cause negative effects in word for computer applications.

Intelligent communication is one of the new trends about computing environment construction. In providing the natural intelligent human machine interaction, natural language expressions play an essential role. Let us now attempt to extend the observation into the frameworks of natural language processing, in viewpoints of input and output aspects, text input and sentence generation provide the main natural language interfaces between users and machines. Therefore, the semantic extraction and generating of natural language processing plays more essential roles for human machine interactions. Actually, we should now look more carefully into the results obtained in text input and natural language generating. Since the accuracy of text input is not near to perfect, it will cause the natural language misunderstanding. The spelling correction is one of the most important modules for natural language processing. The related applications including web search query, writing systems, recommend systems, document mining and typos checking before printing are very close to spelling correction.

There are many research effort developed for spelling error detection and correction recently. Sun et al. (2010) explore the phrase-based spelling error models from the clickthrough data by measuring the edit distance between an input query and the optimal spelling correction. Ahmad and Kondrak (2005) also have learned a spelling error model from search query logs to improve the quality of query. Li et al. (2006) applied distributional similarity based models for query spelling correction. Gao et al. (2010) Employed the ranker-based approach that contains a surface-form similarity, phonetic-form similarity,

*Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7)*, pages 43–48,
Nagoya, Japan, 14 October 2013.

entity, dictionary, and frequency features for large scale web search. Besides, Ahmad and Kondrak (2005) adopted EM algorithm to enhance the performance of spelling error detection. There are some works tried to build a transformation model like machine translation, the noisy channel model was one of the selected to describe the spelling error correction. Hidden Markov Models (HMMs) are used to correct Spelling errors for search queries and developed a system called as CloudSpeller (Li et al. 2011). Considering of the domain specific domain, Bao et al. (2011) employed graph theory to correct the error in word and query levels. Cucerzan and Brill (2004) used domain knowledge to exploit the spelling correction as an iterative process. For single word, context-sensitive spelling correction and rich morphology are proposed by Ingason et al. (2009). Mitton (2010) survey the spelling checking algorithm and systems developed for writing systems in the past five decades. Huang et al. (2010) proposed a system framework integrating n-gram models and internet knowledge resources to detect spelling errors in printer driver module. Actually, some application interface (API), tools and knowledge bases are useful for spelling error detection and correction. Google (2010) has developed a Java API for Google spelling check service. Microsoft (2010) also provides Microsoft web n-gram services. An online keyword typo generates tool, Seobook (2010), was developed for generating the corpus. Considering of lexicon and ontology, WordNet and FrameNet are both the main knowledge representations for English (Christiane 1998)). Correspondingly, HowNet and E-Hownet are lexicon ontologies for simple and traditional Chinese separately (Li et al. 2011;Dong and Dong 2006). According to the word expression in E-Hownet, lexical senses are described as two aspects: entities and relations. Thus, all the taxonomic relations of lexical senses can be identified according to their definitions in E-Hownet.

Since Word spelling is the essential for natural language processing, spelling correction is a common an essential task in written language automatically to detect and correct human errors. However, spelling check in Chinese is very different from that in English or other alphabetic languages. Therefore, a novel spelling error detection and correction method based on N-gram ranked inverted index is proposed in this paper. Considering of context information such as those in a sentence or long phrase with a certain meaning, N-gram scores are used to arrange the rank of nodes in the inverted index linked list. Besides word N-gram, character frequencies are also used herein first for errors result phonologically similar or visually similar characters (Liu et al. 2011). Both character and word information are used in the proposed approach to achieve the performance of spelling correction.

The rest of this paper is organized as follows. Section 2 describes the proposed method and the related important modules in spelling correction in system framework. Next, we also present the detail description about the proposed method especially in N-gram ranked inverted index list. Experiments to evaluate the proposed approach and the related discussion are presented in Section 3. Concluding remarks and findings are finally made in Section 4.

## 2 The proposed system framework

In this section, we want to illustrate the proposed system framework to detect and correct the spelling errors. Our goal is to find the locations and correct the corresponding error character in input Chinese sentences. For more clear presentation, herein, the system framework is divided into two parts: training and generation phases as described in Section 2.1 and 2.2 respectively. Actually, similar pronunciation and shape dictionary and N-gram ranked inverted index list are constructed in the training phase and adopted in test phase.

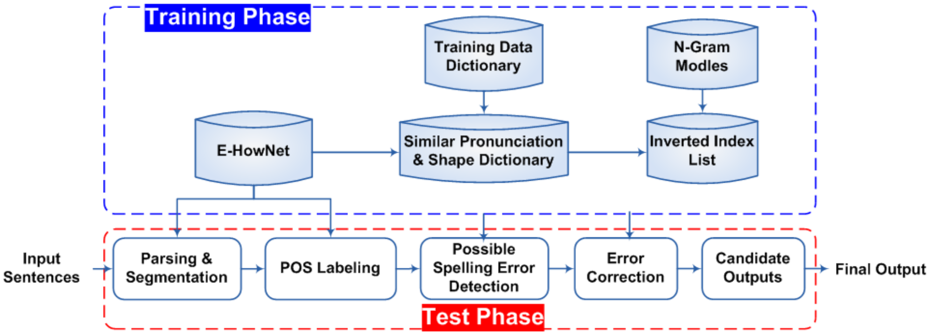

Figure 1. The proposed system framework

### 2.1 Training phase

The aim of training phase is to construct the dictionary containing similar pronunciation and shape information for each Chinese character. E-Hownet and pre-trained N-gram models are further used to be the ranking score construct the inverted index list. Finally, the candidate outputs are generated according to the N-gram ranked inverted index list. More detail illustration is described in the follows.

As shown in Figure 1. First, we are going to preprocess the sentences we got from the SIGHAN-7 organizer. In this step, we have to remove each sentence NID number in the input file. Then we will have the sentences that without the NID number as our output according to the traditional Chinese parser that was developed by academia Sinica, Taiwan. The results are further fed into the tool, CKIP Autotag, to do word segmentation and part-of-speech tagging based on E-Hownet. Since the corresponding part-of-speech (POS) of each word is obtained in the sentences. Each word is given a part of speech at the end of a word in parentheses. In the second step, for convenience, we are going to remove unessential blank spaces and parentheses. This way will let us more conveniently in the following file operations. In fact, these processes are also adopted in test phase.

For obtaining the correction of each possible spelling error, the similar pronunciation and shape dictionary are constructed here. Since typos usually resulted from similar pronunciation or character shape, we constructed the index for each work from its confusing set including similar pronunciation and character shape. Since the pronunciation of each Chinese character is composed of syllable and tone. Four categories of pronunciation similar confusing set those are potential correction in pronunciation, are gathered: same pronunciation, the same syllable with a different tone, similar syllable with the same tone and similar syllable with a different tone. The corresponding posterior probability is obtained by the confusing matrix used in speech recognition engine constructed by HTK. Considering of the corresponding characters, are called as potential correction in shape, those shapes are similar to that of possible spelling error character. Length based Cangjie code similarity measure is used to estimate the posterior probability for the shape confusing character set.

Since the potential correction either in pronunciation or shape are gathered and defined in the dictionary described in the previous paragraph, the competing candidates are obtained by replacing the possible spelling error using potential correction. One inverted index list for each possible spelling error is constructed according to the corresponding potential correction. Considering of efficiency, the node order is arranged according to the character frequency in initial state. Word based N-gram scoring is further used for re-sorting the node in the inverted index list. Herein, back off base tri-gram models are used to estimate the probability of the contextual information.

### 2.2 Test phase

As described in previous sections, the inverted index lists with N-gram ranking are built in the training phase. The spelling correction problem is formulated as the ranking of the potential corrections and original possible spelling error in the contextual score in the test phase. Since the word segmentation and part-of-speech (POS) labeling is the same as those in the training phase. Here, we begin the processes with the third step. Third, we are going to find the wrong word from the sentences. After we have the POS parsing result, we choose the word that consists of two characters from the POS result and find it with the words in E-HowNet. If we cannot find it in E-HowNet, then we regarded it as possible suspicious word and enumerate it in suspicious list. We saved its word and POS in a text file named find_wrong. E-HowNet is a lexical knowledge based evolved from HowNet and created by the CKIP group. Then we filter some words in this step in order to remove some words by mistake. Those filtered out words may be words consist of more than four characters with POS of VH …etc., words consist of more than three characters with POS of ‘Nb’, ‘VA’, ‘Nc’, ‘VE’ …etc and POS of ‘Neu’, ‘Neqa’, ‘Nf’, ‘VB’, ‘Ncd’, ‘VK’, ‘Nh’, ‘P’ …etc. And we also filter out the following words contain “到”(to), “過”(through), “亂”(disorder) and “年級”(grade) …etc. We show some of the suspicious word list.

The fourth step, we are going to do the error correction on those incorrect words. We choose one of a character in the word that to the suspicious list and refer to the similar pronunciation and similar shape dictionaries provided by the SIGHAN-7 organizer. For example, ‘挫’折 (setback) and 挫‘折’ (setback). We want to find out the pronunciation of the character. It may be same pronunciation with the same tone, same pronunciation without same tone, similar pronunciation with the same tone, similar pronunciation without same tone and same radical with same strokes. And we combine the character with a similar shape character into a new word. Then we find each new word in E-HowNet to verify if there is exist or not. If the new word was not found in E-HowNet, then we will save it into wrong dictionary. After fixing the error, we saved it into the correct dictionary. If the new word was found in E-HowNet, then we will skip to the next character combination word. And so on…. After finishing the word pronunciation part, then we do the same way in word shape part. Fifth, we have to remove the duplicate words in the wrong dictionary. And remove it in wrong and correct dictionary synchronously. As a result, we can prevent doing the same thing twice. Sixth, we use the words in the wrong dictionary to find in the sentences. If we found it, that is to say, the sentence contains this error. Then we replace the error with the corresponding correct word in the correct dictionary and calculate the error location in the output. Seventh, we have several different potential corrections and original possible spelling error, then re-ranking the order according to the N-gram language models in the optimization step. Finally, we can output the best result with the highest N-gram score to the output file.

## 3 Experimental results

This goal of this study is spelling error detection and correction according Chinese spelling check competition in SigHan. The aim of the subtask 1 is to find out the location of the spelling error in the sentences. On the other hand, the subtask 2 aims at finding out the error location and do the error correction. All sentences at least contain more than one error. In this bake-off, the evaluation includes two sub-tasks: error detection and error correction. The errors are collected from students' written essays. Since there are less than 2 errors per essay such as described in (Chen et al. 2011), in this bake-off the distribution of incorrect characters will match the real world error distribution in the sub-task one. The first subtask aims at the evaluation of error detection. The input sentences might consist of no error to evaluate the false-alarm rate of a system (Wu et al. 2010). The second sub-task focuses on the evaluation of error correction. Each sentence includes at least one error. The ability to accomplish these two sub-tasks is the complete function of a spelling checker.

### 3.1 Spelling Error Detection

The training data and test data consist of 350 and 1000 sentences separately. Both of them are provided by the SIGHAN-7 organizer.

Table 1. Performance evaluation of the proposed method in subtask 1.

| RUN | 1 | 2 | 3 |
|---|---|---|---|
| False-Alarm Rate | 0.2371 | 0.2129 | 0.0929 |
| Detection Accuracy | 0.738 | 0.761 | 0.825 |
| Error Location Accuracy | 0.623 | 0.652 | 0.748 |
| Detection Precision | 0.5514 | 0.5850 | 0.7451 |
| Detection Recall | 0.68 | 0.70 | 0.6333 |
| Detection F-score | 0.609 | 0.6374 | 0.6847 |
| Error Location Precision | 0.2405 | 0.2813 | 0.4431 |
| Error Location Recall | 0.2967 | 0.3367 | 0.3767 |
| Error Location F-score | 0.2657 | 0.3065 | 0.4271 |

According to the results shown in Table 1, the suitability of the subtasks 1 is high enough. Compared to other approaches, we consider the mapping between the spelling error and correction more.

### 3.2 Spelling Error Correction

The training data is same as error detection. The test data consists of 1000 sentences those are not the same as the error detection subtask 1.

Table 2. Performance evaluation of the proposed method in subtask 2.

| RUN | 1 | 2 | 3 |
|---|---|---|---|
| Location Accuracy | 0.369 | **0.663** | **0.663** |
| Correction Accuracy | 0.307 | **0.625** | **0.625** |
| Correction Precision | 0.485 | 0.703 | 0.703 |

According to the results shown in Table 2, the suitability of the subtasks 2 is excellent. Due to the proposed approach considers both character confusing set and word contextual information, the performance is able to provide the right information to detect and correct the spelling error for users. Especially, The proposed approach outperforms the other approaches significantly in location accuracy and correction accuracy. These results show the N-gram ranked inverted index list able to obtain improvement for spelling error correction. The performance of run 3 outperforms that of run 2 due to some pruning for the word with more than two characters. According to the observations of the error pattern obtained from the training data, we know the spelling error usually appears with the word with less than three characters. By this, the performance is improved significantly.

## 4 Conclusions

A novel approach to detect and correct the spelling error in traditional Chinese text are proposed in this study. The algorithm is based on the idea of N-gram ranked inverted index list. For detecting the potential correction, the similar patters based on pronunciation and character shape are gathered in a dictionary. To capture the contextual information, the word based N-gram ranking is adopted to arrange the node order in the inverted index list. Finally, the optimal result is selected as the output. The experimental results verified that the proposed approach results in keeping more information either in character or word levels. The performance about the spelling error detection is acceptable and that about correction outperforms other approaches. The experimental results show the proposed method is practice.

### Acknowledgments

The authors would like to acknowledge the National Science Council (NSC 100-2622-E-415-001-CC3) and Ministry of Education of Taiwan for financial support for this research.

## References

Sun, X., Micol, D., Gao, J., Quirk, C., 2010. Learning Phrase-Based Spelling Error Models from Clickthrough Data. *Proceedings of ACL 2010*.

Ahmad, F., and Kondrak, G. 2005. Learning a spelling error model from search query logs. *In HLT-EMNLP*, pp 955-962.

Li, M., Zhu, M., Zhang, Y., and Zhou, M. 2006. Exploring distributional similarity based models for query spelling correction. *Proceedings of ACL 2006*, pp. 1025-1032.

Gao, J., Li, X., Micol, D., Quirk, C., and Sun, X., 2010. A Large Scale Ranker-Based System for Search Query Spelling Correction, The 23rd International Conference on Computational Linguistics 2010 (COLING 2010). Pp. 358–366.

Ahmad, F., and Kondrak, G., 2005. Learning a Spelling Error Model from Search Query Logs, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 955–962.

Li, Y., Duan, H., Zhai, C.X. . 2011. CloudSpeller: Spelling Correction for Search Queries by Using a Unified Hidden Markov Model with Web-scale Resources. *Spelling Alteration for Web Search Workshop 2010*, pp.10-14.

Bao, Z., Kimelfeld, B., Li, Y., 2011. A Graph Approach to Spelling Correction in Domain-Centric Search, , *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL) 2011*, pp. 905–914.

Cucerzan, S., and Brill, E.. 2004. Spelling correction as an that exploits the collective knowledge of Web users. *Proceedging of Conference on Empirical Methods in Natural Language Processing EMNLP*, pp. 293–300.

Ingason, A.K., Johannsson, S.B., Rognvaldsson, E., Helgadóttir, S., Loftsson, H. 2009. Context-Sensitive Spelling Correction and Rich Morphology, *NODALIDA 2009 Conference Proceedings*, pp. 231–234.

Mitton, R. 2010. Fifty years of spellchecking. *Wring Systems Research*, 2:1–7.

Huang, Y.-H., Yen M.-C., Wu, G.-H., Wang, Y.-Y., Yeh, J.-F. 2010. Print Pickets Combined Language Models and Knowledge Resources. ROCLING 2010, pp.297-309.

Google. 2010. A Java API for Google spelling check service.http://code.google.com/p/google-api-spellingjava/.

Microsoft Microsoft web n-gram services. 2010. http://research.microsoft.com/web-ngram

Seobook. 2010. Keyword typo generator. http://tools.seobook.com/spelling/keywordstypos.

Christiane F. 1998. *WordNet: An Electronic Lexical Database* (Language, Speech, and Communication). The MIT Press.

Dong, Z.D., and Dong Q. 2006. *HowNet and the Computation of Meaning*. World Scientific Publishing Co. Pte. Ltd.

Liu, C.-L., Lai, M.-H., Tien, K.-W., Chuang, Y.-H., Wu, S.-H., and Lee, C.-Y. 2011. Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications, *ACM Trans. Asian Lang. Inform. Process.* Vol. 10, No. 2, Article 10 (June 2011), 39 pages.

Chen, Y.-Z., Wu, S.-H., Yang, P.-C., Ku, T., and Chen, G.-D. 2011. Improve the detection of improperly used Chinese characters in students' essays with error model. *Int. J. Cont. Engineering Education and Life-Long Learning*, Vol. 21, No. 1, pp.103-116, 2011.

Wu, S.-H., Chen, Y.-Z., Yang, P.-C., Ku, T., and Liu, C.-L. 2010. Reducing the False Alarm Rate of Chinese Character Error Detection and Correction, *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010)*, pages 54－61, Beijing, 28-29 Aug., 2010.

Chen, W.-T., Lin, S.-C., Huang, S.-L., Chung, Y.-S., and Chen, K.-J. 2010, E-HowNet and Automatic Construction of a Lexical Ontology, *the 23rd International Conference on Computational Linguistics*, Beijng, China.

Bai, M.-H., Chen K.-J., and Chang, J. S. 2008, Improving Word Alignment by Adjusting Chinese Word Segmentation, *Proceedings of IJCNLP2008*.