Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Building a Confused Character Set for Chinese Spell Checking

Lung-Hao LEE[a], Wun-Syuan WU[b], Jian-Hong LI[a], Yu-Chi LIN[c] & Yuen-Hsien TSENG[c*]
[a]*Department of Electrical Engineering, National Central University, Taiwan*
[b]*Department of Chinese, National Taiwan Normal University, Taiwan*
[c]*Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taiwan*
*samtseng@ntnu.edu.tw

Abstract: In this paper, we describe the construction details of a confused character set for Chinese spell checking. The SIGHAN 2013-2015 bakeoff datasets are adopted to measure the performance of correct character suggestions. Our confusion set significantly outperforms the existing confusion set in candidate selection for automatic spelling checkers.

Keywords: Chinese spell checking, confusion set, shape similarity, pronunciation similarity

1. Introduction

Automatic Chinese spell-checking tools (like those for English in MS Word) are valuable for Chinese language learners. However, it is particularly challenging for Chinese, in part, because whether or not a Chinese character is correct depends on its context. For example, the character "偏" ("pian" in pronunciation) is correct in the word "偏見" ("pian jian4"; "prejudice" in English), but it is incorrect in the word "普偏" which instead should be "普遍" ("pu3 pian4"; "common"). Hence, a sentence "學生打工的情況很普偏" is incorrect and its corresponding correction is "學生打工的情況很普遍" (A student having a part-time job is very common). The possible error causes may be similar pronunciation and/or shape which are shared between the characters "偏" and "遍".

Ideally, automatic spelling checkers should be capable of detecting character or word errors and making correction accordingly. In practice, they often first pin-point the locations of various types of spelling errors and then suggest a candidate list of correct characters for the detected error where the correct one should be as ranked as early as possible (favorably top 1 in the list) in order for automatic replacement. In other words, spell checking is often a two-step process: Once erroneous characters are identified, a confused character set (henceforth referred as "confusion set") is applied for error correction/replacement. This paper focuses on the second part of the Chinese spell checking. In particular, we focus on building of a confusion set that is independent of the methods to spot the spell errors and that is better than the existing ones.

Empirical analysis shows that most Chinese spelling errors arise from phonologically similar, visually similar, and semantically confused characters (Liu et al. 2011). A confusion set contains a set of seed characters each accompanying with a list of corresponding confused characters. For example, a character "不" in the set can be used to find its confused character list such as "部, 步, 布, ….". In general, a confusion set is built according to similar shape or pronunciation that are concerned. However, without a proper order of the confused character list, it is inconvenient for an automatic spelling checker to seek possible correction candidates effectively and efficiently. This observation motivates us to build a universal confusion set to facilitate automatic Chinese spell checking.

2. Chinese Confusion Set Construction

A first step for confusion set construction is to determine candidate characters as seeds. The Sinica Corpus is the balanced Chinese corpus with word segmentation and part-of-speech tagging. We obtain a word list with accumulated word frequency in Sinica Corpus 3.0. From this word list sorted in decreasing order of frequency, there are 4,743 distinct and commonly used Chinese characters. These characters are regarded as the seeds and stored according to their frequencies decreasingly.

The next step is to find easily confused characters for each seed character. The following three major methods are used and combined to yield a confused character list for each seed character:

Real-Error Frequency. In SIGHAN 2013-2015 bakeoffs, real spelling errors caused by Chinese native speakers or Chinese-as-a-foreign-language learners were collected and manually corrected to evaluate the performance of automatic spelling checkers (Tseng et al., 2015; Wu et al., 2013; Yu et al., 2014). In the training sets, there are 8,772 spelling errors, in which 3,394 characters are unique. We utilize these spelling errors (and their corrections) to find the confused characters for each corrected character (which is also in seed characters). For example, the most seen character "的" in the spelling errors is frequently misused as "得" (161 cases), followed by "地" (32 cases). In this case, we create a list of confused/misused characters for the character: "的".

Shape Similarity. For this error type, we use common and basic vocabulary, e.g., use the word "unusual" rather than the word "arcane". The Master Ideographs Seeker ("全字庫" in Chinese) developed by National Development Council, Taiwan provides the components of each Chinese character. For example, a character "腦" consists of components "月", "巛", "丿", "口" and "乂". We then measure the shape similarity between two characters using the Jaccard similarity coefficient. For example, the similary of "腦" and "惱" (its components are "忄", "巛", "丿", "口" and "乂") is 0.66 (=4/6). If a character have more than one unique constructed components such as "行" (there are two constructed components: one is "彳", "一" and "丁"; the other is "彳", "亍" and "亅"), each combination will be used to measure the similarity respectively and the larger one will be retained.

Pronunciation Similarity. The Master Ideographs Seeker also provides the pronunciation of each character. If more than one pronunciation is used for a character, the most common usage in the CKIP Electronic Dictionary will be adopted. In Chinese phonetics, a character can be pronounced using initial consonants, vowels, and tones. Because its relative complexity, we describe the three phonetic similarity measures individually and their combination with examples as follows.

For the initial consonants of Chinese phonetics, Table 1 shows points and manners of articulation and the corresponding coordinates. The similarity of any two consonants is regarded as distance calculation in the coordinate system. All distances are normalized into values ranging from 0 to 1. The most similar is 1 when identical consonants exist between two characters. For no-consonant cases, its similarity with any one of consonants is 0. Besides, the similarity of two no-consonant cases is also 1.

Table 1. *Places and Manners of Articulation for Consonants and their Corresponding Coordinates*

Manners \ Places			bilabial	labiodental	alveolar	velar	palatal	post alveolar	dental
status	vocal folds	airflow							
plosive	voiceless	unaspirated	ㄅ(1,1)		ㄉ(3,1)	ㄍ(4,1)			
		aspirated	ㄆ(1,2)		ㄊ(3,2)	ㄎ(4,2)			
affricate	voiced	unaspirated					ㄐ(5,3)	ㄓ(6,3)	ㄗ(7,3)
		aspirated					ㄑ(5,4)	ㄔ(6,4)	ㄘ(7,4)
nasal	voiced		ㄇ(1,5)		ㄋ(3,5)				
lateral approx.	voiced				ㄌ(3,6)				
fricative	voiceless			ㄈ(2,7)		ㄏ(4,7)	ㄒ(5,7)	ㄕ(6,7)	ㄙ(7,7)
	voiced							ㄖ(6,8)	

For vowels (or called finals for more general cases) of Chinese phonetics, they are divided into five types called: 1) prenuclear glides: ㄧ、ㄨ、ㄩ; 2) simple finals: ㄚ、ㄛ、ㄜ、ㄝ; 3) compound finals: ㄞ、ㄟ、ㄠ、ㄡ; 4) nasal finals: ㄢ、ㄤ、ㄣ、ㄥ; and 5) retroflex final: ㄦ. If two prenuclear glides are the same, the similarity is 1, otherwise 0. For the remaining finals, if two finals are the same, the similarity is 1; if they are the same type, the similarity is 0.5; other cases are 0.

For tones of Chinese phonetics, there are five distinct tones, that is, neutral tone, 1st tone, 2nd tone, 3rd tone, and 4th tone. The largest similarity is 1 for identical tones and the lowest similarity is 0 for the pair tone: neutral vs. 4th. The similarity is 0.75 for the following pairs: neutral vs. 1st, 1st vs. 2nd, 2nd vs. 3rd, and 3rd vs. 4th. The similarity is 0.5 for the pairs: neutral vs. 2nd, 1st vs. 3rd, and 2nd vs. 4th. Finally, the similarity is 0.25 for the pairs: neutral vs. 3rd and 1st vs. 4th.

The above similarities of consonants, vowels, and tones are then averaged to yield the final pronunciation similarity. Take the two characters "生" (ㄕㄥ) and "身" (ㄕㄣ) for example. 1) Both initial consonants is "ㄕ" ; 2) both do not contain prenuclear glides; 3) the vowels "ㄥ" and "ㄣ" belong to the same type (i.e., the nasal finals); and 4) both tones are 1st tone. So the pronunciation similarity of "生" and "身" is 0.875, which is the average of similarities: 1, 1, 0.5 and 1.

Finally, to construct a confusion set, each candidate character is used to measure the confused degree with the seed character. Take a character "買" for example, the confused degree with "賣" is 14.59, in which it is calculated by the sum of real-error frequency 13 times, shape similarity 0.66 (exceeds the threshold of 0.5) and pronunciation similarity 0.93 (exceeds the threshold of 0.8), i.e., 13+0.66+0.93=14.59. Finally, as an indexed seed character "買", only a maximum of top 20 confused characters "賣, 乃, 奶, …" ordering by the confused degree are included in our confusion set.

3. Experiments and Evaluation Results

In this section, we evaluate our confusion set, compared with the confusion set provided by Liu et al. (2011) which is the only one we could find. The experimental data came from the SIGHAN 2013-2015 bakeoffs (Tseng et al., 2015; Wu et al., 2013; Yu et al., 2014), including 2,773 Chinese spelling errors. The objective is to measure the effectiveness of confusion sets after erroneous characters have been identified. For this purpose, Mean Reciprocal Rank (MRR), which is the average of reciprocal positions of correct characters in the descending-ordered list of confused characters, is adopted as an evaluation metric. The only previously existing confusion set (Liu et al., 2011) that contains confused characters generating from the combinations of the same/similar in shape/pronunciation/tone/radical/stroke are used for comparisons. Table 2 shows the evaluation results. Our confusion set achieved the best MRR of 0.4184, which significantly outperforms the existing set no matter which similarity measurement is used. This evaluation result supports us to apply the confusion set built by ourselves rather than using the existing one, so that we achieved the second prize in the 2018 Chinese Language Teaching Application Software Competition.

Table 2. *Evaluation on Confusion Sets for Chinese Spell Checking*

	Our Confusion Set	The Existing Confusion Set (Liu et al., 2011)					
		Similar Shape	Same Pron. Same Tone	Same Pron. Diff. Tone	Similar Pron. Same Tone	Similar Pron. Diff. Tone	Same Radical Same Stroke
MRR	*0.4184*	*0.1277*	*0.1989*	*0.0835*	*0.0078*	*0.0093*	*0.0131*

4. Conclusions

This study describes a new confusion set constructed for Chinese spell checking. Base on the experimental results, our set performs better than the previous one. Future work may re-rank the confused candidate list based on the contextual information to achieve better MRR performance.

Acknowledgements

This study was partially supported by the Ministry of Science and Technology, under the grant MOST 106-2221-E-003-030-MY2, 107-2221-E-003-014-MY2, 108-2218-E-008-017-MY3 and 108-2634-F-002-022.

References

Liu, C.-L., Lai, M-H., Tien, K.-W., Chuang, Y.-H., Wu, S.-H., & Lee, C.-Y. (2011). Visually and phonologically similar characters in incorrect Chinese words: analysis, identification, and applications. ACM Transactions on Asian Language Information Processing, 10(2), Article 10.

Tseng, Y.-H., Lee, L.-H., Chang, L.-P., & Chen, H.-H. (2015). Introduction to SIGHAN 2015 bake-off for Chinese spelling check. *Proceedings of SIGHAN'15* (pp. 32-37). Beijing, China: ACL Anthology.

Wu, S.-H., Liu, C.-L., & Lee, L.-H. (2013). Chinese spelling check evaluation at SIGHAN bake-off 2013. *Proceedings of SIGHAN'13* (pp. 35-42). Nagoya, Japan: ACL Anthology.

Yu, L.-C., Lee, L.-H., Tseng, Y.-H., & Chen, H.-H. (2014). Overview of SIGHAN 2014 bake-off for Chinese spelling check. *Proceedings of CLP'14* (pp. 126-132). Wuhan, China: ACL Anthology.