LETTER

Three-Phase Text Error Correction Model for Korean SMS Messages

Jeunghyun BYUN†, *Nonmember*, So-Young PARK††, *Member*, Seung-Wook LEE†, *Nonmember*, *and* Hae-Chang RIM†a), *Member*

SUMMARY In this paper, we propose a three-phase text error correction model consisting of a word spacing error correction phase, a syllable-based spelling error correction phase, and a word-based spelling error correction phase. In order to reduce the text error correction complexity, the proposed model corrects text errors step by step. With the aim of correcting word spacing errors, spelling errors, and mixed errors in SMS messages, the proposed model tries to separately manage the word spacing error correction phase and the spelling error correction phase. For the purpose of utilizing both the syllable-based approach covering various errors and the word-based approach correcting some specific errors accurately, the proposed model subdivides the spelling error correction phase into the syllable-based phase and the word-based phase. Experimental results show that the proposed model can improve the performance by solving the text error correction problem based on the divide-and-conquer strategy.

key words: *text error correction, word spacing errors, spelling errors, SMS messages*

1. Introduction

Text error correction is an essential operation not only for improving text readability but also for obtaining high performance in natural language processing techniques such as part-of-speech tagging, information extraction, and document classification. Most of these techniques are developed under the assumption that input texts do not contain any errors, yet many texts actually have word spacing errors and spelling errors. Furthermore, colloquial style texts such as e-mails, SMS messages, and blogs contain more errors because authors sometimes prefer funny and informal expressions to formal and correct expressions. Thus, text error correction becomes more important than ever before.

However, text error correction is a difficult task because a text can include complex errors. For example, the erroneous text ‘lemme c’ corresponding to the correct text ‘let me see’, contains two spelling errors, ‘lem’ (let) and ‘c’ (see), and a word spacing error, ‘lemme’ (let me). Especially, the word ‘lemme’ is very difficult to be corrected because it contains a spacing and spelling mixed error with a noisy context ‘c’.

Nevertheless, most previous text error correction approaches just focus on either word spacing error correction [1]–[3] or spelling error correction [4]–[8]. Although these approaches can show high accuracy in each field without any counterpart errors, they can not handle any spacing and spelling mixed errors such as ‘lemme’ (let me) at all.

Recently, some approaches have corrected both word spacing errors and spelling errors [9], [10]. Still, a phrase-based model [9] suffers from sparse data because a phrase-based rule $ya \rightarrow you$ extracted from a pair of an incorrect text ‘c **ya**’ and its correct text ‘see **you**’, cannot be applied to a phrase ‘very **ya**ng’ to produce a correct text ‘very **you**ng’ since it is not character-based.

On the contrary, a character-based model [10] with $ya \rightarrow you$ can be applied to ‘very **ya**ng’ (very **you**ng), but the rule incorrectly changes correct words such as ‘**ya**cht’ and ‘Chechn**ya**’ into ‘**you**cht’ and ‘Chechn**you**’. Furthermore, the model suffers from computational complexity because it generates too many candidates based on both word spacing errors and spelling errors at the same time.

In this paper, we propose a three-phase text error correction model based on the divide-and-conquer strategy. First, we try to solve the complex mixed error correction problem by utilizing different statistics and contexts in the spacing error correction phase and the spelling error correction phase. In order to do that, we have separately constructed word spacing error corrected corpus and fully corrected corpus. Furthermore, the spelling error correction phase is divided into the syllable-based spelling error correction phase and the word-based spelling error correction phase. By considering two different correction units, we try to cover many spelling errors and to improve accuracy. We expect that the proposed three-phase error correction model can solve the complex mixed error correction problem step by step.

2. Proposed Model

The proposed three-phase text error correction model consists of a word spacing error correction phase, a syllable-based spelling error correction phase, and a word-based spelling error correction phase as described in Fig. 1. Given a user input sentence ‘lemme c’, the word spacing error correction phase changes it into ‘lem me c’ to correct an word spacing error. And then, the syllable-based spelling error correction phase changes the incorrect part ‘m’ into the correct part ‘t’ in the word ‘lem’. Finally, the word-based spelling error correction phase changes the incorrect word ‘c’ into the correct word ‘see’.

As shown in Fig. 1, each phase is learned from a train-

Manuscript received July 14, 2008.
Manuscript revised December 22, 2008.

†The authors are with Korea University, Korea.
††The author is with Sang-Myung University, Korea.
a) E-mail: rim@nlp.korea.ac.kr

DOI: 10.1587/transinf.E92.D.1213

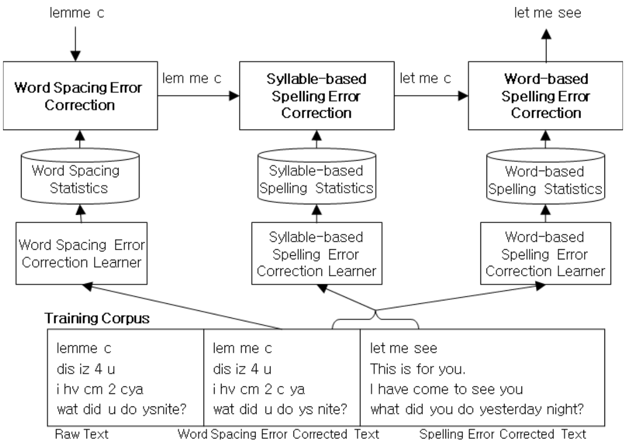
lemme c
let me see
Word Spacing Error Correction
lem me c
Syllable-based Spelling Error Correction
let me c
Word-based Spelling Error Correction
Word Spacing Statistics
Syllable-based Spelling Statistics
Word-based Spelling Statistics
Word Spacing Error Correction Learner
Syllable-based Spelling Error Correction Learner
Word-based Spelling Error Correction Learner
Training Corpus
lemme c
dis iz 4 u
i hv cm 2 cya
wat did u do ysnite?
lem me c
dis iz 4 u
i hv cm 2 c ya
wat did u do ys nite?
let me see
This is for you.
I have come to see you
what did you do yesterday night?
Raw Text
Word Spacing Error Corrected Text
Spelling Error Corrected Text

Fig. 1 Three-phase text error correction model.

ing corpus. The corpus consists of many triples of a raw sentence, a word spacing error corrected sentence, and a spelling error corrected sentence where the word spacing error corrected sentence may include some spelling errors.

The word spacing error correction phase is trained from the word spacing error corrected corpus. And then, both of two spelling error correction phases are trained from the differences between the word spacing error corrected corpus and the spelling error corrected corpus as represented in Fig. 1.

2.1 Three-Phase Text Error Correction Model

Given a user input sentence S^0 such as ‘lemme c’, the error correction problem can be defined as the task of selecting the right candidate S^3 such as ‘let me see’ based on its probability estimation among every possible generated candidate as described in the Eq. (1).

$$\underset{S^3}{argmax}\, P(\, S^3 \mid S^0\,) \tag{1}$$

In order to represent the proposed three-phase model to generate a corrected sentence per phase, the Eq. (1) can include two intermediate corrected candidate sentences, S^1 such as ‘lem me c’ and S^2 such as ‘let me c’ as shown in the Eq. (2). And then, the Eq. (3) generalizes multiple events by the chain rule and the assumption that the i-th corrected sentence S^i only depends on its previous corrected sentence S^{i-1}.

$$\underset{S^3}{argmax}\, P(\, S^3, S^2, S^1 \mid S^0\,) \tag{2}$$

$$\approx \underset{S^3}{argmax} \prod_{i=1}^{3} P(\, S^i \mid S^{i-1}\,) \tag{3}$$

For the purpose of reducing the complexity of the error correction problem, each phase generates its own best corrected sentence S^i as shown in the Eq. (4). Ultimately, the final corrected sentence S^3 is produced based on these best corrected sentences, S^2 and S^1.

$$\underset{S^i}{argmax}\, P(\, S^i \mid S^{i-1}\,) \tag{4}$$

$$= \underset{f^i_{1n}}{argmax}\, P(\, f^i_1, f^i_2, \cdots, f^i_n \mid f^{i-1}_1, f^{i-1}_2, \cdots, f^{i-1}_n) \tag{5}$$

$$= \underset{f^i_{1n}}{argmax} \prod_{j=1}^{n} P(\, f^i_j \mid f^{i-1}_{1n},\ f^i_{1j-1}\,) \tag{6}$$

In order to alleviate the sparse data problem, the i-th corrected sentence S^i is replaced by a sequence of sentence fractions as described in the Eq. (5) where f^i_j indicates the j-th sentence fraction in the sentence S^i, and f^{i-1}_j indicates the sentence fraction corresponding to f^i_j in the sentence S^{i-1}.

2.2 Word Spacing Error Correction Phase

With the aim of applying the Eq. (6) to the word spacing error correction problem finding a sequence of appropriate word spaces from a sequence of the given syllables, the sentence fraction f^i_j can be replaced by a pair of a syllable s^i_j and a word spacing tag t^i_j. Particularly, the word spacing tag t^i_j indicates whether a word space exists between the syllable s^i_j and the next syllable s^i_{j+1} or not. Furthermore, the word spacing error correction phase [1] focuses on finding an unknown sequence of word spacing tags rather than a sequence of the given syllables as described in the Eq. (7); because s^{i-1}_{1n} is the same as s^i_{1n}, and t^{i-1}_{1n} is disregarded as described in the Eq. (8). Finally, the Eq. (8) is derived by applying the chain rule and the independence assumption.

$$\underset{f^i_{1n}}{argmax} \prod_{j=1}^{n} P(\, f^i_j \mid f^{i-1}_{1n},\ f^i_{1j-1}\,)$$

$$= \underset{t^i_{1n}}{argmax} \prod_{j=1}^{n} P(\, s^i_j,\ t^i_j \mid s^{i-1}_{1n},\ t^{i-1}_{1n},\ s^i_{1j-1},\ t^i_{1j-1}\,) \tag{7}$$

$$\approx \underset{t^i_{1n}}{argmax} \prod_{j=1}^{n} \{P(\, t^i_j \mid t^i_{j-1},\ s^i_{j-1}) \times P(\, s^i_j \mid t^i_j,\ s^i_{j-1})\} \tag{8}$$

For example, a word spacing error corrected candidate sentence ‘lem me c’ can be represented as a sequence of syllables, ‘lemmec’, and a sequence of word spacing tags, ‘$\phi\phi$␣ϕ␣␣’ where a word spacing tag ‘ϕ’ indicates no word space, and a word spacing tag ‘␣’ indicates a word space. Given j indicating 3, the probabilistic term $P(\, t^i_3 \mid t^i_2,\ s^i_2) \times P(\, s^i_3 \mid t^i_3,\ s^i_2)$ becomes $P($ ␣ $\mid \phi,\ e) \times P(\, m \mid$ ␣$,\ e)$.

2.3 Syllable-Based Spelling Error Correction Phase

To get a correctly revised sentence from a given sentence, the syllable-based spelling error correction phase [4] replaces the sentence fraction f^i_j into few syllables. As shown in the Eq. (9), the phase is simplified according to the assumption that a sentence fraction f^i_j only depends on the sentence fraction f^{i-1}_j, its left three syllables ($f^{i-1}_{j-3}, f^{i-1}_{j-2}$, and f^{i-1}_{j-1}), and its right three syllables ($f^{i-1}_{j+1}, f^{i-1}_{j+2}$, and f^{i-1}_{j+3}).

$$\underset{f_{1n}^{i}}{argmax} \prod_{j=1}^{n} P(\ f_j^i \mid f_{1n}^{i-1},\ f_{1\,j-1}^{i}\ )$$
$$\approx \underset{f_{1n}^{i}}{argmax} \prod_{j=1}^{n} P(\ f_j^i \mid f_{j-3}^{i-1}, f_{j-2}^{i-1}, f_{j-1}^{i-1}, f_{j}^{i-1}, f_{j+1}^{i-1}, f_{j+2}^{i-1}, f_{j+3}^{i-1}) \tag{9}$$

When the syllable-based spelling error correction phase generates a candidate sentence ‘let me c’ from a given sentence ‘lem me c’, we can assume that f_1^{i-1}, f_2^{i-1}, $\cdots$, and f_8^{i-1} indicate ‘l’, ‘e’, ‘m’, ‘␣’, ‘m’, ‘e’, ‘␣’, and ‘c’ respectively. Given $j = 3$, the probabilistic term $P(\ f_3^i \mid f_0^{i-1}, f_1^{i-1},\ f_2^{i-1},\ f_3^{i-1},\ f_4^{i-1},\ f_5^{i-1},\ f_6^{i-1}\ )$ becomes $P(\ t \mid \phi, l, e, m, \text{␣}, m, e\ )$ where ϕ indicates no syllable symbol.

This phase can perform insertion, deletion or m-to-n substitution operations because f_j^i and f_j^{i-1} can become no syllable, one syllable, or a few syllables. For example, the probabilistic term $P(duce|t, r, o, \phi, \text{␣}, 2, u)$ is applied to generate a sentence “lemme intoduce 2u” from a given sentence “lemme into 2u” by inserting ‘duce’. Also, the probabilistic term $P(\phi|\ g, o, o, oo, d, \text{␣}, d)$ is applied to generate a sentence “good day!” from a given sentence “gooood day!” by deleting ‘oo’. Besides, the probabilistic term $P(th|r, d, \text{␣}, d, a, t, \phi)$ is applied to generate a sentence “i heard that” from a given sentence “i heard dat” by substituting ‘th’ for ‘d’.

2.4 Word-Based Spelling Error Correction Phase

As shown in the Eq. (10), the word-based spelling error correction phase replaces the sentence fraction f_j^i into a word w_j^i. Like the syllable-based spelling error correction phase, the phase is simplified according to the assumption that a word w_j^i only depends on the word w_j^{i-1}, its left one word w_{j-1}^{i-1}, and its right one word w_{j+1}^{i-1}.

$$\underset{f_{1n}^{i}}{argmax} \prod_{j=1}^{n} P(\ f_j^i \mid f_{1n}^{i-1},\ f_{1\,j-1}^{i}\ )$$
$$\approx \underset{w_{1n}^{i}}{argmax} \prod_{j=1}^{n} P(\ w_j^i \mid w_{j-1}^{i-1}, w_{j}^{i-1}, w_{j+1}^{i-1}\ ) \tag{10}$$

For example, the word-based spelling error correction phase can generate a candidate sentence “let me see” from a given sentence “let me c”. Given $j = 3$, the probabilistic term $P(\ w_3^i \mid w_2^{i-1},\ w_3^{i-1},\ w_4^{i-1}\ )$ becomes $P(\ see \mid me, c, \phi\ )$ where ϕ indicates no syllable symbol.

3. Experiments

For the purpose of examining the error correction performance of the proposed model, we have tested the model on a Korean SMS corpus which is divided into 90% for the training set and 10% for the test set. The corpus consists of 109,084 triples of a raw sentence, a word spacing error corrected sentence with some spelling errors, and a spelling error corrected sentence without any word spacing error where a sentence indicates a message. In the corpus, a fully corrected message is composed of 3.15 words and 11.98 syllables in average while a raw sentence consists of 1.46 words and 7.46 syllables in average.

Furthermore, we utilize the following five performance measures in order to analyze each phase’s precision in detail. First, *ic* indicates the correction ratio of the number of words, which are changed from “incorrect” to “correct”, to the total number of words. Second, *cc* indicates the ratio of the number of correct words which are not changed, to the total number of words. Third, *ii* indicates the ratio of the number of incorrect words which are not changed, to the total number of words. Fourth, *ii** indicates the ratio of the number of incorrect words, which are changed from “incorrect” to “other incorrect”, to the total number of words. Fifth, *ci* indicates the ratio of the number of words, which are changed from “correct” to “incorrect”, to the total number of words. Furthermore, we also use *accuracy* indicating the ratio of correct constituents in the total constituents after the input constituents are processed by the correction model.

3.1 Performance of Phase Combinations

In order to evaluate the correction effect of combinations, we try to combine the word s**p**acing error correction phase(p), the **s**yllable-based spelling error correction phase(s), and the **w**ord-based spelling error correction phase(w) as shown in the left column of Table 1. Particularly, the combined order is represented as the sequence of p, s, and w such as *ps*, *pw*, *sp*, *wp*, *psw*, and *pws*. For example, *pw* indicates to perform the word s**p**acing error correction phase before the **w**ord-based spelling error correction phase.

Table 1 shows that the *raw* text in the corpus is very erroneous. 83.36% raw words include more than one error. Among the words of the fully corrected corpus, only 7.2% words exist in the raw text. Since most errors are related to word spacing errors, the word spacing error correction phase improves 57.92% *word*-unit *accuracy*, and this improvement is much better than the spelling correction phase. It is also noticeable that the rates of *cc* and *ii* related to the not-changed-words are very high in the word-based spelling error correction phase(i.e. w) because the spelling corrector precisely corrects errors in this phase. On the contrary, since the syllable-based phase actively corrects errors, its *ic*,

Table 1 Performance of phase combinations.

	word						*sentence*
	ic	*cc*	*ii*	*ii**	*ci*	*accuracy*	*accuracy*
raw	·	16.64	**83.36**	·	·	**7.20**	1.16
p	58.40	6.74	5.51	27.59	1.76	**65.12**	27.97
s	**8.62**	15.74	43.24	**31.95**	**0.46**	10.80	6.61
w	7.87	**16.03**	**73.17**	2.74	0.19	10.69	5.79
ps	79.45	6.69	1.04	**11.01**	1.82	**86.09**	**72.02**
pw	79.38	6.74	2.02	10.10	1.76	86.09	71.57
sp	65.59	6.64	1.73	23.07	2.96	68.56	51.71
wp	54.91	6.96	5.55	29.55	3.03	55.89	33.10
psw	79.90	6.68	0.90	**10.70**	1.82	86.55	**73.10**
pws	79.65	6.72	1.15	10.70	1.79	86.33	72.28

ii*, and *ci*, related to the changed-words, are higher than the word-based phase.

Moreover, the spelling correction phase as the second phase improves about 20% *word*-unit *accuracy* by correcting spelling errors that the previous word spacing error correction phase did not revise. As shown in Table 1, the models *ps* and *pw* have higher accuracy than the models *sp* and *wp*. This is because the spacing correction phase is applied before the spelling correction phase. It is also remarkable that the *sentence*-unit *accuracy* is improved about 1.1% by utilizing two different spelling error correction phase step by step. We assume that different kinds of errors are corrected by applying different units of correction rules. Besides, we found that the model *psw* corrects some errors which cannot be corrected by the model *ps*.

3.2 Comparison with Previous Models

For the comparison with previous models in the same test environment, we have reimplemented a syllable-based spelling error correction model (*Byun 2007*) [4], a word spacing error correction model (*Lee 2007*) [1], a simultaneous text error correction model (*Noh 2007*) [10], a phrase-based statistical model (*Aw 2006*) [9], and a model combined (*Lee 2007*) and (*Aw 2006*). Then these models are applied to the same Korean SMS corpus as shown in Table 2.

Table 2 shows that (*Noh 2007*) improves by over 20% *sentence*-unit *accuracy* by correcting both spelling errors and word spacing errors as compared with either the spelling error correction model (*Byun 2007*) or the word spacing error correction model (*Lee 2007*). As compared with the phrase-based model (*Aw 2006*) designed for English SMS text with some word spaces, (*Lee 2007*), one of the best word spacing error correction models, is more suitable for the Korean SMS text with very few word spaces because the phrase-based model is too difficult to correct errors without phrases divided by word spaces.

It is noticeable that the error correcting performance is improved by combining two models such as (*Lee 2007*) + (*Aw 2006*). As compared with (*Noh 2007*), (*Lee 2007*) + (*Aw 2006*) shows higher accuracy even though (*Aw 2006*) shows much lower accuracy. It shows that a step by step model such as (*Lee 2007*) + (*Aw 2006*) or the proposed model is better than a simultaneous model such as either (*Noh 2007*) or (*Aw 2006*) because the simultaneous model suffers from computational complexity by generating too many candidates composing of all kinds of errors at the same time. Furthermore, the proposed model shows higher accuracy than (*Lee 2007*) + (*Aw 2006*) because each phase in the proposed model focuses on its own error type unlike (*Aw 2006*). It shows that the divide-and-conquer strategy is quite effective for improving the performance.

Table 2 Comparison with previous models.

	accuracy (*word*)	*accuracy* (*sentence*)
Raw text	7.20	1.16
(*Byun 2007*)	10.80	6.61
(*Lee 2007*)	65.12	27.97
(*Noh 2007*)	70.29	49.02
(*Aw 2006*)	32.00	12.78
(*Lee 2007*) + (*Aw 2006*)	78.98	53.34
Proposed Model	86.55	73.10

4. Conclusion

In this paper, we propose a text error correction model consisting of a word spacing error correction phase, a syllable-based spelling error correction phase, and a word-based spelling error correction phase. The proposed model has the following characteristics.

First, the proposed model can reduce the text error correction complexity by solving the text error correction problem step by step based on the divide-and-conquer strategy.

Second, the proposed model can handle word spacing errors, spelling errors, and the complex mixed errors. As compared with a model correcting either word spacing errors or spelling errors, the proposed model improves accuracy by over 20% with a word-unit by correcting both word spacing errors and spelling errors.

Third, a compared to the model with a single spelling error correction phase, the proposed model can improve the sentence-unit accuracy by utilizing two different spelling correction phases step by step.

Acknowledgement

This work was supported by the Second Brain Korea 21 Project and WCU (World Class University) program through the Korea Science and Engineering Foundation funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0).

References

[1] D.G. Lee, H.C. Rim, and D. Yook, "Automatic word spacing using probabilistic models based on character n-grams," IEEE Intelligent Systems, vol.22, no.1, pp.28–35, 2007.

[2] S.S. Kang, "Eojeol-block bidirectional algorithm for automatic word. spacing of hangul sentences," J. Korea Information Science Society, vol.27, no.4, pp.441–447, 2000.

[3] J.H. Choi, "Automatic korean spacing words correction system with bidirectional longest match strategy," Proc. 9th Conference of hangul and Korean Information Processing, pp.145–151, Korean Information Science Society, 1997.

[4] J. Byun, S.Y. Park, and H.C. Rim, "Automatic spelling correction rule extraction and application for spoken-style korean text," Proc. 6th International Conference on Advanced Language Processing and Web Information Technology, pp.195–199, 2007.

[5] Y.S. Lee, Y.J. Park, and M. suk Song, "Spelling correction in korean using the 'eojeol' generation dictionary," KIPS Journal, vol.8, no.1, pp.98–104, Feb. 2001.

[6] K. Toutanova and R.C. Moore, "Pronunciation modeling for improved spelling correction," ACL '02: Proc. 40th Annual Meeting on Association for Computational Linguistics, pp.144–151, Association for Computational Linguistics, Morristown, NJ, USA, 2001.

[7] E. Brill and R.C. Moore, "An improved error model for noisy channel spelling correction," ACL '00: Proc. 38th Annual Meeting on

Association for Computational Linguistics, pp.286–293, Association for Computational Linguistics, Morristown, NJ, USA, 2000.

[8] H. Lim and U. Kim, “A spelling correction system based on statistical data of spelling errors,” KIPS Journal, vol.2, no.6, pp.839–846, Nov. 1995.

[9] A. Aw, M. Zhang, J. Xiao, and J. Su, “A phrase-based statistical model for sms text normalization,” Proc. COLING/ACL 2006 Main Conference Poster Sessions, pp.33–40, Association for Computational Linguistics, Sydney, Australia, July 2006.

[10] H. Noh, J. Cha, and G.G. Lee, “A joint statistical model for word spacing and spelling error correction simultaneously,” J. Korea Information Science Society, vol.34, no.2, pp.131–139, 2007.