Improvement of Korean Proofreading System Using Corpus and Collocation Rules

Chae Young-Soog[*]

Systems Engineering Research Institute

This paper presents the techniques of correcting for spelling errors, orthographical errors, and grammatical errors in computer-based text. And this paper addresses an extension that goes beyond normal checking of isolated single word by taking multi-words as well as a sentence. The candidate words for spelling errors are created by applying function of rules and correction rule table that contains heuristic information of collocation. To prevent excessive creation of candidate words and improve accuracy, we use the high frequency word dictionary that contains 300,000 words derived from corpus. For constituent errors, by applying grammar based partial parsing rules, collocation words errors between the words can be found. We make an experiment with correction techniques on corpora that are the final result of SERI's research, texts, newspaper materials, and public materials. The system has 98% accuracy rate when the 8.5% errors caused by unregistered words were excluded. The average number of prospective candidates suggested by the system is 1.12.

I. Introduction

With the increase use of word processing systems, the speller become their essential function. Spellers for English detect not only errors in word level but also the grammar and style errors in sentence level. There are a number of programs that perform some kind of grammar and style checking on the software market. While there are many systems developed for English, there has been not a helpful proofreading software for Korean yet. Now, some Korean spellers are being published.

This paper describes the improvement of a Hangul Speller that consists of spell, grammar, and style checker and corrector with a help message for errors. By collecting error types found in documents, this Korean Proofreading System classifies the errors by each cause. It is not only processing errors for orthographical rules and spelling, but also it uses a sentence and writing style as units for finding out errors in element of sentences. That is, in this system, the errors are classified into a category of orthographical rules, proper spacing, spelling errors including the changes in phonemes in a word, and errors in a sentence and writing system among words. According to the classified fault types, the expanded rules and knowledge base are established as a scope of consideration by using a morpheme and sentence as a unit. They increase the ways to improve the accuracy of clauses, and minimize the number of candidates. In addition, it is designed in a way that can suggest prospective clauses for errors found in sentence elements and writing style created by collocations of words.

The system uses different contrasting method for finding out standard language errors according to misused part of speech, and it executes morphemic changes simultaneously. The candidate words for spell, grammar, and style errors are created by applying function of rules and correction rule table that contains heuristic information. To prevent excessive creation of candidate words and improve accuracy, we use the high frequency word dictionary that contains 300,000 words derived from corpus and part of speech pattern. For constituent errors, by applying grammar based partial parsing rules, collocation words errors between the words can be found.

II. Characteristics of Korean

The writing tool of Korean language has two sets of characters--Hangul and Hanja. Hangul is

[*] Natural Language Information Processing Department, Systems Engineering Research Institute, yschae@seri.re.kr

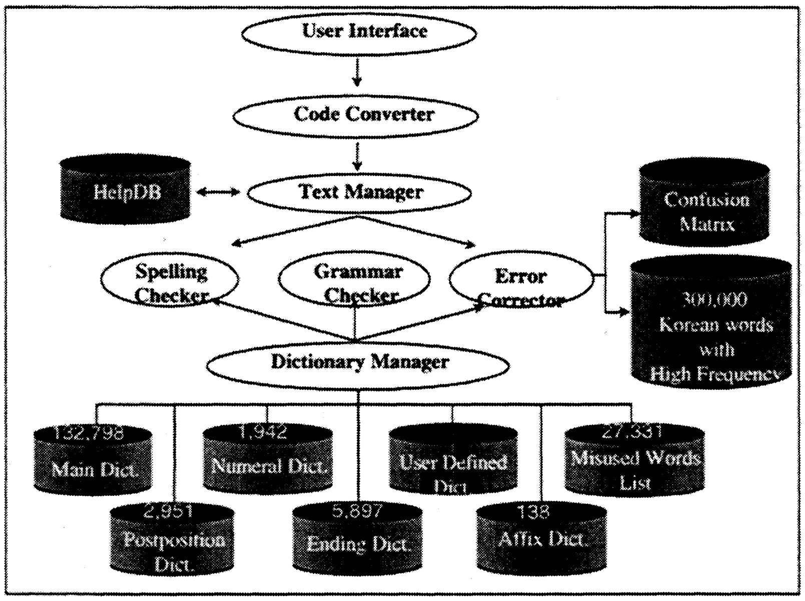
the Korean Alphabets which can make 11,720 syllables. A syllable consists of three phonemic symbols that are initial consonant, medial vowel, and final consonant. Hanja (Chinese Characters used in Korea) character set contains several tens of thousands of characters, but only 2,884 characters are in common use. Hangul appears in Korean texts with Hanja and English alphabet. So we need to separate modules to check and to correct erroneous words by different character sets.

In linguistic genius aspect, Korean is well known as an agglutinative language. A word can be constructed by a content part and an inflectional part. In Korean, a word means a spacing unit. A content part can have multiple parts of speech(noun, proper noun, verb, and so on) and an inflectional part can consist of multiple function morphemes like postpositions(the nominal suffix compounds) or endings (verbal suffix compounds). So, more than 10 parts of speech (morphemes) can make a word and a verb can conjugate more than 20 thousands different forms. Much the worse, the irregular verbs and adjectives make the morphological analyses of Korean more difficult.

III. System Overview

We improve this system by following steps. Firstly, we classified the error patterns by surveying errors in various types if texts such as textbooks, editorials, essays, newspapers, etc. Secondly, we make confusion sets where each confusion set has words or parts of speech that would be frequently confused wth each other. The patterns of phrase related to syntactic and stylistic errors are also classified. Thirdly, we construct the semantic-syntactic rules to detect and correct errors by a confusion set. The rules related to the patterns of error phrases are coded. Finally, we add detail help messages to the help message dictionary to address reasons of errors.

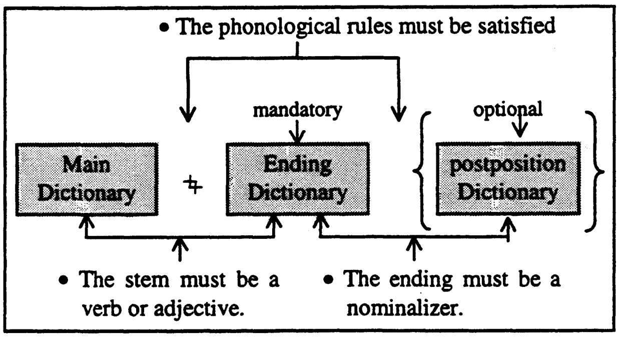
Our system consists of seven modules(User Interface, Code Converter, Text Manager, Spelling Checker, Grammar Checker, Spelling Corrector, Dictionary Manager), Help DB, Confusion Matrix, Frequency Word List, and seven dictionaries. Input is a Korean text mixed with Korean, English, and Chinese characters. The system can analyze a text by not only an interactive mode but also a batch mode. In the interactive mode, a user can refer to a help window to correct errors. However in the batch mode, he can not do it. But we support the position where the error occurs and its candidate words.



[Fig. 1] The Architecture of Speller

- Spelling Checker

The spelling checker detects the erroneous words in a text. This module is built upon a morphological analyzer, which can analyze a content part and inflectional part. This module receives a word separated from Text Manager and divides it into morphemes using dictionaries. Our system follows the dictionary-based approach. That is, cascading dictionaries control the basic morphotatics of a Korean word. [Fig. 2] shows an example.

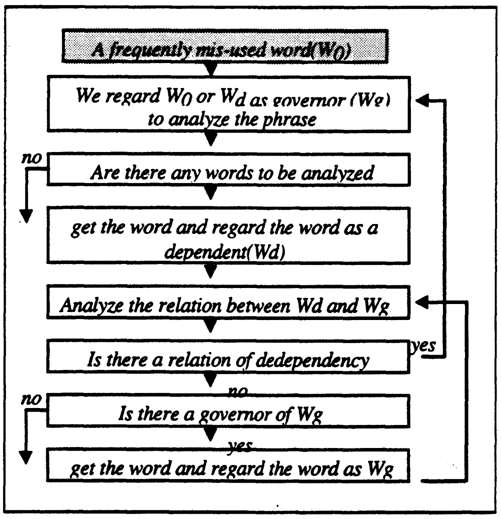

• The phonological rules must be satisfied
mandatory
optional
Main Dictionary
+
Ending Dictionary
postposition Dictionary
• The stem must be a verb or adjective.
• The ending must be a nominalizer.

[Fig. 2] An Example of Cascading of Dictionaries

First of all, the main dictionary has all the conjugated forms of irregular verbs and adjectives. The ending dictionary and the postposition dictionary have all different forms of endings and postpositions too. The decision whether a verb and an ending can combine is decided by the phonological rules. So, our system does not have any routine to deal with the irregular verbs or adjectives.

Additional morphotatic constraints are also applied to constrain the combination of morphemes more elaborately. One of them is that the postpositions can be attached to the ending when the ending is a nominalizer. Another rule is that the ending "*neun*[adnominalizer]" can only combine with a verb not with an adjective. Our system has many specific rules for morphotactics and these rules are being continuously extended to elaborate the system. The dictionary-based approach used in our system makes our system both very efficient in processing and easy to codify new constraints.

- Grammar Checker


A frequently mis-used word(W0)
We regard W0 or Wd as governor (Wg) to analyze the phrase
no
Are there any words to be analyzed
get the word and regard the word as a dependent(Wd)
Analyze the relation between Wd and Wg
yes
Is there a relation of dedependency
no
no
Is there a governor of Wg
yes
get the word and regard the word as Wg

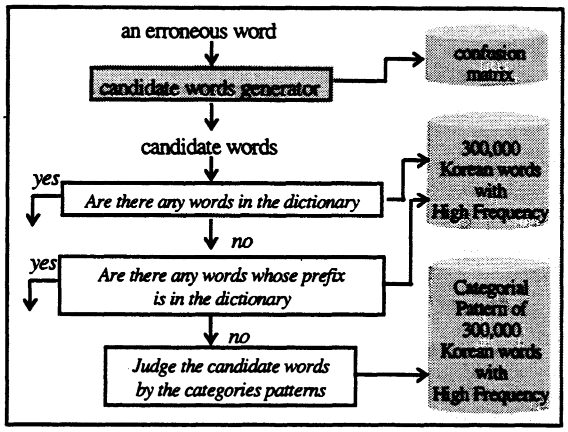
When a frequently mis-used word with grammar errores appears in a text, our system checks that the word is correct or not. One clue about the identity of a frequently mis-used word comes from the words around it and context of the sentence. If an error occurs, we select the governor of a phrase' to check the grammatical errors and to correct them. And the system suggests a correct word and gives a help message. To improve the processing speed, we used demon-based programming where each entry in a dictionary has rule names related to that entry. We use to analyze a sentence is based on the dependency grammar. The dependency grammar is widely used for analyzing Korean language because it is a partially free word language.

[Fig. 3] The Process of Partial Parsing

[Wd : the word that we regard as the dependent in a phrase, Wg : the word that we regard as the governor in a phrase

Wo : the frequently mis-used word]

- Spelling Corrector


an erroneous word
candidate words generator
confusion matrix
candidate words
yes
Are there any words in the dictionary
300,000 Korean words with High Frequency
no
yes
Are there any words whose prefix is in the dictionary
no
Judge the candidate words by the categories patterns
Categorial Pattern of 300,000 Korean words with High Frequency

[Fig. 4] Selection of candidate words using data from a corpus

The spelling corrector tries to find the most likely correct word and suggests the candidate words and a help to aid the user to know the reason of errors and give a hint for correcting errors. We classify the spelling errors into three types: orthographic errors, typographical errors, and errors caused by the similarity of pronunciation. The orthographic errors include the errors using non-standard words or dialects, errors violating spacing rules, errors violating phonological rules, etc. The errors using non-standard word or dialects are corrected by listing all the pairs of the non-standard part of speech and standard part of speech in the dictionary. The incorrect forms that are frequently violated are also listed in the dictionary. The other orthographic errors are corrected by special rules that codify the error patterns and their correcting rules.

The typing errors and errors caused by the similarity of pronunciation of characters are corrected by using the confusion matrix. The confusion matrix has pairs of a character string that is frequently in an incorrect word and it's original character string.

One serious problem is that the number of candidate words corrected by the confusion matrix is generally too big and some of them are not acceptable as Korean words. The reason to generate unacceptable words is that the spell checking routine is over-accept words. But it is very difficult to minimize the over-acceptance.

To improve the reliability of correction, we use the data from a corpus. From 50 million Korean words, we got about 2 millions and 500 thousands different Korean words. But the 300 thousands different Korean word with high frequency covers 91% of the corpus. We also analyze how the words in the dictionary are composed of with respect to the categories of morphemes in each word.

By using this result, we elaborate the correction by the matrix as follows. First, the candidates that are in the dictionary with 300 thousand words with high frequency have priority. When the system can not find a candidate in the dictionary, the candidate whose prefix is in the dictionary is selected. If the two approaches fail, the categories of morphemes in the candidate words are evaluated.

The performance of the correction varies very much with the priorities of the correction rules. Our system gives priorities as shown in [Table 1].

Rank	Correction Rules
1	Errors by using non-standard words
2	Errors violating orthographic rules except spacing errors
3	Spacing errors
4	Errors caused by the similarity of pronunciation of characters
5	Typing errors

[Table 1] The Priorities of Correction Rules

IV. Evaluation Results

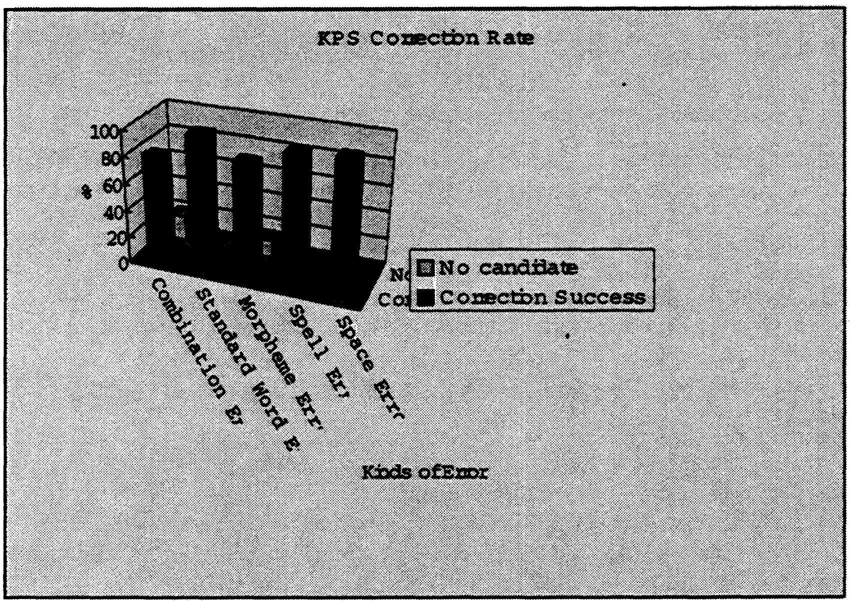
We experiment our system by documents in a corpus with various genres. The documents are selected at random from the valanced corpus which is developed by SERI(Systems Engineering

Research Institute). We make an experiment with correction techniques on corpora that are the final result of SERI's research, texts, newspaper materials, and public materials. The system has 98% accuracy rate when the 8.5% errors caused by unregistered words were excluded. In terms of grammatical errors, the Error Corrector proved 98.5% correction rate for word spacing, correction rate 91.4% for phoneme and syllable changes, and correction rate of 99% for orthographical rules errors. The average number of prospective candidates suggested by the system is 1.12.

Genre	# of words	# of errors	Performance of Corrector(%)		Types of Correcting		
			Precision	recall	I	II	III
Thesis	90,386	77	100	97.85	40	6	14
Novel	69,221	71	98.6	84.5	33	12	26
Text-book	70,841	4	100	100	4	0	0
Article	58,126	93	95.7	95.7	85	1	2
Essay	53,032	30	100	100	14	3	12
Total	341,606	275	99.05	95.98	180	26	62

(I : violating spacing rules, II : using a non-standard word, III : others)

[Table 2] The Result of Correction



[Fig. 5] The Result of Correction per Error Type

V. Conclusion

We have developed a practical Speller for Korean text proofreading. This paper describes the design and implementation of Speller that consists of Spell and Grammar Checker and Spelling Corrector. Spelling Checker detects the erroneous words in a text mixed with Korean, English, and Chinese characters. This is done by cascading dictionaries by Korean morphotatics. Spelling Corrector tries to find the most likely correct word and suggests the candidate examples to substitute and a help message appropriate for the types of error. The spelling checker has about 99% of accuracy to detect spelling errors and the spelling corrector has 96% of correction rate.

We limited our research to word level and phrase level. So we are developing an advanced Speller with a context-sensitive grammar and style checker in a phrase and sentence level. The advanced Speller will check and correct errors caused by syntactic and semantic reasons. And we will suggest more detailed help context which adapts to user model. From now on, more researches on developing techniques that can generalize the established rules, and techniques that can strengthen the sentence analysis will be needed. To strengthen the role of perfect writing support instrument, that supports the user's error finding process, a separate supplementary function will be needed. The future system requires having not only the spell checking and correcting, but also it requires to have a search function for many areas such as synonyms, antonyms, familiarity of words by different age level, the frequency of word usage, meaning of the words, and examples of word usage. In addition, as a writing style search function, the future system is require to have a function that can provide information on sentences that are difficult to comprehend by just reading them.

Reference

[1] N.H. Macdonald, L.T.Frase, P.Gingrich, and S.A.Keenan, "The WRITER'S WORKBENCH : Computer aids for text analysis", *IEEE Trans. Communication*, COMM-30, No. 1, 1982, 105-110,.

[2] G.E.Heidorn, Jensen, L.A. Miller, R.J.Byrd, and M.S.Chodorow, "The EPISTLE Text-Critiquing System:, *IBM System Journal*, 21(3), 1982, 305-326.

[3] Karen Jensen, George E.Heidorn, Stephen D.Richardson, Norman Haas, "PLNLP, PEG, and CRITIQUE : Three Contributions to computing in the humanities", *IBM Research Report*, 1986.

[4] James L. Peterson, "Computer Programs for detecting and Correcting Spelling Errors", *Communications of the ACM*, 23(12), 1980, 676-687.

[5] Koichi Takeda, Emiko Suzuki, Tetsuro Nishino, Tetsunosuke Fujisaki, "CRITAC-An experimental system for Japanese text proofreading", *IBM J. RES. Development*, 32(2), 1988.

[6] Y.J Kim, etc, "A Korean Polishing System by collocation of Words and Partial Parsing", Cognitive Science, 1997.

[7] James J. Peterson, "Computer Programs for Detecting and Correcting Spelling Errors", *Computing Practices*, 23(12), 1980.

[8] Thomas N.Turba, "Checking for Spelling and Typographical Errors in Computer-Based Text", *SIGPLAN-SIGOA*, 1981, 51-60.

[9] J.R.Ullmann, "A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words", Computer Journal, 20(2), 1975 , 155-161.