2012 International Conference on Information and Computer Applications (ICICA 2012)
IPCSIT vol. 24 (2012) © (2012) IACSIT Press, Singapore

Error Decision Rules to Detect Misspelled Predicate

Gil-Ja So [1+] and Hyuk-Chul Kwon [2]

[1] Department of Game Contents, University of Young-San
99 Pilbong-gil Haeundae-Gu, Busan, 612-743, Korea

[2] School of Electrical & Computer Engineering, Pusan National University,
Busandaehak-ro 53 beon-gil, Geumjeong-gu, Busan 609-735 Korea

Abstract. Korean grammar checkers typically detect context-dependent errors by employing heuristic rules that are manually formulated by a language expert. However, such grammar checkers are not consistent. In order to resolve this shortcoming, we propose new method for generalizing error decision rules to detect the above errors. For this purpose, we use an existing thesaurus KorLex, which is the Korean version of Princeton WordNet. Through the Tree Cut Model and the MDL(minimum description length) model based on information theory, we extract noun classes from KorLex and generalize error decision rules from these noun classes. In conclusion, the precision of our grammar checker exceeds that of conventional ones by 6.2%.

Keywords: Grammar Checker, Context dependent error detection, Selectional constraint noun classes, Generalization of an error decision rule, Minimum Description Length

1. Introduction

Grammar checker is the system to detect spell error, syntax error, semantic error in text document. Spell error can be detected using a word, but syntax and semantic error can't. Later errors are called context-sensitive or context-dependent errors. Korean grammar checkers typically detect context-dependent errors by employing heuristic rules that are manually formulated by a language expert[1]. These rules are appended each time a new error pattern is detected. However, such grammar checkers are not consistent. Specially, if error word is a predicate, there can be many rules, depending on the nouns which can be an object or subject of the predicate. In order to resolve this shortcoming, we propose new method for generalizing error decision rules to detect the misspelled predicate. For this purpose, we use the selectional constraints of the predicate. Selectional constrains mean the semantic restrictions that a word imposes on the environment in which it occurs. In case of the predicate, an noun classes of an object or a subject consisting of semantically similar senses can be a selectional constraint. In this paper, selectional constraints are noun classes which can be located as an object or a subject of the predicate.

For this purpose, we use an existing noun thesaurus, KorLex, which is the Korean version of Princeton WordNet. KorLex has hierarchical word senses for nouns, but does not contain any information about the relationships between cases in a sentence. There is a problem how to determine noun classes which is suitable for the selectional constrains of the predicate. Through the Tree Cut Model and the MDL(minimum description length) model based on information theory, we extract noun classes from KorLex and generalize error decision rules from these noun classes. In order to verify the accuracy of the new method in an experiment, we extracted nouns used as an object of the four predicates usually confused from a large corpus, and subsequently extracted noun classes from these nouns. We found that the number of error decision rules generalized from these noun classes has decreased to about 64.8%. In conclusion, the precision of our grammar checker exceeds that of conventional ones by 6.2%.

[+] Corresponding author.
E-mail address: kjso@ysu.ac.kr.

2. Extraction Selectional Constraints of the Predicate from KorLex through the MDL

2.1. KorLex

The KorLex is the large corpora which has been made from 2004, referencing the Princeton WordNet(PWN). Korlex 1.5 contains currently about 130,000 synsets and 150,000 word senses for nouns, verbs, adjectives, adverbs, and classifiers[13]. Concepts in KorLex are represented as sets of synonyms. A word sense in KorLex is a word-concept pairing, e.g. given the concepts bae^1=<*bogbu*("abdomen"), *bae*("belly")>, bae^2=<*sunbak*("vessel"),*bae*("ship")> and bae^3=<*baesu*("times")> we can say that *bae* has three senses, bae^1, bae^2, bae^3. KorLex can represent an unambiguous concept like this[2].

2.2. MDL

The availability of lexical databases, such as WordNet and Euro WordNet, appears to be useful for many different research areas including real word error detection and correction task. Many researchers have proposed various methods about acquiring selectional preferences of verbs. In this paper, we apply MDL-based model which is proposed by Li and Abe[3]. MDL is the statistical inference model to find the regularity of the data set by viewing learning as data compression[4]. MDL model tries to find the best hypothesis H that most compress data set. This idea can be applied to a model selection problem. MDL select the model that can compress the model and the data most. The amount of bits to compress the model itself is the "model description length" and the amount of bits to compress the data set is the "data description length". The total amount of information is the sum of these two.

Total Information(L(TOTAL)) =

Model Description Length(L(MOD)) + Data Description Length(L(DATA))

$$L(MOD) = \frac{k}{2} * \log |S|$$

$$L(DATA) = -\sum_{n \in S} \log P(n|v,r)$$

k is the number of the synsets that are included in the model. S is the argument list. C(s) is the size of S, the frequency of the argument that is founded in the corpus. The model of which value of L(TOT) is low is selected in the model selection process.

2.3. Determining a suitable class for a selectional constraints from KorLex

We adopted the Li and Abe's Tree Cut Model to determine a suitable class for a selectional constraints from KorLex. Li and Abe shows a thesaurus as a tree like a Fig.1[3].

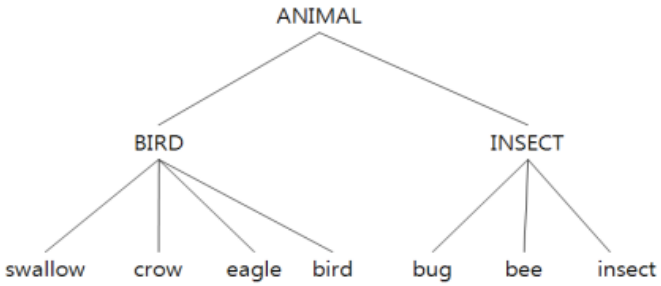

Fig.1 An example thesaurus.

A cut in a tree is any set of nodes in the tree that defines a partition of the leaf nodes. Tree cut model M consists of a tree cut Γ and probabitltiy parameter vector θ. That is: M=(Γ, θ).

$\Gamma=[C_1,C_2,\ldots,C_{k+1}]$, $\theta=[P(C_1|v,r),P(C_2|v,r)\ldots,(C_{k+1}|v,r)]$ Where C_i,(i=1,…,k+1) is a cut, and $P(C_i|v,r)$ is a conditional probability for each C_i, v is verb and r is a argument name[3].

We have extracted object list of the predicate from the corpus. Each object can be mapped to the synset of which concept word is consisted of the lexeme of the object. We can make many tree cut models in the KorLex. We can select a model of which Total Information as described earler is the lowest. The selected model have suitable classes for a selectional constraints of the predicate, given v. Fig.2 shows the algorithms to determine a suitable class for a selectional constraints from a KorLex.

```
function Find_Class(c, node_depth)
   node_depth++
   if(node_depth > leaf_level && is_found_in_corpus(c)) then
      return leaf_node;
   else
      for each hyponym s(i) of c
         ret = Find_Class(s(i), node_depth)
         if ret==LEAF_NODE or ret == GENERALIZE_SUCCESS then
            frequency of c += frequency of s(i) sub_list = append(s(i))
         else if ret == GENERALIZE_FAIL then
            return ret //Generalization fail!!
         endif
      loop
      if LTOT(c) <= LTOT(sub_list) then
         return GENERALIZE_SUCCESS
                                              else
         return GENERALIZE_FAIL
      endif
   endif
end function
```

Fig.2 An alogrithms to determine a selectional constraints

3. Generalizing Error Decision Rules Using Selectional Constraints

To detect a misspelled predicate, decision rules used in the korean grammar checker is like a Fig.3.

P is the erroneous predicate in the sentence. R is the replace word of the predicate P. If the Grammar Checker determines that there is an error in the sentence, P is replaced with the R. Anti Selectional constraints is the argument list which can not be used with the predicate P but can be used with the R.

```
1.  P ga-ri-ki-da ("point out")     // Misspelled predicate
2.  M verb                          // Morphem of predicate
3.  R ga-reu-chi-da ("teach")       // Replace word
4.  D forward                       // Direction to parse
5.  Anti Selectional constraints // anti co-location argument of the predicate P
      ①  A OBJ                      //  Argument to be checked ( subject or object)
      ②  C 6225142                  //  C is a noun class id in the KorLex
```

Fig.3 An error decision rules of predicate.

Fig.3 shows a decision rule to detect an error between "*gareuchida*(to teach)" and "*garikida*(to point)" . These two verbs are frequently misused because of their analogous word spell. Noun class "knowledge", of which KorLex synset id is 6225142, can not be an object of the predicate "*garikida*(to point)". The Grammar checker shall notify the error to the user when theses decision rules are met in a sentence.

4. Experiment

In order to verify the accuracy of the new method in an experiment, we extracted nouns used as an object of the four predicates usually confused from a 37 million words articles, and subsequently extracted noun classes from these nouns. We generalized error decision rules using these noun classes. To evaluate of our new grammar checker, we have executed the grammar checker for three predicates in the large corpora. Precision is a commonly used metric for the evaluation of the grammar checker system. Precision measures the accuracy of retrieval. This is calculated by dividing the number of erros detected by the total of errors. Tab.1 shows the result of the new grammar checker we propose. We compared the precision of the conventional grammar checker with the one of ours. The conventional grammar checker has used a noun instead of noun class as a selectional constraints of the predicate. The precision of our grammar checker exceeds that of conventional ones by 6.2%.

5. Summary

Korean grammar checkers typically detect context-dependent errors by employing heuristic rules that are manually formulated by a language expert.Howerver, this method is not consistent and time consuming. In order to resolve this shortcoming we use the selectional constraints of the predicate, that is, noun classes of

an object or a subject to restrict the means of the predicate. We extracted selectional constraints from existing noun thesaurus, KorLex, which is the Korean version of Princeton WordNet. The decision rules used in the grammar checker is generalized by selectional constraints. In Experiment, the precision of our grammar checker exceeds that of conventional ones by 6.2%.

Tab.1 Evaluation of the new grammar checker

Commonly misspelled Predicate and replace word	**Precision**	
	Conventional grammar checker	**Our new grammar checker**
Garikida / gareuchida (*point out* / *teach*)	84%	93%
Neul-ri-da/neul-i-da (*increase/roll out*)	65.4%	82.8%
Deu-reonaeda/deul-eonaeda (*reveal / catch out*)	92.7%	76.5%
Ma-chi-da / maj-hi-da *(finish / hit)*	65.4%	80.1%
Average	76.87	83.1

6. Acknowledgments

This research is supported by Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Joint Research Center Project 2010.

7. References

[1] M. Y. Kang, A. S. Yoon, H. C. Kwon, "Improving partial parsing based on error-pattern analysis for Korean grammar-checker," ACM Transactions on Asian Language Information Processing, vol. 2, no. 4, pp. 301-323, 2003.

[2] Aesun Yoon, Soonhee Hwang, E. Lee, Hyuk-ChulKwon. 2009. Consruction of Korean WordNet ‘KorLex 1.5’, JourNal of KIISE: Sortware and Applications, Vol 36: Issue 1:92-108.

[3] H. Li and N. Abe, "Generalizing case frames using a thesaurus and the MDL principle," Computational Linguistics, vol. 24 no. 2, pp. 217-244, 1998.

[4] Peter GrÄunwald, “A Tutorial Introduction to the Minimum Description Length Principle”