 

# RECALL AND PRECISION IN GRAMMAR AND STYLE CHECKING

**Javier Gómez Guinovart**

*Computational Linguistics Group – University of Vigo*
*http://www.uvigo.es/webs/h06/webh06/sli/index.html – jgomez@uvigo.es*

Abstract: This paper presents a contribution to the methodology for the evaluation of the performance of syntax and style checkers, at the level of both error detection and error correction. The model proposed is based on recall and precision, as used in the evaluation of information retrieval systems.

Keywords: computational linguistics, evaluation of natural language processing systems, syntax checking, style checking, recall, precision.

## 1. INTRODUCTION

This paper presents a contribution to the methodology for the evaluation of the performance of syntax and style checkers, at the level of both error detection and error correction. The model proposed (see also Gómez-Guinovart, 1996a, b) is a "black-box" (Palmer and Finin, 1990) based on recall and precision, as used in the evaluation of information retrieval systems (Salton, 1989). The originality of this presentation lies in the extension of this evaluation model to the error-correction component of these systems of linguistic verification, once its applicability to the detection level has been proved by Atwell and Elliott (1987) and Atwell (1987) within the CLAWS project at the University of Lancaster. The paper also offers some reflections on the limits in the application of this methodology to style checkers, and on the conditions of acceptance of language checking systems by their users.

On the other hand, this model of evaluation focuses on measuring the system performance (that is, what the system does) with respect to error detection and correction, and it does not

ISBN: 0 08 043 438X

 

address how the system works. In this sense, the language verification system is considered as a "black box", and the model as a "black-box evaluation", following the ideas expressed by Palmer and Finin in their work on evaluation of natural language processing systems:

> In general, it should be possible to perform black box evaluation without knowing anything about the inner workings of the system —the system can be seen as a black box, and can be evaluated by system users. (1990, pp. 176–177)

## 2. RECALL AND PRECISION IN SYNTAX CHECKING

We can define the recall of a grammar checker at error-detection level (RDG) as the ratio between the errors correctly detected (DG) and the total number of errors that should be detected (EG); on the other hand, precision at the same level (PDG) can be defined as the ratio between the errors correctly detected (DG) and the total number of errors detected (TG), both correctly and incorrectly (that is to say, including "false alarms" or sequences identified as errors by the system, but which are actually not errors), that is,

$$(1) \qquad RDG = \frac{DG}{EG}$$

$$(2) \qquad PDG = \frac{DG}{TG}$$

Thus defined, recall would evaluate, at detection level, the ability of a system to identify all the errors in a text, whereas precision would evaluate its ability to detect only the errors, without raising false alarms. For example, given a text with ten syntactic errors, 100% recall would be obtained by a checker which recognised those ten errors; but for a checker to attain 100% precision, not only must it detect those ten errors but it must also not identify any grammatical sequence as an error (Example 1). Should that checker identify ten other errors incorrectly in addition to the ten errors correctly identified, its precision would be reduced to 50% (Example 2), while its recall would remain the same. Finally, if the checker detected only five out of the ten syntactic errors in the text, and at the same time it did not detect any spurious errors, its recall would be 50% but its precision would reach 100% (Example 3), as can be seen in Table 1:

ISBN: 0 08 043 438X

 

Table 1: RDG and PDG

| Detection level | Ex. 1 | Ex. 2 | Ex. 3 | Ex. 4 | Ex. 5 | Ex. 6 |
|---|---|---|---|---|---|---|
| Errors in text (EG) | 10 | 10 | 10 | 10 | 10 | 10 |
| Total identified (TG) | 10 | 20 | 5 | 10 | 40 | 10 |
| Rightly detected (DG) | 10 | 10 | 5 | 5 | 10 | 0 |
| Recall (RDG) | 100% | 100% | 50% | 50% | 100% | 0% |
| Precision (PDG) | 100% | 50% | 100% | 50% | 25% | 0% |

The same considerations can be applied at error-correction level, by replacing the mathematical variables for the detection level with the ones corresponding to the error-correction component. Since detection is a process which is previous to correction, it is obviously necessary to evaluate correction on the basis of the detection output, ruling out the correction failures caused by previous failures in the detection process. Thus, the recall of a grammar checker at error-correction level (RCG) is defined as the ratio between properly detected errors which are rightly corrected (WG) and the total number of errors that should be corrected (DG) (that is to say, the number of errors correctly detected by the system), that is,

$$(3) \qquad RCG = \frac{WG}{DG}$$

This means that recall at error-correction level evaluates the ability of a grammar checker to properly correct all the errors rightly detected by the system, disregarding, on the one hand, the false alarms (errors which are incorrectly detected and thus are wrongly corrected) and, on the other hand, the ratio between the number of errors corrected and the total number of errors which should have been detected. This is so because, as we said before, the correction component should not be held responsible for detection failures.

As regards precision at error-correction level (PCG), it can be defined as the ratio between the number of errors properly detected which are then rightly corrected (WG) and the total number of properly detected errors which are corrected either rightly or wrongly (CG) (including "false solutions", i.e. errors correctly detected and wrongly corrected), as follows:

$$(4) \qquad PCG = \frac{WG}{CG}$$

Thus, the precision of a checker at correction level evaluates the ability of the checker to offer only adequate corrections (and not false solutions) on the basis of the errors which are properly identified by the detection component. For example, given a text in which the

ISBN: 0 08 043 438X

 

detection component has correctly identified ten syntactic errors, a checker which corrected those ten errors properly would attain 100% recall and 100% precision at correction level (Example 7). However, both recall and precision would be reduced to 50% if the correction component provided ten solutions and only five of them were adequate (Example 8). Finally, the checker would attain 100% precision but only 50% recall, if it corrected five out of the ten errors which were rightly detected but all of the solutions provided were appropriate (Example 9), as shown in Table 2:

Table 2: RCG and PCG

| Correction level | Ex. 7 | Ex. 8 | Ex. 9 | Ex. 10 | Ex. 11 | Ex. 12 |
|---|---|---|---|---|---|---|
| Rightly detected (DG) | 10 | 10 | 10 | 20 | 20 | 10 |
| Total corrected (CG) | 10 | 10 | 5 | 10 | 20 | 10 |
| Rightly corrected (WG) | 10 | 5 | 5 | 5 | 5 | 0 |
| Recall (RCG) | 100% | 50% | 50% | 25% | 25% | 0% |
| Precision (PCG) | 100% | 50% | 100% | 50% | 25% | 0% |

As can be observed by comparing the two tables above, the mathematical properties of RDG and PDG are somewhat different from those of RCG and PCG, since whereas PDG can be equal to, greater or smaller than RDG, PCG can never be smaller than RCG, that is, $PCG \geq RCG$.

The reason for this disparity is that the total number of errors which are corrected (CG), including the false solutions, can never be greater than the number of errors which should be corrected (DG) —that is, $CG \leq DG$—, given the fact that we consider the correction process to be based on the detection process and that, therefore, there is a coincidence between 1. the errors which should be corrected (DG) and the errors rightly detected, and 2. the number of errors corrected (CG) and the number of errors rightly detected which are corrected.

Conversely, the reason that PDG may be equal to, greater or smaller than RDG is that the total number of errors detected (TG), including the false alarms, may be equal to, greater or smaller than the total number of grammatical errors in the text (EG). From that it follows that the relationship between RDG and PDG is variable and may reveal the degree of mathematically inverse dependence reflected in the results provided by Atwell and Elliott (1987, pp. 135–138).

Atwell and Elliott's techniques for detecting syntactic errors are based on discovering "unusual" tag-pairs. A sequence of two words is considered "unusual" if the likelihood of its occurrence in the corpus, from the point of view of their grammatical word-class, is below a certain "unusualness" threshold. Atwell and Elliott evaluate their error-detecting methodology by means of different "unusualness" thresholds, and state the existence of a trade-off between recall and precision: "by raising the threshold it is possible to improve the precision score, but only at the expense of the recall score" (1987, p. 138).

ISBN: 0 08 043 438X

 

These results show that a grammar checker with very restrictive detection mechanisms typically achieves a very high precision and a low recall, whereas a system with scarcely restrictive detection mechanisms attains a very high recall and a low precision. The same kind of inverse relationship between recall and precision can be observed in the information retrieval systems, when one tries to improve their performance by altering the number and specificity of the keywords used to index the documents in the textual database. As Salton points:

> In practice, a compromise must be reached because simultaneously optimizing recall and precision is not normally achievable. Indeed when the indexing vocabulary is narrow and specific, retrieval precision is favored at the expense of recall, since many extraneous items are then rejected, but many useful ones are as well. The reverse obtains when the indexing vocabulary is broad and nonspecific; in that case recall is favored at the expense of precision. (1989, p. 278)

## 3. RECALL AND PRECISION IN STYLE CHECKING

The evaluation model proposed for grammar checkers can be perfectly adapted to style checkers, by simply making the necessary modifications derived from the different scopes of both systems. Thus, recall at the detection level of a style checker (RDS) may be defined as the ratio between the number of stylistic improprieties that it detects correctly (DS) and the total number of improprieties that it should detect (ES). Accordingly, precision at the same level (PDS) would be the ratio between the number of improprieties correctly detected (DS) and the total number of improprieties detected (TS), both correctly and incorrectly (that is to say, including "false alarms" or sequences identified as stylistic improprieties by the system, but which are actually not stylistic improprieties), as shown here:

$$(5) \qquad RDS = \frac{DS}{ES}$$

$$(6) \qquad PDS = \frac{DS}{TS}$$

Likewise, recall at the correction level of a style checker (RCS) can be defined as the ratio between the number of improprieties rightly detected which the system is able to correct well (WS), and the total number of improprieties which should be corrected (that is to say, the number of improprieties rightly detected by the system) (DS). Accordingly, precision at the same level (PCS) will be defined as the ratio between the number of stylistic improprieties rightly detected and properly corrected (WS) and the total number of improprieties rightly

ISBN: 0 08 043 438X

 

detected and corrected either rightly or wrongly (CS) (including "false solutions" or improprieties rightly detected but wrongly corrected), as shown below:

$$(7) \qquad RCS = \frac{WS}{DS}$$

$$(8) \qquad PCS = \frac{WS}{CS}$$

Nevertheless, the application of this evaluation model to style checkers is somewhat limited on account of the present incipient state of research on style, especially as regards tipology and the objective characterization of the different stylistic norms of a language. As we argue in (Gómez-Guinovart, 1996a), in order to be able to clearly determine the impropriety of a linguistic-stylistic feature of a text, it is first necessary to establish 1. the linguistic-stylistic varieties of a language, and 2. the linguistic-stylistic features defining the stylistic norm for each variety. This should be done as objectively as possible —for instance, by means of the statistical analysis of wide and representative corpora, as proposed by Biber (1989). For the moment being, the scientific community has not reached a reasonable consensus as far as these two questions are concerned. Therefore, the values given to the variables ES, DS, WS and CS are only relatively reliable, rather remaining a matter of opinion for the system evaluator.

## 4. FINAL REMARKS

Obviously, the degree of satisfaction obtained by the users of a language checking system will be optimum when its performance level attains 100% recall and 100% precision, but these figures can only be achieved in controlled-language checking environments. For example, the results of the evaluation of the error-detection component in the BSEC (Boeing Simplified English Checker) show 79% precision and 89% recall (Wojcik, *et al.*, 1993), whereas the same component of the SECC (Simplified English Checker/Corrector) obtains 87% precision and 93% recall (Adriaens and Macken, 1995), and IBM EasyEnglish attains 81% precision and 87% recall (Bernth, 1997).

Nevertheless, as for unrestricted-language checking systems, a system with intermediate recall and precision levels can better contribute to enhance computer word processing than a system with high recall and low precision, or a system with high precision and low recall. Similarly, after considering operational information retrieval systems, Salton reaches a similar conclusion:

> In many circumstances, an intermediate performance level, at which both the recall and the precision vary between 50 and 60 percent, is more satisfactory

ISBN: 0 08 043 438X

 

> for the average user than either of the limiting performance levels that favor high recall or high precision exclusively. (1989, p. 278)

When that compromise between recall and precision is not achievable, word processing users would prefer a language checking system with high precision and low recall to a system with high recall and low precision, since computer users are usually more inclined to tolerate errors produced by the omission of a necessary action (for instance, when the system doesn't detect an existing syntactic error) than to tolerate errors produced by the execution of wrong actions (as "false alarms" at detection level, or "false solutions" at correction level). This consideration agrees with the view expressed by Adriaens and Macken with respect to the acceptance of controlled-language checking systems:

> Spurious errors are at least misleading, often irritating (especially if there are many of them), and in the worst case they lead to a total rejection of the tool (when the user is sure that the errors are spurious). Missed errors are not so bad (from a user's point of view): mostly, the user will never know there were missed errors at all. For one thing, if he knew what errors he made, he would not need the tool; for another, missed errors by definition do not show up in the system output (so they cannot be a source of irritation). (1995, p. 128)

Users lose confidence in the system when confronted with "false alarms" and "false solutions". These "wrong actions" are unacceptable in Computer-Aided Second Language Learning, where students rely on the program "as judge and jury" (Pennington and Brock, 1992, pp. 96–98). In contrast, Richardson and Braden-Harder remark that "false alarms" and "false solutions" are best accepted by professional users, since they can be easily rejected:

> We have found, however, that professionals seem much more forgiving of wrong critiques, as long as the time required to disregard them is minimal. This is similar to using spelling checkers, which wrongly highlight many proper names, acronyms, etc., but are considered quite useful. (1988. p. 201)

## REFERENCES

Adriaens, G. and L. Macken (1995). Technological Evaluation of a Controlled Language Application: Precision, Recall, and Convergence Tests for SECC. In: *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, I, pp. 123–141.

Atwell, E. (1987). How to Detect Grammatical Errors in a Text without Parsing it. In: *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, pp. 38–45.

Atwell, E. and S. Elliott (1987). Dealing with Ill-Formed English Text. In: *The Computational Analysis of English* (R. Garside, G. Leech and G. Sampson (Eds.)), pp. 120–138. Longman, London.

Bernth, A. (1997). EasyEnglish: A Tool for Improving Document Quality. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 159–165.

Biber, D. (1989). A Typology of English Texts. *Linguistics*, 27, pp. 3–43.

ISBN: 0 08 043 438X

 

Gómez-Guinovart, J. (1996a). *Fundamentos y límites de los sistemas de verificación automática de la sintaxis y el estilo*. Ph. D. thesis. University of Santiago de Compostela.

Gómez-Guinovart, J. (1996b). Aportaciones a la metodología de evaluación de los sistemas de verificación de la sintaxis. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, **19**, pp. 7–13.

Palmer, M. and T. Finin (1990). Workshop on the Evaluation of Natural Language Processing Systems. *Computational Linguistics*, **16** (3), pp. 175–181.

Pennington, M. C. and M. N. Brock (1992). Process and Product Approaches to Computer-Assisted Composition. In: *Computers in Applied Linguistics: An International Perspective* (M. C. Pennington and V. Stevens (Eds.)), pp. 79–109. Multilingual Matters, Clevedon.

Richardson, S. and L. Braden-Harder (1988). The Experience of Developing a Large-Scale Natural Language Processing System: Critique. In: *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 195–202.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading.

Wojcik, R. H., P. Harrison and J. Bremer (1993). Using Bracketed Parses to Evaluate a Grammar Checking Application. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 38–45.

ISBN: 0 08 043 438X