APORTACIONES A LA METODOLOGÍA DE EVALUACIÓN DE LOS SISTEMAS DE VERIFICACIÓN AUTOMÁTICA DE LA SINTAXIS

Javier Gómez Guinovart
Universidad de Vigo
uvifejgg@cesga.es

Abstract

This paper presents a methodology for the evaluation of the performance of syntax 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.

1. Introducción

Dos de los parámetros utilizados con mayor frecuencia para evaluar la eficacia de los sistemas de recuperación de la información textual son la cobertura (*recall*) y la precisión. La cobertura se define como la proporción de información relevante que se ha logrado recuperar, mientras que la precisión determina la proporción de información recuperada que se considera relevante (Salton 1989:248). Por ejemplo, la cobertura del resultado de una consulta a un sistema de gestión documental (SGD) sería la proporción de documentos relevantes que se han incluido en la selección ofrecida al usuario, con respecto al total de documentos relevantes que se encuentran en la base de datos. Su precisión, en cambio, sería la proporción de documentos relevantes recuperados, en relación al total de documentos incluidos en la selección. Los documentos no relevantes que, accidentalmente, se incluyeran en esta selección constituirían el «ruido» (o información no pertinente) de la selección. De acuerdo con estas premisas, es posible evaluar la eficiencia de un SGD con respecto a una determinada consulta, en términos de su cobertura y precisión, mediante el siguiente cálculo (Salton 1989:277-278):

$$\text{cobertura} = \frac{\text{número de documentos relevantes recuperados}}{\text{número total de documentos relevantes de la base de datos}}$$

$$\text{precisión} = \frac{\text{número de documentos relevantes recuperados}}{\text{número total de documentos recuperados}}$$

En virtud de estas fórmulas, tanto la cobertura como la precisión pueden obtener un valor comprendido en una escala del 0 al 1 (o, traducido en porcentajes, del 0% al 100%), siendo mayor la eficiencia de un SGD cuanto más elevados sean los valores obtenidos por estos dos parámetros de estimación.

Este modelo de evaluación puede adaptarse con algunas modificaciones a la evaluación de los sistemas de verificación automática de la sintaxis, siempre que se tenga en cuenta que para valorar el rendimiento de estos sistemas hay que considerar, por una parte, su capacidad para detectar los errores y, por otra, su habilidad para corregirlos o para sugerir su corrección —en caso de que el sistema examinado incluya esta función—[1]. La originalidad de esta comunicación radica precisamente en la extensión de este modelo de evaluación al componente de corrección de errores de la verificación sintáctica, habiéndose demostrado ya la posibilidad de su aplicación al nivel de la detección en Atwell (1987) y en Atwell y Elliott (1987)[2].

2. Evaluación de sistemas de verificación automática de la sintaxis

Hechas estas consideraciones, podemos definir la cobertura en el nivel de detección de errores de un sistema de verificación de la sintaxis (C_d) como la proporción de errores que detecta correctamente (D) en relación con el número total de errores que debería detectar (E); mientras que definiremos su precisión (P_d), en ese mismo nivel, como la proporción de errores que detecta correctamente (D) en relación con el número total de errores que detecta, tanto correcta como incorrectamente (es decir, incluidas las «falsas alarmas» o secuencias señaladas como errores por el sistema, pero que en realidad no son tales errores) (T), según las ecuaciones:

$$C_d = \frac{\text{Número de errores que el sistema detecta correctamente (D)}}{\text{Número total de errores que el sistema debería detectar (E)}}$$

$$P_d = \frac{\text{Número de errores que el sistema detecta correctamente (D)}}{\text{Número total de errores detectados correcta o incorrectamente (T)}}$$

Así definidas, la cobertura evaluaría, en el nivel de la detección, la capacidad de un sistema para localizar *todos* los errores de un texto, en tanto que la precisión valoraría su habilidad para detectar *sólo* los errores, sin provocar falsas alarmas. Por ejemplo, dado un texto con diez errores sintácticos, un sistema de verificación que reconociese los diez errores poseería un 100% de cobertura, pero sólo alcanzaría el 100% de precisión si, además, sólo detectase esos diez errores y no señalara como incorrecta ninguna secuencia gramatical (Caso 1). Si ese mismo verificador señalase incorrectamente otros diez errores, además de los diez correctamente señalados, su grado de precisión se reduciría al 50% (Caso 2), permaneciendo inalterada su cobertura. Por otra parte, si el verificador sólo detectase cinco de los diez errores sintácticos del texto y, al mismo tiempo, no detectase ningún falso error, su cobertura sería del 50%, pero su precisión alcanzaría el 100% (Caso 3), como puede observarse en la tabla siguiente:

---

[1] Para la aplicación de este modelo de evaluación a los sistemas de verificación automática del estilo, véase Gómez Guinovart (1996).

[2] Siguiendo las ideas de Atwell y Elliott, Wojcik et al. (1993) y Adriaens y Macken (1995) han aplicado los conceptos de cobertura y precisión a la evaluación del componente de detección de sus respectivos sistemas de verificación sintáctico-estilística del «inglés simplificado» (*Simplified English* o *SE*).

Nivel de detección	*Caso 1*	*Caso 2*	*Caso 3*	*Caso 4*	*Caso 5*	*Caso 6*
Errores del texto (E)	10	10	10	10	10	10
Total señalados (T)	10	20	5	10	40	10
Bien detectados (D)	10	10	5	5	10	0
Cobertura (C_d)	100%	100%	50%	50%	100%	0%
Precisión (P_d)	100%	50%	100%	50%	25%	0%

Las mismas medidas pueden aplicarse al nivel de corrección de un verificador, sustituyendo de manera adecuada los operadores de cálculo referidos al nivel de detección por los correspondientes factores relativos a la corrección. Siendo el proceso de detección un proceso lógicamente previo al proceso de corrección, se hace necesario evaluar el nivel de corrección de un verificador gramatical sobre la base del resultado de la detección, descartando los fallos de la corrección motivados por desaciertos previos del proceso de detección. De este modo, la cobertura en el nivel de la corrección de un sistema de verificación de la sintaxis (C_C), sería la proporción de errores correctamente detectados que corrige bien (B), en relación con el número total de errores que debería corregir (o sea, el número de errores que el sistema detecta correctamente) (D), según la siguiente ecuación:

$$C_C = \frac{\text{Número de errores detectados correctamente y bien corregidos (B)}}{\text{Número de errores que el sistema detecta correctamente (D)}}$$

Es decir, la cobertura de un verificador gramatical, en el nivel de la corrección, evalúa su capacidad para corregir con acierto *todos* los errores correctamente detectados por el sistema, sin tomar en consideración ni la proporción de errores incorrectamente detectados (falsas alarmas) que corrige (evidentemente, mal), ni la proporción de errores que corrige en relación con el total de errores que deberían haber sido detectados, ya que, como acabamos de indicar, no podemos hacer responsable a la corrección de fallos producidos durante el proceso de detección.

Por otra parte, la precisión de la corrección de un verificador gramatical (P_C) sería la proporción de errores correctamente detectados que corrige bien (B), en relación con el número total de errores correctamente detectados corregidos (incluidas las «falsas soluciones» o errores correctamente detectados mal corregidos) (G), de acuerdo con la siguiente fórmula:

$$P_C = \frac{\text{Número de errores detectados correctamente y bien corregidos (B)}}{\text{Número de errores detectados correctamente y corregidos bien o mal (G)}}$$

Así, la precisión de un sistema de verificación, en el nivel de corrección, valoraría la capacidad de ofrecer *solamente* correcciones adecuadas, y no falsas soluciones, a partir de los errores correctamente señalados durante la fase de detección. Por ejemplo, dado un texto en el que el componente de detección ha reconocido correctamente diez errores gramaticales, un sistema de verificación que corrigiese bien estos diez errores poseería un 100% de cobertura y un 100% de precisión en el nivel de corrección (Caso 1). Si, en cambio, de las diez soluciones ofrecidas, sólo cinco fueran acertadas, su precisión se vería reducida al 50%, lo mismo que su cobertura (Caso 2). Por último, si sólo corrigiese cinco de los diez errores correctamente detectados, pero las cinco soluciones de corrección fueran acertadas, obtendría un 100% de precisión, pero sólo un 50% de cobertura (Caso 3), como se ejemplifica en la tabla siguiente:

Nivel de corrección	*Caso 1*	*Caso 2*	*Caso 3*	*Caso 4*	*Caso 5*	*Caso 6*
Bien detectados (D)	10	10	10	20	20	10
Total corregidos (G)	10	10	5	10	20	10
Bien corregidos (B)	10	5	5	5	5	0
Cobertura (C_c)	100%	50%	50%	25%	25%	0%
Precisión (P_c)	100%	50%	100%	50%	25%	0%

Como se puede observar comparando las dos tablas anteriores, las propiedades matemáticas de C_d y P_d son algo distintas que las de C_c y P_c, ya que mientras que P_d puede ser igual, mayor o menor que C_d, P_c no puede tener nunca un valor menor que C_c, es decir, $C_c \leq P_c$. La razón de esta disparidad radica en que el número total de errores corregidos (G), incluidas las falsas soluciones, nunca puede ser superior al número de errores que deben ser corregidos (D) —es decir, $G \leq D$—, ya que consideramos que, a efectos de su evaluación, el proceso de corrección se basa en el de la detección y, por tanto, hacemos coincidir los errores que se deben corregir (D) con los errores correctamente detectados, y el número de errores corregidos (G) con el de errores correctamente detectados corregidos.

Por el contrario, la razón de que P_d pueda ser igual, mayor o menor que C_d radica en que el número total de errores detectados (T), incluidas las falsas alarmas, puede ser igual, mayor o menor que el número total de errores gramaticales presentes en el texto (E). De ahí que la relación entre C_d y P_d sea tan variable y pueda manifiestar el grado de dependencia matemáticamente inversa que reflejan los resultados ofrecidos por Atwell y Elliott (1987:135-138)[3]. Estos autores evaluaron su técnica de detección de secuencias agramaticales, basada en la identificación de los pares de

[3] Aun siendo matemáticamente inversa, la relación entre C_d y P_d que se observa en los resultados de los análisis llevados a cabo por Atwell y Elliott no es siempre proporcionalmente inversa; más bien, parece tratarse de un tipo de relación matemática inversa no lineal.

categorías contiguas «inusuales», utilizando distintos «umbrales de normalidad», y descubrieron que elevando el valor del umbral de normalidad se incrementaba la precisión del sistema, al mismo tiempo que disminuía su cobertura, y viceversa:

> These results illustrate the 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. (1987:138)

De los resultados descritos por Atwell y Elliott se desprende que un sistema de verificación gramatical con mecanismos de detección muy restrictivos posee típicamente una precisión muy elevada y una cobertura reducida, mientras que un sistema con mecanismos de detección poco restrictivos poseería una cobertura muy amplia acompañada de una precisión escasa. Se trata de la misma relación inversa entre cobertura y precisión que se manifiesta en los sistemas de recuperación de la información cuando se intenta mejorar su rendimiento alterando el número y la especificidad de los descriptores empleados para indexar los documentos de la base de datos:

> 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. (Salton 1989:278)

3. Conclusiones

Por supuesto, el grado de satisfacción de los usuarios de un sistema de verificación sintáctica será óptimo cuando sus resultados alcancen simultáneamente el 100% de cobertura y el 100% de precisión. Sin embargo, no es realista suponer que, en el estado actual de las tecnologías utilizadas por estos sistemas, se pueda llegar a conseguir tal nivel de eficacia en un plazo relativamente cercano[4]. Lo que sí parece claro es que, en general, un sistema de verificación gramatical con unos niveles medios de cobertura y precisión puede contribuir en mayor medida a la mejora del entorno informático del procesamiento de textos que un sistema con un alto índice de cobertura y un bajo índice de precisión, o que un sistema con un elevado índice de precisión y un pobre índice de cobertura. A esta misma conclusión llega Salton tras analizar el rendimiento de los sistemas de recuperación de la información operativos:

> In many circumstances, an intermediate performance level, at which both the recall and the precision vary between 50 and 60 percent, is more satisfactory for the average user than either of the limiting performance levels that favor high recall or high precision exclusively. (1989:278)

---

[4] La situación es muy diferente en los sistemas de verificación de lenguajes controlados, donde las características de su ámbito de aplicación permiten alcanzar resultados mucho más satisfactorios. Por ejemplo, los resultados de la evaluación del componente de detección del verificador de inglés controlado BSEC (*Boeing Simplified English Checker*) ofrecidos por Wojcik et al. (1993) indican un nivel de precisión del 79% y un nivel de cobertura del 89%, mientras que los ofrecidos por Adriaens y Macken (1995) para el módulo de detección del sistema SECC (*Simplified English Checker/Corrector*) alcanzan unos niveles del 87% de precisión y del 93% de cobertura.

De no lograr este compromiso entre cobertura y precisión, un sistema de verificación gramatical con altos índices de precisión y bajos índices de cobertura podría llegar a obtener una mejor aceptación entre los usuarios de aplicaciones de procesamiento de textos, que un sistema de verificación gramatical con elevados índices de cobertura y bajos índices de precisión, ya que —en general— los usuarios de aplicaciones informáticas suelen mostrar una mayor tolerancia con los fallos de los programas debidos a la omisión de una acción necesaria (como el cometido al no detectar un error sintáctico existente), que con los fallos de los programas debidos a la realización de una acción equivocada (como las «falsas alarmas» en la fase de detección, o las «falsas soluciones» en la fase de corrección)[5]. Estos últimos tienden a minar la confianza del usuario en el sistema y son inaceptables, por ejemplo, en contextos de uso similares al de la enseñanza de segundas lenguas asistida por ordenador, en los que el estudiante suele aceptar como criterio de autoridad las indicaciones del programa[6].

4. Referencias bibliográficas

ADRIAENS, Geert; y MACKEN, Lieve (1995), «Technological Evaluation of a Controlled Language Application: Precision, Recall, and Convergence Tests for SECC», *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Centre for Computational Linguistics, Lovaina, vol. I, pp. 123-141.

ATWELL, Eric (1987), «How to Detect Grammatical Errors in a Text without Parsing it», *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, Universidad de Copenhague, Copenhague, pp. 38-45.

ATWELL, Eric; y ELLIOTT, Stephen (1987), «Dealing with Ill-Formed English Text», en GARSIDE, Roger; LEECH, Geoffrey; y SAMPSON, Geoffrey *eds.* (1987), *The Computational Analysis of English: A Corpus-Based Approach*, Longman, Londres-Nueva York, pp. 120-138.

GÓMEZ GUINOVART, Javier (1996), *Fundamentos y límites de los sistemas de verificación automática de la sintaxis y el estilo*, tesis doctoral, Universidad de Santiago de Compostela, Santiago de Compostela.

[5] Lo mismo opinan Adriaens y Macken con respecto al rendimiento de los sistemas de verificación de lenguajes controlados: «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)» (1995:128).

[6] Véanse, por ejemplo, los casos de «reliance on the program as judge and jury» relatados por Pennington y Brock (1992:96-98), en una aplicación del verificador sintáctico-estilístico Critique (de IBM) a la enseñanza de inglés como segunda lengua a alumnos de nivel universitario. Richardson y Braden-Harder observan, sin embargo, que las falsas alarmas y las falsas soluciones no son tan mal aceptadas por los usuarios profesionales de estos sistemas, ya que pueden ser rechazadas sin demasiadas molestias: «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» (1988:201).

PENNINGTON, Martha C.; y BROCK, Mark N. (1992), «Process and Product Approaches to Computer-Assisted Composition», en PENNINGTON, Martha C.; y STEVENS, Vance *eds.* (1992), *Computers in Applied Linguistics: An International Perspective*, Multilingual Matters, Clevedon, pp. 79-109.

RICHARDSON, Stephen; y BRADEN-HARDER, Lisa (1988), «The Experience of Developing a Large-Scale Natural Language Processing System: Critique», *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin-Marriot at the Capitol, Austin, pp. 195-202

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

WOJCIK, R. H.; HARRISON, P.; y BREMER, J. (1993), «Using Bracketed Parses to Evaluate a Grammar Checking Application», *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, ACL, Columbus, pp. 38-45.