Método de Corrección Ortográfica Basado en Trigramas y Distancia de Edición

Bordignon, Fernando R. A.; Tolosa[1], G. H.; Peri, Jorge y Barrientos, Diego

Universidad Nacional de Luján
Departamento de Ciencias Básicas
Laboratorio de Redes de Datos
{bordi, tolosoft, jperi}@unlu.edu.ar, dc_barrientos@yahoo.com.ar

[1] Becario de Investigación. Secretaría de Investigación y Postgrado. Universidad Nacional de Luján

Resumen

En este trabajo se exponen los primeros resultados obtenidos de evaluación de un método de corrección ortográfica. Éste permite identificar errores y generar una lista de posibles reemplazos ordenada de acuerdo a la distancia que las sugerencias mantienen con la palabra incorrecta. El método opera en dos etapas de procesamiento. Primero, mediante la utilización de un filtro basado en trigramas se construye una lista de términos candidatos; luego, se ordena la lista utilizando la métrica distancia de edición.

Los primeros resultados muestran el método basado en trigramas es una alternativa válida para la corrección de errores de ortografía, alcanzando un rendimiento cercano al 81%. Especialmente, se debe considerar que se trata de un corrector de ortografía de propósito general basado en palabras aisladas y sin ningún tipo de información del contexto.

Introducción

La corrección de ortografía es el proceso de verificar si una palabra cualquiera está escrita correctamente, tomando como elemento de comparación un diccionario. La tarea de detectar errores ortográficos y corregirlos automáticamente es un problema que ha sido investigado por muchos años. Sus soluciones tienen múltiples aplicaciones potenciales en la edición de textos, reconocimiento óptico de caracteres, traducciones automáticas, procesamiento de lenguaje natural, recuperación de información, sistemas de comunicación para discapacitados, aprendizaje de lenguaje y demás [2, 3, 4, 12].

Para cada una de estas aplicaciones se desarrollaron soluciones particulares, algunas vinculadas a la fonética, como por ejemplo Soundex [8, 11] y Metaphone [9, 10] y otras orientadas a cubrir errores de reconocimiento óptico basadas en comparación de cadenas [6, 12]. El área de las bases de datos también ha colaborado con el desarrollo de algoritmos de corrección de ortografía ya que algunas funciones de comparación de nombres son – en esencia – algoritmos de corrección [7].

Por otro lado, con la masificación de las comunicaciones y el aumento de la cantidad de información disponible en formato digital, el área de corrección ortográfica se expandió a nuevas aplicaciones. Por ejemplo, algunos motores de búsqueda – como Google – poseen asesores de reemplazo de expresiones de consulta que contienen un error de ortografía o su frecuencia en el uso

es baja. Esta funcionalidad tiende a superar barreras de comunicación debido a que algunas investigaciones plantean que cerca del 10% de las consultas están mal escritas [3, 5].

Método Propuesto

El nuevo método de corrección ortográfica se basa – como la mayoría – en utilizar como base un diccionario. El proceso comienza con una búsqueda sobre éste, si la palabra no existe entonces se la considera un error y se sugiere una lista de posible reemplazos. Para generar la lista de sugerencias se obtiene una medida de distancia entre cada palabra del diccionario y la palabra evaluada, y se toman solo aquellas que estén bajo un valor umbral. Sin embargo, este proceso es costoso computacionalmente si se lo realiza en forma secuencial sobre todo el diccionario. Para reducir la cantidad de palabras a evaluar, se propone un método de filtrado previo.

Para la implementación del filtro mencionado se descomponen las palabras del diccionario en n-gramas (en este trabajo se utilizaron trigramas) y se almacenan en una estructura de datos que soporte acceso directo por n-grama. Esta estructura es un índice invertido donde la clave es una n-grama y tiene asociada una lista de palabras que los contienen (*posting list*).

Luego, por cada palabra a evaluar, se toman sus n-gramas y se buscan en el índice aquellas palabras que contienen una cierta cantidad de n-gramas en común con la evaluada. Los elementos de este subconjunto son comparados con la palabra considerada incorrecta utilizando una métrica de distancia. En este trabajo se utilizó la distancia de edición o distancia de Levenstein. La lista final de sugerencias surge de tomar las palabras que menor distancia tienen con la palabra evaluada, respecto del valor umbral mencionado.

En resumen, el método consta de dos etapas: primero, el “filtrado”, que permite reducir el conjunto de palabras del diccionario a evaluar respecto de la palabra errónea; y segundo, la “selección”, que permite definir la lista de posibles reemplazos a partir de comparar las palabras filtradas con la errónea mediante la distancia de edición. El aporte principal de este trabajo es la combinación de las técnicas descriptas anteriormente con el objetivo de generar un nuevo método de corrección ortográfica.

Metodología experimental y resultados obtenidos

A los efectos de evaluar el método propuesto se realizaron tests de eficiencia en la corrección ortográfica. Se tomó como referencia un bloque de pruebas, con palabras en inglés, provenientes del corrector ortográfico Aspell [1]. Éste consta de 547 palabras incorrectas con sus respectivas palabras correctas asociadas. Además, se tomaron como referencia los resultados de eficiencia de las pruebas realizadas por el equipo de desarrollo de Aspell sobre diferentes versiones de su sistema y otros clásicos como Ispell, Word97, Soundex, Metaphone y Speedcop.

En la tabla 1 se presentan los resultados de la eficiencia de los diferentes métodos evaluados sobre el juego de pruebas mencionado. La columna *Score* corresponde al porcentaje de palabras encontradas por cada uno entre las sugerencias. El método propuesto se identifica como Secor-Trigramas. Además, se implementaron los métodos Metaphone y Speedcoop identificados también con el prefijo Secor.

En el gráfico 1 se muestran los mismos resultados separados de acuerdo a la posición dentro de la lista de sugerencias donde se encontró la palabra buscada. Nótese que para el caso del método propuesto el rankeo de la lista se realiza por el valor de la distancia de edición, pero para un conjunto de palabras con el mismo valor no se considera – en esta versión – ordenamiento o rankeo alguno.

Método	Score	Total no encontrado	Total encontrado	Posición
aspell-.50.3	89,76	491	56	1
Secor-Trigramas	80,80	442	88	2
aspell-.20	79,34	434	89	3
Word 97	68,01	372	158	4
Secor-Soundex	68,01	372	150	5
Ispell 3.1.20 w/ -S option	51,92	284	239	6
Secor-Metaphone	41,68	228	295	7
Secor-Speedcop	9,69	53	474	8

Tabla 1 – Resultados del test de eficiencia ordenados por *Score*

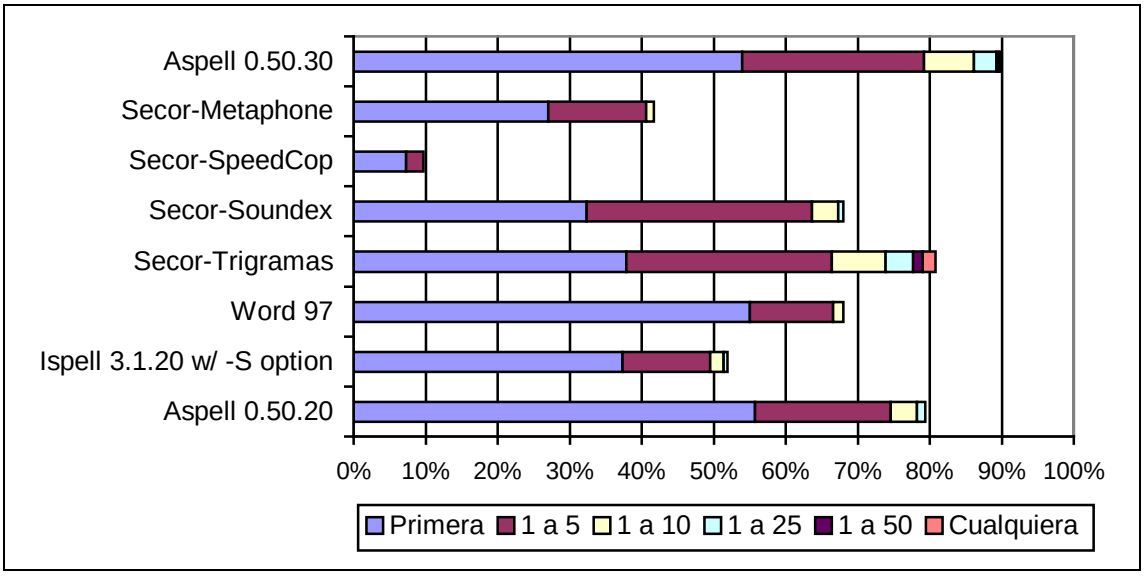

Gráfico 1 – Resultados de eficiencia por posición en lista de sugerencias

Como se puede ver en la tabla 1, el método de Trigramas se ubicó en segundo lugar detrás de Aspell 0.50.30, superando a Word97, Ispell y a la versión 0.50.20 de Aspell, por lo que resulta un rendimiento interesante para una primera versión. Si se toma en cuenta sólo el porcentaje de palabras correctas que aparecen como primera sugerencia, el método cae al cuarto lugar. Esto ocurre debido a que – como se mencionó – no existe rankeo para un conjunto de palabras con el mismo valor de distancia. No obstante la probabilidad de que la palabra correcta esté dentro de las sugerencias es alta ya que si las comparadas tienen – al menos – un trigrama en común, la correcta estará en la lista de las preseleccionadas. Esta situación se aprecia en el crecimiento en eficiencia del método luego de considerar solamente la primera palabra de la lista.

Una cuestión importante a destacar es que algunos métodos presuponen que el primer caracter de la palabra a evaluar es correcto ortográficamente. Esta restricción no es aplicable al

método de trigramas lo que lo hace más robusto. Además, hay que considerar que en el juego de pruebas utilizado solo el 22% de las palabras tienen el primer caracter erróneo.

Consideraciones y trabajos futuros

A partir de estos primeros resultados se puede deducir que el método de trigramas es una alternativa válida para la corrección de errores de ortografía. Esto se fundamenta en el rendimiento obtenido (cercano al 81%), en que no requiere que el primer caracter sea ortográficamente correcto y – además – en que es independiente del lenguaje.

Especialmente, se debe considerar que se trata de un corrector de ortografía de propósito general basado en palabras aisladas y sin ningún tipo de información del contexto.

Una aplicación posible del método consiste en un sistema de sugerencias para un motor de búsquedas web, similar al servicio "*Quizás quiso decir*" de Google. Para validar esta aplicación se trabajó sobre el sitio web oficial de la Universidad Nacional de Luján, tomando el vocabulario del índice invertido de su motor de consultas como lista de palabras correctas. Cuando un usuario ingresaba una palabra a buscar, se verificaba contra el vocabulario. Si existía, continuaba el proceso de búsqueda. De lo contrario se generaban las sugerencias y se presentaban al usuario. Nótese que todos los términos de la lista de sugerencias pertenecen al vocabulario del sitio web en cuestión.

Un punto importante a tratar en trabajos futuros consiste en definir y evaluar un criterio para el rankeo de la lista de sugerencias con el objetivo de que la palabra correcta se encuentre en los primeros lugares.

Bibliografía

[1] Atkinson, K. *GNU Aspell Version 0.50.3*.2002.
http://aspell.net.

[2] Crowell, J.; Zeng, Q. y Kogan, S. "*A technique to improve the spelling suggestion rank in medical queries*". En: Proceedings of AMIA Symposium. 2003.

[3] Dalianis, H. "*Evaluating a spelling support in a search engine*". En: Proceedings of NLDB-2002, 7th International Workshop on the Applications of Natural Language to Information Systems, Junio, 2002.

[4] Davidson, L. "*Retrieval of mis-spelled names in an airline passenger record system*". Communications of the ACM, 5(3):169:171, Marzo, 1962.

[5] Dean, J. "*Google's future plans*". Google press release. Diciembre, 2002. http://www.searchengineshowdown.com/newsarchive/000611.shtml

[6] Divita, G.; Browne, A.; Tse, T., Cheh, M.; Loane, R. y Abramson, M. "*A spelling suggestion technique for terminology servers*". En: Proceedings of AMIA Symposium. 2000.

[7] Eriksson, K. "*Approximate Swedish name matchingsurvey and test of different algorithms*", Reporte TRITA- NA-E972, Royal Institute of Technology, NADA, Estocolmo, 1997.

http://www.nada.kth.se/ theory/projects/swedish.html

[8] Knuth, D. E. "*The Art of Computer Programming: Vol. 3 Sorting and Searching*". Addison-Wesley, 1973.

[9] Philips, L. "*Hanging on the metaphone*". Computer Language, 7(12): 39-43, 1990.

[10] Philips, L. "*The double-metaphone search algorithm*". C/C++ User's Journal, 18(6), Junio, 2000.

[11] Russell, R. "*INDEX (Soundex Patent)*". US Patent No. 1,261,167, pp.1:4. 1918. http://patimg2.uspto.gov/.piw?Docid=01261167&idkey=E7455D7DADEF

[12] Taghva, K. y Stofsky, E. "*OCRSpell: an interactive spelling correction system for OCR errors in text*". International Journal of Document Analysis and Recognition, 3:125:137, 2001.