ALGORITMO PARA CORREGIR ERRORES ORTOGRAFICOS EN ESPAÑOL RELACIONADOS AL ACENTO

Igor A. Bolshakov
Centro de Investigación en Computación del
Instituto Politécnico Nacional
Unidad Profesional Adolfo López Mateos
México, D.F. 07738, México
Igor@pollux.cenac.ipn.mx

Sofia N. Galicia Haro
Centro de Investigación en Computación del
Instituto Politécnico Nacional
Unidad Profesional Adolfo López Mateos
México, D.F. 07738, México
sofia@pollux.cenac.ipn.mx

RESUMEN

Se propone un principio simple para detectar y corregir errores ortográficos, ya sean tipográficos o por desconocimiento, en textos en español, en pares de palabras casi homónimas como **genero** y **género**. Este tipo de errores se produce por la omisión de la marca de acento en el sustantivo o adjetivo; esta omisión ocasiona una transformación a una forma verbal correcta, pero totalmente fuera de su contexto.

Se describe el funcionamiento del programa que se construyó para detectar este tipo de error; en el cuál se incluyó la característica de solicitar la decisión del usuario en todos los casos donde reconoce este error de contexto. Esta misma idea podría ser una buena adición a los correctores ortográficos más comunes para el español, que actualmente no detectan este tipo de errores.

Se presentan unos ejemplos y la lista de las palabras casi-homónimas.

INTRODUCCIÓN

Como es bien conocido, los subsistemas ortográficos de los procesadores modernos de palabra se predestinan para el descubrimiento de cadenas de letras en textos de lenguajes naturales, que no son palabras correctas del lenguaje específico. Estas cadenas se consideran, cada una a la vez, separadas de su contexto y no se toman en cuenta sus palabras vecinas para encontrar errores y corregirlos en forma automatizada. Por lo que no pueden descubrir errores que transformen una palabra correcta hacia otra palabra correcta, pero totalmente inoportuna en el contexto específico.

En la escritura del español, en la cual se usan signos de acentuación, pueden aparecer errores del tipo indicado, que se conectan al acento. Un extranjero u otra persona, que no conoce bien la escritura del español, puede omitir el acento en combinaciones de palabras como: *este artículo*, *la práctica* o *páginas siguientes*. Ningún procesador común de palabras detecta todos estos vocablos como errores porque los considera como las formas personales de los verbos *articular*, *practicar* o *paginar*, respectivamente.

El primer autor de este artículo, como extranjero, cometió más de 60 errores, conectados al acento, en su primer artículo científico en Español. Y más de la mitad de ellos estuvieron en casi homónimos como *artículo* (*s*) *vs. articulo* (*v*), *fórmula* (*s*) *vs.* formula (v), género (s) vs. genero (v); donde (s) y (v) significan sustantivo y verbo respectivamente. La herramienta de ortografía de Word para Windows-95 [1], opción español, no pudo detectar ninguno de los errores en estos casi homónimos.

Mientras que un redactor humano elimina fácilmente tales errores apoyándose en el contexto, una computadora podría hacer casi las mismas operaciones correctivas, si existiera un programa de análisis sintáctico (parser, [2]). Pero por el momento, los procesadores comunes de palabras no tienen un analizador sintáctico de español.

Se propone en este artículo, una forma posible de descubrir la mayor parte de esos errores con un programa simple sin parser ni diccionario y se muestran ejemplos basados en un texto que se presenta en el Apéndice 1. En este texto, las palabras que aparecen en negritas corresponden a las palabras con errores de tipo casi homónimos y las palabras que

aparecen en letras cursivas corresponden a los elementos del grupo que se analizó en cada caso. Se trató de corregir la ortografía de este texto con el procesador de palabras Word (Microsoft), el más empleado actualmente, y se encontró que para él no contiene ningún error ortográfico del tipo indicado aunque realmente este texto contiene 31 errores del tipo casi homónimos. Y aún utilizando otro procesador de palabra como Word Perfect , que tiene integrados otros métodos más complejos, con este mismo texto solamente indicó la posibilidad de 9 errores de ese tipo.

Descripción del método considerado

Todos los pares de palabras casi homónimos considerados tienen una relación: sustantivo o adjetivo vs. forma verbal, por lo que son estas categorías gramaticales las que se intentan reconocer para detectar los posibles errores. Considerando además, que los errores de ortografía que más frecuentemente se cometen se relacionan al olvido o desconocimiento de escritura de palabras acentuadas, la tarea se centró en determinar el sustantivo o adjetivo sin acento. Para tener la posibilidad de detectar estos errores sin un diccionario, se emplea la concordancia (uno de los medios gramaticales principales de relación interna) de igualdad de género y número entre sustantivo y adjetivo. La principal regla de concordancia considerada en este trabajo es que cuando el adjetivo (o artículo) se refiere a un solo sustantivo, concuerda con él en género y número.

Debido a que las palabras que se quieren detectar son sustantivos o adjetivos, y considerando las clases de sustantivos y adjetivos en el español [3], para analizar la concordancia, se formaron cuatro grupos: femenino singular, femenino plural, masculino singular y masculino plural. En los cuatro grupos se consideraron, mediante la terminación de la palabra o en su forma completa, los artículos definidos e indefinidos correspondientes, las preposiciones: con, de, por, sin, los adjetivos completos: buen, gran, mal, primer, cuyas terminaciones no corresponden a las consideradas en cada grupo, y pronombres posesivos y demostrativos como antecesores del grupo que contiene a la palabra correcta, pero totalmente inoportuna en el contexto.

En el grupo masculino singular se consideró la terminación *-o* que abarca muchos de los sufijos masculinos, como los que a continuación se presentan, para detectar posibles adjetivos y sustantivos,

- *-ismo,*
- *-ado,*
- *-amiento,*
- *-andero,*
- *-ario,*
- *-edero,*
- *-atorio,*
- *-ero,*

etc.

También se consideraron las palabras con terminación en *-al* (sufijo con significado: "relativo a") y la terminación en *e* para considerar los siguientes sufijos:

- *-aje,*
- *-ante,*
- *-ente,*
- *-iente,*
- *-eje,*

etc.

como posibles palabras del grupo relacionado a la palabra casi homónima.

En el grupo masculino plural se tomaron en cuenta las terminaciones en *-os* y *-es* para determinar plurales de las terminaciones singulares de adjetivos y sustantivos masculinos.

En el grupo femenino singular se consideraron las terminaciones en *-e*, y *-a* que abarcan muchos de los sufijos de adjetivos y sustantivos femeninos, como los siguientes:

- *-ante,*
- -anza,
- *,-ería,*
- *-iente,*
- *-atoria*
- *-ida,*
- *-ancia,*
- *-osa,*

etc.

así como las terminaciones *-ión* y *al* para detectar palabras relacionadas al grupo de la palabra casi homónima.

En el grupo femenino plural las terminaciones tomadas en cuenta fueron *-as*, y *-es* como plurales de las terminaciones singulares de adjetivos y sustantivos femeninos.

Estructura del programa y ejemplos de operación

Se construyó un programa que se apoya en una lista de casi homónimos, ya que existen decenas de palabras que podrían ocasionar el problema descrito. Empezamos con una lista de 120 palabras pero realmente existen más palabras que ocasionan este tipo de errores. Gracias al trabajo que desarrolla el Dr. Alexander Gelbukh, sobre morfología del español, y a su ayuda, esta lista cuenta actualmente con 3

00 palabras y se presenta en su totalidad, en su forma acentuada, en el Apéndice 2.

Como ejemplo estan los siguientes casos:

- *ánimo* (valor, energía; alma o espíritu),
- *ánima* (alma),
- *ánimas* (plural de ánima)

con sus contrapartes:

- *animo,*
- *animas,*
- *anima*

(conjugaciones del verbo *animar* en presente indicativo para las tres personas del singular). Para estos casos y el resto de la lista, el programa intenta reconocer en las palabras vecinas, las palabras que permitan identificar a estas casi homónimas como sustantivos o adjetivos en textos arbitrarios, de acuerdo a los conocimientos de concordancia descritos.

Para las palabras del ejemplo anterior, a continuación, se presentan las siguientes líneas que corresponden al fragmento inicial de la lista de palabras casi homónimas en el programa (y cada una de ellas corresponde a una constante estructurada en el programa):

palabra = 'anima', género = femenino, número = singular,

palabra = 'animas', género = femenino, número = plural,

palabra = 'animo', género = masculino, número = singular,

. . .

donde palabra contiene la forma verbal correspondiente: anima, animas, animo y los valores de género y número corresponden a los sustantivos: *ánima, ánimas, ánimo.*

El programa que se construyó para este algoritmo contiene 900 líneas de código en Pascal para computadoras personales, 27 subrutinas y requiere que la entrada sea en modo texto. La "ventana" que se busca para analizar cada una de las palabras que pudieran ser casi homónimas consta de dos partes: una precedente a la palabra casi homónima y una segunda parte posterior a ella. La primera parte consta de una sola palabra y la segunda parte de dos palabras.

El programa principal revisa el archivo de entrada, separándolo por oraciones y éstas a su vez se fraccionan en grupos de cadenas para determinar la ventana. Ya definidos estos grupos, se compara cada una de las palabras contra la lista considerada. Al descubrir que existe una palabra de la lista de casi homónimas, se procede a verificar la concordancia con la dos partes de la ventana para establecer si se trata de una palabra inoportuna en el contexto, y por lo tanto que se puede sustituir.

Esta revisión permite en la mayoría de las ocurrencias de estos casi homónimos detectar los casos relevantes. En el texto que se da en el Apéndice 1, la palabra *anima*, del verbo *animar* no se reconoce como adjetivo o sustantivo y por lo tanto no lo detecta como posible error en el siguiente fragmento:

*... peor aún cuando uno **anima** a los jóvenes a evitar el uso del auto ...*

En cambio, la palabra *animo* (sustantivo acentuado) se reconoce como posible error en el siguiente fragmento:

*... Pero el **animo** no debe decaer ...*

Debido a la sencillez de la revisión de concordancia que se propone, el programa falla en dos ocasiones en el texto presentado como Apéndice 1. El primer caso, en el fragmento:

*... y que es importante que cada uno **participe** en la educación de los ...*

donde detecta como posible error a la palabra *participe*. Debido a la consideración de la terminación *-o* para adjetivos sustantivos masculinos, la palabra *uno* cae en este filtro y permite detectar a *participe* como sustantivo o adjetivo masculino y no como conjugación del verbo participar.

El segundo caso, en el fragmento:

*... que no sólo consideran totalmente **validas** si no **revalidas**. ...*

el programa no detecta estos errores: *válida* y *reválidas* porque las reglas de concordancia consideradas no incluyen el caso de adverbio adjetivo ni el caso conjunción adjetivo.

Se deja finalmente al usuario la última decisión de sustituir una palabra debido a la incapacidad de asegurar con estas reglas simples una única acepción a una palabra de un lenguaje natural. Cuando el usuario acepta la modificación sugerida, se sustituye su contraparte acentuada en el texto, lo que permite, como en los procesadores comunes de palabras ya mencionados, tener un archivo de salida con todas las palabras corregidas.

En general, se observa que con este método es posible detectar correctamente los errores del tipo indicado, de acuerdo a los siguientes tres casos:

1) Los casos donde la palabra casi homónima es un sustantivo o adjetivo y las palabras vecinas permiten identificarlo con el método de concordancia simple. Por ejemplo:

... es un **critico** implacable de ...

donde: *un* es artículo e *implacable* es un adjetivo, exactamente como se considera en el método.

2) Los casos donde la palabra casi homónima es una forma verbal y las palabras vecinas permiten identificar que no se trata de sustantivo o adjetivo. Por ejemplo:

... algunos días de la semana **estimulo** mi circulación ...

donde: *semana* es sustantivo (considerado en el método) y no coordina en género y número con *estimulo* y *semana* no puede coordinar con la palabra *mi.*

3) Los casos donde la palabra casi homónima es un sustantivo o adjetivo pero no se puede detectar con el método simple de concordancia, sin embargo, las terminaciones de las

palabras vecinas se ajustan al método por coincidencia. Por ejemplo:

... finalmente **ilegitimo** ...

donde: *finalmente* es adverbio (no previsto en el método) pero la terminación *-ente* está considerada en el método.

Para probar el método se consideraron, como entrada al programa, textos técnicos incluso capítulos de la tesis de Maestría de la segunda autora y aunque el vocabulario era más especializado, aparecieron errores de este tipo casi homónimos en porcentajes de 10 a 30 % del total de los errores ortográficos. En textos semejantes al presentado en el Apéndice 1, la porción de los errores del tipo considerado descubiertos por el programa es próxima al 95%.

Conclusión

Los errores del tipo descrito son comunes y difíciles de detectar de forma absoluta. En algunos casos las palabras podrían identificarse ya sea como sustantivos o adjetivos o como formas verbales por su uso más común, por ejemplo: *úlcera* y *ulcera* (poco empleada en un texto como forma verbal), pero no es el caso de la mayoría de las palabras consideradas. De la lista completa que aparece en el Apéndice 2, puede observarse que son palabras comunes y que su uso es generalizado. El método que se presenta ayuda a detectar este tipo de errores de palabras casi homónimas, y dada su sencillez se considera que no requiere demasiado esfuerzo incluirlo en los procesadores comunes de palabra, como es el caso del procesador de palabra *Word* para *Windows*, cuyo empleo está muy extendido en diversos ámbitos como el académico y el comercial.

Apéndice 1

Texto considerado para los ejemplos

Los cambios en la ciudad donde vivo han sido notables. Cuando *circulo* por la colonia Roma, recuerdo mi mochila de cuero (hace 30 años, *un* ***articulo*** *demasiado usual*), la poco frecuentada casa *de* ***prestamos***, *el* ***trafico*** *escaso* y sobre todo el cielo azul, lleno *de* ***oxigeno***. Ahora con tanto smog, algunos días de la semana *estimulo* mi circulación sanguínea con ejercicio muscular y creo *que oxigeno algo* mis pulmones. Debería hacerlo diariamente pero *el* ***transito*** *vehicular exagerado, las* ***fabricas*** *contaminantes* y otras cosas más producen mucha contaminación y no es posible ejercitarse al aire libre (***magnifica*** *forma de* mejorar nuestra salud).

En esa época era frecuente ver a los vecinos jugando fútbol en las calles, ahora *cuando transito por* colonias populares los *únicos* ***participes*** *dinámicos* soOn los autos, *con* ***matriculas*** *de diferentes* lugares, circulando *sin* ***limite*** *de velocidad,* es *una* ***lastima*** como un modo de transporte pasa a dominar una ciudad. Además, cada conductor es *un* ***critico*** *implacable de* reglamentos y cree conocer *la* ***formula*** *precisa* en vialidad pero olvida *como invalida las reglas, como lastima a los* peatones con sus actitudes irresponsables, sólo cree en su derecho, *finalmente* ***ilegitimo***. Esto ocasiona un *gran* ***desanimo***, peor cuando *uno anima a los* jóvenes a evitar el uso del auto y sólo responden: ¡No *me limites!*, es *un* ***dialogo*** *de sordos*. Creen que *uno invalida sus razones* a las que no sólo consideran *totalmente* ***validas*** *sino* ***revalidas***. Y como es *muy* ***incomodo*** no ir en auto y no es palpable e inmediato el daño ocasionado por la contaminación, seguimos ignorando este problema. Pero *el* ***animo*** no debe decaer, debemos disminuir *el* ***trafico*** *vehicular excesivo.*

Otro ***capitulo*** *aparte* es el exceso de habitantes, en este aspecto, son importantes *las* ***platicas*** *para* los más jóvenes, a quienes se les debería pedir casi como *una* ***suplica*** que se den *una* ***prorroga*** como pareja sin hijos, que eviten tener hijos como *un* ***tramite*** *forzoso.* Insistir en que ser padres responsables no sólo es una *satisfacción* ***intima*** sino una manera de vivir mejor, que no se trata sólo de querer serlo sino de aprender y que no hay *un* ***limite***, que es importante que cada *uno participe en la* educación de los hijos, que no existen *tareas* ***especificas*** como madre o como padre.

En estas grandes ciudades a todo se acostumbra uno y es mejor aceptar y tomar *con* ***jubilo*** que, como *el* ***titulo*** *de* un programa de televisión, «Aquí nos tocó vivir».

APÉNDICE 2

Lista de palabras casi homónimas

acídula
acídulas
acídulo
adúltera
adúlteras
adúltero
ágora
álabe
alígera
alígeras
alígero
ánima
ánimo
apócope
apócopes
ápoda
ápodas
ápodo
apóstata
apóstatas
apóstrofe
apóstrofes
apóstrofo
árbitra
árbitras
árbitro
artículo
auténtica
auténticas
auténtico
báscula
básculas
beatífica
beatíficas
beatífico
cálamos
cálculo
cántara
cántaras
capítulo
cápsula
cápsulas
catálogo
célebre
célebres
centrífuga
centrífugas
centrífugo
círculo
cláusula
cláusulas
coágulo
cómputo
cópula
cópulas
crítica
críticas
crítico
cronómetro
décimos
decrépita
decrépitas
decrépito
décupla
décuplas
décuplo
depósito
desánimo
diagnóstica
diagnósticas
diagnóstico
diálogo
doméstica
domésticas
doméstico
dómine
dómines
ejército
émbolo
émula
émulas
émulo
epílogo
equívoca
equívocas
equívoco
específica
específicas
específico
espontánea
espontáneas
espontáneo
estímulo
estípula
estípulas
estómago
estrépito
fábrica
fábricas
filósofa
filósofas
filósofo
fórmula
fórmulas
gárrula
gárrulas
gárrulo
género
gráfica
gráficas
gráfico
gránulo
hábito
hidrógeno
homóloga
homólogas
homólogo
idólatra
idólatras
ilegítima
ilegítimas
ilegítimo
ímproba
ímprobas
ímprobo
incómoda
incómodas
incómodo
íncubo
íntegra
íntegras
íntegro
interlínea
intérprete
intérpretes
íntima
íntimas
íntimo
inválida
inválidas
inválido
júbilo
lágrima
lágrimas
lámina
láminas
lápida
lápidas
lástima
lástimas
légamos
legítima
legítimas
legítimo
letífico
líbero
lícita
lícitas
lícito
límite
límites
línea
líquida
líquidas
líquido
lúbrica
lúbricas
lúbrico
mácula
máculas
magnífica
magníficas
magnífico
máquina
máquinas
matrícula
matrículas
módulo
monólogo
náufraga
náufragas
náufrago
nómina
nóminas
núcleo
número
ópera
óptima
óptimas
óptimo
órbita
óvalo
óvulo
óxido
oxígeno

pacífica	prédicas	recíproca	síncope	título
pacíficas	préstamos	recíprocas	síncopes	tráfago
pacífico	pródiga	recíproco	síndico	tráfico
página	pródigas	réplica	solícita	trámite
páginas	pródigo	réplicas	solícitas	trámites
pálpito	prólogo	réproba	solícito	tránsito
páramos	pronóstico	réprobas	sólida	trépano
partícipe	prórroga	réprobo	sólidas	triángulo
partícipes	prórrogas	reválida	sólido	úlcera
pátina	próspera	reválidas	subtítulo	úlceras
pátinas	prósperas	rótula	súplica	última
pérdida	próspero	rótulas	súplicas	últimas
pérdidas	pública	rúbrica	tálamos	último
petróleo	públicas	rúbricas	témpano	válida
plática	público	simultáneo	témpera	válidas
pláticas	púrpura	simultánea	témperas	válido
práctica	púrpuras	simultáneas	término	vínculo
prácticas	purpúrea	simultáneo	térrea	vómito
práctico	purpúreas	síncopa	térreas	
prédica	purpúreo	síncopas	térreo	

Referencias

[1] Word para WINDOWS 95. Guía del usuario. Microsoft Corp., 1995.

[2] Allen, James. Natural Language Understanding. Benjamin Cummings Publ., 1995

[3] Bolshakov, Igor A. El Modelo Morfológico Formal para Sustantivos y Adjetivos. Computación y Sistemas, No. 1,1996.

Igor A. Bokshakov. *Nacionalidad rusa, obtuvo el doctorado en Ciencias Técnicas en el instituto "Vympel". Sus áreas de interés son: la lingüística aplicada, algoritmos y programación de computadoras especializadas en automatización de oficinas en estandarización de softeware. Actualmente es profesor titular del Centro de Investigación en Computación del Instituto Politécnico Nacional, México.*

Sofía Natalia Galicia Haro. *Nacionalidad mexicana, obtuvo el grado de maestro en ciencias en la maestria en Ciencias de la Computación de la UNAM (IIMAS-UACPyP del CCH), esta como alumna en el laboratorio de Lenguaje Natural del Centro de Investigación en Computación del I.P.N. Sus áreas de interés son: ligüistica computacional, compiladores. Actualmente esta cursando el doctorado.*