Desarrollo de un corrector ortográfico para aplicaciones de conversión texto-voz

Ana Armenta, José Gregorio Escalada, Juan María Garrido, Miguel Ángel Rodríguez
División de Tecnología del Habla
Telefónica I+D
Emilio Vargas 6, MADRID 28043
aalv@tid.es, jges@tid.es, jmgarri@tid.es, miguel@tid.es

Resumen: En Telefónica I+D hemos desarrollado un corrector ortográfico automático especialmente orientado a la adaptación de textos de Internet para su lectura por nuestro conversor texto-voz. El modelo de corrección detecta las palabras erróneas de acuerdo con el diccionario de formas correctas, y genera alternativas de corrección a través de transformaciones elementales de caracteres (inserción, borrado, sustitución o trasposición). La elección de la mejor de las alternativas se apoya en la probabilidad de ocurrencia (según su frecuencia en el corpus de entrenamiento), la probabilidad de la transformación elemental (las matrices de confusión), el contexto (según el modelo de lenguaje) y el grado de corrección del texto (factor de confianza).
Palabras clave: modelo de corrección, matrices de confusión, diccionario, modelo de lenguaje, abreviaturas con expansiones múltiples, factor de confianza

Abstract: Telefónica I+D has developed an automatic correction method specially focused on the adaptation of Internet text in order to be read by our text-to-speech system. The correction model detects errors by searching words in a dictionary, and then (if the word isn´t found) correction candidates are generated by making elemental transformations (such as insertion, deletion, substitution or transposition). The selection of the best candidate takes into account the word probability (its frecuency in the training corpus), the probability of the elemental transformation (confusion matrices), the context (language model) and the text correction level (reliability factor)
Keywords: correction model, confusion matrices, dictionary, language model, reliability factor

1 Introducción

Las aplicaciones en las que se utiliza un conversor texto-voz (CTV) para suministrar información al usuario o cliente final, a través del teléfono, son muy dependientes del grado de corrección en la escritura de los textos que reciben como entrada. Si el texto está mal escrito, mal puntuado, mal acentuado... la lectura que hará el CTV posiblemente será incorrecta.

El usuario puede sentirse defraudado por esa "mala lectura" que percibe; sobre todo porque él no ve el texto y no entiende la razón de esa lectura deficiente. Incluso puede que el usuario ni siquiera pueda entender de manera adecuada la información recibida. Es decir, un servicio que funciona bien puede no resultar satisfactorio para el usuario por culpa de unos textos mal escritos.

En la división de Tecnología del Habla de Telefónica I+D hemos desarrollado una serie de técnicas de corrección que constituyen un ejemplo de aplicación de estrategias de procesamiento del lenguaje natural a las necesidades de productos y servicios industriales.

Aunque existen en el mercado multitud de correctores ortográficos de altas prestaciones, el tipo de aplicaciones en los que nuestra empresa demanda la

integración de un módulo corrector exige que el funcionamiento del mismo sea completamente automático, sin ninguna interacción con el usuario.

Además, el módulo corrector debe ser fácilmente adaptable a otros idiomas, o incluso a las características de los textos de unas y otras aplicaciones en las que se pueda integrar. Por este motivo se han desarrollado módulos que se encargan del preproceso y adaptación de los mensajes cortos y los mensajes de correo electrónico, para de esta forma conseguir una lectura más clara e inteligible.

Las tareas del corrector las hemos dividido en dos etapas. Una primera etapa de filtrado en la que se tratan aspectos determinísticos del texto de entrada ("emoticonos", conversión de tipos y normalización de la escritura), con técnicas tradicionales de preproceso de texto. Y una segunda etapa de corrección ortográfica, basada en criterios estadísticos, con un modelado del lenguaje de los mensajes cortos y de los correos electrónicos y sus errores. En esta comunicación nos centraremos en la descripción del desarrollo de esta segunda etapa, esto es, en el modelo de corrección que hemos diseñado y desarrollado.

2 Modelo de corrección

2.1 El modelo inicial

Se diseñó un primer modelo de corrección ortográfica basado en la detección y tratamiento de palabras aisladas. El funcionamiento del sistema es, a grandes rasgos, el siguiente. Cada palabra se busca en un diccionario que contiene todas las palabras que no precisan corrección ortográfica. Esto incluye tanto las consideradas válidas en el idioma (válidas como entradas de un diccionario) como las formas conjugadas, los diminutivos, los neologismos, palabras extranjeras frecuentes, etc.

Para completar nuestro diccionario con todas estas formas que no suelen aparecer en los diccionarios convencionales, se revisaron semiautomáticamente las palabras de varios corpus, incorporando al diccionario toda forma que se considerara correcta. Para ello utilizamos un corpus de texto periodístico en formato electrónico, un corpus de mensajes de correo electrónico obtenido de diversas listas de distribución y un corpus de mensajes cortos recogido entre el personal de la empresa.

Cuando una palabra no aparece en el diccionario (es lo que llamaremos en adelante un 'error no ambiguo' porque no hay ambigüedad en su incorrección), el modelo de corrección ortográfica establece que se realicen transformaciones simples en la palabra (inserción, borrado, substitución o trasposición de caracteres), y cada una de las formas generadas se comprueba frente al diccionario. De todas las formas generadas que aparecen en el diccionario, el sistema elige como correcta aquélla que tenga mayor probabilidad, obtenida como el resultado de multiplicar la "probabilidad a priori" [1] de la forma generada (probabilidad de la palabra en el idioma), por la "probabilidad de canal" de la transformación (probabilidad de que un carácter se substituya por otro, desaparezca, etc, en función de la identidad del carácter y de su contexto, reflejada en las matrices de confusión).

De acuerdo con el teorema de Bayes:

$$P(corr_i / error) = \frac{P(corr_i) * P(err / corr_i)}{P(err)} \quad (1)$$

donde $P(corr_i / error)$ es la probabilidad de que la forma correcta sea $corr_i$, condicionada a que aparezca la forma *err;* $P(corr_i)$ es la probabilidad a priori de la forma $corr_i$; $P(err / corr_i)$ es la probabilidad de que se haya producido la forma *err*, condicionada a la forma correcta $corr_i$; y $P(err)$ es la probabilidad de que aparezca err, que será igual para todas las alternativas de corrección. Por tanto, para maximizar la probabilidad de encontrar la forma correcta, suponiendo

que se ha detectado el error de manera fiable, basta con encontrar el máximo de la ecuación *(2)*:

$$P(corr_i / error) = P(corr_i) * P(err / corr_i) \qquad (2)$$

En una primera versión, la "probabilidad a priori" se estimaba directamente como la frecuencia de aparición de la palabra en un corpus, y se guardaba asociada a la entrada de la palabra en el diccionario. Posteriormente se incorporó un modelo de lenguaje que permite incluir información contextual en la estimación de la probabilidad "a priori", y así poder afrontar la corrección de lo que en adelante llamaremos errores ambiguos (palabras que vistas de forma aislada son correctas, pero que no lo son si se tiene en cuenta el contexto).Las "probabilidades de canal" se calculan a partir de las matrices de confusión, que, a su vez, se estiman como el número de veces que se ha producido cada uno de los errores básicos (inserción, borrado, sustitución o trasposición) en el corpus de entrenamiento [2].

El entrenamiento se realiza con técnicas de *boot-strapping*, utilizando el propio corrector ortográfico para detectar formas erróneas y su correspondiente forma correcta. Se obtiene el alineamiento óptimo [3] (con menor número de errores elementales) entre las dos formas, y se incrementan las cuentas correspondientes de las matrices de confusión. Con estas nuevas matrices se vuelve a procesar el corpus para identificar de nuevo los errores. Inicialmente, estas matrices se entrenaron con pares "forma errónea-forma correcta", repasadas manualmente, para acelerar la convergencia del método.

En el Anexo 1 mostramos nuestras matrices de confusión con la intención de aclarar tanto el método de construcción de las matrices como su significado En la matriz de inserción $Ins[x, y]$ se computa el número de veces que 'x' fue escrito como 'xy'; en la matriz de borrado $Borr[x, y]$, las veces que los caracteres 'xy' fueron escritos como 'x'; en la matriz de sustituciones $Sust[x, y]$, las veces que 'x' fue escrito como 'y'; y finalmente, en la matriz de trasposiciones $Tras[x, y]$, el número de veces que 'xy' fue escrito como 'yx'. Las matrices de confusión constituyen un modelado del tipo de errores ortográficos que se dan en el corpus de entrenamiento, y son la base para el cálculo de las probabilidades de canal.

De esta forma, la probabilidad de canal $P(err / corr_i)$, se calcula como:

- $Ins[x, y] / cuenta[x]$ en caso de inserción
- $Borr[x, y] / cuenta[y]$ en caso de borrado
- $Sust[x, y] / cuenta[x]$ en caso de sustitución
- $Tras[x, y] / cuenta[x, y]$ en caso de trasposición

donde $cuenta[x]$, $cuenta[y]$ y $cuenta[x, y]$ representan el número de apariciones de x, y, y xy respectivamente en el corpus de entrenamiento.

Esta aproximación permite adaptar el comportamiento del módulo corrector a distintos tipos de contenidos, e incluso a distintos idiomas. Entrenando el diccionario de formas y las matrices de confusión con textos procedentes de un tipo de contenidos determinado, tendremos un modelo de corrección adaptado al entorno en cuestión.

2.2 Corrección de uno o dos errores simples

En un primer momento, el módulo corrector intentaba tratar, para cada palabra considerada incorrecta (no presente en el diccionario), un solo error simple de los recogidos en las matrices de confusión. Esta limitación se impuso para mantener unas buenas prestaciones de tiempo de respuesta. Por cada letra de la forma

"errónea", se generan unas 50 alternativas, por lo que en promedio se generan por cada palabra unas 200 formas que hay que consultar en el diccionario. Si se intentan corregir dos errores, se pasa a unas 40.000 consultas por palabra, y para tres errores, unos 8.000.000.

Para poder obtener mejores tiempos de respuesta, y por tanto aumentar el número de errores tratados, se implementaron, en una segunda versión, técnicas de acceso al diccionario basadas en estructuras *trie* [4][5]. También se utilizaron técnicas de poda que reducen el número de consultas. De esta forma, se consiguieron unos tiempos de respuesta aceptables para la corrección de dos errores simples (similares a los de la versión anterior para un único error).

Otro inconveniente de incrementar el número de errores tratados es un incremento de la probabilidad "de falsa alarma" (modificar palabras correctas que no han sido incluidas en el diccionario). Si se intenta una sola transformación elemental, es posible que no se genere ninguna forma recogida en el diccionario, y por tanto se mantiene la forma original. Pero si se acumulan más transformaciones, es seguro que en algún momento se generarán formas recogidas en el diccionario, y por tanto se modificará la palabra.

	1error simple	2 errores simples
Corrección de errores no ambiguos	86,80(%)	87,63(%)
Corrupción de palabras correctas	0,18(%)	0,31(%)
Expansión correcta de abreviaturas	76,71(%)	77,48(%)

Tabla 1. Comportamiento del corrector ante una y dos transformaciones elementales, en ambos casos se utiliza el modelo de lenguaje

En la Tabla 1 se presentan resultados del corrector cuando corrige con una o dos transformaciones elementales. Como se ve, al tratar dos errores simples aumenta la tasa de corrección de errores no ambiguos, pero aumenta también la tasa de "corrupción" de palabras correctas.

2.3 Abreviaturas con expansiones múltiples

Una característica muy frecuente de las abreviaturas o códigos que se utilizan en los mensajes cortos de móviles (y cada vez más en los mensajes de correo electrónico) es que a menudo tienen varias lecturas posibles, de manera que queda a la inteligencia del lector el encontrar la expansión correcta.

Para el tratamiento automático de estos mensajes, tan pernicioso puede resultar el no expandir estas abreviaturas como expandirlas con una opción incorrecta.

Por esta razón, uno de los aspectos abordados en nuestro desarrollo ha sido el tratamiento de abreviaturas con varias posibilidades de expansión.

Para ello, manteniendo como referencia el modelo seguido en el corrector, se optó por una representación muy sencilla para cada abreviatura:

abr ¬ exp1(freq1) ¬ exp2(freq2) ¬ exp3(freq3) ...

Con esta representación, una abreviatura *abr* puede ser sustituida por *exp*1, o *exp*2, o exp3, o... El número entre paréntesis representa la frecuencia con la que se da esa expansión. Si no se indica ningún número, esta frecuencia se estima a partir de la recogida en el diccionario para cada una de las formas *exp*1, *exp*2, etc.

Cuando se encuentra en el texto una forma que aparece en la tabla de abreviaturas, se proponen todas las posibles expansiones, y será la evaluación proporcionada por el modelo de lenguaje la que determine cuál es la forma correcta.

3 Modelo de lenguaje

3.1 Un primer modelo de lenguaje

Como ya hemos mencionado, en una segunda fase de desarrollo del corrector nos planteamos la inclusión en el modelo

de corrección de información contextual a través de un modelo de lenguaje. Puesto que se puede considerar que el alfabeto de símbolos coincide con el de las palabras recogidas en el diccionario de formas correctas (unas 100.000), no parecía adecuado utilizar un modelo de lenguaje basado directamente en la forma de las palabras. Se optó entonces por la utilización de categorías gramaticales, enriquecidas con información adicional que recogiera las relaciones entre las palabras (información de género, de número, verbo auxiliar, etc). Se diseñó así un conjunto de 75 etiquetas.

Se empleó el módulo de proceso lingüístico del CTV para procesar los distintos corpus manejados (texto periodístico, mensajes de correo electrónico y mensajes cortos corregidos) y generar los pares "palabra<>etiqueta", utilizando el módulo de categorización gramatical del conversor. A partir de esta secuencia de pares se obtuvo tanto la cuenta de trigramas que constituye el modelo de lenguaje como las etiquetas correspondientes a cada forma del diccionario (una misma forma puede funcionar con distintas etiquetas en distintos contextos; por ejemplo, 'ría' puede ser verbo o nombre).

Sobre la cuenta de trigramas (aparecieron 59.245 secuencias, de las 421.875 posibles) no se utilizó ninguna técnica de *discounting*, pues se observó que al conservarse la decisión local, palabra a palabra, no se perdía ninguna precisión al rechazar las formas con una etiqueta que nunca se dio en el contexto correspondiente. La probabilidad asociada a cada trigrama fue incorporada entonces al modelo de corrección.

En la Tabla 2 se presentan: en A, los resultados del corrector antes de la incorporación del modelo de lenguaje de categorías (cuando solamente se tenían en cuenta la probabilidad a priori y las matrices de confusión); en B, los resultados del corrector cuando funciona con la probabilidad correspondiente al modelo de lenguaje de categorías.

	A	B
Corrección de errores no ambiguos	83,30(%)	86,80(%)
Corrupción de palabras correctas	0,50(%)	0,18(%)
Expansión correcta de abreviaturas	74,34(%)	76,71(%)

Tabla 2. Comportamiento del corrector sin y con modelo de lenguaje de categorías (A y B, respectivamente)

3.2 Mejoras y ajustes del modelo de lenguaje

Cuando se ha pretendido extender el uso del modelo de lenguaje para decidir entre las distintas expansiones de una abreviatura (y para decidir frente a formas ambiguas, potencialmente correctas), se ha comprobado que muchas expansiones (las de las abreviaturas más frecuentes), se corresponden a la misma categoría gramatical, o se utilizan en contextos gramaticales muy similares. Por ejemplo, algunas de las expansiones más habituales de 't' son 'te' y 'tú', y ambas aparecen frecuentemente ante verbo.

Sin embargo, 'tú' aparecerá casi exclusivamente frente a formas en segunda persona del singular, mientras que 'te' admite una distribución más amplia. Una posibilidad era incluir la persona de la conjugación de los verbos como un elemento más en la etiqueta de categoría, pero este camino conducía a una explosión del número de etiquetas, que difuminaba la capacidad aglutinante del modelo, y la posibilidad de su correcto entrenamiento.

Por esta razón se optó por complementar el modelo de etiquetas de categorías con un modelo de formas de palabras. Además de los trigramas de categorías ya descritos, el modelo incluye ahora trigramas de palabras. Para evitar la explosión del número de trigramas posibles

(para unas 100.000 palabras en el diccionario, se tendrían 1.000.000.000.000.000 de posibles trigramas), se limitó el modelo a los trigramas en los que los tres elementos tienen una frecuencia muy alta (entre las 5.000 palabras más frecuentes). Aunque esta aproximación es muy poco ortodoxa, nos facilita un mecanismo suficiente para resolver los contextos más frecuentes (que es donde más suelen utilizarse las abreviaturas).

Las probabilidades de los dos submodelos se combinan con un peso que se ha ajustado para que en los casos correctos ambos submodelos queden equilibrados.

3.3 Tratamiento de errores en formas ambiguas

Un número muy elevado de los errores presentes en los mensajes analizados transformaban la palabra correcta en otra forma que coincidía con otra palabra correcta. Se trata de formas ortográficamente correctas cuyo uso es incorrecto en ese contexto (aproximadamente en la mitad de los errores analizados se daba esta circunstancia). Este fenómeno es el que se ha denominado "errores en formas ambiguas", y no eran tratables en la primera versión del corrector, pues sólo se intentaba corregir una palabra si no aparecía en el diccionario de formas correctas. Puesto que este tipo de errores generan palabras que sí van a aparecer en el diccionario de formas correctas, estas palabras se daban por buenas, y no se intentaban transformar.

Dada la importancia de este volumen de errores, y confiando en las posibilidades del modelo de lenguaje para detectar contextos anómalos (que marcarán posibles errores en formas ambiguas), se decidió afrontar su tratamiento.

Un elemento nuevo es el del grado de confianza del texto original y de las formas propuestas. Puesto que se trata de modificar una palabra que en principio podría ser correcta por otra distinta, se incluyó un procedimiento en la interfaz, que permite especificar un nivel de confianza (de 0 a 100) del texto que se está tratando. Si la confianza en la corrección inicial del texto es alta, sólo se aceptarán transformaciones que supongan un claro incremento de la probabilidad resultante. Si la confianza es baja (el usuario supone que es muy probable que el texto contenga formas erróneas), se aceptarán las transformaciones propuestas con umbrales más bajos de probabilidad. Hay valores especiales del factor de confianza: el 110 se emplea para corregir únicamente mediante expansión de abreviaturas e inserciones simples; el 120 se emplea para solamente expandir abreviaturas.

4 Resultados

Hemos presentado un módulo corrector, desarrollado en Telefónica I+D, completamente automático, cuyos componentes (modelo de lenguaje, matrices de confusión, diccionario, tabla de abreviaturas) se pueden adaptar fácilmente al idioma y tipo de texto según sea la aplicación a la que se quiera incorporar.

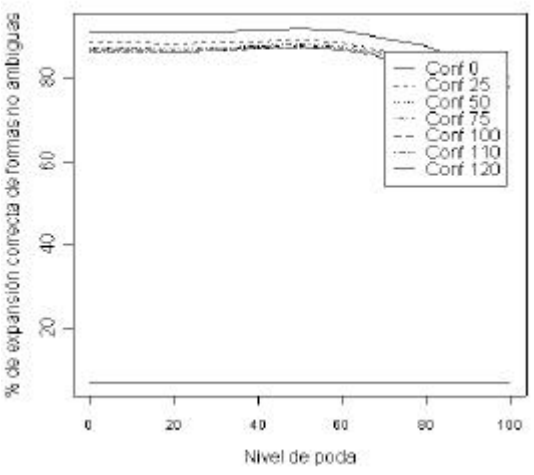


Figura 1. Variación del porcentaje de expansión correcta de formas no ambiguas en función del factor de confianza

En concreto, nosotros lo hemos aplicado a textos de Internet en español

(mensajes cortos y mensajes de correo electrónico), ya que estos van a ser leídos por el CTV, que exige como entrada texto correctamente escrito.

Las figuras 1, 2 y 3 describen el comportamiento de nuestro corrector de mensajes cortos según se trate de errores no ambiguos, errores ambiguos o abreviaturas, para distintos niveles de poda y del factor de confianza.

Como vemos, el comportamiento del corrector se mantiene constante hasta que se aplican podas del 50-60%, punto en el que comienza a decaer la tasa de correcciones acertadas. También queda patente el distinto comportamiento del corrector ante variaciones del factor de confianza según se trate de errores no ambiguos, errores ambiguos o abreviaturas. El corrector debe ser configurado con las opciones más apropiadas para cada servicio en el que se vaya a utilizar.

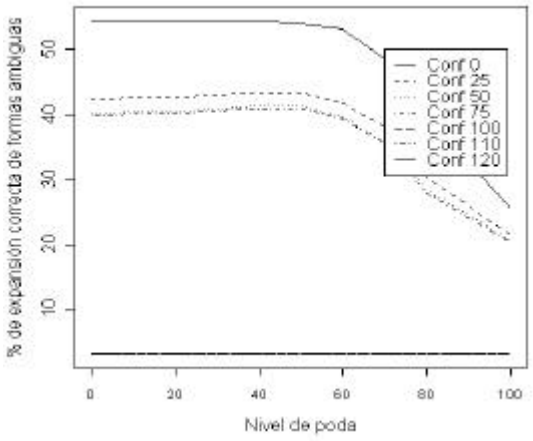


Figura 2. Variación del porcentaje de expansión correcta de formas ambiguas en función del factor de confianza

El módulo corrector que hemos descrito está siendo utilizado en diversos servicios de lectura de mensajes cortos en castellano que utilizan el CTV.

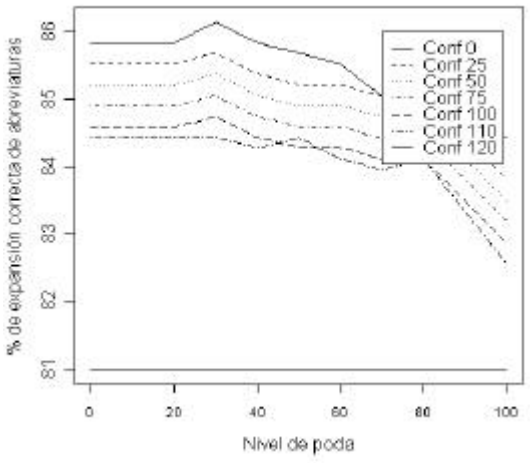


Figura 3. Variación del porcentaje de expansión correcta de abreviaturas en función del factor de confianza

Bibliografía

[1] Shannon, C.E. 1948. A mathematical theory of communication, *Bell System Technical Journal*, vol.27:379-423 y 623-656. .

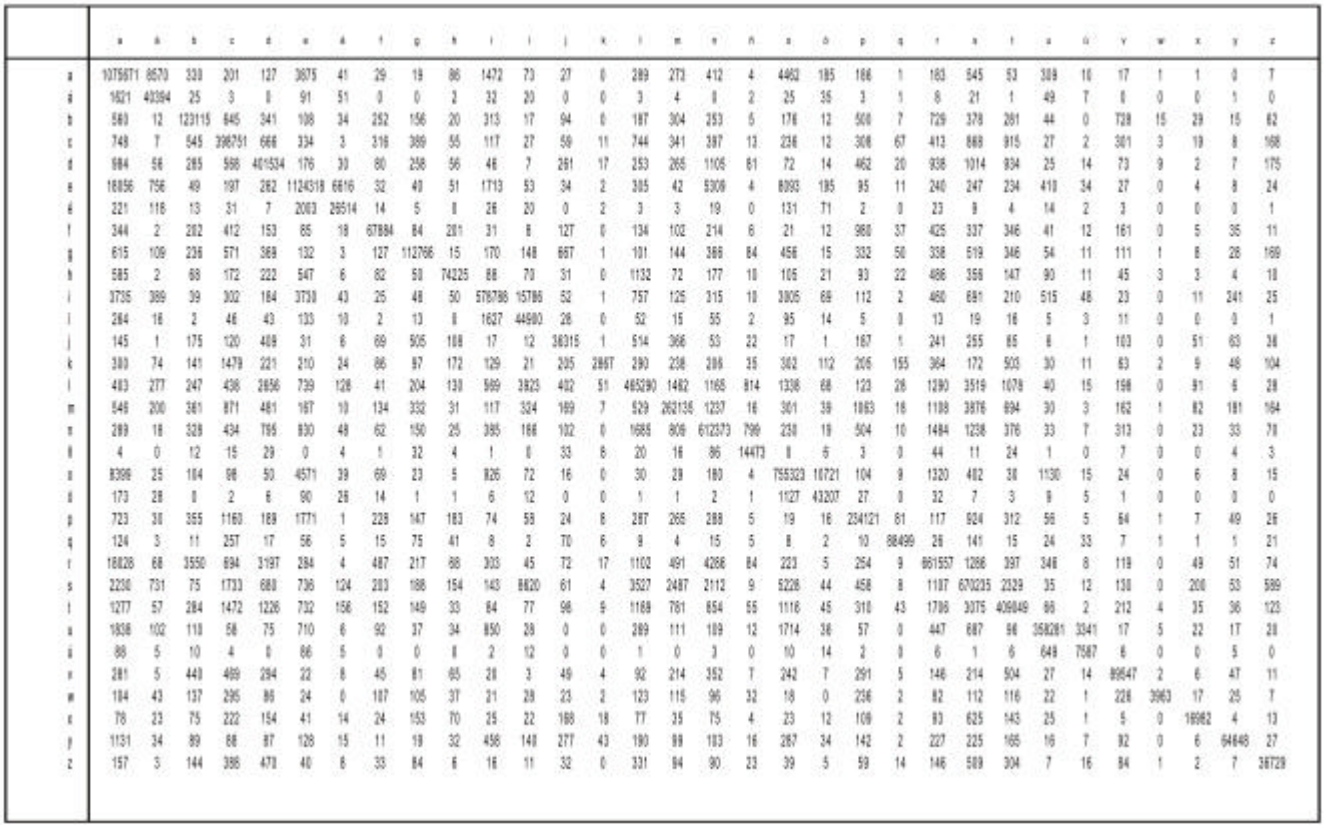
[2] Church, K. y W. Gale. 1991. Probability scoring for spelling correction. *Statistics and Computing I: 93-100.*

[3] Wagner, R.A. y M.J. Fisher. 1974. The String-to-String Correction Problem . *Journal of the asociation for Computing Machinery* , vol 21, No 1: 168-173.

[4] Muth, F. y A.L Tharp, . 1977. Correcting human error in alphanumeric terminal input. *Information Processing and Management*, 13(6): 329-337.

[5] Knuth, D.E. 1968 . *The Art of Computer Programming*, Volume 3 / Sorting and Searching

A Anexo 1: Matrices de confusión



	a	á	b	c	d	e	é	f	g	h	i	í	j	k	l	m	n	ñ	o	ó	p	q	r	s	t	u	ú	v	w	x	y	z
a	1075671	8570	330	201	127	3675	41	29	19	86	1472	73	27	0	289	273	412	4	4462	185	186	1	183	545	53	389	10	17	1	1	0	7
á	1621	40394	25	3	0	91	51	0	0	2	32	20	0	0	3	4	0	2	25	35	3	1	6	21	1	49	7	0	0	0	1	0
b	593	12	123115	645	341	108	34	252	156	20	313	17	94	0	187	304	253	5	176	12	508	7	729	378	261	44	0	728	15	29	15	62
c	748	7	545	396751	666	334	3	316	389	55	117	27	59	11	744	341	397	13	236	12	308	67	413	868	915	27	2	301	3	19	8	168
d	984	56	285	598	401534	176	30	80	258	56	46	7	261	17	253	265	1105	61	72	14	462	20	936	1014	934	25	14	73	9	2	7	175
e	16856	756	49	197	262	1124318	6616	32	40	51	1713	53	34	2	385	42	5309	4	8093	195	95	11	240	247	234	410	34	27	0	4	8	24
é	221	116	13	31	7	2083	26514	14	5	8	26	20	0	2	3	3	19	0	131	71	2	0	23	9	4	14	2	3	0	0	0	1
f	344	2	282	412	153	85	18	67884	84	281	31	8	127	0	134	102	214	6	21	12	980	37	425	337	346	41	12	161	0	5	35	11
g	615	109	236	571	369	132	3	127	112766	15	170	148	667	1	101	144	386	64	456	15	332	50	338	519	346	54	11	111	1	8	28	169
h	585	2	68	172	222	547	6	82	50	74225	88	70	31	0	1132	72	177	10	105	21	93	22	486	356	147	90	11	45	3	3	4	18
i	3735	389	39	302	184	3733	43	25	48	50	576788	15786	52	1	757	125	315	10	3805	69	112	2	460	691	210	515	48	23	0	11	241	25
í	284	16	2	46	43	133	10	2	13	8	1627	44980	28	0	52	15	55	2	95	14	5	0	13	19	16	5	3	11	0	0	0	1
j	145	1	175	120	489	31	6	69	505	188	17	12	36315	1	514	366	53	22	17	1	187	1	241	255	85	8	1	103	0	51	63	38
k	380	74	141	1479	221	210	24	86	97	172	129	21	205	2867	290	238	286	35	362	112	206	155	364	172	503	30	11	83	2	9	46	104
l	483	277	247	436	2656	739	128	41	204	130	969	3823	402	51	465290	1482	1165	814	1336	68	123	26	1290	3519	1079	40	15	198	0	91	6	28
m	546	200	361	871	481	167	10	134	332	31	117	324	169	7	529	262135	1257	16	301	39	1863	16	1108	3876	694	30	3	162	1	82	181	164
n	289	18	328	434	765	830	48	62	150	25	385	166	102	0	1666	909	612573	799	238	19	504	10	1484	1238	376	33	7	313	0	23	33	78
ñ	4	0	12	15	29	0	4	1	32	4	1	0	33	8	20	16	86	14473	8	6	3	0	44	11	24	1	0	7	0	0	4	3
o	8399	25	184	98	50	4571	39	69	23	5	826	72	16	0	30	28	180	4	755323	16721	104	9	1320	482	38	1130	15	24	0	6	8	15
ó	173	28	8	2	6	90	26	14	1	1	6	12	0	0	1	1	2	1	1127	43207	27	0	32	7	3	9	5	1	0	0	0	0
p	723	30	355	1160	189	1771	1	228	147	183	74	58	24	8	287	265	288	5	19	16	234121	61	117	924	312	56	5	64	1	7	49	26
q	124	3	11	257	17	56	5	15	75	41	8	2	70	6	9	4	15	5	8	2	10	88499	26	141	15	24	33	7	1	1	1	21
r	18828	68	3550	694	3197	284	4	487	217	88	303	45	72	17	1102	491	4288	84	223	5	264	9	861557	1286	397	348	8	119	0	49	51	74
s	2230	731	75	1733	680	736	124	283	188	154	143	8620	61	4	3527	2487	2112	9	5228	44	458	8	1197	670235	2329	35	12	130	0	200	53	599
t	1277	57	284	1472	1226	732	156	152	149	33	84	77	98	9	1169	761	854	55	1116	45	310	43	1786	3075	406949	86	2	212	4	35	36	123
u	1838	102	118	58	75	710	6	92	37	34	850	28	0	0	289	111	189	12	1714	36	57	0	447	687	96	358281	3341	17	5	22	17	28
ú	88	5	10	4	0	86	5	0	0	8	2	12	0	0	1	0	3	0	10	14	2	0	6	1	6	648	7587	6	0	0	5	0
v	281	5	449	489	294	22	8	45	81	65	28	3	49	4	92	214	352	7	242	7	291	5	146	214	504	27	14	89647	2	6	47	11
w	194	43	137	295	86	24	0	187	105	37	21	28	23	2	123	115	96	32	18	0	236	2	82	112	116	22	1	228	3963	17	25	7
x	78	23	75	222	154	41	14	24	153	70	25	22	168	18	77	35	75	4	23	12	109	2	93	625	143	25	1	5	0	16962	4	13
y	1131	34	89	68	87	128	15	11	19	32	456	148	277	43	190	88	183	16	267	34	142	2	227	225	165	16	7	92	0	6	64648	27
z	157	3	144	386	473	40	8	33	84	6	16	11	32	0	331	84	90	23	39	5	59	14	146	589	304	7	16	84	1	2	7	36729

Figura 5. Matriz de sustituciones Sust[x, y]

	a	á	b	c	d	e	é	f	g	h	i	í	j	k	l	m	n	ñ	o	ó	p	q	r	s	t	u	ú	v	w	x	y	z
a	178	6	25	46	36	196	5	9	31	41	375	15	17	37	96	149	366	1	54	0	18	10	914	575	99	228	2	2	19	2	72	23
á	43	0	0	0	0	335	0	0	0	5	14	1	0	1	1	3	11	0	4	0	0	0	10	10	0	4	0	0	0	0	0	1
b	66	0	28	2	6	29	1	19	3	0	74	1	15	3	63	5	4	1	32	0	4	0	56	27	1	28	2	6	0	0	5	3
c	52	0	0	227	9	21	4	2	1	133	129	1	1	33	35	6	4	1	116	1	21	0	30	63	21	76	0	11	13	6	2	8
d	33	1	3	9	32	82	5	8	15	12	211	7	7	6	1498	6	671	3	54	1	4	0	135	52	97	32	3	8	23	1	18	16
e	93	6	20	45	54	187	4	14	6	6	304	236	9	9	1169	133	84	461	10	73	6	21	1	367	681	196	194	4	1	3	62	27	7
é	4	1	0	0	0	12	0	0	0	0	5	0	0	0	1	1	3	0	0	0	0	15	6	6	3	18	0	0	0	0	0	0
f	16	0	0	4	7	28	0	88	2	4	16	1	1	1	12	0	2	0	5	0	1	0	10	4	0	13	0	0	0	23	0	0
g	6	2	1	0	15	10	0	5	7	5	35	0	0	0	4	5	6	3	39	1	2	0	48	0	6	48	0	0	0	0	1	0
h	12	1	1	1	3	22	0	1	0	1	27	0	1	0	3	6	3	0	49	0	2	0	56	13	10	9	0	3	4	3	4	3
i	122	19	40	119	61	129	2	15	10	10	186	105	8	20	61	78	345	13	53	0	42	0	85	154	156	42	0	25	0	8	4	5
í	8	2	0	9	4	3	0	0	2	8	44	6	1	5	1	0	8	0	6	1	1	8	0	8	6	3	0	0	1	0	1	0
j	1	0	3	0	0	1	0	0	0	9	14	39	0	2	0	16	1	0	0	9	2	1	0	6	25	4	2	0	0	0	4	1	4
k	0	0	0	1	0	7	0	0	0	0	0	1	0	0	0	1	0	8	0	1	2	1	0	1	2	4	6	0	0	1	0	0	0
l	174	7	5	17	22	102	8	18	8	8	141	5	15	55	297	17	11	2	21	0	19	0	9	60	165	18	3	17	3	1	32	6
m	48	4	21	7	5	7	8	3	0	10	64	8	17	3	81	92	285	1	26	0	16	0	8	12	16	10	0	4	5	0	0	6
n	153	2	3	53	73	130	8	1	106	86	199	2	11	36	9	72	185	0	105	6	4	0	19	349	162	40	0	14	0	0	39	10
ñ	1	0	0	0	1	1	0	1	0	27	15	1	0	0	6	0	93	0	1	0	0	0	1	8	2	0	0	0	0	0	109	0
o	124	0	63	23	15	201	0	5	24	9	214	3	3	43	127	99	221	0	192	1	73	0	92	634	111	435	0	30	70	1	28	3
ó	0	0	1	0	0	0	0	4	6	0	2	0	0	0	0	3	2	1	56	0	3	0	0	5	0	0	4	1	0	0	0	0
p	51	1	0	1	6	37	2	4	1	27	39	5	1	11	38	1	4	2	39	1	55	0	78	7	40	21	0	5	7	2	4	0
q	0	0	0	0	0	0	0	0	2	0	2	0	0	0	1	0	0	0	14	0	4	1	1	3	2	0	0	0	1	0	3	0
r	171	9	18	20	82	1218	8	20	6	16	194	66	1	33	396	72	95	1	552	1	6	1	337	622	126	66	0	79	3	7	43	10
s	219	0	1	46	36	306	17	5	20	114	260	3	5	55	102	33	49	0	74	3	18	3	89	614	238	65	4	9	6	21	25	7
t	19	1	8	12	5	136	1	0	5	108	123	17	3	11	103	24	8	5	32	0	1	0	387	29	137	29	3	3	0	44	14	28
u	42	0	27	16	10	38	9	3	6	1	49	0	2	12	21	3	51	1	11	0	4	0	45	31	5	37	1	4	0	2	13	2
ú	1	0	0	0	0	0	0	0	0	0	2	0	0	1	7	1	0	0	0	0	1	0	0	0	0	6	0	0	0	0	0	0
v	3	0	1	2	3	6	2	0	1	0	37	1	1	19	3	3	2	0	10	0	0	0	4	7	2	2	0	1	0	0	3	1
w	1	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0
x	0	0	2	2	0	0	0	4	2	22	4	0	0	4	1	1	0	0	1	0	0	0	2	2	10	2	0	0	2	0	0	0
y	1	0	3	0	1	0	0	0	1	0	0	0	0	6	3	0	0	1	1	1	0	0	0	11	4	6	0	0	5	0	0	0
z	18	1	0	1	0	1	0	3	1	31	11	0	0	0	3	0	3	0	2	0	0	2	2	2	7	1	0	0	1	5	0	16

Figura 6. Matriz de inserciones Ins[x, y]