Estado del arte en...

Corrección ortográfica automática

Lic. José Francisco Zelasco - Matrícula COPITEC 71 / Ing. Andrés Hohendahl / Lic. Judith Donayo

El presente trabajo presenta un exhaustivo estudio del estado del arte de los sistemas automatizados orientados a detectar y corregir errores de texto en lenguaje natural. Comprende aquellos errores que requieren una apropiada identificación y restauración, sean estos de origen ortográfico o tipográfico. Está orientado al idioma español, lo que lo hace particularmente interesante por ser especialmente compleja su morfología y debido a su amplia difusión (~350.10⁶ hispano-parlantes/2010).

1.1.- TECNOLOGÍA EXISTENTE

En esta sección se evaluarán los antecedentes vinculados al desarrollo de la tecnología electrónica y de computación a lo largo de las últimas décadas para luego mostrar la problemática ligada a la cuestión. No se incluirán temas que no resulten de aplicabilidad directa.

1.1.1.- TECNOLOGÍA HARD

La tecnología electrónica ha traído consigo un sinnúmero de cambios en los últimos tiempos. Hoy es casi impensable un dispositivo sin algún componente electrónico. Las comunicaciones, los medios masivos como la radio, la televisión, internet, satélites, comunicación móvil, microondas, láser, fibras ópticas, etc.; rigen, mueven y conectan hasta lo impensado del mundo de hoy.

La penetración de pico, micro y mini computadoras en sus diversas formas en la vida cotidiana es cada día mayor, no es algo que se vaya a detener. Ya sea en forma de un reloj, un juguete (*gameboy, playstation*), un celular, tablet, note/net/ultrabook o una desktop PC, un servidor, el lavarropas o la heladera inteligente y tantos aparatos más. Tarde o temprano habrá numerosas computadoras por persona, los dispositivos más inesperados serán inteligentes y se comunicarán con los usuarios y/o entre sí, conformando picoredes para mejorar nuestro confort, eficiencia, seguridad y en definitiva contribuyendo a mejorar la calidad de vida.

1.1.2.- TECNOLOGÍA SOFT "LA INTELIGENCIA ARTIFICIAL"

En las primeras películas futuristas de ciencia ficción, las computadoras imaginadas ya "*pensaban*" y tomaban decisiones, con caras de lata inexpresivas, un show de cintas de grabación girando con luces destellantes, a modo de ojos. No pocas veces en estas ficciones, se ha cuestionado éticamente el tema, planteándose el dilema casi bizantino de la "*inteligencia de las máquinas*" versus la del hombre. Todavía continúa el dilema si algún día próximo o no, la inteligencia humana será superada por las máquinas, llamando a este evento: "*la singularidad*" por algunos futurólogos. Otros predicen que las personas se unirán a la tecnología, solo el tiempo lo dirá.

Parece desmedida la expresión Inteligencia Artificial, dado lo poco que hoy poseen de "*inteligencia real*" todos estos complicadísimos sistemas como las computadoras hogareñas, portátiles, smartphones, etc., frente a la inmensa inteligencia biológica de que se observa en la complejidad de las funciones una proteína que interviene en procesos biológicos básicos involucrados en la vida misma.

Los mecanismos actuales, si bien revelan logros importantes, no han conseguido reducir el lapso comunicacional entre el usuario y la máquina, logrando que pueda ejecutar una orden verbal no preestablecida o memorizada, e interactúe de un modo más simple y natural.

Desde un principio se preveían operarios es-

pecializados presionando botones para lograr algún objetivo específico. Esto no cambió en 50 años, es impensable no tener que presionar botones en los teclados y/o superficies sensibles "*touch*". Lo cierto es que se debe ser un experto mecanógrafo si se necesita ingresar o solicitar información de algún sistema. El foco se orienta al mecanismo de ingreso y el manejo de errores cometidos.

Los teclados de hoy día, solo han evolucionado en sus materiales y estética, pero no hay cambio substancial respecto de las máquinas de escribir precursoras.

Además de la actividad de mecanografiado, es necesario conocer el funcionamiento de ciertos sistemas, el manejo de tipos de ventanas, también se debe saber cómo dar órdenes por medio del teclado, usando, en ciertos casos textos relativamente crípticos.

A esto se agrega conocer el uso del ratón, y el resultado de los estímulos que se realizan por medio de pantallas táctiles e incluyendo un sinnúmero de símbolos, imágenes e íconos danzando con atractivos colores y sonidos, al compás de la tecnología que las mueve.

Estas pantallas y sus métodos de interacción cambian incansablemente con cada generación, con mucha mayor velocidad de la que se pueden contabilizar o aprender

1.1.3.- COMPLEJIDAD INNECESARIA

Conforme se avanza en estas interacciones, se torna lamentablemente más complejo realizar actividades o tareas extremadamente simples: como escribir una carta y enviarla a un amigo, guardar y/o imprimir una foto, hacer que una agenda o celular avise que se tiene cita con el médico tal día y cómo llegar.

Subyace un tema no menos importante y apenas solucionado, que es la manera de lograr relacionarse con sistemas de complejidad creciente, tendiendo a ser hoy de cuestionable usabilidad.

Esto se observa en que las rampas de aprendizaje son cada vez más empinadas. Hay una necesidad imperiosa de volver a lo simple, contraria a la dirección de la presión de la tecnología, que debe justificar el porqué de procesadores y hardware cada vez más rápidos y con requisitos de memoria crecientes.

Lamentablemente, las personas o profesionales, quienes crean y/o diseñan estos sistemas de ingreso de datos, son muchas veces los mismos desarrolladores o empresas creadoras de la tecnología que subyace. En casos, terminan creando soluciones a la medida de ellos mismos y no al alcance de la gente común, gente de edad avanzada o niños.

1.2.- PROCESAMIENTO DE LENGUAJE NATURAL

Las ciencias que sin duda arribarán pronto a estas y otras nuevas metas de relación hombre-máquina, los cuales son los terrenos de la hoy en día "*mal*" llamada Inteligencia Artificial (*IA*) son el nuevo campo llamado "*Procesamiento de Texto Natural*"*(NLP: Natural Language Processing)* y tal vez se anexen numerosas ramas como una recientemente creada llamada computación con palabras (*CW: Computing with Words*) u otros derivados.

1.2.1.- TEXTO, PERO BIEN ESCRITO

No importando cual es el método de entrada, si deletreamos o hablamos, el sistema elegido para la comunicación de la inteligencia humana, sigue siendo el texto, y si ese texto no está bien formulado, la comunicación falla y por ende los sistemas pasan a ser problemáticos y hasta inútiles.

Las personas ejercen desde siempre y naturalmente la capacidad de leer correctamente textos en mal estado, escuchar bien las palabras entrecortadas, faltantes y hasta mezcladas con ruidos de toda índole. Si vamos en el futuro, a interactuar más inteligentemente con las máquinas, es hora que estas aprendan las cosas que las personas hacen naturalmente: ***corregir errores***.

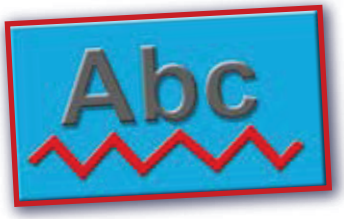
De hecho el poder determinar si un apellido o nombre propio está mal escrito en una base de datos, es un tema abierto y los sistemas actuales si bien rara vez revisan esos campos de texto por no disponer de herramientas ni algoritmos cuando lo hacen o los dejan con errores o tal cual se ingresaron, generando problemas cuando esos datos son usados posteriormente. Esto se agrava en ámbitos sensibles como los son los dominios legales, administrativos, o de salud, cruzado de información, minería de datos, operación bancaria, crediticia, o fiscal, entre muchas otras situaciones.

2.-LIMITACIONES TECNOLOGÍA, DIFICULTADES Y SOLUCIONES PROPUESTAS

El ingreso de texto en los sistemas informáticos, ha resultado una de las principales actividades masivas de las últimas décadas; impulsada recientemente por la inmensa conectividad masiva y la creciente interacción de los usuarios en internet con la web 2.0, foros, blogs, redes sociales, el chat y los mensajes cortos de los teléfonos móviles y en cada vez más electrónica de consumo como las smart-TV, que permiten ingreso de texto.

2.1.- RAZONES DE PESO

La constante miniaturización de la electrónica, acompañada por su baja de costos y el aumento paulatino de su potencia de cómputo (*ley de Moore*) permitieron a la industria crear nuevos dispositivos electrónicos cada vez más potentes para toda clase de usos.



El sinnúmero de nuevas funciones de estos dispositivos requiere del ingreso de órdenes cada vez más complejas, muchas de las cuales se realizan en forma mecánica, usando sensores especializados (*mouse/trackball/tacto/luz*), mientras que el habla directa con conversión voz a texto aún no es lo suficientemente robusto para ser usado masivamente.

Queda por último el teclado alfanumérico como el elemento contemporáneo más usual y robusto para el ingreso de datos y órdenes precisas en forma de texto.

La difusión de estas nuevas tecnologías y el grado de compromiso que presentan por su reducido tamaño, hacen que el ingreso de texto en estos dispositivos resulte cada vez más engorroso y problemático, multiplicando el número de errores que aparecen asociados. Por esto se requiere de nuevas estrategias eficientes y métodos embebidos en su electrónica para abordar los complejos problemas derivados de los errores de escritura. Esto se hace más destacado cuando se procesa lenguaje natural, pues torna potencialmente problemáticos los errores de escritura, como un ejemplo simple: los comandos de sistemas operativos y lenguajes de computación tradicionales, no toleran el más mínimo error.

2.2.-CELULARES Y LA PREDICCIÓN DE TEXTO

Un sistema de ingreso de datos para texto debiera de ser tolerante e inteligente, en especial si está orientado a lenguaje natural. Demostrado está el éxito y difusión de técnicas predictivas como el T9, usadas para el ingreso de texto en teclados numéricos de celulares, controles remotos, televisores inteligentes, reproductores y cámaras de foto, aunque un cierto porcentaje de usuarios lo deshabilitan.

2.3.- ANTECEDENTES Y NECESIDAD

Por muchos motivos que se irán analizando, la problemática a resolver para corregir errores de texto es de complejidad llamada ***NP duro***, lo cual plantea un desafío, agravado por la restricción de recursos aplicables, que se impone naturalmente.

En 1949 *Shannon* [7] publicó su tratado de modelos de comunicación de información como canales ruidosos, aplicable claramente al texto en los libros y numerosos autores como *Kernighan*[17] han usado estas teorías para detectar y corregir los errores.

En el año 1964 *Demerau*[9] ya planteó técnicas para detectar y corregir errores de escritura, en los años 1983 y 1984 numerosos autores, entre ellos *Angell R. & al.*[10] y *Daelmans W. & al.*[11]; presentaron técnicas estadísticas atacando la misma problemática, la mayoría sobre idiomas como el inglés y el alemán.

En el año 1984 *Pollock y Zamora*, [20] presentan un extenso informe usando el sistema SPEEDCOP sobre las estadísticas de los errores y sus clases en texto inglés.

Otros autores que han hecho a su vez muy buenas síntesis del problema de ortografía por computadora conjuntamente con las técnicas de corrección y detección en cada momento de la historia; son los trabajos de *Karen Kukick*. [15] en el año 1990 y anteriormente, en 1980 *Peterson*[8].

Autores citados, como *Karen Kukich* [15], en los años 1990, luego de analizar las diversas técnicas publicadas, denotaron claramente que había necesidad de mejores métricas realizando un exhaustivo estudio de los métodos de distancia de edición entre otros había hasta ese entonces, clasificando los errores en tres tipos: no-palabras, palabras aisladas y palabras en contexto.

La misma autora luego en 1992 publicó un artículo muy interesante y completo en la *ACM* [16] que explica y recaba el estado del arte conjuntamente con la efectividad de las diversas aproximaciones al problema hasta ese entonces con las limitaciones existentes junto a las soluciones creadas.

Son numerosas las publicaciones, [17] habiendo diversos enfoques del tratamiento de errores de texto como canal ruidoso. Hay algunos trabajos de *Brill* [18] del año 2000 que se basa en este enfoque para lograr detectar errores y luego tratar de reparar la ortografía.

Al igual que trabajos publicados, también registraron innumerables patentes en Estados Unidos como las US565771/1995 y la US5907839/1996 entre muchas otras, reivindicando algoritmos y métodos para la corrección de ortografía automática, la mayoría basados en estadísticas con un mix de complejos procesamientos de corpus con altas dosis de heurísticay complejas metodologías.

En el año 2000 *Yarowsky* también hizo una clara y concisa reseña del estado de este arte, en el Cap 5 del libro [12].

La hipótesis de la necesidad de corrección de texto se refuerza en el 2003 con los trabajos de *Armenta & all* [6]; donde pone el foco en la interface texto-voz.

A pesar de todos los desarrollos y patentes creados, *David Yarowskyen* sus trabajos de lingüística del año 2011 [3],aún plantea la necesidad de reconstrucción ortográfica automática. Siendo aún hoy un tema no satisfactoriamente resuelto.

2.4- IMPLEMENTACIONES COMERCIALES

El primer procesador de texto comercial muy difundido, fue el *WordStar* para CP/M el cual incluía un sistema muy básico de revisión ortográfica. Le sucedieron muchos otros como el *WordPerfect* entre tantos que saltaron a efímeras famas, cada uno en su época.

El gigante del software de los 90, *Microsoft sin dudas* embebió en su procesador de textos *MS-Word* un corrector ortográfico propietario desde el principio, del cual no hay especificación, mediciones de calidad

ni documentación disponible, salvo promesas.

Si bien este procesador de texto *MS-Word* es bastante bueno, no ha sobrevivido mucha competencia con que comparar; habiendo claros indicios de sus flaquezas: con frecuencia el corrector automático termina desactivado, por resultar contraproducente.

En el 2004 Martín Reynaert [4] desarrolla un sistema corrector para inglés y francés; logrando cifras de mérito interesantes basadas en corrección bajo contexto, comparando contra el *MS-Word y el ISpell y sus derivados*. En este trabajo la novedad consiste en usar estadísticas de bigramas de palabras y una función de dispersión especial para determinar rápidamente si una palabra pertenece al diccionario de palabras existentes y ayuda a hallar un candidato en caso de no pertenecer. Esto, en castellano sería de realización más compleja.

En 2007, *Peter Norvig* [13] describió en forma clara y simple como construir un corrector ortográfico básico, dando ejemplos utilizando una aplicación desarrollada en lenguaje ***pyhton***.

Hay un trabajo reciente del 2010 de *R. Mitton* [22] que analiza los diferentes métodos utilizados desde el principio del problema mismo, citando un gran número de trabajos y autores que se remontan a 1962.

Considerando el estado del arte actual, en lengua española, los productos y bibliotecas más importantes del mundo en esa lengua, aún no poseen incorporados correctores ortográficos satisfactorios pero tampoco hay consensuada métrica ni claros estándares especificados.

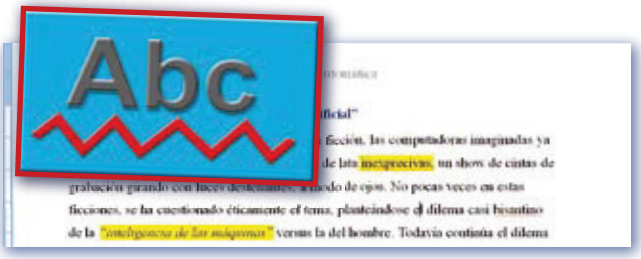
Recientemente, [23] se desarrolló una notable mejora en el estado del arte, aumentando 300% la calidad y velocidad de la corrección automática respecto al MS-Word y los demás exponentes del actual estado del arte, incluyendo reconstrucción fonética.

2.5.- AMBIENTES DE DESARROLLO Y CÓDIGO ABIERTO

Investigadores del *Politécnico de Cataluña*, en España (*UPC*) [5] crearon un producto para el tratamiento de texto en lenguaje natural. Contiene un analizador morfológico que usa muchos recursos computacionales, siendo el único de desarrollo abierto, completo y aplicable al español.

Hasta mediados del año 2013, su versión 3.0 no poseía módulo de corrección de errores de ortografía. En su lugar realiza un etiquetado de palabras desconocidas, basado en sufijos máximos, que falla cuando un error de ortografía está justamente en el sufijo y no realiza corrección alguna.

Hay otros pocos ambientes de desarrollo basados en código abierto, como el proyecto GATE de la *Universidad de Sheffield*[2], UK realizado en 1995; el cual posee numerosísimos módulos y plug-ins, pero la parte de corrección ortográfica es pobre y no posee recursos aplicables para el idioma español.



Los proyectos de código abierto que han atacado efectivamente el problema de errores de ortografía, son los aquí mencionados: *ASpell*[1], *ISpell, Hunspell, MySpell, OpenOffice, LibreOffice* y sus derivados

2.6.- CORRECCIÓN DE CÓDIGO LIBRE

Desde los años 90, pocos productos en software libre han trascendido, uno emblemático se llama *MySpell* .

Basándose en este proyecto, *László Németh* en hungría y con el auspicio de la "*Budapest University Media Research Centre*" (*BME-MOKK*) creó un producto junto a una librería **C++**, llamados: ***Hunspell*** la cual se usa ampliamente en *OpenOffice*. Otros correctores ortográficos como el GNU-***ASpell***, derivaron de ellos y poseen módulos agregables como "plug-in" para numerosos procesadores y navegadores.

Estos correctores son usados en procesadores de texto de amplia difusión como el ***OpenOffice***, ***Libre Office***. También se usan como *plug-ins*,para navegadores como el ***FireFox*** y el ***google-Chrome*** y muchos otros del mundo de software libre.

Todos ellos son de uso libre pero de mediana calidad, basándose la mayoría de ellos en compresión morfológica de afijos sumada a una búsqueda simple de errores restringidos a pocas letras y con algunos aditivos fonéticos en los más avanzados.

Otro aspecto a considerar es que la calidad va acorde al recurso lingüístico. El costo de crear su diccionario asociado y mantenerlo es muy alto, por eso se hacen productos compatibles con los diccionarios morfológicos originados en el proyecto ***MySpell***, enriquecidos comunitariamente a lo largo de décadas.

La conclusión es que casi todos estos productos poseen cifras de mérito similares, fuertemente dependientes del diccionario de base.

2.7.- RESTAURACIÓN DE ACENTOS Y DIACRITOS

Dentro de todos los posibles errores de ortografía, los más frecuentes en idiomas con marcas diacríticas (*acentos, diéresis, etc.*) son justamente la omisión o confusión de esas marcas. Diversos autores afirman que la prevalencia de este tipo de errores se sitúa entre el 80 y 90% de todos los tipos de errores. Esto, ciertamente, ha generado una creciente preocupación por este tipo de errores, buscando métodos para su detección y posible corrección en forma automática.

En numerosos trabajos se describen infinidad de técnicas específicas para *restaurar acentos* y si bien esto es importante, es tan solo una parte del problema de los errores en los textos. Hay un trabajo

del año 1994 [21] que utiliza listas de decisión para resolver esto para español y francés. Uno de los últimos trabajos muy interesantes en el tema es el publicado en 2012 por *Atserias & all*. [14].Se enfoca en el español, además de resumir y citar una gran cantidad de fuentes útiles; las analiza y explica presentando y comparando cifras de mérito de diversos métodos, y sus respectivos productos comerciales como ***Word2007*** y algunos on-line como ***Google Docs*** y ***correctorortografico.com***.

Las evaluaciones citadas [14], arrancaban en cifras de mérito o calidad de restauración de acentos, usando una medida estadística llamada *F* del 82% en métodos que usan técnicas estadísticas, contra apenas 30% medidos en el *MS-Word 2007*, cayendo a menos del 1% para los servicios online!

3.- TENDENCIAS

Una fuerte tendencia es que los procesadores de textos y 'front-ends' de sistemas lingüísticos, incorporando analizadores morfológicos robustos, que son sistemas capaces de reconocer y etiquetar partes constitutivas de un texto cualquiera (ver en glosario). Por lo general, los analizadores morfológicos que hay comercialmente y los más conocidos de código abierto cono ***FreeLing***, poseen un léxico acotado, buscan exactitud, no hallan información más allá de la incluida puramente en forma "dura" en sus bases de datos, y responden pobremente ante ambigüedades. La próxima generación de estos sistemas deberá poseer un léxico amplio y extensible automáticamente, poseer robustez y capacidad de inferencia, realizar corrección ortográfica y fonética en contexto semántico con análisis de alternativas, hacer manejo natural de términos multi-palabras y abreviaturas, ser capaces de extraer información semántica aproximada, ante la falta de datos, inferir significados, manejar palabras parasintéticas con criterio robusto, realizar la identificación de idiomas, detección de pronunciabilidad de supuestas palabras para descartarlas y ver las como ruido 'noise', detección de formulismos: matemáticos, químicos, computacionales entre otros.

Se observa que Google, hace uso de inteligencia artificial, colectiva y lingüística predictiva para 'adivinar' y 'sugerir' texto cuando estima que está mal escrito. También el gigante informático Microsoft, en su producto estrella: *Windows 7* incluyó una línea de búsqueda textual inteligente en su menú.

Surgirá una mayor necesidad de estas técnicas cuando los sistemas puedan dialogar con el usuario en lenguaje natural, habiendo numerosas evidencias de esta tendencia.

Recientemente en el 2010 salió al mercado un sistema llamado ***Siri*** para los productos de *Apple* como el ***iPhone***, ***iPod e iPad*** el cual ofrece interacción conversacional con algunas aplicaciones, como los recordatorios, el tiempo, la bolsa, la mensajería, el e-mail, el calendario, los contactos, las notas, la música y los relojes.

Es un asistente con reconocimiento y respuesta por voz, útil para responder cosas sencillas y realizar algunas operaciones con los dispositivos. Está basado principalmente en procesamiento remoto de habla y diálogo en lenguaje natural. En ***Siri*** la voz es digitalizada en el celular, luego se envía a los servidores centrales de Apple por la red móvil, quienes realizan el reconocimiento de habla y la lógica de diálogo para la determinación de la respuesta, enviando al celular del usuario la respuesta en forma de voz sintetizada junto a las órdenes asociadas.

Este esquema de tele-procesamiento, claramente se fundamenta en la escasa capacidad de proceso y almacenamiento disponibles en los dispositivos móviles. A pesar de ser impresionantemente grande, aún no alcanza para el procesamiento de habla, junto a los complejos sistemas de diálogo, necesarios en lenguaje natural.

Estimamos que dada la dificultad de comparar métodos y técnicas de corrección, aún hacen falta métricas y estándares claros y corpus anotados en numerosos idiomas, para lograr una medición de la calidad, bondad y el adecuado contraste entre todos los sistemas de corrección existentes y los por venir.

El porcentaje de errores existentes y presentes en la información contenida en casi todos los sistemas informáticos es considerable puesto que casi la totalidad de datos fueron ingresados por individuos.

En un futuro Futuro Próximo el volumen de textos ingresado en computadoras y en dispositivos móviles, tanto sea mediante SMS, chat, MSN-messenger, skype, whatsup, twitter, blogs como facebook, blogger, etc.; además de algunos sistemas incipientes de diálogo como Siri de Apple crecerá imponiendo la tendencia de comunicación en lenguaje natural.

Las redes sociales, la web 2.0 y sus derivadas, producen abundante información de texto que tiene particular interés para realizar diversos análisis que permitan obtener estadísticas sobre opiniones, productos, tendencias, temas de discusión, etc. Este interés que puede ser satisfecho mediante minería de textos y procesamiento de lenguaje, requiere de un mecanismo automatizado para la reconstrucción de errores.

Simultáneamente en el mercado están apareciendo un grupo creciente de soluciones basadas en procesamiento inteligente del habla y otras usando diálogo en lenguaje natural.

4.- CONCLUSIÓN

Todo esto refuerza y fundamenta aún más la necesidad de nuevos desarrollos de calidad, con aplicación de todas las herramientas de la ingeniería en el área, enfocados a las necesidades que surgen de los diferentes idiomas como el español y otros lenguajes flexivos, de reconocida problemática.

Tan sólo en los últimos pocos años se han realizado notables avances, teniendo en cuenta que se trata de un problema acuciante; confirmando la regla empírica de las tecnologías exponenciales, vislumbrada por el cofundador de Intel Gordon Moore ya en 1965.
En la medida que se avance en este tema espinoso, se posibilitarán más y mejores servicios generando, sin dudas, un incremento substancial en la calidad de vida de las personas en su interacción con las computadoras y máquinas inteligentes; hasta que, en un futuro cada vez más cercano, conversaremos con nuestros robots ¿tal vez HAL-9000? [23]

BIBLIOGRAFIA:

[1] "ASPELL",http://aspell.sourceforge.net/man-html/Affix-Compression.html
[2] "GATE: General Architecture for Text Engineering, http://gate.ac.uk/, 6/2011
[3] Yarowsky D. (publicaciones) http://www.cs.jhu.edu/~yarowsky/pubs.html, 06/2011
[4] Reynaert M., 'Proceedings of the 20th international conference on Computational Linguistics' 2004 (COLING 2004) Geneva (pp. 834-840)
[5] Freeling http://nlp.lsi.upc.edu/freeling/ , 06/2011
[6] Armenta, A., Escalada, J. G., Garrido, J. M. & Rodríguez, M. A. (2003). "Desarrollo de un corrector ortográfico para aplicaciones de conversión texto-voz." In Procesamiento del Lenguaje Natural, 31, pp. 65-72.
[7] Shannon, Claude. "Communication in the presence of noise" 1949 Proceedings of the IRE.-37(1):10-21.
[8] Peterson, J.L. (1980) "Computer programs for detecting and correcting spelling errors" .Communications of ACM, 23, pp. 676-687.
[9] Damerau, F.J. (1964) "A technique for computer detection and correction of spelling errors". Communications of ACM, Vol. 7, No. 3, pp. 171-177, March, 1964.
[10] Angell, R.C., Freund, G.E. & Willett, P. (1983) Automatic spelling correction using a trigram similarity measure. Information Processing & Management, 19, 255-261
[11] Daelemans, W., Bakker, D. & Schotel, H. (1984) "Automatische detectie en correctie van spelfouten". Informatie, Vol. 26, pp. 949-1024.
[12] Jurafsky, Dan and James H. Martin. 2000. Speech and Language Processing. Prentice-Hall. Chapter 5.
[13] Peter Norvig. 2007. How to Write a Spelling Corrector. On http://norvig.com.
[14] Jordi Atserias, Maria Fuentes, Rogelio Nazar, Irene Renau "Spell-Checking in Spanish: The Case of Diacritic Accents" 2012 Proceedings LREC, IStambul, Turkey. ELRA isbn: 978-2-9517408-7-7
[15] Kukich, K., "A comparison of some novel and traditional lexical distance metrics for spelling correction", lNNC-90-Paris, pp. 309-313 (1990)
[16] Alberga, C.N. String Similarity and Misspelling, In Communications of ACM, Vol. 10, No. 5, pp. 302-313, May, 1967.
[17] Kernighan et. al. "A Spelling Correction Program Based on Noisy Channel Model", In Proceedings of COLING-90, The 13th International Conference On Computational Linguistics, Vol 2. 1990
[18] Brill, E. and Moore, R. C. An Improved Error Model for Noisy Channel Spelling Correction. In proceedings of 38th Annual meeting of Association for Computational Linguistics, pp. 286-293, 2000.
[20] Pollock, J. J. & Zamora, A. (1984). "Automatic spelling correction in scientific and scholarly text." In Communications of the ACM, 27 (4), pp. 358-368.
[21] Yarowsky, D."Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French." In Proceedings. ACL. 1994
[22] Mitton, Roger (2010) Fifty years of spellchecking. Writing Systems Research 2 (1), pp. 1-7. ISSN 1758-6801
[23] Hohendahl, Andrés (2014) Reconocimiento y Corrección de Errores de Ortografía en Dispositivos Electrónicos. FI-UBA, (Tesis-electrónica)

GLOSARIO:

En esta sección se explican algunos términos específicos y que son importantes de entender en detalle para comprender este trabajo. Se describen lo suficiente, tratando de enunciar las principales ventajas y limitaciones prácticas en su utilización actual y, en donde sea posible, se cita el estado del arte o las "prácticas frecuentes" en el tema.

ANALIZADOR MORFOLÓGICO: Consiste en determinar la forma, clase o categoría gramatical de cada palabra de una oración. El análisis morfológico se refiere al hecho de determinar la categoría gramatical a la cual un determinada término pertenece, ya sea un pronombre, adjetivo, sustantivo, determinante, etc., pero además se utiliza en cuanto al análisis en la estructura de las palabras, de tal manera de estudiar su lexema, cuántos fonemas posee, sus variaciones (las flexiones de número o género, entre otras cosas). Lo complejo de este análisis, va a depender del término en cuestión y de su etimología; sirve entonces para poder referirnos con un criterio mayor a la estructura, forma y posibilidades de variación de cada palabra en una frase.
GLN / NLG (*Natural Language Generation, Generación de Lenguaje Natural*): Es la generación de lenguaje basándose en datos y lógica gramatical específica (ATG). Lo existente es radicalmente complejo, suele ser exclusivo para el inglés y no se obtienen resultados robustos, son mecanismos utilizando Planning (GraphPlan y sus variantes) como base, sintetizando estructuras con granularidad gradual, pero olvidan por lo general la semántica subyacente y la retórica asociada al diálogo.
LENGUAJE NATURAL: Traducido del inglés (Natural Language), es la lengua o idioma hablado o escrito para propósitos generales de comunicación. Son aquellas lenguas que han sido generadas espontáneamente en un grupo de hablantes con propósito de comunicarse, a diferencia de otras lenguas, como puedan ser una lengua construida, los lenguajes de programación o los lenguajes formales usados en el estudio de la lógica formal.
NLP: (Natural Language Processing = Procesamiento de Lenguaje Natural). El procesamiento del lenguaje natural es una subrama de la inteligencia artificial y de la lingüística. El fin del NLP es construir sistemas y mecanismos que permitan realizar el análisis científico del texto generado por las personas en forma natural. Una finalidad es para mejorar la comunicación entre personas y máquinas por medio de lenguajes naturales, donde se agrega la necesidad de la Generación de Lenguaje Natural, llamado NLG (ver). Además, se trata de que los mecanismos que permitan esa comunicación sean lo más eficaces posibles, computacionalmente hablando.
NP DURO (NP-HARD): Proviene del inglés: *Number Permutations Hard*, es cuando la longitud del proceso de un algoritmo que resuelve un problema depende de una magnitud combinatoria, ésta suele crecer sin límites y termina siendo irresoluble por la imposible cantidad de operaciones a realizar para completar el algoritmo, tornándolos del tipo irresoluble en tiempos polinomiales o razonables; en esto se basa la criptografía, como ejemplo.
NO-PALABRA: Se trata de formas escritas factibles de ser pertenecientes de un idioma, conteniendo solamente los caracteres permitidos por éste, ser pronunciables y responder a criterios morfológicos y fonéticos del idioma, es decir permiten una correcta conversión a fonemas, factible se ser articulados. Hay distintas clases de No-Palabras:

1. Palabra Fuera del Vocabulario (*Out Of Vocabulary OOV*): es más permisiva e incluye principalmente estas formas, candidatas a ser reconocibles por una persona e inclusive inferido su significado. En otras palabras, una no-palabra es una forma escrita (combinación válida de caracteres) pronunciable cuyo significado no pueda ser inferido o reconocida por constituyentes morfológicos. Suelen ser candidatas a ser expresamente incluidas en diccionarios. De hecho cada tantos años la Real Academia Española, así como la Academia Argentina de Letras, entre otras muchas autoridades en letras, incluyen palabras nuevas, muchas de ellas importadas de otros lenguajes, inventadas o emergentes de la misma sociedad y de uso suficientemente frecuente y con uno o más significados concretos.
2. Palabra Fuera de Diccionario: Es una forma escrita no incluida expresamente en el diccionario del contexto, no es una definición global sino acotada a un contexto de análisis morfológico en determinado idioma. Debe reunir ciertas características como ser pronunciable y que pueda ser el resultado de una derivación o flexión morfológica a partir de una palabra existente, pertenecientes al diccionario del contexto. Esta derivación puede o no tener sentido semántico correcto. Toda no palabra es una palabra fuera de diccionario, mientras que no toda palabra fuera de diccionario es una no-palabra.

PALABRA: En el contexto lingüístico podríamos definir una palabra como toda forma escrita que satisfaga una de las siguientes características: pertenecer a un idioma, diccionario o a un listado de palabras aceptadas como tal. Se diferencian de las no-palabras en que deben necesariamente estar en uno de esos grupos y cumplir criterios mínimos de morfología y fonética.
PARSERS: Un parser es un programa que analiza un texto para determinar su estructura sintáctica. El análisis sintáctico convierte el texto de entrada en otras estructuras (comúnmente árboles), que son más útiles para el posterior análisis y capturan la jerarquía implícita de la entrada. Un analizador léxico crea tokens de una secuencia de caracteres de entrada y son estos tokens los que son procesados por el analizador sintáctico para construir la estructura de datos, por ejemplo un árbol de análisis o árboles de sintaxis abstracta.
TOKENIZADOR: Pieza de software que implementa un segmentador (cortadores en partes = tokens) de texto crudo (secuencia de caracteres o 'string') obteniendo segmentos como formas escritas, números, fórmulas matemáticas, químicas, etc. Se usa siempre que se debe segmentar una 'string' en palabras, antes de usar analizador morfológico.
Por lo general son autómatas finitos, mayormente generados por software mediante un script escrito en alguna gramática regular. También se los llama Lexers que son basados en Expresiones regulares.