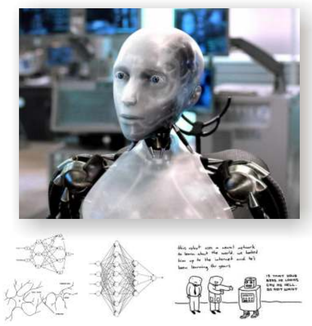
Interfase Hombre-Máquina con Inteligencia Artificial

Procesamiento de Lenguaje Natural Robusto



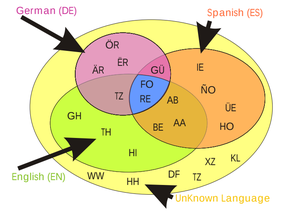
Andres T. Hohendahl [1,2] ***José F. Zelasco*** [1,3]

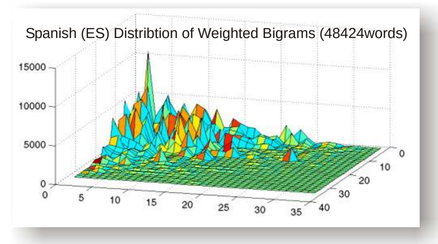
andres.hohendahl@fi.uba.ar *jfzelasco@fi.uba.ar*

1 Laboratorio de Estereología y Mecánica Inteligente. Dto. de Ing. Mecánica de la Facultad de Ingeniería, U.B.A.
2 Instituto de Ingeniería Bio Médica, Universidad de Buenos Aires, Paseo Colón 850, (1065), Buenos Aires, Argentina.
3 INTIA, Facultad de Ciencias Exactas, UNCPBA, Universidad del Centro de la Provincia de Buenos Aires, Campus Universitario, Paraje Arroyo Seco (B7000) Tandil, Prov. de Buenos Aires, Argentina.

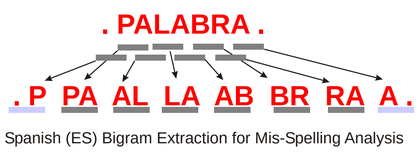
Desafíos

En computación, el ingreso de datos y los errores son un problema grave, para lidiar con esto hay que vencer numerosos escollos.
Se debe de reconocer y etiquetar ese texto '*sucio*', estimando el idioma y sabiendo si '*eso*' es pronunciable o si es solamente ruido del tipo '*un gato caminando en el teclado*' , o si son siglas, etc.
Deben identificarse las palabras, locuciones y términos, números, horas y fechas, siglas, acrónimos y abreviaturas, unidades, monedas, fórmulas matemáticas y otros fenómenos escritos.
Debe analizar flexiones (*conjugaciones*) y estimar aceptablemente palabras parasintéticas y desconocidas (*fuera de vocabulario*).
Necesita corregir la mayoría de errores de ortografía más comunes y se debieran de ponderar las alternativas más '*humanas*'.
Es necesario reconocer un número importante de palabras, siendo a la vez ampliable en palabras y reglas. (*2M/español 200k/inglés*).
Nota: Todo esto debe funcionar rápido, ser portable y conservativo en términos de recursos, además de ser confiable y fácil de ampliar.

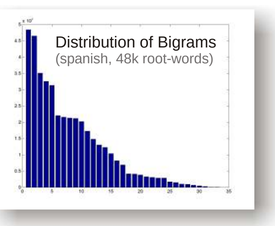




La distribución estadística de bigramas entre varios idiomas, provee valiosa información para estimar el idioma de una palabra o frase desconocida.



Spanish (ES) Bigram Extraction for Mis-Spelling Analysis



La distribución de bigramas y trigramas es una característica que se repite de forma similar en casi todos los idiomas. Por la forma en que se presenta, permite ser usada para determinar no solo el idioma sino para restaurar errores.

Sistemas Usados

Procesamento de afijos: Algoritmo de análisis de Prefijos, Sufijos e Infijos combinados basados en estándares y diccionarios de libre diponibilidad (*Open Office /ASpell / Ispell*)
Identifiación Estadística de Idioma [(1)]: Provee medidas para poder reconocer el idioma y si las palabras son pronunciables.
Correción de Ortografía y Restaurador Léxico: Usamos algoritmos estadísticos con heurística para identificar palabras erradas y brindar un 'paquete' de palabras posibles, incluyendo una estimación de cuán viable es cada una (*de ser la correcta*), para su uso y posible discriminación contextual, en una etapa posterior.
Restauración de palabras por Similitud Fonética: Basado en un algoritmo propio[(2)], se utiliza como parte integral del mecanismo de restauración léxica, incluyendo la medida de verosimilitud de cada una de las formas escritas o grafemas, en medida humana.
Analizador, Flexionador y Derivador Morfológico: Analiza, Compila, Expande, Minimiza y Edita las Palabras y sus Reglas .

Segmentación y Lematización

```
Hubo 28ºC bajo cero, dijo el dr. jones hoy 5:49 hs.

hubo  [VAIS3SF p:0.5 | VAIS3SM p:0.5]
28_DC0MP(Determinante,Cardinal,Masculino,Plural)
ºC_Zu(Cifra,Unidad+thermodynamic+temperature)
bajo [VIIP1SM p:0.029 | VTIP1SM p:0.029 | SPSMS p:0.74 | AQ0MS p:0.09
     | NCMS p:0.059 | bajó_VTIS3SM<*bajar,regular>~0.86 p:0.029
     | bajó_VIIS3SM<*bajar,regular>~0.86 p:0.029]
cero [NCMS p:0.6 | AQ0MS p:0.2 | DC0MS p:0.2]
._Fc(COMMA) dijo[VTIS3SM p:0.5 | VIIS3SM p:0.5]
el [DA0MS  | él_PP3MS~0.55 p:0.1] dr[Zu  | doctor_Yt<&dr>]
jones [Jonas_NPMS~0.55 p:0.0018 | nones_RN<*non>~0.97 p:0.0062
   | nones_NCCP<*non>~0.96 p:0.11 | nones_AQ0CP<*non>~0.96 p:0.035
   | Jonel_NPFS~0.55 p:0.0018 | jones_Xen<*jones>~0.93]~0.55
hoy_RT(Adverbio,Tiempo) 5:49_Wt(Fechahora,Hora)
hs [NCMSU p:0.028 | Zu p:1] ._Fp(DOT)
```

La separación en sílabas, palabras y símbolos de un texto no es una tarea trivial, más aún si se desean averiguar los lemas a la vez de tolerar errores y dar alternativas ortográficas. Normalmente esto es realizado mediante autómatas finitos (*llamados tokenizadores o lexers*) basados en lenguajes del tipo *expresión regular*, para reconocer y aislar las secuencias de caracteres que forman las palábras, siglas y números.
Para lograr esto, se ha portado *JLex* de *Java* a **C#** bajo **.NET 2.0**, creando una herramienta para Visual Studio.
La lematización con correccion ortográfica, en idiomas flexivos como el español es un problema ***NP-duro***, debido a la combinación posible de afijos, amentada por los errores.
La contribución es lograr esto en forma efectiva yeficiente.

Características Obtenidas

✓ Módulo analizador de multi-etiquetado robusto con corrector ortográfico, aceptando múltiples idiomas en una misma instancia.

✓ Etiqueta palabras, símbolos y locuciones mediante un estándar internacional, expandido semanticamente basado en *EAGLES 2.0* [(3)] incuyendo una salida adicional en norma *Penn TreeBank* (***en***).

✓ Reconoce números en formatos enteros, científicos y hasta hexadecimales, incluyendo Números Romanos como: *MCMXIV*

✓ Identifica Números dichos con palabras como '*dos mil ciento treinta con 30 céntimos*' tanto en inglés como en español. Adicionalmente cuando etiqueta cardinales, numerales y ordinales, provee el cardinal del cual provienen, muy útil para comprensión artificial basado en semántica, luego del procesamiento gramatical.

✓ Acepta en español e inglés, 925 unidades (SI, CGS, MKS, etc.) ~260 países ~1100 monedas (ISO 4127), ~6000 locuciones (***es***), fecha/hora bajo múltiples formatos y normas (ISO 8601, etc.)

✓ Controlado por aprox. 65 parámetros especialzados.

✓ Analiza la expasión de 'jerga' donde hay palabras mezcladas con números, en mensajes cortos y chat; por ej: **salu2 = saludos**

✓ La Restauración y Medición de Similitud Fonéticas para el Español están basados en estándares fonéticos Sampa[(2)], con reglas especiales para incluir palabras adoptadas de otros idiomas.

✓ Incluye índice fonético estándar *Soundex*, *Metaphone y Double-Metaphone*, útiles para Inglés y escasamente para el Español.

✓ Estima el idioma de palabras desconocidas, determinando si son pronunciables, siglas, abreviaturas, acrónimos o simple 'basura'.

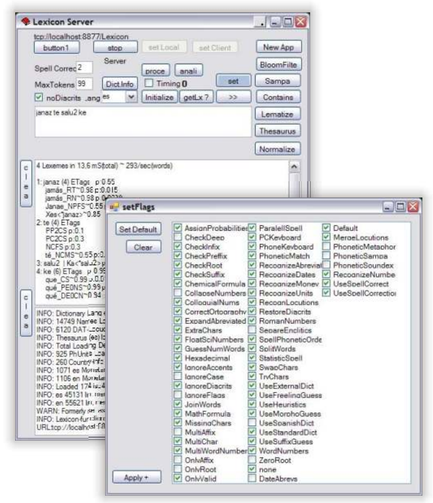
✓ Provee varias alternativas ortográficas, aún si no hay errores.

✓ La Lematización, Búsqueda y Restauración son independientes de las marcas diacríticas, siendo éstas un motivo frecuente de errores logrando un aumento radical de la velocidad del proceso.

✓ Escaso uso de RAM ~20Mb para 80k raíces y 4k reglas; reconoce aprox. $3x10^6$ palabras y casi ilimitadas en modo parasintético con palabras fuera de vocabulario OOV. (*Out Of Vocabulary*).

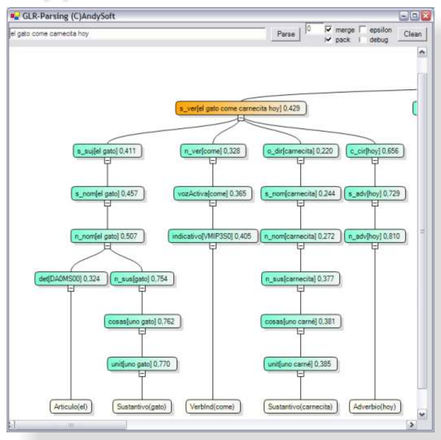
✓ La velocidad resultante[(5)] del analizador morfologico es de 80 mil palabras por segundo, bajando a 2 mil/segundo cuando se pide el árbol de flexión/derivación, con todas las alternativas y raíces.
La restauración fonética logra ~1000/sec. y ~30/sec para errores.

App.: Servidor Léxico



El servicio creado permite distribuir los recursos de un sistema de procesamiento de lengua, sistema mediante un servidor remoto quien mantiene actualizadas sus bases de datos automáticamente en un solo lugar, proveyendo de funciones léxicas a todo módulo que lo solicite, mediante protocolos de alta performance, evitando la duplicación de recursos, dado que el consumo de memoria de las aplicaciones lingüísticas suelen ser considerables.
A la vez permite configurar cada instancia de su uso en forma independiente, mediante un conjunto de parámetros especializados enviados en cada requerimiento, pudiendo de este modo procesar múltiples idiomas y modos de operación en forma eficiente.

App.: Parser GLR Robusto



El análisis gramatical y semántico suele ser un problema complejo cuando el número de terminales (*tokens*) es alto.
Nuestro modelo para Español posee más de 700, dada la gran cantidad de conjugaciones verbales y flexiones posibles. Hemos modelizado esta gramática española con un parser de tipo GLR el cual hace un buen análisis sin perder generalidad ni velocidad, pese a ser NP-duro.

Educación y Cultura

Diccionario y Traductor x SMS

Se ideó un diccionario y traductor español-inglés, que reconoce conjugaciones y provee aproximación ante palabras mal escritas.
Enviando un mensaje con ENES ESES y ESEN y la palabra al 55588, disponible en Argentina y Ecuador.

App: Enseñar a Escribir

VAHIEMA → BALLENA (d~0.69 dT: 1 ms)
VAYEMA → BALLENA (d~0.29 dT: 1 ms)

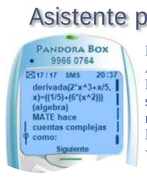
La Similitud Fonética tiene un costado inesperado: permite diseñar algoritmos eficientes para corregir palabras mal escritas por personas que recién estan aprendiendo la relación entre los sonidos y las letras. Esto se basa en la ´proximidad´ fonética con palabras reales del diccionario. Esto combinado con sistemas pedagógicos, consitye un poderoso sistema de Enseñanza y Asistencia Pedagógica en donde las computadoras ya tendrán razgos mas humanos y afables, reduciendo la brecha cultural y digital

Aplicaciones Móviles

CHE: Asistente Educativo

Hay una prueba-demo de un Asistente Virtual, capaz de responder preguntas simples, hacer cuentas de matemáticas, dar algunas definiciones de palabras, mediante un mensaje al 55588

Asistente para Matemáticas x SMS

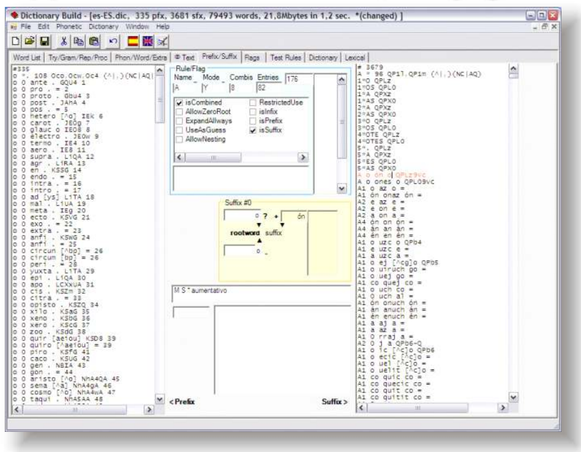


Desarrollamos un asistente para Matemáticas, Análisis, Álgebra, Probabilidad y Estadística, etc. Posee una gran librería de funciones científicas y se usa mandando un mensaje de texto. Como respuesta agrega enseñanzas relacionadas.
Enviando MATE y la expresión al 55588.
+info: www.pandorabox.com.ar

App.: Restauración de Ortografía

La reparación de formas mal escritas o corrección de ortografía, normalmente se basa en permutación y sustitución de letras, pero es difícil saber cual de todas esas permutaciones conviene hacer antes, pues el número de permutaciones y sustituciones es un problema NP-duro. Hemos avanzado creando algoritmos de predicción '*blandos*' que ponen el ojo en donde es más probable que haya un error, similar a lo que un '*humano*' siente intuitivamente.

App.: Diccionario Flexivo/Conjugador

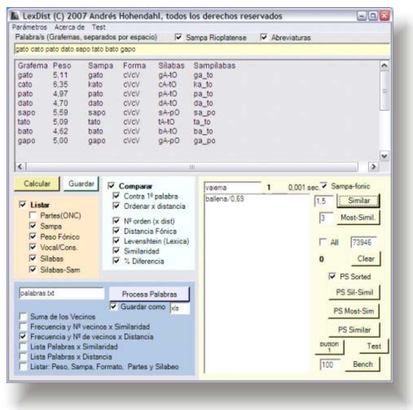


La creación y edición de un diccionario morfológico es una tarea compleja, pues requiere de ingresar las palabras, sus etiquetas, expresar sus reglas morfológicas en forma precisa, luego probar el resultado, importar y analizar listados de palabras por lotes, etc.
Se han creado utilitarios muy específicos para poder crear, editar y enriquecer estos recursos con un mínimo esfuerzo.

Búsqueda Inteligente en Bibliotecas

Se ideó un sistema de ***Búsqueda de Referencia Inteligente para Bibliotecas*** y otras fuentes de cultura y conocimiento, permitiendo búsquedas inteligentes con capacidad conceptual, prescindiendo de ortografía estricta inclusive para nombres propios.
Se usaron los algoritmos desarrollados: restauración fonética y ortográfica, lematización, tesauros ontológicos, a la vez de valoración por Tf*Idf, etc.
El acrónimo IOPAC proviene de "*Internet Open Public Access Catalog*"

App.: Similaridad Fonética



Se han creado ´utilitarios´ académicos para poder trabajar con los diferentes algoritmos de similitud existentes, compararlos entre sí, modelar sus múltiples factores y generar luego las planillas de excel para realizar el trabajo estadístico y gráficas.

Ej.: Análisis Morfológico Robusto

```
vimo hoi ezpe kavrom kon eza vayema ozpitalaria ke nempekapo salu2 klhdrdzkuio

vimo vino_VPIS3SM<*venir>~0.98 p:0.00000019 | vimo_Xes<*vimo>~0.83 p:0.036
     vino_VIIS3SM<*venir>~0.98 p:0.00000019 | vino_NCMS~0.98 p:1
hoi  hoy_RT~0.99 p:0.0000048 | ohm_NCMS0h~0.95 p:0.0000048 | Ho_NPMS~0.55 p:0.9 |
     hot_AQ0MS~0.89 p:0.0000048 | hoi_Xen<*hoi>~0.84 p:0.9
ezpe éste_PD0MS~0.98 p:0.03 | este_DD0MS~0.98 p:0.9 | este_NCMS~0.98 p:0.0075 |
     este_PD0MS~0.98 p:0.03 | Xes<*ezpe>~0.93 p:0.0075
kavrom cabrón_AQ0MS~0.96 p:0.76 | cabrón_NCMS~0.96 p:0.23 | Xde<*kavrom>~0.84 p:0.076
kon_con_SPSMS<%mode>~1 p:0.08 | kon_Xde<*kon>~0.9]~0.997
eza esa_PD0FS<*eso> p:0.076 | esa_DD0FS<*ese> p:0.9 | Xes<*eza>~0.93 p:0.076
vayema [ballena_NCFS0z~0.99 p:0.19 | ballena_AQ0FSL<*balleno>~0.99 p:0.62 |
       ballena_NCFS0L<*balleno>~0.99 p:0.19 | vayema_Xes<*vayema>~0.88 p:0.062
ozpitalaria [hospitalaria_AQ0FS<*hospitalario>~1 p:0.039 | ozpitalaria_Xes<*ozpitalaria>~0.88]~0.995
ke  _que_CS~0.99 p:0.011 | qué_PE0NS~0.99 p:0.032 | qué_DE0CN~0.94 p:0.000017 |
       qué_PT0CNN~0.94 p:0.0015 | que_PR0CNN~0.94 p:0.009 | Xes<*ke>~0.55
nempekapo  mentecato_AQ0MS~0.95 p:0.76 | mentecato_NCMS~0.95 p:0.23 |
       Xen<*nempekapo>~0.789 p:0.076
salu2  saludos_NCMP<*saludo> | p:0.09 Ka<*salu2> p:0.9
klhdrdzkuio  _B(BadWord)~0.875
```

Se muestra el resultado de un análisis morfológico de palabras muy mal escritas, las etiquetas son EAGLES 2.0, adicionadas con semántica, verosimilitud y probabilidad. La capacidad de restauración parece "casi humana", hasta con 50-60% de letras mal.

Líneas de Trabajo

Compilador para motor de Diálogo Natural (DDL + C# > MSIL)
Modelización de procesos cognitivos para diálogo (Dialog Objects)
Parsing Condicional (Enriquecido con ontologías y contexto)
Compilador/Parser/Chunker NLP/GLR+ (Con Scrödinger Tokens)
Modelos Cognitivos (Matemáticas, Conjuntos, Lógica & Unidades)
Ambiente de Ejecución Cognitivo (DDL Dialog Objects Runtime)
Comprensión Artificial Superficial por Inferencia Morfológica.
Resolver Deixis, Co-Referencia y Anáforas (yo/mi/mío, tú/su/suyo..)

Referencias

(1) *A. Hohendahl, J. Zelasco "Efficient algorithm for fast language detection applied to Natural Language Processing"* WICC2006 - ISBN 950-9474-35-5 Art 694. Univ. Morón, Bs.As. Arg.

(2) *Hohendahl, A.T.; Zanutto, B. S.; Wainselboim, A. J.; "Development of a Phonetic Similarity Algorithm for written text"* SLAN2007. X Latinoamerican Congress of Neuro-Psycology 2007

(3) Our Language detection (1) provides useful information to determine if it is worth trying to correct misspelling or if the input is in some foreign language or is simply garbbage.

(4) EAGLES 2.0 / Parole enhanced by means of adding infix semantic classification and with combination of more than 200 specialized tags, useful for text understanding.

(5) Desktop System: XP Sp3 .NET 2.0 2Gb RAM 2.6Ghz i386 E5200 (using single processor core)

Aplicaciones Potenciales

Asistencia/Enseñanza: Diálogo Robusto con los más Pequeños.
Reducir la Brecha Digital: Sistemas de Referencia Inteligentes.
Medicina: Historias Clínicas Médicas Digitales Interactivas.
Bases de Datos: Limpieza de Errores y Registros Duplicados.
Comprensión Artificial: Extracción/Inferencia Semántica.
Síntesis/Reconocimiento de Voz (usando diccionario morfológico).
Analisis Gramatical y Semántico Robusto (deep parsing).
P.L.N. Robusto, incluyendo OOV (Out Of Voccabulary words).