Reconocimiento y Corrección de Errores por Ingreso de Lenguaje Natural en Dispositivos Electrónicos

Andrés T Hohendahl [1,2]

Director Dr. José F. Zelasco [1]

Codirector Dr. Ing. Silvano B. Zanutto [2]

1 Laboratorio de Estereología y Mecánica Inteligente
Facultad de Ingeniería, U.B.A.

2 Instituto de Ingeniería BioMédica,
Facultad de Ingeniería, U.B.A.

Sistema de Comunicaciones
Componentes

- Datos: Palabras en un Lenguaje Natural
  - Formatos: fonético y grafémico
- Emisores
  - Moduladores: Voz, Escritura, Pantallas, Máquinas, etc.
- Transmisión
  - Medios: Luz, Sonido , Relieves de Superficie, etc.
- Receptores
  - Demoduladores: Oido, Vista, Tacto (biológicos)
- Almacenamiento
  - Artificial: Escritura (físico), Audio/Video (electrónico)
  - Memoria Humana: Neurológico (electroquímico)

→ Existen Errores! Y de muchos tipos diferentes!

Sistema de Comunicaciones
Análisis del tipo de Dato

- Datos
  - Palabras en un determinado Lenguaje
    - Formatos
      - Fonético (sonidos característicos)
      - Grafémico
        - » Alfabético (Letras que los Representan Combinadas)
        - » Ideográfico (Símbolos Visuales Complejos: Chino)
        - » Icónico (representación más compleja)
    - Códigos

      Reglas, Lógica, Cognitiva (hipótesis)

      ¿ Son sistemas Robustos ?

      ¿ Hay posibilidad de Métricas ?

Sistema de Comunicaciones
Tipos de Emisores de Datos

- Emisores

  Considerando como Fuente Originaria el Humano

  Cognición
  - Generación de Lenguaje Natural
    - Redacción del texto (ideas → secuencia de palabras y signos)
    - Reglas del lenguaje, escritura y habla

  – Moduladores Naturales
  (Los más Comunes)
  - Aparato Fonoarticulador (habla)
  - Mano (escritura caligráfica)
  - Manos (escritura mecanográfica)
  - Cuerpo y Manos (gestos: amslan)

Introducción de Datos

Interfaz Hombre-Máquina (HCI)

Directo

Movimiento & Posición

- Teclado (QWERTY, Numérico, etc.)
- Mouse (tradicional, posición y accionamiento: clic)
- Táctil (Touch Screen / Force Feedback)
- Acelerómetros, Brújula-3D, GPS

Indirecto

Reconocimientos (algoritmos ~ Inteligencia Artificial)

- Lectura de Texto (OCR)
- Auditivo: ASR (Voz a Texto), Entonación, Emoción
- Visual: Caras, Gestos y emoción, etc.
- Lápiz Electrónico: Escritura, Dibujo, Firma y Estilo
- Biométrico: Huella digital, Iris, etc.
- Químico: Gases, Humo, Olores, etc.

Sistema de Comunicaciones
Transporte y Almacén

- Transmisión
  - Física: Luz, Sonido y Tacto (braile)

    (No nos adentraremos en esto)
- Almacenamiento y Reproducción

  (No nos adentraremos en esto)

  - Humano
    - Memorias Léxica
    - Memoria Fonética
  - Artificial

    - Libros y Escritos
      - » Grafemas e Imágenes estáticas
    - Audio y Video
      - » Instantáneas de su emisión.

Sistema de Comunicaciones

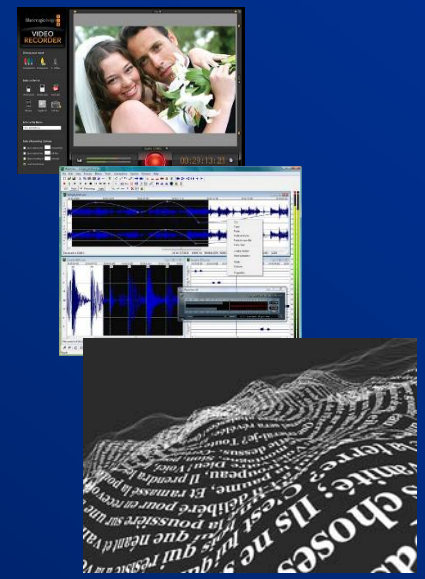
Reproducción

Mecanismos = Físicos

- Video: Luces, CRT/LCD, Proyección, algún día.. 3D real.
- Audio: Sonidos, Música, Efectos especiales: 3D (HRTF)
  - → Habla: TTS (texto a voz), con prosodia y emoción.
- Mecánicos: Vibración, Movimiento & Accionamientos
- Químicos: Perfumes, Olores..?

Elaboración ? Mímica !

- Fotos & Videos
  - Realidad (Memoria Visual y Auditiva)
  - Realidad Virtual (Simulación Visual y Auditiva)
  - Realidad Aumentada (Suma Datos Representados)
- Música & sonidos
  - Pre-grabados/compuestos y/o sintetizados.
- Textos
  - previamente escritos... (por otros humanos)
  - GLN (Generación de Lenguaje Natural)



Sistema de Comunicaciones

Mecanismos de Recepción

- Receptor
  - Multi-Etapa (cascada)
    - Demodulador (parte 1)
      - Oido (acústico-biológico)
      - Vista (visual-biológico)
      - Tacto (sensitivo-biológico)
    - Demodulador (parte 2)
      - Reglas Automáticas (neuro-biológicos)
    - Demodulador (parte 3)
      - Reglas Cognitivas (neuro-psicológicas)
        - Lógica y Sentido Común
        - Concordancia: Género, Número, Tiempo
        - Ubicación
          - » Deixis, Anáfora, Coreferencia

Sistema de Comunicaciones
Teorías a Analizar

- Texto Escrito
  - Análisis de códigos y robustez
    - Propios del Lenguaje y la Escritura
  - Caracterización de los Errores
    - Usando modelos y técnicas de telecomunicaciones
  - Medición
    - Similitud Fonética
    - Similitud Léxica
    - Matriz de Confusión
    - Estimación de Idioma y su Pronunciabilidad
    - Diversas Tasas de Error

Qué cosa quiso poner...?

Los Errores de Ortografía

Todo segmento de texto no es más que una chorrera de Letras...
Si no está en el diccionario → No se sabe nada a priori, por ej.:

K A V R H O M (7 letras)

~ Hay 35 letras diferentes c/signos diacríticos (acentos/diéresis/eñe)
- » Cambiar 1 letra ~ 7 x 34 ~ 234 intentos de búsqueda en diccionario
- » Cambiar 2 letras ~ 7 x 34 x 6 x 34 ~ 5 x 10^4 intentos
- » Cambiar 4 letras ~ 1.2 10^{12} intentos (sin contar eliminaciones ni inserciones)
- » Cambiar 7 letras ~ 9 x 10^{15} intentos (incluyendo duplicados)

Es un problema complejo y de orden combinatorio!

→ **NP Duro** = IRRESOLUBLE EN TIEMPOS RAZONABLES (POLINOMIALES)

Tampoco sabemos
- » El idioma (hoy todo es multilingual-mixto)
- » Si existe, en cuál diccionario está, si es un apellido o nombre propio.
- » Cuáles letras están falladas / faltan o sobran (debo probar cada letra)
- » Si se Invirtieron algunas letras y cuáles

→ No sabemos si se puede reparar en un tiempo/costo razonable
→ Si Hallamos varias alternativas (es probable)
- → No sabremos jamás cual de ellas puede ser la más apropiada

→ Pero: un Humano resuelve esto en forma intuitiva y sin pensar mucho!

¿ Será **CABRÓN** ? (6 letras: 4 cambios + 1 eliminación)

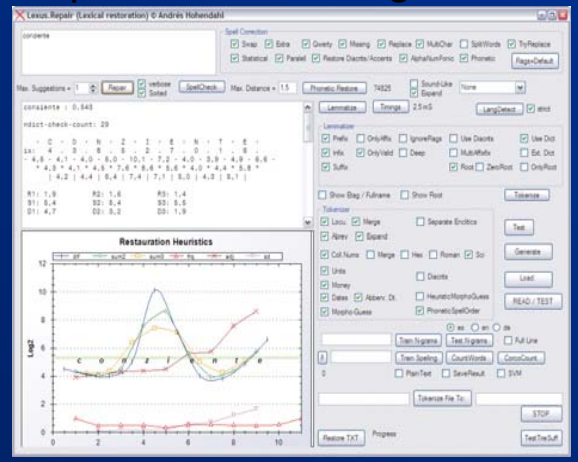
Inteligencia Artificial I

Humanizando la PC

Hooolaa

"tal vez quiso escribir 'hola'.."

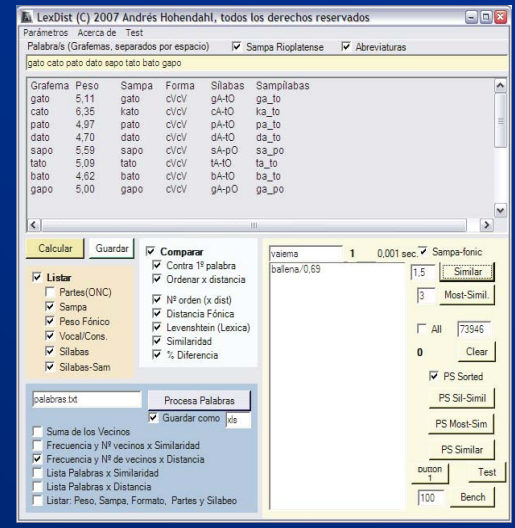
Reparación de Ortografía



vahyema

"esta suena como ballena.."

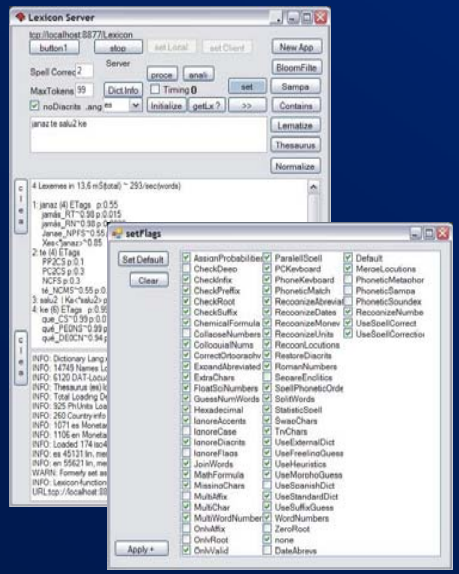
Similitud Fonética



Hoy a las 2 y media am

"puso una fecha/hora.."

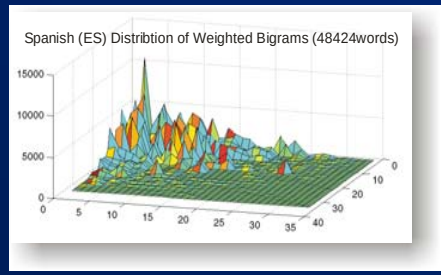
Servidor Léxico



Hohendahl

"esto parece alemán.."

Estimador de Idioma



Sistema de Comunicaciones
Implementación

- Algoritmos Analizador y Corrector
  - Conocer Informción a priori
    - Tipo de Dato: numérico, palabras, frases, etc.
    - Idioma
  - Corrección de Errores
    - Usando modelos y técnicas de telecomunicaciones
    - Imitando el modelo de restauración humano
    - Prediciendo la más probable (cognitivamente)
- Programación de Módulos
  - Lenguajes Modernos, Portátiles y Eficientes
    - Compatibles con Hardware (embedded)
    - Optimización de velocidad vs. uso de recursos (RAM/CPU)

Inteligencia Artificial II

Reparando Horrores de Ortografía

Análisis Morfológico y Lematización Robusta

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

Lineas de Desarrollo

Human.**C**omputer.**I**nterface.

- Modelización Cognitiva para reglas complejas
  - Concordancias, Coreferencias, Deixis
- Sistemas de Diálogo Artificiales
  - Compilador GLR Robusto (Tomita c/Scrödinger Tokens)
  - Run-Time Cognitivo
    - Implied verbal Logic (Math, Set & Boolean Logic)
    - Simple Scientific Math (numeric + algebraic)
    - Scientific Units Cognitive Operations
    - Artificial Shallow Understanding
  - Extracción de Información en OOV. & mistyped words (morphologically correctly constructed, even with errors)
  - Resoluciónd e Espacio de Conversación (yo, tu, él, aquello)
  - Navegación Ontológica y Reglas de Sentido Común

Preguntas?

Gracias!

Andrés T. Hohendahl
andres.hohendahl@fi.uba.ar

blog: web.fi.uba.ar/~ahohenda