Investigación + Desarrollo
Asistiendo a la Enseñanza
con
Procesamiento de Lenguaje Natural

Andrés T. Hohendahl [1,2,3,4]

1 Laboratorio de Estereología y Mecánica Inteligente
Facultad de Ingeniería (tesista), F.I.U.B.A.

2 PandoraBox, (director) www.pandorabox.com.ar

3 IByME (Instituto de Biología y Medicina Experimental
U.B.A. + Conicet) (investigador externo)

4 IIBM (Instituto de Ingeniería Bio Médica, F.I.U.B.A.)
(ex-integrante)

Introducción

- Computadoras
  - Video/Audio/Interacción
- Interfaz Hombre-Máquina
  - Teclado, Mouse, Pantalla y Parlantes
- Objetivos
  - Reducir la brecha de aprendizaje.
  - Mejorar la Efectividad de la Tecnología para la Enseñanza
- Fallas frecuentes por
  - Interacción compleja (click!, selecc. opc. A en menu2..)
  - Mala Ortografía y Escaso Conocimiento de Palabras (Léxico)
  - Si hay Help, generalmente no explica lo que se busca..!

Niños y la Enseñanza

Los Niños Aprendiendo

Importante Espíritu Lúdico

- Suelen querer Jugar todo el tiempo.

Distracción Frecuente

- Es difícil mantenerlos con la atención puesta en algo.

Son Excelentes Descubridores

- Pero piensan como niños, por eso fallan.

No hay conocimiento “de “base”

- Suelen “probar” y si no funciona enseguida, pierden el interés.

Necesitan Estímulo y Desafío

- Nada es más aburrido que leer un libro o consultar una enciclopedia (aunque sea multimedia) – fallo: MS-Encarta!

La Enseñanza con Informática Hoy

La interacción actual, se basa en “cosas pre-armadas”

- Libros Digitales

  Si bien representan el futuro, no alcanzan para motivar.

- Libros Interactivos

  Son pocos, suelen ser para los muy chicos (hasta 4/5años), básicamente poseen con interacción gráfica y personajes.

- Enciclopedias Digitales

  Son inmensos repositorios de conocimiento, desperdiciados por pobres interfaces con el usuario. Es complicado usarlas. Quien busca debe saber exactamente como hacerlo.

- Juegos de Computadoras

  Pocas veces persiguen fines didácticos, suelen ser de acción.

Un Problema Clásico

Los Diccionarios, Enciclopedias y Sistemas de Búsqueda por lo general sólo reconocen palabras en modalidad raíz (lema) o verbos en infinitivo. Ejemplo:

Un niño busca una palabra que no entiende en el diario:

..y el haz lo cegó rápidamente mientras...

Va a la computadora y pone en el teclado la palabra nueva

CEGO

Un sistema normal no puede entender esto ¿saben de alguno?

Bien, esto sólo es resoluble por un Lematizador Morfosintáctico

Hemos desarrollado algo que reconoce sin lugar a duda:

CEGÓ [CEGAR] verbo intransitivo, 3º persona del pretérito. corrigiendo de paso la falta del tilde.

Otro Problema Frecuente

Las Computadoras sólo hallan lo correctamente escrito, o en su defecto, que empiece con las mismas letras.
Ejemplo:

Un niño oye en un programa de TV “las Ballenas del Sur”
Va a la computadora y pone en el teclado la palabra nueva

VAYEMAZ

Un sistema normal no puede entender esto ¿saben de alguno?
Bien, esto sólo es resoluble por aproximación fonética.
> Hemos logrado un sistema que reconoce sin lugar a duda **BALLENAS** Corrigiendo hasta 5 de 8 letras (62% de error) en apenas 0.001 seg. en una computadora de escritorio.

Una Propuesta Diferente

Que las Computadoras sean mas Humanas y amenas

- Diálogo Natural, como una persona

  Que el niño pueda “preguntar algo” igual como al docente.

- Respuesta Elaborada

  La respuesta debe adaptarse a la pregunta y no forzar la pregunta para obtener una respuesta pre-armada.

- Personalidad Virtual

  Un sistema debiera de responder como un humano, y con paciencia, pero con límites acorde a las circunstancias.

- Un Personaje Virtual Es mas Interesante

  Si el niño descubre que no está hablando con una máquina “tonta” sino con un ente inteligente que posee virtudes casi humanas, pondrá interés, por el mero hecho del desafío.

Sistemas Propuestos I

Diccionarios a los que se les pueda “preguntar” de igual modo que a una persona o profesor. Ej:

- Diccionario Inteligente

  ¿Qué es una cabreadita?

  CABREADITA es el diminutivo de CABREADA que significa...

  Traducción y Detección de Idiomas

  ¿Qué significa worried?

  WORRIED es el pasado de WORRY (en inglés) que significa..

  ¿Cómo se dice perrita en inglés?

  PERRITA es el diminutivo femenino de PERRO, que en inglés se dice “dog”. En inglés los diminutivos se forman agregando adjetivos que indican la condición de pequeño: “little dog” o “small dog”, menos veces cambiando la palabra: “doggy”...

Sistemas Propuestos II

Que se les pueda “preguntar” como a una persona.

- Profesor de Matemáticas Virtual

  Cuánto es 2 * 3 + 5 / 3

  Derivada de 2*x^2 +3/(x+1) respecto a x, evaluada en 3

  Quiero un número primo de 8 dígitos

  Cuánto es el factorial de 23

  Graficar y = 2*x+cos(1/x) entre 3 y 5.5

  Cuánto es MXMCII + 50

  Cuánto es la raíz cuadrada de 500

  Que dé respuestas concretas, reconociendo errores de sintaxis en fórmulas e indicando sugerencias a preguntas mal redactadas, enseñando algo con cada respuesta.

Sistemas Propuestos III

Juegos de Rol Didácticos e Inteligentes

Desafíos y metas, con los que se deba dialogar, haciendo consultas y respondiendo a preguntas del juego, mejorando la experiencia interactiva y la fijación en memoria con fotos, audio y videos. Ej:

- Geografía: Descubrir lugares remotos en forma virtual, averiguando cosas, leyendo mapas y viendo fotos, tal como una visita real. Inclusive hablando con personajes lugareños.
- Historia: Participar de la historia, hablar con los próceres, tener diversas “misiones”, lograr metas con puntajes conforme se respondan las preguntas y enigmas de la historia.
- Ciencias: Personificar por ejemplo a un Naturalista, quien descubre y clasifica especies nuevas, dialogando con asistentes y colegas averiguando cosas afines, viendo imágenes, etc.
- Matemáticas: Resolver problemas simples planteados por situaciones reales teniendo un asistente con quien dialogar.

Desarrollos I

Lematizador y Flexionador Morfológico

- Reconocimiento de Idioma (Alemán/Inglés y Español)
- Unidades, Monedas, Fechas, Números
- Etiquetado estadístico de palabras desconocidas
- Reconoce > 3M de palabras, incluyendo parasíntesis
- Etiquetado múltiple por reglas externas.
- Alta Velocidad: >3000 palabras/seg. con .NET 2.0
- Provee ranking de etiquetas (Verosimilitud)

Editor de Diccionarios Morfológicos y Flexivos:

- Compilado de Diccionarios (listas de Palabras)
- Importación de Formatos GNU ASPELL e ISPELL
- Genera y prueba reglas de flexión combinables
- Manipulación de diccionarios por combinación de Etiquetas EAGLES y Morfología usando expresiones regulares
- Importa Formatos de Flexión IFFIX (.AFF)
- Soporta Etiquetas EAGLES 2.0 ampliadas y sin restricciones.

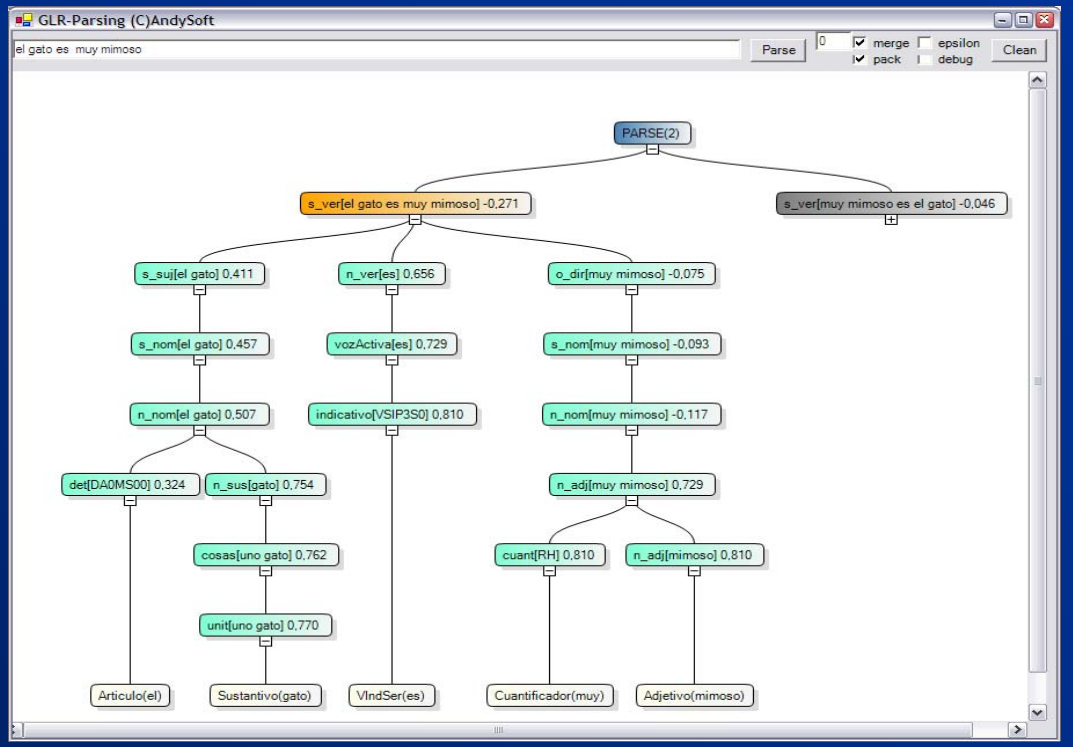
Desarrollos II

AndyBot: Sistema de Desarrollo (herramienta) para crear Agentes Inteligentes

- Permite Crear Personajes para dialogar en Lenguaje Natural (cotidiano) creando: asistentes, profesores, juegos, etc.
- Utiliza reconocimiento lingüístico complejo, con corrección de errores automática y análisis de múltiples alternativas viables.
- Simple: Factible de configurar sin conocimientos especiales.
- Seguimiento de hilos de conversación complejos.
- Sistema completo de desarrollo vía WEB (online)
- Dialoga con los motores más comunes de Chat: MSN. GTalk, ICQ
- Disponible como servicio on-line, las 24 horas. 24x7x365
- Útil: Capacidad de miles de usuarios simultáneos.
- Veloz: Responde generalmente en fracciones de segundo.
- Ideal para asistencia a la enseñanza y juegos didácticos.
- Conectable a bases de datos externas.
- Personalizado: recuerda datos específicos de cada usuario.

Desarrollos III

(1) Prototipo Funcionando



Parser de Español (1)

Etiquetado P.O.S. múltiple.
Reconoce Idiomas (A.L.R.)
Ranqueo (N-best) basado en heurística+contexto del A.S.T.
Reconocimiento de entidades complejas, etiquetado de palabras desconocidas.
Operaciones Cognitivas básicas: (asociación, operaciones matemático-lógicas), Reconoce Unidades Científicas (convierte y opera). Entiende y resuelve fechas gregorianas (relativas y absolutas), reconoce notación horaria (iso.8601). Entiende y maneja números en formato romano, científico, hexadecimal, con palabras y combinados.
Usa una gramática española-castellana importante (~500 producciones en ABNF)

Desarrollos IV

Aplicaciones Comerciales Funcionando

- Diccionario y Traductor inglés/español por SMS c/reconcimiento de idioma y en caso de error, dá una lista de palabras sugeridas que "suenan parecido". Traduce 87k lemas expandidos a más de 1.8 millones de términos y locuciones. Hace cuentas científicas, números romanos y traduce literales numéricos, reconoce errores brindando alternativas.

  Funcionando en Argentina (55588) y en Ecuador (8086)

  Keywords → ESES, INES, ESIN y DIC al 55588 (Argentina)

- Profesor de Matemáticas por SMS opera con 150 dígitos exactos (enteros), operaciones con fracciones, vectores, escalares y matrices. Resuelve y Simplifica Ecuaciones literales, Halla integrales y derivadas literales y definidas, hace expansiones, etc. (*+provee enseñanzas de matemáticas con cada respuesta!)*

  Keyword → MATE al 55588 (Argentina)

Líneas de Trabajo en Curso I

- Reconocimiento de Texto (Ordenes, Diálogo, Frases)
  - Aplicación en ASR (corrección de errores e inferencia)
  - Aplicación en TTS Mejora de Calidad (envolventes prosódicas)
- Modelos Cognitivos
  - Modelo de Objetos Gramático-Semánticos
  - Resolución de Concordancia y Anáfora en Contexto
  - Solución a Concordancias de orden superior (Ontológicas)
- Gramática, Sintaxis y Semántica Española
  - Definición del español en BNF + extensiones (concordancia, etc.)
  - Lexicografía (diccionario flexivo con >2 M palabras para POS)
  - Integración de Ontologías al Parser (CYC, WordNet, otras)
  - Parsing + WSD basado en combinación de técnicas.

Líneas de Trabajo en Curso II

- Comprensión y Compilación de Lenguaje Natural
  - Parser Estadístico/Especulativo + No-Determinístico (Híbrido)
    - Aprendizaje no Supervisado de modelos de Gramáticas (on-the-fly)
    - Parsing Sujeto a contexto con mejora en POS y WSD
  - Modelo de Ideas como Objetos Operativos
  - Modelo Matemático-Cognitivo (Ideas)
  - Creación de un Run Time (RT) Cognitivo

    Con modelos matemático + estadístico-especulativo subyacentes

- Respuestas a Preguntas (Question Answering) Q.A.
  - Análisis de estructuras gramaticales emergentes de A.S.R.
  - Identificación de entidades (NE) y lógica de preguntas
  - Conversión de Texto a Ideas (TTI) e Ideas a Texto (ITT)
  - Vínculos con bases de datos, mecanismos de inferencia
  - Indexado de conceptos basados en texto (Ideas)
  - Creación de Bases de Conocimiento para Q.A.
  - Reconocimiento y Extracción de Ontologías

Andrés Tomás Hohendahl

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

www.**PandoraBox**.com.ar

andres.hohendahl@fi.uba.ar