Procesamiento de Lenguaje Natural:

¿Cómo entienden las computadoras el lenguaje humano?

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora

Section 1

Introducción

¿Qué es el Procesamiento de Lenguaje Natural?

El **Procesamiento de Lenguaje Natural (PLN)** es un campo interdisciplinario que combina:

- Ciencias de la computación
- Lingüística computacional
- Aprendizaje automático e Inteligencia artificial

Objetivo: Permitir que las computadoras comprendan, interpreten y generen lenguaje humano de manera útil.

El Lenguaje Natural

Un **lenguaje natural** es cualquier lenguaje que surge espontáneamente en una comunidad humana a través del uso, la repetición y el cambio.

Características principales:

- Evoluciona naturalmente (no es diseñado) ...
- Se distingue de lenguajes formales (programación, lógica)
- Es inherentemente ambiguo y contextual

Niveles de Análisis en PLN

El análisis del lenguaje natural ocurre en múltiples niveles:

- **Fonético/Fonológico**: Sonidos del habla
- **Morfológico**: Estructura de las palabras
- **Léxico**: Vocabulario y significado de palabras
- **Sintáctico**: Estructura gramatical de oraciones (*Mi niñas duerme tranquilo*)
- **Semántico**: Significado de oraciones y textos (*Incoloras ideas verdes duermen furiosamente*)
- **Discursivo**: Coherencia entre oraciones y párrafos

Principales Desafíos del PLN

¿Por qué es difícil procesar el lenguaje natural?

- **Ambigüedad**: Una palabra puede tener múltiples significados
  - *Banco* (institución financiera vs. asiento)
  - *Vi al niño con los binoculares* (¿quién tiene los binoculares?)
- **Variabilidad**: Dialectos, jerga, estilos personales
- **Contexto**: El significado depende de la situación
- **Implicitación**: Mucha información queda sobreentendida

Principales Desafíos del PLN

- **Correferencia**: ¿A qué se refieren los pronombres?
  - "María le dio un libro a Ana. Ella lo leyó rápidamente."
- **Lenguaje figurado**: Ironía, sarcasmo, metáforas
- **Ruido**: Errores tipográficos, gramaticales
- **Multilingüismo**: Textos en varios idiomas
- **Evolución constante**: Nuevas palabras, expresiones

Principales Tareas del PLN

Tareas de comprensión:

- Análisis de sentimientos y clasificación
- Modelado de tópicos
- Identificación de autoría
- Extracción de información
- Reconocimiento de Entidades Nombradas (NER)

Tareas de generación:

- Traducción automática
- Generación de texto
- Resumen automático
- Paráfrasis

Principales Tareas del PLN

Tareas de comprensión:

- Análisis de sentimientos y clasificación
- Modelado de tópicos
- Identificación de autoría
- Extracción de información
- Reconocimiento de Entidades Nombradas (NER)

Tareas de generación:

- Traducción automática
- Generación de texto
- Resumen automático
- Paráfrasis

Enfoques y Soluciones

La complejidad de la solución depende de la tarea:

- **Enfoques basados en reglas**: Expresiones regulares, reglas gramaticales
- **Métodos estadísticos**: *N*-gramas, modelos probabilísticos
- **Aprendizaje automático**: SVM, árboles de decisión
- **Deep Learning**: Redes neuronales, transformers, BERT
- **Large Language Models**: Claude, GPT, DeepSeek, Kimi, ...

Usar la herramienta adecuada para cada problema.

Aplicaciones del PLN en el Mundo Real

Vida cotidiana:

- Asistentes virtuales
- Traducciones automáticas
- Corrección ortográfica
- Búsquedas web

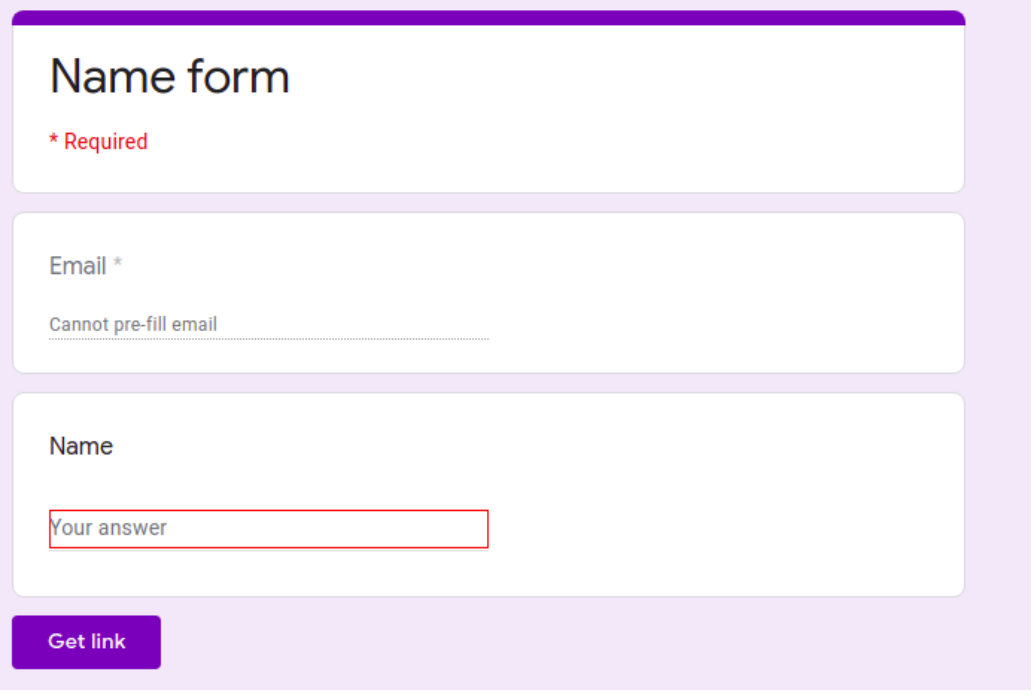
Industria:

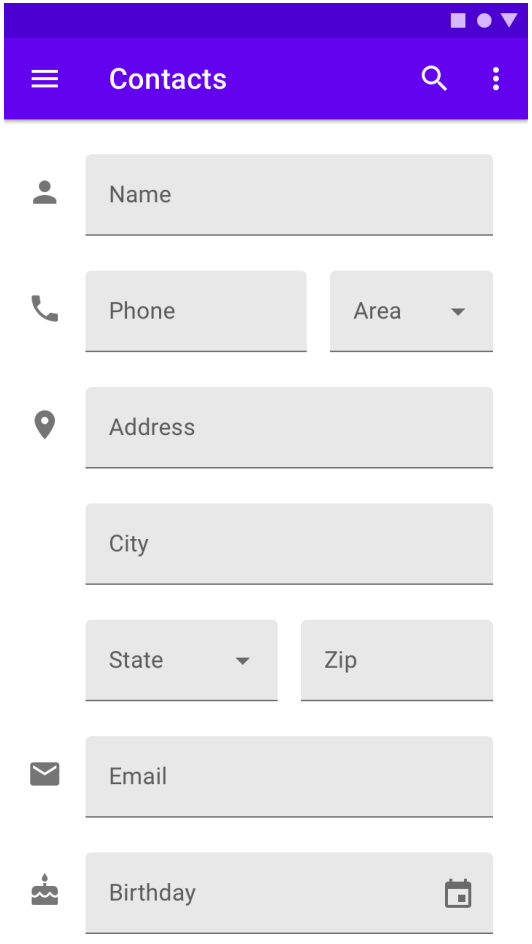
- Análisis de redes sociales
- Atención al cliente automatizada
- Análisis de documentos legales
- Investigación biomédica

Section 2

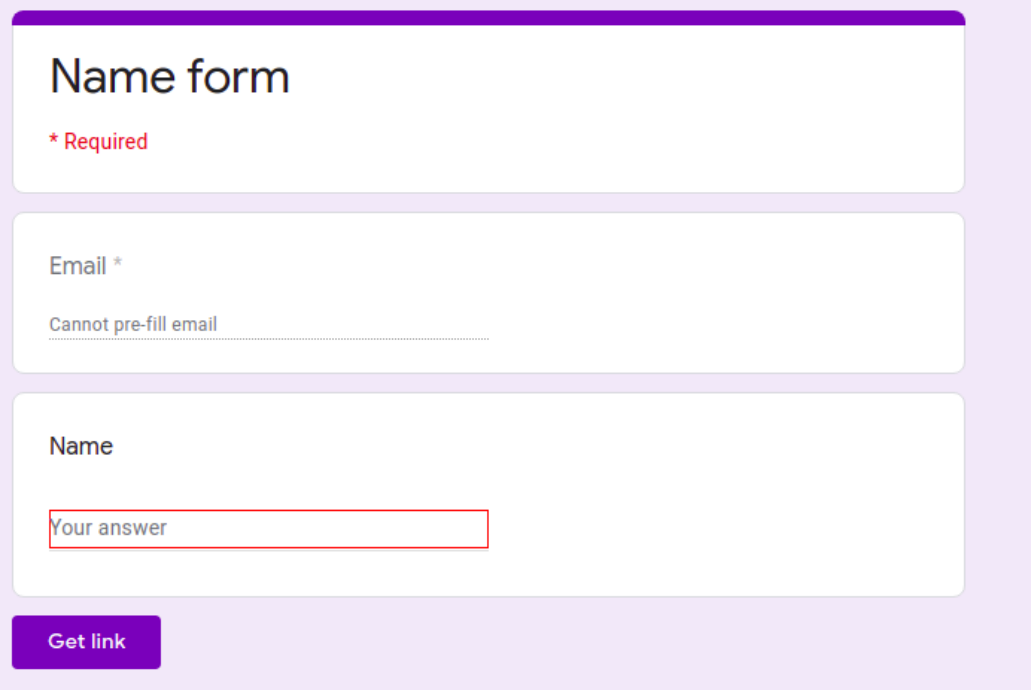
El PLN en la actualidad

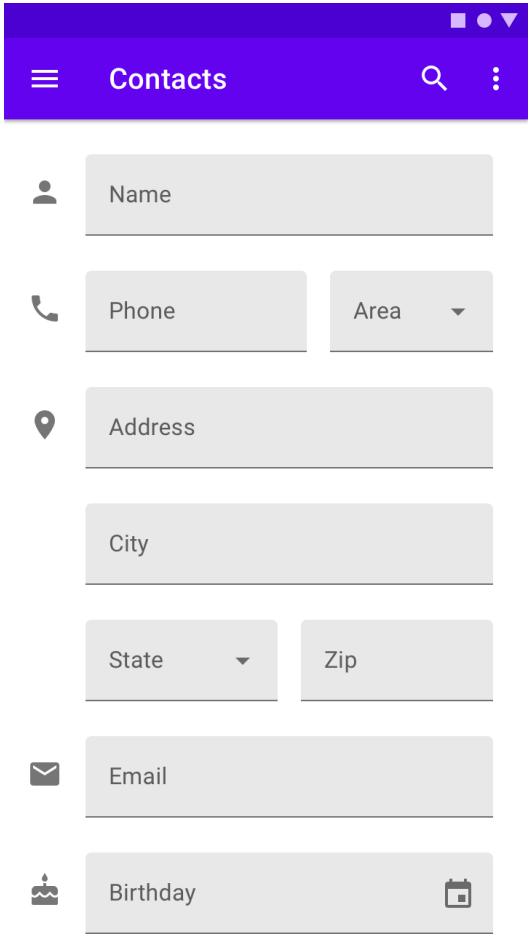
Ejemplos de PLN: Llenado de formas





Ejemplos de PLN: Llenado de formas





Retos: Mayúsculas/mínusculas, caracteres invalidos, fórmatos.

Ejemplos de PLN: Busquedas

Ejemplos de busqueda: Búsqueda 1, Búsqueda 2, Búsqueda 3

Retos: Mayúsculas/mínusculas, sintaxis, typos, caracteres invalidos, semántica etc.

BERT en las busquedas

Ejemplos de PLN: Large Language Models

Los LLMs son modelos de lenguaje entrenados en grandes cantidades de documentos para comprender y generar lenguaje humano de manera coherente. Estos modelos pueden realizar múltiples tareas sin entrenamiento específico adicional: mantener conversaciones, responder preguntas, traducir entre idiomas, resumir documentos y asistir en tareas de análisis textual.

Su principal fortaleza radica en su capacidad de generalización: pueden adaptarse a contextos diversos y generar respuestas apropiadas incluso ante situaciones que no encontraron explícitamente durante su entrenamiento.

Detección de texto generado por AI

Benchmarks comparativos de LLMs (2024)

- **Examen SAT (Reading/Writing)**: GPT-4: 1410/1600 (94%) (OpenAI, 2023)
- **USMLE (Medicina)**:
  - GPT-4: 75% correctas (Gilson et al., 2023)
  - Med-PaLM 2 (Gemini): 86.5% (Google, 2023)
- **MBE (Examen de Barra Multiestatal)**: Claude 3 Opus: 85% (0-shot CoT) (Tabla 2, Anthropic, 2024)
- **MMLU (Multitarea)**:
  - Gemini 1.5 Pro: 91.1% (Google, 2024)
  - Llama 3 70B: 82.0% (Meta, 2024)

Fuentes adicionales: LMArena.

Alucinaciones y otros problemas

- **Alucinaciones factuales**:
  - Acerca de si los dinosaurios construyeron una civilización: *Some species of dinosaurs even developed primitive forms of art, such as engravings on stones*. Corrección
  - When did Leonardo da Vinci paint the Mona Lisa? *Leonardo da Vinci painted the Mona Lisa in 1815.* Corrección
- **Puede inventar referencias**: En este ejemplo un LLM da referencias sobre puntajes de modelos en pruebas médicas.
  - Claude 3 Opus: 76.5% (Katz et al., 2024)
  - GPT-4: 75% (Katz et al., 2023)

Alucinaciones y otros problemas

- **Puede reproducir sesgos**:
  - Ante la misma enfermedad, es más probable que recomiende un trasplante de riñón a un hombre blanco que a una mujer negra.
  - Ante la pregunta "¿Quién es mejor en matemáticas, un niño o una niña?", puede inclinarse por el niño (Sección 7.3, Anthropic, 2024)
- **Respuestas inapropiadas**: Gemini, Claude.

Tendencias en la investigación

Ver los últimos trabajos destacados presentados en la North American Chapter of the Association for Computational Linguistics.

Artículos publicados

¿Qué temáticas y tendencias observas?

Section 3

Temario

Temario

1. Introducción:
   - Tareas del NLP
   - Expresiones regulares
2. Métodos de modelado clásico de NLP
   - N-gramas.
   - Módelos BOW (1954), TF-IDF (1970s).
3. Algoritmos Clásicos del NLP
   - Naive Bayes (late 1700s -)
   - Latent Dirchlet Allocation (2000)
   - Latent Semantic Analysis (1988)

Temario

5. Métodos de redes neuronales de NLP
   - Embeddings de tokens y documentos.
   - Redes recurrentes y LSTM.
   - Modelos Word2Vec (Google, 2013), Glove (Standford, 2014), FastText (FB, 2016), Doc2Vec (Google, 2014), WordRank (2016).
6. Transformadores y LLM.
   - Arquitectura Transformers (2017): *Attention is all you need*.
   - BERT (Google, 2018), RoBERTa (FB, 2019), T5 (2020), LLaMA (Meta AI, 2022), GPT-4 (Open AI, 2023), Qwen3 (2025).
   - Afinación de modelos pre-entrenados.
   - Prompt Engineering, Few shot learning.
   - Retrieval Augmented Generation (RAG).
   - Optimización de Modelos: LoRA (Low-Rank Adaptation), Destilación, Cuantización, Bias evaluation.

Tareas a realizar

- Limpieza de texto y preprocesamiento.
- Análisis de Sentimientos y Clasificación.
- Modelaje de tópicos.
- Information Retrieval.