E-BOOK

Un poco más de conversación, un poco más de acción

Construya una infraestructura de datos para la IA conversacional

NetApp

Contenido

¿Demasiadas charlas banales? Conviértalas en algo grande.

PLN: también llamado procesamiento del lenguaje natural. También llamado IA conversacional. También llamado *robots que hablan.*

Da igual cómo lo llame, un sistema de IA conversacional habla como un ser humano, entiende los distintos contextos y ofrece respuestas inteligentes gracias a unos enormes avances en aprendizaje profundo, que hacen que los sistemas de IA sean más naturales y menos transaccionales.

El aprendizaje profundo no solo hace que la IA sea más intuitiva, sino que también elimina la necesidad de tener un ser humano en el back-end que conozca en profundidad la lingüística y las técnicas basadas en reglas. El aprendizaje profundo abre la puerta a que los sectores que utilizan un lenguaje específico y más complejo (como los servicios financieros, la sanidad y las ciencias biológicas, el gobierno, el sector automovilístico, las manufacturas y el comercio al por menor) puedan adoptar soluciones de PLN.

Los datos son la clave para mejorar la conversación

Estos modelos de IA pueden ser enormes y muy complejos. Necesitan una gran cantidad de datos que se muevan a la velocidad del pensamiento. Para ser competente, una infraestructura de PLN debe ser capaz de:

1. Subir el volumen
2. Afinar su voz
3. Responder en un abrir y cerrar de ojos

PLN: ya no es solo para bots de chat

Desde los asistentes inteligentes hasta los motores de búsqueda y el texto predictivo, el PLN es el nuevo lenguaje global. Está por doquier, incluso donde menos se lo espera.

Evaluación de solvencia

El PLN se puede usar para generar calificaciones crediticias basadas en datos como la geolocalización, la actividad en redes sociales, la conducta de navegación y las redes de pares, entre otros.

Asignación de los ensayos clínicos

Puede ser difícil obtener la participación de pacientes en los ensayos clínicos, sobre todo porque la gente no sabe que dichos ensayos están disponibles. Con el PLN, los investigadores y los fabricantes pueden asignar pacientes automáticamente a los ensayos clínicos.

Cuerpos de seguridad

Las jefaturas de policía utilizan el PLN para identificar los móviles de los crímenes y así mantener seguras a las personas, reducir la violencia y hacer la vigilancia policial más comprensiva y receptiva.

Mantenimiento de vehículos

El PLN hace que mantener sus vehículos en plena forma sea más fácil para los conductores. En lugar de hojear un manual de usuario bien gordo, los propietarios solo tendrán que preguntarle al vehículo: «¿Qué piloto se ha encendido?» o «¿Cómo cambio un fusible?».

Reparación de aeronaves

El PLN ayuda a los mecánicos a sintetizar la información de los gruesos manuales de servicio y así aumentar su comprensión de los problemas de los que informan los pilotos.

1. Subir el volumen

Llevar a cabo un PLN correctamente requiere una cantidad absurda de datos. Piense en todas las palabras que se han dicho en la historia y estará cerca.

El PLN tiene que ser capaz de procesar, entender y cotejar la entrada de voz con una biblioteca de datos inmensa para crear una respuesta inteligible en milisegundos.

Este requisito es especialmente difícil de cumplir debido a la complejidad del lenguaje humano, que está lleno de reglas y excepciones, y se vuelve aún más complicado cuando se tienen en cuenta los matices de las expresiones, el sarcasmo y el humor. Puede que los modelos creados específicamente para un sector también requieran información específica acerca de un campo, una empresa o unos productos en concreto.

Por eso, el tamaño de los modelos de IA conversacional ha aumentado hasta alcanzar los millones o miles de millones de parámetros. Normalmente, cuanto más datos tenga, más preciso será el modelo. Los modelos de formación de este tamaño pueden llevar semanas de tiempo de computación y necesitar los mejores marcos de aprendizaje tanto profundo como automático.

Google Translate

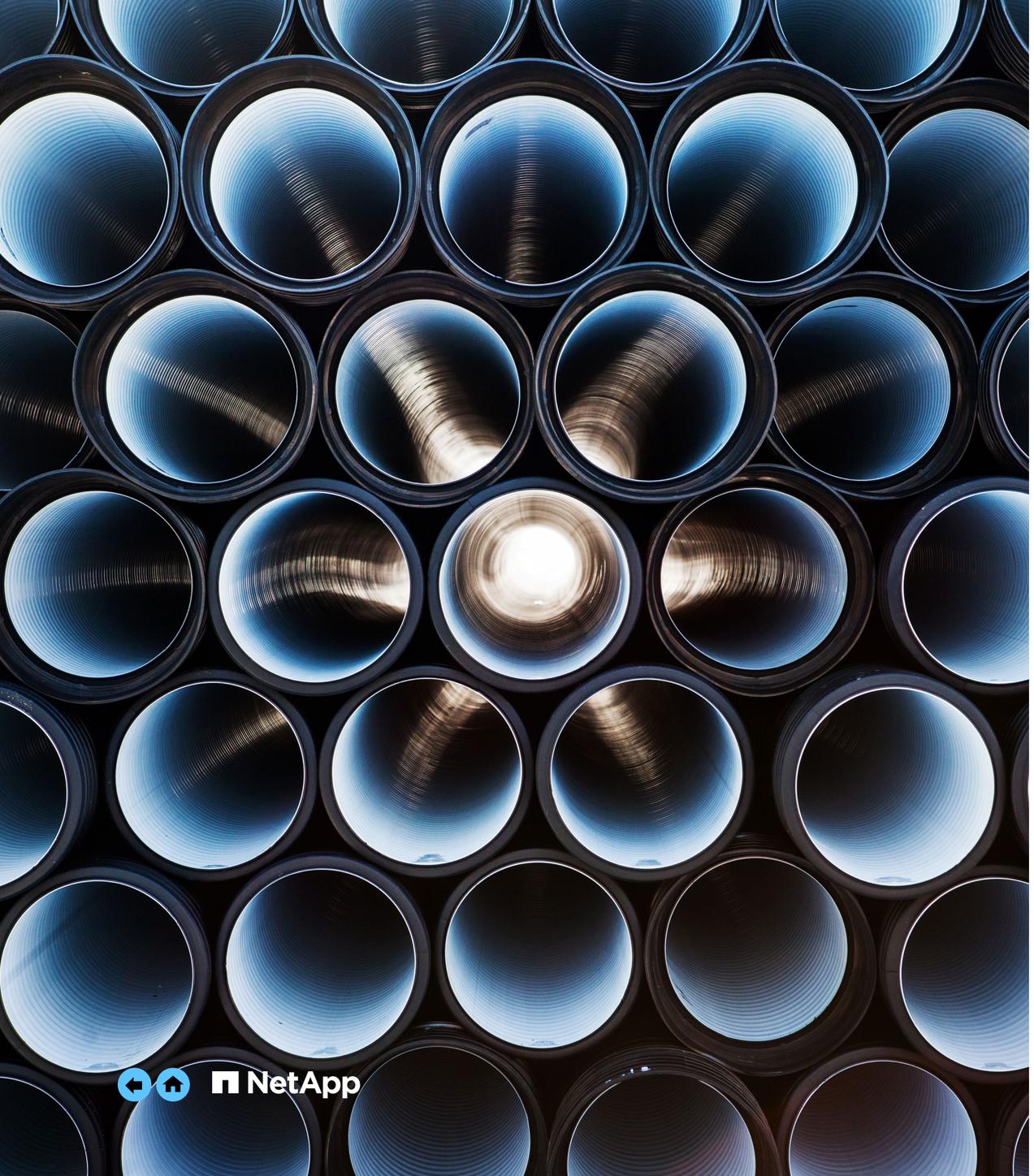
Google Translate da cabida a más de 100 idiomas y está abierto a colaboración para ayudar a contrastar y mejorar las traducciones y las formaciones de los modelos para idiomas con corpus de formación limitados (lo que viene a decir conjuntos de datos de fuentes). Google Translate procesa 140 000 millones de palabras todos los días. Eso se corresponde con el trabajo de 70 millones de traductores humanos. *Todos los días*.

Google BERT

Google BERT es un popular modelo PLN que tiene 340 millones de parámetros. BERT representa un avance en cuanto al PLN porque va más allá de las interfaces de voz transaccionales, como los algoritmos de árbol de los móviles, para convertirse en algo conversacional de verdad. Puede leer mensajes y responder preguntas con una gran precisión.

BioMegatron

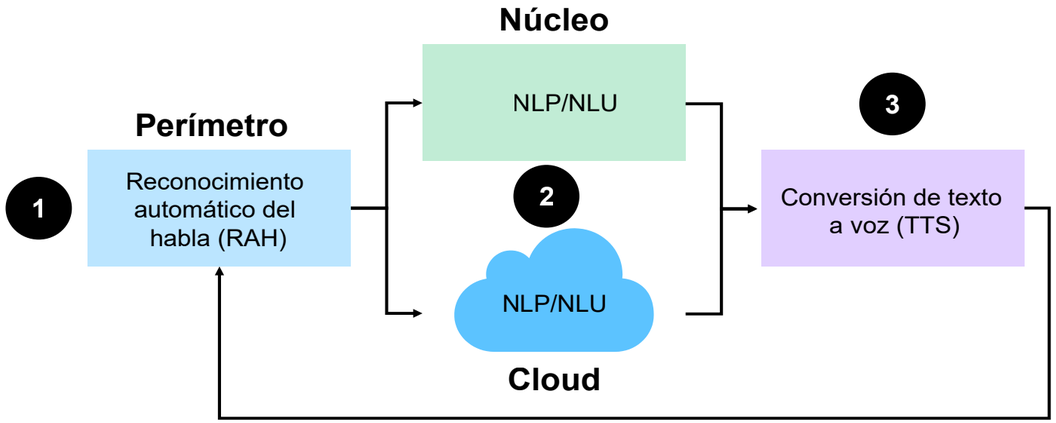
BioMegatron es el modelo de lenguaje biomédico basado en transformadores más grande que se haya formado jamás. Tiene hasta 1200 millones de variantes de parámetros. Se formó con 6100 millones de palabras de PubMed, un repositorio de conceptos abstractos y artículos completos de revistas especializadas en temas biomédicos.



2. Afinar su voz

Un PLN rápido y efectivo necesita una canalización de datos que cubra todo el ecosistema, desde la ingesta y el reconocimiento hasta la síntesis de voz. Los datos deben fluir rápida y libremente a través de las distintas etapas de la canalización para impulsar el procesamiento del lenguaje en tiempo real.

La típica canalización de PLN consta de tres fases:



En una infraestructura de PLN moderna, miles de ubicaciones del perímetro reúnen terabytes de datos todos los días. Cuando el acceso a estos datos se ve limitado por una infraestructura de silos, el aprendizaje profundo solo los utiliza de manera superficial.

3. Responder en un abrir y cerrar de ojos

Para que la IA replique el habla humana, tiene que operar a la velocidad del cerebro humano o incluso más rápido. Cuanto mayor es un modelo, mayor es la espera desde que un usuario formula la pregunta hasta que la IA responde. Para que suene natural, toda la computación debe suceder en un lapso de 300 milisegundos.

Este proceso consta de varias etapas:

1. Convertir el habla del usuario a texto
2. Comprender el significado del texto
3. Buscar la respuesta más adecuada según el contexto
4. Ofrecer una respuesta por voz

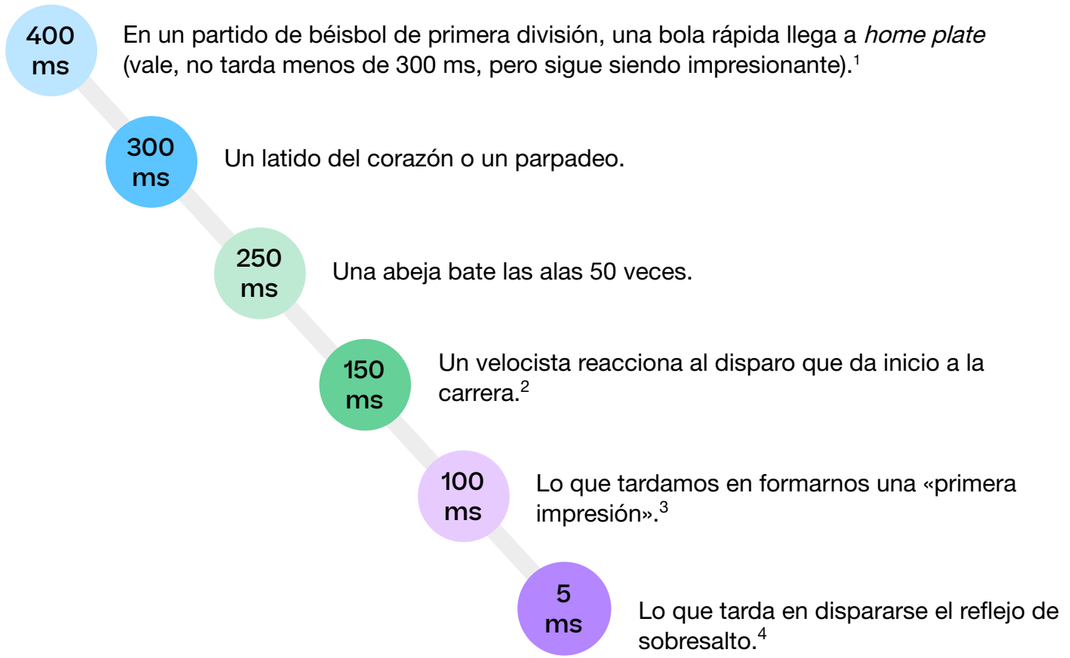
Con unos requisitos de latencia tan ajustados, a menudo los desarrolladores de IA conversacional tienen que compensar. Un modelo complejo de alta calidad podría tardar más en estar listo que un modelo de procesamiento del lenguaje menos abundante que ofreciera resultados rápidamente, pero cuyas respuestas tuvieran menos matices.

Igual que un solicitante de empleo que está nervioso, un asistente de voz podría atascarse durante un rato en una conversación y llenar el silencio incómodo con sonidos o frases como «Ahora mismo lo busco». Sin embargo, la IA conversacional ideal (es decir, el Santo Grial del PLN) es lo suficientemente sofisticada como para entender con precisión las dudas de una persona y responder rápidamente con un lenguaje natural fluido.

¿De qué velocidad estamos hablando?

Normalmente, el PLN necesita menos de 300 milisegundos (ms), que son 0,3 segundos, de latencia para crear una respuesta en tiempo real. ¿Cómo de rápido es eso? Muy rápido.

Estas son algunas cosas que pasan en 300 ms o menos:



NetApp utiliza su lenguaje

Con ONTAP® AI de NetApp®, impulsado por los sistemas NVIDIA DGX y los sistemas de almacenamiento all-flash conectados a cloud de NetApp, se pueden formar y optimizar modelos de lenguaje enormes y de vanguardia para llevar a cabo inferencias rápidamente. La solución de Data Fabric de NetApp simplifica la gestión de datos de la canalización de sus datos de IA, desde el perímetro hasta el núcleo y cloud.

- **Las soluciones de IA de NetApp AI** eliminan los cuellos de botella para permitir una recogida de datos más eficiente, unas cargas de trabajo de IA más aceleradas y una integración de cloud más fluida.
- **Las soluciones de gestión de datos unificadas de NetApp** permiten efectuar un movimiento de datos fluido y rentable entre entornos híbridos multicloud.
- **El ecosistema de partners de clase mundial de NetApp** proporciona integraciones totalmente técnicas con líderes de IA, partners de canal e integradores de sistemas, proveedores de hardware y software, y partners de cloud. Crean soluciones de IA inteligentes, potentes y fiables que le ayudan a cumplir sus objetivos empresariales.
- **Los servicios profesionales de NetApp** proporcionan la experiencia especializada que necesita para reducir la complejidad, ampliar sus oportunidades de IA y alcanzar el éxito.

Por cierto, NetApp se ha posicionado como un líder de IDC MarketScape en todo el mundo por su almacenamiento de escalado horizontal y basado en archivos.[5] Es importante porque las cargas de trabajo de visión artificial (sí, lo ha adivinado) son de escalado horizontal y están basadas en archivos.

Haga felices a sus científicos de datos

Cinco veces

Ejecute cinco veces más datos a través de su canalización de IA

Menos de 60 segundos

Copie conjuntos de datos en cuestión de segundos en lugar de horas o días

Aproxima-damente 20 minutos

Configure su infraestructura de IA con integración en Ansible en aproximadamente 20 minutos

El asistente de ventas de NetApp: una receta para el éxito

Mediante el uso de Jarvis de NVIDIA, que es un marco integral para construir servicios de IA conversacional, NetApp y NVIDIA han construido un asistente de ventas virtual que acepta entradas tanto de voz como de texto y responde preguntas sobre el tiempo, los puntos de interés y los precios de stock mediante una conexión a las API de Weatherstack y Yelp Fusion, y al kit de desarrollo de software con Python de eBay. Échele un vistazo

El asistente de ventas de NetApp (NARA) está construido en:

- **Jarvis de NVIDIA.** Jarvis ofrece servicios acelerados por GPU para la IA conversacional que usan una canalización de aprendizaje profundo integral que ha sido optimizada para que mantenga baja la latencia.
- **ONTAP AI de NetApp** Esta arquitectura contrastada combina los sistemas NVIDIA DGX y el almacenamiento all-flash de NetApp. ONTAP AI optimiza el flujo de datos de manera fiable, pues le permite formar y ejecutar modelos conversacionales complejos sin superar los requisitos de latencia.
- **NeMo de NVIDIA.** NeMo es un kit de herramientas de Python para construir, formar y ajustar los modelos de IA conversacional acelerados por GPU que permite construir modelos con API fáciles de usar, que incluyen aplicaciones de reconocimiento automático del habla (RAH), procesamiento del lenguaje natural (PLN) y conversión de texto a voz (TTS).

¿Se une al PLN?

Bueno, ya me conoce. ¿Qué vendrá a continuación? ¿Mantener conversaciones con animales del bosque? No podemos enseñarle a hablar a una ardilla. Sí podemos enseñarle a usted a construir una infraestructura de IA adecuada para el PLN.

Obtenga más información acerca de las soluciones de IA de NetApp:

- IA de NetApp
- ONTAP AI
- Soluciones de NetApp para PLN

¿Preguntas? Nuestros expertos en soluciones de IA están a la espera.

1. O'Neill, Shane. «Real-time bidding: What happens in 200 milliseconds?». Nanigans.
2. Welsh, Tim. «Exactly how long does it take to think a thought?». The Christian Science Monitor. 1 de julio de 2015.
3. Wargo, Eric. «How Many Seconds to a First Impression?». Association for Psychological Science. 1 de julio de 2006.
4. Wise, Jeff. «What Is the Speed of Thought?». New York Magazine. 19 de diciembre de 2016.
5. Potnis, Amita. IDC MarketScape: Worldwide Scale-Out File-Based Storage 2019 Vendor Assessment. IDC. Diciembre de 2019.

Acerca de NetApp
En un sector lleno de generalistas, NetApp es un especialista. Nos centramos en una cosa: ayudar a que su empresa aproveche al máximo sus datos. NetApp incorpora a cloud los servicios de datos de clase empresarial en los que confía, y lleva la sencilla flexibilidad de cloud al centro de datos. Nuestras soluciones líderes del sector funcionan en diversos entornos del cliente y en los sistemas de cloud públicos más grandes del mundo.

Como empresa de software centrado en datos y orientado a cloud, solo NetApp puede ayudar a crear su Data Fabric exclusivo, a simplificar y conectar su cloud, y a proporcionar con seguridad los datos, los servicios y las aplicaciones correctos a las personas adecuadas en cualquier momento y lugar.

Si desea obtener más información, visite **www.netapp.com/es**