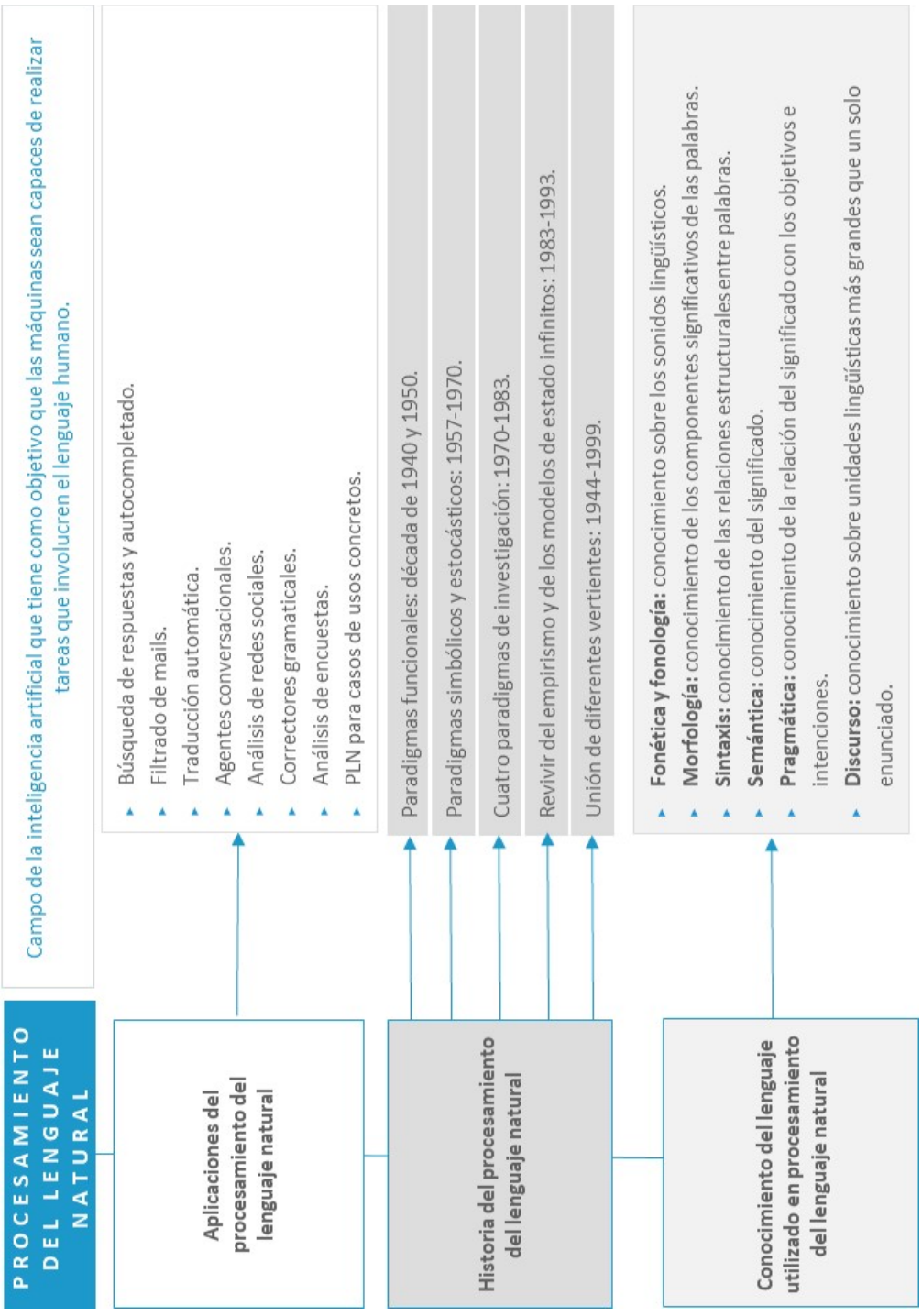
Procesamiento del Lenguaje Natural

Tema 1. Introducción al procesamiento del lenguaje natural

Índice

Esquema



1.1. Introducción y objetivos

Comenzaremos introduciendo de manera general el campo de la inteligencia artificial denominado **procesamiento del lenguaje natural (PLN).** La idea de que las máquinas sean capaces de poseer habilidades de comunicación y lenguaje es tan antigua como la propia aparición de los primeros ordenadores.

Se describirán diferentes **aplicaciones** que utilizan **técnicas basadas en el procesado del lenguaje natural.** Por ejemplo: los sistemas de búsqueda de respuestas, el filtrado de mails, la traducción automática, los agentes conversacionales, el análisis de redes sociales, los correctores gramaticales, el análisis de encuestas y, finalmente, la utilidad del PLN para algunos casos de uso concretos.

Algunas de estas aplicaciones, como los agentes conversacionales o la traducción automática, entre otras, se verán con detalle en el futuro. Junto con ello, en los próximos apartados, se presentarán la historia del procesamiento del lenguaje natural y se puede observar cómo ha evolucionado dicho campo y los diferentes enfoques que ha ido tomando a lo largo de la historia. Por último, analizaremos cómo el conocimiento del lenguaje es imprescindible para el correcto funcionamiento de los sistemas de procesamiento de lenguaje natural, describiendo los diferentes tipos de conocimiento del lenguaje.

Objetivos

- Definir el concepto de procesamiento del lenguaje natural.
- Identificar distintas aplicaciones del mundo real donde se usa el PLN como parte fundamental.

Ideas clave

- Narrar la historia del PLN.
- Definir los diferentes tipos de conocimiento del lenguaje que es necesario para correcto funcionamiento de los sistemas de PLN.

1.2. Procesamiento del lenguaje natural

El procesamiento del lenguaje y del habla ha sido tratado históricamente de forma muy diferente en la informática, la ingeniería, la lingüística, la psicología o la ciencia cognitiva. Hoy en día se concibe el procesamiento del lenguaje natural como **área que abarca varios campos diferentes y diversos pero superpuestos**.

Por ello, el procesamiento del lenguaje natural es un campo interdisciplinario que une a informáticos, ingenieros electrónicos y de telecomunicaciones con lingüistas, sociólogos y psicólogos.

El reconocimiento de la voz, que incluye tareas del procesamiento de la señal, se ha tratado tradicionalmente en la ingeniería electrónica y de telecomunicaciones. El análisis sintáctico y la interpretación semántica de las palabras y frases son áreas tradicionales del procesamiento del lenguaje natural que se estudian en el campo de la informática. La morfología, la fonología y la pragmática son tareas de investigación en la lingüística computacional. Psicolingüistas y sociolingüistas estudian respectivamente los mecanismos cognitivos para la adquisición del lenguaje y cómo la sociedad influye en el uso de la lengua.

El ancho espectro que abarca el campo del procesamiento del lenguaje natural hace que se conozca con diferentes nombres debido a las diferentes vertientes involucradas. Algunos de estos nombres, que provienen de estas diferentes facetas, serían procesamiento del lenguaje y del habla, tecnología del lenguaje, procesamiento del lenguaje natural, lingüística computacional o reconocimiento y síntesis del habla.

Ideas clave

En la inteligencia artificial, el procesamiento del lenguaje natural es un campo que tiene como objetivo que las máquinas sean capaces de realizar tareas que involucren el lenguaje humano.

Algunas de las tareas que debe realizar una máquina para ser capaz de procesar el lenguaje natural incluyen funcionalidades tales como la de habilitar a la máquina de habilidades para comunicarse con personas, la de mejorar la comunicación entre humanos o, simplemente, la de procesar un texto o el habla.

1.3. Aplicaciones del PLN

El PLN es una de las áreas más importantes dentro del campo de la IA y de la Ciencia de Datos. Por este motivo, actualmente existen multitud de aplicaciones basadas en IA de las que el PLN es una **pieza clave.** Entre estas aplicaciones destacan algunas como las siguientes:

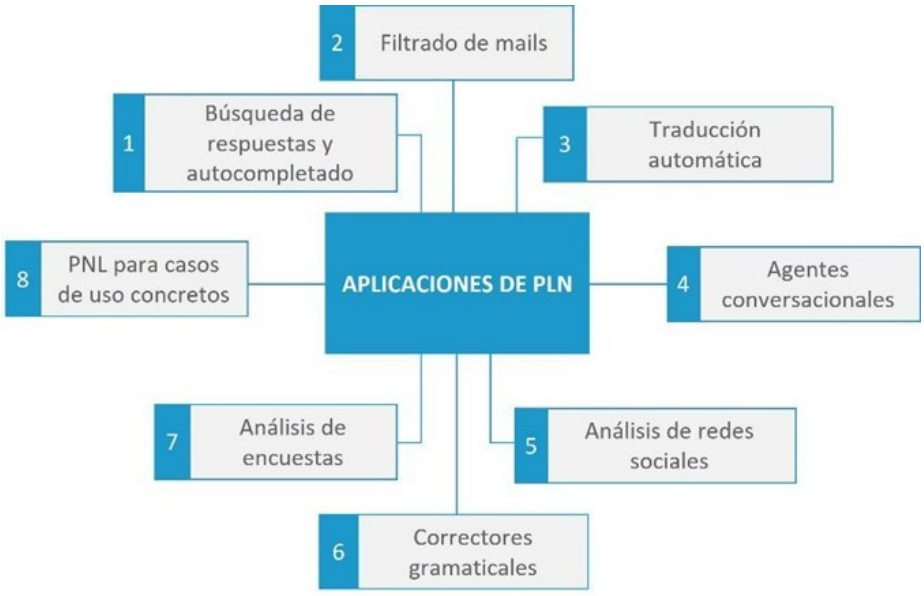


Figura 1. Aplicaciones del procesamiento del lenguaje natural. Fuente: elaboración propia.

Búsqueda de respuestas y autocompletada

En inglés, *Question Answering* (QA) es un tipo de recuperación de la información basado en el lenguaje natural. Por lo tanto, es una **extensión de una búsqueda simple de información en la web**, pero en lugar de solo escribir palabras clave, un usuario puede hacer preguntas completas. Los motores de búsqueda pueden responder preguntas sobre fechas y ubicaciones sin necesidad de aplicar procesamiento del lenguaje natural. Sin embargo, para responder a preguntas más complejas se requiere alguno de estos aspectos:

- La extracción de información o de un fragmento de texto en una página web.
- Hacer inferencia, es decir, sacar conclusiones basadas en hechos conocidos.
- Sintetizar y resumir información de múltiples fuentes o páginas web.

Los sistemas de búsqueda de respuestas, que requieren la comprensión de la información, se componen de diferentes elementos tales como un módulo de extracción de información, un módulo para resumir de forma automática o un módulo de desambiguación del sentido de las palabras.

Junto a la búsqueda de respuestas, el PLN se usa también para el autocompletado de textos, de manera que a medida que un usuario vaya escribiendo una secuencia de palabras, el sistema tratará de inferir las palabras que vendrían después.

Filtrado de mails

Uno de los sistemas de filtrado de mails más conocidos es **el filtro de spam,** que en muchas ocasiones se basa en el uso de técnicas de PLN junto con **modelos de aprendizaje automático.** Con ello, en función de determinados elementos del correo (remitente, asunto, contenido, etc.), se podrán filtrar de manera automática aquellos mails que son potencialmente spam.

Ahora bien, esta aproximación no sirve sólo para la detección de spam, sino que también sirve para **clasificar mails en otro tipo de categorías** (promociones, asuntos importantes, etc.).

Traducción automática

Es otra tarea relacionada con el procesamiento del lenguaje natural. El objetivo de la traducción automática es **traducir automáticamente un documento de un idioma a otro.** Además, se incluye en este ámbito no solo la traducción de documentos de texto, sino también la traducción de forma automática del habla de un lenguaje o idioma a otro.

Los mejores resultados se obtienen utilizando métodos estadísticos basados en frases. En estos métodos se utiliza básicamente el aprendizaje automático para analizar grandes conjuntos de datos y realizar traducciones que no contemplen las cuestiones gramaticales. En esta se requieren también herramientas para solventar la ambigüedad de las palabras como serían los algoritmos de desambiguación.

Agentes conversacionales

También llamados sistemas de diálogo. Son **programas que conversan con las personas a través del lenguaje natural.** Los *chatbots* son uno de los tipos más avanzados de los agentes conversacionales porque permiten mantener conversaciones no estructuradas, una característica de las conversaciones entre personas. Los agentes conversacionales pueden interactuar con el humano ya sea a través de la voz (hablando con el usuario), de texto, en el caso que la conversación se lleve a cabo a través de un chat, o utilizando ambas modalidades a la vez.

Los agentes conversacionales se caracterizan por ser **sistemas que toman turnos para conversar,** por lo que aparte de tener que analizar el lenguaje natural durante la conversación, deben tener en cuenta el turno de palabra. Por lo tanto, los agentes conversacionales, además de tratar con tareas básicas del procesamiento del lenguaje natural como son el reconocimiento de palabras y frases o la semántica de

estas, deben mantener el estado de la conversación y ser capaces de generar nuevas frases que continúen la conversación que mantienen con la persona.

Análisis de redes sociales

Gran parte de la población usa algún tipo de red social, de manera que estas recogen mucha información de usuarios de distintas partes del mundo. De esta manera, por ejemplo, si una empresa lanza un nuevo producto al mercado, las redes sociales tendrán información muy valiosa de la opinión de gran parte de los consumidores sobre ese nuevo producto. Ahora bien, esta información suele quedar reflejada en las redes sociales como textos (ej., un *tweet*). Para poder analizar grandes volúmenes de información de este tipo de manera automática se usan precisamente técnicas de PLN, muchas veces en combinación con **algoritmos de aprendizaje automático.** Por ejemplo, podemos analizar el sentimiento de los usuarios hacia una marca o producto concreto en un momento dado para ver si este es positivo, negativo o neutro, utilizando técnicas de PLN que extraigan información relevante de los textos, junto con modelos de aprendizaje automático que clasifiquen esa información en una de las tres categorías.

Correctores gramaticales

Así como el PLN sirve para tareas como el autocompletado de texto (sugiriendo posibles maneras de continuación de una frase), también **es un elemento clave en los correctores gramaticales.** Teniendo en cuenta las palabras previas escritas, el conocimiento de la lengua, la información sintáctica y semántica. Se pueden tanto corregir errores ortográficos, así como proponer maneras más correctas de escribir una frase.

Análisis de encuestas

Es habitual que muchas empresas y organismos realicen encuestas para **conocer la opinión de los usuarios respecto de distintos temas.** Ahora bien, puede ocurrir que estas encuestas contengan *feedback* de los usuarios expresado en texto libre,

de manera que, si el número de encuestas es muy elevado, puede ser muy laborioso hacer un análisis manual de todas ellas. Aquí también entran las técnicas de PLN para poder procesar esos textos y extraer la información relevante que allí se contiene.

PLN para casos de uso concretos

A modo de ejemplo, otro de los sectores donde el PLN está siendo de gran utilidad es en el ámbito de Recursos Humanos (RR. HH.) para tareas como, por ejemplo, la criba curricular. Si la cantidad de CV que recibe la empresa es muy elevada, y si además los formatos de estos no son homogéneos, el uso de técnicas de PLN puede ayudar a ordenarlos según cómo estén de alineados con la descripción del puesto de trabajo. Esto también se puede utilizar para ver si un candidato está más alineado para otra vacante disponible en lugar de a la que había aplicado.

Este es un ejemplo, pero hay otros casos de uso específicos para distintos dominios (ej., analizar automáticamente historias clínicas en el ámbito médico).

1.4. Historia del procesamiento del lenguaje natural

La historia del procesamiento del lenguaje natural se puede dividir en diferentes etapas. Desde la aparición de las bases o paradigmas fundacionales en la década de 1940 hasta la explotación de los paradigmas más modernos para desarrollar hoy en día aplicaciones más inteligentes.

1940-1950	Paradigmas fundacionales
1957-1970	Paradigma simbólico y estocástico
1970-1983	Cuatro paradigmas de investigación
1983-1993	Revivir del empirismo y los modelos de estados finitos
1994-1999	Unión de las diferentes vertientes
2000	Auge del aprendizaje automático

Tabla 1. Cronología de las diferentes etapas de la historia del PLN. Fuente: elaboración propia.

Paradigmas fundacionales: década de 1940 y 1950

El principio del PLN data del período justo después de la Segunda Guerra Mundial, cuando se dio el **origen del ordenador.** En este período, desde la década de 1940 hasta el final de la década de 1950, se trabajó intensamente en dos paradigmas fundacionales:

- Autómatas.
- Modelos probabilísticos o de teoría de la información.

Autómatas

El autómata surgió en la década de 1950 a partir del famosísimo estudio publicado por **Turing** (1936), Los números computables, con una aplicación al Entscheidungsproblem, y que se considera como la base de la informática moderna. El trabajo de Turing condujo primero a un modelo simplificado de la neurona como elemento de computación que podría describirse en términos de lógica proposicional (McCulloch y Pitts, 1943) y luego a la definición de los autómatas finitos y las expresiones regulares (Kleene, 1951).

Shannon aplicó las cadenas de Markov, un modelo de proceso estocástico discreto en el que la probabilidad de ocurrencia de un evento depende solo del evento inmediatamente anterior, a los autómatas para el lenguaje (Shannon, 1948). Aprovechando las ideas del trabajo de Shannon, Chomsky fue el primero en considerar las máquinas de estados finitos como una forma de caracterizar una gramática y definió el lenguaje de estados finitos como un lenguaje generado por una gramática de estados finitos (Chomsky, 1956).

Estos primeros modelos llevaron a la aparición de la teoría del lenguaje formal, que utilizó el álgebra y la teoría de conjuntos para definir los lenguajes formales como secuencias de símbolos. Esta teoría incluye las gramáticas libres de contexto, un concepto que Chomsky definió en 1956 para las lenguas naturales, pero que también fue descubierto de forma independiente por Backus y Naur en su descripción del lenguaje de programación ALGOL (Backus, 1959) (Naur et al., 1960).

Teoría de la información

El segundo paradigma fundamental de este período fue el desarrollo de **algoritmos probabilísticos para el procesamiento del lenguaje y del habla.** Esta idea proviene de la otra gran contribución de Shannon, el teorema de codificación de canal en la teoría de la información, y que muestra que es posible la **transmisión y decodificación del lenguaje** a través de un **canal de comunicación ruidoso**.

Shannon tomó prestado el concepto de **entropía de la termodinámica** como una forma de medir la capacidad de información de un canal o el contenido de información de un idioma, y realizó la primera medida de la entropía del inglés utilizando técnicas probabilísticas.

También durante este período inicial se desarrolló el **espectrógrafo de sonido** y se realizó investigación fundamental en la fonética instrumental, lo que sentó **las bases para el reconocimiento de la voz.** La primera máquina capaz de realizar el reconocimiento de la voz apareció a principios de los años cincuenta. En 1952 investigadores de Bell Labs construyeron un sistema estadístico basado en correlación que podía reconocer con un 97-99 % de precisión cualquiera de los diez primeros números a partir de unos patrones grabados con los sonidos de las vocales de un único hablante (Davis, Biddulph y Balashek, 1952).

Paradigma simbólico y estocástico: 1957-1970

A fines de la década de 1950 y comienzos de la década de 1960, el procesamiento del habla y el lenguaje se había dividido muy claramente en dos paradigmas:

- Simbólico.
- Estocástico.

Paradigma simbólico

Apareció de dos líneas de investigación: la teoría del lenguaje formal y la inteligencia artificial.

La primera línea se basaba en el trabajo que realizaron Chomsky y sus colaboradores **en la teoría del lenguaje formal y la sintaxis generativa,** además de en el trabajo de muchos lingüistas e informáticos que estudiaban los algoritmos de análisis, inicialmente de arriba hacia abajo *(top-down) y* de abajo hacia arriba *(bottom-up)* y luego a través de la programación dinámica. Uno de los primeros sistemas de análisis fue TDAP (Transformations and Discourse Analysis Project) que implementó Zelig Harris entre junio de 1958 y julio de 1959 en la Universidad de Pennsylvania.

La segunda línea de investigación del paradigma simbólico fue el nuevo campo de la **inteligencia artificial.** En el verano de 1956, John McCarthy, Marvin Minsky, Claude Shannon y Nathaniel Rochester congregaron durante dos meses a un grupo de investigadores para realizar un simposio sobre lo que decidieron llamar «inteligencia artificial».

Aunque el incipiente ámbito de la inteligencia artificial incluía una minoría de investigadores centrados en algoritmos estocásticos y estadísticos (incluidos los modelos probabilísticos y las redes neuronales), este nuevo campo se enfocó básicamente en el razonamiento y la lógica, representados por el trabajo de Newell y Simon en los programas de ordenador Logic Theorist y General Problem Solver (Newell, Shaw y Simon, 1959).

Ideas clave

En los principios de la inteligencia artificial fue cuando se construyeron los primeros **sistemas de comprensión del lenguaje natural.** Estos sistemas simples estaban diseñados para trabajar en un dominio concreto y basaban su funcionamiento en la búsqueda de patrones y heurística de palabras clave para realizar el razonamiento y la búsqueda de respuestas. Fue a finales de la década de 1960 cuando se desarrollaron algunos sistemas lógicos más formales.

Paradigma estocástico

Se desarrolló principalmente por parte de investigadores de los departamentos de estadística y de ingeniería electrónica. A fines de la década de 1950 comenzó a aplicarse el método bayesiano al problema del reconocimiento óptico de caracteres. Por ejemplo, Bledsoe y Browning construyeron un **sistema bayesiano** de reconocimiento de texto que calculaba la probabilidad de una secuencia de letras dadas las palabras de un diccionario.

En la década de 1960 aparecieron también los primeros **modelos psicológicos para el PLN** basados en **gramáticas transformacionales y los primeros corpus disponibles online.** Un ejemplo es el Brown Corpus, un corpus del inglés americano desarrollado en 1963 por la Brown University y que contenía una colección de un millón de palabras extraídas de 500 textos de diferentes géneros: periódicos, novelas, no ficción, académico, etc. (Kucera y Francis, 1967).

Cuatro paradigmas de investigación: 1970-1983

En el siguiente período, en la década de 1970 y a principios de la década de 1980, se produce una explosión de la investigación en el procesamiento de lenguaje y del habla. Es en esta época cuando se desarrollan una serie de paradigmas de investigación que todavía hoy dominan el campo:

Figura 2. Paradigmas de investigación dominantes. Fuente: elaboración propia.

E l **paradigma estocástico** jugó un papel muy importante en el desarrollo de algoritmos de reconocimiento de voz para los que se utilizaban modelos ocultos de Markov (HMM) y el teorema de codificación de canal de Shannon. Algunos de los investigadores que trabajaron en este ámbito fueron Jelinek, Bahl, Mercer y sus socios del Thomas J. Watson Research Center de IBM, Baker en la Carnegie Mellon University, e investigadores de los Bell Laboratories de AT&T.

E l **paradigma basado en lógica** surgió del trabajo de Alain Colmerauer y sus colaboradores en la década de 1970 que desarrollaron Q-system, un analizador *bottom-up* basado en una **serie de reglas con variables lógicas** y que permitían la traducción del inglés al francés (Colmerauer, 1978). En este paradigma basado en lógica se engloba también la gramática de cláusulas definidas (Definite Clause Grammar, DCG), una forma de expresar la gramática en un lenguaje de programación lógico como por ejemplo Prolog (Pereira y Warren, 1980).

De forma independiente, apareció también el trabajo de Kay (1979) sobre la gramática funcional y, poco después, el trabajo de Joan Bresnan y Ronald Kaplan (1982) sobre la gramática léxico funcional (Lexical functional grammar, LFG), una gramática generativa que se centra en la investigación de la sintaxis del lenguaje natural.

Ideas clave

El **paradigma de la comprensión del lenguaje natural** apareció durante la década de 1970 con la aparición del sistema SHRDLU (Winograd, 1972). Este sistema simulaba un robot que movía bloques y era capaz de recibir comandos del lenguaje natural en formato de texto, como por ejemplo «mueve el bloque rojo que se encuentra en la parte superior del verde más pequeño». Este sistema, complejo y sofisticado para su época, también fue el primero en construir una gramática relativamente extensa del inglés.

Los avances en el modelo de análisis del lenguaje de Winograd dejó claro que la investigación debía comenzar a enfocarse en la semántica y los modelos de discurso. Roger Schank y sus colaboradores, conocidos como la Escuela de Yale, construyeron una serie de programas de comprensión del lenguaje que se centraron en el conocimiento humano, por ejemplo, en planes y objetivos, y la organización de la memoria humana (Schank y Riesbeck, 1981). Estos trabajos, por ejemplo, el realizado por R. F. Simmons (1973), usaban una representación del conocimiento en forma de red, lo que se conoce como redes semánticas (Quillian, 1968), y comenzaron a incorporar en sus representaciones la idea de gramática de casos (Fillmore, 1968), por la cual se establece una relación entre un verbo y múltiples papeles temáticos que serían los sintagmas nominales.

El paradigma basado en lógica y el paradigma de la comprensión del lenguaje natural se unificaron en sistemas que usaban la lógica de predicados como una representación semántica, como el sistema de búsqueda de respuestas LUNAR (Woods, 1967).

El **paradigma del modelado del discurso** se centró en las cuatro áreas clave del discurso. Grosz y sus colaboradores introdujeron el estudio de la estructura del discurso y el enfoque del discurso (Grosz, 1977) (Grosz y Sidner, 1986). Una serie de investigadores comenzaron a trabajar en la resolución de forma automática de referencias en el discurso (Hobbs, 1978 y 1979). Por último, se desarrolló el modelo

BDI (creencias-deseos-intenciones en inglés), un marco de trabajo basado en la lógica de actos de habla (Perrault y Allen, 1980) (Cohen y Perrault, 1979).

Revivir del empirismo y los modelos de estados finitos: 1983-1993

A partir de 1983 volvieron dos clases de modelos que habían perdido popularidad a finales de la década de 1950 y principios de la década de 1960 debido a los argumentos teóricos en su contra (Chomsky, 1959).

Figura 3. Modelos de estados finitos y modelos empíricos. Fuente: elaboración propia.

Los **modelos de estados** finitos revivieron en esta época y comenzaron a recibir atención porque se usaron en la fonología y la morfología (Kaplan y Kay, 1981), y en la sintaxis (Church, 1980).

La segunda tendencia en este período fue lo que se ha llamado **el retorno del empirismo,** marcada por el aumento de los modelos probabilísticos para el procesamiento del lenguaje y del habla. Cabe destacar la influencia del trabajo de los investigadores del Thomas J. Watson Research Center de IBM sobre modelos probabilísticos en el reconocimiento de voz. Estos métodos probabilísticos y otros enfoques basados en los datos se extendieron al etiquetado morfosintáctico (POS tagging), al análisis y resolución de ambigüedades y a la semántica.

Este paradigma empírico vino acompañado por el enfoque de la evaluación de los modelos basado en los datos. Entonces en este período se desarrollaron métricas cuantitativas para la evaluación y se enfatizó en la comparación del rendimiento de estas métricas con los resultados de las investigaciones previas. Además, en este período, se trabajó considerablemente en la generación de lenguaje natural.

Unión de las diferentes vertientes: 1994-1999

En los últimos cinco años del pasado milenio, el campo del procesamiento del lenguaje natural sufrió grandes cambios.

En primer lugar, los modelos probabilísticos y los modelos basados en datos se volvieron estándares para el procesamiento del lenguaje natural. Los algoritmos de análisis, de etiquetado morfosintáctico, de resolución de referencias y de procesamiento del discurso empezaron a incorporar probabilidades y adoptaron metodologías de evaluación provenientes de los ámbitos del **reconocimiento de la voz y la recuperación de información.**

En segundo lugar, el aumento en la velocidad y la memoria de los ordenadores permitió la explotación comercial de varias áreas del procesamiento del lenguaje y del habla, en particular el reconocimiento de la voz y la revisión de la ortografía y la gramática. Además, los algoritmos de procesamiento del lenguaje y del habla comenzaron a aplicarse a la comunicación aumentativa para ayudar a personas con algún tipo de discapacidad. Por último, el aumento de la web enfatizó la necesidad de la recuperación y la extracción de información basada en el lenguaje natural.

Auge del aprendizaje automático: 2000

Las tendencias empiristas que marcaron la última parte de la década de 1990 se aceleraron a un ritmo asombroso en el nuevo milenio. Esta aceleración fue impulsada en gran parte por el auge del **aprendizaje automático.**

Figura 4. Aprendizaje automático. Fuente: elaboración propia.

Grandes cantidades de material hablado y escrito se pusieron a disposición de los investigadores a través de organizaciones tipo el Linguistic Data Consortium (LDC). Estos materiales eran fuentes de texto estándar que venían anotados con información sintáctica, semántica y pragmática. Entre estos materiales caben destacar las primeras colecciones anotadas: el Penn Treebank, el Prague Dependency Treebank, que anota la estructura de las dependencias, y las anotaciones semánticas del PropBank.

La existencia de estos recursos anotados promovió la tendencia de atacar los problemas más complejos del procesamiento del lenguaje natural, tipo el análisis sintáctico y semántico, como problemas de aprendizaje automático supervisado. Por ejemplo, se empezaron a aplicar las máquinas de vectores de soporte (SVM), el principio de máxima entropía, la regresión logística multinomial y los modelos bayesianos en la lingüística computacional.

Entonces, este mayor enfoque en el aprendizaje automático llevó a una interacción más seria de los lingüistas computacionales con la comunidad estadística.

El coste y la dificultad de producir corpus anotados se convirtió en un factor limitante del uso de los enfoques supervisados para muchos problemas del procesamiento del lenguaje. Por lo que a partir de 2005 aparece una nueva tendencia hacia el uso de técnicas de **aprendizaje no supervisado** en el procesamiento del lenguaje natural. Entonces se empezaron a construir algunas aplicaciones lingüísticas a partir de datos sin anotación alguna, por ejemplo, para la traducción automática o para el modelado de temas.

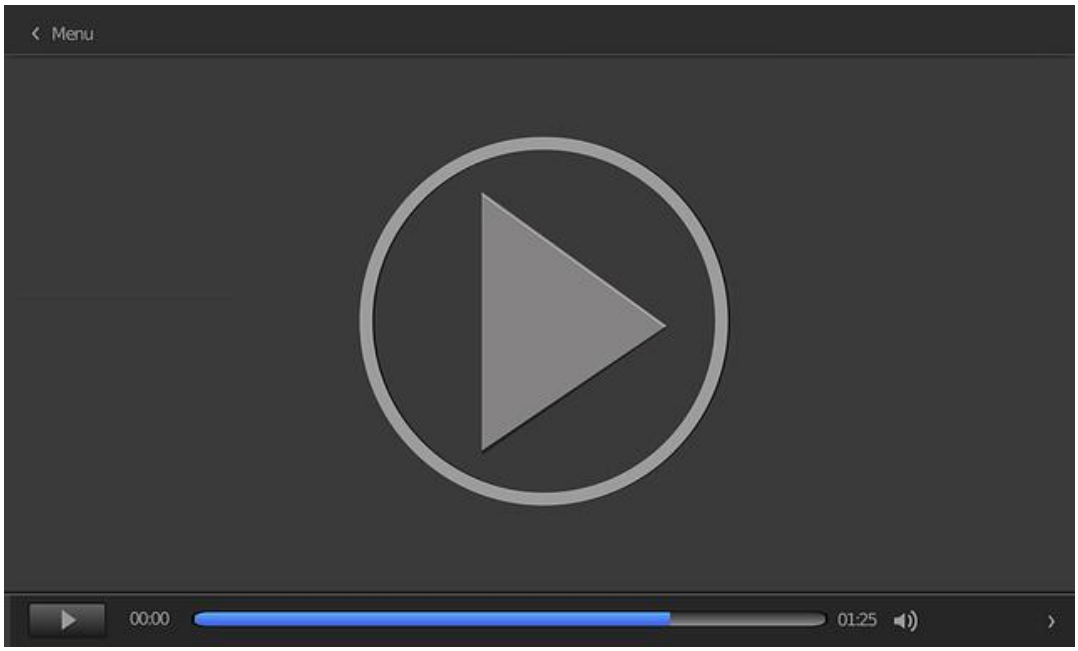
Los algoritmos de aprendizaje no supervisado **se han usado** en el **etiquetado morfosintáctico** (POS *tagging)* para agrupar palabras en las correspondientes partes del lenguaje (Goldwater y Griffiths, 2007) (Sirts, Eisenstein, Elsner y Goldwater, 2014). Además, las técnicas del aprendizaje no supervisado se han usado también para el etiquetado semántico, donde se han creado conjuntos de roles semánticos a partir de las características sintácticas (Titov y Klementiev, 2012) (Lang y Lapata, 2014).

Ideas clave

En 2006, Geoffrey Hinton acuña el término ***deep learning*** (aprendizaje profundo). Con el auge de este tipo de redes neuronales en la década de 2010, estas redes de neuronas artificiales profundas se empezaron a usar en diferentes ámbitos del procesamiento del lenguaje natural. Las redes neuronales recurrentes se están utilizando como una alternativa a los modelos ocultos de Markov (HMM) en análisis morfosintáctico y en el análisis sintáctico (Chen y Manning, 2014) (Dozat, Qi, y Manning, 2017). Además, las redes neuronales profundas también se están utilizando para el etiquetado semántico (Collobert et al., 2011) (Foland Jr. y Martin, 2015). De hecho, el *deep learning* es la base de los modelos de secuencia a secuencia (seq2seq) que se utilizan en los agentes conversacionales y *chatbots* actuales.

Ideas clave

En el vídeo *Aplicaciones del procesamiento del lenguaje natural* se realizará un análisis de la importancia del PLN en el mundo actual, y cómo es un elemento clave en muchas aplicaciones.



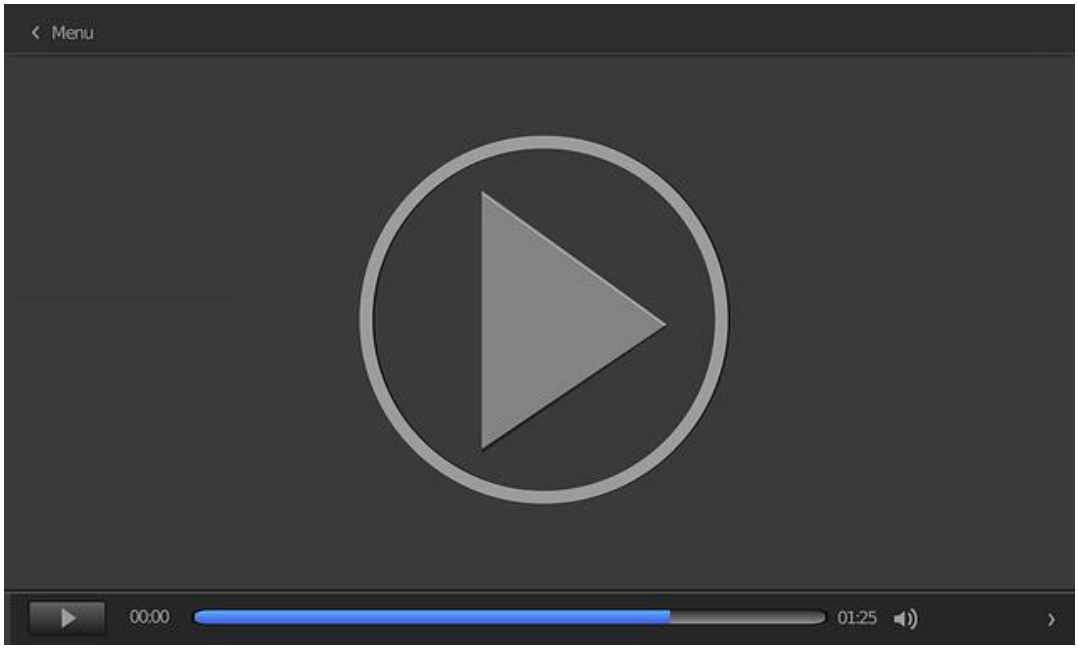
01.01. Aplicaciones del procesamiento de lenguaje natural

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=b037af01-ef52-426a-9654-af5b0149cd5e

Ideas clave

En el vídeo *Historia del procesamiento del lenguaje natural* se resume la historia de este, al tiempo que destaca los hitos más importantes.



01.02. Historia del procesamiento del lenguaje natural

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=8a2b6d13-252e-46f7-80e5-ac6b00c2e82f

1.5. Conocimiento del lenguaje utilizado en el PLN

Lo que distingue a las aplicaciones de procesamiento de lenguaje de otros sistemas de procesamiento de datos es **su uso del conocimiento del lenguaje.** Por ejemplo, un programa que cuenta el número de bytes, palabras y líneas en un archivo de texto podemos considerarlo como una aplicación de procesamiento de datos ordinaria.

Sin embargo, si para contar las palabras en el archivo de texto se requiere conocimiento sobre lo que significa ser una palabra porque se quieren contar el número de pronombres que aparecen en el archivo, ese programa se convierte en un sistema de procesamiento de lenguaje natural.

Por supuesto, este sistema de procesamiento de lenguaje es extremadamente simple en comparación con los agentes conversacionales, los sistemas de traducción automática y los sistemas de búsqueda de respuestas, que requieren un conocimiento mucho más amplio y profundo del lenguaje.

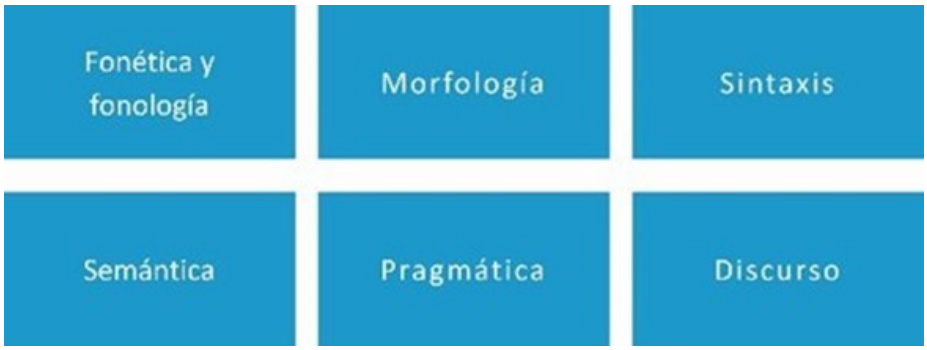


Figura 5. Conocimientos necesarios para tareas complejas de PLN. Fuente: elaboración propia.

Un agente conversacional necesita reconocer las palabras de una señal de audio y generar una señal de audio de una secuencia de palabras. Estas tareas de

reconocimiento de la voz y del habla son tareas de síntesis que requieren conocimiento sobre **fonética y fonología.** Es decir, conocimiento de cómo se pronuncian las palabras a partir de la secuencia de sonidos y cómo se generan cada uno de estos sonidos acústicamente.

Fonética y fonología: conocimiento sobre los sonidos lingüísticos

El agente conversacional también necesita **reconocer y saber producir variaciones de las palabras.** Por ejemplo, reconocer que «puertas» es el plural de una palabra y saber generar una frase en la que esta palabra se utilice en plural. Entonces, estas tareas requieren conocimiento sobre morfología, es decir, la forma en que las palabras se descomponen en partes que tienen un significado como puede ser la raíz de la palabra y una terminación que indique que es un plural.

Morfología: conocimiento de los componentes significativos de las palabras

Si vamos más allá de las palabras como elementos aislados y entendidas de forma individual, un agente conversacional debe usar conocimiento estructural para encadenar las palabras que constituyen su respuesta. El agente debe saber identificar que una secuencia de palabras no tiene sentido a pesar de que el conjunto de palabras original pudiera tener sentido si se ordenaran las palabras de otra forma. El conocimiento necesario para ordenar y agrupar palabras se llama sintaxis.

Sintaxis: conocimiento de las relaciones estructurales entre palabras

Para responder a una pregunta, el agente conversacional puede necesitar saber algo sobre la semántica léxica, es decir, sobre el significado de cada una de las palabras, así como sobre semántica composicional o el significado de varias palabras que se utilizan de forma conjunta en una combinación de palabras.

Semántica: conocimiento del significado

Además de comprender el significado de las palabras, el agente conversacional puede necesitar saber algo sobre la relación de las palabras con la estructura

sintáctica: si un sintagma es un complemento circunstancial de tiempo o un complemento del nombre.

Por lo tanto, el conocimiento sobre la sintaxis y el conocimiento sobre la semántica se van a utilizar de forma conjunta en el procesado del lenguaje natural.

El agente conversacional necesita determinar también el tipo de expresión que le ha interpuesto el usuario. Necesita saber si la expresión con la que este le acaba de interpelar es una pregunta a la que debe dar una respuesta hablada, una solicitud para que realice una acción o un simple enunciado o declaración sobre un hecho. Además, el agente puede determinar usar expresiones más formales y educadas en función de su interlocutor y cómo este le haya preguntado. Entonces, el agente necesita conocimiento sobre el diálogo o la pragmática para poder identificar la intención que tiene el usuario al interpelarle y dar una respuesta acorde.

Pragmática: conocimiento de la relación del significado con los objetivos y las intenciones

Por último, el agente conversacional necesita interpretar palabras o expresiones que hacen referencia a términos que han aparecido anteriormente en la conversación. Sería el caso de pronombres o sintagmas nominales con determinantes que se refieren a partes previas del discurso. Es por eso por lo que el agente examina las preguntas anteriores que se formularon previamente y utiliza el conocimiento sobre el discurso previo para resolver las referencias cruzadas.

Discurso: conocimiento sobre unidades lingüísticas más grandes que un solo enunciado

En conclusión, para realizar tareas complejas de PLN se necesitan **diferentes tipos de conocimiento del lenguaje,** concretamente conocimiento sobre la fonética y fonología, la morfología, la sintaxis, la semántica, la pragmática y el discurso.

1.6. Referencias bibliográficas

Backus, J. W. (1959). *The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM Conference.* International Business Machines Corp., New York, USA.

Bresnan, J. y Kaplan, R. M. (1982). Introduction: Grammars as mental representations of language. En Bresnan, J. (Ed.), *The Mental Representation of Grammatical Relations.* MIT Press.

Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. *En Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 740-750). Doha, Catar: Association for Computational Linguistics.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2*(3), 113-124.

Chomsky, N. (1959). A review of B. F. Skinner's «Verbal Behavior». Language, 35, 26-58.

Church, K. W. (1980). *On memory limitations in natural language processing* (Tesis de maestría, Massachussetts Institute of Technology). https://www.researchgate.net/publication/230875927_On_Memory_Limitations_in_Natural_Language_Processing

Cohen, P. R. y Perrault, C. R. (1979). Elements of a planbased theory of speech acts. *Cognitive Science, 3*(3), 177–212.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. y Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine*

Learning Research, 12, 2493–2537.

Colmerauer, A. (1978). Metamorphosis grammars. En L. Bolc (Ed.), *Natural Language Communication with Computers, Lecture Notes in Computer Science 63* (pp. 133-189). Springer Verlag.

Davis, K. H., Biddulph, R. y Balashek, S. (1952). Automatic recognition of spoken digits. *Journal of the Acoustical Society of America, 24(6*), 637-642.

Dozat, T., Qi, P., y Manning, C. D. (2017). Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. En *Proceedings of the CoNLL 2017 Shared Task* (pp. 20-30). Vancouver, Canadá: Association for Computational Linguistics.

Fillmore, C. J. (1968). The case for case. En E. W. Bach y R. T. Harms, (Eds.), *Universals in Linguistic Theory* (pp. 1-88). Nueva York, Estados Unidos: Holt, Rinehart & Winston.

Foland Jr., W. R. y Martin, J. H. (2015). Dependency-based semantic role labeling using convolutional neural networks. En *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics* (*SEM 2015) (pp. 279-289). Denver, Estados Unidos: *SEM 2015 Organizing Committee.

Goldwater, S. y Griffiths, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *Annual Meeting of the Association of Computational Linguistics* ACL-07, 744-751. http://aclweb.org/anthology/P07-1094

Grosz, B. J. (1977). The representation and use of focus in a system for understanding dialogs. *International Joint Conference on Artificial Intelligence,* 67-76.

Grosz, B. J. y Sidner, C. L. (1986). *Attention, intentions, and the structure of discourse. Computational Linguistics, 12*(3), 175-204.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua,* 44, 311-338.

Hobbs, J. R. (1979). *Coherence and coreference. Cognitive Science,* 3, 67-90.

Kaplan, R. M. y Kay, M. (1981). Phonological rules and finite-state transducers. *Annual meeting of the Linguistics Society of America.* Linguistic Society of America.

Kay, M. (1979*). Functional grammar. Proceedings of the Annual Meeting of the Berkeley Linguistics Society, Estados Unidos.* http://linguistics.berkeley.edu/bls/previous_proceedings/bls5.pdf

Kleene, S. C. (1951). *Representation of events in nerve nets and finite automata.* Santa Mónica, Estados Unidos: RAND Corporation. https://www.rand.org/pubs/research_memoranda/RM704.html

Kucera, H. y Francis, W. N. (1967). *Computational analysis of present-day American English.* Brown University Press.

Lang, J. y Lapata, M. (2014). Similarity-driven semantic role induction via graph partitioning. *Computational Linguistics, 40*(3), 633-669.

McCulloch, W. S. y Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5,* 115-133.

Naur, P., Backus, J. W., Bauer, F. L., Green, J., Katz, C., McCarthy, J., Perlis, A. J., Rutishauser, H., Samelson, K., Vauquois, B., Wegstein, J. H., van Wijnagaarden, A. y Woodger, M. (1960). Report on the algorithmic language ALGOL 60. *Communications of the ACM, 3*(5), 299–314.

Newell, A., Shaw, J. C. y Simon, H. A. (1959). Report on a general problem-solving program. En UNESCO (Ed.), *Information processing: proceedings of the International Conference on Information Processing* (pp. 256-264). Oldenbourg Verlag.

Pereira, F. C. N. y Warren, D. H. D. (1980). Definite clause grammars for language analysis: a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence, 13*(3), 231-278.

Perrault, C. R. and Allen, J. (1980). A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics, 6*(3-4), 167-182.

Quillian, M. R. (1968). Semantic memory. En M. Minsky (Ed.), *Semantic Information Processing* (pp. 227–270). MIT Press.

Schank, R. C. y Riesbeck, C. K. (Eds.). (1981). *Inside computer understanding: five programs plus miniatures.* Lawrence Erlbaum.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379-423.

Simmons, R. F. (1973). Semantic networks: Their computation and use for understanding English sentences. En R. C. Schank y K. M. Colby (Eds.), *Computer Models of Thought and Language* (pp. 61-113). San Francisco, Estados Unidos: W.H. Freeman and Co.

Sirts, K., Eisenstein, J., Elsner, M., y Goldwater, S. (2014). POS induction with distributional and morphological information using a distance-dependent Chinese restaurant process. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,* 265-271. http://www.aclweb.org/anthology/P14-2044

Titov, I. y Klementiev, A. (2012). A Bayesian approach to unsupervised semantic role induction. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics,* 12–22. https://dl.acm.org/citation.cfm?id=2380821

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, 42(*1), 230-265.

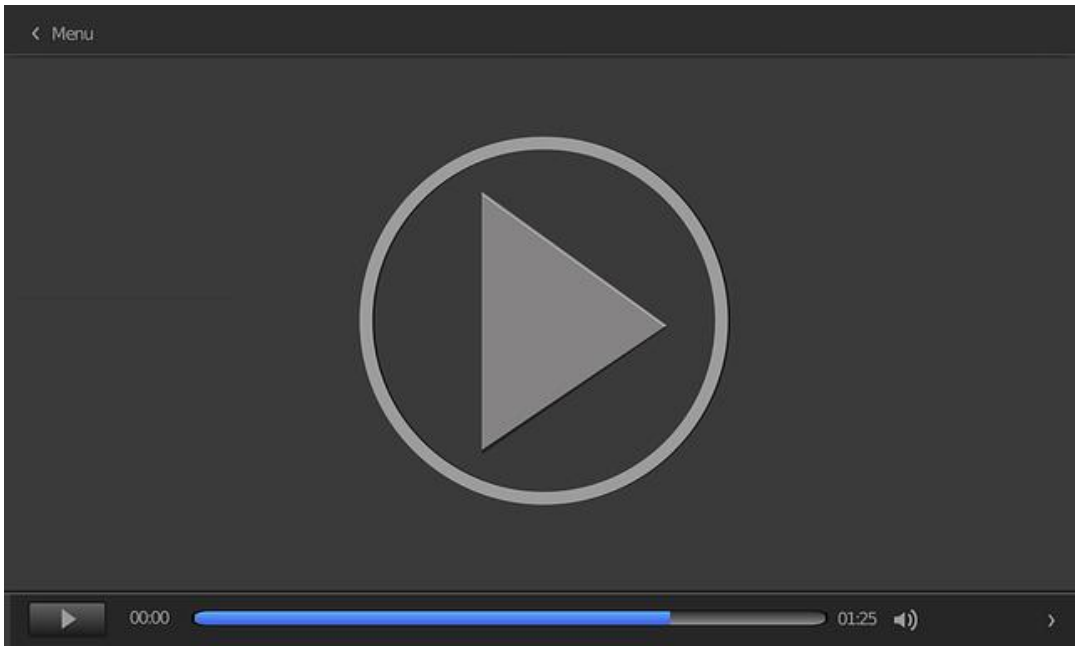
Winograd, T. (1972). *Understanding natural language. Cognitive Psychology, 3*(1), 1-191.

Woods, W. A. (1967). *Semantics for a Question-Answering System.* Garland Publishing.

Sistemas de diálogo

Mooc Málaga. (2015). *Sistemas de diálogo.* https://www.youtube.com/watch?v=9tfhNWqsx-8

Este vídeo muestra una breve introducción a los sistemas de diálogo hablado que son una de las aplicaciones del lenguaje natural que más interés suscitan.



Sistemas de diálogo

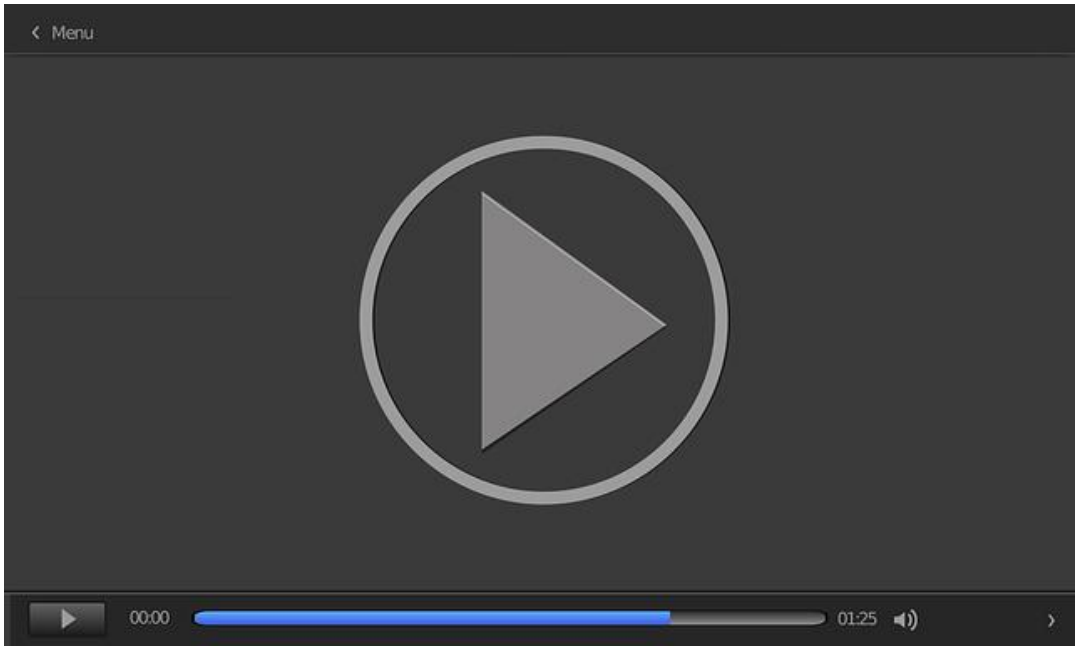
Accede al vídeo:

https://www.youtube.com/embed/9tfhNWqsx-8

Búsqueda de respuestas

Mooc Málaga. (2017). *Búsqueda de respuestas.* https://www.youtube.com/watch?v=49_KA89mPHg

Este vídeo muestra una breve introducción a los sistemas de búsqueda de respuestas donde el PLN permite que estos sistemas vayan más allá que los sistemas de recuperación de información.



Búsqueda de Respuestas

Accede al vídeo:

https://www.youtube.com/embed/49_KA89mPHg

1. Indica las afirmaciones correctas sobre el procesamiento del lenguaje natural:

A. También se llama reconocimiento de la voz.

B. Es un campo de la inteligencia artificial que tiene como objetivo que las máquinas realicen tareas que involucren el lenguaje humano.

C. Es un campo interdisciplinar que involucra disciplinas tan diversas como el procesamiento de la señal, el análisis sintáctico y semántico, la morfología, fonología y pragmática, la psicolingüística y la sociolingüística.

D. Es un campo en el que trabajan informáticos, ingenieros, lingüistas, sociólogos y psiquiatras.

2. Indica las afirmaciones correctas sobre el procesamiento del lenguaje natural en las décadas de 1940 y 1950:

A. El procesamiento del lenguaje natural aparece justo antes de la Segunda Guerra Mundial.

B. Uno de los paradigmas fundacionales del procesamiento del lenguaje natural se basa en uso de autómatas finitos y más concretamente cadenas de Markov.

C. Uno de los paradigmas fundacionales del procesamiento del lenguaje natural se basa en las ideas de Shannon descritas en la teoría de la información.

D. En esa época aparece la primera máquina capaz de reconocer la voz.

3. Indica las afirmaciones correctas sobre el paradigma simbólico en el período de 1957 a 1970:

A. Una de sus líneas de investigación se basa en la teoría del lenguaje formal y la sintaxis generativa.

B. Una de sus líneas de investigación se basa en la inteligencia artificial.

C. Aplicaba el método bayesiano.

D. Utilizaba los primeros corpus *online.*

4. Indica las afirmaciones correctas sobre los cuatro paradigmas de investigación que se dieron en el período de 1970 a 1983:

A. El paradigma estocástico utilizaba modelos ocultos de Markov para el reconocimiento de la voz.

B. El paradigma del modelado del discurso se basó en la estructura y el enfoque del discurso.

C. El paradigma basado en lógica utilizaba la lógica de predicados.

D. En el paradigma de la comprensión del lenguaje natural se utilizaban las redes semánticas.

5. Indica las afirmaciones correctas sobre los cambios que sufrió el campo del procesamiento del lenguaje natural a finales del pasado milenio:

A. Se recuperaron los modelos de estados finitos y el empirismo

B. Se volvieron populares los modelos probabilísticos.

C. Se volvieron estándares los modelos basados en datos.

D. Se empezaron a comercializar algunos productos con tecnologías del procesamiento del lenguaje natural.

6. Indica las afirmaciones correctas sobre como el auge del aprendizaje automático ha cambiado el procesamiento del lenguaje natural desde el año 2000:

A. El *deep learning* es un ámbito por explotar que se va a estudiar en los próximos años.

B. El aprendizaje supervisado se ha usado en el análisis semántico.

C. El aprendizaje no supervisado se ha usado en el análisis sintáctico y semántico.

D. El aprendizaje no supervisado es una técnica más eficiente que el aprendizaje supervisado al no requerir anotaciones de los corpus.

7. Indica cuales de las siguientes aplicaciones utilizan técnicas del procesamiento del lenguaje natural:

A. Los agentes conversacionales.

B. La corrección ortográfica.

C. La búsqueda de respuestas.

D. La traducción automática.

8. Si un agente conversacional tiene que crear una respuesta en la conversación y para ello encadena diferentes palabras, ¿qué tipo de conocimiento necesita para realizar esta tarea? Indica las respuestas correctas:

A. Conocimiento sobre el significado de las palabras.

B. Conocimiento sobre las relaciones estructurales entre palabras.

C. Conocimiento fonético.

D. Conocimiento sintáctico.

9. Si un agente conversacional necesita interpretar si el usuario le ha hecho una pregunta o simplemente le ha contado un hecho, ¿qué tipo de conocimiento necesita para realizar esta tarea? Indica las respuestas correctas:

A. Conocimiento pragmático.

B. Conocimiento sobre el diálogo.

C. Conocimiento sobre el discurso.

D. Conocimiento sobre conocimiento sobre la relación del significado con los objetivos y las intenciones.

10. Si un agente conversacional necesita generar una frase en la que el número del nombre en el sujeto debe ser un plural para que concuerde con el verbo, ¿qué tipo de conocimiento necesita para realizar esta tarea? Indica las respuestas correctas:

A. Conocimiento morfológico.

B. Conocimiento semántico.

C. Conocimiento sintáctico.

D. Conocimiento sobre los componentes significativos de las palabras.