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

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.

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

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 que el agente examina las preguntas anteriores que se formularon previamente y utiliza el conocimiento sobre el **discurso** previo para resolver las referencias cruzadas.

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.

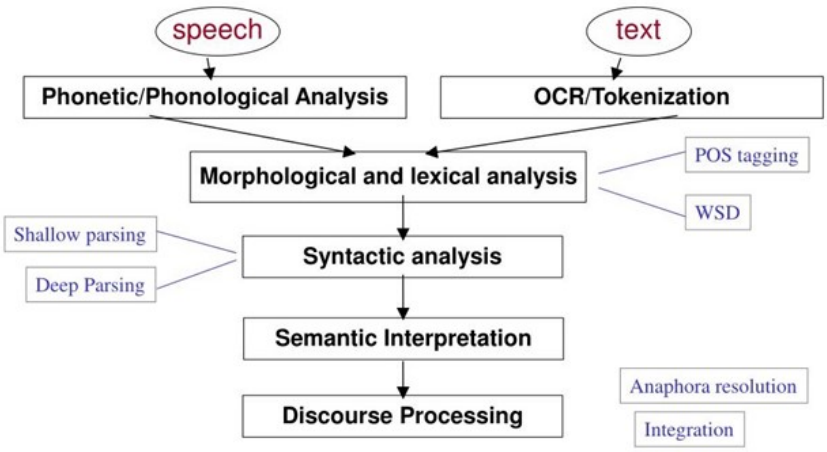

Figura 7. Pipeline o proceso de PLN. Fuente: SlidePlayer, s.f.

Como veremos en el siguiente capítulo, estos encadenamientos de análisis e interpretaciones permiten crear *pipelines* que sólo deben ser ajustados en los parámetros necesarios de cada uno de ellos para que puedan dar el mejor resultado.

Lingüística computacional. Experto en PLN

Si ampliamos, por tanto, el área de actuación concreta, el **procesamiento del lenguaje natural** (NLP) resulta de una combinación de *la informática, las ciencias de la información, la IA y la lingüística computacional*.

Aunque las computadoras, y la informática, destacan en el manejo de grandes **conjuntos de datos estructurados**, requieren un poco de **ayuda cuando se trata de lenguajes humanos** con cientos de idiomas y dialectos diferentes, todos con su propio conjunto de reglas gramaticales, jerga, términos y sintaxis.

¿Alguna vez te has preguntado cómo Google o Alexa pueden entender lo que dices? Algunos asistentes inteligentes ya conocen multitud de lenguas. Por ejemplo, Siri habla 20 idiomas y *Google Translate* supera el centenar, siendo la aplicación que más lenguas incluye. Sin embargo, existen entre 5.000 y 7.000 lenguas en el mundo, lo que hace mucho más complejo extender el PLN a todos estos idiomas".

La respuesta nos la encontramos dentro de una de las profesiones más demandadas en este campo: la **lingüística computacional**. Se trata de una disciplina reciente que se ha incorporado a los planes de estudio de las licenciaturas y grados de formación lingüística y que consiste, a grandes rasgos, en el estudio de la lengua y el desarrollo de aplicaciones lingüísticas con medios computacionales. Es decir, que son responsables de PLN y de que las máquinas (como los asistentes de voz) nos entiendan cuando hablamos con ellas.

"El Lingüista Especializado en Procesamiento de Lenguaje Natural será el experto humanista que conoce los modelos lingüísticos en profundidad y da apoyo al equipo de desarrollo de software relacionado con el procesamiento del lenguaje. Deberá conocer de primera mano los diferentes modelos para el procesamiento del lenguaje natural; desde los modelos lógicos que reconocen los patrones estructurales del idioma, hasta los modelos probabilísticos, que incluyen algoritmos que estudian las colecciones de ejemplos y datos recogidos, e infieren las respuestas basándose en la probabilidad que tienen de aparecer en un determinado contexto. Además, deberá tener nociones de programación para poder asistir a los ingenieros en el proceso de desarrollo de las aplicaciones". Elena Ibáñez, fundadora de Singularity Experts

Es cierto, que para trabajar como lingüista computacional (también denominado experto en PLN) **no solo hay que saber de morfología, sintaxis y semántica, sino también de programación y algo de código.** Todos estos profesionales han tenido que aprender ciertos lenguajes de programación para hacer su trabajo; en algunos casos de forma profunda pues son colaboradores en el desarrollo junto a ingenieros y matemáticos; en otros, profundizando en materias para poder tener una interlocución fluida con éstos.

Las capacidades que deben tener estos perfiles incluyen, aparte de **conocimientos lingüísticos** propios de carreras filológicas o de "letras", aprender **programación** sin límites, conocimientos de estadística y tener un buen nivel de determinados idiomas

(el inglés como base).

En la siguiente figura podemos ver las características que nos propone freelancermap.com como componentes de este puesto de experto en PLN.

Figura 8. Lingüista computacional o experto en PLN. Fuente: Sensoricx, s.f.

Como nos presenta esta empresa se trata de una profesión con unas determinadas funciones, conocimientos imprescindibles y una formación específica en áreas de letras y ciencias informáticas.

Funciones de un lingüista computacional o experto en procesamiento de lenguaje natural:

- Diseño y desarrollo de sistemas de procesamiento del lenguaje natural.
- Definición de conjuntos de datos apropiados para el aprendizaje de idiomas.
- Uso de representaciones de texto efectivas para transformar el lenguaje natural en características útiles.

- Entrenamiento del modelo desarrollado y realización de experimentos de evaluación.
- Implementar los algoritmos y herramientas adecuados para las tareas de NLP.
- Realizar análisis estadísticos de los resultados y perfeccionar los modelos existentes.
- Mantener las bibliotecas y marcos de NLP.
- Implementar los cambios según sea necesario y analizar los errores.
- Estar al día en novedades relativas a inteligencia artificial y/o big data.

Habilidades necesarias para un ingeniero/a especialista de NLP

- Conocimientos de NLP y NLU (Natural language Understanding).
- Comprensión de las técnicas de representación de texto, algoritmos, estadísticas.
- Experiencia en traducción automática y compilación de datos.
- Conocimiento de los marcos de aprendizaje automático y librerías de *Deep learning.*
- Familiaridad con los marcos de Big Data como Spark, Hadoop.
- Conocimientos de programación: Python, Java y/o R, así como de sus librerías específicas de PLN.
- Conocimientos avanzados de scripting y *class-based programming*.
- Experiencia en transfer learning BERT o GPT.
- Fuerte capacidad de resolución de problemas y de trabajo en equipo con metodologías ágiles.
- Análisis sintáctico y semántico.
- Conocimiento de los oleoductos de CI/CD.

- Fuertes habilidades de comunicación.

Por último, destacar que existen numerosos empleos no sólo en el desarrollo de software relacionado con PLN, sino que también nos encontramos nuevas profesiones que permiten a la empresa reducir costes y optimizar sus procesos no productivos.

Un ejemplo es la nueva profesión de *botmaster* que es el responsable de los *chatbots* y sus interacciones. El coste de solucionar dudas de los clientes por teléfono o de manera presencial es 40 veces más caro que con una máquina. Éstas hablan varios idiomas, pueden atender a miles de clientes simultáneamente, son multiproceso y no se toman vacaciones ni días libres.

Un lingüista es quien enseña a hablar a la máquina con el cliente. La creación de la herramienta es fundamental: cómo debe funcionar la máquina, el diccionario, definen los algoritmos... Se trata de un “educador” de la máquina, la mejora, y le proporciona nuevas funcionalidades.

3.4. Aplicaciones del Procesamiento del Lenguaje Natural

El PLN cubre varias disciplinas que proporcionan funcionalidades a muchos sistemas que actualmente estamos utilizando en nuestro entorno social y de empresa. Veremos las grandes líneas inicial que nos llevan a aplicaciones particulares en la empresa de alto impacto en la actualidad.

¿Qué podemos hacer?

1. **Reconocimiento de patrones de lenguaje.** Al procesar grandes cantidades de documentos, el reconocimiento de patrones permite filtrar datos importantes en cadenas de texto de forma inmediata. Es el primer paso para que la recuperación de información y la clasificación de textos sea posible.

2. **Recuperación de información.** El reconocimiento de patrones de lenguaje hace sencilla la tarea de encontrar un fragmento en particular dentro de cantidades de texto inmanejables para el ser humano en poco tiempo. No inventa palabras o frases nuevas, sino que identifica la información valiosa.

3. **Traducciones automáticas de idiomas.** Ya sea con voz o texto, esta capacidad utiliza datos que se procesan por la lingüística computacional y está en un proceso constante de mejora y aprendizaje con el uso de *Deep learning*.

4. **Clasificación de información.** Gracias a la aplicación de palabras clave, la información puede categorizarse para que su consulta sea más eficiente.

5. **Resumen de textos.** Al igual que con la clasificación, resumir un documento de gran extensión se apoya en ciertas palabras o frases clave. Otro de los usos que se le da a esta función de clasificación, es la de detección de *spam*.

6. **Generación de lenguaje natural (GLN o NLG).** Es uno de los más ambiciosos de esta lista, pues es lo que permite que una máquina responda las interacciones de un

humano con frases nuevas y no como lo haría un *chatbot* (que elige de una lista establecida la opción que mejor se adapta).

7. **Comprensión del lenguaje natural (CLN o NLU).** Para comprender ciertos mensajes con su intencionalidad, el procesamiento de lenguaje natural está impactando en el análisis de las emociones que se expresan a través de frases dentro de contextos de conversación o diálogo. Formas más complejas utilizan técnicas de análisis de sentimientos. Áreas de marketing la utilizan para saber qué sienten los usuarios sobre una marca, producto o servicio, desde datos de entrada como mensajes, comentarios o reacciones en diferentes redes sociales.

¿Para qué utilizarlo en la empresa?

Aunque podría clasificarse en tres grandes disciplinas como son los agentes conversacionales (o sistemas de diálogo), la traducción automática, la búsqueda de respuestas o la corrección ortográfica y la verificación gramatical, nos encontramos en un mundo en constante innovación.

Actualmente existen muchas aplicaciones en modo piloto o en laboratorio, y aparecen continuamente *startups* con soluciones a casi todo tipo de problemas, dando como resultado un mundo en constante evolución.

Vamos a detallar varias de ellas, de entre muchas otras, animando al lector a trasladar estas funcionalidades a nuevas áreas creando soluciones 'nuevas' y sistemas orientados a tareas específicas. Como podréis ver muchos de ellos están interrelacionados en soluciones más complejas.

Agentes conversacionales

También llamados sistemas de diálogo son programas que conversan con las personas a través del lenguaje natural. En general, 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 las mismas deben mantener el estado de la conversación y ser capaces de generar nuevas frases que continúen la conversación que mantienen con la persona.

Los **asistentes virtuales** permiten pedirle a Siri, Alexa o Google que encienda las luces de tu cocina o cambie la canción que estás escuchando. Según esté programados estos asistentes virtuales reaccionan a comandos de voz por una orden establecida y analizan las palabras que escuchan para realizar una búsqueda de información dentro de sus dispositivos —como encontrar la dirección de un contacto guardado— o en internet.

Figura 9. Altavoces con Siri integrado. Fuente: Apple2fan, 2018.

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. Además, tienen la capacidad de aprender nuevas interacciones, según el nivel de sofisticación de cada uno, y son muy útiles para atender consultas sencillas, pero muy comunes dentro de ámbitos como el de atención sanitaria inicial, resolución de problemas técnicos básicos, redirección de

llamadas a atención al cliente e incluso trámite inicial en Administración Pública o consultas a bufetes de abogados.

Es muy interesante realizar seguimiento de las múltiples soluciones porque pueden aportar grandes ideas en el desarrollo de nuestra propia solución en un segmento de mercado distinto al que fue desarrollado. Recomendamos esta lista que sugiere (https://www.userlike.com/es/blog/los-mejores-chatbots) los 8 mejores en función utilidad, facilidad de uso y conversacionalidad (también muestra cómo crear uno para usar con API o desde 0):

- Seguridad sanitaria mundial. No es conversacional, pero ha permitido difundir información contrastada de forma masiva.

Figura 10. Chatbot de la OMS para el COVID-19 (Who Health Alert). Fuente: WhatsApp, s.f.

- Erica de Bank of America. Acercar la terminología bancaria a la conversación normal de un cliente. El resultado: más de 14 millones de personas que recurrieron al *bot* para gestionar sus finanzas personales.

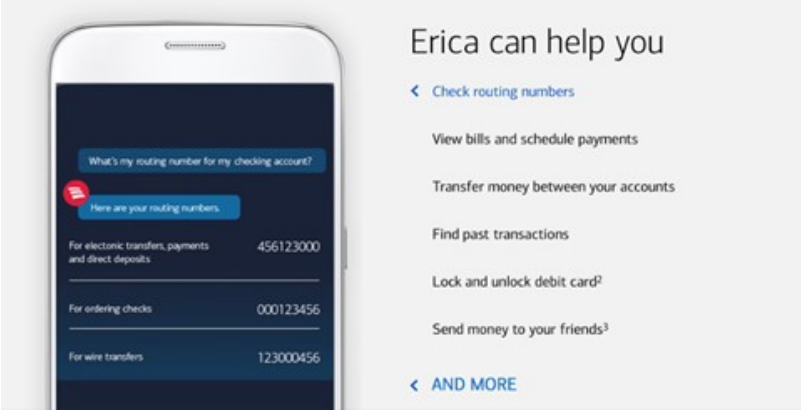

Figura 11. Erica, la asistente virtual de Bank of America. Fuente: Bank of America, s.f.

- Domino's Dom. Hecha en colaboración con los creadores de Siri, no sólo hace pedidos – su función principal- sino que utiliza el humor para acercar la conversación.

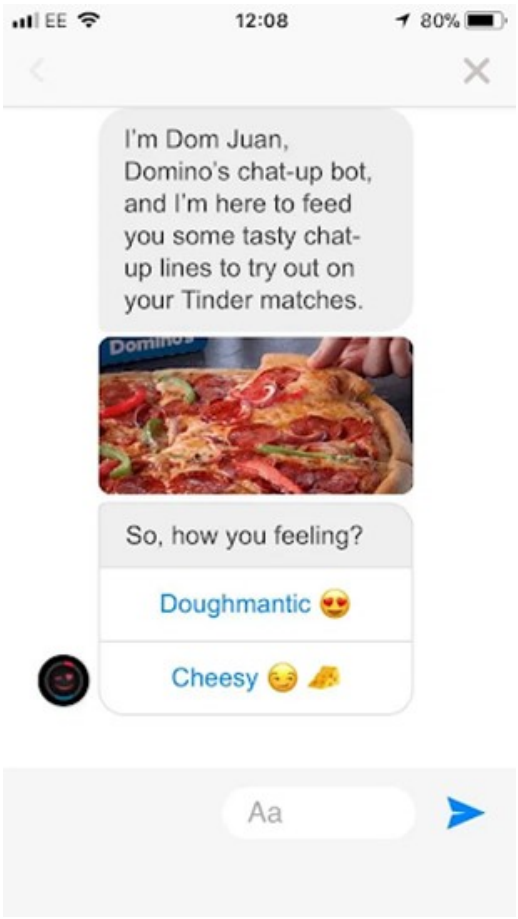

Figura 12. Dom, la felicitación de San Valentín del robot pizzero. Fuente: The Sun, 2019.

- Juliet de Westjet. Un *chatbot* que permite resolver gran cantidad de dudas que tienen los pasajeros de la compañía aérea Westjet para agilizar la atención más sencilla o concreta, derivando las preguntas complejas a personal de atención.

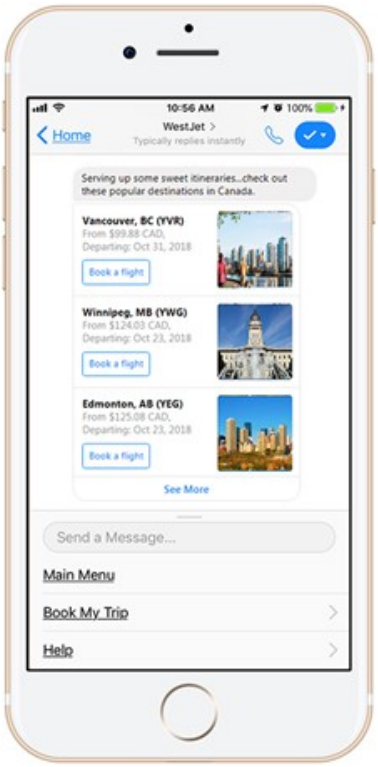

Figura 13. Juliet, asistente de viajes digital de WestJet. Fuente: WestJet, s.f.

- DoNotPay. La aplicación DoNotPay es el hogar del primer robot abogado del mundo que presta servicios legales a residentes en Estados Unidos y en el Reino Unido.

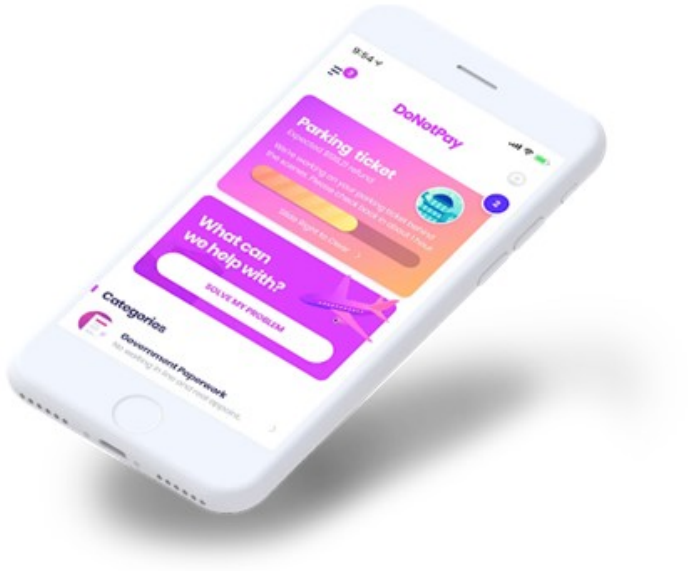

Figura 14. Aplicación DoNotPay, primer abogado 'robot'. Fuente: Periodismo, s.f.

- AskBenji. En este caso nos encontramos con una solución original creada por los mismos estudiantes que necesitaban un 'ayudante' para entender cómo solicitar las becas y préstamos que proporciona la Ayuda Federal. Es un primer paso porque trabaja sólo con palabras clave, pero dispone ya de uno de los primeros criterios de cualquier proyecto de chatbot (e incluso AI): ser útil y económico.

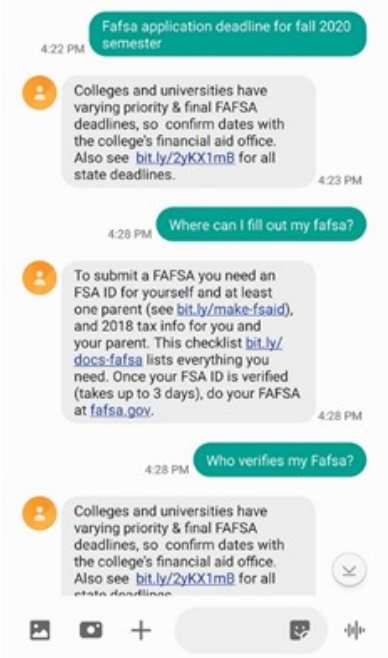

Figura 15. Ask Benji, chatbot de Arizona. Fuente: Ask Benji, s.f.

- Andy. Este *bot* permite no sólo la realización de ejercicios y demás utilidades de uso normal en la enseñanza de idiomas, sino la de tener conversaciones informales en función del usuario. Además realiza correcciones, lo que incrementa la complejidad del sistema.

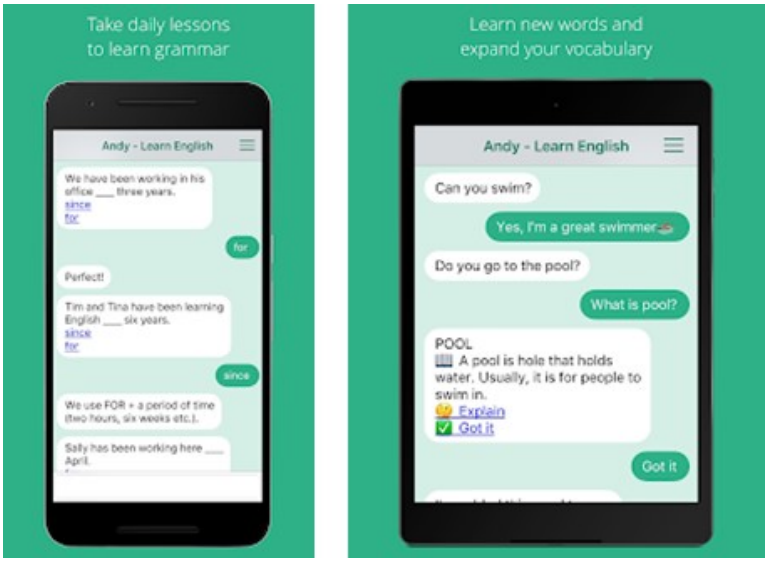

Figura 16. Andy - Bot de habla inglesa. Fuente: Google Play, s.f.

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, 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**.

Por ejemplo Google es una de las empresas que más ha invertido en **sistemas de traducción automática,** con su traductor que utiliza un motor estadístico propio. Los sistemas de autocorrección y autocompletado de texto, también utilizan Procesamiento del Lenguaje Natural (PLN o NLP). Hace poco superó la traducción

automática de más de 100 idiomas en su plataforma, tanto en voz como en texto. Con el paso del tiempo, Google Translate ha podido mejorar los alcances de la herramienta con el uso de *deep learning* y con ello mejorar su capacidad de comprender algunos juegos de palabras, metáforas e intenciones, aunque está todavía lejos de ser perfecto.

Otro de los grandes 'players en esta área de la traducción es la empresa DeepL que no sólo permite su uso web sino la integración mediante un API, y con mejora como son la adaptación de la traducción a **diferentes grados de formalidad** e incluso con diccionarios personalizados.

Por otro lado, los **sistemas de reconocimiento de voz** procesan los mensajes en voz humana, reconoce la separación de palabras e incluso los signos gramaticales. Después los transforman en texto, los interpretan y comprenden la intencionalidad de los mismos, y tras la generación de la respuesta en texto, se vuelve a transformar en voz humana a través de la síntesis de voz. La síntesis del habla o de voz, es la que capacita a la máquina para poder generar y reproducir habla en lenguaje natural, añadiendo riqueza al discurso e incluso entonación adecuada.

Recuperación de la información

Otro campo que está aportando gran capacidad de síntesis a las empresas es la **recuperación de información – RI** (*en inglés* ***Information Retrieval*** - IR), es el campo dentro del PLN que se encarga de procesar textos de documentos, para poder recuperar partes específicas en base a palabras clave. No genera nuevas frases, por lo que no necesita utilizar reglas gramaticales.

Esta área está siendo impulsada de forma creciente ya que los agentes de los mercados financieros buscan tener acceso a las más avanzadas técnicas de NLP para automatizar tareas sofisticadas como **clasificación de documentos**, **extracción de información**, **generación de insights financieros** y resumen de contenido en tiempo casi real. No es tan "inteligente" como la Generación del

Lenguaje Natural, pero, por ejemplo, puede llegar a 'redactar' informes concretos en el ámbito financiero después de 'leer' tendencias económicas, políticas y múltiples documentos escritos y/o hablados (podcasts, telediarios, debates económicos, …).

Búsqueda de respuestas

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.

Los motores de búsqueda (o buscadores) que son la principal aplicación permiten una comprensión alta de la intención de las consultas. Lo notas cuando te equivocas en una letra o quizá en la escritura de un nombre que no sabes cómo se escribe del todo, y el buscador te corrige o muestra la escritura correcta (con la opción de presentar los resultados con el término o frase que se ingresó originalmente) como podemos ver en la siguiente figura.

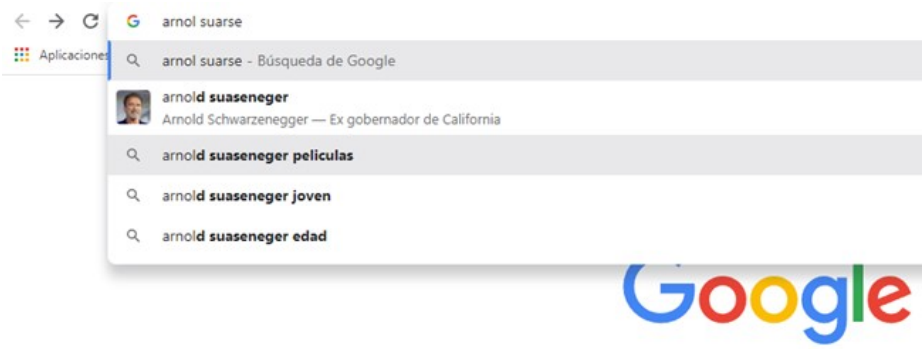

Figura 17. Ayuda a la búsqueda de respuestas en Google. Fuente: elaboración propia.

Y cuando se trata de imágenes o videos, ¿qué utiliza? Para eso sirven los alt-text, que describen las imágenes con precisión para dos razones principales: que las personas con problemas de visión sepan de qué se trata el archivo multimedia que acompaña un texto que leen en su dispositivo gracias a la síntesis de texto a voz, y para que los buscadores detecten mejor el contenido y lo relacionen con las consultas de los usuarios.

Modelado de temas

Un campo que ayuda en gran medida a la empresa es el denominado **Modelado de Temas o *Topic Modeling***. Es una capacidad importante para que los sistemas empresariales le den sentido a la información no estructurada, desde tickets abiertos con comentarios y quejas en sistemas de soporte a clientes así como la revisión de documentos empresariales.

El modelado de temas ayuda en muchos aspectos como:

- automatizar procesos para enrutar los comentarios y el correo de los clientes;
- categorizar y luego responder de manera efectiva a publicaciones en redes sociales, reseñas y otro contenido generado por el usuario de los distintos canales.
- responder más rápidamente a los elementos críticos al comprender los temas en las

interacciones omnicanal entrantes, así como responder de manera más efectiva al enrutar los materiales a las áreas de atención más apropiadas.

- enriquecer los sistemas internos de gestión del conocimiento de RRHH
- agilizar el seguimiento de la marca en los departamentos de marketing

Detección de spam

En la **detección de *spam*** en correo electrónico y sus adjuntos, el uso de palabras clave son las que permiten que el sistema de un correo electrónico logre clasificar ciertos mensajes como no deseados, además de realizar un análisis de seguridad en búsqueda de software que tiene potencial de dañar un dispositivo o robar información sensible. La clasificación es una de las capacidades más potentes del PLN.

Es lo que también permite que podamos clasificar el correo en favorito, publicidad o social, tal como lo hacen los actuales gestores de correo para una mejor gestión de los mensajes.

Autocorrección de texto

Por último, un área más sencilla aunque no menos interesante y útil en múltiples aplicaciones de procesamiento de texto es la **autocorrección de texto**. Mientras escribimos en un procesador de textos (incluidos los gestores de correo como Gmail), puedes activar la función de autocompletar o autocorregir para hacer la escritura más ágil y con menos errores. Por detrás existe un diccionario integrado en uno o más idiomas que identifica errores ortográficos, gramaticales o hasta detecta frases hechas para añadirlas rápidamente. Lo más impactante de esta herramienta es que puede alimentarse según lo permita el usuario, porque agrega o elimina palabras al diccionario para un trabajo más eficiente y personalizado (uno de los objetivos generales de la IA).

3.5. Entornos PLN para crear sistemas

Como vimos en la lección anterior, detrás de las aplicaciones basadas en el PLN hay varios procesos en los que es necesario realizar un tratamiento de la lengua, escrita o hablada, para que pueda ser analizada por una computadora. Los procesos realizados **tratan de simular el proceso que realizamos las personas para comprender e interpretar el lenguaje.** La diferencia entre este proceso y el efectuado por los seres humanos actualmente, es que una computadora puede analizar enormes masas de datos a velocidades muy rápidas, aunque no de manera tan exacta y precisa como lo hacemos los humanos.

Sistemas orientados a tareas de PLN

En la práctica, hasta ahora y previo a la irrupción de los algoritmos de redes neuronales artificiales profundas, el diseño de sistemas hacía uso de diferentes recursos. La IA no es la única aproximación utilizada pero ahora mismo es

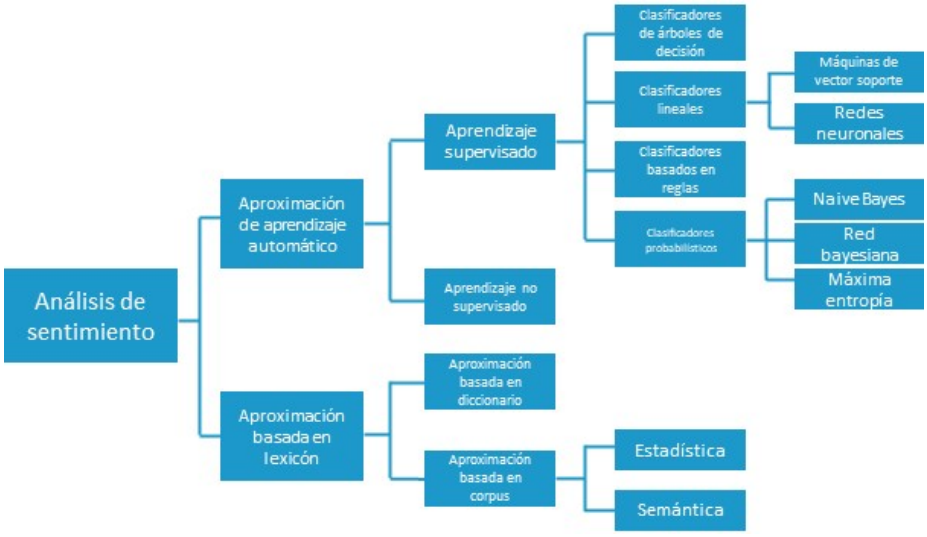

Figura 18. Aproximaciones técnicas a soluciones de análisis de sentimiento. Fuente: elaboración propia.

Por ejemplo, en la figura anterior el análisis de sentimientos tan de moda en el estudio de la opinión pública en RRSS (Twitter, Instagram, Facebook,…) puede hacerse según diferentes líneas de trabajo (o aproximaciones técnicas). Todas ellas requieren una serie de componentes que es muy conveniente conocer y manejar:

Recursos y ejemplo de procesamiento

Corpus lingüísticos

Recopilación de textos relacionados con el campo que queremos analizar. Servirán de conjunto de datos para aprender e inferir. Deben ser:

- *Variados*: diversos tipos de textos.
- *Representativos*: muestra de tamaño suficiente.
- *Equilibrados*: neutrales, con material de todos tipos en la misma proporción.
- *Tamaño adecuado*: no deben exceder el tamaño conveniente para poder trabajar con el conjunto de textos de manera efectiva.
- *Manejables por ordenador*: Su formato y dimensiones deben hacerlos operativos para las aplicaciones que queramos usar sobre el corpus

Tokenizadores

Son analizadores léxicos o segmentadores de palabras. Consisten en pequeñas unidades con sentido semántico. En este ejemplo se han creado *tokens* separados por espacios en blancos en la frase:

Frase:	La abuela juega con la aguja
Tokens:	'La' // 'abuela // 'juega' // 'con' // 'la' // 'aguja

Figura 19. Representación de tokens. Fuente: elaboración propia.

Hay una serie de dificultades a las que se enfrentan las aplicaciones que construyen los *tokens*:

- *Fronteras ambiguas*: no en todas las lenguas el espacio en blanco es una regla de separación de sintagmas con un único significado. En alemán, por ejemplo, las palabras se pegan unas a otras para formar palabras compuestas.
- *Formatos*: hay muchos formatos para estipular las fronteras de aplicación de las fechas, etc.…
- *Abreviaturas, siglas y apóstrofes*: Este tipo de partículas lingüísticas dificultan aún más la separación del texto en tokens.

N-gramas

Son agrupaciones de *n* tokens. Se combinan en el texto todos los posibles n-gramas de un determinado concepto, así determinamos estructuras de más de una palabra que se repiten con sentido en el mismo.

Tokens
'El // 'balón // 'de' // 'baloncesto // 'se' // 'perdió'
Unigramas
'El // 'balón // 'de' // 'baloncesto // 'se' // 'perdió'
Bigramas
'El balón' // 'balón de' // 'de baloncesto // 'baloncesto se' // 'se perdió'
Trigramas
'El balón de' // 'balón de baloncesto' // 'de baloncesto se' // 'baloncesto se perdió'

Figura 20. Representación de N-gramas. Fuente: elaboración propia.

Etiquetadores de partes de oraciones

Clasifican las palabras morfológicamente según sean verbos, adjetivos, preposiciones... Sirven para adjudicar la funcionalidad estructural de cada palabra en la estructura de la frase.

Lematizadores

Adjudican a cada palabra o token la raíz léxica de la misma en el diccionario. De este modo se simplifica el análisis semántico de las expresiones, por ejemplo:

Supresión de palabras funcionales

Consiste en eliminar en el texto las palabras sin léxico para crear esquemas de significado de las expresiones, por ejemplo:

El monitor del ordenador vale caro

Quitando las palabras funcionales quedará:

'monitor' // 'ordenador' // 'caro'

Ejemplo sobre librería NLTK

En la práctica este proceso permite modelizar un proceso que consigue realizar PLN sobre textos y llegar al análisis de sentimientos. En la siguiente figura podemos ver los pasos que sigue de forma resumida un proceso de aprendizaje automático para PLN.

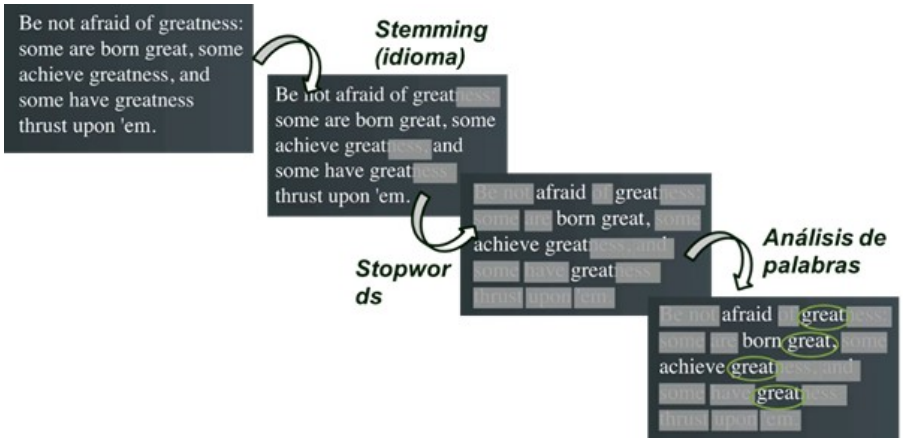

Figura 21. Pasos estructurados para analizar textos y extraer palabras para analizar.

Al final logramos reducir nuestro problema de forma considerable y tener datos que nos puedan dar sentido al análisis final del texto (sentimientos, tendencias, …).

Se propone al lector revisar la página web: Python NLTK Demos for Natural Language Text Processing and NLP (text-processing.com) en la que ver cómo utilizar la librería NLTK y sus funciones principales.

Mecanismo de Recuperación de Información (RI)

La recuperación de información o RI (en inglés **IR – Information Retrieval**), consiste en el aporte de documentos con información relacionados con una secuencia de palabras dada. Es el caso más básico de NLP, y con el que se está más familiarizado pues es el proceso que se realiza cada vez que buscamos en un motor de búsqueda en internet.

Cuando se habla del **TF*IDF** aplicado al SEO, los usuarios de herramientas SEO buscan la creación de textos únicos para mejorar el posicionamiento del sitio web en los resultados de búsqueda. Hasta ahora, la densidad de término se ha usado como único referente para la optimización de textos, sin embargo, la fórmula **TF*IDF** ofrece un modo mucho más preciso de optimizar el contenido.

Dado que los motores de búsqueda analizan la relación semántica entre los términos es muy importante optimizar semánticamente el contenido del sitio web. Este proceso se llama *Indexación de la Semántica Latente*.

Proceso

Para el proceso RI es clave la creación de listas de palabras relevantes 0 frecuentes en los documentos sobre los que se realiza la búsqueda.

El método más extendido es el **TF-IDF**, que proviene de ***Term Frequency-Inverse Document Frequency***, y es la combinación de algoritmos:

- **TF:** calcula frecuencia relativa de una palabra comparado con el nº total de palabras en un texto y también la de todas las palabras usadas en el texto
- **IDF:** determina la relevancia de un texto con respecto a una palabra clave específica (busca disminuir el peso de las repetitivas y con menor valor).

Generamos **listas de palabras clave** con una calificación o peso que indica cómo es de relevante respecto al documento seleccionado y al corpus en general.

Además, estas listas permiten **calificar a los documentos** del corpus con base en estas palabras clave, es decir, si las palabras clave tienen un gran peso, entonces el documento está más relacionado con ellas que uno con las mismas palabras clave pero con menor peso. Por tanto, cuando un usuario ingrese una consulta, los documentos que tengan las palabras de esa consulta con mayor peso serán los que muestre el sistema de búsqueda de información.

¿Cómo leerlo?

- Un **peso alto en tf-idf** se alcanza con una elevada frecuencia de término (en el documento dado) y una pequeña frecuencia de ocurrencia del término en la colección completa de documentos. Como el cociente dentro de la función logaritmo del idf es siempre mayor o igual que 1, el valor del idf (y del tf-idf) es mayor o igual

que 0.

- Cuando un término aparece en muchos documentos, el cociente dentro del logaritmo se acerca a 1, ofreciendo un valor de idf y de tf-idf cercano a 0.

Figura 22. Ejemplo de recuperación de información (RI).

En este método surge un problema natural: *textos con mayor longitud tendrán un peso mayor sobre la respuesta aunque tengan menor relación a la pregunta*.

Esto se soluciona **normalizando los textos del corpus**, esto es, modificándolos para que tengan longitudes y estructuras de dimensiones parecidas y comparables. Existen diversos criterios de normalización como:

- Normalización coseno.
- Normalización por pivote.
- Normalización por máximo TF.

Mecanismo de Extracción de Información (EI)

Este proceso de **extracción de la información o EI (**en inglés **IE – Information Extraction)** consiste en seleccionar estructuras y combinar datos que son encontrados, de manera explícita o implícita, en uno o más textos.

- Toma un texto en lenguaje natural de un documento fuente, y extrae los **hechos**

esenciales acerca de uno o más tipos de hechos predefinidos.

- Representa cada uno de los hechos como una plantilla donde los espacios son llenados con base en lo encontrado en el texto. **Sólo una pequeña parte** de la información encontrada es relevante para llenar la plantilla, el resto puede ser ignorada.

El propósito de la EI es **estructurar el texto posiblemente no estructurado**. Con esto se pueden obtener resultados que pueden ser otorgados de manera directa al usuario o ser modificados para que puedan ser insertados en una base de datos.

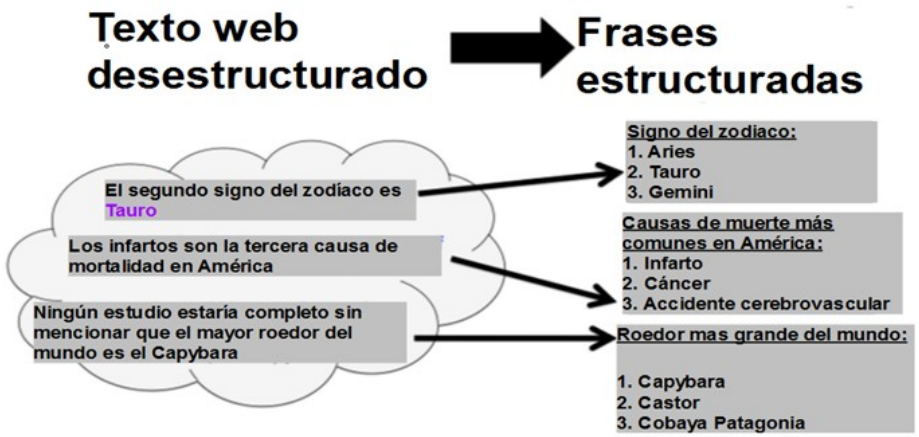

Figura 23. Propósito de la Extracción de Información.

Entornos de desarrollo y librerías específicas

Al tratarse de una rama de la inteligencia artificial como veíamos en la Figura 18, disponemos de entornos o módulos de desarrollo dentro de las grandes plataformas de IA en la nube y en concreto de machine learning como son entre otras muchas:

- **Microsoft Azure Cognitive Services – Language Understanding (LUIS)**. LUIS es un servicio de inteligencia artificial de reconocimiento del lenguaje natural (NLU) que permite a los usuarios interactuar con sus aplicaciones, bots y dispositivos IoT usando el lenguaje natural.

- **AWS – Amazon Comprehend.**Amazon Comprehend es un servicio de procesamiento de lenguaje natural (NLP) que utiliza el machine learning para descubrir información en datos no estructurados. En lugar de revisar los documentos, se simplifica el proceso y la información que no se ve es más fácil de entender.

- **Google Cloud Platform – Natural Language AI**. Natural Language utiliza el aprendizaje automático para mostrar la estructura y el significado de los textos. Puedes extraer información sobre personas, lugares o eventos, así como comprender mejor las opiniones en las redes sociales y las conversaciones de los clientes. Esta herramienta te permite analizar textos e integrarlos en tu almacenamiento de documentos de Cloud Storage.

- **IBM Watson IA.** Dentro de la plataforma Watson IA disponemos de diferentes módulos para traducción, speech <-> voice, comprensión de lenguaje natural, clasificación de lenguaje natural, etc…

Podemos también trabajar con APIs para las diferentes áreas de NLP que hemos visto en el capítulo anterior que exponen los principales fabricantes de *software* de IA como son, además de las que se encuentran en los 4 grandes, las siguientes:

- Aylien. Aprovechando el contenido de noticias con NLP, se trata de un API SaaS que utiliza aprendizaje profundo y PNL para analizar grandes volúmenes de datos basados en texto, como publicaciones académicas, contenido en tiempo real de medios de comunicación y datos de redes sociales.

- **NLTK**. La biblioteca (kit de herramientas) de Python más popular. Con una estructura modular, NLTK proporciona muchos componentes para tareas de PNL, como tokenización, etiquetado, derivación, análisis y clasificación, entre otros.

- **MonkeyLearn**. PNL simplificado. Es una plataforma fácil de usar impulsada por PNL que le ayuda a obtener información valiosa a partir de sus datos de texto. Dispone de modelos entrenados previamente para introducirnos en el conocimiento de estas

herramientas.

- Stanford Core PNL. Potente conjunto de herramientas rápido y robusto de la universidad de Stanford, escrito en Java.
- TextBlob. Una interfaz intuitiva que nos acerca a la complejidad inicial de NLTK
- SpaCy. Biblioteca en Python ultrarrápida para tareas avanzadas de PNL. Esta biblioteca es una opción alternativa excelente si desea preparar texto para el aprendizaje profundo y sobresale en las tareas de extracción.
- GenSim. Es una biblioteca de Python altamente especializada (modelado de temas) que se ocupa en gran medida de tareas de modelado de temas utilizando algoritmos como Latent Dirichlet Allocation (LDA).

Por último, aunque no menos importante, son todos los *frameworks* de desarrollo *open source* especializados en PLN y que surgen en ricas comunidades de desarrollo que permiten la realización de proyectos de bajo coste con altísimos niveles de complejidad. La curva de aprendizaje puede ser mayor, pero nos permite un nivel de personalización mucho más rico. Además de NLTK, SpaCy y GenSim podemos considerar principalemente:

- Apache OpenNLP
- AllenNLP.
- Intel NLP Architect
- Flair

Deep learning para el PLN

Como hemos visto, hasta hace poco todas las técnicas que se utilizaban en PLN provienen del *machine learning* clásico apoyándose en modelos lineales como la regresión logística o las SVM (Máquinas de Vector Soporte o *Support Vector*