Tecnologías emergentes y datos abiertos:
PROCESAMIENTO DE LENGUAJE NATURAL

AWARENESS

Contenido elaborado por Alejandro Alija, experto en Transformación Digital y datos abiertos.

Abril 2020

AWARENESS

¿Qué son las Tecnologías del Lenguaje?

Las **tecnologías digitales del lenguaje** son aquellas capacidades, herramientas informáticas y algoritmos que hacen posible que las **máquinas puedan entender y generar expresiones en lenguaje humano** (escrito y hablado) en **múltiples idiomas**. El conjunto de tecnologías digitales del lenguaje ocupa un lugar preferente y de actualidad en los principales *hubs* y espacios de innovación de las empresas e instituciones académicas. También es habitual que el apoyo al desarrollo de estas tecnologías se materialice en forma de planes y programas impulsados por las instituciones estatales, como en el caso de España con el **Plan de Impulso a las Tecnologías del Lenguaje.**

El siguiente cuadro muestra el marco teórico establecido por el Ministerio de Asuntos Económicos y Transformación Digital a través del citado Plan, donde se habla de *Procesamiento del Lenguaje Natural (PLN[1]) y Traducción Automática (TA).*

[1] En el Plan de impulso a las tecnologías del lenguaje, elaborado por el Ministerio de Asuntos Económicos y Transformación Digital, las siglas utilizadas para referirse al Procesamiento del Lenguaje Natural son PLN. Sin embargo, en este informe utilizaremos su forma en inglés NLP (Natural Language Processing). Esto se debe a que el uso de estas siglas en inglés está mucho más extendido. De esta forma, si un lector interesado en el tema, realiza una búsqueda en Google del término PLN, no encontrará ninguna entrada relevante en la primera página de resultados. De lo contrario, cuando buscamos NLP, rápidamente accedemos a la información general sobre Procesamiento del Lenguaje Natural.

Las tecnologías de Procesamiento del Lenguaje Natural (PLN) y Traducción Automática (TA) son las tecnologías que hacen posible analizar textos y facilitar su explotación en aplicaciones informáticas de uso muy común en sectores tan dispares como la Sanidad, la Educación o el Turismo.

Por ejemplo, la detección de entidades nombradas (nombres propios de personas o empresas, marcas de productos o topónimos), filtrado y clasificación de documentos, creación de resúmenes automáticos, extracción de información, análisis de sentimientos, minería de opinión, seguimiento y monitorización de la reputación en los medios sociales, corrección ortográfica y gramatical, búsqueda inteligente y optimizada, sistemas de respuesta automática a preguntas y asistentes personales, la traducción automática de textos, etc.

Todas estas aplicaciones se pueden resumir como la explotación de información no estructurada que mejora la comprensión de textos en corpora documentales.

La visión tan multidisciplinar del conjunto de tecnologías del lenguaje se representa bien en la siguiente figura:

Figura 2. *¿Cuál es el lugar del Procesamiento del Lenguaje Natural como campo de la ciencia y la tecnología? Fuente: https://www.plantl.gob.es/tecnologias-lenguaje/catalogo-TL/Paginas/clasificacion-TL.aspx.*

De acuerdo con este mismo marco teórico, las principales aplicaciones de este conjunto de tecnologías son:

- Optimización de procesos industriales de gestión lingüística de documentación: traducción de documentos y herramientas de autor (correctores, generación de documentos, etc.)
- Comunicación y asistencia personal (asistentes virtuales, comunicación persona-máquina para coches, atención al cliente e interacción con robots; buscadores inteligentes y respuesta automática de preguntas).
- Procesamiento inteligente de información y conocimiento (extracción y minería de información de textos y contenidos, clasificación de documentos, resumen automático, etc.).
- Asistencia en el aprendizaje de lenguas.

El Procesamiento del Lenguaje Natural

En particular, este informe se enfoca fundamentalmente sobre el campo del **Procesamiento del Lenguaje Natural**, que tiene **múltiples aplicaciones en nuestra vida cotidiana**.

De una forma sencilla podemos decir que el **Procesamiento del Lenguaje Natural (NLP)** es hacer que las máquinas (los ordenadores) **entiendan el lenguaje humano**: hablado o en forma de texto. Más formalmente, como hemos introducido anteriormente, el Procesamiento del Lenguaje Natural (NLP) es un campo híbrido entre la informática y la lingüística, que utiliza diferentes técnicas, algunas de ellas basadas en Inteligencia Artificial, para interpretar el lenguaje humano.

Procesamiento de Lenguaje Natural (NLP)

Formalmente, el NLP es un campo interdisciplinar que trata de hacer que las máquinas, mediante programas de software, sean capaces de leer, entender y derivar el significado del lenguaje humano escrito. Las aplicaciones del NLP es la vida cotidiana son múltiples. Algunos ejemplos populares son:

- **Autocompletar y predicción de texto:** en motores de búsqueda (por ejemplo: Google, Bing) y más recientemente en los clientes de correo electrónico.
- **Revisión ortográfica:** en casi todas partes, en el navegador, en el procesamiento de textos (por ejemplo: Microsoft Office u Open Office) en las aplicaciones de mensajería instantánea.
- **Análisis de revisiones y comentarios**: Analizar automáticamente las opiniones (webs de recomendación sobre productos, restaurante, viajes, etc.) de los clientes es una de las mayores aplicaciones de NLP.

Existe bastante ambigüedad en la bibliografía existente sobre la definición y el uso del término NLP (Procesamiento del Lenguaje Natural, en español).

- Multitud de referencias[2] utilizan NLP para referirse, de forma general, al conjunto de tecnologías que hacen posible que seamos capaces de comunicarnos con una máquina independientemente del idioma y el canal que utilicemos para ello. Por lo tanto, este uso del término NLP agrega todas aquellas tecnologías de traducción automática (TA), asistentes conversacionales y sistemas de conversión de lenguaje hablado a texto y viceversa.
- En otras ocasiones, sin embargo, encontramos usos más concretos del término NLP, refiriéndose estrictamente a aquellas tareas que tienen que ver con el análisis de las oraciones escritas para su simplificación y posterior clasificación.

Sea como fuere, el propósito de este informe es introducir al lector en el campo de las tecnologías digitales que, de una forma u otra, sin entrar en demasiados tecnicismos, son responsables de que las máquinas sean capaces de entender el lenguaje humano para multitud de aplicaciones.

Precisamente, uno de los objetivos de este informe es plantear de forma clara y cercana los múltiples casos de uso que se apoyan en este conjunto de tecnologías. En este sentido, a lo largo del documento se utiliza el concepto de NLP desde una perspectiva amplia para atribuir a un solo término las múltiples aplicaciones de este grupo de tecnologías.

[2] Otros ejemplos:

- https://www.datacentric.es/blog/business/procesamiento-lenguaje-natural-revolucion-futuro/
- https://www.sas.com/es_ar/insights/analytics/what-is-natural-language-processing-nlp.html

La importancia de los datos abiertos en el Procesamiento del Lenguaje Natural

Los **algoritmos de NLP son grandes consumidores de datos** que sirven de entrada para el **entrenamiento de los modelos de Inteligencia Artificial** que hacen posible el entendimiento del lenguaje humano por parte de las máquinas. Además, la particularidad del procesamiento de lenguaje natural frente a otros campos de la ciencia de datos es su **variabilidad idiomática**. Esto es, los modelos han de ser entrenados con conjuntos de datos en cada idioma para obtener resultados óptimos. Veamos un ejemplo para hacernos una idea de la magnitud de los datos necesarios.

Uno de los últimos algoritmos de NLP publicados en 2019, GPT-2, ha sido entrenado con **40GB de textos disponibles en Internet**. Por comparación, una copia de *El Quijote* de Miguel de Cervantes en formato *pdf* ocupa aproximadamente un 1MB de espacio en disco. De forma ilustrativa, el algoritmo GPT-2 ha sido entrenado con 40.000 obras del tamaño de *El Quijote.* Es evidente, que tal cantidad de texto escrito, **necesita necesariamente del uso de datos abiertos en forma de textos**. Algunos repositorios de datos abiertos están especialmente preparados para albergar textos que sirvan como recursos lingüísticos de calidad para entrenar algoritmos de NLP. En el reciente informe *Estudio sobre datos reutilizables como recursos lingüísticos* (2019) se describe una relación de recursos lingüísticos por temática y organización de origen.

En el momento de escribir este informe, el mundo entero se enfrenta a una de las mayores pandemias de la era moderna. **La crisis del Covid-19** eclipsa cualquier otra noticia de interés. Incluso en esta grave situación de emergencia sanitaria mundial, las **tecnologías de NLP junto con los datos abiertos juegan un papel fundamental** para ayudar a la sociedad en la lucha contra el virus. Así, La Casa Blanca junto con una coalición de grupos de investigación líderes han preparado el conjunto de datos abiertos sobre la investigación del COVID-19 (CORD-19). El conjunto de datos CORD-19 es un recurso de más de 44,000 artículos académicos, incluidos más de 29,000 con texto completo, sobre COVID-19, SARS-CoV-2 y coronavirus relacionados. Este conjunto de **datos de libre acceso** se proporciona a la comunidad de investigación global para aplicar los avances recientes en el **Procesamiento del Lenguaje Natural** y otras técnicas de IA para generar nuevas ideas en apoyo de la lucha continua contra esta enfermedad infecciosa.

¿Cómo hacemos que las máquinas entiendan el lenguaje humano?

El ser humano ha sido muy hábil con el desarrollo de los ordenadores y la ciencia de la computación moderna. Un ordenador convencional basado en tecnología del silicio es una máquina, que, a pesar de su complejidad, se basa en el simple principio de **codificar y decodificar información digital binaria** basada en ceros y unos. Por lo tanto, parece lógico pensar que, para hacer que una máquina *entienda* nuestro lenguaje, debemos de convertir el texto en códigos binarios. Esto se conoce como **codificación de texto o *text encoding***.

Es importante destacar que la máquina, y en particular, los modelos (algoritmos) de Procesamiento del Lenguaje Natural no *entienden,* en sentido estrictamente humano, el significado de nuestro lenguaje. En realidad, estos modelos lo que hacen es **mapear la**

estructura estadística del lenguaje escrito. Habitualmente esta fórmula es suficiente para resolver muchas tareas textuales simples como las que hemos citado en párrafos anteriores.

Veamos un ejemplo sencillo para entender cómo funciona el *text enconding.* Por simplicidad vamos a analizar una frase **a nivel de sus palabras**. Podríamos utilizar caracteres individuales u otras estructuras como conjuntos de varias palabras o expresiones. En NLP, a los conjuntos de palabras que forman una expresión se les conoce como N-gram donde N representa el número de (en este caso) palabras que tiene la expresión.

Utilizando esta aproximación del análisis de las palabras que forman una oración, tomemos la siguiente frase: "El gato se sentó sobre el libro." Veamos qué pasos son necesarios para ejecutar el proceso del *text encoding*.

Lo primero que hacemos es **asignar un índice** a cada palabra de la siguiente forma:

El	gato	se	sentó	sobre	el	libro.
1	2	3	4	5	6	7

Una vez hecho esto **generamos un array** (en este caso una lista de vectores) **que representa el índice de cada palabra y su posición de ocurrencia**. En este caso la única palabra que se repite es "el".

El resultado del *encoding* es el siguiente en forma de lista de 7 vectores (array).

```
, , 1

   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    0    0    0    0    0    0    0    0     0

, , 2

   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    1    0    0    0    0    0    0    0     0

, , 3
```

```
   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    0    1    0    0    0    0    0    0     0

, , 4

   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    0    0    1    0    0    0    0    0     0

, , 5

   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    0    0    0    1    0    0    0    0     0

, , 6

   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    0    0    0    0    1    0    0    0     0

, , 7

   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    0    0    0    0    0    1    0    0     0
```

La palabra “el” aparece en la posición 1 y 6 de la oración (que se representan en la figura anterior como ,,*1* y ,,6). Se han generado dos vectores distintos, aunque se trata de la misma palabra. Así, la forma de representar esta ocurrencia de palabras en el array de resultados es como se muestra en la siguiente figura.

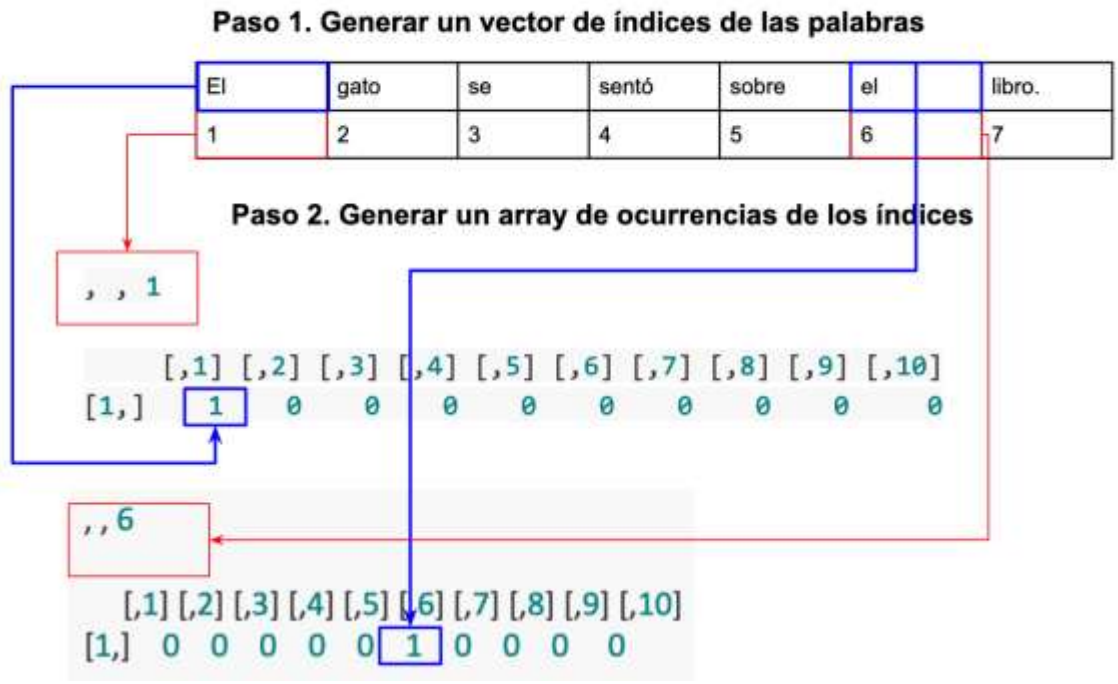

Figura 3. *Proceso de word-encoding explicado de forma gráfica.*

La técnica que hemos visto aquí es la más sencilla para codificar texto en representaciones numéricas. Esta técnica se denomina ***one-hot-encoding*** y a pesar de ser ampliamente utilizada presenta las siguientes **limitaciones**:

- Se **utilizan grandes cantidades de memoria** para almacenar información en su mayoría nula o inservible, puesto que la mayoría de las posiciones del array multidimensional que hemos visto anteriormente son cero.
- Se **pierde la información relativa a la similitud de las palabras,** lo que va en detrimento de la capacidad de entender su significado.

Alternativamente a este método, podríamos dar un paso más y generar un vector denso, asignando a cada palabra en la frase un índice arbitrario único. Por ejemplo, ordenaríamos alfabéticamente el vocabulario de la frase: *El gato libro se sentó sobre.* Asignamos un índice numérico a cada palabra del vocabulario ordenado. Finalmente, construimos la oración colocando el índice en la posición de ocurrencia [1,2,4,5,6,1,3]

La alternativa al método *one-hot-encoding* es el mecanismo conocido como ***word embeddings***. Mientras que los vectores obtenidos a través de la codificación ***one-hot-encoding*** son **binarios**, **dispersos** (la mayoría de los valores son cero) y de muy **alta dimensión** (como vimos en el ejemplo, para una sencilla frase se generaban 7 vectores, uno por palabra), el resultado del ***word embeddings*** son vectores de más **baja dimensión** (vectores densos), a diferencia de los vectores dispersos.

Para grandes vocabularios (por ejemplo, 20.000 palabras), el método *one-hot-encoding* generaría arrays de dimensión 20.000 mientras que el método *word embeddings* generaría arrays de 256 o 512 dimensiones. Por lo tanto, una de las principales ventajas de utilizar *word embeddings* es su capacidad **para comprimir la misma información en muchas menos dimensiones**.

Finalmente, otra gran ventaja del método *word embeddings* es que los vectores densos se aprenden (se generan) de los datos de entrada, mientras que en el *one-hot-encoding* la asignación del índice es arbitraria. Es decir, es posible generar un ***espacio* de *word embeddings* determinado para un conjunto de textos de entrada, un idioma determinado y un objetivo concreto**. Es decir, tienen en cuenta las relaciones entre las palabras. Por ejemplo, París, Grenoble y Francia, tienen similitud en el contexto de países y ciudades, y por lo tanto los números que representan estas palabras serán similares entre sí. Otro ejemplo, un espacio *word embeddings* concreto es aquel que se ha generado a partir de una base de datos de críticas de películas de cine en inglés. El objetivo de este espacio es servir como base para analizar qué películas han gustado más y cuáles menos. Una vez generado este espacio, puede ser utilizado por cualquier aplicación similar a la anterior. De esta forma, sería como tener un modelo de Inteligencia Artificial pre-entrenado al que solo le tenemos que suministrar los nuevos datos de entrada.

En resumen, las diferencias entre el *one-hot- encoding y el word embeddings* son:

ONE-HOT-ENCODING	WORD EMBEDDINGS
Binario: vector formado por 0 y 1.	**Continuo:** vector formado por números reales.
Disperso: la mayoría de los valores son cero.	**Denso:** los valores son números reales.
Alta dimensión: se generan un gran número de vectores.	**Baja dimensión:** permite comprimir la misma información en menos vectores.
Codificación impuesta: los índices se establecen manualmente de manera arbitaria.	**Codificación aprendida:** los valores de los vectores se aprenden de los datos.

Figura 4. *Comparativa entre el método one-hot-encoding y el word embeddings.*

La complejidad técnica que subyace debajo del Procesamiento del Lenguaje Natural hace que sea imposible cubrir con mayor nivel de detalle los procesos de generación de espacios de *word embeddings.* Si bien es cierto que la introducción incluida en párrafos anteriores será muy valiosa para seguir de forma fluida el ejemplo completo de la sección *Action*, incorporar más complejidad técnica a esta informe queda fuera de su alcance. Sin embargo, el lector más atrevido puede consultar la sección *Próxima parada* donde encontrará multitud de enlaces a referencias que extienden con creces el contenido técnico de este informe.

Un poco de historia

La historia del Procesamiento del Lenguaje Natural abarca el periodo que va desde el fin de la Segunda Guerra Mundial hasta nuestros días. Por tanto, cuenta ya **con 75 años de largo y arduo** recorrido.

Alan Turing, conocido como uno de los padres de la Inteligencia Artificial y de los antepasados de los ordenadores, publicó en 1950 un artículo titulado "Computing Machinery and Intelligence", que puede considerarse el texto que inaugura la historia del NLP. Merece la pena citar el comienzo del artículo:

> *I PROPOSE to consider the question: "'Can machines think?"*
>
> *This should begin with definitions of the meaning of the terms 'machine' and 'think'.*
>
> ***TRADUCCIÓN:*** Propongo considerar la pregunta: "¿Pueden las máquinas pensar?" Esto debería comenzar con definiciones del significado de los términos "máquina" y "pensar".

No deja de ser sorprendente que alguien comenzara así un artículo sobre la inteligencia de las máquinas hace 70 años, teniendo en cuenta que lo más parecido a un ordenador era un engendro mecánico del tamaño de una habitación.

Figura 5. *De Karl Baron from Lund, Sweden - Vacuum tube computer: Uploaded by shoulder-synth, CC BY 2.0,* https://commons.wikimedia.org/w/index.php?curid=8828095

Hasta la década de 1980, la mayoría de los sistemas de procesamiento de lenguaje natural se basaban en conjuntos complejos de reglas pre-definidas, coincidiendo a su vez con el auge de los denominados sistemas expertos[3].

A partir de finales de los **años 80**, se produce la primera revolución en el campo del Procesamiento del Lenguaje Natural. Gracias al **aumento en la capacidad de cálculo** de los ordenadores siguiendo la Ley de Moore, comienzan a introducirse **estrategias basadas en la**

[3] Los sistemas expertos son programas informáticos que contienen reglas lógicas que codifican y parametrizan el funcionamiento de sistemas sencillos. Por ejemplo, un programa informático que codifica las reglas del juego de ajedrez pertenece al tipo de programas que conocemos como sistemas experto.

estadística avanzada (primeros algoritmos de machine learning) para el procesamiento del lenguaje. Algunos de estos antiguos algoritmos de machine learning, como los árboles de decisión, producían sistemas de reglas estrictas similares a las diseñadas manualmente en la década anterior. Con la progresiva **democratización de los ordenadores personales**, se generaron **más y más datos digitales** de entrada para entrenar a estos algoritmos, mejorando de forma continua su precisión en tareas como la clasificación de textos, dando como resultado los filtros anti-spam, por ejemplo.

El siguiente hito importante en el campo del procesamiento del lenguaje se produce en el año **2013**, cuando el grupo de investigación de Google dirigido por Tomas Mikolov inventan el **algoritmo Word2vec**. A partir de la existencia de algoritmos como word2vec y otros posteriores como Glove o FastText que pueden ser pre-entrenados con grandes volúmenes de datos, el campo del NLP sufre una gran democratización, permitiendo a los desarrolladores de software crear aplicaciones que utilizan el Procesamiento del Lenguaje Natural como entrada o salida de la funcionalidad de dicha aplicación.

En resumen, la historia del Procesamiento del Lenguaje Natural es muy amplia, como resume la siguiente imagen:

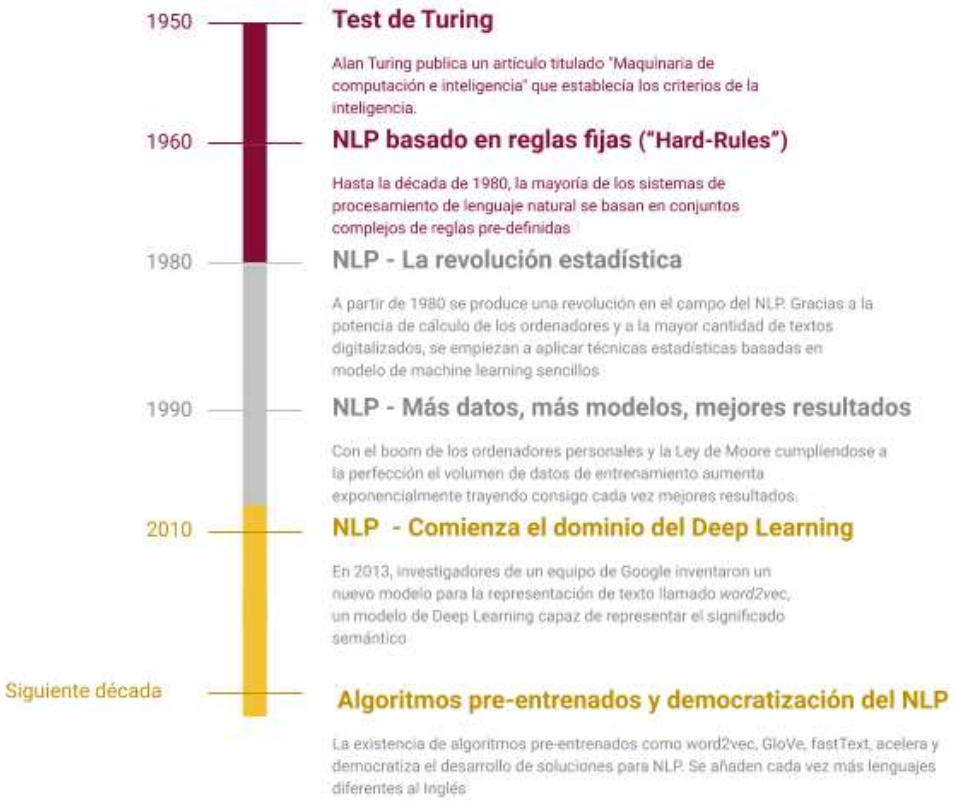

Figura 6. *Línea temporal que destaca los hitos más importantes en el desarrollo del NLP desde sus inicios hasta nuestros días.*

Impacto

Desde un punto de vista de impacto tangible, el NLP ha realizado grandes avances en los últimos años, impulsado por la ola de digitalización que vivimos y navegando por el infinito océano de Internet. A continuación, revisamos los principales hitos del NLP divididos por su grado de madurez:

Tareas casi resueltas completamente por NLP

Detección de spam. ¿Desde cuándo no revisas el Spam en tu cuenta de correo electrónico? No hace muchos años que había que comprobar varias veces al día que un correo electrónico importante no había terminado en la bandeja del Spam de tu cliente o servidor de correo. En la actualidad, la probabilidad de que un algoritmo de detección de spam produzca un falso positivo es realmente baja.

Detección de partes de las oraciones (POS, del inglés *Part of Speech*). Son algoritmos que, dada una oración, tratan de determinar qué tipo de palabras se encuentran en ella, por ejemplo, cuáles son nombres, verbos, adjetivos, etc.

Detección y reconocimiento de entidades (NER, del inglés *Name Entity Recognition*). Son algoritmos que, dada una oración, tratan de determinar a qué tipo de entidades corresponden los nombres que se encuentran. Por ejemplo, nombres de personas, ubicaciones, organizaciones.

Tareas que demuestran un rápido y satisfactorio avance en NLP

Análisis de sentimientos. Dada una oración, el algoritmo trata de determinar su polaridad (por ejemplo: positiva, negativa, neutral) o emoción (por ejemplo: feliz, triste, sorprendida, enojada). Esta tarea tiene una gran importancia en el análisis de opinión que, lógicamente, es de crucial importancia para empresas de productos, servicios, medios de comunicación, etc.

Detección de referencias cruzadas. Para hacer que una máquina entienda el lenguaje humano es necesario detectar qué palabras hacen referencias unas a otras. Por ejemplo, en una oración en la que hay un nombre propio y más adelante se usa un pronombre para referirse al nombre anterior, es necesario detectar que ambos, nombre propio y pronombre, hacen referencia a lo mismo dentro del significado de la oración.

Desambiguación del sentido de las palabras (WSD, del inglés *Word Sensing Desambiguation*). En el lenguaje humano, muchas palabras tienen más de un significado. Para entender el significado de una oración en particular, es necesario seleccionar el significado que más sentido tenga en el contexto de dicha oración.

Tareas de NLP cuyo grado de madurez es todavía limitado

Asistentes de diálogo y chat-bots. Aunque su evolución ha sido notoria en los últimos años, todavía son tecnologías de uso muy restringido y limitado a dominios muy específicos (medicina, asistente de call-center, acciones rutinarias con el smartphone, etc.).

Asistentes de pregunta-respuesta. Su capacidad de entender el sentido de la pregunta especialmente en lenguaje hablado es bajo y las acciones de call-back (es decir, aquellas acciones que el asistente tiene que realizar cuando no encuentra lo esperado por la persona usuaria o no ha entendido el sentido de la pregunta) son muy rudimentarias.

Generación de resúmenes. La generación de resúmenes de texto pertenece a un subdominio del NLP conocido como generación de lenguaje natural (NLG). El grado de desarrollo del NLG es bajo fuera de dominios específicos y entornos con condiciones muy controladas (como, por ejemplo, resúmenes de eventos deportivos basados en estadísticas, o resúmenes meteorológicos).

NLP para idiomas de bajos recursos. Se estima que en el mundo hay unos 7.000 idiomas hablados y, sin embargo, la mayoría de estos idiomas son residuales e incapaces de generar suficiente material escrito para poder entrenar los algoritmos de procesamiento. Las últimas investigaciones en NLP ponen el foco en estos idiomas con nuevas técnicas que permiten mitigar el efecto de disponer de recursos escasos para el procesamiento.

Ahora que ya conocemos los principios básicos del Procesamiento del Lenguaje Natural (incluso nos hemos asomado ligeramente a sus bases técnicas), su historia y sus principales aplicaciones, es el momento de activar nuestra inspiración visitando juntos algunos casos de uso de mucha actualidad que se sustentan sobre las bases del NLP.