Lingüística computacional

En breve

ENG *Computational linguistics* **CAT** *Lingüística computacional* **GAL** *Lingüística computacional* **POR** *Lingüística computacional*

otros nombres

El término *procesamiento del lenguaje natural* (PLN) se usa frecuentemente como sinónimo del término *lingüística computacional.* Sin embargo, históricamente su uso ha sido distinto, como explicamos en la introducción de esta entrada.

resumen

La lingüística computacional (LC) es un campo interdisciplinar en el que se encuentran la inteligencia artificial, la lingüística, la ciencia cognitiva y la informática, dedicado al estudio del lenguaje natural desde una perspectiva computacional. Los principales objetivos teóricos y prácticos de la LC son la comprensión del lenguaje natural y el consiguiente desarrollo de herramientas y sistemas que puedan procesar y generar lenguaje natural, respectivamente. Esta entrada se centra en la LC aplicada a la modalidad escrita de los lenguajes naturales y cubre los siguientes aspectos:

(i) los métodos más comunes de la LC aplicados a los diferentes niveles de análisis del lenguaje natural (morfología, sintaxis, semántica y pragmática);

(ii) una reseña histórica de la LC a través de los principales enfoques computacionales utilizados en este campo y el impacto de dichos enfoques en la relación entre la LC y otras áreas de la lingüística;

(iii) los diferentes tipos de recursos lingüísticos utilizados en la LC y su papel;

(iv) una visión general de las aplicaciones más destacadas de la LC;

(v) un resumen de las tendencias actuales de la CL y sus áreas de aplicación.

ficha

Antonio Toral Ruiz & Tommaso Caselli

2022

Toral Ruiz, Antonio & Tommaso Caselli. 2022. "Lingüística computacional" @ *ENTI* (*Enciclopedia de traducción e interpretación*).

https://doi.org/10.5281/zenodo.6366264

https://www.aieti.eu/enti/computational_linguistics_SPA/

Entrada

ENG *Computacional linguistics* **CAT** *Língüística computacional* **GAL** *Lingüística computacional* **POR** *Lingüística computacional*

contenido

Introducción

La lingüística computacional (LC) es un campo interdisciplinar en el que se encuentran la inteligencia artificial, la lingüística, la ciencia cognitiva y la informática, dedicado al estudio del lenguaje natural desde una perspectiva computacional. Los principales objetivos teóricos y prácticos de la LC son la comprensión del lenguaje natural y el consiguiente desarrollo de herramientas y sistemas que puedan procesar y generar lenguaje natural, respectivamente.

Históricamente, el término LC suele utilizarse para referirse a los objetivos teóricos de la disciplina, es decir, el desarrollo de modelos computacionales del lenguaje natural y la comprensión de cómo el ser humano procesa el significado. A su vez, el término procesamiento del lenguaje natural (PLN) suele utilizarse para referirse a los aspectos de ingeniería, es decir, al desarrollo de programas informáticos que procesan el lenguaje natural. Sin embargo, a menudo ambos términos se utilizan como sinónimos en contextos intercambiables.

Figura 1. *Desarrollando sistemas para generar lenguaje natural.*

Esta entrada se centra en la LC aplicada a la modalidad escrita del lenguaje y está estructurada como se detalla a continuación. En primer lugar, tratamos los métodos más comunes de la LC aplicados a diferentes niveles de análisis del lenguaje natural. A continuación, describimos los principales enfoques computacionales utilizados en la LC. Posteriormente, introducimos los recursos lingüísticos (RL) y explicamos el papel que desempeñan en esta disciplina. También ofrecemos una visión general de las aplicaciones más destacadas de la LC. Por último, concluimos con un resumen de las tendencias actuales de la LC y sus ámbitos de aplicación.

cabecera

¶ Niveles de análisis

Las lenguas naturales son sistemas complejos. De hecho, el estudio científico del lenguaje (lingüística) ha identificado diferentes subcampos que incluyen el estudio de la forma (fonología y morfología), las estructuras (sintaxis), el significado (semántica) y el uso (pragmática). El resto de esta sección ilustra cómo la LC se relaciona con cada uno de estos subcampos (excepto la fonología), por medio de diferentes niveles de análisis, o tareas. La fonología es relevante para la modalidad oral, por lo que se excluye, ya que esta entrada se centra en la modalidad escrita. Además, debido a las limitaciones de espacio, sólo tratamos las principales tareas o niveles de análisis. Para una descripción más exhaustiva de las tareas de la LC, se remite al lector interesado a Jurafsky y Martin (2020) o a Eisenstein (2019).

La primera tarea consiste en la identificación de las cadenas que componen un *token*. Un token puede entenderse generalmente como una cadena que corresponde a una palabra (o a un símbolo de puntuación) y se considera la unidad básica de análisis en la LC. Esta capa de análisis se conoce como tokenización y la manera en que se lleva a cabo depende de la lengua. Por ejemplo, para las lenguas indoeuropeas la tokenización corresponde a la introducción de espacios entre las palabras y los símbolos de puntuación, mientras que para las lenguas sinotibetanas, en las que no hay una separación explícita (espacios en blanco) entre las palabras, la tokenización implica identificar los límites de las palabras.

Los *tokens* deben analizarse y clasificarse en las diferentes categorías gramaticales, por ejemplo, nombres, verbos, adjetivos y artículos, entre otros. Esta tarea se conoce como etiquetado gramatical (*part-of-speech tagging* o, abreviado, *PoS tagging* en inglés). El objetivo es asignar a cada *token*, o combinación de *tokens* (por ejemplo, en el caso de expresiones multipalabra), una etiqueta que contenga información morfosintáctica básica que distinga si el *token* representa, por ejemplo, un sustantivo o un verbo, si es, por ejemplo, singular o plural, si expresa un tiempo finito, entre otras características. A lo largo de los años se han desarrollado diferentes repositorios de etiquetas gramaticales específicos para cada idioma. Una propuesta que es independiente de la lengua es el conjunto de etiquetas UD.

Otra tarea común relacionada con la morfología es la lematización. El objetivo consiste en asignar a cada *ocurrencia* (*token)* su lema, es decir, su forma canónica. En castellano, por ejemplo, las palabras *escribió, escribimos* y *escribirán* tienen el mismo lema, *escribir.* Una de las motivaciones de esta tarea es reducir la dispersión debida a la presencia de diferentes formas flexionadas del mismo lema, especialmente durante el entrenamiento (véase Enfoques computacionales para más detalles). Una tarea relacionada con esto es el análisis morfológico, cuyo objetivo es la identificación y explicitación de los morfemas que componen cada token y de sus características morfológicas, por ejemplo, género, número, caso, tiempo. El análisis morfológico es especialmente útil cuando se trabaja con lenguas con morfología productiva (por ejemplo, las lenguas túrquicas y finoúgricas).

La identificación de cómo las palabras pueden combinarse para formar estructuras gramaticales se aborda mediante el análisis sintáctico. De acuerdo con los marcos teóricos y los formalismos propios de la lingüística, hay dos formas habituales de abordar esta tarea. La primera, el análisis

sintáctico de constituyentes, se inspira en el paradigma de Chomsky de las gramáticas sin contexto y se basa en la identificación de constituyentes (por ejemplo, frases sustantivas y verbales) y en reglas para su combinación. El segundo, el análisis sintáctico de dependencias, inspirado en la obra del lingüista francés Tesnière, determina la estructura sintáctica mediante relaciones asimétricas, llamadas dependencias, entre los *tokens*. En ambos formalismos, cada relación sintáctica se da entre un par de *tokens*. En la LC, el análisis sintáctico de dependencias es actualmente el paradigma más utilizado para representar la estructura sintáctica, ya que ha resultado tener algunas ventajas (Jurafsky & Martin 2020 - capítulo 15) entre las que destacan el hecho de que puede aplicarse fácilmente a diferentes familias de lenguas, su facilidad de anotación, y la facilidad de procesamiento (en términos del formato de representación). La figura 2 ilustra gráficamente las diferencias entre estos dos paradigmas de representación sintáctica para una misma frase inglesa: *"John hit the ball" ("Juan golpeó la pelota").*

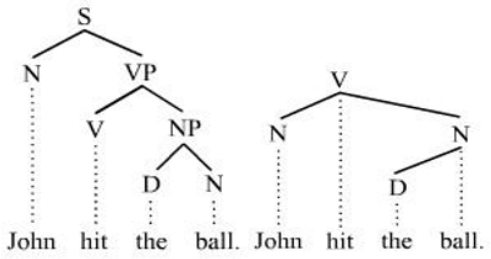

Figura 2. *Ejemplos de análisis sintáctico de constituyentes (izquierda) y de dependencias (derecha) para una frase. (Fuente: Wikipedia CC-BY-SA 3.0). S significa oración, NP frase nominal, VP frase verbal, D determinante, N nombre y V verbo.*

Mientras que la sintaxis trata el estudio de las relaciones *formales* entre palabras (o constituyentes), la semántica aborda la representación de lo que una palabra denota en el mundo, así como la manera en que diferentes palabras se combinan entre sí para dar lugar a expresiones con significado. El estudio de la semántica está intrínsecamente ligado al problema de las representaciones del significado, es decir, a la relación entre signos y significantes, referencia y denotación. En la LC hay diferentes tareas que pertenecen a esta área de la lingüística. En esta entrada describiremos brevemente dos de ellas: la desambiguación del sentido de las palabras (comúnmente abreviada como WSD, del inglés *Word Sense Disambiguation*) y el etiquetado de roles semánticos (comúnmente abreviado como SRL, del inglés Semantic Role Labelling).

La WSD es la tarea consistente en identificar qué sentido tiene una palabra en su contexto y se utiliza habitualmente para evaluar modelos de comprensión del lenguaje natural. Tradicionalmente, la investigación en WSD se centraba en el uso de recursos léxicos computacionales elaborados manualmente (por ejemplo, WordNet) que enumeran los diferentes sentidos de cada palabra, así como sus posibles relaciones semánticas (por ejemplo, sinonimia, hiperonimia/hiponimia). Más recientemente, se ha aplicado con éxito una nueva metodología inspirada en la hipótesis distribucional (Firth 1957), denominada semántica distribucional (Baroni & Lenci 2010). En este caso, los significados y las relaciones entre las palabras se infieren a partir de grandes cantidades de datos textuales procedentes de corpora. Desarrollos posteriores del enfoque semántico distribucional son los llamados *word embeddings* (Mikolov, Sutskever, Chen *et al*. 2013), un conjunto de técnicas de aprendizaje que asignan los significados de las palabras a un espacio vectorial de números reales.

El SRL contribuye a formalizar las llamadas *representaciones semánticas superficiales* a nivel de oración. En concreto, el SRL hace explícito cuál es el rol semántico de los diferentes argumentos asociados a un predicado (ya sea un verbo o un nombre). Por ejemplo, en la frase *«María se comió una manzana»*, se entiende que *María* es el **agente** de la acción, mientras que *la manzana* es el

tema de la acción, es decir, el argumento que se ve afectado. Los sistemas de SRL asignan los roles a partir de una lista codificada en recursos específicos. Dos de los recursos más utilizados para el inglés son PropBank y FrameNet, este último está también disponible para una lengua ibérica: el castellano. Los roles semánticos también se denominan roles temáticos (Fillmore 1968).

El estudio del significado de fragmentos de texto más amplios, más allá de palabras sueltas y oraciones aisladas, ha dado lugar a una rica gama de enfoques y teorías en lingüística y filosofía, como la semántica formal, la semántica condicional de la verdad y la semántica cognitiva. La LC se centra en el desarrollo de procedimientos automáticos, o algoritmos, para construir representaciones del significado de las expresiones del lenguaje natural, que posteriormente pueden utilizarse para llevar a cabo un razonamiento automático. Algunos de los métodos utilizados para desarrollar estos enfoques automáticos se basan en las teorías anteriores, mientras que otros siguen enfoques basados en datos. Los enfoques tradicionales, vinculados a teorías composicionales del significado, conciben la oración como la unidad básica del significado, y el discurso como la intersección entre los valores semánticos de las oraciones constituyentes. Esta visión ha sido cuestionada por las teorías dinámicas de la semántica (por ejemplo, DRT, SDRT, lógica dinámica de predicados) que promueven una visión en la cual las oraciones son elementos dinámicos que reflejan relaciones de un contexto del discurso a otro. La noción de contexto de discurso varía y puede ser una estructura de representación (por ejemplo, DRT), un conjunto de funciones de asignación, o un conjunto de pares de asignación de modelos. Aunque el estudio de la semántica se remonta a Aristóteles, la semántica del discurso y el modelado computacional del discurso representan desarrollos bastante recientes tanto en el área de la lingüística (Mann & Thomson 1988) como en la de la LC (Grosz & Sidner 1986; Webber 1988; Gardent & Webber 1988).

Para concluir esta sección, mencionamos una tarea que conecta los subcampos de la semántica y la pragmática: la resolución de la anáfora. En lingüística, la referencia en un texto a una entidad que ha sido previamente introducida en el discurso se denomina anáfora, mientras que la expresión referente se llama antecedente. Por ejemplo, en la frase *«A María le encanta el helado, pero Juan lo odia»*, el pronombre *«lo»* es una anáfora del nombre *«helado»*, el cual se denomina antecedente. Los enfoques de la LC para la resolución de anáforas tienen como objetivo identificar automáticamente las menciones de anáforas y enlazarlas con su antecedente. Para ello es necesario que el sistema sea «consciente» de la estructura del discurso y de los factores pragmáticos que pueden influir en la resolución de la tarea.

En general, en la LC, las tareas descritas previamente se han considerado tradicionalmente como si siguieran una estructura jerárquica en la que la información fluye de una tarea a otra, como en los dos ejemplos contemporáneos siguientes, que siguen flujos secuenciales:

- UD Pipe (Straka & Straková 2017), en el que el texto de entrada se tokeniza en primer lugar (nivel superficial), posteriormente se etiqueta con PoS y se lematiza (nivel morfológico) y, por último, se analizan las dependencias (nivel sintáctico).
- The Parallel Meaning Bank (Abzianidze, Bjerva, Evang *et al*. 2017): dado un texto de entrada, se ejecuta el siguiente flujo secuencial: tokenización (nivel superficial), análisis de constituyentes (nivel sintáctico), seguido de etiquetado semántico (nivel semántico), simbolización, etiquetado de roles semánticos (nivel semántico), desambiguación del sentido

de las palabras (nivel semántico), resolución de anáforas (nivel semántico/pragmático) y representación del discurso (nivel semántico).

Un ejemplo de cómo estas tareas, vinculadas a los diferentes niveles de análisis lingüístico, desempeñan un papel en una aplicación de la LC es en la traducción automática (TA), en la cual la traducción puede realizarse, por ejemplo, (i) a nivel superficial, (ii) utilizando información superficial y morfológica o (iii) información superficial, morfológica y sintáctica.

cabecera

¶ Enfoques computacionales

En general, podemos identificar tres grandes etapas de enfoques computacionales que han conformado el desarrollo de herramientas para la comprensión, generación y traducción del lenguaje natural. La evolución de cada una de estas etapas ha estado muy influida por los avances tecnológicos y la potencia de cálculo del *hardware* disponible en su momento. Dado que estas oleadas se produjeron de manera secuencial, a través de su panorámica emerge de manera natural un relato histórico de la disciplina, que presentamos a continuación.

1. **Basado en reglas**, también denominado basado en el conocimiento. En este enfoque, especialistas en aspectos lingüísticos escriben reglas manualmente. Los enfoques varían desde simples reglas de correspondencia de patrones hasta conocimiento basado en marcos teóricos de la lingüística formal (por ejemplo, gramáticas generativas, DRT).
2. **Estadístico**. Se trata de un enfoque basado en datos, en el que un sistema computacional induce el conocimiento a partir de datos (por ejemplo, textos). Se atribuye el origen de esta metodología al campo del procesamiento del habla y posteriormente fue aplicada a texto escrito. Las siguientes citas son un ejemplo del estado de ánimo en el momento de su introducción en la LC: « *Anytime a linguist leaves the group the [speech] recognition rate goes up»* (Jelinek 1988) y *«there is no data like more data»* (comentario inédito de Mercier en 1985, citado en Jelinek 2004).
3. **Neuronal**. Este enfoque utiliza redes neuronales artificiales. Al igual que el enfoque estadístico, también se basa en datos. Sin embargo, hay una diferencia clave entre ambos enfoques: el tipo de representaciones que se suelen utilizar (Jurafsky & Manning 2020 - Capítulo 6), discretas (dispersas) en el enfoque estadístico y continuas (densas) en el neuronal.

Cabe señalar que la distinción entre estos enfoques no es tan clara en la práctica, ya que existen interacciones y combinaciones entre ellos, lo que da lugar a lo que puede denominarse sistemas híbridos. El término híbrido se utiliza normalmente en este contexto para referirse a una combinación de métodos basados en reglas y en datos (véase Klavans & Resnik 1996 para una visión general y un debate sobre sus problemas y soluciones). Una tendencia reciente son los sistemas híbridos que combinan redes neuronales y métodos basados en el conocimiento (Gómez-Pérez, Denaux & García-Silva 2020).

Dentro de los enfoques basados en datos, existe otra distinción adicional en función de si los datos utilizados para entrenar un modelo (es decir, los datos de entrenamiento) están etiquetados o no. En el primer caso, los sistemas se denominan

Figura 3. *Las redes neuronales artificiales están revolucionando el procesamiento del lenguaje natural.*

supervisados, mientras que en el segundo se llaman **no supervisados**. Un ejemplo de etiqueta podría ser la categoría gramatical (en inglés part-of-speech o, abreviado, PoS) de cada palabra; un sistema supervisado para predecir automáticamente la categoría gramatical de cada palabra (comúnmente denominado PoS tagger) se entrenaría con datos en los que cada palabra ha sido anotada manualmente con su categoría gramatical. Por otro lado, los sistemas no supervisados se basan en grandes cantidades de datos no etiquetados. Una técnica común para el aprendizaje no supervisado es el agrupamiento, es decir, la agregación de puntos de datos que son similares según algún criterio. Por ejemplo, el etiquetado de PoS se ha abordado en el paradigma no supervisado, reformulándose como Inducción de PoS (Biemann 2011; Stratos, Collins & Hsu 2016; Cárdenas, Lin, Ji *et al*. 2019). El paradigma no supervisado se ha aplicado también en otras áreas de la LC como la inducción del sentido de las palabras (Navigli 2009) y la traducción automática (Artetxe, Labaka, Agirre *et al*. 2018).

A menudo se da el caso de que se dispone de pequeñas cantidades de datos etiquetados para una tarea determinada (es decir, datos que podrían utilizarse para entrenar un sistema supervisado), mientras que hay grandes cantidades de datos sin etiquetar (es decir, datos que podrían utilizarse para construir un sistema no supervisado). Para aprovechar al máximo ambos tipos de datos, se pueden utilizar enfoques **semisupervisados**.

En la comunidad de la LC se percibe que la amplia adopción de métodos basados en datos, junto con el uso de técnicas de aprendizaje automático, han contribuido a cambiar profundamente el campo. En primer lugar, los modelos basados en datos han demostrado ser más eficientes que los sistemas basados en reglas en cuanto al tiempo requerido para su desarrollo, y han obtenido resultados competitivos o incluso mejores. En segundo lugar, el uso de enfoques supervisados basados en datos ha impulsado el desarrollo de datos anotados con sus correspondientes esquemas de anotación, la mayoría de los cuales se basan en teorías o marcos lingüísticos. Por último, cabe destacar que se está produciendo una creciente distensión del diálogo entre la LC y otras áreas de la lingüística (Baldwin & Kordoni 2011).

cabecera

¶ Recursos lingüísticos

La European Language Resources Association define el término *recurso lingüístico* como "un conjunto de datos y descripciones del habla o de la lengua en un formato legible por ordenador". Los recursos lingüísticos ocupan un papel central en la LC, ya que cualquiera de sus enfoques (véase la sección Enfoques computacionales) requiere algún tipo de datos. A continuación, establecemos una distinción entre tres tipos de recursos lingüísticos, sobre los que profundizamos y de los que damos ejemplos para lenguas ibéricas.

Corpora

Un corpus es "una colección de fragmentos de lengua que se seleccionan y ordenan según criterios explícitos para ser utilizados como muestra representativa de la lengua" (Sinclair 1996). Se pueden

hacer varias distinciones más, por ejemplo, según la modalidad (escrita frente a hablada; en el caso de la segunda, los datos pueden proporcionarse en forma de audio y/o transcrita), según la anotación (texto plano, es decir, sin anotaciones, frente a anotado; en el caso de las anotaciones pueden referirse a muchos aspectos diferentes, como distintos niveles de análisis o metadatos), o según las lenguas incluidas (monolingüe frente a multilingüe; en el caso de los corpus multilingües pueden ser paralelos o comparables).

Un conocido corpus anotado para el castellano y el catalán es AnCora, que contiene 500.000 palabras para cada lengua con varias anotaciones a nivel morfológico, sintáctico y semántico. CORILGA es un corpus hablado del gallego anotado con metadatos del hablante. TweetMT es un corpus paralelo de tweets para el castellano emparejado con el catalán, el gallego, el euskera y el portugués. caWaC es un corpus de texto plano del catalán recopilado de la web. En Sketch Engine se pueden encontrar otros corpus para lenguas ibéricas, por ejemplo, para el portugués.

Conjuntos de datos lingüísticos estructurado

Utilizamos este término para referirnos a un conjunto variado de recursos lingüísticos básicos que, a diferencia de los corpus, tienen cierto grado de estructura. Entre ellos se encuentran los léxicos computacionales, las ontologías (tanto las fundacionales como las específicas de un dominio) y las terminologías. Estos recursos se han popularizado en la LC con el desarrollo de diccionarios legibles por ordenador. Aunque su uso se limitó inicialmente al estudio del léxico, desde entonces han cobrado popularidad distintos métodos para su aplicación al análisis de textos. Estos conjuntos de datos pueden describirse como repositorios estructurados de información lexicográfica y/o conocimiento del mundo. Un repositorio de este tipo muy popular para las lenguas ibéricas es el Multilingual Central Repository (MCR), que integra redes semanticas basadas en WordNet para el castellano, el catalán, el euskera, el gallego, el portugués y el inglés, así como la ontología Adimen-SUMO. BabelNet es un diccionario enciclopédico multilingüe y una red semántica que se construyó enlazando Wikipedia con WordNet. Contiene datos lingüísticos de todas las lenguas oficiales ibéricas así como del aragonés, el asturiano y el extremeño.

Un tema de investigación de gran interés en este ámbito es la interoperabilidad y reutilización de estos recursos, dado el tiempo y el esfuerzo humano que supone su creación. Se han promovido diferentes iniciativas de normalización (por ejemplo, el grupo de trabajo ISO TC37). La promoción de la web semántica ha impulsado la comunidad de Linked Open Data, la cual ha encontrado sus proponentes, métodos y aplicaciones en el ámbito de los léxicos y las ontologías, dando lugar al movimiento Linguistic Linked Open Data.

Conjuntos de datos de referencia

Bajo este término se engloban una serie de conjuntos de datos desarrollados específicamente para evaluar el rendimiento de múltiples sistemas de PLN. La principal diferencia entre un conjunto de datos de referencia y un corpus es que un conjunto de referencia se centra en un fenómeno lingüístico específico con el que se evalúa el rendimiento de los sistemas.

El uso de datos de referencia ha ganado popularidad gracias a las campañas de evaluación, que pueden describirse como series de talleres o conferencias centradas en una o más tareas (es decir, el fenómeno de interés) y un calendario común para la publicación de los datos de entrenamiento y evaluación, y para la presentación de las predicciones de los sistemas desarrollados por los

participantes. Una característica distintiva de las campañas de evaluación es que los participantes no tienen acceso al etiquetado humano de los datos de evaluación hasta el final del llamado "periodo de evaluación".

Figura 4. *Liberar el código es importante para garantizar la reproducibilidad.*

Una de las campañas de evaluación más importantes en el ámbito de la LC es SemEval, que se centra en aspectos semánticos. Hoy en día también se organizan campañas de evaluación a nivel nacional y regional, como forma de valorar el estado de los sistemas de PLN para lenguas específicas. Algunos ejemplos son IberLEF para lenguas ibéricas, EVALITA para el italiano y GermEval para el alemán.

Hasta ahora hemos seguido una definición estricta de los recursos lingüísticos, limitada a los conjuntos de datos. Sin embargo, este término puede tener una interpretación más amplia que incluya también las herramientas. Hoy en día es muy común en el ámbito de la LC liberar no sólo los datos sino también el código utilizado para ejecutar los experimentos, a menudo bajo licencias libres. Esta práctica se considera muy importante para garantizar la reproducibilidad (Pedersen 2008).

La disparidad en la disponibilidad de recursos lingüísticos entre las distintas lenguas es un problema conocido y un cuello de botella para la LC. Durante el proyecto META-NET se llevó a cabo una revisión del estado de cobertura de los recursos lingüísticos para 32 lenguas de la UE, lo que dio lugar a la publicación de una serie de libros blancos. Cada una de estas 32 lenguas fue evaluada en cuanto al soporte de tecnologías lingüísticas en cuatro áreas diferentes: traducción automática, interacción del habla, análisis de textos y disponibilidad de recursos lingüísticos. Como hemos ilustrado, la disponibilidad de corpora o conjuntos de datos anotados desempeña un papel crucial para el desarrollo de sistemas de PLN. Las lenguas con pocos recursos (por ejemplo, el euskera y el catalán, entre muchas otras) sufren esta disparidad. Esto también puede tener efectos negativos en la sociedad: por ejemplo, la falta de sistemas de TA adecuados está representando un problema para garantizar el acceso a la información correcta durante la pandemia de COVID-19.

cabecera

¶ Aplicaciones

Son muchas las **tareas** que se investigan en el ámbito de la LC. Éstas se suelen clasificar según el nivel de análisis lingüístico en el que operan (véase la sección Niveles de análisis). Además de las tareas, también hay aplicaciones. Mientras que el objetivo principal de las tareas es abordar computacionalmente un fenómeno concreto de una lengua, el objetivo principal de una aplicación es abordar un problema del mundo real que implique al lenguaje natural, y que normalmente es más amplio que una sola tarea. A continuación, describimos brevemente un conjunto de aplicaciones populares de la LC por orden alfabético:

- El **análisis de sentimientos** tiene por objeto determinar los estados afectivos expresos en el texto. Un ejemplo común es la clasificación de los mensajes en las redes sociales como positivos, negativos o neutros.

- Los **asistentes de escritura** tratan de mejorar la calidad de la redacción del usuario, por ejemplo, corrigiendo los errores ortográficos y gramaticales y proporcionando sugerencias de autocompletado.
- La **atribución de autoría** tiene como objetivo determinar quién es el autor de un texto cuya autoría se desconoce.
- La **búsqueda de respuestas** tiene como objetivo encontrar respuestas a preguntas formuladas por el usuario en grandes colecciones de texto.
- El **reconocimiento óptico de caracteres**, comúnmente abreviado como OCR (del inglés *Optical Character Recognition*) tiene como objetivo convertir imágenes de texto, que pueden ser mecanografiadas o manuscritas, en texto codificado por la máquina.
- El **resumen de textos** tiene como objetivo acortar los textos de entrada preservando la información más importante.
- La **traducción automática** aborda la traducción de textos entre diferentes lenguas naturales de manera automatizada.

cabecera

Potencial para la investigación

La aplicación de métodos y sistemas de la LC puede repercutir en muchos ámbitos diferentes, desde las tecnologías educativas hasta las actividades cotidianas (por ejemplo, los asistentes virtuales como Siri o Alexa), por mencionar solo dos. Además, están surgiendo nuevos campos que ganan cada vez más popularidad, como la analítica de las redes sociales, en la que las técnicas de la LC se aplican a los mensajes de dichas redes para, por ejemplo, monitorizar la popularidad de productos o personas, o la sociolingüística computacional. La creciente fiabilidad de los modelos de LC ha hecho que éstos se apliquen en ámbitos como las Humanidades y la Medicina. Una aplicación de gran éxito de la LC en estos ámbitos es el análisis de cantidades masivas de datos de texto para identificar información novedosa potencialmente relevante que pueda necesitar una investigación más profunda. Un ejemplo es el proyecto BiographyNet, en el que se han utilizado tecnologías de análisis de texto para documentos históricos en coordinación con historiadores. Más recientemente, el estallido de la pandemia del COVID-19 ha impulsado el desarrollo de aplicaciones de extracción de información a partir de publicaciones académicas relacionadas con el COVID. En definitiva, la disponibilidad de las tecnologías de PLN está teniendo un gran impacto en la sociedad, y se prevé un mayor crecimiento.

Recientemente, la última oleada de sistemas de PLN basados en modelos del lenguaje (LM) preentrenados, junto con la creciente presencia de código listo para usar, ha impulsado una mayor democratización de la LC y el PLN. Estos LM utilizan grandes redes neuronales y están preentrenados con grandes cantidades de datos, lo que implica grandes requisitos computacionales. El coste de entrenar el GPT-3, el más reciente de estos modelos en 2021, se estima en 12 millones de dólares. Esto implica que solo un reducido número de grandes empresas puede entrenar los modelos de los que depende toda la comunidad de la LC. Dicho esto, el éxito de estos modelos es innegable: en numerosas y diferentes tareas de PLN, los LM preentrenados están logrando resultados superiores a los de otras técnicas. Su éxito ha llevado a la comunidad investigadora a estudiar qué tipo de conocimiento lingüístico está realmente codificado en estos modelos, y también qué significa que estos modelos "entiendan el lenguaje" o su capacidad de generalización. En términos más generales, la interpretabilidad de las representaciones ha

suscitado el interés de una creciente comunidad investigadora. El interés en esta área también se ve impulsado por los requisitos de interpretabilidad y explicabilidad del reglamento GDPR de la UE.

Otros aspectos que suscitan un interés y una atención crecientes en las comunidades de la LC y el PLN son la relación entre los datos, los algoritmos y la ética. Una prolífica área de investigación se centra en la comprensión de los sesgos en los datos que pueden ser transferidos a los sistemas. Por ejemplo, Bolukbasi, Chang, Zou *et al*. (2016) han demostrado que los *word embeddings* entrenados en Google News no están libres de estereotipos de género femenino/masculino.

Cerramos esta última sección con una breve exposición sobre la relevancia de las técnicas de la LC y el PLN en la traducción automática. Mientras que en el anterior enfoque estadístico era habitual que las distintas aplicaciones de PLN utilizaran diferentes técnicas computacionales, en el actual enfoque neuronal existe una tendencia convergente, de modo que la arquitectura central, por ejemplo, Transformer (Vaswani, Noam, Parmar *et al.* 2017), se comparte entre distintas aplicaciones. Esto ha acercado a las comunidades investigadoras que trabajan en diferentes aplicaciones del PLN (incluida la traducción automática), fomentando la fertilización cruzada de ideas. La mayoría de los sistemas de traducción automática que se construyen hoy en día solo utilizan texto plano, por lo que emplean conocimientos lingüísticos implícitos. Sin embargo, aumentar estos sistemas con conocimiento lingüístico explícito es útil en algunos escenarios, por ejemplo, la segmentación morfológica para las lenguas aglutinantes (Ataman, Negri, Turchi *et al.* 2017) y la información sintáctica para lenguas con pocos datos (Li, Xiong, Tu *et al.* 2017).

cabecera

Referencias

Abzianidze, Lasha; Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen & Johan Bos. 2017. “The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations” @ Lapata, Mirella; Phil Blunsom & Alexander Koller (eds.) 2017. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 242-247. Valencia: ACL. [+info] [quod vide]

Artetxe, Mikel; Gorka Labaka, Eneko Agirre & Kyunghyun Cho. 2018. “Unsupervised neural machine translation” @ *Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018)*. Vancouver. [+info] [quod vide]

Ataman, Duygu; Matteo Negri, Marco Turchi & Marcello Federico. 2017. “Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English” @ *The Prague Bulletin of Mathematical Linguistics* 108, 331-342. [+info] [quod vide]

Baldwin, Timothy & Valia Kordoni. 2011. “The Interaction between Linguistics and Computational Linguistics” @ *Linguistic Issues in Language Technology* 6/1, 1-6. [+info] [quod vide]

Baroni, Marco & Alessandro Lenci. 2010. “Distributional Memory: A General Framework for Corpus-Based Semantics” @ *Computational Linguistics* 36/4, 673-721. [+info] [quod vide]

Biemann, Chris. 2011. *Structure Discovery in Natural Language* Berlin: Springer. [+info]

Bolukbasi, Tolga; Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama & Adam T. Kalai. 2016. “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings” @ *Advances in Neural Information Processing Systems* 29. [+info] [quod vide]

* Eisenstein, Jacob. 2019. *Introduction to Natural Language Processing*. Cambridge: The MIT Press. [+info]

Fillmore, Charles J. 1968. “The case for case” @ Bach, Emmon & Robert T. Harms (eds.) 1968. *Universals in Linguistic Theory*, 1-88. New York: Holt, Rinehart & Winston. [+info] [quod vide]

Firth, John R. 1957. “A synopsis of linguistic theory 1930-1955” @ *Studies in linguistic analysis*, 1-32. Oxford: Blackwell. [+info]

Gardent, Claire & Bonnie Webber.1988. “Describing discourse semantics.” @ *Fourth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+ 4)*, Philadelphia: ACL, 50-53. [+info] [quod vide]

* Goldberg, Yoav. 2017. *Neural Network Methods in Natural Language.* San Rafael: Morgan & Claypool Publishers. [+info]

Gómez Pérez, José Manuel; Ronald Denaux & Andrés García Silva. 2020. *A Practical Guide to Hybrid Natural Language Processing*. Berlin: Springer. [+info]

Grosz, Barbara J & Candace L. Sidner. 1986. “Attention, Intentions, and the Structure of Discourse” @ *Computational Linguistics* 12/3, 175-204. [+info] [quod vide]

* Jurafsky, Dan & James H. Martin. (in progress). *Speech and Language Processing*. 3rd edition. Stanford: Stanford University. [+info]

Klavans, Judith L. & Philip Resnik (eds.) 1996. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge: The MIT Press. [+info]

Li, Junhui; Deyi Xiong, Zhaopeng Tu, Muhua Zhu, Min Zhang & Guodong Zhou. 2017. “Modeling source syntax for neural machine translation” @ *Proceedings of ACL*. [+info] [quod vide]

Mann, William C. & Sandra A. Thompson. 1988. “Rhetorical structure theory: Toward a functional theory of text organization” @ *Text* 8/3, 243-281. [+info]

* Manning, Christopher & Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press. [+info]

Mikolov, Tomas; Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. “Distributed representations of words and phrases and their compositionality” @ *Advances in Neural Information Processing Systems* 26. [+info] [quod vide]

Navigli, Roberto. 2009. “Word sense disambiguation: A survey” @ *ACM computing surveys (CSUR)* 41/2. [+info] [quod vide]

Pedersen, Ted. 2008. “Last Words: Empiricism Is Not a Matter of Faith” @ *Computational Linguistics* 34/3, 465-470. [+info] [quod vide]

Schubert, Lenhart. 2014. “Computational Linguistics” @ *Stanford Encyclopedia of Philosophy*. Stanford: Stanford University. [+info]

Straka, Milan & Jana Straková. 2017. “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe” @ *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88-99. Vancouver: ACL. [+info] [quod vide]

Stratos, Karl; Michael Collins & Daniel Hsu. 2016. “Unsupervised part-of-speech tagging with anchor hidden markov models” @ *Transactions of the Association for Computational Linguistics* 4, 245-257. [quod vide]

Vaswani, Ashish; Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. “Attention is All you Need” @ *Advances in Neural Information Processing Systems* 30, 5998-6008. [+info] [quod vide]

Webber, Bonnie Lynn. 1988. “Tense as Discourse Anaphor” @ *Computational Linguistics* 14/2, 61-73. [+info] [quod vide]

Créditos

Antonio Toral Ruiz

Profesor titular en tecnología lingüística en la Universidad de Groninga (Países Bajos). Se doctoró en Lingüística Computacional en la Universidad de Alicante (España) y lleva investigando en el ámbito de la traducción automática (TA) desde 2010. Sus intereses de investigación incluyen la aplicación de la TA a los textos literarios, la TA para las lenguas con pocos recursos y el análisis de las traducciones producidas por máquinas y humanos.

Tommaso Caselli

Profesor titular en semántica computacional en la Universidad de Groninga (Países Bajos). Se doctoró en Lingüística Computacional por la Universidad de Pisa (Italia). Sus principales áreas de investigación son la extracción y representación de eventos, el procesamiento temporal y la extracción de tramas. Es uno de los promotores de la serie de talleres *Event and Stories in the News*, y actualmente trabaja en el desarrollo de modelos computacionales y herramientas de PLN para extraer estructuras argumentales de las noticias.