Generación del Lenguaje Natural: retos y desafíos científicos

Lloret, Elena; Suárez, Armando; Ferrández, Antonio; Navarro, Borja; Martín, Tania J.; Vicente, Esther Marta; Miró, María; Moreda, Paloma.; Romá, María Teresa; Palomar, Manuel.

GPLSI. Grupo de investigación en Procesamiento del Lenguaje Natural.
Universidad de Alicante

Resumen

En este informe científico se identifican los principales retos científicos en los cuales está enmarcado el grupo de investigación en procesamiento del lenguaje y sistemas de información (GPLSI) de la Universidad de Alicante (UA). El objetivo de este informe es establecer un plan de investigación basado en la experiencia y participación en proyectos de investigación, y subrayar los principales retos de futuro que nos permitan definir nuevos proyectos de investigación y desarrollo en los próximos años.

La Generación Automática de Lenguaje Natural (GLN) es una tarea compleja que se apoya en numerosos recursos y herramientas del Procesamiento del Lenguaje Natural (PLN). La GLN comienza a dar resultados satisfactorios, al menos para ciertos dominios y objetivos, como se puede ver de los resultados obtenidos por parte del GPLSI. No obstante, requiere un importante esfuerzo investigador.

En este artículo planteamos, desde la perspectiva que da la investigación realizada hasta ahora en grupo de investigación Procesamiento Natural y Sistemas de Información (GPLSI), futuras líneas de investigación en GLN, incorporando diferentes disciplinas y metas.

Introducción

La generación automática de lenguaje natural es un gran desafío dentro del PLN. Más allá de la comprensión del lenguaje natural (CLN), significa la producción de texto o habla de forma coherente, inteligible y adecuada para el objetivo de comunicación deseado.

A la hora de diseñar un sistema de GLN se han de tener en cuenta tres aspectos principales: las etapas abordadas, la información de entrada y el objetivo comunicativo. Así, tradicionalmente, el proceso de generación se ha abordado en tres grandes etapas: planificación del documento (macroplanificación), microplanificación y realización lingüística (Reiter y Dale, 2000). En la fase de macroplanificación, el sistema debe decidir qué información debe incluirse en el texto y cómo organizarla en una estructura coherente, dando lugar a un plan de documento. A partir de este plan de documento, en la etapa de microplanificación se generará un plan discursivo, donde se decidirá qué palabras y

expresiones referenciales son apropiadas para construir las oraciones. Finalmente, la etapa de realización lingüística genera el texto final con la información y la estructura seleccionadas.

En todos estos aspectos y módulos de un sistema GLN hay muchas tareas por desarrollar y se abren múltiples caminos y aproximaciones que investigar, siendo como es la GLN un área del PLN en expansión tanto por los resultados obtenidos hasta ahora como por las necesidades de la sociedad de la información cada vez más patentes.

Este documento profundizará en algunas de esas propuestas que, pensamos, impulsarán la consecución de sistemas GLN más precisos y, finalmente, útiles. En primer lugar, revisaremos los trabajos que para nosotros son más relevantes en el área. A continuación, haremos una breve reseña de dos proyectos de investigación recientes y cercanos al tema que nos ocupa. Más tarde, expondremos algunas de esas nuevas líneas de investigación mencionadas. Finalmente, expondremos las conclusiones que, a nuestro entender, se extraen del trabajo presentado.

Estado de la cuestión

Cada una de las etapas en un sistema GLN mencionadas anteriormente aborda objetivos y tareas diferentes. Algunas investigaciones se centran en una de estas etapas de forma independiente. Como ejemplos, destacan SimpleNLG (Gatt y Reiter, 2009), que prioriza la etapa de realización; AIGRE (Smith y Lieberman, 2013), cuyo enfoque se centra en la tarea de generación de expresiones de referencia; o el enfoque propuesto por Gardent y Pérez-Beltrachini (2017), que se centra exclusivamente en la etapa de microplanificación.

También ha habido algunos intentos de abordar el proceso de GLN como un todo, principalmente utilizando técnicas clásicas de aprendizaje automático. Es el caso de Duma y Klein (2013), quienes propusieron un método simultáneo para obtener plantillas de forma automática, seleccionar el contenido de estas, y determinar la estructura y el vocabulario del texto a generar. O el de Konstas y Lapata (2013), que analizaron varios mecanismos para asociar información procedente de una base de datos de pronósticos meteorológicos a oraciones en lenguaje natural. No obstante, estos sistemas están generalmente limitados a un único dominio, tipo de texto y objetivo comunicativo.

Atendiendo a la información de entrada de los sistemas actuales de GLN, ésta puede ser datos o texto, con enfoques “de concepto a texto” (*concept-to-text*) o “de texto a texto” (*text-to-text*), respectivamente. Generalmente, los enfoques “de concepto a texto” toman datos numéricos como entrada, pero también no numéricos procedentes de bases de datos, corpus anotados o registros (logs). Ejemplos de este tipo de enfoques incluyen el sistema de recomendaciones desarrollado en (Lim-Cheng et al., 2014) y el sistema de generación de historias propuesto en (Laclaustra et al., 2014).

El otro tipo de enfoques, “de texto a texto”, parte de un texto del que se extraen las ideas relevantes, y se genera uno nuevo, con variaciones a partir de las palabras originales o aportando nueva información relacionada. Los enfoques de resúmenes de texto son el ejemplo más representativo de los sistemas de este tipo (Mille et al., 2016), pero hay otros, como los sistemas de búsqueda de respuestas (Mazidi y Tarau, 2016) o sistemas de simplificación de textos (Saggion et al., 2016).

Debido a que los recursos necesarios son costosos y difíciles de obtener para las aproximaciones anteriores, los enfoques "de base de conocimiento a texto" (*knowledge base-to-text*) han comenzado a utilizarse en los sistemas de GLN (Gyawali y Gardent, 2014; Pérez-Beltrachini et al., 2016). Estos enfoques parten de ontologías o conjuntos de datos en formato RDF, y generan texto en lenguaje natural.

Por otro lado, para PLN en general, incluyendo CLN y GLN, la investigación y desarrollo de sistemas más flexibles se ha convertido en una de las prioridades, como se puede observar por la gran cantidad de esfuerzo invertido en técnicas de inteligencia computacional estadística (Bellegarda y Monz, 2016). Los proyectos construidos sobre estas premisas avanzan bajo el supuesto de que el aprendizaje dinámico, a partir de datos de ejemplo, incrementa la adaptabilidad del sistema a diferentes contextos. Esto es ahora posible gracias al progreso de la tecnología y a la gran cantidad de información heterogénea disponible (Lohr, 2012).

Además, la incorporación de información contextual está impulsando el progreso hacia métodos más sofisticados y avanzados. Ejemplos de tales métodos incluyen algoritmos de aprendizaje profundo (*deep learning*) que entrenan redes neuronales para producir resultados más adecuados según el contexto, y redes que incorporan capacidades de memoria para soportar procesos de inferencia. Concretamente, en tareas de GLN, los trabajos pioneros en este tipo de técnicas investigaron modelos de lenguaje basados en caracteres de tal manera que, para generar texto, se predecía el siguiente carácter a uno dado (Sutskever et al., 2011). Actualmente, se han investigado distintos tipos de arquitecturas basadas en redes neuronales en el contexto de la GLN. Por ejemplo, (Dong et al., 2017) aplica satisfactoriamente modelos de codificación-decodificación para parafrasear una oración, mientras que (Costa et al., 2018) utilizan redes LSTM para generar explicaciones en lenguaje natural. Otros tipos de arquitecturas que están cobrando cada vez una mayor relevancia para la generación de textos son las redes generativas antagónicas (GAN) (Subramanian et al., 2017). En todos los casos, los resultados positivos obtenidos hacen que estos tipos de algoritmos de aprendizaje sean muy prometedores para el futuro de los enfoques de GLN.

Resultados previos

En esta sección se exponen proyectos de investigación (INTEGER, FIRST y RESCATA) precursores de las futuras líneas de investigación.

Proyecto Generación Inteligente de Textos[1]

Este proyecto se propuso, y se está desarrollando, bajo la hipótesis de que la integración del objetivo comunicativo en los sistemas de generación automática del lenguaje natural (GLN) propiciarán sistemas con alta flexibilidad en cuanto a su dominio de aplicación y resultados. Es el primer paso para obtener una aproximación holística al problema. Además, que

[1] INTEGER, *Intelligent Text Generation RTI*2018-094649-B-I00

mediante el uso de técnicas de aprendizaje profundo se puedan producir varios tipos de textos partiendo de fuentes de información heterogéneas.

Como validación de tales hipótesis, se planteaban las siguientes cuestiones: ¿es posible obtener un enfoque holístico capaz de resolver automáticamente diferentes problemas de GLN? ¿Se puede generar dinámicamente una amplia variedad de textos de gran calidad para diversos propósitos? La obtención de la respuesta a estas preguntas es el objetivo de este proyecto, un nuevo paradigma de GLN que nos permita un enfoque inteligente, capaz de producir textos de distintas naturaleza y finalidad de uso.

Según uno de los referentes fundamentales del área (Reiter y Dale, 2000), a la hora de definir formalmente las entradas de un sistema GLN, además de la fuente de conocimiento, es necesario definir un objetivo comunicativo que condicione la generación para que el resultado informe. Objetivos comunicativos son entretener, persuadir, explicar o recomendar, pero también muchos otros según se entienda la necesidad de nuevo texto o habla.

Históricamente, la mayor parte de la investigación en PLN se ha centrado en la comprensión del lenguaje natural, relegando la tarea de GLN a la mera extracción literal de fragmentos de texto (Narayan et al., 2018), el uso de técnicas “copia-pega” (Jing y McKeown, 2000), plantillas (Mitchell et al., 2014) o enfoques *ad hoc* para un dominio que generan lenguaje mediante vocabularios y gramáticas específicas, restringiendo su variabilidad y variedad (Bouayad-Agha et al., 2012; Androutsopoulos et al., 2013).

Debido a la complejidad del proceso de generación, el objetivo comunicativo se suele asumir en el diseño del sistema. Así pues, la meta comunicativa permanece invariable o restringida a un conjunto reducido de opciones previamente determinadas por la aplicación que se va a construir. Son ejemplos de sistemas que sólo cumplen un único objetivo de comunicación SumTime (Reiter et al., 2005) —informes sobre predicciones meteorológicas—, SkillSum (Williams y Reiter, ¡2008) —informes de evaluación académica para la retroalimentación de los estudiantes— PersuAIDE! (Munigala et al., 2018) — oraciones persuasivas para captar la atención de los usuarios ante determinados productos del dominio de la moda—, y Shed (Lim-Cheng et al., 2014) — recomendaciones para sugerir dietas personalizadas basadas en el perfil del usuario—.

Casi ningún sistema, excepto el sistema Personage (Mairesse y Walker, 2011)aborda más de un objetivo comunicativo. En este sistema, la recomendación y la comparación son los dos objetivos comunicativos que se pueden elegir. A pesar de que sus autores lo describen como un "generador de lenguaje altamente parametrizable", la mejora y adaptación del sistema para otros objetivos comunicativos no sería tan directa y supondría un elevado coste.

Por todo ello, nuestra hipótesis de partida es que es posible integrar el objetivo comunicativo en el proceso, flexibilizando los sistemas de GLN de forma que, con un enfoque holístico, se contemplen todas las etapas del proceso de GLN. Además, que la producción de varios tipos de textos se puede hacer partiendo de fuentes de información heterogéneas.

INTEGER parte y continúa a partir de los resultados obtenidos en proyectos previos de nuestro grupo de investigación. Cabe mencionar entre ellos, “FIRST: A flexible interactive reading support tool” (FP7-287607) [1], “LEGOLang: Técnicas de deconstrucción aplicadas a las Tecnologías del Lenguaje Humano” (TIN2012-31224) [2], y “RESCATA: Representación

Canónica y Transformaciones de los textos aplicado a las Tecnologías del Lenguaje Humano" (TIN2015-65100-R) [3]

El proyecto FIRST investigó técnicas de PLN —centrándose únicamente en la parte de CLN— para producir versiones simplificadas o enriquecidas de textos de diferente naturaleza y dominio, de manera que dichos textos fueran más fáciles de comprender. LEGOLang y RESCATA avanzaron en un sistema canónico de representación del lenguaje a partir del cual se pudieran obtener diferentes flexiones de un texto, entendidas como versiones o transformaciones del texto en función de las necesidades del usuario como, por ejemplo, texto enriquecido con imágenes o definiciones o texto simplificado.

En base a esta experiencia, el proyecto INTEGER es un paso más, ahora centrado en la GLN, aprovechando y ampliando los resultados obtenidos. Entre los objetivos específicos del proyecto mencionaremos

1. Analizar y definir qué es un objetivo comunicativo y su implicación en la producción de lenguaje.
2. Caracterizar los objetivos comunicativos y analizar su relación con los géneros textuales.
3. Investigar, proponer y desarrollar enfoques novedosos para generar modelos comunicativos de lenguaje utilizando técnicas de PLN y algoritmos avanzados de aprendizaje automático, como el aprendizaje profundo (deep learning).
4. Investigar, proponer y desarrollar un método holístico de GLN, flexible y adaptativo dinámicamente, que utilice el conocimiento obtenido de los modelos comunicativos de lenguaje.

En el momento en el que se está redactando este documento, el proyecto se encuentra en su tercer y último año, pudiendo decir que los objetivos 1 y 2 se han cumplido. Sin embargo, los objetivos 3 y 4 siguen sin ser alcanzados plenamente. Las nuevas propuestas que se van a mostrar aquí supondrán ese impulso adicional y necesario para satisfacer esos objetivos al tiempo que ampliarán el alcance de los resultados esperados.

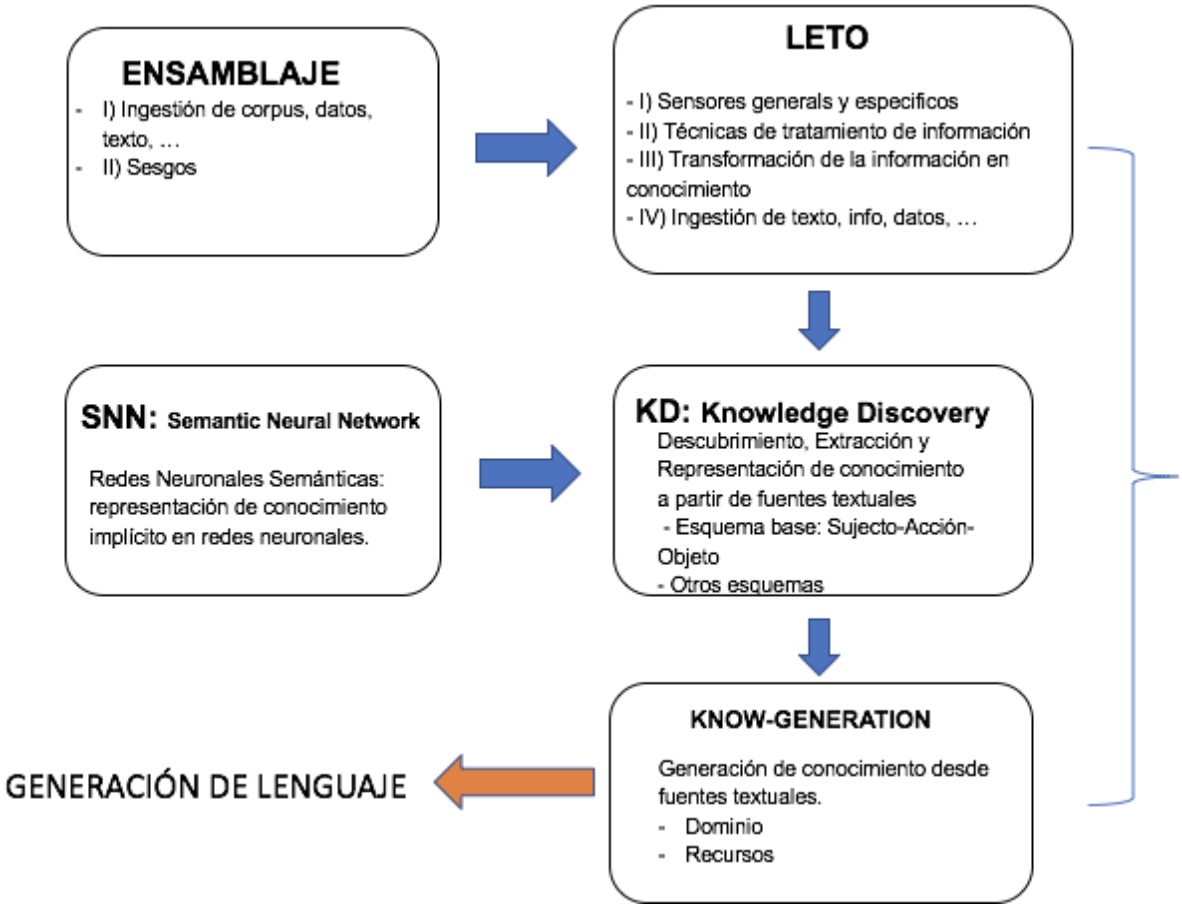
Desafíos y oportunidades científicas

En este apartado se presenta diferentes desafíos y oportunidades para el tratamiento de la generación del lenguaje. Desafíos y oportunidades que son consecuencia de la investigación desarrollada en el seno del grupo de investigación.

1. Preparación de datos y textos, construcción y ensamblaje de un data-text-lake

En el siguiente gráfico se establece un enfoque basado en el sistema LETO de extracción y tratamiento de la información, su representación y descubrimiento de nuevo conocimiento que permita planificar la generación del informe. Las etapas están descritas en el siguiente gráfico.

1. Preparación de datos y textos, construcción y ensamblaje de un data-text-lake
2. Preparación del texto, descubrimiento de conocimiento y red semántica
3. Planificación de la generación del lenguaje



A continuación, nos centraremos en la descripción de la primera etapa.

La sociedad se encuentra en un momento de revolución tecnológica y de la información, caracterizado por la velocidad sin precedentes a la que los cambios tecnológicos se producen, y por el aumento exponencial de la cantidad de información disponible (WSIS, 2017; Telefónica, 2016). En esta actual Sociedad de la Información, una gran parte de la información está disponible en formato digital (Brynjolfsson y McAfee, 2012), de ahí que también se hable de Sociedad Digital. En ella no sólo encontramos contenido estático, sino también contenidos dinámicos generados por las personas usuarias, en diferentes formatos —texto, imágenes, audio, vídeo, etc. (McKenzie et al., 2012)— en diferentes registros (Mosquera y Moreda, 2011) —formal, informal, entre otros—, e incluso en diferentes idiomas (Navigli y Ponzetto, 2012).

Esta sobreabundancia y heterogeneidad de la información hace más complicado, si cabe, el análisis y procesamiento de la misma (1) sin hacer uso de la tecnología existente puesto que se hacen necesarias habilidades cognitivas relacionadas con nuevas formas de acceder, procesar y generar conocimiento (Donald, 2014). Por tanto, si bien en un principio la tecnología, y más en concreto las tecnologías de la información, sólo jugaban el papel de ser el medio a través del cual se podía acceder a toda esta información, las nuevas necesidades

de la sociedad han hecho que ese papel cambie, convirtiéndola no sólo en el medio de acceso sino también en el instrumento que debe proporcionar los mecanismos adecuados para garantizar la disponibilidad, la accesibilidad y la asequibilidad de dicha información.

En consecuencia, el análisis, procesamiento y comprensión de la información documental mediante las Tecnologías del Lenguaje Humano (TLH) está cobrando cada vez mayor importancia. Los avances llevados a cabo a través de sus aplicaciones, tales como la recuperación de información (Vila et al., 2013), sistemas de búsquedas de respuesta (Moreda et al. 2011), sistemas de minería de opiniones (Fernández et al., 2013), o generadores de resúmenes (Vodolazova et al., 2013), entre otros, intentan hacer una aportación a la comprensión de un texto. Esta necesidad de mejorar la comprensión de los textos mediante técnicas de TLH, se manifiesta año tras año en los congresos internacionales más relevantes como ACL (2), NAACL (3), ASSETS (4); a través de la organización de talleres específicos como el "SEM 2015: The Fourth Joint Conferences on Lexical and Computational Semantics" (5) o el "Workshop on Innovative Use of NLP for Building Educational Applications" (6); o mediante la participación y el intercambio de conocimiento a través de grupos de interés, como son el "Special Interest Group on Computational Semantics" (SIGSEM) (7) o el "Special Interest Group on Accessible Computing" (SIGACCESS) (8). Todas estas iniciativas forman un paraguas bajo el cual se recogen las investigaciones más recientes y punteras relacionadas con la comprensión del texto utilizando las TLH.

Y todo ello tiene lugar con el objetivo principal de cubrir necesidades del usuario/a. Sin embargo, los avances realizados hasta el momento han tenido su centro en el *qué* y muy pocos esfuerzos se han dirigido hasta ahora al *cómo*. La mayoría de las investigaciones abordadas hasta la fecha en TLH intentan resolver *qué* hacer para buscar información, *qué* hacer para interpretar las opiniones, o *qué* hacer para generar un resumen, por poner algunos ejemplos, pero poco o casi nada se ha hecho en cuanto a *cómo* mostrar todo el conocimiento que estos procesos generan, a las usuarias y usuarios.

No todo el mundo demanda la información de la misma manera y para garantizar la accesibilidad a la información a cualquier persona, independientemente de su formación, condición social, sexo o capacidades cognitivas, es imprescindible integrar las necesidades de la persona como parte transversal e intrínseca en cualquier tarea de TLH. A día de hoy, éste sigue siendo otro de los grandes retos pendientes para las TLH. Generalmente, las necesidades de las personas usuarias se han tenido en cuenta sobre todo en aspectos de personalización, a nivel de visualización principalmente, según las preferencias que una persona pueda tener (Tanca et al., 2011; Jayanthi y Rathi, 2014), y no tanto en cuestiones relacionadas con la accesibilidad y asequibilidad de la información.

Investigaciones recientes en TLH empiezan a abordar aspectos relacionados con las necesidades cognitivas y de formación de las personas a través de las tareas de simplificación y enriquecimiento de textos, las cuales persiguen facilitar la comprensión de los textos a un determinado público objetivo. Por ejemplo, si tomamos como referencia la tarea de generación de resúmenes, las necesidades de usuario/a abordadas hasta el momento van en la línea del tipo de resumen que se desee generar —si es un resumen genérico o centrado sobre un aspecto concreto en el que la persona está interesada; si queremos obtener un resumen de un documento en particular o de varios, conteniendo éstos información relacionada con lo que necesitamos conocer en el resumen—, la longitud del mismo, el idioma, etc. Sin embargo, desde la simplificación o el enriquecimiento de textos, las

necesidades de las personas son más amplias. Desde esta perspectiva se plantea la eliminación de posibles obstáculos relativos al significado de un texto, como en el caso del vocabulario especializado, que pueden dificultar su comprensión a diferentes tipos de usuario/a, como las personas con autismo (Martín-Valdivia et al., 2014), con síndrome de down (Stajner y Saggion, 2013), con dislexia (Rello et al., 2013) o personas que están aprendiendo un idioma (Azab et al., 2015).

Y si son pocos los estudios que tienen en cuenta las capacidades cognitivas de las personas, menos son aún los que tienen en cuenta las desigualdades por razón social o de sexo. Hasta la fecha no hay ningún trabajo en el área de las TLH en el que se haya abordado de manera explícita cómo la información tratada por sistemas de búsqueda de respuesta, de minería de opiniones o de generación de resúmenes, entre otros, ha de facilitar la información a la persona de manera que dicha información carezca de sesgos de ningún tipo. Podemos concluir, por tanto, que no hay herramientas suficientes que aborden el problema de la personalización de la información desde una perspectiva única, contemplando los procesos de comprensión, interpretación, seguimiento y generación del lenguaje que eviten sesgos no sólo cognitivos y de formación, sino también sesgos relativos a la condición social y al sexo de las personas.

La disponibilidad de la información pública de forma inclusiva, igualitaria y accesible es un derecho fundamental de la sociedad del siglo XXI y debe garantizar el uso de un lenguaje inclusivo, igualitario y accesible.

Se considera que el lenguaje es inclusivo cuando expresa la diversidad y contribuye a eliminar los ejes básicos de discriminación, como el género, la raza o la condición social. El lenguaje inclusivo debe reconocer la dignidad y derechos de todas las personas sin importar sexo, edad, raza, etnia, nacionalidad, orientación sexual, credo político o religioso. Uno de los principales elementos discriminatorios del lenguaje es la invisibilidad del colectivo.

Por otra parte, se puede considerar que el lenguaje es igualitario cuando, además de inclusivo para cada uno de los colectivos, es capaz de guardar el equilibrio entre ellos sin generar predominancias (Díaz Hormigo, 2007). Desde el punto de vista de los diferentes modelos del lenguaje para cada colectivo, se trataría de garantizar que cada modelo estuviera representado en la misma proporción. De esta manera, se consideraría lenguaje igualitario de género cuando no existe ninguna predominancia entre los modelos de lenguaje asociados al colectivo masculino o femenino.

Por último, se considera que el lenguaje es accesible si la forma de comunicar los contenidos se hace teniendo en cuenta las limitaciones que pueda tener cada uno de los colectivos, evitando la exclusión generada por la incapacidad para alcanzar la información (W3C, 2008).

En este sentido, las Tecnologías del Lenguaje Humano suponen un impulso definitivo para la creación de recursos y herramientas destinadas a favorecer el uso, identificación, corrección, transformación y generación del discurso inclusivo, igualitario y accesible (LIIA), tal y como ya se ha demostrado previamente en otros ámbitos del lenguaje citados previamente: TLH para trastornos del espectro autista (Martín-Valdivia et al., 2014), síndrome de down (Stajner y Saggion, 2013), dislexia (Rello et al., 2013) o aprendizaje de nuevos idiomas (Azab, Hokamp, & Milhacea, 2015).

Por otra parte, el uso de técnicas de perfilado de la personalidad (*user personality profiling*), que ya empiezan a desarrollarse en el ámbito de la inferencia de conocimiento para Big Data (Buraya et al., 2017), permitirá distinguir las características personales, sus intereses y preferencias generales, sus competencias o nivel de experiencia respecto de un tema, sus objetivos y qué capacidades físicas y cognitivas están vinculadas al individuo. De esta manera se pueden determinar de manera detallada cuáles son sus necesidades lingüísticas, y por tanto los servicios de información que deberían serle proporcionados.

Creación de un data-text-lake libre de sesgos

El objetivo es la aplicación de las Tecnologías del Lenguaje Humano para la creación, transformación y generación de un data-text-lake libre de sesgos para el aprendizaje automático. Un data-text-lake con datos y textos libre de sesgos, inclusivo, igualitario y accesible (IIA) para modelos de lenguaje universales. Este objetivo general deberá ser tratado a partir de los siguientes objetivos particulares.

- Estudio y caracterización de los modelos de lenguaje. Construcción de recursos y uso de tecnologías para el aprendizaje automático. Aprendizaje de los modelos discriminatorios para perfiles de usuario/a.
- Creación de un modelo de lenguaje universal: inclusivo, igualitario y accesible, que permita la fusión de los modelos de lenguaje centrados en perfiles de usuario/a garantizando el equilibrio entre éstos, y eliminando planteamientos discriminatorios.
- Desarrollo de técnicas de recuperación y clasificación de información relevante para la sociedad IIA: búsqueda de información de acuerdo con perfiles de usuario/a basado en el modelo de sus necesidades lingüísticas.
- Desarrollo de técnicas para la identificación y corrección del lenguaje discriminatorio de acuerdo con los modelos de perfiles.
- Desarrollo de técnicas de transformación y generación del lenguaje adaptado al perfil del usuario/a.

Análisis, recopilación y adquisición de recursos, herramientas y técnicas existentes.

Búsqueda de córpora existentes para analizar su utilización y/o adaptación. Además de córpora, se analizarán herramientas y técnicas de TLH que se puedan reutilizar en base a las capacidades y rendimiento que ofrezcan (por ejemplo, analizadores lingüísticos).

Construcción o adaptación de córpora específico.

Para poder aplicar técnicas de aprendizaje automático y para poder evaluar las herramientas generadas, es necesario disponer de córpora que se corresponda con las características de los modelos de lenguaje basados en perfiles que se determinan en la tarea anterior. Si los corpus analizados son insuficientes para lograr los objetivos del proyecto, será necesario construir nuestros propios corpus o adaptar los ya existentes para su utilización en técnicas de aprendizaje automático aplicadas a la finalidad del proyecto. En caso de tener que adaptar corpus que ya se hayan recopilado o anotado, así como en el caso de tener que recopilar y construir nuevos, se seguirá una metodología estándar basada en procesos de compilación y anotación de corpus que ya existan (Zafra, Gómez-Soriano, & Navarro-Colorado, 2017), y que se tomarán como referencia para definir el esquema de anotación más adecuado en cada caso. También se trabajarán técnicas para la construcción automática de corpus para aquellos casos donde sea muy costoso obtener un amplio volumen de muestras (Canales et al., 2017).

Ensamblaje de corpus

Diseñar y validar una estrategia de conjunto para extender automáticamente los corpus, particularmente orientados hacia las NER y las tareas de extracción de relaciones, a partir de las salidas de los sistemas.

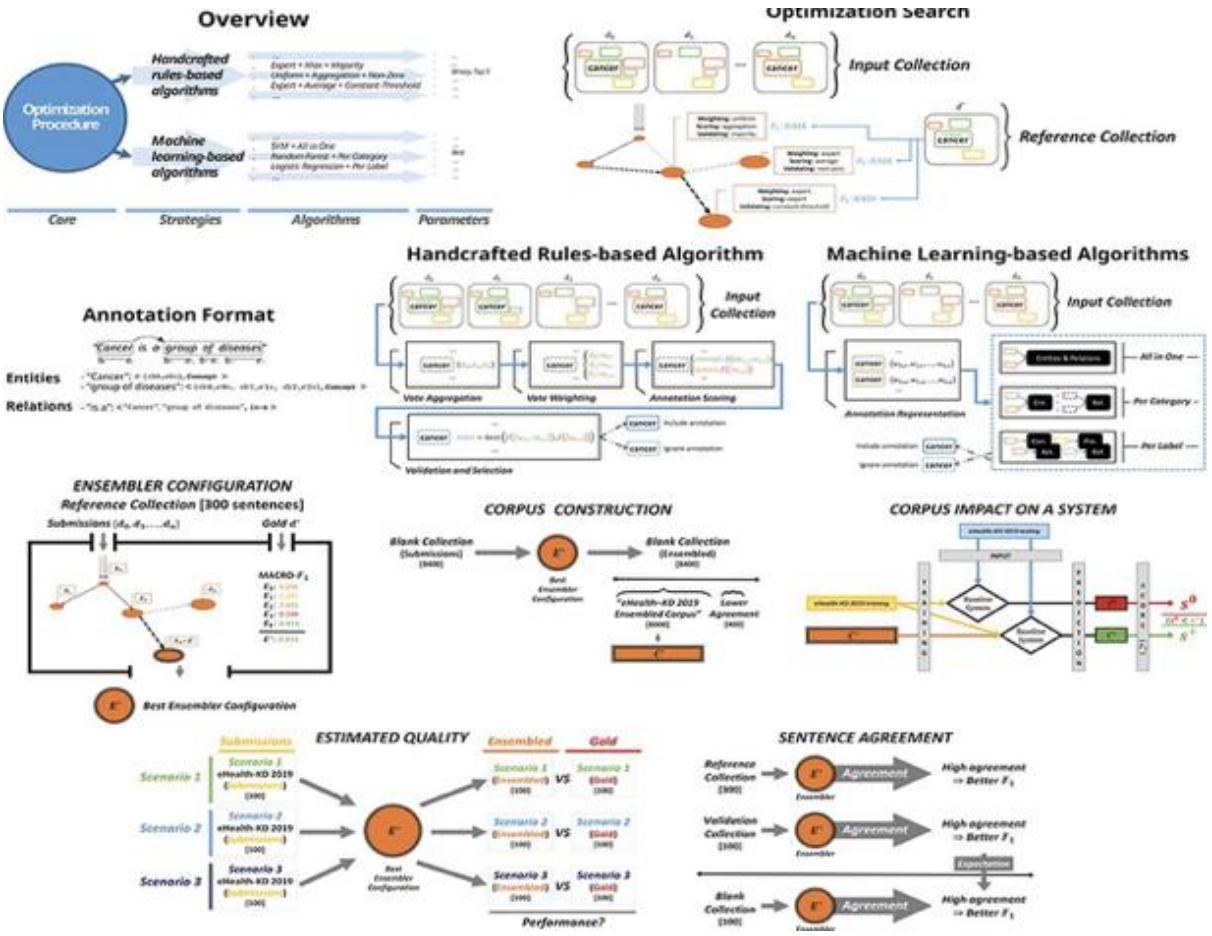
Este tipo de extensión de corpus se asemeja a una forma mejorada de hacer la extensión de corpus a través de bootstrapping: en lugar de usar un solo modelo para anotar las oraciones no etiquetadas de un corpus, las predicciones (salidas) de varios modelos se ensamblan para producir una única, más robusta (versión anotada-etiquetada).

Investigar en métodos de conjunto para la construcción de corpus etiquetados como una mejor alternativa al clásico bootstrapping.

Reducir sesgos existentes en datos y en corpus.

Construir corpus libres de sesgos con mayor calidad para el desarrollo de algoritmos de aprendizaje y desarrollo de tareas y aplicaciones

A continuación se muestra una visión general extraída de Consuegra-Ayala, J.P., Gutiérrez, Y., Piad-Morffis, A., Almeida-Cruz, Y., Palomar, M., “Automatic extension of corpora from the intelligent ensembling of eHealth knowledge discovery systems outputs“, in: Journal of Biomedical Informatics 116, 2021



2. Creación de corpus y anotación de la intención

Las vertientes derivadas de la lingüística computacional conocidas como Procesamiento y Generación del Lenguaje Natural (PLN y GLN) tienen como objetivo procesar los elementos pertenecientes a cada nivel lingüístico para elaborar sistemas que puedan etiquetar el lenguaje de forma automática y posteriormente generar texto con la condición de que sea natural. No obstante, de entre todos los niveles del lenguaje (fonológico, morfosintáctico, semántico…) que se deben procesar para generar dichos mensajes, la pragmática se ha visto en la mayoría de los casos relegada a un segundo plano. Esto se debe, de acuerdo con Cherpas (1992), a la clara preferencia de los sistemas de procesamiento por una progresión desde el nivel de análisis más bajo según los medios disponibles de cada investigación, debido a su mayor facilidad de implementación. No obstante, la pragmática está considerada como la rama de la lingüística que estudia el significado de cualquier mensaje teniendo en cuenta su contexto (Resende et al., 2020), y que, sin el cual, sería complicado conseguir la «naturalidad» de dichos mensajes, ya sea por la capacidad de la pragmática de identificar las intenciones de los hablantes o su conocimiento compartido previo, o el contexto sociocultural en el que se genera dicho mensaje, entre otros aspectos.

Como consecuencia de esta situación actual, el estudio de la pragmática desde una perspectiva computacional se ha convertido en una necesidad que, a pesar de los pequeños avances que se van reflejando en la comunidad investigadora, tal y como la creación de las ramas de estudio conocidas como Pragmática Computacional (Sayers et al., 2021) o Pragmalingüística (Esenova, 2018), todavía queda mucho por hacer, debido sobre todo a la corta existencia de esta disciplina (Bublitz y Norrick, 2011) y a la variedad de definiciones que existen actualmente dependiendo del enfoque desde el que se pretende investigar (Jucker et al., 2018), dando lugar a campos de estudio tan diferentes como la pragmática clínica, neuropragmática, pragmática cultural, pragmática variacional, entre muchas otras (Huang, 2017).

A pesar de las dificultades que empañan la consideración de la pragmática dentro de los estudios de lingüística computacional, son muchos los autores que han puesto en valor la importancia de esta rama para avanzar en la investigación dentro tanto de PLN como de GLN (Mann, 1980; Herring, 2013; Bonial et al., 2020). De ahí que actualmente exista un gran número de investigadores que con sus estudios están aportando conocimiento pragmático a los sistemas de PLN/GLN para que sean más eficientes, dividiendo la investigación de ambos campos en dos vertientes muy definidas. Por un lado, a partir de los estudios dedicados a la creación de interfaces inteligentes entre usuarios y robots con aptitudes conversacionales funcionales (Griol et al., 2016; Trott et al., 2016; Budzynska et al., 2014). Por otro lado, encontramos el área de estudio dedicada a la *computer-mediated communication* (o comunicación mediada por ordenador) (Georgakopoulou, 2011), que incluye todos los medios de comunicación englobados en la Web 2.0, como son los blogs, las webs personales o redes sociales como Twitter o Facebook. Algunos de los objetivos de la investigación focalizada en estos medios de comunicación son analizar los sentimientos de los usuarios (Tian et al., 2017), detectar las noticias falsas en periódicos digitales (Parikh et al., 2018) o identificar las intenciones que guardan los mensajes de los usuarios (Saha et al., 2019) mediante la Teoría de los Actos de Habla (SAT) fundada por Austin (1962) y definida por Searle (1969; 1985). De acuerdo con la SAT, cualquier mensaje que emitimos incluye tres dimensiones o actos: el locucionario (el simple hecho de emitir un mensaje), el perlocucionario (el efecto que tiene dicho mensaje en el receptor) y el ilocucionario (o la intención que guarda el mensaje emitido). Siguiendo esta última dimensión, los mensajes se pueden clasificar en cinco tipos de intenciones, dependiendo del propósito del hablante al emitir cada mensaje, lo que ha suscitado un gran interés en la comunidad lingüística y más concretamente, en el área de investigación de PLN y GLN (Allen, 1980; Briggs, 2013), para conseguir desarrollar sistemas que identifiquen de forma automática la intención de los mensajes y así crear en un futuro programas que logren generar texto automáticamente con una intención comunicativa concreta.

Dados los diferentes modelos de anotación que se han adaptado a cada una de las investigaciones desarrolladas en este ámbito, uno de los focos de estudio actuales en este campo es la ampliación de las guías de anotación de intenciones comunicativas para corpus lingüísticos de otros géneros textuales y que, además, estén disponibles para idiomas diferentes del inglés, donde ya existen diversos modelos referidos sobre todo a la anotación de diálogos (Leech, 2013; Kirk, 2016). Bien es cierto que ya existen algunos ejemplos de guías de anotación en español, aunque recogen transcripciones de diálogos de dominios muy concretos, como es el caso del FerroviELE (Caballero et al., 2014), creado a partir de diálogos transcritos del personal de atención al público de Renfe con clientes sobre cuestiones

administrativas y dudas sobre los servicios. Otro claro ejemplo, en este caso dentro del ámbito de la salud, es la guía de anotación de datos clínicos (Paúls, 2015), que se centra en la clasificación pragmática acorde al modelo de análisis del déficit lingüístico del Perfil PerLa de Evaluación Pragmática (Paúls, 2006).

Por lo tanto, uno de los desafíos actuales a abordar en el grupo de investigación es la creación de una guía de anotación de intenciones comunicativas que sirva de modelo para la anotación lingüística de diferentes tipologías textuales y, a ser posible, en diversos idiomas. De este modo, se conseguiría un modelo de anotación adaptable a diferentes objetivos de investigación con el que poder enriquecer con información pragmática los programas de PLN mediante la identificación automática de las intenciones de un texto. Además, la mejora de dichos programas a partir de una mayor tipología de información anotada y su posterior implementación en sistemas de GLN propiciaría la generación automática de texto teniendo en cuenta sus objetivos comunicativos, consiguiendo así un mayor grado de la naturalidad tan perseguida en esta área de investigación.

3. Generación controlada, generación guiada por el conocimiento, y su aplicación a generación de resúmenes, de historias y búsqueda de respuestas.

La creciente popularidad de los modelos de lenguaje contextuales basados en "Transformers", tales como GPT-2[2] (Radford et al., 2018), GPT-3[3] (Brown et al., 2020), T5[4] (Raffel et al., 2020), etc., ha supuesto un cambio de paradigma en la generación de lenguaje. En teoría, este tipo de modelos se definen como muy potentes y versátiles, en el sentido de que son capaces de generar textos de diferentes naturaleza -poesía (Santillan y Azcarraga, 2020), noticias periodísticas (Suraperwata y Suyanto, 2020), resúmenes (Wang et al., 2020) y dominios (Kurup et al., 2021), que además son gramaticalmente correctos y presentan una mayor fluidez en comparación con los modelos tradicionales. Sin embargo, a pesar de las aparentes ventajas que ofrecen estos modelos, en la práctica, los textos generados presentan ciertas deficiencias que ponen de manifiesto algunas de las debilidades y limitaciones respecto al uso de estos modelos. Además de seguir proporcionando una generación basada en palabras (palabra a palabra), que impide tener una visión global del contexto en el que se enmarca la información a producir, los textos generados suelen contener información repetida, manifestándose también problemas respecto a la coherencia discursiva entre las frases o párrafos, entre otros problemas. De todos los problemas que presentan estos nuevos modelos de lenguaje, hay uno que es especialmente relevante por las implicaciones y el uso malicioso que puede conllevar: la producción de texto sin ser consciente de lo que se genera, esto es, sin la capacidad de entender realmente qué es lo que se está generando. Por tanto, a pesar de que este tipo de modelos del lenguaje son los que predominan en la actualidad, no hay garantías de que la información generada de forma automática sea factualmente correcta y precisa. Esto es peligroso ya que podría dar lugar al fenómeno conocido como "*hallucination*"[5] (Rohrbach et al., 2018), es decir, la generación de información ficticia, que, a

[2] https://openai.com/blog/better-language-models/
[3] https://openai.com/blog/openai-api/
[4] https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html
[5] https://ehudreiter.com/2018/11/12/hallucination-in-neural-nlg/

su vez, podría conllevar a interpretaciones erróneas, difusión de información falsa, manipulación, sesgos, … Aunque este problema no es exclusivo de los modelos de lenguaje basados en redes neuronales, como los Transformers, y podría potencialmente ocurrir con cualquier método de GLN estadística sin ningún tipo de supervisión posterior, al ser los modelos neuronales más complejos y menos transparentes que los tradicionales, es mucho más difícil de controlar, porque aparentemente la calidad del texto generado es buena.

De hecho, modelos como GPT-3 no están exentos de controversia por su posible uso con fines malintencionados y no éticos (Floridi y Chiriatti, 2020), y sobre esto, existen varias discusiones al respecto[6],[7],[8],[9],[10]. Por otra parte, pruebas realizadas con GPT-3 en el contexto de la búsqueda de respuestas[11] confirman que aunque este modelo siempre dé una respuesta, no es capaz de razonar sobre si lo que responde es correcto o no, si la pregunta formulada tiene sentido o no, o simplemente si no sabe cuál es la respuesta. La Figura 1 ilustra dos ejemplos de preguntas, en las que se puede observar los problemas anteriormente descritos. Tal y como se muestra, mientras que para la primera pregunta, la respuesta que se obtiene es correcta, para la segunda, al seguir el mismo patrón de respuesta, el modelo da una respuesta (la que mayor probabilidad obtenga) cuando en realidad la pregunta formulada no tiene sentido y por lo tanto, lo más adecuado hubiera sido que el modelo hubiera contestado algo así como *"un pie no tiene ojos"*, *"lo siento, esa pregunta no tiene respuesta"*, o *"la pregunta formulada no tiene sentido en el mundo real"*.

```
Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many eyes does my foot have?
A: Your foot has two eyes.
```

Figura 1. Ejemplos de respuestas proporcionadas por el modelo GPT-3. Fuente: https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

La identificación de dichas debilidades es también una oportunidad para la creación y avance de nuevas líneas de investigación que analicen y planteen cómo conseguir una generación automática de textos de calidad. Sobre cierta intuición inicial de que estos nuevos tipos de modelos de lenguaje son capaces de capturar conocimiento, pero actualmente el que capturan lo hacen a un nivel muy superficial (Zellers et al., 2019) o puede estar sesgado (Shwartz y Choi, 2002), el principal reto en torno a esta línea de investigación sería cómo dotar a estos modelos de la capacidad de razonar, para conseguir una adquisición, descubrimiento e integración efectiva de conocimiento externo o sentido común en los sistemas de GLN. Esto contribuiría significativamente a mejorar la calidad semántica del texto

[6] https://towardsdatascience.com/creating-fake-news-with-openais-language-models-368e01a698a3
[7] https://www.oreilly.com/radar/ai-powered-misinformation-and-manipulation-at-scale-gpt-3/
[8] https://www.technologyreview.es/s/13206/tr10-gpt-3-representa-lo-mejor-y-lo-peor-de-la-ia-actual
[9] https://www.nature.com/articles/d41586-021-00530-0
[10] https://www.nbcnews.com/tech/tech-news/have-you-read-something-written-gpt-3-probably-not-it-n1240384
[11] https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

generado, evitando, entre otros fenómenos, la inclusión de hechos inventados no acordes a la realidad. De hecho, se dice que que el razonamiento del sentido común es la nueva frontera de la Inteligencia Artificial ---*"commonsense reasoning is the new frontier of artificial intelligence*"[12],[13], [14].

En base a esta idea, la integración de conocimiento del mundo en dichos modelos daría lugar a una GLN guiada por el conocimiento (también conocida en inglés como *knowledge-guided generation*) que permitiría una generación precisa y correcta del contenido a comunicar respecto a cómo son los hechos en realidad.

Esto sería posible ya que los Transformers, el tipo de arquitectura "end-to-end" en la que se basan los modelos de lenguaje actuales, se pueden entrenar y adaptar (*fine-tuning*) para que: 1) aprendan a generalizar e identificar información implícita; y 2) tengan en cuenta ciertas características que debe reflejar el texto generado, tales como estructura, longitud, estilo, formalidad, etc. permitiendo de esta manera una generación más controlada. Este tipo de ajustes, que se enmarcan dentro de la GLN controlada, ayudaría por un lado a generar textos de forma más natural, más diversos y con más riqueza semántica, y por otro, a detectar patrones extraños o sesgos que no tuvieran que aparecer en el texto generado. Además, el control de diferentes aspectos del texto es clave para que los sistemas de GLN se puedan aplicar y transferir con éxito a escenarios reales relevantes para la industria y la sociedad (Len et al., 2020).

El conocimiento del mundo (Yu et al., 2020), ya sea en su forma estructurada (por ejemplo, a partir de bases de datos, bases de conocimiento o redes semánticas) o no (información disponible en Internet), adquirido de manera externa gracias a distintos recursos lingüísticos y bases de conocimiento disponible, tales como ConceptNet[15], Wikipedia[16], DBPedia[17], Babelnet[18],Wordnet[19], se pueden integrar directamente en los modelos de GLN "end-to-end" así como utilizar en algunas de las fases del proceso de generación si se opta por una arquitectura modular ("*pipeline*") en la que en cada uno de los módulos se emplean modelos neuronales o se combinan modelos de lenguaje pre-entrenados conocidos como modelos de lenguaje "Plug and Play" (Dathathti et al., 2020), con clasificadores específicos en algunas de dichas fases. Cualquiera de estos tipos de enfoques puede ser válido, pues ambos presentan tanto ventajas como desventajas. Según el estudio realizado por Ferreira et al. (2019), mientras que los sistemas "end-to-end" dotan de mayor fluidez al texto generado, las arquitecturas modulares generan textos más correctos desde el punto de vista semántico y presentan mayores capacidades a la hora de generalizar sobre información no vista anteriormente. Como se ha comentado previamente, el problema de la "alucinación"

[12] https://homes.cs.washington.edu/~msap/acl2020-commonsense/slides/01%20-%20Intro.pdf
[13] https://www.zdnet.com/article/the-next-frontier-for-artificial-intelligence-learning-humans-common-sense/
[14] https://www.accenture.com/_acnmedia/PDF-154/Accenture-Artificial-Geneal-Intelligence-Report-v2-tbd.pdf
[15] https://conceptnet.io/
[16] https://es.wikipedia.org
[17] https://es.dbpedia.org/
[18] https://babelnet.org/
[19] https://wordnet.princeton.edu/

generando textos imprecisos o incorrectos desde el punto de vista semántico es uno de los problemas actuales que presentan los sistemas neuronales “end-to-end”.

De hecho, la incorporación de conocimiento en los modelos de lenguaje posibilitaría y/o mejoraría la calidad de numerosas aplicaciones de GLN que podrían ser aplicadas a multitud de dominios y ámbitos (financiero, periodístico, educativo, etc.). Entre ellas, cabe destacar:

- Generación de resúmenes abstractivos → la integración de conocimiento en el proceso de construcción de resúmenes contribuiría por un lado a producir resúmenes estructurados detectando conceptos relevantes, incluso cuando éstos se describen a través de eventos complejos en varias frases no consecutivas y por otro, a conseguir parafrasear el texto de los documentos a partir de los cuales se va a generar el resumen de manera fidedigna.
- Generación de historias (narrativa) → además de que el uso de conocimiento del mundo ayudaría a una mejor comprensión del hilo conductor y mejoraría la narración de la historia, esta tarea también se beneficiaría de los mismos aspectos que los comentados anteriormente para la generación de resúmenes.
- Búsqueda de respuestas → los hechos almacenados en bases de conocimiento ayudarían a completar la información y permitiría a los sistemas elaborar respuestas más detalladas.

4. Detección de relevancia y modelos de lenguaje posicionales

La generación de lenguaje natural abarca un amplio espectro de tareas cuyas entradas, salidas y objetivos comunicativos difieren en gran medida, dando lugar a una variabilidad difícilmente abordable en una sola propuesta o en un único estudio. Sin embargo, una serie de funcionalidades comunes se han definido como marco general de las aplicaciones desarrolladas en el seno de esta disciplina (Reiter y Dale, 2000), que se adaptan a las diferentes estrategias planteadas. Por un lado, una serie de tareas se refieren a la selección de los mensajes que se quieren transmitir, así cómo al ordenamiento de los mismos en la salida. Por otro lado, otra serie de acciones son requeridas para conseguir que tales mensajes se transformen en el texto final, el resultado del proceso de generación, que implican decisiones respecto al idioma, las palabras que se van a utilizar, la flexión de las mismas y la agregación o división de oraciones, entre otras. El primer conjunto de tareas se suele denominar *macroplanificación* del texto mientras que el segundo se ha dado en llamar etapa de *realización* del texto (Reiter y Dale, 2000). La investigación y desarrollo de aplicaciones de generación puede tener como objetivo el proceso completo que engloba a ambas partes, ya sea considerando arquitecturas integrales (también llamadas “end-to-end”), ya sea considerando arquitecturas modulares (también llamadas “pipeline”); o puede centrarse en alguna de esas subtareas en concreto. En este sentido, tareas más específicas como la generación de expresiones referenciales (Krahmer y Van Deemter, 2012) o la generación de texto desde representaciones semánticas (Zhou y Lampouras, 2020); han dado lugar a importantes subdisciplinas respaldadas por activas comunidades que consiguen avanzar el estado del arte apoyándose en los denominados *challenges* o competiciones en torno a propuestas concretas, los datos que en ellos se producen y la evaluación que en su entorno se desarrolla (Belz et al., 2009; Gardent et al., 2017).

Ahora bien, con el éxito de los modelos pre-entrenados en diversas tareas del procesamiento de lenguaje natural se ha producido un repunte de ciertos planteamientos integrales o "end-to-end" (aquellos que realizan el proceso de generación aprendiendo directamente la relación entre la entrada y la salida en un único paso). No obstante, este tipo de enfoques están siendo analizados muy críticamente desde la comunidad de generación de lenguaje natural. Por un lado, esto se debe a razones que comparten en mayor o menor medida los enfoques basados en datos que requieren corpus alineados (corpus paralelos) para aprender, dado que específicamente en el ámbito de generación, conseguir tales corpus y, más aún, corpus que respondan a la exigencia de tamaño de planteamientos deep learning, es una tarea harto difícil, y onerosa. Pero por otro lado, los planteamientos "end-to-end" están bajo la atenta observación de la comunidad porque, aunque han demostrado una significativa capacidad para generar texto fluido, presentan una serie de limitaciones, insalvables por el momento (Faille et al., 2020). Éstas se refieren, entre otras, i) a la imposibilidad de controlar el proceso de generación de modo que la creación se pueda condicionar por el objetivo comunicativo o las preferencias del usuario, por ejemplo; ii) a la opacidad de los procesos de generación que no permite determinar qué funcionalidad está mejorando y cuál introduce errores; o iii) a la desconexión entre la entrada y lo expresado en el texto de salida que, a pesar de ser fluido, puede incluir hechos falsos (se dice de tales sistemas que alucinan), hechos que no se corresponden con lo expresado en la entrada u obviar otros que deberían aparecer. Situaciones todas que, además, pueden implicar una serie de consecuencias éticas que han de ser consideradas, llegando incluso a plantearse el hecho de que cierto tipo de aplicaciones no deberían ser desarrolladas, tras el análisis de sus potenciales salidas (Zellers et al., 2019).

Mucha investigación está siendo desarrollada para aliviar tales desventajas y mejorar los sistemas basados en técnicas de aprendizaje. Se estudian estrategias para aumentar la cantidad de datos (Sha et al., 2018; Arun et al. 2020), por ejemplo, o para incrementar la transparencia y explicabilidad de los sistemas, abogando por una mejora en el proceso de creación que vuelve a arquitecturas secuenciales, a la introducción de módulos individuales que introduzcan la granularidad necesaria para que el proceso sea más transparente y condicionable, dado que los diferentes módulos realizan tareas diferenciables, las entradas pueden incluir información relacionada con la intención y el contexto, y las salidas de cada uno de ellos pueden indicar qué partes del proceso se han de reforzar y cuáles están funcionando mejor (Moryossef et al., 2019; Ferreira et al., 2019). En línea con esta idea, se ha trabajado en una propuesta de generación que se centra en uno de esos módulos o etapas funcionales mencionadas anteriormente, la macroplanificación o etapa encargada de la selección y organización de los mensajes que han de ser generados. En particular, la propuesta se centra en un tipo de generación que parte de una entrada con forma de discurso, entendido éste como un texto compuesto por una serie de oraciones que están relacionadas entre sí mediante elementos cohesivos (por ejemplo, correferencia) que confieren coherencia al texto. La propuesta ha sido desarrollada con el objetivo de vertebrar sistemas de generación flexibles, con el fin de que se emplee como un componente implícito en el proceso de generación. Una serie de trabajos han demostrado la viabilidad de tal propuesta y su capacidad de adaptación no solo a diferentes géneros (noticias, críticas, etc) y diferentes tareas de generación (creación de cuentos, generación de resúmenes y titulares), sino también a tareas que requieren para su resolución de la comprensión del discurso, como la detección de la relación entre un titular y el cuerpo de la noticia que introduce. Para conseguir tal versatilidad, la propuesta se ha basado en un tipo de modelo de lenguaje que considera tanto los elementos relevantes del texto, como la distribución de los mismos en el conjunto

del texto, evitando obviar su estructura, que es uno de los problemas más recurrentes en planteamientos estadísticos.

Los modelos posicionales son calculados considerando un vocabulario o un conjunto determinado de elementos que, hasta el momento, han permitido incluir aspectos semánticos en el proceso, considerando niveles de abstracción basados en conjuntos de sinónimos o en la detección de entidades nombradas. Sin embargo, el proceso se enriquecería al incluir elementos pragmáticos relacionados con el objetivo comunicativo o las preferencias de usuario. Mientras este último aspecto abre un capítulo nuevo y lo consideramos como siguiente paso en este trabajo, en el presente se está llevando a cabo una investigación en relación a los objetivos comunicativos a partir de su relación con los géneros textuales y la identificación de patrones que contribuyen a su definición. Esta investigación comprende la detección de diversas características comunes a documentos pertenecientes a corpus de tres géneros narrativos (noticias, críticas y cuentos) así como el análisis de su relación con los objetivos comunicativos asociados a estos géneros, de modo que puedan ser más adelante introducidos como parámetros de sistemas de generación favoreciendo, de esta manera, el diseño y desarrollo de sistemas de generación capaces de adaptar su proceso y resultados a las necesidad del contexto de producción.

5. Análisis computacional de eventos en textos narrativos, generación de resúmenes narrativos. generación de texto rítmico.

Más allá de cuestiones literarias, el texto narrativo es el tipo de texto más común en la comunicación humana. Se puede encontrar en diferentes ámbitos: desde en una conversación cara a cara, hasta en textos periodísticos, políticos, publicitarios, económicos, educativos, etc.; así como en ámbitos creativos como la literatura, el cine o los videojuegos. Su relevancia social ha sido determinada desde la filosofía o la lingüística (Ochs, 2000) o la neurociencia y la biología evolutiva (Gottschall, 2012). Este último, por ejemplo, considera que el ser humano es un “animal narrativo” (*storytelling animal*), es decir, que la narratividad es inherente al ser humano y lo caracteriza como especie. La narración es un recurso cognitivo humano que nos ayuda a comprender la realidad y a afrontar la sociedad compleja en la que vivimos (Gottschall, 2012). Es, sobre todo, su capacidad para empatizar con el receptor y transmitir emociones lo que hace del texto narrativo especialmente relevante (Ochs, 2000).

Frente a otros tipos de texto, el texto narrativo tiene una estructura especial. Podemos definirlo como aquél en el que se enuncian una serie de acontecimientos relevantes, reales o ficticios, ocurridos a unos personajes, en un lugar más o menos definido y con una organización temporal determinada, formando un todo textual coherente. Estos “acontecimientos” se manifiestan en el texto en forma de *eventos*. Desde un punto de vista general, un evento es algo que ocurre en el mundo (real o imaginario) que denota acciones, procesos o estados (Mani et al., 2005). Desde un punto de vista lingüístico, todo evento tiene una estructura determinada que incluye a los participantes del evento, el tiempo y lugar en

que se produce, instrumentos, elementos afectados por la acción, etc. (Hovav et al., 2010; Levin y Hovav, 2005).

En el ejemplo (1) podemos ver la mención del evento “land”, de cuya estructura eventiva se menciona a un participante (“the Airbus A380”), un lugar (“in the United States of America”) y un momento o tiempo (“on Monday”). El ejemplo está formalizado según el estándar ISO TimeML Working Group (I.T.W. Group, 2008) .

(1) [ING] <A0>The Airbus A380, the world’s largest passenger plane</A0>, was set to <EVENT eid=“e78”> land </EVENT> <LOC>in the United States of America</LOC> <TIMEX>on Monday</TIMEX> after a test flight.

El procesamiento automático de textos narrativos implica, por tanto, la detección e interpretación de estas estructuras eventivas. Si bien la extracción de eventos y la detección de eventos correferentes es una tarea con larga tradición en PLN (Schank y Abelson, 1977; Humphreys et al., 1997; Doddington et al., 2004), ha sido durante la última década cuando se han propuesto diferentes aproximaciones centrados en textos narrativos. La tarea se ha centrado sobre todo en detectar cadenas de eventos (sucesión lineal de evento según determinadas relaciones (Chambers y Jurafsky, 2009)) y en la correferencia de eventos (detección de dos o más eventos textuales que refieren a un mismo evento real (Lu et al., 2018)). Así, Chambers y Jurafsky (2009) proponen un modelo para extraer esquemas narrativos, esto es, secuencias de eventos relacionadas con determinados personajes (entidades persona dentro de la estructura eventiva). Otros trabajos en esta misma línea son McIntyre y Lapata (2010), Elsner (2012) o Sprugnoli y Tonelli (2017) entre otros. Destaca la propuesta de Mostafazadeh (2017), que no solo detecta eventos en textos narrativos sino que también intenta deducir el evento lógico siguiente. En esta línea hay otras propuestas como Frermann et al. (2018). Sims et al. (2019), finalmente, desarrolla un modelo para textos literarios de mayor complejidad en cuanto a estructuras eventivas.

Salvo este último trabajo, la mayoría de texto narrativos tratados en Procesamiento del Lenguaje Natural son texto periodísticos o textos con una estructura narrativa línea y en cierto modo simple. Es muy común, sin embargo, la presencia de texto narrativos que, sin llegar a ser texto literario, presentan estructuras narrativas (y dentro de ellas estructuras eventivas) bastante complejas, sobre todo en el habla cotidiana. Siguiendo la propuesta de Sims et al. (2019), una forma seria de abordar el tratamiento computacional del texto narrativo es precisamente desarrollar modelos para el tipo de texto narrativo más complejo: la novela y, en general, la prosa literaria. En este tipo de texto se hallan estructuras eventivas complejas, cuyo estudio y modelización computacional permitirían desarrollar modelos computacionales capaces de procesar cualquier tipo de texto narrativo.

Para desarrollar esta línea de investigación se necesita un corpus de novela amplio y bien editado. En este punto la *European Literary Text Collection* (corpus ELTeC (Odebrecht et al., 2019)) es quizá hoy día el recurso apropiado. El corpus ELTeC es una colección de más de 500 novelas europeas escritas en diferentes idiomas europeos (inglés, español, francés, portugués, alemán, etc.) y publicadas entre 1840 y 1920 (época dorada de la novela). Este

corpus ha sido desarrollado por miembros de este equipo dentro del proyecto europeo *Distant Reading for European Literary History* (COST Action CA16204): la selección de los textos se ha realizado de tal manera que se permitan realizar análisis comparados multilingües, y todos los textos han sido anotados con el estándar XML-TEI para facilitar su procesamiento automático. El objetivo de la línea, por tanto, es analizar y anotar las estructuras eventivas de parte de este corpus, tanto de modo manual como automático. El corpus sería así tanto el objeto de análisis como el corpus de entrenamiento y/o *gold standard* del modelo automático. A partir de la anotación, se podría desarrollar ya modelos computacionales con las diferentes técnicas de aprendizaje automático actuales.

Una vez generado el modelo de extracción y análisis de eventos completos en textos narrativos, se propone dar un paso más: desarrollar un modelo de resumen automático de textos narrativos complejos en la línea planteada en Barros et al. (2019). Esta propuesta se fundamenta en dos partes: una primera fase de selección de los eventos más relevantes de un texto para, a partir de ellos, generar en una segunda fase un resumen de tipo abstractivo.

6. *Fake news*, metáfora y GLN.

Las líneas de investigación de *Fake News* y el *Estudio de la Metáfora en Dominios Específicos en inglés* tienen impacto en el área de la lingüística y la psicolingüística, y ciencias sociales. En concreto, ambas líneas buscan mejorar la comunicación, eliminar posibles obstáculos y hacerla más eficaz para determinados usuarios como, por ejemplo, personas con distintos niveles de comprensión de idioma inglés, como lengua extranjera, que necesitan ser operacionales con este idioma en un contexto determinado, ya sea profesional, social, etc.

La primera de estas líneas, *Fake News,* explora los criterios que podrían influir en la difusión de las noticias falsas a través de redes sociales, tomando caso de estudio la población universitaria. Desde las elecciones estadounidenses del 2016, la definición de noticias falsas ha llegado a entenderse como la diseminación intencionada de información falsa imitando los estándares periodísticos tradicionales (Lazer et al., 2018). Según Talwar et al. (2019), el aumento de casos que comparten noticias falsas maliciosas en las redes sociales, se ha convertido en una gran preocupación, especialmente porque un número notable de usuarios confían en las redes sociales para informarse, por ejemplo, hasta un 62% según el estudio de Gottfried y Shearer (2016). Concretamente, nuestro estudio investiga qué criterios podrían influir en los comportamientos que impactan la difusión de las noticias falsas. Este trabajo investiga un aspecto de "digital literacy", es decir la alfabetización digital, y así descubrir posibles estrategias de intervención para abordar el problema de forma específica y más eficaz. La motivación subyacente de la investigación es trabajar hacia una sociedad menos polarizada y manipulada por la desinformación, donde haya más reflexión, ética, crítica, y donde el debate sea objetivo y equilibrado.

La segunda línea de investigación, *Estudio de la Metáfora en Dominios Específicos*, en inglés se fundamenta en la teoría de la metáfora conceptual, desarrollado en la investigación de Lakoff y Johnson (1980). En el dominio financiero, por ejemplo, podemos hallar frases como *"bear market bounce"—incremento puntual del mercado* -- o *"dead cat bounce"—último intento de revivir para un mercado ya muerto/en declive*--, entre muchos otros, que tienen su propio significado ligado al dominio. Entender el significado de estas metáforas de dominio facilita la inclusión de comunidades ajenas. Según Rai y Chakraverty (2020), entendemos

que la necesidad del momento es procesar metáforas en un lenguaje común para todas las comunidades, que a menudo es ambiguo, y que requiere un conocimiento global actualizado para entender su significado y propósito. El procesamiento de las metáforas requiere de la incorporación de pragmática, percepción y conocimiento dinámico del mundo. En el mundo de los mercados financieros se detectan nuevas metáforas conceptuales a través de las redes sociales, como por ejemplo "*diamond hands*"—, tiene mucho valor, no vendas, guárdalo-- o "*paper hands*"—no tiene valor, véndelo--. La presente línea de investigación tiene como motivación fomentar la inclusividad, en el área conocida como inglés para fines específicos, y facilitar que personas puedan asimilar información abstracta dentro de un contexto desconocido para ellas. Tiene efecto directo en, por ejemplo, la gestión de ahorros e inversiones personales, en el caso de querer invertir en mercados financieros liderados por culturas anglosajonas.

Detalles del Trabajo en Curso.

1. Fake News

    1.1. Hemos desarrollado el cuestionario en tres fases desde la 1) *face validity fase*; 2) piloto; hasta 3) el cuestionario final.

1.2. Resultados

    1.2.1. Los resultados se analizaron mediante métodos estadísticos descriptivos y el estudio de la relación entre variables.

    1.2.2. La mayoría de los participantes fueron incapaces de detectar las dos noticias falsas incluidas en el cuestionario.

    1.2.3. El comportamiento ideal de reflexionar, verificar y, si es necesario, exponer antes de compartir una noticia es el modus operandi de una minoría muy pequeña.

1.3. Actualmente, estamos trabajando en un *major review* por petición del editor de la revista *Computers & Education*. Piden, entre otras cosas, por ejemplo, que se proporciona ""validity and reliability assessment" del cuestionario.

2. Metáfora:

    2.1. Hemos descargado todos los artículos del *WSJ (Wall Street Journal*) y los tweets de los influencers en los mercados financieros que tengan más de 50,000 seguidores por el periodo del estudio (01/01/2020 – 31/03/2021).

    2.2. Hemos empezado a descargar los artículos del *FT (Financial Times).*

    2.3. Los resultados preliminares indican:

        2.3.1. Se detectan diferencias en las metáforas entre los medios *mainstream (WSJ)* y Twitter.

        2.3.2. Se detecta mucha variación por mes a lo largo del periodo estudiado, y entonces tendremos que mapearlo y ver si hay una conexión con eventos tanto políticos o económicos y la frecuencia o tipo de metáfora.

7. Arquitectura de componentes de refuerzo del aprendizaje de lengua de señas empleando proximidad fonológica

En el proceso de traducción automática se genera texto en un nuevo idioma a partir de un texto de entrada. Dicha traducción se complica mucho más cuando la entrada/salida supera el nivel textual, para incluir información del movimiento de manos y gestos, tal y como ocurre en el lenguaje de signos (LS). Para superar este obstáculo se suele recurrir a la transcripción de dicho lenguaje de signos a un lenguaje textual en el que se codifica cada movimiento o gesto. En nuestro caso, dentro de la problemática de la traducción de lengua de signos, nos hemos centrado en el proceso educativo de la lengua de signos a oyentes, el cual puede resultar especialmente complejo (Kelly, 2010) debido a las dificultades inherentes del lenguaje de signos, y a la falta de expertos en ambas lenguas, de ahí la necesidad de incorporar herramientas tecnológicas de refuerzo del aprendizaje.

La propuesta de nuestra investigación se centra en una arquitectura modular portable para el aprendizaje de la LS basada en proximidad fonológica (Naranjo-Zeledón et al., 2020; Naranjo-Zeledón et al. 2021). Este concepto de proximidad fonológica trabaja sobre la transcripción del lenguaje de signos en un vector de 29 columnas con 26 parámetros codificados en forma de valores numéricos para identificar una medida de similitud:

TSE,For,1;2;2;2;1;5;2;3;2;4;9;2;2;1;1;1;1;1;1;1;1;3;3;2;4;22;1;2;1;1,1,2,2,2,1,5,1,1,1,1,1,1,2,
3,2,3,3,2,4,9,2,2,4,22,1,2

TEC,South,2;2;2;2;1;5;4;3;2;4;15;1;2;1;1;1;1;1;1;1;1;3;3;2;4;22;1;2;1;1,2,2,2,2,1,5,1,1,1,1,1,1,4,
3,2,3,3,2,4,15,1,2,4,22,1,2

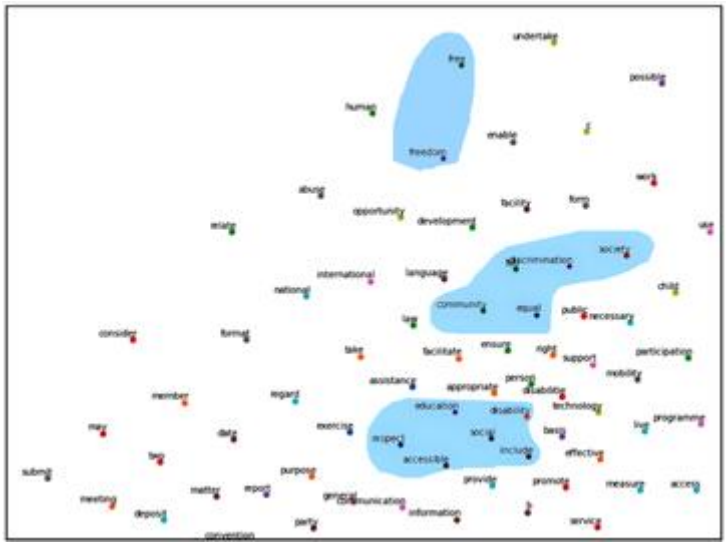
SICID,CR,4;4;4;4;1;4;3;3;2;4;15;1;2;1;1;1;1;1;1;1;1;3;3;2;4;22;1;2;1;1,4,4,4,4,1,4,1,1,1,1,1,1,3,3,
2,3,3,2,4,15,1,2,4,22,1,2

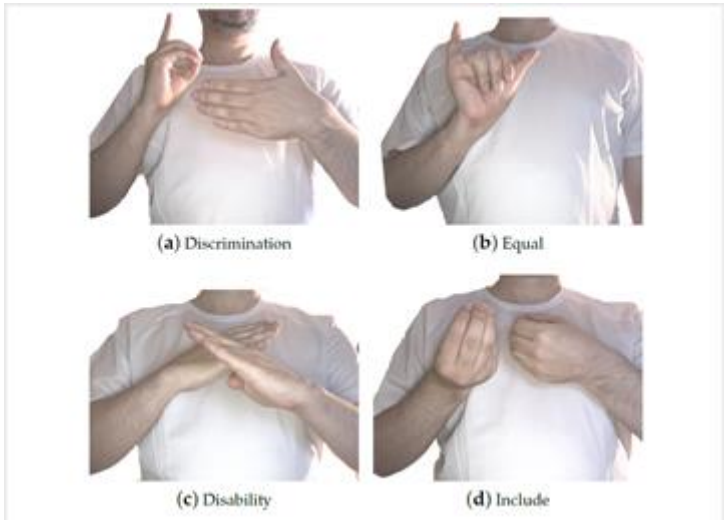
- Form:
    1. Index
    2. Medium
    3. Annular
    4. Little finger
    5. Separation
    6. Thumb
- Orientation:
    1. Rotation
    2. Wrist posture
    3. Intentionality
- Location:
    1. Space laterality
    2. Space height
    3. Depth
    4. Arm contact

Entre las diferentes medidas de cálculo de similitud (Naumann y Herschel, 2010; Bisandu et al., 2019), la medida de similitud del coseno entre dichos vectores ha sido la que mejor ha funcionado sobre los LS, permitiéndonos ser capaces de identificar signos homónimos y parónimos, los cuales hemos comprobado que guían el proceso de aprendizaje de la LS para

hacerlo más eficiente, tal y como hemos analizado dicho impacto sobre la LS costarricense (LESCO).

Asimismo, hemos incorporado Word2Vec para detectar grupos de signos con similitud semántica que también guíen el aprendizaje. Esto se ha automatizado para diferentes textos de aprendizaje, aunque para esta investigación se ha utilizado el texto en LS de la convención de los derechos de las personas con discapacidad, tal y como se aprecia en la siguiente figura en la que se observan los clústeres de signos semánticamente similares, aunque tengan signos fonológicamente distantes (p.ej. “discriminación” e “igual”):





Como resultado de esta línea de investigación, se ha implementado dicha arquitectura sobre LESCO, traduciendo dichos grupos de signos a través de un avatar:

Se ha analizado el impacto de dicha herramienta sobre el aprendizaje de LESCO a través de encuestas cualitativas y cuantitativas sobre 12 alumnos de dicha LS, obteniendo en el test SUS de usabilidad una puntuación de 89 (alto nivel) (Lewis y Sauro, 2018), y en el "phonological proximity score", obteniendo 91.0 (*the tool is useful to our objective of reinforcing sign language learning*):

LESCO learning reinforcement test, using similar signs

*Mandatory

Please rate your level of agreement with each of the following statements: *

	1 - Totally disagree	2 - Somewhat disagree	3 - Neither agree nor disagree	4 - Somewhat agree	5 - Totally agree
I think I would like to use this tool frequently.	○	○	○	○	○
I found this tool unnecessarily complex.	○	○	○	○	○
I thought this tool was easy to use.	○	○	○	○	○
I think I would need help to be able to use this tool	○	○	○	○	○
I found the various functions of this tool to be well integrated.	○	○	○	○	○
I thought there was too much inconsistency in this tool.	○	○	○	○	○

2 **Opinion/Judgment of learning improvement by using phonological proximity**

Participant	Opinion / Judgment	Tone
p1	'It helps me to locate possible signs with which I could be confused or improve the context of use'.	Positive
p2	'To be able to distinguish signs that can be confused at the time of a conversation'.	Positive
p3	'It seems very valuable to me, since with the same form, several signs can be practiced, then the hand becomes more skilled'.	Positive
p4	'I believe that having similar signs can make it easier for me to learn new signs, as long as I am interested in reviewing and comparing them thoroughly. Once you have mastered a sign, making the move to a similar one is much easier'.	Positive
p5	'At the beginning there are words that are similar and when you see them after a long period of time, these differences are not very noticeable, so it is useful to remember precise words well'.	Positive
p6	'The implementation of this section is useful and does not hinder the use of the application if it generates confusion as to why some are similar but this does not affect the flow of use'.	Positive
p7	'Yes, I think it is quite useful because it helps to know vocabulary similar to a letter / word, and it also helps to see the differences between each one so as not to be mistaken'.	Positive
p8	'The implementation of this section is useful and does not hinder the use of the application. It creates confusion as to why some are alike but this does not affect the flow of use'.	Mainly positive
p9	'I think it has facilitated because one sign is memorized and the small differences are noticed with respect to another'.	Positive
p10	'It has facilitated [sic] because by presenting similar signs, it helps to make their difference'.	Positive
p11	'Seeing similar signs allows me to learn some other signs more quickly, since I can associate them'.	Positive
p12	'It seems to me that in learning signs, having similar signs at hand allows me to find a better way to express the message with the correct signs. In addition, identify the differences and be able to apply it when using a specific sign'.	Positive

Bibliografía

Allen, J. F., & Perrault, C. R. (1980). Analyzing intention in utterances. *Artificial intelligence*, 15(3), 143-178. https://doi.org/10.1016/0004-3702(80)90042-9

Androutsopoulos, I., Lampouras, G., & Galanis, D. (2013). Generating natural language descriptions from OWL ontologies: the NaturalOWL system. *Journal of Artificial Intelligence Research*, *48*, 671-715. https://doi.org/10.1613/jair.4017

Arun, A., Batra, S., Bhardwaj, V., Challa, A., Donmez, P., Heidari, P., ... & White, M. (2020). Best Practices for Data-Efficient Modeling in NLG: How to Train Production-Ready Neural Models with Less Data. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, 64-77. https://arxiv.org/abs/2011.03877

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Barros, C., Lloret, E., Saquete, E., & Navarro-Colorado, B. (2019). NATSUM: Narrative abstractive summarization through cross-document timeline generation. *Information Processing & Management*, *56*(5), 1775-1793. https://doi.org/10.1016/j.ipm.2019.02.010

Bellegarda, J. R., & Monz, C. (2016). State of the art in statistical methods for language and speech processing. *Computer Speech & Language*, *35*, 163-184. https://doi.org/10.1016/j.csl.2015.07.001

Belz, A., Kow, E., Viethen, J., & Gatt, A. (2009). Generating referring expressions in context: The GREC task evaluation challenges. In: Krahmer, E., Theune, M. (eds.) *Empirical methods in natural language generation (EACL and ENLG 2009)*. Lecture Notes in Computer Science, vol. 5790, 294-327. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15573-4_15

Bisandu, D. B., Prasad, R., & Liman, M. M. (2019). Data clustering using efficient similarity measures. *Journal of Statistics and Management Systems*, *22*(5), 901-922. https://doi.org/10.1080/09720510.2019.1565443

Bonial, C., Donatelli, L., Abrams, M., Lukin, S., Tratz, S., Marge, M., ... & Voss, C. (2020). Dialogue-amr: abstract meaning representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 684-695. https://aclanthology.org/2020.lrec-1.86

Bouayad-Agha, N., Casamayor, G., Mille, S., & Wanner, L. (2012). Perspective-oriented generation of football match summaries: Old tasks, new challenges. *ACM Transactions on Speech and Language Processing (TSLP)*, *9*(2), 1-31. https://doi.org/10.1145/2287710.2287711

Briggs, G., & Scheutz, M. (2013). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 27(1). https://ojs.aaai.org/index.php/AAAI/article/view/8471

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin (Eds.). *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 1877-1901. *https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf*

Bublitz, W., & Norrick, N. R. (Eds.). (2011). *Foundations of pragmatics* (Vol. 1). Walter de Gruyter. https://doi.org/10.1515/9783110214260

Budzynska, K., Janier, M., Kang, J., Reed, C., Saint-Dizier, P., Stede, M., & Yaskorska, O. (2014). Towards argument mining from dialogue. In S. Parsons, N. Oren, C. Reed & F. Cerutti. (Eds.),

Computational Models of Argument, 185-196. IOS Press. https://doi.org/10.3233/978-1-61499-436-7-185

Caballero, M., Díaz, L., Taulé, M. (2014). *Guía de anotación del corpus FerroviELE*. Working paper 2: Diana-Construcciones. Universitat de Barcelona, Barcelona. http://clic.ub.edu/sites/default/files/pagines/GUIA-FERROVIELE_2.0.pdf

Chambers, N., & Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 602-610. https://aclanthology.org/P09-1068.pdf

Cherpas, C. (1992). Natural language processing, pragmatics, and verbal behavior. *The Analysis of verbal behavior*, *10*(1), 135-147. https://doi.org/10.1007/BF03392880

Consuegra-Ayala, J. P., Gutiérrez, Y., Piad-Morffis, A., Almeida-Cruz, Y., & Palomar, M. (2021). Automatic extension of corpora from the intelligent ensembling of eHealth knowledge discovery systems outputs. *Journal of Biomedical Informatics*, *116*. https://doi.org/10.1016/j.jbi.2021.103716

Costa, F., Ouyang, S., Dolog, P., & Lawlor, A. (2018). Automatic generation of natural language explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI '18 Companion)*. https://doi.org/10.1145/3180308.3180366

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., ... & Liu, R. (2020). Plug and play language models: A simple approach to controlled text generation. In the *Eighth International Conference on Learning Representations*. https://arxiv.org/abs/1912.02164

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., & Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, *2*(1), 837-840. http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf

Dong, L., Mallinson, J., Reddy, S., & Lapata, M. (2017). Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, 875-886. https://arxiv.org/abs/1708.06022

Duma, D., & Klein, E. (2013). Generating natural language from linked data: Unsupervised template extraction. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers* (pp. 83-94). https://aclanthology.org/W13-0108.pdf

Elsner, M. (2012). Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 634-644. https://aclanthology.org/E12-1065.pdf

Esenova, K. U., & Ismayilova, F. K. (2018). Major units in the notion of pragmalinguistics. *European Journal of Natural History*, (6), 45-52. https://world-science.ru/en/article/view?id=33944

Faille, J., Gatt, A., & Gardent, C. (2020). The Natural Language Generation Pipeline, Neural Text Generation and Explainability. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, 16-21. https://www.aclweb.org/anthology/2020.nl4xai-1.5

Ferreira, T. C., van der Lee, C., Van Miltenburg, E., & Krahmer, E. (2019). Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *2019 Conference on*

Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, 552-562. https://arxiv.org/abs/1908.09022

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681-694. https://doi.org/10.1007/s11023-020-09548-1

Frermann, L., Cohen, S. B., & Lapata, M. (2018). Whodunnit? Crime drama as a case for natural language understanding. *Transactions of the Association for Computational Linguistics*, *6*, 1-15. https://doi.org/10.1162/tacl_a_00001

Gardent, C., & Pérez-Beltrachini, L. (2017a). A statistical, grammar-based approach to microplanning. *Computational Linguistics*, *43*(1), 1-30. https://doi.org/10.1162/COLI_a_00273

Gardent, C., Shimorina, A., Narayan, S., & Pérez-Beltrachini, L. (2017b). The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, 124-133. https://doi.org/10.18653/v1/W17-3518

Gatt, A., & Reiter, E. (2009). SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)* (pp. 90-93). https://aclanthology.org/W09-0613.pdf

Georgakopoulou, A. (2011). Computer-mediated communication. In J. O. Östman, & J. Verschueren (Eds.). *Pragmatics in practice, 9*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gottfried, J., & Shearer, E. (2016). *News use across social media platforms 2016*. Pew Research Center. *http://assets.pewresearch.org/wp-content/uploads/sites/13/2016/05/PJ_2016.05.26_social-media-and-news_FINAL-1.pdf*

Gottschall, J. (2012). *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt.

Griol, D., & Callejas, Z. (2016). A neural network approach to intention modeling for user-adapted conversational agents. *Computational intelligence and neuroscience*, *2016*. https://doi.org/10.1155/2016/8402127

Gyawali, B., & Gardent, C. (2014). Surface realisation from knowledge-bases. In the *52nd Annual Meeting of the Association for Computational Linguistics*, 424-434. https://hal.archives-ouvertes.fr/hal-01021916/

Herring, S. C., Stein, D., & Virtanen, T. (Eds.). (2013). *Pragmatics of computer-mediated communication* (Vol. 94). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110214468

Hovav, M. R., Doron, E., & Sichel, I. (Eds.). (2010). *Lexical semantics, syntax, and event structure* (No. 27). Oxford: Oxford University Press.

Huang, Y. (Ed.). (2017). *The Oxford handbook of pragmatics*. Oxford University Press.

Humphreys, K., Gaizauskas, R., & Azzam, S. (1997). Event coreference for information extraction. In *Proceedings of the Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, 75-81. https://aclanthology.org/W97-1311.pdf

I.T.W. Group. (2008) *ISO TimeML TC37 draft international standard DIS 24617-1*. Retrieved from https://semantic-annotation.uvt.nl/ISO-TimeML-08-13-2008-vankiyong.pdf

Jing H., K.R. McKeown (2000). Cut and Paste Based Text Summarization. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics conference (NAACL 2000)*, 178-185. https://aclanthology.org/A00-2024.pdf

Jucker, A. H., Schneider, K. P., & Bublitz, W. (Eds.). (2018). *Methods in pragmatics* (Vol. 10). Walter de Gruyter GmbH & Co KG. https://doi.org/10.1515/9783110424928

Kelly, D. (2010). *Computational Models for the Automatic Learning and Recognition of Irish Sign Language* [Doctoral dissertation, National University of Ireland Maynooth]. MURAL - Maynooth University Research Archive Library. http://mural.maynoothuniversity.ie/2437/

Kirk, J. M. (2016). The pragmatic annotation scheme of the SPICE-Ireland corpus. *International Journal of Corpus Linguistics*, 21(3), 299-322. https://doi.org/10.1075/ijcl.21.3.01kir

Konstas, I., & Lapata, M. (2013). Inducing document plans for concept-to-text generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1503-1514. https://aclanthology.org/D13-1157.pdf

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173-218. https://doi.org/10.1162/COLI_a_00088

Kurup, L., Narvekar, M., Sarvaiya, R., & Shah, A. (2021). Evolution of Neural Text Generation: Comparative Analysis. In S.K. Bhatia, S. Tiwari, S. Ruidan, M. C. Trivedi & K. K. Mishra. (Eds.). *Advances in Computer, Communication and Computational Sciences*, *1158*, 795-804. Springer, Singapore. https://doi.org/10.1007/978-981-15-4409-5_71

Laclaustra, I. M., Ledesma, J., Méndez, G., & Gervás, P. (2014). Kill the Dragon and Rescue the Princess: Designing a Plan-based Multi-agent Story Generator. In *Proceedings of the 5th International Conference on Computational Creativity,* 347-350. http://computationalcreativity.net/iccc2014/wp-content/uploads/2014/06//15.7_Laclaustra.pdf

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By.* Chicago and London: University of Chicago Press.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

Leech, G., & Weisser, M. (2013). The spaadia annotation scheme. *http://martinweisser.org/publications/SPAADIA_Annotation_Scheme.pdf*

Len, Y., Portet, F., Labbé, C., & Qader, R. (2020, December). Controllable Neural Natural Language Generation: comparison of state-of-the-art control strategies. In *WebNLG+: 3rd Workshop on Natural Language Generation from the Semantic Web*. *Retrieved from https://hal.archives-ouvertes.fr/hal-03082599*

Levin, B. & Hovav, M. K. (2005). *Argument Realization*. Cambridge: Cambridge University Press.

Lim-Cheng, N. R., Fabia, G. I. G., Quebral, M. E. G., & Yu, M. T. (2014). Shed: An online diet counselling system. In *DLSU research congress*, 1-7. http://xsite.dlsu.edu.ph/conferences/dlsu_research_congress/2014/_pdf/proceedings/FNH-IV-031-ft.pdf

Lohr, S. (2012). The Age of Big Data. The New York Times. http://www.nytimes.com

Lu, J., & Ng, V. (2018). Event Coreference Resolution: A Survey of Two Decades of Research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 5479-5486. https://dl.acm.org/doi/10.5555/3304652.3304787

Mairesse, F., & Walker, M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, *37*(3), 455-488. https://doi.org/10.1162/COLI_a_00063

Mani, I., Pustejovsky, J., & Gaizauskas, R. (Eds.). (2005). *The language of time: a reader*. Oxford University Press on Demand.

Mann, W. C. (1980). *Toward a Speech Act Theory for Natural Language Processing*. University of Southern California Marina del Rey Information Sciences Inst. https://apps.dtic.mil/sti/citations/ADA087250

Mazidi, K., & Tarau, P. (2016). Automatic Question Generation: From NLU to NLG. In A. Micarelli, J. Stamper & K. Panourgia (Eds.), *Intelligent Tutoring Systems. ITS 2016. Lecture Notes in Computer Science.* Volume 9684 (pp. 23-33). Springer, Cham. https://doi.org/10.1007/978-3-319-39583-8_3

McIntyre, N., & Lapata, M. (2010). Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1562-1572. https://aclanthology.org/P10-1158.pdf

Mille, S., Ballesteros, M., Burga, A., Casamayor, G., & Wanner, L. (2016). Towards multilingual natural language generation within abstractive summarization. In A. Nebot, X. Binefa & R. López de Mántaras (Eds.). *Artificial Intelligence Research and Development*, 309-314. IOS Press.

Mitchell, M., Bohus, D., & Kamar, E. (2014). Crowdsourcing language generation templates for dialogue systems. In *Proceedings of the INLG and SIGDIAL 2014 Joint Session*, 16-24. https://aclanthology.org/W14-5003.pdf

Moryossef, A., Goldberg, Y., & Dagan, I. (2019). Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2267-2277. https://arxiv.org/abs/1904.03396

Mostafazadeh, N. (2017). *From Event to Story Understanding*. [Doctoral Dissertation, University of Rochester]. ProQuest Dissertations Publishing. https://www.proquest.com/openview/6b0aa13253abf0131ca50ddd37e73aa3/1?pq-origsite=gscholar&cbl=18750

Munigala, V., Mishra, A., Tamilselvam, S. G., Khare, S., Dasgupta, R., & Sankaran, A. (2018). Persuaide! An adaptive persuasive text generation system for fashion domain. In *Companion Proceedings of the The Web Conference 2018*, 335-342. https://doi.org/10.1145/3184558.3186345

Naranjo-Zeledón, L., Chacón-Rivas, M., Peral, J., & Ferrández, A. (2020). Phonological Proximity in Costa Rican Sign Language. *Electronics*, *9*(8), 1302. https://doi.org/10.3390/electronics9081302

Naranjo-Zeledón, L., Chacón-Rivas, M., Peral, J., & Ferrández, A. (2021). Architecture design of a reinforcement environment for learning sign languages. In *PeerJ*, In Press, Corrected Proof. F. I.

Narayan, S., Cohen, S. B., & Lapata, M. (2018). Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1, 1747-1759. https://doi.org/10.18653/v1/N18-1158

Narayan, S., & Gardent, C. (2020). Deep learning approaches to text production. *Synthesis Lectures on Human Language Technologies*, *13*(1), 1-199. https://doi.org/10.2200/S00979ED1V01Y201912HLT044

Naumann, F., & Herschel, M. (2010). An introduction to duplicate detection. *Synthesis Lectures on Data Management*, *2*(1), 1-87. https://doi.org/10.2200/S00262ED1V01Y201003DTM003

Ochs, E. (2000). Narrativa. In T. A. van Dijk (Coord.). *El discurso como estructura y proceso*. Barcelona, Gedisa.

Odebrecht, C., Burnard, L., Navarro-Colorado, B. Eder, M. & Schöch, C. (2019) The European Literary Text Collection (ELTeC). In *Digital Humanities Conference (DF2019)*. https://dev.clariah.nl/files/dh2019/boa/0715.html

Parikh, S. B., & Atrey, P. K. (2018). Media-rich fake news detection: A survey. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, 436-441. IEEE. https://doi.org/10.1109/MIPR.2018.00093

Paúls, B. G. (2006). Capítulo 2: Más allá de las palabras y la estructura: las categorías del componente pragmático. In E. Garayzábal (Ed.). *Lingüística clínica y logopedia*, (p. 81). Antonio Machado.

Paúls, B. G. & Fernández-Urquiza, M. (2015). Etiquetado pragmático de datos clínicos. *E-Aesla*, 1. https://cvc.cervantes.es/lengua/eaesla/pdf/01/33.pdf

Pérez-Beltrachini, L., Sayed, R., & Gardent, C. (2016). Building RDF content for data-to-text generation. In *The 26th International Conference on Computational Linguistics (COLING 2016)*, 1493-1502. https://hal.inria.fr/hal-01623800/

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training (preprint). https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research 21* (140), 1-67. https://arxiv.org/abs/1910.10683

Rai, S., & Chakraverty, S. (2020). A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, *53*(2), 1-37. https://doi.org/10.1145/3373265

Reiter, E., & Dale, R. (2000). *Building Natural Language Generation Systems* (Studies in Natural Language Processing). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519857

Reiter, E., Sripada, S., Hunter, J., Yu, J., & Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, *167*(1-2), 137-169. https://doi.org/10.1016/j.artint.2005.06.006

Resende de Mendonça, R., Felix de Brito, D., de Franco Rosa, F., dos Reis, J. C., & Bonacin, R. (2020). A framework for detecting intentions of criminal acts in social media: A case study on Twitter. *Information*, 11(3), 154. https://doi.org/10.3390/info11030154

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object Hallucination in Image Captioning. In *CoRR*. https://arxiv.org/abs/1809.02156

Saggion, H., Bott, S., & Rello, L. (2016). Simplifying words in context. Experiments with two lexical resources in Spanish. *Computer Speech & Language*, 35, 200-218. https://doi.org/10.1016/j.csl.2015.02.001

Santillan, M. C., & Azcarraga, A. P. (2020). Poem Generation using Transformers and Doc2Vec Embeddings. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1-7. IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207442

Saha, T., Saha, S., & Bhattacharyya, P. (2019). Tweet Act classification: A deep learning based classifier for recognizing speech acts in twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. https://doi.org/10.1109/IJCNN.2019.8851805

Sayers, D., Sousa-Silva, R., Höhn, S., Ahmedi, L., Allkivi-Metsoja, K., Anastasiou, D., ... & Yayilgan, S. Y. (2021). The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies. https://hal.archives-ouvertes.fr/hal-03230287

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge University Press.

Searle, J. R. (1985). *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: an inquiry into human knowledge structures*. New Jersey: Lawrence Erlbaum Associates.

Shah, P., Hakkani-Tur, D., Liu, B., & Tur, G. (2018). Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, 41-51. https://aclanthology.org/N18-3006.pdf

Shwartz, V., & Choi, Y. (2020). Do Neural Language Models Overcome Reporting Bias?. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6863-6870. https://doi.org/10.18653/v1/2020.coling-main.605

Sims, M., Park, J. H., & Bamman, D. (2019). Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3623-3634. https://doi.org/10.18653/v1/P19-1353

Smith, D., & Lieberman, H. (2013). Generating and interpreting referring expressions as belief state planning and plan recognition. In *Proceedings of the 14th European Workshop on Natural Language Generation* (pp. 61-71). https://aclanthology.org/W13-2107

Sprugnoli, R., & Tonelli, S. (2017). One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, *23*(4), 485-506. https://doi.org/10.1017/S1351324916000292

Subramanian, S., Rajeswar, S., Dutil, F., Pal, C., & Courville, A. (2017). Adversarial generation of natural language. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 241-251. https://aclanthology.org/W17-2629.pdf

Suraperwata, R. H., & Suyanto, S. (2020). Language Modeling for Journalistic Robot based on Generative Pretrained Transformer 2. In *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 1-6. IEEE. https://doi.org/10.1109/ICoICT49345.2020.9166359

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 1017-1024. https://icml.cc/2011/papers/524_icmlpaper.pdf

Talwar, S., Dhir, A., Kaur, P., Zafar, N., & Alrasheedy, M. (2019). Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services*, *51*, 72-82. https://doi.org/10.1016/j.jretconser.2019.05.026

Tian, Y., Galery, T., Dulcinati, G., Molimpakis, E., & Sun, C. (2017). Facebook sentiment: Reactions and emojis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 11-16. https://doi.org/10.18653/v1/W17-1102

Trott, S., Eppe, M., & Feldman, J. (2016). Recognizing intention from natural language: clarification dialog and construction grammar. In *Workshop on Communicating Intentions in Human–Robot Interaction*. http://publications.eppe.eu/Trott_etal_Language_Intention_Ro-Man2016.pdf

Veale, T., Shutova, E. & Klebanov, B. (2016) Metaphor: A Computational Perspective. *Synthesis Lectures on Human Language Technologies, 9*(1). Morgan and Claypool Publishers. https://doi.org/10.2200/S00694ED1V01Y201601HLT031

Wang, Z., Duan, Z., Zhang, H., Wang, C., Tian, L., Chen, B., & Zhou, M. (2020). Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 485-497. https://doi.org/10.18653/v1/2020.emnlp-main.35

Williams, S., & Reiter, E. (2008). Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, *14*(4), 495-525. https://doi.org/10.1017/S1351324908004725

Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., & Jiang, M. (2020). A Survey of Knowledge-Enhanced Text Generation. In *CoRR*. https://arxiv.org/abs/2010.04389

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? In *CoRR*. https://arxiv.org/abs/1905.07830

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019b). Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. https://arxiv.org/abs/1905.12616

Zhou, G., & Lampouras, G. (2020). Generating Safe Diversity in NLG via Imitation Learning. In *CoRR*. https://arxiv.org/abs/2004.14364

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia, Innovación y Universidades del Gobierno de España con el proyecto de investigación Generación Inteligente de Textos (INTEGER, RTI2018-094649-B-I00). Igualmente, por el proyecto Tecnologías del Lenguaje Humano para una Sociedad Inclusiva, Igualitaria, y Accesible (SIIA, PROMETEU/2018/089), dentro del programa PROMETEO de la Generalitat Valenciana. Por la COST Action *Distant Reading for European Literary History* (CA16204 - Distant-Reading). More information: http://www.distant-reading.net - https://www.cost.eu/actions/CA16204/