

[Inicio](#) > Nace 'AINA', el proyecto del Govern para garantizar el catalán en la era digital

Nace 'AINA', el proyecto del Govern para garantizar el catalán en la era digital

El proyecto está impulsado por el Departamento de Políticas Digitales de la Generalitat con la colaboración del BSC.



El proyecto AINA generará los recursos digitales y lingüísticos necesarios para facilitar el desarrollo de asistentes de voz, traductores automáticos o agentes conversacionales en catalán

El objetivo último es que la ciudadanía pueda participar en catalán en el mundo digital al mismo nivel que los hablantes de una lengua global, como el inglés, y evitar así la extinción digital de la lengua

El proyecto tiene un presupuesto de 13,5M €, que se financiarán con los fondos NextGenerationEU, y arranca con una aportación inicial de 250.000 € del Departamento de Políticas Digitales

El primer recurso generado es el 'corpus' del catalán para entrenar los algoritmos de Inteligencia Artificial (IA), el más grande creado hasta el momento, con 1.770 millones de metadatos asociados a palabras

El próximo paso será generar los modelos de la lengua, modelos del habla y modelos para la traducción utilizando redes neuronales multicapa

Dotar al catalán de recursos digitales y lingüísticos para que se convierta en una lengua competitiva en el mundo digital y asegurar así su supervivencia futura es el objetivo del proyecto AINA, que el consejero de Políticas Digitales y Administración Pública, Jordi Puigneró, ha presentado hoy en rueda de prensa, acompañado por la directora general de Sociedad Digital, Joana Barbany; el director asociado del Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC), Josep Maria Martorell, y la investigadora y colíder de la unidad de minería de datos del BSC, Marta Villegas, responsable del proyecto.

Impulsado por el Departamento de Políticas Digitales, con la colaboración del BSC, el proyecto AINA generará corpus y modelos informáticos de la lengua catalana para que las empresas que crean aplicaciones basadas en inteligencia artificial (IA), como asistentes de voz, traductores automáticos, agentes conversacionales, etc., puedan hacerlo fácilmente en catalán.

Presupuesto y alcance del proyecto

El proyecto AINA tiene un presupuesto global de 13,5 millones de euros para el período 2020 a 2024 y es uno de los proyectos priorizados por el Departamento de Políticas Digitales para ser financiado con los fondos europeos NextGenerationEU. De momento, arranca con una aportación inicial de 250.000 € que el Departamento de Políticas Digitales ha asignado al BSC para ampliar los corpus de la lengua catalana y así obtener modelos lingüísticos que abarquen las diferentes variantes y registros.

El BSC ya dispone de un primer corpus textual del catalán, consistente en 1.770 millones de palabras, reunidas en 95 millones de frases. Este corpus, el más grande que se ha hecho hasta ahora en lengua catalana, se ha obtenido a base de descargar textos de diferentes fuentes digitales (páginas web, archivos, etc), limpiarlos y borrar duplicidades.

La Generalitat de Catalunya ha facilitado toda la información de sus páginas web y del DOGC, lo que ha supuesto el 33% de todos los contenidos descargados, y han sido necesarias 2.000 horas de procesadores del superordenador MareNostrum para revisar los datos obtenidos, eliminar duplicidades y todo lo que no fueran propiamente oraciones en catalán.

Este primer corpus se llevó a cabo con la financiación del Plan de Impulso de las Tecnologías del Lenguaje, de la Vicepresidencia de Asuntos Económicos y Transformación Digital del Gobierno español.

Ahora, con el impulso del Departamento de Políticas Digitales, se crearán nuevos corpus para incorporar las diferentes variantes dialectales del catalán, diferentes registros lingüísticos (coloquial, literario, administrativo, etc.) y archivos de voz e imagen. La Corporación Catalana de Medios Audiovisuales aportará todo su repositorio documental.

Con toda esta información, el siguiente paso será entrenar redes neuronales multicapa para que "aprendan el catalán" y generen modelos de la lengua, modelos del habla y modelos para la traducción. Estos modelos también son muy costosos de hacer porque necesitan gran capacidad de cálculo (lo que se está construyendo en base al primer corpus textual utilizará 9.000 horas de GPU), y serán las bases sobre las que se podrán desarrollar aplicaciones basadas en IA, como asistentes de voz, predictores y correctores lingüísticos, chatbots, aplicaciones de resumen automático, búsquedas inteligentes, aplicaciones para el análisis de sentimientos o motores de traducción y subtítulos automática, entre otros.

Todos los modelos que creará el BSC estarán a disposición de todas aquellas empresas o entidades que las quieran utilizar, ya que se publicarán en abierto y con licencias permisivas.

El mundo digital, una oportunidad y un reto para el catalán

Esto debe permitir que el catalán dé un salto cualitativo y cuantitativo en el ecosistema digital. De hecho, el mundo digital es hoy una oportunidad y un reto para la lengua catalana. Actualmente, las tecnologías de voz y las aplicaciones e interfaces de voz para el acceso al mundo digital son estratégicas para el desarrollo pleno de la lengua en todos los sectores. La interacción entre las personas y la tecnología ha entrado en una nueva fase donde cada vez menos se hace a través de dispositivos como el teclado, el ratón o las pantallas táctiles, para dar paso a una nueva forma de interacción más natural a través de la voz y el habla. Y este hecho otorga especial relevancia a la lengua, que se convierte en uno de los principales vehículos de interacción.

Esta nueva interacción debe poder hacerse, también, en catalán. En este sentido, el Govern tiene el firme propósito de garantizar que la ciudadanía pueda hablar e interactuar en catalán en el mundo digital al mismo nivel que los hablantes de otras lenguas, como el inglés o el castellano, que, hoy por hoy, tienen garantizada su supervivencia digital, porque detrás han tenido Estados que han invertido para dotar de recursos suficientes en cuanto a las técnicas de aprendizaje y redes neuronales en Inteligencia Artificial.

Un estudio realizado en 2011 por la red europea de excelencia META-NET, gracias a más de 200 expertos en Tecnologías del Lenguaje, advierte que más de 20 lenguas europeas, entre ellas el catalán, se enfrentan a la extinción digital si no reciben más apoyo tecnológico en cuatro áreas: la traducción automática, la interacción con la voz, el análisis textual y la disponibilidad de recursos lingüísticos.

La IA y las Tecnologías del Lenguaje

Las Tecnologías del Lenguaje son aquellas que ya utilizamos en nuestro día a día cuando corregimos de forma automática un texto en el correo electrónico, utilizamos un buscador web en Internet, traducimos de forma automática una página web, damos órdenes de voz en el teléfono móvil, interactuamos con asistentes virtuales o seguimos las indicaciones del navegador GPS, entre otros. Y son las tecnologías que permitirán dialogar con ordenadores, aparatos domésticos e incluso con nuestro vehículo de manera natural.

Las nuevas tecnologías de Inteligencia Artificial y las Tecnologías del Lenguaje se basan en la aplicación de algoritmos sobre grandes conjuntos de datos de calidad, pero los conjuntos de datos sobre los que se entrenan los algoritmos son específicos para cada lengua.

En este sentido, las grandes multinacionales como Google, Apple y Microsoft utilizan los recursos digitales del inglés creados por la Agencia de Proyectos de Investigación Avanzada del Departamento de Defensa de EEUU (DARPA), que han sido la base lingüística de la IA a nivel mundial, ya que generar estos mismos recursos por parte de una empresa resultaría muy costoso.

Un proyecto estratégico

El proyecto AINA se enmarca en la estrategia digital del Govern, a través de dos iniciativas lideradas por el Departamento de Políticas Digitales: el Consejo de Dirección interdepartamental para la promoción del catalán en Internet y en las tecnologías digitales avanzadas, aprobado en diciembre de 2018, y la Estrategia de Inteligencia Artificial de Catalunya (Catalonia.AI), aprobado en febrero de 2020.

El primero cuenta con la participación de la Dirección General de Sociedad Digital, la Dirección General de Política Lingüística, la Dirección General de Medios de Comunicación, la Agencia de Ciberseguridad de

Catalunya, la Dirección General de Atención Ciudadana y la Fundación puntCAT, y tiene entre sus objetivos impulsar la presencia del catalán en los asistentes de voz. Por su parte, uno de los ejes prioritarios de la estrategia Catalonia.AI está vinculado al uso normalizado de la lengua catalana en las interfaces como elemento clave en el desarrollo de la IA, ya que la lengua es el elemento de comunicación de base para acceder, utilizar e interactuar con estas tecnologías.

AINA, un nombre donde confluyen lengua y tecnología

El proyecto ha sido bautizado con el nombre de AINA en homenaje a la filóloga menorquina Aina Moll, figura central de la promoción y la normalización del catalán y primera Directora General de Política Lingüística de la Generalidad de Cataluña de 1980 a 1988. Ella fue artífice del lanzamiento, en 1982, de la primera campaña institucional de sensibilización sobre el uso de la lengua 'el català, cosa de tots', que, con el popular personaje de “Norma, al capdavant”, tenía como objetivo concienciar a la sociedad sobre la situación sociolingüística del catalán. Un año después, se aprobaba la primera ley de normalización lingüística.

AINA contiene una referencia a la tecnología (AI: Inteligencia Artificial) que hará posible su normalización en el ámbito digital.

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

Source URL (retrieved on 17 Oct 2024 - 08:10): <https://www.bsc.es/es/noticias/noticias-del-bsc/nace-aina-el-proyecto-del-govern-para-garantizar-el-catal%C3%A1n-en-la-era-digital>