El proyecto AINA busca millones de voces para que la tecnología entienda y hable el catalán

AINA es un proyecto basado en tecnologías de datos e Inteligencia Artificial impulsado por la Vicepresidencia del Gobierno y el BSC para hacer posible que las máquinas entiendan y hablen el catalán.

La campaña 'Nuestra lengua es tu voz' invita a la ciudadanía de todas las variedades dialectales del catalán a enseñar su habla mediante la lectura de unos textos

La Vicepresidencia destinará este año 3M€ al proyecto AINA para, entre otros objetivos, crear el primer corpus de voz del catalán y generar la segunda versión, enriquecida, del corpus de texto

AINA está construyendo los recursos digitales del catalán necesarios para que cualquier empresa o entidad pueda utilizarlos para desarrollar soluciones o servicios como traductores, asistentes personales o agentes conversacionales en catalán

Bajo el lema 'Nuestra lengua es tu voz', el G*overn* de Catalunya lanza este 17 de febrero una campaña de captación de voces para generar el primer corpus o "diccionario" de voz del catalán. La campaña se inscribe en el proyecto AINA, impulsado por el Departamento de la Vicepresidencia y de Políticas Digitales y Territorio en colaboración con el Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS) para hacer que la tecnología entienda y hable el catalán.

El proyecto AINA está construyendo los corpus (conjuntos masivos de datos) y los modelos de la lengua catalana de forma que cualquier empresa u organización pueda utilizarlos para desarrollar sus soluciones o servicios específicos (traductores, asistentes personales, sintetizadores de voz, clasificadores de textos, etc.) para poder relacionarnos con las máquinas en catalán.

En definitiva, para enseñar catalán a las máquinas de modo que la ciudadanía pueda relacionarse con ellas y participar en el mundo digital en catalán al mismo nivel que los hablantes de una lengua global, como el inglés, y evitar así, la extinción digital de la lengua catalana.

La participación ciudadana en la campaña de recogida de voces 'Nuestra lengua es tu voz' se hará a través de la iniciativa de Common Voice de Mozilla por el catalán, una plataforma donde todo el mundo que lo quiera podrá leer y grabar un número ilimitado de frases (agrupadas de 5 en 5 pero sin límite) para ayudar a las máquinas a aprender cómo hablamos las personas.

Aunque esta colaboración se puede realizar de manera totalmente anónima y sin ningún registro previo, conocer los parámetros de género, edad y variante dialectal de la persona “donante” de voz facilita mucho el trabajo de clasificar los datos de voz obtenidos y, al mismo tiempo, permite saber si se está contemplando toda la diversidad lingüística del catalán. Por eso, la campaña anima a la ciudadanía a registrarse y crear un perfil en la plataforma para avanzar más rápidamente en los objetivos del proyecto AINA.

Enseñar catalán a las máquinas, todo un reto

“Enseñar” una lengua a las máquinas de modo que sean capaces no sólo de entendernos cuando les hablemos sino de respondernos de forma coherente a lo que les hemos preguntado o pedido es hoy un reto.

Si queremos que los ordenadores, asistentes de voz y otros sistemas informáticos hablen y entiendan el catalán, es necesario conseguir datos masivos de la lengua (en formato de texto y de voz). Estos datos se pasan a una red neuronal profunda que va aprendiendo cómo se combinan las palabras hasta generar un modelo de la lengua capaz, por ejemplo, de distinguir los diferentes significados de la palabra “banco” gracias a los diferentes contextos en los que se utiliza.

Para construir el corpus de la lengua (conjuntos de datos) que necesita una máquina, es necesario tener millones de textos y millones de horas de audio y vídeo en esa lengua y, además, que esos millones y millones de datos representen toda la riqueza de la lengua incluyendo, por ejemplo, grabaciones de voz de personas de diferentes géneros, distintas franjas de edad y diferentes variantes dialectales y registros.

Obtener este volumen y concreción de datos es especialmente difícil para las lenguas minoritarias a escala mundial como el catalán, ya que lenguas mayoritarias como el inglés tienen fácilmente a disposición toda esta información: sólo hace falta ir a Internet para encontrar millones y millones de textos , audios y vídeos en inglés.

Por este motivo, la campaña 'Nuestra lengua es tu voz' invita a la ciudadanía de habla catalana de todas las edades, géneros, condiciones y procedencias a “dar” su voz, con el objetivo de obtener unos contenidos de voz que capten toda la riqueza del catalán oral, con todos sus registros y variedades dialectales. Actualmente, el perfil de voz mayoritario en la plataforma Common Voice de Mozilla es la de hombres de entre 30 y 50 años hablantes de catalán central.

Crear el primer corpus de voz en catalán, hito de AINA para 2022

La creación de la primera versión del corpus de voz del catalán es uno de los principales hitos del proyecto AINA para este 2022. Este corpus se nutrirá de los contenidos obtenidos a través de la plataforma de Common Voice de Mozilla, pero también de la aportación del repositorio documental de la Corporación Catalana de Medios Audiovisuales (CCMA) o el Consejo del Audiovisual de Catalunya (CAC), entre otros.

En paralelo, el proyecto se marca también como objetivo este año la creación de la segunda versión del corpus de texto del catalán. A día de hoy, el proyecto dispone de un primer corpus textual, consistente en 1.770 millones de palabras reunidas en 95 millones de frases, que se ha obtenido en base a descargar textos de diferentes fuentes digitales en catalán (planas web, archivos, etc.), limpiarlos y borrar duplicidades. Ahora, se seguirá trabajando en este corpus de texto para generar una segunda versión mejorada y enriquecida que recoja todos los matices de la lengua escrita, ya sean variantes dialectales o registros lingüísticos, como el coloquial, el literario o el administrativo.

Otros objetivos destacados en la hoja de ruta del proyecto AINA para este 2022 son:

- Crear tres servicios lingüísticos básicos (de anonimización, de clasificación de documentos y de identificación de entidades y conceptos clave) necesarios para construir futuras aplicaciones y soluciones para el usuario final
- Crear modelos (cursos) de la lengua especializados en un ámbito concreto (por ejemplo, el de la salud o el jurídico) o en una tarea concreta (por ejemplo, traducción de textos), para ayudar a las máquinas a entender ya analizar mejor los matices y el contexto de las palabras en un texto o conversación.
- Crear un motor de traducción catalán-castellano para mejorar la calidad de los motores actualmente disponibles
- Implementar un caso de uso de impacto en la Administración Pública catalana para mostrar el potencial y la integración en aplicaciones reales de las diferentes piezas desarrolladas por el AINA.

3M€ de presupuesto para 2022 para un proyecto estratégico

Para hacer posible esta hoja de ruta, el Departamento de la Vicepresidencia y Políticas Digitales y Territorio destinará este año 3 M€ de su presupuesto al proyecto AINA mediante una subvención directa al BSC, que será el encargado de ejecutarlo. Con esta aportación, que multiplica por 12 el presupuesto destinado por la Generalitat en 2021, el *Govern* refuerza su firme apuesta por este proyecto estratégico que tiene como objetivo último 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, lenguas que, por ahora, tienen garantizada su supervivencia digital porque detrás han tenido Estados que han invertido en dotarlas de recursos suficientes en cuanto a las técnicas de aprendizaje y redes neuronales en Inteligencia Artificial.

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

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---