GAMES: Generación automática de metadato y contenido para medios y archivos en euskera

GAMES: Automatic generation of metadata and multimedia content for media and archives in Basque

Aitor Álvarez,[1] Ander González-Docasal,[12] Aitor García Pablos,[1] Elena Zotova,[1] Montse Cuadros,[1] Haritz Arzelus,[1] Alaitz Artolazabal,[3] Joxe Rojas,[3] Josu Azpillaga,[4] Iban Arantzabal[5]

[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián (España)
[2]Universidad de Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza (España)
[3]Tokikom, Bilbao Lanekintza 10, 48004 Bilbao (España)
[4]Codesyntax, Azitaingo Industrialdea 3, 20600 Eibar (España)
[5]Goiena, Otalora Lizentziaduna 31, 20500 Arrasate-Mondragón (España)
{aalvarez, agonzalezd, agarciap, ezotova, mcuadros, harzelus}@vicomtech.org
{aartolazabal, jrojas}@tokikom.eus, jazpillaga@codesyntax.com, iarantzabal@goiena.eus

Resumen: El ingente volumen de contenido multimedia obliga a los medios a contar con soluciones efectivas de metadatado que permitan su etiquetado y recuperación automática. En este contexto presentamos GAMES, una plataforma orientada al metadatado y generación de contenido en euskera. Además de la arquitectura, se describen los módulos tecnológicos y su evaluación sobre contenidos del dominio.
Palabras clave: Metadato, recuperación de la información, aprendizaje profundo.

Abstract: The increasing volume of multimedia content is pushing the media to seek for effective solutions for the automatic generation of metadata to facilitate the tagging, indexing and retrieval of contents. In this context, GAMES is presented as a platform focused to metadata and content generation in Basque. In addition to the main architecture, the technological modules and their evaluation are presented.
Keywords: Metadata, information retrieval, deep learning.

CEUR-WS.org/Vol-2968/paper3.pdf

1 Financiación y participantes

GAMES es un proyecto de Investigación y Desarrollo de carácter competitivo financiado por el Gobierno Vasco[1] a través de la convocatoria Hazitek de la Agencia Vasca de desarrollo empresarial Spri[2]. Su principal objetivo ha sido la implementación de la primera plataforma de extracción automática de metadatos y generación de material audiovisual sobre contenidos en euskera en el sector de medios de comunicación vascos.

El proyecto ha tenido una duración de 33 meses (abril 2018 - diciembre 2020), y ha contado con el siguiente consorcio: (1) Tokikom[3], red que gestiona un total de 66 medios locales en todos los soportes: papel, radio, televisión y digital; (2) Goiena[4], empresa de servicios de comunicación con un importante desarrollo en comunicación comarcal; (3) CodeSyntax[5], compañía experta en consultoría y servicios de Internet y TICs; y (4) Vicomtech[6], como centro de investigación aplicada experto en tecnologías del habla y del lenguaje basadas en Inteligencia Artificial (IA).

2 Estado del arte y motivación

El crecimiento imparable de contenido multimedia junto con los reciente avances en tecnologías IA están impulsando a los archivos y medios de comunicación a la incorporación de soluciones que permitan una identificación y descripción eficientes de sus contenidos a través de metadatos. Sin embargo, el euskera no está habitualmente soportado en estas soluciones y, de estar incluido, no alcanza la calidad esperada sobre contenidos media. Como solución más cercana cabe destacar GEPSA (San Vicente, Saralegi, y Zubia, 2021),

[1]https://www.euskadi.eus/
[2]https://www.spri.eus/es/ayudas/hazitek/
[3]https://tokikom.eus/
[4]https://goiena.eus/
[5]https://www.codesyntax.com/
[6]https://vicomtech.org/

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

una plataforma de seguimiento de medios escritos en euskera y castellano que incorpora tecnología IA de detección y clasificación de entidades nombradas (NERC), extracción de palabras clave, clasificación de temática y retos de la sociedad.

Con el ánimo de cubrir este hueco en el mercado, nació el proyecto GAMES, una solución para la generación automática de metadatos sobre contenidos multimedia (video, audio y texto) en euskera compuesta por 6 módulos tecnológicos IA.

3 *Arquitectura general*

La plataforma GAMES se ha diseñado y desarrollado para ser un servicio *back-end* fácilmente integrable en cualquier solución de terceros, permitiendo ofrecer un servicio de metadatado en euskera no intrusivo en estas soluciones. Está fundamentada en arquitectura REST[7] y está compuesta por una capa de servicio basada en la solución Traefik[8], que ofrece un proxy inverso y balanceador de peticiones. Este componente se comunica con las APIs de cada módulo tecnológico mediante servicios web basados en protocolo HTTP. Estos servicios permiten recuperar el estado de cada proceso, obtener un resultado o lanzar a procesar una tarea en cualquier módulo IA. Los resultados son devueltos en objetos JSON (JavaScript Object Notation), disponiendo cada módulo tecnológico de su formato específico de resultado. Estos módulos IA son fácilmente combinables para la generación de *pipelines* tecnológicos y esta orquestación se realiza desde el cliente. En la Figura 1 se muestra la arquitectura de la solución.

4 *Principales módulos IA*

Todos los módulos integrados en GAMES están basados en técnicas de Deep Learning y están entrenados para procesar contenidos multimedia (video, audio y texto) en euskera, una lengua aglutinante con unas propiedades lingüísticas singulares, una rica morfología y un orden de palabras relativamente libre.

4.1 Transcripción habla-texto

Este módulo permite convertir a texto los contenidos de video y audio en euskera a través de tecnología neuronal de transcripción del habla. El motor de reconocimiento de

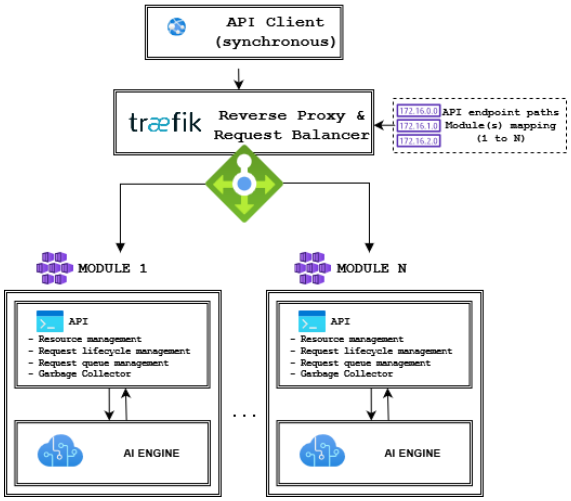

Figura 1: Arquitectura de GAMES.

habla, construido sobre la herramienta Kaldi (Povey et al., 2011), está compuesto por un modelo acústico híbrido DNN-HMM basado en redes neuronales retardadas factorizadas configuradas como se describe en (Alvarez et al., 2021) para los sistemas en castellano. El modelo acústico está entrenado con 645 horas compuestas por 482 horas del corpus de mintzai-ST (Etchegoyhen et al., 2021) del dominio del Parlamento Vasco y 163 horas de contenidos televisivos transcritos a través de la plataforma *Idazle*[9], una solución de Vicomtech para la subtitulación automática integrada en los flujos de trabajo de la televisión pública vasca EiTB. El modelo de lenguaje corresponde a un 3-grama entrenado con 27,7M de palabras compuestas principalmente por las transcripciones y noticias web extraídas de diarios digitales.

La salida cruda del reconocedor es enriquecida con capitalización y signos de puntuación con un modelo *Transformer* entrenado a partir del modelo BERTeus (Agerri et al., 2020). Con las marcas de tiempo y puntuaciones de confianza obtenidas por palabra, este módulo permite generar ficheros para indexación y búsqueda (XML, JSON), transcripción (TXT) o subtitulación (SRT, VTT).

4.2 Síntesis texto-habla

Este módulo convierte en habla cualquier texto de entrada en euskera. En GAMES se entrenó un modelo de síntesis neuronal basado en la arquitectura Tacotron-2 (Wang et al., 2017) para convertir el texto de entra-

[7] https://restfulapi.net/
[8] https://traefik.io/
[9] https://www.idazle.eus/

da en espectrogramas. Estos espectrogramas son posteriormente transformados en onda acústica a través del vocoder de Nvidia basado en el modelo neuronal generativo Waveglow (Prenger, Valle, y Catanzaro, 2019). El modelo Tacotron-2 fue entrenado con un *corpus* de 20,37 horas de diferentes locutores en euskera, posteriormente ajustado (*fine-tuning*) con las 3,44 horas de una sola locutora, utilizando como base un primer modelo entrenado con 21,25 horas en castellano. La síntesis del habla tiene dos aplicaciones principales en GAMES: la generación automática de *podcasts* y video-noticias en euskera.

4.3 NERC

Este módulo realiza la detección y clasificación de entidades nombradas sobre un texto de entrada en euskera. Para ello se entrenó un modelo de etiquetado secuencial basado en BERTeus (Agerri et al., 2020) usando el conjunto de datos de Egunkaria y las particiones estándar de *train* y *test* presentado en (Alegria et al., 2004). Este corpus distingue las categorías de locativo (LOC), organización (ORG), persona (PER) y otros (OTH). Los resultados de *micro-Precision* (miP), *Recall* (miR) y *F1-score* (miF) obtenidos en el subconjunto de *test* se presentan en la Tabla 1.

Tabla 1: Etiquetas anotadas y resultados en *train* y *test* del módulo NERC.

Categ.	Train	Test	miP	miR	miF1
LOC	1968	440	0.887	0.844	0.865
ORG	1937	394	0.816	0.816	0.816
PER	1497	382	0.903	0.949	0.925
OTH	294	53	0.367	0.367	0.367

Como puede observarse en la Tabla 1, los mejores resultados alcanzan un *micro-F1* de 0.925 para la categoría PER, mientras que la categoría OTH obtiene una mayor confusión, probablemente, por su baja representación en las particiones y por ser una categoría anotada con menor precisión.

4.4 Clasificación de noticias

Este componente permite clasificar en categorías del ámbito periodístico las noticias en euskara. El módulo IA está igualmente basado en un modelo *Transformer*, usando BERTeus (Agerri et al., 2020) como base (Devlin et al., 2018).

El entrenamiento y evaluación del sistema se realizó sobre un corpus compilado en el dominio de las noticias provistas por los medios del consorcio con la distribución presentada en la Tabla 2, en la que además se muestran los resultados de *micro-Precision*, *Recall* y *F1-score* por categoría. Destaca la categoría de deportes, mientras se observa mayor confusión entre categorías semánticamente próximas (e.g. sociedad y cultura). Las puntuaciones más bajas las reciben las categorías menos representadas en los datos de entrenamiento.

Tabla 2: Instancias anotadas de *train* y *test*, y resultados por categoría.

Categoría	Train	Test	miP	miR	miF1
Sociedad	2694	630	0.776	0.805	0.790
Deportes	1604	412	0.956	0.959	0.957
Cultura	1134	262	0.770	0.793	0.781
Política	514	134	0.843	0.767	0.803
Opinión	456	124	0.868	0.918	0.892
Euskara	274	64	0.729	0.797	0.761
Economía	203	48	0.706	0.393	0.505
Educación	202	15	0.706	0.500	0.585
Entorno	107	15	0.429	0.400	0.414

4.5 Resumen abstractivo de textos

Este módulo permite sintetizar de manera abreviada un texto de entrada; una línea tecnológica incipiente y que para el euskera supone un reto mayor por la falta de datos de entrenamiento o de recursos lingüísticos específicos. En GAMES se ha experimentado con un modelo *Encoder-Decoder* inicializado con tres modelos pre-entrenados: IXAmBERT (Agerri et al., 2020), RoBasquERTa (Suárez, Romary, y Sagot, 2020) y el modelo multilingüe mT5-small (Xue et al., 2020). Los modelos se entrenaron con un *corpus* de noticias en euskera de los medios del consorcio que consta de 73.773 documentos, seleccionando 2,000 para validación y otros 2,000 para la evaluación final. Dada la inexistencia de un resumen literal, se asumió como resumen el titular y la entradilla de cada noticia.

Los modelos han sido evaluados con la métrica ROUGE, donde ROUGE-N mide la superposición de n-gramas entre la referencia y la hipótesis, y ROUGE-L evalúa las subsecuencias comunes más largas. Como se observa en la Tabla 3 el modelo basado en IXAmBERT muestra el mejor resultado, probablemente por una representación más amplia del euskera en el modelo preentrenado.

4.6 Generación de vídeo-noticias

Este componente es un ejemplo de *pipeline* tecnológico en el que se integran diferentes

Tabla 3: Resultados de los modelos de resumen automático sobre la partición de *test*.

Model	Rouge-1	Rouge-2	Rouge-L
IXAmBERT	27.33	11.92	22.64
RoBasquERTa	22.12	8.76	17.69
mT5-small	19.69	7.49	15.66

módulos de la solución, y su función es la generación automática de video-noticias partiendo del título, noticia y una foto relacionada. La voz es generada con la síntesis de habla, la noticia puede ser sumarizada para generación de videos cortos y, además, se incluyen subtítulos sincronizados con el audio para sumar accesibilidad.

5 Conclusiones

En este trabajo se ha presentado el proyecto GAMES, una plataforma *back-end* para la generación automática de metadatos y contenido audiovisual en euskara. Esta solución está integrada en los flujos de trabajo de las empresas del consorcio y tendrá continuidad en un nuevo proyecto en el que se incorporarán nuevos módulos tecnológicos como la traducción automática basada en el activo *Itzuli*[10], biometría vocal y análisis de imagen.

Agradecimientos

Este trabajo contó con el apoyo del Departamento de Desarrollo Económico, Sostenibilidad y Medio Ambiente del Gobierno Vasco en el marco del proyecto GAMES (ZL-2020/00074).

Bibliografía

Agerri, R., I. S. Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, y E. Agirre. 2020. Give your text representation models some love: the case for basque. *arXiv preprint arXiv:2004.00033*.

Alegria, I., O. Arregi, I. Balza, N. Ezeiza, I. Fernandez, y R. Urizar. 2004. Design and development of a named entity recognizer for an agglutinative language. En *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition*.

Alvarez, A., H. Arzelus, I. G. Torre, y A. González-Docasal. 2021. The vicomtech speech transcription systems for the albayzín-rtve 2020 speech to text transcription challenge. En *Proceedings of IberSPEECH2020*.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Etchegoyhen, T., H. Arzelus, H. Gete Ugarte, A. Alvarez, A. González-Docasal, y E. Benites Fernandez. 2021. mintzai-st: Corpus and baselines for basque-spanish speech translation. En *Proceedings of IberSPEECH2020*.

Povey, D., A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, y others. 2011. The kaldi speech recognition toolkit. En *IEEE 2011 workshop on automatic speech recognition and understanding*, numero CONF. IEEE Signal Processing Society.

Prenger, R., R. Valle, y B. Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. En *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 3617–3621. IEEE.

San Vicente, I., X. Saralegi, y N. Zubia. 2021. GEPSA, a tool for monitoring social challenges in digital press. En *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, páginas 46–50, Kyiv, Abril. Association for Computational Linguistics.

Suárez, P. O., L. Romary, y B. Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv preprint arXiv:2006.06202*.

Wang, Y., R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, y others. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, y C. Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer.

[10] https://itzuli.vicomtech.org/es/api/