Miguel A. Alonso, Margarita Alonso-Ramos, Carlos Gómez-Rodríguez, David Vilares, Jesús Vilares (Eds.)

Pre-conference Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations (SEPLN-PD 2022)

co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022)

Actas precongreso del congreso anual de la Sociedad Española para el Procesamiento del Lenguaje Natural 2022: Proyectos y Demostraciones (SEPLN-PD 2022)

Celebrado conjuntamente con el congreso anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2022)

A Coruña, Spain, September 2022 / A Coruña, España, septiembre de 2022

https://sepln2022.grupolys.org/

Preface for SEPLN-PD 2022 Pre-conference Proceedings

The aim of SEPLN-PD is to offer, both the scientific community and companies, a forum to present, share and publicize real applications in the field of Natural Language Processing, as well as R&D projects.

There were 23 papers submitted for peer-review to SEPLN-PD with descriptions of projects and demonstrations. The papers were evaluated on the basis of originality, quality, relevance, and structure and presentation of the paper. Every paper was reviewed by 3 reviewers and the review comments were shared with the authors for incorporating the suggestions and comments. Finally, 22 papers were accepted for this volume, 9 for projects and 13 for demonstrations.

SEPLN-PD 2022 has been supported partially by the Vice-Rectorate for Science Policy, Research and Transfer of the University of A Coruña, with co-funding from the R&D Agreement on Strategic Actions for 2022 between the Department of Culture, Education and University of the Xunta de Galicia and the University of A Coruña.

Prefacio de las actas precongreso de SEPLN-PD 2022

El objetivo de SEPLN-PD es ofrecer, tanto a la comunidad científica como a las empresas, un foro para presentar, compartir y dar a conocer aplicaciones reales en el campo del Procesamiento del Lenguaje Natural, así como proyectos de I+D+i.

Se presentaron 23 trabajos con descripciones de proyectos y demostraciones para su revisión por pares por parte de SEPLN-PD. Los trabajos se evaluaron en función de su originalidad, calidad, relevancia, estructura y presentación. Cada artículo fue revisado por tres revisores y los comentarios de la revisión se compartieron con los autores para que incorporaran las sugerencias y comentarios. Finalmente, se aceptaron 22 trabajos para este volumen, 9 con descripciones de proyectos y 13 con demostraciones.

La realización de SEPLN-PD 2022 ha contado con la aportación del Vicerrectorado de Política Científica, Investigación y Transferencia de la Universidad de A Coruña, con cofinanciación del Convenio de Acciones Estratégicas I+D+i para 2022 entre la Consellería de Cultura, Educación y Universidad de la Xunta de Galicia y la Universidad de A Coruña.

Editors of the Pre-conference Proceedings /

Editores de las actas precongreso

Miguel A. Alonso (Universidade da Coruña and CITIC, Spain)

Margarita Alonso-Ramos (Universidade da Coruña and CITIC, Spain)

Carlos Gómez-Rodríguez (Universidade da Coruña and CITIC, Spain)

David Vilares (Universidade da Coruña and CITIC, Spain)

Jesús Vilares (Universidade da Coruña and CITIC, Spain)

Programme Committee / Comité de Programa

Miguel A. Alonso (*Chair*, Universidade da Coruña and CITIC, Spain)

Margarita Alonso-Ramos (Universidade da Coruña and CITIC, Spain)

Xabier Arregi Iparragirre: (UPV/EHU, Spain)

Manuel de Buenaga Rodríguez (Universidad de Alcalá, Spain)

Sylviane Cardey-Greenfield (Centre de recherche en linguistique et traitement automatique des langues, Lucien Tesnière. Besançon, France).

Irene Castellón Masalles (Universidad de Barcelona, Spain)

José Camacho Collados (Cardiff University, UK)

Arantza Díaz de Ilarraza (UPV/EHU, Spain)

Antonio Ferrández Rodríguez (Universidad de Alicante, Spain)

Alexander Gelbukh (Instituto Politécnico Nacional, Mexico)

Koldo Gojenola Galletebeita (UPV/EHU, Spain)

Carlos Gómez-Rodríguez (Universidade da Coruña and CITIC, Spain)

Xavier Gómez Guinovart (Universidad de Vigo, Spain)

José Miguel Goñi Menoyo (Universidad Politécnica de Madrid, Spain)

Ramón López-Cozar Delgado (Universidad de Granada, Spain)

Mariana Lara Neves (German Federal Institute for Risk Assessment, Germany)

Elena Lloret (Universidad de Alicante, Spain)

Bernardo Magnini (Fondazione Bruno Kessler, Italy)

Nuno J. Mamede (Computadores Investigação e Desenvolvimento em Lisboa, Portugal)

External Reviewers / Revisores externos

Laura Alonso Alemany (Universidad Nacional de Córdoba, Argentina)

Ana-Maria Bucur (University of Bucharest, Romania)

Óscar Araque Iborra (Universidad Politécnica de Madrid, Spain)

Marco Casavantes (INAOE, Mexico)

Riccardo Cervero (Universitat Politècnica de València, Spain)

Elisabet Comelles (Universitat de Barcelona, Spain)

Laritza Coello (INAOE, Mexico)

Víctor Manuel Darriba Bilbao (Universidade de Vigo, Spain)

Agustín Daniel Delgado Muñoz (UNED, Spain)

Andrés Duque (UNED, Spain)

Miguel Angel García Cumbreras (Universidad de Jaén, Spain)

José Antonio García-Díaz (Universidad de Murcia, Spain)

Juan Luis García Mendoza (INAOE, Mexico)

Delia Irazú Hernández-Farias (Universidad de Guanajuato, Mexico)

Salud María Jiménez-Zafra (Universidad de Jaén, Spain)

Arturo Montejo-Ráez (Universidad de Jaén, Spain)

Arantxa Otegi (UPV/EHU, Spain)

David Owen (Cardiff University, UK)

José M. Perea-Ortega (Universidad de Extremadura, Spain)

Flor-Miriam Plaza-del-Arco (Universidad de Jaén, Spain)

Francisco J. Ribadas-Pena (Universidade de Vigo, Spain)

Giulia Rizzi (Università degli studi di Milano-Bicocca, Italy)

Juan Fernando Sánchez Rada (Universidad Politécnica de Madrid, Spain)

David Vilares (Universidade da Coruña and CITIC, Spain)

Table of Contents / Índice

Projects / Proyectos

Demonstrations / Demostraciones

Organization / Organización

Sponsors / Patrocinadores

Projects
Proyectos

Explorando la generación de contenido *online* por el usuario y su influencia predictiva en la Calidad Relacional. Aplicación al sector hotelero de Andalucía

Exploring the generation of online content by users and its predictive influence on the Relational Quality. Application to the Andalusian hotel sector

M. J. Sánchez-Franco U. de Sevilla majesus@us.es

José A. Troyano U. de Sevilla troyano@us.es

Fermín Cruz U. de Sevilla fcruz@us.es

M. Alonso-Dos-Santos U. de Granada manuelalonso@ugr.es

Resumen: Nuestro proyecto se centra en la identificación de factores competitivos de los establecimientos hoteleros en Andalucía. Para ello usaremos técnicas de Procesamiento del Lenguaje Natural aplicadas a las opiniones *online* publicadas por los usuarios, a través de plataformas de infomediación como TripAdvisor. Estamos especialmente interesados en la identificación de métricas que permitan cuantificar el concepto de Calidad Relacional. El equipo de investigadores del proyecto está compuesto tanto por expertos en Marketing Relacional, como por expertos en Tecnologías del Lenguaje. Ambas visiones complementarias serán de gran ayuda, tanto en el diseño experimental como en la transferencia de resultados al sector turístico.
Palabras clave: Contenidos generados por usuarios, calidad relacional, tecnologías del lenguaje, sector turístico

Abstract: Our project focuses on the identification of competitive factors of hotel establishments in Andalusia. We will use Natural Language Processing techniques applied to the online reviews published by users through infomediation platforms such as TripAdvisor. We are especially interested in the identification of metrics that allow us to quantify the concept of Relational Quality. The research team of the project is composed of both experts in Relationship Marketing and experts in Language Technologies. Both complementary visions will be of great help both in the experimental design and in the transfer of results to the tourism sector.
Keywords: User generated content, relationship quality, language technologies, tourism sector

1 Introducción

El desarrollo de Internet ha supuesto un cambio de paradigma en la forma en la que se inspiran, se contratan, se organizan y se viven los viajes turísticos, y ha cambiado la manera en la que los turistas toman sus decisiones, debido al intercambio de opiniones y experiencias mediante el uso de aplicaciones, redes sociales y otros espacios en Internet (Consejería de Turismo, 2016). En particular, los viajeros buscan asesoramiento desinteresado antes de reservar un hotel, y consultan las opiniones emitidas y las valoraciones de los establecimientos hoteleros en plataformas de infomediación (Tripadvisor, Expedia, Yelp, Booking, ...). El entorno global en que participa el sector, y las transformaciones singulares debidas a la irrupción de las tecnologías de la información y la comunicación junto a los sistemas de información asociados, provocan un tránsito desde modelos tradicionales *offline* de búsqueda de servicios hacia otros modelos de consulta, reserva o compra basados en la publicación de contenidos generados por el propio usuario en sistemas de reservas *online* o de recomendación (Raguseo, Neirotti, y Paolucci, 2017). Las revisiones *online* son espontáneas, esclarecedoras, e incluso apasio-

nadas, fácilmente accesibles desde cualquier lugar y en cualquier momento (Guo, Barnes, y Jia, 2017) son recuerdos o reconstrucciones cognitivas de un viaje o una estancia que reducen significativa y aparentemente el riesgo potencial de la compra (Sparks, So, y Bradley, 2016). Si bien los contenidos compartidos que recrean las experiencias del huésped pueden estar intencionadamente distorsionados, la información se percibe como creíble y desinteresada, y se convierte en la clave que sustenta sus decisiones (Sánchez-Franco, Navarro-García, y Rondán-Cataluña, 2016).

En suma, las opiniones *online* creadas por el usuario desempeñan un papel crucial en la construcción de la reputación de los hoteles, y consecuentemente en la atracción de usuarios y su retención. El estudio de los contenidos creados es además necesario para intensificar la calidad relacional. Precisamente este concepto de calidad relacional es concebido como una propuesta adecuada para explicar y predecir el éxito de una relación entre el establecimiento hotelero y el usuario medido habitualmente a través de la lealtad, y basado en la teoría del compromiso en las relaciones (Parasuraman y Grewal, 2000).

Nuestro proyecto de investigación propone una aproximación basada en técnicas de Procesamiento del Lenguaje Natural (PLN) para extraer las cualidades de los productos hoteleros a partir de la experiencia comunicada del turista durante sus estancias hoteleras (por ejemplo, la localización del establecimiento, la calidad del servicio y el valor percibido, la atmósfera del hotel, ...), y por otro lado, predecir a partir de ellas el cumplimiento de las expectativas desde el punto de vista del usuario.

1.1 Experiencia del equipo investigador

En este apartado es de destacar el aspecto multidisciplinar del equipo de investigadores y colaboradores del proyecto. Se ha perseguido la integración de investigadores expertos y conocedores de los campos disciplinares asociados principalmente a las áreas de conocimiento de Comercialización e Investigación de Mercados, y Lenguajes y Sistemas Informáticos, vinculados a cuatro universidades distintas.

Es de esperar que, durante el desarrollo del proyecto, aparezcan sinergias entre investigadores procedentes de muy diversos campos, con experiencias tanto teóricas como experimentales en investigaciones relacionadas con el Marketing y en el Procesamiento del Lenguaje Natural.

1.2 La Calidad Relacional

La satisfacción de los clientes es habitualmente algo difícil de medir, pero es aspecto fundamental para cualquier empresa. En especial para mantener un alto grado de retención de clientes. En este proyecto nos basaremos en el concepto de calidad relacional (Fogg y Eckles, 2007) como un instrumento para cuantificar ese grado de satisfacción. Trabajaremos con tres escalas que dan estructura a este concepto:

- Satisfacción: sentimiento del cliente de que sus expectativas se han cumplido.
- Confianza: certeza del cliente de que la empresa cumplirá sus expectativas.
- Compromiso: deseo del cliente de mantener su relación con la empresa.

La investigación tradicional se basa en cuestionarios para evaluar las distintas escalas de la calidad relacional. Esto supone un alto coste, además de presentar distintos inconvenientes como son la introducción de sesgos a través de la redacción de las preguntas, el esfuerzo que se le exige al cliente, o la imposibilidad de adaptarse a sectores tan dinámicos como el turismo por la escasa frecuencia con la que se actualizan los propios cuestionarios.

El reto que nos planteamos en este proyecto es precisamente el de construir una alternativa, a este modelo clásico de evaluar la calidad relacional, usando textos publicados por los propios usuarios y automatizando el proceso mediante el uso de tecnologías del lenguaje.

2 Objetivos

El objetivo general del proyecto es validar empíricamente hipótesis del ámbito del Marketing Relacional, mediante la aplicación de técnicas de Procesamiento del Lenguaje Natural. En concreto se pretende:

- Aplicar un marco global de interpretación de los contenidos publicados en las plataformas de infomediación basado en la disciplina de Marketing Relacional. En particular centrándonos en la fase del

compromiso verdadero (Fogg y Eckles, 2007).

- Aplicar técnicas de Procesamiento del Lenguaje Natural para identificar métricas que permitan valorar, a través de la calidad relacional, la imagen de los establecimientos hoteleros y por extensión de Andalucía como destino turístico global.

No hemos encontrado trabajos previos que aborden el análisis de la calidad relacional desde la perspectiva del Procesamiento del Lenguaje Natural.

3 Metodología y resultados esperados

En este apartado, resumiremos nuestra aproximación metodológica y los resultados esperados, tanto a nivel experimental como de transferencia de conocimiento.

3.1 Metodología

La figura 1 muestra los elementos más significativos del proceso metodológico que pretendemos seguir en el proyecto. En el diagrama se reflejan los dos tipos de perfiles investigadores mediante un ordenador (experto en tecnologías del lenguaje) y una persona (experto en el dominio de marketing). Bajo cada una de las tres fases del proceso se indican, con estos iconos, qué perfiles de investigadores estarán involucrados.

En la primera fase de descubrimiento de datos y análisis exploratorio, ambos perfiles son necesarios. El principal resultado de esta fase serán las preguntas de investigación identificadas, que han sido consideradas como interesantes por parte de los expertos en el dominio, y valoradas como viables por los expertos en PLN.

La segunda fase es de corte totalmente experimental, y en ella los expertos en PLN convertirán las preguntas en tareas evaluables, y desarrollarán sistemas para resolver dichas tareas.

En la fase final de análisis, vuelven a participar ambos perfiles. Es el momento de extraer conclusiones e interpretar los resultados obtenidos.

El proceso implica una revisión continua, que se refleja con los dos bucles que permiten identificar nuevas preguntas de investigación interesantes en todo momento.

En cuanto a las técnicas a aplicar, aparte de las herramientas PLN para el desarrollo de *pipelines* clásicos, nuestra intención es basar los experimentos principalmente en modelos pre-entrenados de transformers. Para tareas semánticas como las que pretendemos resolver, son claramente la mejor alternativa.

3.2 Resultados experimentales

La evaluación de las distintas tareas PLN identificadas, nos permitirá medir la eficacia de los sistemas desarrollados a la hora de responder a las preguntas de investigación.

Hay algunas tareas definidas de antemano, en especial las que están relacionadas con la definición de una métrica para dar respaldo al concepto de calidad relacional. Pero más allá de eso, no nos cerramos a usar ninguna de las técnicas y herramientas disponibles en el área del PLN. Desde soluciones léxicas para extraer términos relevantes (Peñas et al., 2001), pasando por modelos basados en bolsas de palabra para tareas de clasificación (Alam y Awan, 2018), *pipelines* clásicos para tareas de extracción de información (HaCohen-Kerner, Miller, y Yigal, 2020), hasta las más recientes técnicas basadas en redes profundas (Li et al., 2020) o *transformers* (Munikar, Shakya, y Shrestha, 2019).

Los corpus que utilizaremos serán principalmente de desarrollo propio. De hecho, invertiremos un porcentaje importante de las horas de dedicación del proyecto a la recopilación y etiquetado de corpus específicos, que nos permitan entrenar y evaluar nuestros sistemas.

3.3 Resultados de transferencia

El hecho de que el equipo de investigación sea multidisciplinar, y que desde el principio participen en la definición de las tareas expertos en el dominio de marketing, hace que la trasferencia de conocimiento sea un paso natural en nuestro proyecto.

Estamos convencidos de que muchas de las tareas que identifiquemos pueden dar lugar a productos software que tengan un claro interés gerencial, tanto desde la perspectiva general del marketing relacional, como en el dominio específico de la gestión hotelera.

Nuestra intención es difundir nuestros resultados, tanto entre agentes públicos como privados, con idea de encontrar oportunidades de transferencia. La revisión de los resultados experimentales desde la perspectiva

Figura 1: Proceso metodológico

de los gestores, nos dará un punto de vista mucho más aplicado y nos permitirá refinar nuestros experimentos para orientarlos al desarrollo de productos de utilidad para el sector hotelero.

Agradecimientos

Financiado por el proyecto US-1380960 de la convocatoria de proyectos de I+D+i en el marco del Programa Operativo FEDER. Consejería de Transformación Económica, Industria, Conocimiento y Universidades. Junta de Andalucía.

Bibliografía

Alam, T. M. y M. J. Awan. 2018. Domain analysis of information extraction techniques. *Int. J. Multidiscip. Sci. Eng*, 9:1–9.

Consejería de Turismo, Regeneración, J. y. A. L. 2016. Plan general de turismo sostenible de andalucía horizonte 2020. Informe técnico, Junta de Andalucía.

Fogg, B. y D. Eckles. 2007. The behavior chain for online participation: how successful web services structure persuasion. En *International Conference on Persuasive Technology*, páginas 199–209. Springer.

Guo, Y., S. J. Barnes, y Q. Jia. 2017. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism management*, 59:467–483.

HaCohen-Kerner, Y., D. Miller, y Y. Yigal. 2020. The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, 15(5):e0232525.

Li, W., L. Zhu, Y. Shi, K. Guo, y E. Cambria. 2020. User reviews: Sentiment analysis using lexicon integrated two-channel cnn–lstm family models. *Applied Soft Computing*, 94:106435.

Munikar, M., S. Shakya, y A. Shrestha. 2019. Fine-grained sentiment classification using bert. En *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volumen 1, páginas 1–5. IEEE.

Parasuraman, A. y D. Grewal. 2000. The impact of technology on the quality-value-loyalty chain: a research agenda. *Journal of the academy of marketing science*, 28(1):168–174.

Peñas, A., F. Verdejo, J. Gonzalo, y others. 2001. Corpus-based terminology extraction applied to information access. En *Proceedings of corpus linguistics*, volumen 2001, página 458.

Raguseo, E., P. Neirotti, y E. Paolucci. 2017. How small hotels can drive value their way in infomediation. the case of 'italian hotels vs. otas and tripadvisor'. *Information & Management*, 54(6):745–756.

Sánchez-Franco, M. J., A. Navarro-García, y F. J. Rondán-Cataluña. 2016. Online customer service reviews in urban hotels: A data mining approach. *Psychology & Marketing*, 33(12):1174–1186.

Sparks, B. A., K. K. F. So, y G. L. Bradley. 2016. Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. *Tourism Management*, 53:74–85.

LIVING-LANG: Living Digital Entities by Human Language Technologies

LIVING-LANG: Tecnologías del lenguaje humano para entidades digitales vivas

L. Alfonso Ureña-López[1], Estela Saquete[2], María-Teresa Martín-Valdivia[1], Patricio Martínez Barco[2]
[1]Computer Science Department, SINAI, CEATIC
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{laurena, maite}@ujaen.es
[2]Department of Software and Computing Systems,
University of Alicante, Spain
{stela, patricio}@dlsi.ua.es

Abstract: This project pursues the dynamic modeling at a spatial-temporal level of digital entities in social media for predicting their behavior. Firstly, digital entities are modelled by identifying the characteristics of individuals through their language and footprint on the network. Then, the extraction of relationships between digital entities is one of the nuclear challenges of the project. The proposal pursues this objective on a semantic level, structuring the information into representations of knowledge suitable for logical processing. Considering the heterogeneous nature of the sources to be dealt with, filtering of information is fundamental, using metrics and quality criteria. This spatial-temporal characterization, together with screening processes, will allow us to study high-performance predictive strategies in the evolution of digital entities. This project is coordinated by the SINAI and GPLSI research groups.
Keywords: Natural Language Processing, Sentiment Analysis, Emotion Mining, Sentiment Enrichment.

Resumen: Este proyecto persigue el modelado dinámico a nivel espacio-temporal de las entidades digitales en los medios sociales para predecir su comportamiento. En primer lugar, se modelan las entidades digitales identificando las características de los individuos a través de su lenguaje y su huella en la red. A continuación, la extracción de relaciones entre entidades digitales es uno de los retos nucleares del proyecto. La propuesta persigue este objetivo a nivel semántico, estructurando la información en representaciones de conocimiento aptas para el procesamiento lógico. Teniendo en cuenta la heterogeneidad de las fuentes a tratar, es fundamental el filtrado de la información, a partir de métricas y criterios de calidad. Esta caracterización espacio-temporal, junto con los procesos de cribado, nos permitirá estudiar estrategias predictivas de alto rendimiento en la evolución de las entidades digitales. Este proyecto está coordinado por los grupos de investigación SINAI y GPLSI.
Palabras clave: Procesamiento Lenguaje Natural, Análisis Sentimientos, Minería Emociones, Enriquecimiento Semántico.

1 Introduction

Human language is the result of human social evolution, and thanks to it we can conceptualize reality, generating abstractions of it at different levels of complexity, which has given us a great capacity for reasoning. Language determines the way in which we relate to one another and, according to some authors, even how we think about and conceive the reality in which we live (Grace, 2016). In this way, language becomes a very valuable resource for the cognitive modelling of an individual as studied in psycholinguistics (Rommetveit, 2014), but also for understanding social interactions and communities in what is known as Computational Social Sciences (Wallach, 2016). This emerging discipline is fuelled by the arrival of great volumes of information,

primarily from the social web. We exchange a vast amount of information on the web. At the same time, our habits regarding information consumption are at a critical time of transformation. Digital media, as the preferred source of information, already threatens traditional written press. Young people choose social networks as their means of communication. Furthermore, this change in habit does not only affect the format or the means where the information is found, we are also changing the speed and type of content. According to Turkle (Arnd-Caddigan, 2015), we have gone from "I think, therefore I am" to "I share, therefore I am", reducing the quality of our "conversations" and, at the same time, creating the vague illusion of never being alone, referred to by the term "echo chamber". Technology also implies changes in the way we act. An example would be the way in which we read (Eshet, 2012) (Salyer, 2015). When we read digital media we "scan" rather than read. Short and simple content are almost the only element of consumption (titles, captions, highlighted sentences...) (Hayles, 2010), and we are often carried away by our emotions when we decide what to read or where we read it. There are new challenges in this new digital paradigm that must dealt with, derived from our inability to adapt to this new scenario and often resulting in the deterioration of our cognitive abilities (Bauerlein, 2008). This enormous amount of information and digital connectivity entails the development of a technology capable of modelling the new paradigm, as well as determining the relationships that arise, their evolution in time and the ability to interfere with or predict their behaviour in the future.

Our previous project set out to identify digital entities, considered to be any entity in the real world (people, companies, organisations, tourist attractions...) with presence in the digital world and from which we can obtain a complete profile from their activity in such an environment. This profile is generated by processing unstructured content (web pages, articles, comments...) using human language technologies. However, the present "digital" situation requires us to go a step further and attempt to answer the following questions: a) How can we ensure the social contextualization of these entities, and model situations that change from day to day? b) How can we deduce new semantic relationships between entities? c) How can we guarantee that captured knowledge is real and contrasted by multiple sources? d) How can we guarantee the coexistence of knowledge in the long term?

This project aims to take this several steps further. In this way and thanks to these characteristics, we can establish relationships between the entities from a social and human perspective, improve the comprehension of the content exchanged, create new knowledge in the analysis of these relational structures and eventually, characterise and predict these networks between entities on a human language level by using temporal dimension, behaviours or phenomena.

This ability to understand language, model it and analyse its changes in time will allow us to face new challenges in the digital society in which we live. By measuring the veracity and credibility of the relationships extracted, we can confront phenomena such as fake news, defined as a deliberate distortion of a reality with the objective of creating and shaping public opinion and influencing social attitudes. Thanks to this project, tasks such as fact-checking, the automatic detection of ideological or confirmation bias, and the detection of clickbaits can be handled automatically, as well as other post-truth problems that are difficult to detect and treat because of their "viral" content. Furthermore, language modelling and new knowledge about these dynamic relationships and their evolution over time will allow us, through the application of diverse techniques, to identify new characteristics and make inferences that provide predictions of future behaviours of digital entities. These predictions can be used for the early detection of problems associated with violence, mental health problems such as suicides, inappropriate behaviours and other security and health risks.

2 Objetives

The project started in 2018 and will be completed in 2022, and it involves a number of specific challenges and objectives of the overall project in the field of NLP research, which are detailed below:

OBJ1. *Generation of the human language models used by digital entities* through recognition of their primary characteristics (linguistic, cognitive, social, cultural and emo-

tional) and independent of the domains and scenarios in which they act.

OBJ2. *Use of the knowledge generated by digital entities and discovery of the semantic relationships between them.* All available sources of information (unstructured, structured and open linked data), extraction mechanisms, identity enrichment, and other inference mechanisms will be taken into account. This will enable the integration of information related to an identity, determining the roles and properties associated to a space-time framework. It also enables the definition of relationships between identities using dynamic aspects such as context, temporary nature or importance.

OBJ3. *Use of knowledge of relationships to determine the coherence, quality and contrast of the semantic relationships extracted.* For this, we will use veracity assessment techniques, emotion analysis and subjectivity, as well as the detection of bias in the information to guarantee and contrast the information that arises from the relationship.

OBJ4. *Prediction of future behaviour of digital entities* by discovering potential future semantic relationships between them, through the analysis of pre-existing networks and based on previously detected relationships.

In summary, this project contributes to the Spanish national Plan for the Promotion of Human Language Technologies, which has aimed to promote the development of natural language processing since 2015.

To achieve the above global objective and the specific objectives of the global project, the coordination of two complementary subprojects is proposed, whose specific objectives will cover the global objectives proposed, and whose reunification will provide the added value sought by the coordination.

3 Results and conclusions

This section describes the most significant results of the project.

Results regarding OBJ1: In this project, the domains to be worked on are mainly health and education, as well as the following scenarios: fake news, knowledge extraction, violence and hate speech (del Arco et al., 2021; del Arco et al., 2020), studying the characteristics of the different scenarios in order to model the language in each of them. Resources associated with the different scenarios and domains defined have been created and used to train machine learning systems.

Results regarding OBJ2: The project has worked on various techniques for knowledge extraction in the different domains and scenarios defined, as well as on the organisation of workshops such as eHealth-KD 2020 to model human language in health documents in Spanish (Piad-Morffis et al., 2020). In addition, knowledge discovery techniques are being applied to the health domain (López-Úbeda et al., 2021; **?**). In addition, work has been done on the discovery of temporal information to enrich the entities by automatically extracting timelines from the documents and generating summaries from these timelines (Barros et al., 2019).

Results regarding OBJ3: In relation to this objective, a systematic study of the state of the art in this matter has been carried out(Saquete et al., 2020) and, based on this study, work has been done to determine both the veracity of the news and its parts and to study the detection of satire, achieving an architecture capable of determining 74 % accuracy (Bonet-Jover et al., 2021). Within this task, progress has been made in the detection of incongruent headlines as well as in fact-checking tasks, as part of the disinformation detection architecture (Sepúlveda-Torres et al., 2021). In addition, work has been done on emotion detection (Canales et al., 2020) (Canales et al., 2019) and negation (Jiménez-Zafra et al., 2020).

Results regarding OBJ4: Regarding this objective, the project focused on the discovery of virality patterns, applying opinion mining techniques that enable us to structure the information based on the polarity of the messages and the emotions they contain (Saquete et al., 2022). After transforming the information from an unstructured textual representation to a structured one, association rules mining were used, concluding that messages with a high-negative polarity and a very high emotional charge, especially emotions that have intensified with the COVID-19 pandemic, such as fear, sadness, anger and surprise are more likely to go viral in social media.

All publications related to the project can be found on the project website[1].

[1] https://livinglang.gplsi.es/

Acknowledgements

This research work has been supported by LIVING-LANG Project (RTI2018-094653-B-C21/C22), funded by FEDER/Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación. It is a coordinated project with SINAI and GPLSI as participating research groups.

References

Arnd-Caddigan, M. 2015. Sherry turkle: Alone together: Why we expect more from technology and less from each other.

Barros, C., E. Lloret, E. Saquete, and B. Navarro-Colorado. 2019. Natsum: Narrative abstractive summarization through cross-document timeline generation. *Inform. Proc. Manag.*, 56(5):1775–1793.

Bauerlein, M. 2008. *The Dumbest Generation—How the Digital Age Stupefies Young Americans and Jeopardizes Our Future*. Jeremy P. Tarcher / Penguin, New York.

Bonet-Jover, A., A. Piad-Morffis, E. Saquete, P. Martínez-Barco, and M. Á. G. Cumbreras. 2021. Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. *Expert Syst. Appl.*, 169:114340.

Canales, L., W. Daelemans, E. Boldrini, and P. Martínez-Barco. 2019. Emolabel: Semi-automatic methodology for emotion annotation of social media text. *IEEE Trans. Affect. Comput*, early access:1–1.

Canales, L., C. Strapparava, E. Boldrini, and P. Martínez-Barco. 2020. Intensional learning to efficiently build up automatically annotated emotion corpora. *IEEE Trans. Affect. Comput.*, 11(2):335–347.

del Arco, F. M. P., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2020. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Trans. Internet Techn.*, 20(2):12:1–12:19.

del Arco, F. M. P., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Syst. Appl.*, 166:114120.

Eshet, Y. 2012. Thinking in the digital era: A revised model for digital literacy. *Issues in informing science and information technology*, 9(2):267–276.

Grace, G. W. 2016. *The linguistic construction of reality*. Routledge.

Hayles, N. K. 2010. How we read: Close, hyper, machine. *ADE*, 150(18):62–79.

Jiménez-Zafra, S. M., R. Morante, M. T. Martín-Valdivia, and L. A. Ureña-López. 2020. Corpora annotated with negation: An overview. *Comput. Linguistics*, 46(1):1–52.

López-Úbeda, P., M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. U. López, and M. T. Martín-Valdivia. 2021. Automatic medical protocol classification using machine learning approaches. *Comput. Methods Programs Biomed.*, 200:105939.

Piad-Morffis, A., Y. Gutiérrez, Y. Almeida-Cruz, and R. Muñoz. 2020. A computational ecosystem to support ehealth knowledge discovery technologies in spanish. *J. Biomed. Informatics*, 109:103517.

Rommetveit, R. 2014. *Words, Meaning, and Messages: Theory and Experiments in Psycholinguistics*. Academic Press.

Salyer, D. 2015. Reading the web: Internet guided reading with young children. *The Reading Teacher*, 69(1):35–39.

Saquete, E., D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert Syst. Appl.*, 141(C), mar.

Saquete, E., J. Zubcoff, Y. Gutiérrez, P. Martínez-Barco, and J. Fernández. 2022. Why are some social-media contents more popular than others? opinion and association rules mining applied to virality patterns discovery. *Expert Syst. Appl.*, 197:116676.

Sepúlveda-Torres, R., M. E. Vicente, E. Saquete, E. Lloret, and M. Palomar. 2021. Headlinestancechecker: Exploiting summarization to detect headline disinformation. *J. Web Semant.*, 71:100660.

Wallach, H. 2016. Computational social science. *Comput. Soc. Sci.*, 307.

ESAN: Automating medical scribing in Spanish

ESAN: Automatización de la toma de notas clínicas

Naiara Perez,[1] Aitor Álvarez,[1] Arantza del Pozo,[1] Andrés Arbona,[2] Oihane Ibarrola,[2] Marta Suarez,[2] Pedro de la Peña Tejada,[3] Itziar Cuenca[3]
[1]Vicomtech Foundation, {nperez,aalvarez,adelpozo}@vicomtech.org
[2]Biokeralty Research Institute AIE, {andres.arbona,oihane.ibarrola,marta.suarez}@keralty.com
[3]Instituto Ibermática de Innovación, {pm.delapena,ia.cuenca}@ibermatica.com

Abstract: The ESAN research project aims at developing a Spanish digital scribe that reduces the administrative workload of clinicians and enhances the quality of the data collected in the medical records by automatically transcribing and structuring doctor-patient conversations. At present, the main goal of the consortium consists in collecting and annotating the data necessary for training and adapting speech and natural language processing models based on deep learning architectures.
Keywords: clinical data, EHR, speech recognition, data mining

Resumen: El proyecto ESAN pretende desarrollar un asistente digital que reduzca la carga administrativa de los clínicos y mejore la calidad de los datos recogidos en sus historias mediante la transcripción y estructuración automática de las conversaciones entre médicos y pacientes. En esta actuación inicial del consorcio, el objetivo principal es recopilar y anotar datos para entrenar o adaptar modelos de procesamiento del habla y del lenguaje natural basados en aprendizaje profundo.
Palabras clave: datos clínicos, HCE, reconocimiento del habla, minería de dato

1 Introduction

The past few decades have seen a worldwide, steady growth in the adoption of electronic health record (EHR) systems, with the ultimate goal of improving the efficiency and quality of the provided care. In spite of their many virtues, EHRs have also increased the administrative workload of healthcare professionals, to the point of having been identified as a direct cause of burnout and lack of meaningful doctor-patient eye contact (Sinsky et al., 2016, among others).

Meanwhile, the accumulation of massive amounts of digitised health records in the era of Big Data has boosted the pursuit of public policies aimed at accelerating the advent of new healthcare paradigms such as personalised medicine. Yet Big Data is no more profitable than the quality of the data allows. Currently, much of the data collected in EHRs is in the form of free text written in haste. It makes irregular use of grammar, standard medical terminology, and of the EHR structure itself. It may omit information that is not of evident immediate value. Moreover, it is barely codified (if at all), all of which hinders its automated exploitation.

More recently, the major and rapid advances of Deep Learning have prompted a surge of interest in the application of artificial intelligence to medical conversations, so much so that several tech giants have recently launched a workshop exclusively focused on this research topic (Shivade et al., 2021).

In this context we present the ESAN project (from "EStructuración de conversaciones en el ámbito SANitario" or *Structuring conversations in the health sector* in Spanish, but also "esan" or *say*, *tell* in Basque). ESAN is the first step of a joint, long-term effort towards alleviating the above introduced problems through the research and development of a Spanish digital scribe.

ESAN is partially funded by the Basque Government through the Elkartek 2021 program of the SPRI Group and will run from 04/2021 to 12/2023. The consortium includes the Vicomtech[1] research centre, Grupo Keralty's R&D division BioKeralty Research Institute[2], and Grupo Ibermática's R&D business unit and project leader Instituto Ibermática de Innovación or i3B[3].

[1]https://www.vicomtech.org
[2]https://biokeralty.com
[3]https://ibermatica.com/en/innovacion/

2 Goals and expected results

The long-term, main technical objective of the ESAN consortium is to develop a Spanish digital scribe. A digital scribe is, in short, a program capable of documenting the encounter between a patient and their doctor or nurse. It involves Automatic Speech Recognition (ASR) to transcribe the conversations, and Natural Language Processing (NLP) to understand and transform those transcripts as necessary (e.g., extract relevant information and classify it into EHR sections).

At this early stage, the identified challenges of the project (see §3) point primarily to the need for problem-specific data and the lack thereof. Thus, the focus of this initial venture of the ESAN consortium is set on building a new corpus. The expected results of this line of work are:

- 150 hours of anonymised recordings (∼1K encounters) in 4 medical specialities, along with their manual, enriched transcripts and the corresponding written medical notes, all in Spanish.
- Guidelines for the annotation of the dialogues regarding the information extraction (IE) and classification tasks, as well as the manual annotations resulting from their application.

Second, we plan to train benchmark models for enriched ASR and IE adapted to the application scenarios of ESAN, exploiting primarily the aforesaid corpus. The specific expected results in this regard are:

- Robust neural models for enriched ASR adapted to face-to-face clinician-patient conversations in Spanish, including automatic capitalisation and punctuation, and supervised diarisation.
- Initial neural IE models to transform the dialogue transcripts into structured data that can be fed to an EHR.

The third and final major goal is to flesh out the next steps based on quantitative and qualitative evaluations of the obtained technology. The expected final outcome is then:

- A road map towards productisation, taking into account the performance of the ASR and NLP models and other aspects that are outside the current scope (e.g., usability, communication standards).

3 Challenges

The challenges faced by the ESAN research project are twofold because it must overcome major ethical and legal obstacles in addition to the scientific and technological.

Conversations between patients and their doctors are among the most sensitive pieces of information conceivable. Voice recordings alone qualify largely as personal data according to the many policies that we are subject to, from the international (e.g., the GDPR of the European Union) to the local (e.g., ethics committees). This means that there is no public dataset that we can leverage, and that we must overcome these ethical and legal barriers in order to collect it ourselves.

Regarding the scientific and technological challenges, at this stage of the project, the difficulties of developing a Spanish digital scribe stem also from the nature of the data to be processed, in all its facets:

Genre The input to the scribe is spontaneous speech produced in the context of a dialogue between two or more people. Current ASR technology still struggles in this scenario due to *a)* the difficulty to obtain quality audio, where all the interlocutors are recorded with optimal volume and energy and *b)* linguistic phenomena inherent to spontaneous speech (overlapping, false starts, repetitions, etc.). Human-human dialogues are a serious challenge for NLP systems too for similar reasons. For example, questions may go unanswered or be answered at a later point in the conversation, or relevant information may be transmitted through non-verbal means.

Domain Along with the genre, the highly specialised application domain constitutes the key defining challenge of ESAN. Out-of-the-box, generic ASR and NLP solutions are not viable here simply because they are not prepared to deal with the specialised vocabulary and the extraction or classification targets of the clinical domain. Further, building new solutions and resources requires at least the guidance of expert knowledge.

Register Conversations in consultations present the added difficulty that doctors tend to address their patients in technical terms, while the patients may be less formal and employ more colloquialisms. From the perspective of the technologies involved in the project, this discursive gap is translated into an increased range of vocabulary and semantic

complexity that the automated systems must recognise and understand.

Language The ESAN consortium expects to gather data in—and, ultimately, be able to process—multiple varieties of the Spanish language, including the Colombian. The differences in pronunciation and vocabulary with standard Castilian Spanish pose an added important challenge both to ASR and NLP technologies and serve only to aggravate the problems listed above.

To these concerns, we must add the fact that the errors of the enriched ASR modules are cascaded down the pipeline to the text processing modules. In addition, it is noteworthy that the health sector is most demanding and intolerant of errors, due to the gravity of the consequences that could follow from decisions based on inaccurate data.

4 Approach

4.1 Audio collection

This is the most crucial yet sensitive task of the project. The strategy involves recording real doctor-patient encounters of at least 4 specialities in a private hospital.

Measures have been taken towards minimising the impact that this activity might have on the doctors' primary job, such as training dedicated staff responsible for informing the patients about ESAN and asking for their consent in the waiting rooms.

In addition, we have already tested a variety of commercial microphone arrays both in terms of quality and user-friendliness, so as to ensure their suitability before starting the audio collection campaign. We will make the recordings with the audio software Audacity[4] in PCM WAV format at 44.1kHz and 24 bits.

4.2 Enriched ASR

The ASR models will be trained with the 150 hours of acoustic corpus to be recorded during the project.

This corpus will be manually annotated through the Transcriber 1.5.1 tool[5] with spoken literal transcriptions and speaker turn information. The annotation process will be assisted by ASR technology, which will be iteratively enhanced as new annotated audio sets are generated: the first set of drafts to be post-edited will be created with generic Castilian Spanish recognition models; once each set is manually corrected, new adapted versions of the ASR models will be trained incrementally. This process will be repeated until all hours are manually revised.

The ASR models will be built using the *nnet3* DNN setup of the Kaldi recognition toolkit[6] following our previous approach based on CNN layers and a TDNN-F network (Álvarez et al., 2022). The ASR engine will also include n-gram language models for decoding and re-scoring the initial lattices. The transcriptions will be enriched with capitalisation and punctuation marks generated by the BERT-based AutoPunct system (González-Docasal et al., 2021), which will be also adapted to the domain. Finally, new speaker diarisation models will be trained for the Kaldi X-Vectors-based system (Snyder et al., 2018) to be developed.

4.3 From transcripts to the EHR

The corpus of transcribed dialogues will be manually annotated at a later stage to serve as training and testing data of IE and classification models.

The annotation policy, whose precise definition is another key task of ESAN, will be built around related efforts (Shafran et al., 2020; Magnini et al., 2021, among others). It will define guidelines for the annotation of information at different levels, including mentions of signs and symptoms, disorders, and medications, as well as related attributes (presence or absence, location, dosage, etc.).

The models for the automatic detection and classification of this information will be based on the ubiquitous Transformers architecture (Vaswani et al., 2017). We plan on exploiting the latest neural language models for Spanish and the biomedical domain (Carrino et al., 2021). This line of work will also profit from previous work of consortium members on clinical IE (i.a., Perez et al. (2019), Lima-López et al. (2020)).

4.4 Validation

Each of the above-mentioned technological modules will be assessed in isolation with gold standard data and the appropriate metrics (e.g., WER, F1-score) during their development. We will also measure the impact of the errors propagated from the ASR down the processing pipeline.

[4] https://www.audacityteam.org

[5] http://trans.sourceforge.net

[6] https://kaldi-asr.org

Equally, if not more, important in order to flesh out the productisation road map, we will carry out a qualitative evaluation of the technology as an integrated solution prototype. To that end, we intend to devise an initial integration of all the core modules, and to develop a graphic user interface for demonstration and testing purposes, through which expert testers will be able to identify potential areas of improvement.

5 Conclusions

We have presented the ESAN project, whose aim is to develop a Spanish digital scribe that reduces the administrative workload of clinicians and enhances the quality of the data collected in the EHRs.

The envisaged solution consists of a neural enriched ASR component followed by IE and classification modules, based too on neural architectures. To that end, the consortium will devote significant resources and effort to gathering the data necessary for adapting this technology to the challenging domain that doctor-patient face-to-face conversations pose. This emphasis on data collection and domain adaptation sets ESAN apart from related projects (Vivancos-Vicente et al., 2021, among others).

Acknowledgements

ESAN is partially funded by the Basque Business Development Agency, SPRI, under the grant agreement KK-2021/00117.

References

Álvarez, A., H. Arzelus, I. G. Torre, and A. González-Docasal. 2022. Evaluating novel speech transcription architectures on the Spanish RTVE2020 Database. *Appl. Sci.*, 12(4).

Carrino, C. P., J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas. 2021. Biomedical and clinical Language Models for Spanish. arXiv:2109.03570 [cs.CL].

González-Docasal, A., A. García-Pablos, H. Arzelus, and A. Álvarez. 2021. AutoPunct: A BERT-based automatic punctuation and capitalisation system for Spanish and Basque. *Proces. de Leng. Nat.*, 67:59–68.

Lima-López, S., N. Perez, M. Cuadros, and G. Rigau. 2020. NUBes: A corpus of negation and uncertainty in Spanish clinical texts. In *Proc. of LREC 2020*, pages 5772–5781.

Magnini, B., B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2021. The E3C project: Collection and annotation of a multilingual Corpus of Clinical Cases. In *Proc. of CLiC-it 2020*, pages 1–7.

Perez, N., P. Accuosto, À. Bravo, M. Cuadros, E. Martínez-Garcia, H. Saggion, and G. Rigau. 2019. Cross-lingual semantic annotation of biomedical literature: experiments in Spanish and English. *Bioinformatics*, 36(6):1872–1880.

Shafran, I., N. Du, L. Tran, A. Perry, L. Keyes, M. Knichel, A. Domin, L. Huang, Y.-h. Chen, G. Li, M. Wang, L. El Shafey, H. Soltau, and J. S. Paul. 2020. The Medical Scribe: Corpus development and model performance analyses. In *Proc. of LREC 2020*, pages 2036–2044.

Shivade, C., R. Gangadharaiah, S. Gella, S. Konam, S. Yuan, Y. Zhang, P. Bhatia, and B. Wallace, editors. 2021. *Proc. of the Second Workshop on NLPMC*.

Sinsky, C., L. Colligan, L. Li, M. Prgomet, S. Reynolds, L. Goeders, J. Westbrook, M. Tutty, and G. Blike. 2016. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med*, 165(11):753–760.

Snyder, D., D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. 2018. X-Vectors: Robust DNN embeddings for speaker recognition. In *Proc. of ICASSP 2018*, pages 5329–5333.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proc. of NIPS 2017*, pages 6000–6010.

Vivancos-Vicente, P. J., J. A. García-Díaz, J. S. Castejón-Garrido, and R. Valencia-García. 2021. ISMR - Sistema basado en Deep Learning para la transcripción y extracción de conocimiento en entrevistas médico-paciente. In *Proc. of SEPLN-PD 2021*, pages 1–4.

InLIFE. Tecnologías del Lenguaje aplicadas al envejecimiento activo

InLIFE. Language Technologies applied to active aging

Miguel Ángel García-Cumbreras,[1] **Fernando Martínez-Santiago,**[1]
Luis Alfonso Ureña-López,[1] **María Teresa Martín-Valdivia,**[1]
Arturo Montejo-Ráez,[1] **María Rosario García Viedma,**[2]
Manuel García-Vega,[1] **Manuel Carlos Dıaz-Galiano,**[1]
María Dolores Molina-González,[1] **Salud María Jiménez-Zafra,**[1]
Flor Miriam Plaza-del-Arco[1]
[1]Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
[2]Department of Psychology
Universidad de Jaén. Campus Las Lagunillas, E-23071, Jaén, Spain
{magc, dofer, laurena, maite, amontejo, mrgarcia}@ujaen.es
{mgarcia, mcdiaz, mdmolina, sjzafra, fmplaza}@ujaen.es

Resumen: El lenguaje humano determina cómo nos relacionamos e incluso cómo pensamos y concebimos la realidad de la que participamos. Es el principal medio de comunicación con nuestro entorno, y a través del cual se modela cognitivamente cada persona, como estudia la psicolingüística. El objetivo principal de este proyecto es el estudio y desarrollo de un asistente conversacional inteligente, que con base en las Tecnologías del Lenguaje Humano (TLH), permite dialogar con personas de edad avanzada con la finalidad de mantener y mejorar su bienestar social. Se integran estas tecnologías en las rutinas e intereses del mayor: asistencia en el desempeño de tareas domésticas cotidianas y de actividades que ejercitan la memoria a corto, medio y largo plazo. La monitorización de la interacción con el asistente virtual permite una evaluación posterior por parte de profesionales del ámbito de la psicología.
Palabras clave: Asistentes virtuales inteligentes, sistemas de diálogo, envejecimiento activo

Abstract: Human language determines how we relate to each other and even how we think and conceive the reality in which we participate. conceive of the reality in which we participate. It is the main means of communication with our environment, and through which each person is cognitively modeled, as studied by psycholinguistics. person, as studied by psycholinguistics. The main objective of this project is the study and development of an intelligent conversational assistant, based on Human Language Technologies (HLT), which allows dialogue with elderly people in order to maintain and improve their social welfare. These technologies are integrated into the routines and interests of the elderly: assistance in the performance of daily household chores and activities that exercise and activities that exercise short-, medium- and long-term memory. The monitoring of the interaction with the virtual assistant allows for subsequent evaluation by professionals in the field of psychology.
Keywords: Intelligent virtual assistants, dialogue systems, active aging

1 Introducción

Según la Organización Mundial de la Salud (OMS), entre 2020 y 2030, el porcentaje de habitantes del planeta mayores de 60 años aumentará un 34 %[1]. En la actualidad, el número de personas de 60 años o más supera al de niños menores de cinco años, y en 2050, el número de personas de 60 años o más será

[1]https://www.who.int/es/news-room/fact-sheets/detail/ageing-and-health

superior al de adolescentes y jóvenes de 15 a 24 años de edad. Es evidente que la pauta de envejecimiento de la población es mucho más rápida que en el pasado. Esto motiva que los países se tengan que enfrentar a retos para garantizar que sus sistemas sanitarios y sociales estén preparados para afrontar ese cambio demográfico.

El concepto "envejecimiento activo"lo propuso la OMS a finales de los años 90 para sustituir el concepto de "envejecimiento saludable", y se puede definir como "el proceso de optimización de las oportunidades de salud, participación y seguridad con el fin de mejorar la calidad de vida a medida que las personas envejecen"(Fernández-Ballesteros, 2009).

Un asistente virtual es un agente software con capacidad de ayuda en la automatización y asistencia a tareas, con una mínima interacción hombre-máquina. La interacción que se da entre un asistente virtual y una persona debe ser natural, mediante el uso del diálogo por voz.

Un asistente virtual inteligente es una ampliación del concepto de asistente virtual, donde al agente software tiene ya capacidad de búsqueda, procesamiento de información y hasta de razonamiento(Torres y Manjarrés-Betancur, 2020).

El proyecto InLife trata sobre el estudio y desarrollo de un asistente conversacional inteligente que, con base en las Tecnologías del Lenguaje Humano, permitirá dialogar con personas de edad avanzada con la finalidad de mantener y mejorar su bienestar social. Para garantizar que el asistente resulte accesible y atractivo se requiere del uso del perfil del usuario, que incluye el modelo de lenguaje especifico de cada persona.

A partir de una entrada en lenguaje natural, en formato texto, o mediante voz y aplicando un reconocedor de voz (ASR, del inglés Automatic Speech Recognition), un módulo de Tecnologías del Lenguaje Humano (TLH) comprende esa información, obtiene la información necesaria para que el siguiente módulo de gestión del diálogo pueda aplicar diversas estrategias y utiliza información del perfil del usuario y del contexto para generar una respuesta. Dicha respuesta, de nuevo en formato textual o generando voz (TTS, del inglés Text to Speech) y en lenguaje natural, se transmite de nuevo al usuario. La consecución de estos diálogos forma la conversación con el asistente conversacional inteligente propuesto en este proyecto(Allen, 1995)(Martínez-Santiago et al., 2015).

La organización de este trabajo es la siguiente. El Apartado 2 muestra la arquitectura del sistema. En el Apartado 3 se detallan aspectos relacionados con el desarrollo del mismo, y finalmente el Apartado 4 se indican las conclusiones principales y el trabajo futuro.

2 Arquitectura del sistema

La arquitectura del sistema está formada por tres componentes: un skill de Alexa[2]; AWS Lambda[3], el backend de la aplicación o Skill que interactúa con el usuario; TypeDB como sistema de gestión de datos.

De forma previa y automática se realiza una extracción de información y procesamiento del contenido de fuentes online de información local, así como de la parrilla de televisión. Una vez procesada esta información, así como información personalizada del usuario, se incorpora al modelado de datos del usuario. Dicha información será utilizada por el módulo de generación de preguntas para las actividades incorporadas actualmente en el proyecto.

Alexa. Es el servicio de voz ubicado en la nube de Amazon, que está disponible en los dispositivos de Amazon y otros de terceras empresas. Cuenta con funcionalidades o aplicaciones, denominadass Skills. Este servicio basado en voz permite a los usuarios interactuar con distintas tecnologías y servicios utilizando el lenguaje natural.

Alexa Skills Kit. Alexa Skills Kit (ASK) es un conjunto de herramientas, documentaciones, ejemplos de código fuente y API para crear Skills de Alexa.

AWS Lambda. Lambda es un servicio de Amazon Web Services (AWS) que permite ejecutar código que se lanza en los servidores de Amazon. Está orientado al desarrollo de cualquier tipo de backend, y se combina y configura muy bien a la hora de desarrollar y poner en marcha un nuevo Skill de Alexa. Es compatible con multitud de lenguajes de programación, incluyendo Node.js o Python.

TypeDB. TypeDB (previamente llamada Grakn.ai) es un sistema de modelado de datos basado en grafos de conocimiento para

[2]https://developer.amazon.com/es-ES/alexa

[3]https://docs.aws.amazon.com/lambda/index.html

sistemas orientados al conocimiento. Es una evolución de la base de datos relacional, muy útil para datos altamente interconectados ya que proporciona un esquema a nivel de concepto que implementa completamente el modelo Entidad-Relación (ER). Sin embargo, el esquema de TypeDB es un sistema de tipos que implementa los principios de representación y razonamiento del conocimiento. Esto permite que el lenguaje de consulta declarativo proporcione un lenguaje de modelado más expresivo y la capacidad de realizar razonamientos deductivos sobre grandes cantidades de datos complejos. TypeDB es una base de conocimientos para sistemas basados en inteligencia artificial y computación cognitiva.

3 Desarrollo del proyecto

3.1 Modelado de datos

El modelado de datos se ha generado para conectar toda la información relativa a un usuario y los datos de las distintas actividades. Se trata del núcleo central del sistema de gestión del diálogo, que permite trabajar con modelos personalizados así como con ampliaciones de actividades y gestión de las conversaciones. La Figura1 muestra el esquema lógico de datos, que tiene el perfil del usuario y del asistente virtual(Martínez-Santiago et al., 2020).

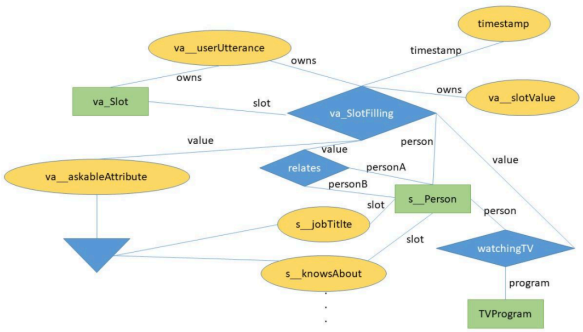


Figura 1: Modelo de datos del proyecto Inlife.

Perfil de usuario. La entidad principal es s__Person, conforme se especifica en schema.org. Está conformada en su mayor parte por atributos, algunos compuestos. Se relaciona con TVProgram a través de WatchingTV, con otras personas a través de personalRelation y sus derivadas: relaciones laborales, de amistad y familiar. Se han implementado reglas para inferir relaciones de parentesco, como hermanos, sobrinos, tíos, abuelos, etc.

Asistente virtual. La entidad principal es va__Slot, que representa un slot como podemos encontrar en Alexa. Cada vez que un usuario interacciona con un slot se crea una instancia de va__SlotFilling. Los valores legales para va__SlotFilling son aquellos objetos (atributos y relaciones) que implementan el rol SlotFilling:value.

3.2 Escenarios y captura de información

La finalidad del proyecto es favorecer el bienestar social de las personas mayores, utilizando el diálogo en los siguientes ámbitos:

- Acceso a información local. El objetivo es doble: en primer lugar, se busca dotar de dinamismo a la conversación, introducir cierto grado de serendipia y sorprender al mayor con datos que quizá no conozca de su entorno; en segundo lugar, estas noticias locales, breves y actuales facilitarán que el mayor viva en contacto con su entorno, le facilite sentirse parte de la sociedad en la que vive y, por ende, empoderar su bienestar social.
- Asistencia en tareas cotidianas. A modo de pequeñas "píldoras", a lo largo del día se recuerda y anima al mayor a asistir a actividades incluidas en su perfil de usuario y que tengan una clara vocación socializadora: gimnasia de mantenimiento, Universidad para mayores, excursiones, etc. Con la finalidad de motivar al mayor, y aprovechando el acceso a las redes sociales de éste, se puede incluir no sólo una descripción del evento si no también informar sobre la participación de personas allegadas al mayor.

Para ambos casos hay una etapa inicial de captura de información, en un ámbito local y en uno nacional (noticias, personajes y lugares famosos, por ejemplo). Se han creado extractores automáticos para la información de la parrilla televisiva y para información sobre localizaciones y rutas (utilizando la API de Google Maps y Here). Así mismo hay un proceso manual de toma de datos personales y de preferencias para cada usuario.

Toda esta información es tratada y cargada en el sistema de gestión de datos TypeDB.

Por otro lado, con la colaboración de una compañera, Doctora en Psicología, se han definido las actividades e interacciones que debe

llevar cada skill, teniendo en cuenta aspectos como la memoria temporal (corto, medio y largo plazo)(García-Viedma, 2006). El procedimiento de programación de actividades ha sido el siguiente:

1. Definición por parte del neuropsicólogo de los pasos que componen las actividades a monitorear (ir a la sala, sentarse, agarrar el control remoto del televisor, etc.) y la relación entre estos pasos y las posibles disfunciones de la memoria (p.ej., no recordar dónde está el mando de la televisión podría ser un signo de déficit de memoria episódico).

2. Diseño e implementación de las interacciones con el altavoz inteligente para detectar el rendimiento cognitivo, en colaboración con un especialista en neuropsicología. Esta interacción toma la forma de pequeñas charlas y juegos simples, con el objetivo de obtener pistas sobre el desempeño cognitivo del usuario.

En cuanto al aspecto funcional, el proyecto incorpora diversos módulos que permite cierto grado de personalización, así como el desarrollo de diversas actividades. La mayor complejidad en desarrollo se ha encontrado a la hora de diseñar una arquitectura que permite incorporar esta flexibilidad en personalización y actividades, así como en el módulo de generación de respuestas y conversaciones con los usuarios finales. Durante los próximos meses se seguirán concretando y evaluando actividades sobre núcleos de población concretos.

3.3 Evaluación del sistema

La información registrada de las interacciones del participante con el sistema se codifica en términos de aciertos, errores (intrusiones y omisiones) y tiempo de respuesta y/o ejecución. De esta forma, al igual que con los datos obtenidos con la batería neuropsicológica se lleva a cabo el análisis estadístico mediante modelos de series temporales y mediante ANOVA factorial mixto. Finalmente, para valorar la sensibilidad y especificidad del sistema se utilizará como método el análisis de curvas ROC.

A la fecha de finalización del proyecto se han finalizado pruebas técnicas en laboratorio. Una vez superadas, está previsto el arranque de pruebas reales, cuya evaluación llevará más tiempo y será realizada por los compañeros de psicología.

4 Conclusiones y trabajo futuro

La finalidad del proyecto Inlife es la adaptación y aplicación de técnicas y herramientas de PLN al envejecimiento activo. El proyecto finalizó en febrero de 2022 con distintas pruebas de laboratorio, pero se sigue trabajando para ponerlo en marcha en situaciones reales, mejorando la interacción mediante diálogo y adquiriendo y procesando de forma automática información relevante para cada usuario final.

Agradecimientos

Este trabajo ha sido parcialmente financiado con los proyectos 1380939 (FEDER Andalucía 2014-2020), P20-00956 (PAIDI 2020, de la Junta de Andalucía), el proyecto LIVING-LANG (RTI2018-094653-B-C21, MCIN/AEI/10.13039/501100011033), ERDF A way of making Europe, siendo el principal financiador el proyecto InLIFE de la Fundación CSIC.

Bibliografía

Allen, J. 1995. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, Inc.

Fernández-Ballesteros, R. 2009. *Envejecimiento activo: Contribuciones de la psicología*. Pirámide Madrid.

García-Viedma, M.-R. 2006. *Valoración del control atencional como marcador cognitivo del inicio de la enfermedad de Alzheimer*. Jaén: Universidad de Jaén.

Martínez-Santiago, F., M. R. García-Viedma, J. A. Williams, L. T. Slater, y G. V. Gkoutos. 2020. Aging neuro-behavior ontology. *Applied Ontology*, 15(2):219–239.

Martínez-Santiago, F., M. Díaz-Galiano, M. García-Cumbreras, y A. Montejo-Ráez. 2015. A method based on rules and machine learning for logic form identification in spanish. *Natural Language Engineering*, 23:1–23, 08.

Torres, M. M. E. y R. Manjarrés-Betancur. 2020. Asistente virtual académico utilizando tecnologías cognitivas de procesamiento de lenguaje natural. *Revista Politécnica*, 16(31):85–96.

Big Hug: Artificial intelligence for the protection of digital societies

Big Hug: Inteligencia artificial para la protección de la sociedad digital

Arturo Montéjo-Ráez[1], María Teresa Martín-Valdivia[1], L. Alfonso Ureña-López[1], Manuel Carlos Díaz-Galiano[1], Miguel Ángel García-Cumbreras[1], Manuel García-Vega[1], Fernando Martínez-Santiago[1], Flor Miriam Plaza-del-Arco[1], Salud María Jiménez-Zafra[1], María Dolores Molina-González[1], Luis-Joaquín García-López[2], María Belén Díez-Bedmar[3]
[1]Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
[2]Department of Psychology
[3]Department of English Studies
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
[1]{amontejo, maite, laurena, mcdiaz, magc}@ujaen.es
[1]{mgarcia, dofer, fmplaza, sjzafra, mdmolina}@ujaen.es
[2,3]{ljgarcia, belendb}@ujaen.es

Abstract: In this paper, we present the Big Hug Project, which aims to claim protect vulnerable citizens and help them and their families to feel more confident when using social media communication platforms. To this end, it proposes activities for building quality data, research in new algorithms to adapt current solutions to the changing nature of colloquial and informal communication, the evaluation of techniques and methods and the development of demonstrators. This project presents an interdisciplinary approach to early detection of young people at high-risk emotional problems. The involvement of colleagues from the Clinical Psychology and Corpus Linguistics fields, furthermore, provides the project with the necessary interdisciplinary to obtain robust results which may be significant to society.
Keywords: Natural Language Processing, NLP, sentiment analysis, Clinical Psychology, early detection.

Resumen: En este artículo presentamos el proyecto Big Hug, que pretende proteger a las personas vulnerables y ayudar a ellos y a sus familias a sentirse más seguros en el uso de medios de comunicación sociales. Para ello propone actividades para la construcción de datos de calidad, la investigación de nuevos algoritmos para adaptar los actuales a la naturaleza cambiante de la comunicación coloquial e informal, la evaluación de técnicas y métodos, y el desarrollo de demostradores. Además, este proyecto presenta un enfoque interdisciplinal en el campo de la detección temprana de los jóvenes con alto riesgo de sufrir problemas emocionales. Además, la participación de profesionales de la Psicología Clínica y de la Lingüística de Corpus dota al proyecto del necesario trabajo interdisciplinal para obtener resultados robustos que pueden ser significativos para la sociedad.
Palabras clave: Procesamiento del Lenguage Natural, PLN, análisis de sentimientos, Psicología Clínica, detección temprana.

1 Introduction

Human language is the main transmission medium involved in social interaction. There are revolutionary Natural Language Processing (NLP) algorithms that can provide means to prevent and predict risky interactions, protecting the most fragile members of our digital societies. Children and adolescents have been identified by the World Health Organization as being at particular risk of psychological distress in these media[1].

Human Language Technologies (HLT) can help us build more confident environments. Thanks to NLP, artificial intelligence solutions are able to model human language and use learned models to extract information and understand the meaning of text flowing through social networks. The combination of deep learning algorithms with linguistic resources and tools, enable the construction of monitoring systems for the early detection of signs of misbehaviours like eating disorders, depression, bullying or suicide tendencies over social media(Losada, Crestani, and Parapar, 2019; Parapar et al., 2021).

To this end, the project proposes two years of activities for building quality data, research in new algorithms to adapt current solutions to the changing nature of colloquial and informal communication, the evaluation of techniques and methods and the development of demonstrators to leverage human-centered solutions that will protect vulnerable citizens and help them and their families to feel more confident when using social media communication platforms. Besides, this project presents an interdisciplinary approach to early detection of young people at high-risk emotional problems. By indicated prevention, scientific community has agreed to name to high-risk individuals who are identified as having some detectable symptoms of emotional disorders but who do not meet criteria or a diagnosis at the current time. The collaboration of colleagues from the Clinical Psychology and Corpus Linguistics fields, furthermore, provides the project with the necessary interdisciplinary approach to obtain robust results which may be significant to society.

Joint efforts of NLP with Corpus Linguistics and Clinical Psychology are sought in this project with a two-fold purpose: a) to analyse the results obtained from the linguistic point of view to fine-tune and complement the NLP findings; and b) to contrast the results with the scientific literature on these disorders in Clinical Psychology.

[1] https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

2 Participants and project funding

The project brings together 3 partners from University of Jaén: SINAI group from Advanced Studies Center in ICT (CEATIC), Department of Psychology and Department of English Studies. This project has been supported by the grant P20_00956 (PAIDI 2020) funded by the Andalusian Regional Government.

3 State of the art

It is estimated 24 million children and young people in the EU suffer from bullying every year, which means that 7 out of 10 suffer some form of harassment or intimidation, whether verbal, physical or through new communication technologies (Cross et al., 2012). Navarro-Gómez (Navarro-Gómez, 2017) stated that social networks allow the viral diffusion of degrading contents. Cyberbullying or electronic aggression has already been designated as a serious public health threat and has elicited warnings to the general public from the Centers for Disease Control and Prevention (CDC) (Aboujaoude et al., 2015).

In another study (Stice and Van Ryzin, 2019), approximately 1 out of 10 people were found to develop some sort of eating disorder, which also caused anxiety, self-harming and a high risk of suicide. May studies have tackled this fact from psychometrics, but better tools for modeling the language used would help (Wang et al., 2017b), even more when eating disorders are rising all around the world. Emotional disorders, like depression and anxiety, affect a quarter of our population during their lifetime (Wang et al., 2017a). Depression can be studied and identified by monitoring users' posts and activity (Losada, Crestani, and Parapar, 2019).

In Spain there are 10 suicides a day, twice as many people die by suicide as by traffic accidents, 11 times more than by homicide and 80 times more than by gender violence. A very complete overview on how computers

and algorithms can help in preventing or detecting suicide risk is the one recently published by Ji (Ji et al., 2020). Recent studies have found that automatic processing of social media communications is an effective way to detect suicidal ideation by applying emotion and sentiment analysis over textual messages (Glenn et al., 2020).

NLP techniques are being applied to the analysis of social media textual data to face new problems like fake-news detection (Monti et al., 2019), offensive language identification (Zampieri et al., 2019), sentiment analysis (Martínez-Cámara et al., 2014), opinion mining and emotion detection (Plaza-del Arco et al., 2020). Social Big Textual Data is challenging, because language varies across time and space, language register is informal, colloquial and full of idioms compared to formal forms of text. Artificial Intelligence has gained a lot of popularity in recent years thanks to advent of Deep Learning techniques (Dean, Patterson, and Young, 2018). Nevertheless, many of the applications and problems overcome where already attempted with traditional algorithms in machine learning, heuristic approaches or knowledge-based systems. The big difference to previous approaches is that current proposals are data-driven: they are able to learn from large amounts of data and build models to perform different tasks with a level of success never reached by other solutions.

This shift has been especially dramatic for NLP. Linguistic-based methods have been surpassed by end-to-end architectures, where no prior knowledge on language is needed (Young et al., 2018), but massive amounts of data are required. During the last two years we have witnessed the birth of amazing models like BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019) or Transformer-XL (Dai et al., 2019), with impressive results in many different tasks. New models seem to learn language linguistic nature from data.

The gross research on NLP is turning towards Transformer based models and exploring how far these architectures are able to learn and perform in human related tasks, being sentiment analysis, emotion detection and hate-speech identification, among them.

There are previous projects in the pursuit of similar goals, like the STOP project (Ramírez-Cifuentes et al., 2020) or MENHIR (Kraus, Seldschopf, and Minker, 2021). The Big Hug project is not only focused in exploring algorithm and models for early detection of disorders, but also in finding effective ways to transfer these systems to real world applications.

4 Objectives of the project

The main objective is clear: a multidisciplinary project for the research on methods and algorithms to analyse textual streams across time and discover patterns for an early detection of potential harmful situations or behaviours. This global goal can be divided into the following sub-objectives: (1) To identify valid technologies for "listening" the interactions in digital environments. (2) To model different forms of aggressive communication or risky situations. (3) To identify young people at high risk, but by the very first time, via a screening of altogether big data, psychological, linguistic variables. (4) To facilitate the replication of the screening protocol based on a well-defined methodology and analysis plan, if the previous objective is met. (5) To enhancement of our capabilities to feed these artificial intelligences with quality data by means of new techniques and methods to process informal language or colloquial expressions. (6) To adapt human language technologies also to the specific one that is usually used to make apologia of those scenarios. (7) To explore practical solutions which may be integrated in the real world.

5 Conclusion

Dispositions for eating, anxiety and depressive disorders, are multifactorial. Big Hug represents a novel approach for mental disorders, integrating mental health, big data and linguistics measures as predictive measures for early diagnosis.

Research on mental health, for the early diagnosis and treatment of emotional mental health problems in the young is fragmented as researchers have traditionally worked in isolation and few studies examined the same or more than a limited set of risk factors, neglecting novel stratification strategies and development of algorithms. The Big Hug project avoids the problems of fragmentation by co-ordinating and developing joint activities related to early identification in order to coordinate high quality transnational research. The different perspectives and especially the different qualifications of

mental-health, applied linguistics and Information and Communication of Technologies (ICT) specialists working in academia could stimulate the discovery of new and creative solutions. Apart from multidisciplinarity, there are relevant transversal aspects in the project.

References

Aboujaoude, E., M. W. Savage, V. Starcevic, and W. O. Salame. 2015. Cyberbullying: Review of an old problem gone viral. *Journal of adolescent health*, 57(1):10–18.

Cross, E., R. Piggin, T. Douglas, and J. Vonkaenel-Flatt. 2012. Virtual violence ii: Progress and challenges in the fight against cyberbullying. *London: Beatbullying*.

Dai, Z., Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Dean, J., D. Patterson, and C. Young. 2018. A new golden age in computer architecture: Empowering the machine-learning revolution. *IEEE Micro*, 38(2):21–29.

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

Glenn, J. J., A. L. Nobles, L. E. Barnes, and B. A. Teachman. 2020. Can text messages identify suicide risk in real time? a within-subjects pilot examination of temporally sensitive markers of suicide risk. *Clinical Psychological Science*, 8(4):704–722.

Ji, S., S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

Kraus, M., P. Seldschopf, and W. Minker. 2021. Towards the Development of a Trustworthy Chatbot for Mental Health Applications. In *MultiMedia Modeling*, pages 354–366. Springer.

Losada, D. E., F. Crestani, and J. Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the CLEF for European Languages*, pages 340–357. Springer.

Martínez-Cámara, E., M. T. Martín-Valdivia, L. A. Urena-López, and A. R. Montejo-Ráez. 2014. Sentiment analysis in twitter. *Natural Language Engineering*, 20(1):1–28.

Monti, F., F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Navarro-Gómez, N. 2017. El suicidio en jóvenes en españa: cifras y posibles causas. análisis de los últimos datos disponibles. *Clínica y Salud*, 28(1):25–31.

Parapar, J., P. Martín-Rodilla, D. E. Losada, and F. Crestani. 2021. eRisk 2021: pathological gambling, self-harm and depression challenges. In *ECIR*, pages 650–656. Springer.

Plaza-del Arco, F. M., M. T. Martín-Valdivia, L. A. Ureña-López, and R. Mitkov. 2020. Improved emotion recognition in spanish social media through incorporation of lexical knowledge. *Future Generation Computer Systems*, 110:1000–1008.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ramírez-Cifuentes, D., A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, D. A. Velazquez, J. M. Gonfaus, J. Gonzàlez, et al. 2020. Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. *Journal of medical internet research*, 22(7):e17758.

Stice, E. and M. J. Van Ryzin. 2019. A prospective test of the temporal sequencing of risk factor emergence in the dual pathway model of eating disorders. *Journal of Abnormal Psychology*, 128(2):119.

Wang, J., X. Wu, W. Lai, E. Long, X. Zhang, W. Li, Y. Zhu, C. Chen, X. Zhong, Z. Liu, et al. 2017a. Prevalence of depression and depressive symptoms among outpatients: a systematic review and meta-analysis. *BMJ open*, 7(8):e017173.

Wang, T., M. Brede, A. Ianni, and E. Mentzakis. 2017b. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International conference on web search and data mining*, pages 91–100.

Young, T., D. Hazarika, S. Poria, and E. Cambria. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

HARTAes-vas: Lexical Combinations for an Academic Writing Aid Tool in Spanish and Basque

HARTAes-vas: Combinaciones léxicas para una Herramienta de ayuda a la redacción de textos académicos en español y en vasco

Margarita Alonso-Ramos,[1] Igone Zabala[2]
[1] Universidade da Coruña (UDC)
[2] Universidad del País Vasco (UPV/EHU)
margarita.alonso@udc.es; igone.zabala@ehu.eus

Abstract: Academic writing has become an object of study because of the need for tools to help novice writers. We focus on routinized lexical combinations that characterise academic discourse in Spanish and Basque. The aim is to extract these combinations from two academic corpora in order to build a writing aid tool serving both languages.
Keywords: Academic writing, collocations, discourse function, writing aid.

Resumen: La escritura académica se ha convertido en un objeto de estudio debido a la necesidad de herramientas de ayuda para los escritores noveles. Nos centramos en combinaciones léxicas rutinarias que caracterizan el discurso académico en español y en vasco. El objetivo es extraer estas combinaciones a partir de sendos corpus académicos con el fin de construir una herramienta de ayuda para las dos lenguas.
Palabras clave: Escritura académica, colocaciones, funciones discursivas, herramienta de ayuda.

1 General description

The HARTAes-vas project is funded by the Ministry of Science and Innovation in the 2019 call for R&D Knowledge Generation Projects. It is a project coordinated between the UPV/EHU and the UDC and, in some objectives, it is a continuation of previous projects related to academic writing in Spanish. In this new project, we are tackling a contrastive approach with two different languages from both a typological and a sociolinguistic point of view. The research team is made up of members of the LyS group at the UDC and the Ixa group at the UPV/EHU together with researchers from the Foundation Elhuyar.

In recent years, academic writing has become a priority object of study, especially in English (Hyland and Shaw 2016, among others). In order for members of the academic community to produce knowledge, they must be able to write in the conventional forms of academic texts. However, when students enter university, they are confronted with new written genres for which they are not provided with tools to facilitate the production of texts. Moreover, university students in Spain must be able to show proficiency in several languages and, paradoxically, Spanish students have more resources to help them with academic English than with the other languages of the state. One of the keys to this competence in writing lies in the mastery of certain routine expressions that give it its specific character: *academic lexical combination* (ALC), ranging from collocations (*extraer conclusiones*, *ondorioak atera* 'draw conclusions'), to discourse markers (*en conclusión*, *ondorioz* 'in conclusion') and also formulas such as *parece razonable concluir que* ('it seems reasonable to conclude that'), *ondorioz esan daiteke* ('consequently we can say'); all these expressions are ALC which we have in order to express a conclusion in Spanish and Basque.

However, before developing the tool that would help the students learn to write in this academic style, a diagnosis of the current written productions of our university students is needed. In previous research we have compiled a corpus of written productions of Spanish academic novices made up of Bachelor's and Master's theses (Alonso-Ramos et al. 2017; García-Salido et al. 2018; hereafter the Spanish novice corpus) and during this project we have compiled a comparable corpus of written productions of academic novices for Basque (hereafter the Basque novice corpus). The different sociolinguistic status of Basque with respect to Spanish forces different strategies: on the one hand, there is no academic corpus of expert academic writing in Basque available as a reference; on the other hand, Basque has not had enough time for the stabilisation of academic registers (Zabala et al. 2021), which suggests as a starting hypothesis that ALCs will have a lower degree of fixation and recurrence. Likewise, the agglutinative nature of Basque poses a challenge to the usual techniques for extracting combinations.

2 Goals

The overall goal is to create a bilingual tool (or two coordinated monolingual tools), focused on the use of ALCs, combining a dictionary and a corpus. We aim to build a tool where the user can choose the language and find help in choosing the appropriate lexical strategies according to different discourse needs.

More specifically, the project aims to: a) develop a model of ALCs that includes the characteristics of agglutinative languages such as Basque where different lexicographic and discursive classifications will be established; b) analyse the learners' use of such combinations in Spanish and Basque; c) investigate what kind of help related to the phenomena of lexical combinations they need when writing; d) develop corpus-based linguistic technologies for the automatic identification of ALCs.

3 Methodology

The project has multiple orientations: lexicological (as far as the linguistic phenomena studied are concerned); corpus linguistics and computational linguistics (insofar as the corpora are the fundamental source of data and the techniques with which they are exploited come from NLP) and didactics (following the approach of so-called *computer-assisted language learning* and, more particularly, the *data-driven learning* methodology).

The agglutinative nature of Basque inspired the design of alternative ALC identification techniques since the usual lexical bundle extraction technique is not suitable in all cases for Basque. The reason is that some formulas are made up of a single word in Basque and it is necessary to take into account the so-called *morphemic bundles* to complement the results obtained with the techniques used for inflectional languages. For example: *en resumen* 'in short' - *laburbilduz* 'short+gather+INSTR'; *por consiguiente* 'therefore'- *ondorioz* 'consequence + INSTR'.

3.1 Extracting academic vocabulary lists with corpus linguistics and NLP techniques

We analysed the Spanish novice corpus morphologically and syntactically to extract collocations with LinguaKit, Freeling and UDPipe, following the same criteria we used in the expert corpus (García-Salido et al. 2018). We extracted the following syntactic patterns: Subject-Verb (*objetivo se centra* 'objective focuses'), Verb-Object (*alcanzar objetivo* 'reach an objective'), Noun-Modifier (*objetivo fundamental* 'main objective'), N of N (*serie de objetivos* 'series of objectives'). We also extracted lists of n-grams, applying criteria of frequency and distribution by scientific domains and assigned the discursive function according to the typology established in García-Salido et al. (2019).

We applied a similar procedure to the Basque novice corpus which was morphologically analysed using Eustagger. We started by extracting an academic vocabulary based on the criteria defined in García-Salido (2021). We have used this word list to identify collocations, without the need to syntactically analyse the corpus (Gurrutxaga and Alegría 2011). We have extracted the following syntactic patterns: Subject-Verb (*datuek erakutsi* 'data show'), Verb-Object (*datuak bildu* 'collect data', *datuetan oinarritu* 'rely on data'), Noun-Modifier (*datu esanguratsu* 'significant data'), N-N (*datu sorta* 'data set', *datu-bilketa* 'data collection'). To obtain the formulas, we extracted lists of n-grams, applying the same criteria of frequency and dispersion and the same typology of

discursive functions described in García-Salido et al. (2019). Once the formula candidates have been validated, the variation was analysed in order to identify prototypical formulas and their variants.

3.2 Testing distributional semantics strategies

Once the two corpora of Spanish and Basque novice academic writing are balanced, we can exploit them as comparable corpora and apply computational techniques of distributional semantics in order to find correspondences between the formulas of the two languages. With the Spanish list, vector representations (embeddings) of each formula can be generated using non-compositional strategies, and we can then use them to identify the Basque single word equivalents of Spanish expressions in a previously obtained cross-linguistic semantic space. In this way, we may be able to relate *por consiguiente* and *ondorioz*, or *para terminar* 'to conclude' and *bukatzeko*, following the non-compositional strategy used by Garcia et al. (2019).

Monolingual distributional models, both monolexical and polylexical, will be generated with *fastText*, and mapped to a multilingual space with *vecmap*. Since we find both compositional and non-compositional expressions among the formulas, we will use equivalent search strategies adapted to each type of structure. For the non-compositional ones, we will represent each formula with a single vector, using the non-compositional method presented in Garcia et al. (2019). We consider that the use of this multilingual strategy can help in the identification of formulas, because if a Basque expression has a high degree of both internal cohesion and distributional similarity with a Spanish formula, the probability that it is indeed a formula in Basque is also very high. Likewise, it seems interesting to explore whether distributional models also identify a more discursive meaning, such as that of the formulas.

4 Results

The quantitative data from the Spanish novice corpus analysis are shown in Table 1. The data are presented with normalised frequency per million words due to the different size of the corpora.

		Types/ M	Tokens/M
Collocations	N-Modif	192	2.724
	N *de* N	85	1.106
	Subject-Verb	39	313
	Verb-object	219	2.753
	Total collocations	536	6.897
Formulae	**Total formulas**	211	20.474

Table 1: The ALC data from the Spanish novice corpus

The results of a contrastive analysis with the expert corpus show that novices use fewer collocations than experts. Also, novices use more collocations belonging to the general language. With respect to formulas, we see that novices use fewer types than experts, but almost as many tokens.

As far as Basque is concerned, we already have a corpus of novice academic writing (Aranzabe et al. 2022). Although its analysis has not yet been completed, we can already observe some characteristics: the ALCs are less stable compared to the Spanish novel corpus and a higher number of ALCs are considered incorrect. By validating the lists of ALCs in the Basque corpus, we will be able to make a more thorough comparison: contrasting formulas by functions and verifying whether the same functions are covered in the two languages and checking whether the equivalent bases are linked to more or fewer collocates in the different languages. This comparison will be vital for the design of the writing aid tool. Pending the aforementioned further analysis, the quantitative data are shown in Table 2.

		Types /M	Tokens/M
Collocations	N-Modif	150	4.024
	N - N	43	1.251
	Subject-Verb	3	58
	Verb-object	108	4.136
	Total collocations	305	9.471
Formulae	**Total formulas**	196	38.171

Table 2: The ALC data from the Basque novice corpus

5 Conclusions and future work

We have presented the main tasks we carried out to obtain the data for an academic writing aid tool. Next, we will explore the transfer strategies for the automatic identification of ALCs in several languages. We start from the hypothesis that a cross-linguistic language model trained to identify the formulas in Spanish could recognise expressions with similar characteristics in Basque. If the results obtained with this strategy are adequate, we could, on the one hand, automatically obtain new formulas in both languages in other corpora and, on the other hand, identify formulas in Basque that could be mapped to those in Spanish. Pending the results of the experiments with distributional semantics techniques, we are making progress in the design of the tool, which must meet two requirements: 1) provide onomasiological access by discursive function; 2) include a field of warnings where examples will be provided as correction models.

Acknowledgments

This work has been supported by the Xunta de Galicia, through grant ED431C 2020/11, and the Spanish Ministry of Science and Innovation through projects PID2019-109683GB-C21 and PID2019-109683GB-C22. I would like to thank Olga Zamaraeva for her valuable and constructive suggestions.

References

Alonso-Ramos, M., M. García-Salido, and M. Garcia. 2017. Exploiting a corpus to compile a lexical resource for academic writing: Spanish lexical combinations. In I. Kosem, et al. (eds.), *Electronic Lexicography in the 21st Century. Proceedings of eLex 2017 Conference*, pages 571–586. Lexical Computing Brno.

Aranzabe, M.J., A. Gurrutxaga, and I. Zabala. 2022. Compilación del corpus académico de noveles en euskera HARTAvas y su explotación para el estudio de la fraseología académica. *Procesamiento del Lenguaje Natural*

Garcia, M., M. García Salido, and M. Alonso-Ramos. 2019. Weighted compositional vectors for translating collocations using monolingual corpora. In *Computational and Corpus-Based Phraseology* (EUROPHRAS 2019). Lecture Notes in Artificial Intelligence, 11755. pages 113-128. Springer.

García-Salido, M., M. Garcia, M. Villayandre, and M. Alonso-Ramos. 2018. A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora. In N. Calzolari et al. (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), pages 260–265. Miyazaki.

García-Salido M., M. Garcia, and M. Alonso-Ramos. 2019. Identifying lexical bundles for an academic writing assistant in Spanish. In G. Corpas Pastor and R. Mitkov eds., *Computational and Corpus-Based Phraseology*. Europhras 2019, Proceedings, volume 11755 of Lecture Notes in Artificial Intelligence, pages 144-158. Springer.

García-Salido, M. (2021). Compiling an Academic Vocabulary List of Spanish. Available at: https://doi.org/10.13140/RG.2.2.27681.33123

Gurrutxaga, A., I. Alegria. 2011. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. In *Proceedings of the Workshop on Multiword Expressions: from parsing and generation to the real world*, 2–7. Association for Computational Linguistics.

Hyland, K., and P. Shaw, P. (eds.). 2016. *The Routledge Handbook of English for Academic Purposes.* London, Routledge.

Zabala, I., Aranzabe, Mª J., and Aldezabal, I. 2021. Retos actuales del desarrollo y aprendizaje de los registros académicos orales y escritos del euskera. *Círculo de Lingüística Aplicada a la Comunicación* 88: 31-50.

Proxecto Nós: Artificial Intelligence at the Service of the Galician Language

Proxecto Nós: Inteligencia artificial al servicio de la lengua gallega

Adina Ioana Vladu[1], Iria de-Dios-Flores[2], Carmen Magariños[1], John E. Ortega[2], José Ramom Pichel[2], Marcos Garcia[2], Pablo Gamallo[2], Elisa Fernández Rei[1], Alberto Bugarín[2], Manuel González González[1], Senén Barro[2], Xosé Luis Regueira[1]
[1] Instituto da Lingua Galega (ILG)
[2] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{adina.vladu, iria.dedios, mariadelcarmen.magarinos, john.ortega, jramom.pichel, marcos.garcia.gonzalez, pablo.gamallo, elisa.fernandez, alberto.bugarin.diz, manuel.gonzalez.gonzalez, senen.barro, xoseluis.regueira}@usc.gal

Abstract: *Proxecto Nós* is an initiative aimed at providing the Galician language with openly licensed resources, tools, and demonstrators in the area of intelligent technologies. The Project has two main scientific and technological objectives: (i) to integrate the Galician language into cutting-edge AI and language technologies, thus enabling the natural use of Galician in human-machine interactions; and (ii) to improve the state of the art of language technologies for Galician.
Keywords: Language technologies, linguistic rights, Galician, low-resource languages.

Resumen: *Proxecto Nós* es una iniciativa cuyo fin es desarrollar recursos, herramientas y demostradores de tecnologías inteligentes del lenguaje para el gallego. Los dos principales objetivos científicos y tecnológicos del proyecto son: (i) integrar el gallego en la vanguardia de la IA y las tecnologías del lenguaje, permitiendo así el uso natural de la lengua gallega en las interacciones hombre-máquina; y (ii) mejorar el estado del arte de las tecnologías lingüísticas para la lengua gallega.
Palabras clave: Tecnologías del lenguaje, derechos lingüísticos, Gallego, lenguas con pocos recursos.

1 Participating Entities and Funding

Proxecto Nós (The *Nós* Project) is an initiative promoted by the Galician Government (Xunta de Galicia), aimed at providing the Galician language with openly licensed resources, tools, demonstrators, and use cases in the area of intelligent technologies. The execution of *Proxecto Nós* has been entrusted to the University of Santiago de Compostela (USC) and is currently being carried out by a research team comprising members of the Instituto da Lingua Galega (ILG) and the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS). The first stage, spanning from the final trimester of 2021 to 2025, will lay the foundations and provide the resources that will help place Galician among the languages that are fully active in the digital society and economy.

2 Context and Motivation

The development of language technologies is a strategic innovation area geared towards the digital society and economy, and it has been a priority in both Spanish (Plan Estatal de Investigación Científica y Técnica y de Innovación, Estrategia Española de Ciencia y Tecnología y de Innovación) and European (Horizon 2020) scientific planning.

Technologies such as machine translation (MT), information extraction (IE), text analytics, and dialogue systems are essential in the digital society, culture, and economy.

Languages in high demand worldwide (especially English) benefit from a large variety of computational resources that can contribute to developing new automatic language processing technologies and tools. Such is the case due to the long-standing research tradition in these areas (e.g., the variety of projects financed by USA's DARPA) and the need to incorporate such languages into the AI applications associated with the latest electronic devices (such as the conversational AI or automatic dictation software developed by Google, Amazon or Apple). Other languages that have joined AI research later, such as Chinese, are currently following in the footsteps of English, through projects such as Baidu's Qian Yan, which improve significantly the computational resources available in their respective language varieties.

Notwithstanding, language technologies are also necessary for languages in lower international demand. Consequently, different languages have developed similar initiatives to Nós. Among others, we can highlight Projecte AINA, which will develop computational resources for Catalan until 2024, or the work carried out at the HiTZ Research Center, focusing on languages technologies for Basque. Other projects, such as CorCenCC (in Great Britain, for Welsh) or UQAILAUT (in Canada, for Inuktitut) were considered success cases in the promotion of the digital use of socially threatened languages.

The democratization of language technologies has a great social and cultural impact on the communities that use them. For instance, MT increases access to contents in different languages, thus facilitating intercultural relations; dialogue systems allow us to communicate with machines in our own language; and semantic technologies enable advances in the automatic comprehension of texts, thus making it possible to process enormous quantities of documents. In the case of Galician, incorporating the language into state-of-the-art AI applications can not only significantly favor its prestige (a decisive factor in language normalization), but also guarantee citizens' language rights and reduce social inequality.

In economic terms, the global Natural Language Processing (NLP) market size was valued at more than USD 10 billion in 2020 and is expected to reach USD 41 billion by 2025 (Aldabe et al., 2021). NLP technologies are used in different areas such as information retrieval, MT, IE (with notable growth in its application in the medical domain during the Covid-19 pandemic), dialogue systems, and automatic text generation, among many others. The capacity to model language, an essential ability for human beings, ensures a promising future for such technologies from both an economic and research and innovation perspective.

3 State of the art: Galician resources and technologies

In 2012, the White Paper *The Galician Language in the Digital Age* (García-Mateo et al., 2012) described Galician as a language with a level of technological support that "gives rise to cautious optimism", while highlighting the need for new resources and tools. Previous research projects on Galician resulted in speech processing resources (COTOVÍA), an annotated reference corpus (CORGA), morphosyntactic lemmatizers and taggers (XIADA, FreeLing, IXA-Pipes), other specialized corpora, both text (CLUVI, CTG, TreeGal) and speech (CORILGA, AGO), MT systems (GAIO, OpenTrad), spellcheckers (OrtoGal), grammar checkers (Avalingua), language analysis and IE tools (Linguakit), language models (SemantiGal, Bertinho), and other resources.

Furthermore, Galician is currently part of multilingual crowdsourced data collection initiatives carried out by important companies on the global IT market, which have resulted in speech databases such as Google's SLR77 (Kjartansson et al., 2020) and Mozilla's CommonVoice 7.0 and 8.0 (Ardila et al., 2020). This situation is reflected in a recent report on the current state of the LT (Language Technology) field for Galician (Ramírez Sánchez & García Mateo, 2022), which informed on the considerable growth in the production of high-quality Galician resources and services, especially text resources.

Despite the quality of these resources, it should be noted that not all are freely and publicly available for the development of LT. The LT field has undergone profound changes over the last few years since the introduction of

neural network systems. Generally, training models using these state-of-the-art technologies requires large quantities of data and has high energetic and computational costs, which continues to be a challenge for low-resource languages. However, as many recent studies show, end-to-end technologies and open-source multilingual pre-trained models created using large quantities of data from high-resource languages (Shen et al., 2018; Baevski et al., 2020; Wolf et al., 2020) can be used, through transfer learning and fine-tuning, to train models in low- or medium-resource languages such as Catalan (Külebi & Öktem, 2018; Külebi et al., 2020) or, in our case, Galician. To this end, the existence of resources and tools that are freely available to the scientific and business community is essential, and that constitutes one of the main objectives of *Proxecto Nós*.

4 Project description

4.1 Organization

The tasks that are to be carried out as part of the Project can be included in the following areas, corresponding to some of the major NLP fields:

(1) Speech synthesis (TTS)
(2) Speech recognition (ASR)
(3) Automatic text generation
(4) Dialogue systems
(5) MT
(6) IE
(7) Opinion mining and fact checking
(8) Language correction and assessment

These broad, mutually interdependent areas fall within the three strategic lines jointly identified by the Project's research team and the Xunta de Galicia (in particular, with the Axencia para a Modernización Tecnolóxica de Galicia): (i) spoken or written conversation with people, (ii) language quality, and (iii) information management.

In accordance with the funding agreement signed by the Xunta de Galicia and the USC, the organization of the tasks included in *Nós* follows a yearly schedule. Each year, resources, language models and demonstrators from different areas will be made publicly available.

4.2 Scientific and Technological Objectives

Proxecto Nós has two main scientific and technological objectives: (i) to integrate the Galician language into cutting-edge AI and language technologies, thus enabling the natural use of Galician in human-machine interactions; and (ii) to improve the state of the art of language technologies for Galician.

For this purpose, resources, tools, and applications will be developed and distributed under open licenses, which will allow them to be integrated into existing devices and services (such as smart speakers or conversational agents) and future technologies. To this end, specific objectives directly related to some of the major tasks of NLP have been established.

Each of these technological objectives will be executed in a different subproject, which will allow the parallel development of different tasks and, overall, a more effective organization of the work. However, a set of general objectives are shared by all the tasks. These objectives are: (i) the compilation of high-quality linguistic resources (annotated reference corpora, web-scale corpora, specialized corpora by tasks and domains, parallel corpora, knowledge bases, dictionaries, etc.); (ii) the elaboration of language and acoustic models (both general-purpose and task-specific models); and (iii) the development of applications based on these models. The project will also have a general coordination mechanism through which resources will be distributed and shared among its subprojects.

The resources and language models developed for each task will be made available to the public, thus allowing their use in all kinds of applications, services, and products, by the scientific community, companies, institutions, and society in general. The results will be disseminated through a repository available at the project's web portal (which can be hosted on internal servers), as well as other established and internationally recognized repositories, such as HuggingFace, GitHub, Zenodo, etc.

Finally, the project contemplates the complete development of applications based on these resources, which will act as visible and accessible demonstrators of the developed technology and will produce a tractor effect that will lead to the development of new products.

5 Conclusion and Future Work

Among the initial results of *Nós*, we can highlight the first crawl of a web-based Galician corpus and a language model based on the CCNet tools and data (Ortega et al., 2022a), and the development and testing of a Spanish-

Galician neural machine translation (NMT) system prototype (Ortega et al., 2022b).

For the current year, *Proxecto Nós* aims to keep generating linguistic and computational resources to explore different subprojects. Specifically, in the first half of 2022 work will be carried out on the design of a high-quality speech corpus of sufficient size so as to allow training TTS state-of-the-art models, to be released in the last trimester. The second half of the year will also see the publication of a speech corpus for ASR. In the same timeframe, the project will publish several text corpora: parallel Galician-Spanish, Galician-English, and Galician-Portuguese corpora; a web-scale Galician text corpus, larger than the one already compiled, to be used in all the subprojects working with written text included in *Nós*; and a domain-specific corpus for automatic text generation. Based on these resources, new language models will be developed using different state-of-the-art techniques, as well as demonstrators or prototypes of a TTS system, NMT system, and automatic text generator for Galician. At the same time, throughout 2022 efforts will focus on extending and improving the first systems developed, and on validating the results obtained via the creation of high-quality gold standards.

Acknowledgements

This research was funded by the project "*Nós*: Galician in the society and economy of artificial intelligence" (*Proxecto Nós: O galego na sociedade e economía da intelixencia artificial* 2021-CP080), agreement between Xunta de Galicia and University of Santiago de Compostela, and grant ED431G2019/04 by the Galician Ministry of Education, University and Professional Training, and the European Regional Development Fund (ERDF/FEDER program).

References

Aldabe, I., Rehm, G., Rigau, G., Way, A. Report on existing strategic documents and projects in LT/AI. European Language Equality (ELE), 2021.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. In: Procs of LREC 2020.

Baevski, A., Zhou, H., Mohamed, A., Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv, 2020, pp. 1–19.

García Mateo, C., Arza Rodríguez, M. (auth.), Rehm, G., Uszkoreit, H. (eds.). The Galician Language in the Digital Age. Springer-Verlag, Berlin Heidelberg, 2012.

Külebi, B., Öktem, A. Building an Open Source Automatic Speech Recognition System for Catalan. In: IberSPEECH, Barcelona, Spain, 2018, pp. 25–29.

Külebi, B., Öktem, A., Peiró-Lilja, A., Pascual, S., Farrús, M. CATOTRON - A Neural Text-To-Speech System in Catalan. In: Procs. of Interspeech 2020.

Kjartansson, O., Gutkin, A., Butryna, A., Demirsahin, I., Rivera, C. Open-Source High Quality Speech Datasets for Basque, Catalan and Galician. In: Procs. of the 1st Joint Workshop on SLTU and CCURL, Marseille, France, 2020, pp.21—27.

Ortega, J.E., de Dios Flores, I., Gamallo, P., Pichel, J.R. A Neural Machine Translation System for Spanish to Galician through Portuguese Transliteration. In: PROPOR 2022, Fortaleza, Brazil.

Ortega, J.E., de Dios Flores, I., Pichel, J.R., Gamallo, P. Revisiting CCNet for Quality Measurements in Galician. In: PROPOR 2022, Fortaleza, Brazil.

Ramírez Sánchez, J.M., García Mateo, C. (auth.), Giagkou, M., Piperidis, S., Rehm, G., Dunne, J. (eds.). Report on the Galician Language (Deliverable D1.15). ELE, 2022.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R.J., Saurous, R.A., Agiomyrgiannakis, Y., Wu, Y. Natural TTS Synthesis By Conditioning Wavenet On Mel Spectrogram Predictions. In: Procs. of ICASSP, 2018.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In Procs. of the 2020 Conference on Empirical Methods in NLP: System Demonstrations, pp. 38–45.

CoToHiLi: Computational Tools for Historical Linguistics

CoToHiLi: Herramientas computacionales para la lingüística histórica

Alina Cristea, Anca Dinu, Liviu P Dinu, Simona Georgescu, Ana Uban, Laurentiu Zoicas
University of Bucharest
Research group: Human Language Technologies Research Center, University of Bucharest
{alina.cristea,ldinu,auban}@fmi.unibuc.ro, {anca.dinu,simona.georgescu,laurentiu.zoicas}@lls.unibuc.ro

Abstract: This project represents a computational framework for historical linguistics. The general purpose of the CoToHiLi project is to integrate expert knowledge and computational power to address cognate identification, cognate-borrowing discrimination, Latin protoword reconstruction and semantic divergence. The goal of the project is twofold: 1) to automate certain parts of the traditional work-flow of the comparative method (such as the collection of data or the automatic alignment based on predefined or inferred rules), and 2) to bring new insights or avenues of investigation, which might not be easily accessible otherwise (e.g., the automatic identification of patterns and regularities in large amounts of data). The project will provide tools for the main Romance kernel group (French, Italian, Portuguese, Romanian, Spanish), as well as Latin. The methodologies and computational tools proposed could also serve as a basis for further development for other comparable language families, including less studied languages, with scarce resources available.
Keywords: historical linguistics, cognates, semantic divergence.

Resumen: El proyecto representa un marco computacional para la lingüística histórica. El propósito general es integrar el conocimiento experto y el poder computacional en la aproximación a la identificación de palabras cognadas, la discriminación entre préstamos y cognados, la reconstrucción de protoformas léxicas y el análisis de la divergencia semántica. Por un lado, nos proponemos automatizar ciertas etapas en el flujo de trabajo tradicional del método comparativo (como la recopilación de datos o la alineación automática) y, por otro lado, pretendemos aportar nuevos conocimientos o vías de investigación que de otro modo no serían fácilmente accesibles (e.g., la identificación automática de patrones y regularidades en un gran caudal de datos). El proyecto se plantea proporcionar herramientas computacionales para el grupo principal del núcleo romance (español, francés, italiano, portugués, rumano), así como para el latín. Las metodologías y herramientas computacionales propuestas podrán servir de base para el estudio comparativo de otras familias lingüísticas, incluso las menos estudiadas, que disponen de recursos limitados.
Palabras clave: lingüística histórica, palabras cognadas, divergencia semántica.

1 Introduction

The general purpose of the CoToHiLi[1] project is to integrate expert knowledge and computational power to address the following topics: cognate identification, cognate-borrowing discrimination, Latin proto-word reconstruction and semantic divergence. Our project is focused on the Romance languages (French, Italian, Portuguese, Romanian, Spanish), and will provide tools for the main Romance kernel group and for Latin. The duration of the project is 3 years, starting from January 2021. The research problems that we address are significant on multiple levels. From a scientific point of view, any advance in historical linguistics is of paramount cultural importance, being inherently connected with human history ("each word a history", cf. (Campbell, 1998)). Longobardi (LanGeLin project, 2012-2018) explored the potential correlation of genetic and linguistic

[1]This is the project's web page, where we will include our results and updates: https://nlp.unibuc.ro/projects/cotohili.html.

distances, starting from what he called Darwin's last challenge (see also (Ritt, 2004)). Given that the socio-economical and cultural factors are some of the motivations for borrowing from one language to another (Campbell, 1998; Epps, 2014), the topic of this research project facilitates reconstructing certain aspects related to society and culture for groups of people speaking a given proto-language, and gaining insights into their past social interactions and into their social and cultural practices (Epps, 2014). From a technological perspective, as linguistic change is the most visible at the lexical and semantic level, computational tools can be designed to serve both aspects. Even though historical lexicology has leveraged technological advances, and some pioneering work was initiated on various steps of the work-flow (cognate identification, proto-word reconstruction), historical semantics has not sufficiently benefited from the advances in computer science. Yet, by drawing special attention to the semantic divergence occurring in pairs of cognates, we could both take a few steps forward towards a unitary theory of semantic change, and improve practical applications such as automatic translation systems or language e-learning systems, aware of false friends and related phenomena.

2 Objectives

The innovation of the project consists in integrating linguists' knowledge with new computational methods in a unified framework, to address important problems from historical linguistics, enabling experts to provide input and feedback throughout the whole development process.

Identification of related words: We aim at going one step further than the current state-of-the-art methods by: a) proposing a more in-depth analysis, by identifying the direction of the borrowings and b) automatizing the whole process as a pipeline that, given a pair of input words, provides an automatic analysis regarding the relationship between them (Ciobanu and Dinu, 2014; Ciobanu and Dinu, 2015; Cristea et al., 2021b).

Latin proto-word reconstruction: To improve previous results, we intend to use more recent techniques (Ciobanu and Dinu, 2018), as conditional random fields (CRF) for sequence labelling and deep learning, in particular character-level neural networks. The alignment technique, which stands at the foundation of our approach, will be improved by an heuristic for choosing the best alignment. We also address the challenging problem of multiple alignment (finding an alignment for more than two words), in order to be able to extract knowledge from cognate sets in multiple languages.

Diachronic semantic divergence: Semantic change is a continuous and complex process ((Campbell, 1998) presents no less than 11 types of semantic change), which has been recently studied in the context of distributional semantics theory. Vector spaces and word embeddings have been used for tracking semantic shifts of words in English across different time periods. Our aim is to exploit vectorial meaning representations to track the semantic change of words in Latin and across multi-languages, in the Romance language family, for the first time, with the substantial purpose of looking for common patterns characterizing the overall semantic divergence cases. Additionally, we intend to explore the statistical properties of the vectorial spaces where the word embeddings reside (Uban, Ciobanu, and Dinu, 2021; Uban et al., 2021).

3 Impact

The methodologies and computational tools we propose could extend their applicability not only to various linguistic branches in the Indo-European family, but also to less studied languages or linguistic families. Such advances could provide new answers in historical and social sciences, given that lexical and semantic change is a key source of clues regarding both the dynamics of cultural interactions between groups in the past, and the technological innovations and exchanges that have taken place across space and time (Epps, 2014). As for the socio-economic impact of the CoToHiLi project, in the context of the increasing number of attempts to create automatic tools designed for linguistic comprehension, our computational devices could support Romance intercomprehension by bringing into light the common linguistic features, as well as the semantic relations between the Romance cognates or borrowings. Such an advance can prove its usefulness in the constant efforts to improve the automatic translation systems.

4 Methodology

For the first two objectives, our methodology is focused on two main aspects: creating clean datasets and developing computational methods for achieving the proposed research tasks. For the Romance languages there are already some existing resources (for cognates, for borrowings and for proto-word reconstruction), but they are scattered, incomplete, or with uncertain availability (cf. (Bouchard-Côté et al., 2013; Ciobanu and Dinu, 2019)). Thus, datasets do not have to be built from scratch, but the data need to be harmonized, verified and enhanced where necessary, in order to become a benchmark in the domain. By using computational tools, corroborated by the direct intervention of classical linguists, we have already built a significant part of the database, representing the starting point for the computational methods that are being developed. We have continued with the alignment of word pairs. Given the lack of an unanimously accepted alignment method (Ciobanu and Dinu, 2019; Kondrak, 2000), we confront a semi-automatic manner of choosing the alignment with the knowledge of classical linguists, in order to establish an heuristic capable of making the best choice. From the alignment, we extract features for machine learning models. We improve current existing computational methods with linguistic features provided by experts. We develop a machine-learning classifiers (using support vector machines), sequential models (using CRF and neural networks) and ensemble techniques. We are currently working with the orthographic form of the words, while for Romanian, Spanish and Italian we are planning to also use the phonetic transcription. For the third objective, in order to identify semantic shifts across time periods as well as languages, we leverage vector space representations of meaning, or word embeddings, relying on traditional models such as word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2016), as well as experimenting with state-of-the-art language models such as BERT (Devlin et al., 2019). The first step consists of building semantic representations for the words in each of the target languages, based on the multilingual corpora. Then, by obtaining a multilingual semantic space, shared between the cognates, we become able to compute the semantic distance between them using their vectorial representations in this space. This result will allow us to analyze both the statistical and the linguistic properties of words whose meanings have diverged. The available corpora are unequal from one language to another; for instance, the Royal Spanish Academy provides an exhaustive diachronic corpus of its language, whereas for Romanian we only have access to a scarce data-base, composed of a fairly limited number of old texts. In order to ensure the accuracy of our analysis, in this stage of the project, we use mainly lexicographic resources, as well as data-bases built for the contemporary stage of each language (for example, multilingual Wikipedia).

5 Current Results

For the first two objectives, we have started building datasets of cognates and borrowed words for the Romance languages.(Cristea et al., 2021a) This first step relies on dictionaries that contain etymological information (e.g., for Romanian we use 13 dictionaries available in digital format). We have proposed a new method automatically discriminating between inherited and borrowed Latin words. We have introduced a new dataset and investigated the case of Romance languages - where words directly inherited from Latin coexist with words borrowed from Latin -, and explored whether automatic discrimination between them was possible. An initial trial was to automatically predict whether a word was inherited or borrowed by simply taking into account its intrinsic structure, given that borrowed words are presumably less eroded than inherited ones, subject to historical sound shifts. We then took a step farther and employed n-gram character features extracted from the word-etymon pairs and from their alignment, which led to considerably better results (Cristea et al., 2021b). For the third objective, a first step has been taken with the investigation of the semantic divergence of cognate pairs in English and Romance languages, by means of word embeddings. To this end, we introduced a new curated dataset of cognates in all pairs of those languages. We described the types of errors that occurred during the automated cognate identification process and manually correct them. Additionally, we labeled the English cognates according to their etymology, separating them into two groups: old borrowings and recent borrowings. On this curated dataset, we

analysed word properties such as frequency and polysemy, and the distribution of similarity scores between cognate sets in different languages. We automatically identified different clusters of English cognates, setting a new direction of research in cognates, borrowings and possibly false friends analysis in related languages (Uban et al., 2021).

6 Conclusions

Drawn within a computational framework, the CoToHiLi project addresses key concerns of historical linguistics centered on the Romance languages, such as cognate identification, cognate-borrowing discrimination, Latin protoword reconstruction and semantic divergence, towards which we have taken a few steps forward by performing various experiments. At this stage of the project, we analyze only the main five Romance languages (French, Italian, Portuguese, Romanian, Spanish), but as we advance we intend to include other Romance idioms as well. We predict that the methodologies and computational tools proposed will also serve as a basis for further development for other comparable language families, including less studied languages, with scarce resources available.

Acknowledgments Research supported by the Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, project number 108/2021, Romania.

References

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information. *TACL*, 5:135–146, 07.

Bouchard-Côté, A., D. Hall, T. L. Griffiths, and D. Klein. 2013. Automated Reconstruction of Ancient Languages Using Probabilistic Models of Sound Change. *PNAS*, 110(11):4224–4229.

Campbell, L. 1998. *Historical Linguistics. An Introduction*. MIT Press.

Ciobanu, A. M. and L. P. Dinu. 2014. Automatic detection of cognates using orthographic alignment. In *Proceedings of ACL 2014, Volume 2*, pages 99–105.

Ciobanu, A. M. and L. P. Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of ACL 2015*, pages 431–437.

Ciobanu, A. M. and L. P. Dinu. 2018. Ab initio: Automatic Latin proto-word reconstruction. In *Proceedings of COLING 2018*, pages 1604–1614.

Ciobanu, A. M. and L. P. Dinu. 2019. Automatic identification and production of related words for historical linguistics. *Computational Linguistics*, 45(4):667–704.

Cristea, A. M., A. Dinu, L. P. Dinu, S. Georgescu, A. S. Uban, and L. Zoicas. 2021a. Towards an Etymological Map of Romanian. In *Proceedings of RANLP 2021*, pages 315–323.

Cristea, A. M., L. P. Dinu, S. Georgescu, M. Mihai, and A. S. Uban. 2021b. Automatic discrimination between inherited and borrowed latin words in romance languages. In *Findings of EMNLP 2021*, pages 2845–2855.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*, pages 4171–4186.

Epps, P. 2014. Historical linguistics and socio-cultural reconstruction. In *The Routledge Handbook of Historical Linguistics*, pages 579–597. London: Routledge.

Kondrak, G. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of ANLP 2000*, pages 288–295.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*, pages 3111—-3119.

Ritt, N. 2004. *Selfish Sounds and Linguistic Evolution. A Darwinian Approach to Language Change*. Cambridge University Press.

Uban, A. S., A. M. Ciobanu, and L. P. Dinu. 2021. Cross-lingual Laws of Semantic Change. *Computational Approaches to Semantic Change*, pages 219–260.

Uban, A. S., A. Cristea, A. Dinu, L. P. Dinu, S. Georgescu, and L. Zoicas. 2021. Tracking semantic change in cognate sets for English and Romance languages. In *Proceedings of LChange 2021*, pages 64–74.

Exploración del conocimiento semántico en modelos vectoriales: polisemia, sinonimia e idiomaticidad

An exploration of the semantic knowledge in vector models: polysemy, synonymy and idiomaticity

Marcos Garcia[1], **Pablo Gamallo**[1], **Martín Pereira-Fariña**[2], **Iria de-Dios-Flores**[1]
[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
[2] Departamento de Filosofía e Antropoloxía
Universidade de Santiago de Compostela
Santiago de Compostela, Galiza
{marcos.garcia, pablo.gamallo, martin.pereira, iria.dedios}@usc.gal

Resumen: En este artículo presentamos el proyecto *Exploración del conocimiento semántico en modelos vectoriales: polisemia, sinonimia e idiomaticidad*, financiado por la Xunta de Galicia dentro del programa "Consolidación e estruturación de unidades de investigación competitivas e outras accións de fomento: Proxectos de Excelencia", con una duración de 5 años (2021-2026). El proyecto tiene como objetivo principal el análisis de los modelos de lengua más recientes en relación con la representacion de varios aspectos de la semántica léxica: polisemia y homonimia, sinonimia e idiomaticidad. Las lenguas en las que trabajaremos son el gallego-portugués (en sus variedades gallega y portuguesa, fundamentalmente), el castellano y el inglés.
Palabras clave: semántica léxica, semántica distribucional, modelos de lengua

Abstract: In this paper, we present the project *An exploration of the semantic knowledge in vector models: polysemy, synonymy and idiomaticity*, funded by the Xunta de Galicia within the program "Consolidación e estruturación de unidades de investigación competitivas e outras accións de fomento: Proxectos de Excelencia", with a duration of 5 years (2021-2026). The main objective of the project is the analysis of the most recent language models regarding the representation of several aspects of lexical semantics: polysemy and homonymy, synonymy and idiomaticity. The languages in which we will work are Galician-Portuguese (in its Galician and Portuguese varieties, fundamentally), Spanish and English.
Keywords: lexical semantics, distributional semantics, language models

1 Introducción y objetivos

El uso de arquitecturas basadas en redes neuronales artificiales se ha convertido en el enfoque más dominante para el procesamiento de las lenguas naturales (PLN) en los últimos años (Collobert et al., 2011), produciendo en muchas áreas resultados significativamente mejores que modelos supervisados diseñados seleccionando propiedades individuales de las tareas a resolver (Schnabel et al., 2015). Este cambio de paradigma ha promovido la popularización de los modelos vectoriales inspirados en la hipótesis distribucional (Harris, 1954; Firth, 1957), que hasta entonces se utilizaban principalmente en la investigación en ciencias cognitivas y lingüística computacional (Miller, 1971; Landauer y Dumais, 1997; Mitchell y Lapata, 2010). En este campo, la implementación de arquitecturas computacionalmente más eficientes, con reducciones drásticas en la dimensionalidad (Mikolov et al., 2013), ha despertado un gran interés en los estudios de semántica distribucional, impulsado también por la codificación que estos modelos realizan de varias regularidades lingüísticas (Mikolov, Yih, y Zweig, 2013). Esta área, anteriormente dominada por metodologías lingüísticamente informadas y más interpretables (por ejemplo, con vectores construidos usando dependencias sintácticas (Padó y Lapata, 2007)), se ha convertido en una de las más productivas en los estudios PLN (Boleda, 2020).

En este sentido, la aparición de técnicas de aprendizaje profundo que utilizan redes neuronales de varias capas de profundidad, con millones de hiperparámetros (y que requieren grandes infraestructuras computacionales) ha provocado la proliferación de modelos de lengua que realizan tareas de PLN con mayor precisión. Entre muchos otros, podemos destacar los modelos públicos ELMo (Embeddings from Language Models (Peters et al., 2018)), o las diferentes variantes de BERT (Bidirectional Encoder Representations from Transformers (Devlin et al., 2019)),

El proyecto aquí presentado encaja en esta nueva línea de investigación, y se centra en el análisis de la capacidad de estos modelos para resolver varios tipos de ambigüedad léxica:[1]

1. Polisemia y homonimia, i.e., una única forma ortográfica que tiene diferentes significados (o sentidos) en función del contexto. Por ejemplo, *escuela* como edificio, como organización, o como conjunto de personas (polisemia), o *banco* como institución financiera, como mobiliario, o como grupo de peces (homonimia).

2. Sinonimia, i.e., diferentes palabras que expresan el mismo sentido en determinados contextos (e.g., *monitor* o *pantalla* para referirse al equipamiento de visualización de un ordenador).

3. Idiomaticidad, i.e., expresiones multipalabra (MWEs) cuyo significado no se corresponde con en de los elementos que la conforman (e.g., *techo de cristal* como barrera social para las mujeres).

En vista de lo anterior, nuestra investigación pretende llenar un vacío de especial importancia en la evaluación de estos modelos computacionales, investigando la presencia de varios tipos de conocimiento relacionados con la semántica léxica en varias lenguas. Así, el objetivo principal del proyecto es explorar los modelos lingüísticos más recientes en relación con la representación de la polisemia y homonimia, la sinonimia y la composicionalidad semántica, así como compararlos con métodos distribucionales y composicionales más interpretables.

[1] En líneas generales seguimos a Cruse (1986) para la definición de los fenómenos aquí mencionados.

Los resultados del presente proyecto permitirán, por un lado, avanzar en la comprensión de la información semántica codificada tanto en representaciones distribucionales estáticas como en los modelos de lengua entrenados con redes neuronales profundas. Además, y aunque el proyecto está enfocado principalmente a la exploración de modelos, tanto los conjuntos de datos que creemos como los resultados de la anotación manual por parte de los participantes serán una contribución importante en relación con la interpretación semántica de la polisemia y homonimia, la sinonimia y la idiomaticidad por parte de hablantes nativos de varios idiomas.

2 Metodología y plan de trabajo

Para el desarrollo de este proyecto utilizaremos la siguiente metodología y técnicas instrumentales, que en general se corresponden al estado del arte en investigación en PLN y lingüística computacional:

En relación con el diseño de experimentos y la recopilación de datos, utilizaremos metodologías propias de los estudios en semántica (Cruse, 1986) y psicolingüística (Goldstone, 1994; Richie et al., 2020), con el fin de generar estímulos controlados. Así mismo, para recopilar anotaciones de informantes humanos utilizaremos métodos *crowdsourcing*, lo que nos permitirá obtener datos de hablantes nativos de forma rápida y eficiente, con control de calidad de las anotaciones (Munro et al., 2010).

A propósito de los modelos computacionales, aquéllos basados en arquitecturas Transformer se implementarán utilizando la biblioteca *transformers*[2], que incluye los últimos modelos basados en aprendizaje profundo. Opcionalmente usaremos otras bibliotecas de acceso abierto que tengan modelos adicionales. Para entrenar y ejecutar modelos estáticos utilizaremos *gensim*[3] y las propias herramientas publicadas por los autores para otros métodos distribucionales basados en dependencias sintácticas interpretables (e.g., Gamallo, de Prada Corral, y Garcia (2021)).

Por último, para comparar las representaciones de los modelos computacionales con los valores obtenidos de las anotaciones de los informantes utilizaremos tres métodos:

[2] `https://github.com/huggingface/transformers`

[3] `https://radimrehurek.com/gensim/`

1. Cálculo de precisión en evaluaciones con valores discretos (por ejemplo, homonimia o sinonimia, y en los resultados de clasificadores lineales).
2. Cómputo de la correlación en las evaluaciones graduales (polisemia o idiomaticidad).
3. Uso de *Representation Similarity Analysis* para observar si los modelos predicen las diferencias relativas entre ejemplos del mismo tipo (e.g., una palabra o MWE con el mismo significado en diferentes contextos) de manera similar a los humanos.

Cabe referir que las cuestiones metodológicas referidas ya han sido utilizadas en trabajos previos, que citamos brevemente a continuación.

2.1 Primeros resultados

A pesar de encontrarnos en una fase inicial disponemos ya de algunos resultados publicados, tanto de investigaciones previas directamente relacionadas con esta propuesta como de trabajos realizados desde el inicio del proyecto. Así, hemos presentado ya diversos conjuntos de datos con anotación de idiomaticidad semántica a nivel de token y de tipo en inglés y portugués, y evaluado varios modelos de lengua en ellos (Garcia et al., 2021a; Garcia et al., 2021b). Además, hemos creado un nuevo *dataset* en gallego-portugués, inglés y español que incluye ejemplos de homonimia y sinonimia en contexto, usado también para comparar diversos modelos y estrategias de contextualización (Garcia, 2021).

Más recientemente hemos comparado modelos Transformers y estrategias distribucionales basadas en dependencias sintácticas en tareas de composicionalidad semántica (Gamallo, de Prada Corral, y Garcia, 2021; Gamallo, Garcia, y de-Dios-Flores, 2022). Por último, hemos participado en la coorganización de la tarea *Multilingual Idiomaticity Detection and Sentence Embedding* (SemEval 2022), en la cual presentamos nuevos recursos con anotación de idiomaticidad semántica en contexto en gallego-portugués e inglés (Tayyar Madabushi et al., 2022).

3 Equipo de trabajo

El proyecto aquí presentado se realiza en el Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) de la Universidade de Santigo de Compostela, y está inserido dentro de su programa científico en Tecnologías de las Lenguas Naturales. En este sentido, miembros del centro pueden colaborar en diferentes tareas de nuestro plan de trabajo que formen parte de sus respectivas áreas de especialización.

Además del investigador principal, el proyecto cuenta con equipos de investigación y trabajo formados por tres doctoras/es, con especializaciones en Lingüística Computacional, Psicolingüística, Lógica y Ciencias de la Computación. En colaboración con un(a) investigador/a predoctoral y de personal técnico que será contratado con fondos propios, estos equipos participan activamente en las diferentes etapas del proyecto. Por último, contamos también con colaboraciones de personal investigador de otras universidades, tanto gallegas como internacionales, con quienes ya hemos participado en iniciativas y proyectos conjuntos de temática similar a la actual.

Agradecimientos

Proyecto financiado por la Xunta de Galicia (*Consolidación e estruturación de unidades de investigación competitivas e outras accións de fomento: Proxectos de Excelencia*, ED431F 2021/01) y por un contrato Ramón y Cajal (RYC2019-028473-I).

Bibliografía

Boleda, G. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, y P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12:2493–2537.

Cruse, D. A. 1986. *Lexical semantics*. Cambridge University Press.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the ACL (NAACL-HLT)*, páginas 4171–4186. ACL.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, páginas 1–32. Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952–1959, London: Longman (1968).

Gamallo, P., M. de Prada Corral, y M. Garcia. 2021. Comparing Dependency-based Compositional Models with Contextualized Word Embeddings. En *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021), Volume 2*, páginas 1258–1265.

Gamallo, P., M. Garcia, y I. de-Dios-Flores. 2022. Evaluating Contextualized Vectors from Large Language Models and Compositional Strategies. *Procesamiento del Lenguaje Natural*, 69.

Garcia, M. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. En *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, páginas 3625–3640. ACL.

Garcia, M., T. Kramer Vieira, C. Scarton, M. Idiart, y A. Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. En *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, páginas 2730–2741. ACL.

Garcia, M., T. Kramer Vieira, C. Scarton, M. Idiart, y A. Villavicencio. 2021b. Probing for idiomaticity in vector space models. En *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 3551–3564. ACL.

Goldstone, R. L. 1994. Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2):178.

Harris, Z. S. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Landauer, T. K. y S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013. Efficient estimation of word representations in vector space. En *Workshop Proceedings of the International Conference on Learning Representations 2013*.

Mikolov, T., W.-t. Yih, y G. Zweig. 2013. Linguistic regularities in continuous space word representations. En *Proceedings of the 2013 Conference of the North American Chapter of the ACL (NAACL-HLT)*, páginas 746–751. ACL.

Miller, G. A. 1971. Empirical methods in the study of semantics. *Semantics, an interdisciplinary reader in philosophy, linguistics, and psychology*, páginas 569–585.

Mitchell, J. y M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Munro, R., S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, y H. Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. En *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, páginas 122–130. ACL.

Padó, S. y M. Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, y L. Zettlemoyer. 2018. Deep contextualized word representations. En *Proceedings of the 2018 Conference of the North American Chapter of the ACL (NAACL-HLT)*, páginas 2227–2237. ACL.

Richie, R., B. White, S. Bhatia, y M. C. Hout. 2020. The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures. *Behavior research methods*, 52(5):1906–1928.

Schnabel, T., I. Labutov, D. Mimno, y T. Joachims. 2015. Evaluation methods for unsupervised word embeddings. En *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, páginas 298–307. Association for Computational Linguistics.

Tayyar Madabushi, H., E. Gow-Smith, M. Garcia, C. Scarton, M. Idiart, y A. Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. En *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Demonstrations
Demostraciones

ALIADA: Artificial Intelligence-based Language Applications for the Detection of Aggressiveness in Social Networks

ALIADA: Aplicaciones del Lenguaje basadas en Inteligencia Artificial para la Detección de la Agresividad en Redes Sociales

José Alberto Mesa-Murgado, Flor Miriam Plaza-del-Arco, Jaime Collado-Montañez, L. Alfonso Ureña-López, M. Teresa Martín-Valdivia
Departamento de Informática, CEATIC, Universidad de Jaén, España
{jmurgado, fmplaza, jcollado, laurena, maite}@ujaen.es

Abstract: In this paper, we present a Web Application Platform for the Detection of Aggressiveness in Social Media using Natural Language Processing and Machine Learning techniques, describing its architecture, the development technologies used and the different language models that have been integrated into the system. Finally, we conclude that the platform is a powerful tool to tackle real time aggressiveness on social media such as sexism or hate speech.
Keywords: Aggressiveness Detection, Web Application, Natural Language Processing, Machine Learning, Deep Learning

Resumen: En este trabajo presentamos una Plataforma de Aplicación Web para la Detección de Agresividad en Medios Sociales utilizando técnicas de Procesamiento del Lenguaje Natural y Aprendizaje Automático, describiendo tanto la arquitectura del sistema como las tecnologías que han sido utilizadas para su desarrollo y los modelos del lenguaje que se integran. Finalmente, concluimos que la plataforma es una herramienta potente para abordar la agresividad en las redes sociales en tiempo real, tales como el sexismo o el lenguaje de odio.
Palabras clave: Detección de la Agresividad, Aplicación Web, Procesamiento del Lenguaje Natural, Aprendizaje Automático, Aprendizaje Profundo

1 Introduction

The misuse of the Internet and specifically of social networks as a powerful tool for dialogue and participation, can lead to the creation, proliferation and dissemination of hate speech. According to the report on the evolution of hate crimes in Spain in 2020[1], Internet (45%) and social networks (22.8%) are the most used means for the commission of hate speech, with messages of ideology, racism/xenophobia, sexual orientation and gender identity showing the highest incidence. Threats, insults and public promotion/incitement to hatred, hostility, discrimination are computed as the most repeated criminal acts. Other communication channels where these acts are committed, but to a lesser extent, are telephony/communications (14.3%) and other sources of social communication (4.2%). The high incidence of these crimes on the Internet and social media shows the high need to combat them. Detecting this phenomenon can help to social media moderators to warn/block bullies and provide support to victims.

In the last years, offensive language research has emerged in the Natural Language Processing (NLP) area seeking to offer solutions to detect automatically this inappropriate behavior on the Web (Fersini, Rosso, and Anzovino, 2018; Aragón et al., 2019). The most recent and best-performing studies offer solutions based on neural networks for the detection of the different phenomena including misogyny and xenophobia (Plaza-del-Arco et al., 2020), sexism (Plaza-del-Arco et al., 2021c), cyberbullying (Elsafoury et al., 2021), aggression (Kumar et al., 2020),

[1] https://bit.ly/3611hm9

or offensive language (Plaza-del-Arco et al., 2021b; Plaza-del-Arco et al., 2021a). Some researches have shown that sentiment and emotion analysis are important features to consider in the detection of these phenomena (Rajamanickam et al., 2020; Plaza-del-Arco et al., 2021a; Plaza-del-Arco et al., 2021b). Although more and more studies are being conducted in this area, the integration of these automatic models to be used in real scenarios by any user is very scarce, especially in languages other than English, such as Spanish.

In this paper we present ALIADA, an artificial intelligence-based language application for the detection of aggressiveness[2] in social media. This application allows real-time monitoring of viral events on the social networks: Youtube and Twitter, integrating trained language models based on NLP solutions to identify aggressiveness on this content and visualizing the outcome to the user. In addition, to overcome the lack of language models available for offensive language research in Spanish, we have taken advantage of the majority of Spanish corpora that have been developed in this area to train different Machine Learning (ML) solutions for the detection of aggression in real-time data.

The rest of the paper is structured as follows: In Section 2 we provide a description of the tool and its architecture. Language models implemented are explained in Section 3. Finally, Section 4 presents conclusions and future work.

2 System Description

The ALIADA Web application consists of five internal modules that interact with each other to attend incoming requests and provide resources to relevant stakeholders (hereinafter, namely, users):

- **Data Storage Module**, based on ELK's Elasticsearch search engine it allows to index data under a non SQL approach.
- **Routing Module**, relies on the FastAPI framework to attend requests asynchronously using Python.
- **User Interface Module**, built using state-of-the-art web technologies such as HTML5, CSS3 (specifically, Bootstrap 5 as CSS framework) and Javascript.
- **Internal Logic Module**, implemented using Python manages data retrievals from social networking sites and the classification of incoming users requests.
- **Artificial Intelligence: Machine Learning Module**, built upon the Torch library for Python, allows to perform inferences in ML and Deep Learning models.

These modules are organized into Backend and Frontend, the former being responsible for routing and associated logic, and the latter of providing a graphical interface to interact with.

2.1 Backend

Encompasses the routing management and handling of incoming endpoint calls:

2.1.1 Stored Data and Storage Process

Information regarding users, their related personalization and data retrieval and classification requests, is stored in an Elasticsearch repository considering:

- The type of the submitted request: either data retrieval or classification.
- Social network used as source: Twitter or Youtube.
- The language model applied.

2.1.2 Data Retrieval and Extraction of New Data

Users can retrieve social data through requests, in which the social network used as source must be specified along with other search parameters: (1) who sent the post or (2) to whom it is targeted at, in which period of time it was published (3) or whether it includes an user provided keyword. Gathered data is anonymized before being stored in the Elasticsearch data warehouse in string format, structured as: (1) source, (2) corresponding source identifier, (3) parent source identifier, whether the publication is a response, (4) release date, and (5) associated textual content (tweet or comment).

Request's retrieved data can be downloaded in comma separated format (.csv) however, importing new data into a request is not allowed. At the same time, a request

[2]We use aggressiveness term to encompass different phenomena such as hate speech, sexism, misogyny, offense.

social data cannot be shared in other requests or by any other users distinct from their original requester who is allowed to run different ML classifying models against a same request in order to collect diverse statistics (e.g: in terms of sexism, offensiveness, hate speech, etc.).

2.1.3 User creation and management

Responses from the server require of authorized credentials that must be granted by an administrator, after requesting access through the contact form on the platform's homepage.

Users must be logged in to request and classify social data, this authorization is sent in each HTTP Request through Javascript Web Tokens (JWT) and serves two purposes: (1) security and (2) personalization.

2.1.4 Request Management

On the one hand, users' requests for data retrieval and classification are segmented into separated queues and serviced according to the date on which they were sent to the server along with a priority value that is reduced progressively as long as no new data is retrieved from the source, helping to determine when a certain topic is no longer relevant. On the other hand, the server traffic is handled asynchronously through FastAPI's uvicorn library which allows to run an ASGI Web server.

2.1.5 Data Classification and Procedure to Add New Models

Classification orders are associated to retrieval requests, they specify which ML model will be applied to the data and internally, they are ordered by the date in which they were sent to the server. Further on, the Pickle and Torch libraries are used to load the trained model architecture and state, as well as its associated vocabulary. Integrating new ML models into the server requires for the uploading of the trained model along with its corresponding word embeddings or bag-of-words structure and a categorical label dictionary to improve the comprehensibility of the model. A new function must be declared inside the Classifying module to load the model and use it against input data.

2.2 Frontend: User Interface

ALIADA provides a minimalistic web interface to make use of all of its features in a fast and intuitive way. Right after logging in from the main webpage, access to all the application's functions is provided: New data retrieval requests, statistics about the classification results, graphs of the total amount of downloaded posts, etc. In the following, these features are further described.

Dashboard. A dashboard (Figure 1) containing the current status of data retrieval and classification requests is displayed. Here, the client can see an ApexCharts' graph[3] that plots the total amount of data downloaded in a given time period, a list containing all active requests and a button to create a new one. Clicking on this button will pop up a form with all the information required to send a new data retrieval request as shown in Figure 2.

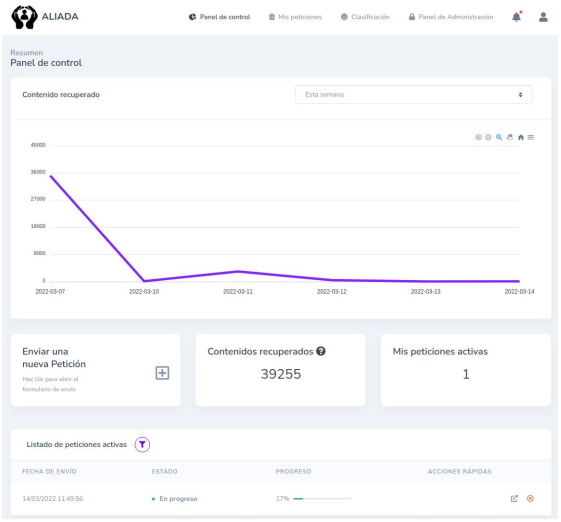


Figure 1: Dashboard.

My requests and classification panel. In order to have a more in-depth view of active and completed requests, two different sections are provided: my requests and classification panel. The former shows the current state (queued, in progress or completed) of each data retrieval request, while the latter shows the classification results in the form of graphs as seen in Figure 3. This section also shows all anonymized texts with their predicted labels, some information about the data retrieval and buttons to both download the full retrieved corpus as a .csv file and reuse the data to infer new labels with a different ML model.

Administrator. Finally, only users with the administrator role have access to the ad-

[3] https://apexcharts.com/

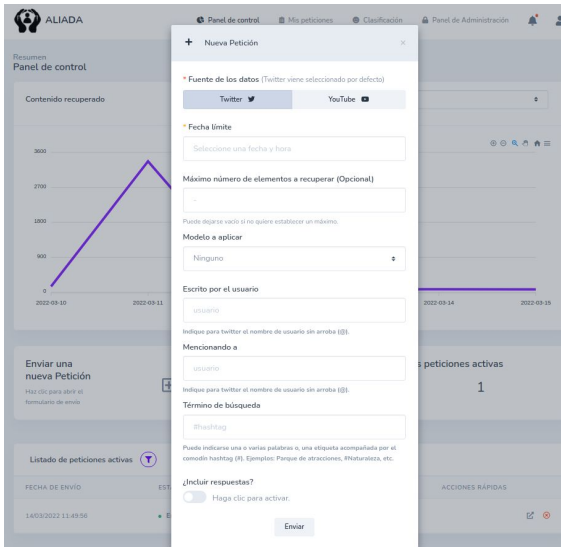


Figure 2: New request form.

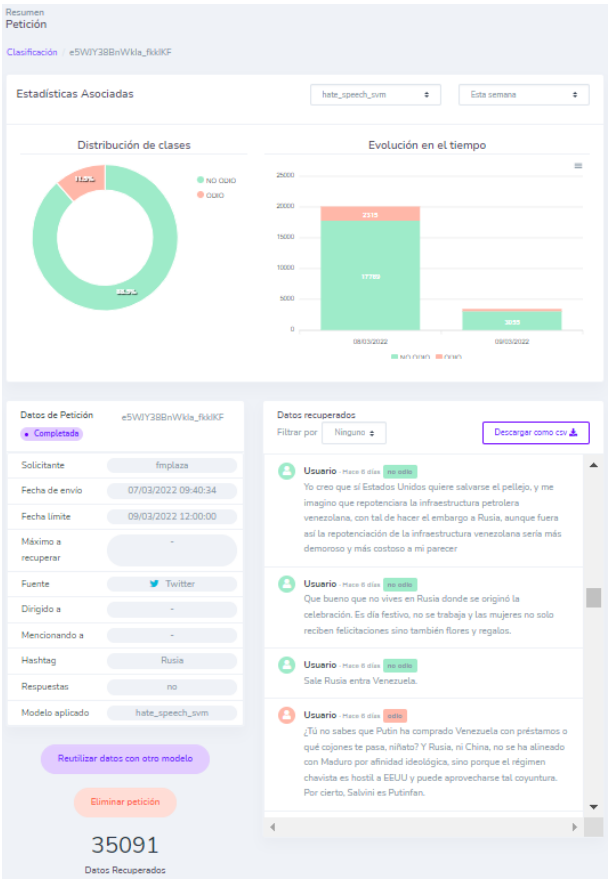


Figure 3: Classification results.

ministration panel. Here, an administrator can see the application's log history or the list of active requests in real-time. Users can also be created and deleted from this panel.

3 Language Models

The main objective of ALIADA is to monitor social media posts for the detection of aggressive content. Therefore, it is necessary to integrate different ML solutions to detect this behavior. Specifically, we have trained different models based on SVM for the detection of three phenomena: hate speech, sexism, and offensiveness.

In order to train these solutions, we have taken into account most of the available corpora generated for aggressiveness detection in Spanish including HatEval (Basile et al., 2019), HaterNet (Pereira-Kohatsu et al., 2019), EXIST (Rodríguez-Sánchez et al., 2021), NewsCom-TOX (Taulé et al., 2021) and OffendES (Plaza-del-Arco et al., 2021b). A total of four models are available in the platform: *hate_speech_svm* has been trained on HatEval, HaterNet and NewsCom-TOX datasets, *offendes_svm* has been trained on the large OffendES dataset, *sexism_svm* is trained on the EXIST dataset and finally *all_concepts_svm* combine all of the datasets.

4 Conclusions and Future Work

ALIADA is a powerful and useful tool to tackle aggressiveness in social networking sites in real-time, allowing for the detection of such attitudes in social publications through ML algorithms. In the near future, we would like to go further and, in addition to post classification, we will develop an explainability tool in order to understand what sections within each post makes it more aggressive than others through what is known as Named Entity Recognition (NER) techniques, and an emotion or performance tool to determine which attitude causes a greater effect in terms of its associated social reactions (namely, likes and retweets).

Acknowledgements

This work has been partially supported by the grants 1380939 (FEDER Andalucía 2014-2020), P20_00956 (PAIDI 2020) funded by the Andalusian Regional Government, LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, and the scholarship (FPI-PRE2019-089310) from the Ministry of Science, Innovation and Universities of the Spanish Government.

References

Aragón, M. E., M. Álvarez Carmona, H. J. Escalante, L. Villaseñor-Pineda, and D. Moctezuma. 2019. Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. page 17.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. Rangel, P. Rosso, and M. Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

Elsafoury, F., S. Katsigiannis, S. R. Wilson, and N. Ramzan, 2021. *Does BERT Pay Attention to Cyberbullying?*, page 1900–1904. Association for Computing Machinery, New York, NY, USA.

Fersini, E., P. Rosso, and M. Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. page 15.

Kumar, R., A. K. Ojha, S. Malmasi, and M. Zampieri. 2020. Evaluating Aggression Identification in Social Media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France, May. European Language Resources Association (ELRA).

Pereira-Kohatsu, J. C., L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados. 2019. Detecting and Monitoring Hate Speech in Twitter. *Sensors*, 19(21):4654.

Plaza-del-Arco, F. M., M. Casavantes, H. Escalante, M. T. Martin-Valdivia, A. Montejo-Ráez, M. Montes-y-Gómez, H. Jarquín-Vásquez, and L. Villaseñor-Pineda. 2021a. Overview of the MeOffendEs task on offensive text detection at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67(0).

Plaza-del-Arco, F. M., S. Halat, S. Padó, and R. Klinger. 2021b. Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language. *CoRR*, abs/2109.10255.

Plaza-del-Arco, F. M., M. D. Molina-González, L. A. U. López, and M. T. Martín-Valdivia. 2021c. Sexism Identification in Social Networks using a Multi-Task Learning System. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 491–499. CEUR-WS.org.

Plaza-del-Arco, F.-M., M. D. Molina-González, L. A. Ureña López, and M. T. Martín-Valdivia. 2020. Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *ACM Trans. Internet Technol.*, 20(2), mar.

Plaza-del-Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021a. A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis. *IEEE Access*, 9:112478–112489.

Plaza-del-Arco, F. M., A. Montejo-Ráez, L. A. Ureña-López, and M.-T. Martín-Valdivia. 2021b. OffendES: A New Corpus in Spanish for Offensive Language Research. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1096–1108, Held Online, September. INCOMA Ltd.

Rajamanickam, S., P. Mishra, H. Yannakoudakis, and E. Shutova. 2020. Joint Modelling of Emotion and Abusive Language Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online, July. Association for Computational Linguistics.

Rodríguez-Sánchez, F., J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. 2021. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67(0).

Taulé, M., A. Ariza, M. Nofre, E. Amigó, and P. Rosso. 2021. Overview of the DETOXIS Task at IberLEF-2021: DEtection of TOXicity in comments In Spanish. *Procesamiento del Lenguaje Natural*, 67.

Exploring gender bias in Spanish deep learning models

Exploración del sesgo de género en modelos de aprendizaje profundo en español

Ismael Garrido-Muñoz,[1] **Arturo Montejo-Ráez,**[2] **Fernando Martínez-Santiago**[3]
CEATIC - Universidad de Jaén, Spain
[1]igmunoz@ujaen.es, [2]amontejo@ujaen.es, [3]dofer@ujaen.es

Abstract: This paper presents a data visualization tool developed during the investigation of the bias present in deep learning language models in Spanish. The tool allows us to explore in detail the outcome of the response of the models we present with a set of template sentences, allowing us to compare the behavior of the models when the templates are presented with a context that alludes to a man or a woman. The exploration of the data in the tool is performed at various levels of detail, from visualizing the model output itself with its weights to visualizing the aggregation of the results by categories. It will be this last visualization that will provide some interesting conclusions about how the models perceive mainly women by their bodies and men by their behavior.
Keywords: bias, gender, deep learning, nlp.

Resumen: En este trabajo se presenta una herramienta de visualización de datos desarrollada durante la investigación del sesgo presente en modelos del lenguaje de aprendizaje profundo en castellano. La herramienta permite explorar con detalle el resultado la respuesta de los modelos a un conjunto de frases plantilla, permitiéndonos comparar el comportamiento de los mismos cuando las plantillas se presentan con un contexto que alude a un hombre o una mujer. La exploración de los datos en la herramienta se realiza a varios niveles de detalle, desde visualizar la salida propia del modelo con sus pesos hasta visualizar el agregado de los resultados por categorías. Será esta última visualización la que aportará unas interesantes conclusiones sobre cómo los modelos perciben preferentemente a las mujeres por su cuerpo y a los hombres por su comportamiento.
Palabras clave: sesgo, genero, aprendizaje profundo, pln.

1 Introduction

In recent years, deep learning models have been gaining popularity, these models are capable of capturing reality with great detail since they are trained from large volumes of data. However, not everything is good in these models, one of their weaknesses is that they work as black boxes. This means that when the model behaves erroneously, it is not possible to correct its behavior or even to know what has caused it or if that error may be occurring with other inputs. Thus the proposed tool fits into the novel fields of explainability, explainable artificial intelligence and fairness. The tool is freely available online[1].

[1]https://isgarrido.github.io/categoryviewer/

2 On biases and fairness

Since these models are so good at capturing reality, they also capture and replicate undesirable stereotypes. One example is the police COMPAS system in the United States. This system assigns detainees a level of risk of recidivism. From an independent analysis, it was discovered that the system failed for both whites and blacks(Julia Angwin, 2016), but the type of error was different. In the case of whites the system would systematically provide a lower level of recidivism risk than the actual level, it was failing in their favor. While in the case of blacks the error was against them, the system assigned a higher level of risk than the actual level. In this case we can talk about a social problem

in which an algorithm can be disruptive in people's lives and simultaneously we also talk about a system whose malfunctioning causes resources not to be allocated where they are really needed(Berkeley et al., 2019). A similar example can be found in a medical system called Optum, which would systematically allocate black patients less resources for their treatment than white patients for the same level of need. This is a case of resource allocation by a biased system can negatively influence people's health. We also have multiple examples in automated recruitment systems such as HireVue(Harwell, 2019) which uses artificial intelligence models to evaluate candidates. However, the system disadvantaged candidates who deviated from the model's definition of normal. This behavior is quite frequent, if the model is trained with examples that are not sufficiently varied, it will not be able to perform adequately when applied to cases for which it has not been trained. In this case it is intuited that HireVue malfunctioned on non-native candidates, since their accent would confuse the model. In itself it is not a problem that a model does not work initially for all cases, the problem comes when the candidate is automatically discarded and does not receive information about the reason. This makes us think that the application of non-explainable models may be unfair in some situations. Amazon also discarded(Dastin, 2018) a similar tool for recruitment, as it was found to be biased against women.

3 The problem of gender bias

In this paper we will focus on the bias in language models, specifically on the bias between men and women (gender bias). There are previous studies that show that language models do indeed capture significant differences between men and women, it is the work of Bolukbasi et al. (2016) the one that makes the first breakthroughs in this area. This work shows that the Word Embeddings model trained from Google News conceives men and women differently. After experimenting with professions, he highlights that the model creates associations such as *Man will be a computer programmer* while *Woman will be a homemaker*. Later the work of Caliskan, Bryson, and Narayanan (2017) will show that this bias is not only present on gender, but also other areas such as race. These types of differences will later be found in more complex models such as BERT(Bender et al., 2021) or RoBERTa(Sharma, Dey, and Sinha, 2021).

4 Proposed tool

The proposal that led to the creation of the proposed tool is the realization of a study on the bias in the main language models in Spanish. The main task is to know if gender bias is present in these models and try to characterize it. For the study we propose a series of template sentences that have a masked word, each template will have a masculine and a feminine version, the model will have to propose a set of words that would replace the masked word, as well as the probability of each word. We will have one set of words for the male version and another for the male version, which will allow us to compare how each version behaves. To focus the study we will use templates that should be completed with an adjective. For example,In the pair of sentences *El alumno es el más <mask>* and *La alumna es la más <mask>* for the first one the model suggests *rápido, inteligente, joven* while for the second template the sugestion are *joven, guapa, votada.*

We will obtain from each model, for each template a result with two metrics. The first is the internal **probability** of the model, the second one is a **RSV** (*ranked status value*) metric that represents the external state of the model, taking in this case the inverse position in the ranking. For example, if we get 5 results for each template, the first result will be the one with the highest probability and its RSV will be 5, the second element will be the one with the second highest probability and its RSV will be 4, and so on. The interest of the first metric is to know precisely the state of the model, while the second metric approximates what happens when a model is applied to a real use case, in which we do use the first N results with the highest probability ordered, independently of the weight of each result.

Subsequently, the adjectives proposed by the template will be categorized and the differences between male and female responses will be studied with the tool. Categories are based on two different classification schemes: the work of Tsvetkov et al. (2014) will appear under the name **Yulia** on the tool, and the work by Wiggins (1979) will be referred

as **Foa & Foa** on the tool.

The results of the analysis are exported to a JSON file and those JSON files are integrated into a web application. The application is a reactive Vue client web application, the tool loads the results of the experimentation and allows to explore graphically its results with help from ChartJs, for generating diagrams and charts.

4.1 Category viewer

From the charts tab you can choose a classification scheme, a model and a variable. Once chosen, the percentage of the words predicted by the model that fall into each category are displayed, in blue are shown the results for men, and in pink those for women. An interesting exploration is to choose the categorization **Yulia** and explore how systematically the value of the category **BODY** is higher for women, while the value of the category **BEHA** (Behaviour) is higher for men. This tells us that the models preferably associate women with attributes of their body while men with their behavior.

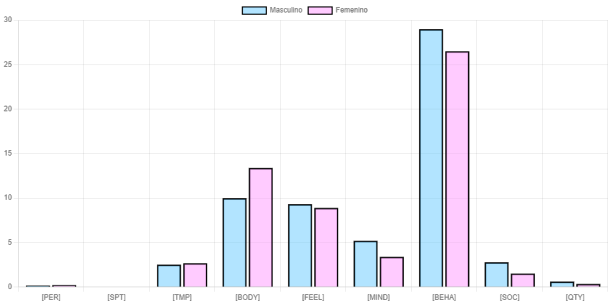


Figure 1: Category viewer

4.2 Tables

In the tables tab you can explore the results of the model from another perspective. In this case we select the categorization, the category to explore and what results we want to show in the table. The most interesting visualization is "M-F Heat" which will show the aggregate value for male minus female and color the table as a heatmap, with the extreme value of each column being red for female and blue for male.

This will allow us to see at a glance whether the leaning in that category is towards male or female, or neither in particular. In addition we will be able to see which models have a higher level of bias given the color intensity. By default we have the RSV and Probability columns that show the external and internal state of the model, this will allow us to appreciate significant differences in some cases. Here we can open the recommended configuration of the table above and see how the *Yulia - Body - M-F Heat* table is mostly red, while the *Yulia - Beha - M-F Heat* table is mostly blue.

Yulia | [BODY] | M-F Heat

	% RSV	% Probability
BSC-TeMU/roberta-base-bne	-3.40	-2.66
BSC-TeMU/roberta-large-bne	-5.86	-4.50
dccuchile/bert-base-spanish-wwm-uncased	-7.69	-13.64
dccuchile/bert-base-spanish-wwm-cased	-9.88	-9.34
mrm8488/electricidad-base-generator	-7.95	-8.07
MMG/mlm-spanish-roberta-base	-3.86	-3.60
bertin-project/bertin-roberta-base-spanish	-0.12	1.97
bert-base-multilingual-cased	-6.18	-6.69
bertin-project/bertin-base-random	-3.22	-0.21
bertin-project/bertin-base-stepwise	-1.96	-2.96
bertin-project/bertin-base-gaussian	-0.12	1.97
bertin-project/bertin-base-random-exp-512seqlen	-3.07	-3.53
bertin-project/bertin-base-stepwise-exp-512seqlen	-1.97	-0.43
bertin-project/bertin-base-gaussian-exp-512seqlen	-3.24	-4.14
amine/bert-base-5lang-cased	-6.23	-7.11
Geotrend/bert-base-es-cased	-7.26	-7.68
BSC-TeMU/RoBERTalex	-0.96	-1.00
Recognai/distilbert-base-es-multilingual-cased	-3.04	-2.70
flax-community/alberti-bert-base-multilingual-cased	-1.10	-5.97
Geotrend/distilbert-base-es-cased	-2.93	-1.38
Min	-9.88	-13.64
Max	-0.12	1.97

Figure 2: Tables snapshot

4.3 Adjective Stats

In the Adjective Stats tab you can study the adjectives obtained over the total number of words proposed by the model. The interest of this tab is simply to be aware that a model yields a very low proportion of adjectives, so we suspect that given the data used in its training it may not allow us to study the bias in the model. On the other hand we can also see which models are the best performing for this type of task, as well as look for significant differences in the number of adjectives proposed by each one.

4.4 Explorer

In the Explorer tab we can explore the adjectives proposed by each model for each sentence, both for the male and female versions.

5 Future work

The tool can be used in different ways. From a research point of view, extending this type of tests to other domains such as race would imply that instead of having two dimensions (male/female) we would have multiple and

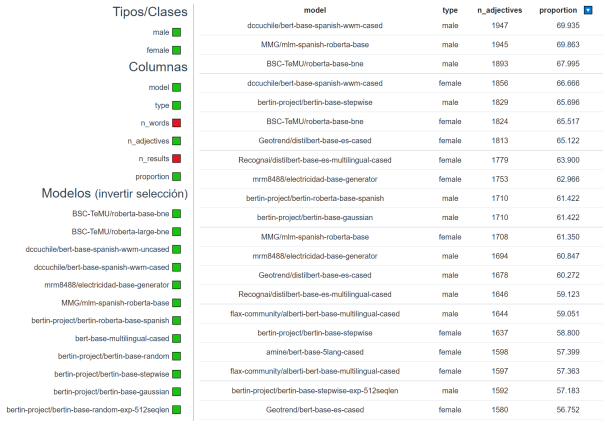


Figure 3: Adjective Stats snapshot

Figure 4: Explorer snapshot

would have to adapt them. It would also be interesting to incorporate capabilities to load results from a remote URL or just drag and drop a local file, allowing that, once the experimental code is released, anyone can use the visualization tool as easily as possible.

Finally, it would be interesting to convert the tool into a complete client side application that puts a GUI not only to the results but also allows to graphically launch experiments through a connection with the experimentation software and to feeds back its results by incorporating them into the visualizations, so to speak, a *no-code* solution for bias analysis.

6 Acknowledgements

This work is partially funded by grant P20_00956 (PAIDI 2020) from the Andalusian Regional Government and by grant RTI2018-094653-B-C21 for project LIVINGLANG by the Spanish Government.

References

[Bender et al.2021] Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

[Berkeley et al.2019] Berkeley, Z. O. U., Z. Obermeyer, U. Berkeley, S. M. U. o. Chicago, S. Mullainathan, U. o. Chicago, and O. M. A. Metrics. 2019. Dissecting racial bias in an algorithm that guides health decisions for 70 million people: Proceedings of the conference on fairness, accountability, and transparency, Jan.

[Bolukbasi et al.2016] Bolukbasi, T., K. Chang, J. Y. Zou, V. Saligrama, and A. Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *CoRR*, abs/1607.06520.

[Caliskan, Bryson, and Narayanan2017] Caliskan, A., J. Bryson, and A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186, 04.

[Dastin2018] Dastin, J. 2018. Amazon scraps secret ai recruiting tool that showed bias against women, Oct.

[Harwell2019] Harwell, D. 2019. A face-scanning algorithm increasingly decides whether you deserve the job, Nov.

[Julia Angwin2016] Julia Angwin, J. L. 2016. Machine bias - there's software used across the country to predict future criminals. and it's biased against blacks., May.

[Sharma, Dey, and Sinha2021] Sharma, S., M. Dey, and K. Sinha. 2021. Evaluating gender bias in natural language inference. *CoRR*, abs/2105.05541.

[Tsvetkov et al.2014] Tsvetkov, Y., N. Schneider, D. Hovy, A. Bhatia, M. Faruqui, and C. Dyer. 2014. Augmenting English Adjective Senses with Supersenses. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4359–4365, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

[Wiggins1979] Wiggins, J. S. 1979. A psychological taxonomy of trait-descriptive terms: The interpersonal domain. 37(3):395–412.

COBATO. Un chatbot orientado a asistir al pequeño comercio

COBATO. A chatbot aimed at assisting small businesses

Clara Díaz-Ruíz[1], Fernando Martínez-Santiago[2], Arturo Montejo-Ráez[2], Maria Teresa Martín-Valdivia[2], L. Alfonso Ureña-López[2], Manuel Carlos Diaz-Galiano[2], Miguel Angel Garcia-Cumbreras[2], Manuel García-Vega[2], Flor Miriam Plaza-del-Arco[2], Salud María Jiménez-Zafra[2], María Dolores Molina-González[2]
Grupo SINAI, Departamento de Informática, Universidad de Jaén
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaen, Spain
[1]cdr00008@red.ujaen.es,
[2]{dofer, amontejo, maite, laurena, mcdiaz, magc}@ujaen.es
[2]{mgarcia, fmplaza, sjzafra, mdmolina}@ujaen.es

Resumen: Se presenta COBATO, un chatbot cuyo dominio es el pequeño comercio y que tiene WhatsApp como canal de comunicación. La finalidad de COBATO es asistir al comercial en aquellas necesidades de información que plantean los clientes y que usualmente se resuelven vía telefónica o algún servicio de mensajería. Así, el asistente virtual facilita al cliente información de productos, horarios, datos de contacto, además de anotar pedidos. En el ámbito del Procesamiento del Lenguaje Natural, se aporta un modelo de datos basado en un grafo de conocimiento que aglutina toda la información que el chatbot requiere del dominio de la aplicación. Una segunda aportación es una representación formal basada en marcos gramaticales del lenguaje que el chatbot conoce. Estos son utilizados para el análisis semántico, así como para generar ejemplos de respuestas de usuario con las que entrenar el modelo de lenguaje usado en el flujo de compresión del lenguaje del chatbot.
Palabras clave: asistente virtual, chatbot, GF, modelos del lenguaje, PLN.

Abstract: COBATO is a chatbot intended by small commerce domain using WhatsApp as a communication channel. The purpose of COBATO is to assist the salesperson in order to provide information that is usually solved via telephone or a messaging service. Thus, the virtual assistant provides the customer with information on products, opening hours, contact details, as well as taking orders. In the field of Natural Language Processing, a data model based on a knowledge graph is proposed, which brings together all the information that the chatbot requires from the application domain. Additionally, a formal representation based on grammatical frameworks of the language that the chatbot knows is obtained. Subsequently, these are used for the semantic analysis of the user response. The fine-tuning of probabilistic language models is achieved by means of examples generated with the grammar.
Keywords: virtual assistant, chatbot, GF, language models, NLP.

1 Introducción

Con frecuencia, el limitado aforo del pequeño comercio conlleva largas esperas y colas para realizar compras domésticas cotidianas. En el contexto del pequeño comercio, el comercio de barrio, un modo de mitigar estas colas es realizar pedidos bien por teléfono o enviando un simple mensaje por WhatsApp, de modo que el cliente se acerca a por el pedido cuando este está confeccionado. Se propone

el desarrollo de un chatbot, denominado COBATO, con el objetivo de apoyar al comercio en tareas propias de la atención al cliente: facilitar información del comercio y de productos así como la elaboración de encargos. COBATO está diseñado como una suerte de intermediario entre el cliente y el dependiente de modo que los tres actores comparten un mismo canal, WhatsApp en este caso. Ya en el ámbito específicamente del Procesamiento del Lenguaje Natural, o PLN, las principales aportaciones de COBATO se resumen en los siguientes puntos:

- Uso de marcos gramaticales (GF, *Grammatical Frameworks*)(Ranta, 2004). Concretamente, los GF proveen de un analizador semántico especializado al dominio de la aplicación. Sin embargo, este analizador, si bien es muy preciso, presenta una cobertura limitada, por lo que es necesario apoyarse en modelos del lenguaje como *word embeddings* o RoBERTa. A estos, la gramática implementada en GF provee de ejemplos de frases de usuario que son generados mediante un proceso denominado "linearización". Estos ejemplos están etiquetados con la acción y las entidades que allí se encuentran, y son utilizados para el ajuste o *fine-tuning* del modelo del lenguaje.
- Grafos de Conocimiento (KG, *Knowledge-Graphs*)(Hogan et al., 2021), como modelo único para la representación del conocimiento, y que encapsula tanto el modelo de datos del dominio de la aplicación como el conocimiento lingüístico, el cual será interpretado por los GF, previa traducción automática de un modelo de datos a otro.

2 Solución propuesta

En la línea de (Fensel et al., 2020a), la energía (Fensel et al., 2020b) y (Christmann et al., 2019), se propone un grafo de conocimiento para enlazar toda la información, como un único formalismo que representa (i) la base de datos, donde se codifica la información estructurada que se desea hacer pública, y (ii) conocimiento lingüístico, que se representa como un conjunto de entidades y conceptos (nodos) y relaciones entre ellos (arcos). Este conocimiento lingüístico es posteriormete trasladado a GF, los cuales proveen de un compilador capaz de realizar análisis semánticos del texto, a la par que generar expresiones que son plausibles conforme la gramática codificada. Como se indicó en la introducción, estas facilidades son aprovechadas tanto para obtener una interpretación muy precisa de la respuesta de usuario como para el ajuste fino del modelo de lenguaje que se requiere en el flujo PLN del chatbot. A continuación se detallan las tecnologías que forman parte de la arquitectura de COBATO (ver Figura 1). COBATO requiere de diversas tecnologías, que se agrupan en servicios de back-end, servicios de front-end, así como ciertos recursos externos utilizados para dotar de conocimiento lingüístico a los servicios de back-end.

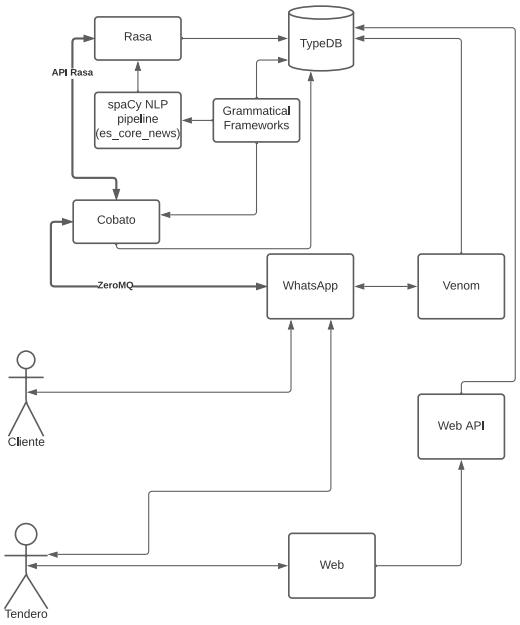


Figura 1: Arquitectura general de COBATO

- Servicios de front-end
  - *React*. Front-end Web para la gestión de la página web que usará el comercio : perfil de negocio, horarios, productos, disponibilidad y precios.
  - *Venom*. Front-end WhatsApp, pasarela entre el chatbot y los dos perfiles de usuario de este, cliente y comercio.
- Servicios de back-end

  - *Node.js*. Actúa como una capa middleware entre la interfaz web y la base de datos, un grafo de conocimiento implementado en TypeDB.
  - *Python*. Es el lenguaje de programación genérico usado para implementar diversas librerías para comunicar los servicios de PLN, front-end y back-end.
  - *Rasa. Framework* para la implementación de asistentes conversacionales basados en texto.
  - *TypeDB*. Gestor de Base de Datos que toma como modelo conceptual de datos el modelo Entidad-Relación, y que se implementa mediante un modelo lógico de datos basado en hipergrafos. A diferencia de otros DBMS basados en hipergrafos, TypeDB requiere de un esquema de datos definido.

- Recursos externos.
  - *Grammatical Frameworks*. Lenguaje de propósito especial que se compila en última instancia en una gramática multilingüe, que consta de una sintaxis abstracta y un conjunto de sintaxis concretas. La sintaxis abstracta define un sistema de árboles sintácticos, y las sintáxis concretas completan la gramática codificando la correspondiente información morfo-sintáctica, morfológica y léxica, particular de cada lenguaje.
  - *Spacy NLP es_core_news*. Modelo de lenguaje usado para el flujo de PLN para procesar la entrada del usuario: extracción de entidades, clasificación de la acción (intención) de usuario, tokenización, similitud semántica de términos, etc.

3 Un caso de uso: las fruterías

En esta primera versión COBATO se ha adaptado al caso concreto de las fruterías, identificándose diez casos de uso principales, relativo a la elaboración de un pedido. Otros casos contemplados refieren el registro y gestión de contenido de comercio a través de la aplicación web, o gestionar ofertas a través de un canal privado entre el comercio y el chatbot. Tomando al cliente como actor principal destaca el registro del canal de WhatsApp y la gestión de pedidos. En relación al canal de WhatsApp para cada cliente, durante el proceso de registro el comercio obtiene un enlace junto con el código QR equivalente. Cuando un cliente desea interactuar con el chatbot solo necesita escanear con su móvil tal código QR. Automáticamente se crea un canal en el cual son miembros el mismo cliente, el comercio y COBATO. Nótese que, en consecuencia, las interacciones cliente-COBATO son igualmente accesibles por el comercio, que puede intervenir en cualquier momento. Un segundo ejemplo destacado desde la perspectiva del cliente es la atención a un pedido (Figura 2). Se corresponde con la transacción a lo largo de la cual el chatbot solicita al cliente qué productos necesita, y en qué cantidad. Una vez el cliente desee finalizar, la transacción queda en estado "pendiente", hasta que, eventualmente, el comercio confirme el pedido en el mismo canal WhatsApp que comparte con el cliente y COBATO, pasando entonces el pedido a "atendido".

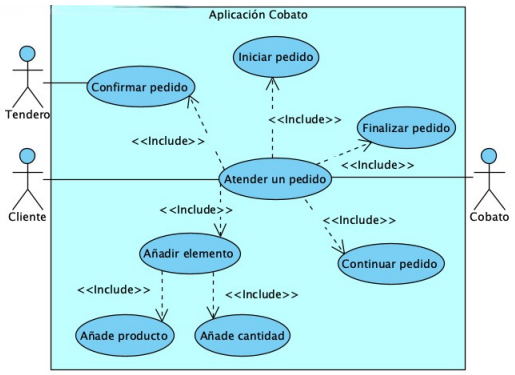


Figura 2: Caso de uso: atender un pedido

Previo al uso de COBATO por parte del cliente, es necesaria su instanciación y puesta en marcha: la BBDD implementada en TypeDB ha sido instanciada con los productos disponibles (frutas y verduras en este caso). También se añade información léxica, sintáctica y semántica del lenguaje entendido por COBATO, a modo de lenguaje controlado en el ámbito de la aplicación, y que representa las posibles interacciones cliente-chatbot expresadas en un lenguaje natural controlado conforme los casos de uso identificados. Ejemplos de expresiones legales en este lenguaje controlado son "cuál es el horario en fin de semana", "¿cuál es la dirección del comercio?", "añade un kilo de producto " o "¿qué precio tiene producto ", don-

de "producto" es cualquier de los productos dados de alta en la BBDD, frutas y verduras en este caso. Posteriormente, este lenguaje controlado es traducido a un programa GF, una gramática, mediante un script escrito en Python. Como se indicó en la sección 2, los GF son muy precisos para el análisis semántico conforme la gramática especificada, pero a su vez limita el lenguaje que es posible analizar. Es por ello que usualmente los *frameworks* tales como Rasa utilizan modelos de lenguaje, de corte probabilístico, en aras de alcanzar una mayor cobertura aun a costa de pérdida de precisión. Aún en este escenario los GF son de gran utilidad: mediante un proceso denominado linearización, GF genera todos los posibles árboles gramaticales plausibles conforme el programa GF escrito. De este modo obtenemos un conjunto de ejemplos con el cual ajustar al ámbito concreto de esta aplicación el modelo de lenguaje que usa Rasa (*es_core_news* en este caso), como parte de su flujo de PLN. Particularmente, estos ejemplos están etiquetados con cada una de las acciones de usuario conforme los casos de uso identificados, así como con las entidades relevantes que allí se encuentran, principalmente frutas y verduras. Finalmente, como parte de esta puesta en marcha de COBATO se han escrito diversos "scripts" para Rasa, de modo que este pueda gestionar la lógica de diálogo y generar las respuestas de usuario necesarias. Una vez COBATO está en funcionamiento, el comercio se ha registrado en el sistema, y el cliente ha usado el código QR correspondiente, este último puede empezar a interactuar con COBATO, y con el dependiente, a través de un canal WhatsApp. Las solicitudes de cliente son, en primera instancia, atendidas por el front-end Venom, el cuál actúa como pasarela entre WhatsApp y Rasa, el framework que finalmente atenderá las solicitudes de usuario. La respuesta que Rasa proporciona al usuario se confecciona conforme la intención de este y el KG almacenado en TypeDB. Para poder identificar la intención de usuario, previamente Rasa es entrenado acorde los guiones escritos a tal efecto y el modelo de lenguaje previamente ajustado con los ejemplos provistos por GF.

4 Conclusiones y trabajo futuro

Se propone el desarrollo de un chatbot cuyo conocimiento se codifica en KG y GF, de modo que toda la información queda en un único repositorio, con las ventajas que ello conlleva desde el punto de vista de la integridad y consistencia de los datos, además de facilitar el mantenimiento, escalado y migración del sistema. El proyecto está en fase de prueba, con lo que el siguiente paso será probarlo en entornos reales. A medio plazo, se pretende avanzar en el uso de los KG. Concretamente, se debe incluir en este conocimiento para la gestión del flujo de diálogo y la generación de las respuestas automáticas. Ambos aspectos actualmente son implementados en Rasa. Separar completamente el *framework* del modelo de conocimiento permitirá el desarrollo de un gestor de diálogo basado íntegramente en grafos.

Agradecimientos

Este proyecto es parcialmente financiado con fondos de la Oficina de Transferencia de Resultados de la Investigación de la Universidad de Jaén y del Instituto de Estudios Giennenses, área de conocimiento de Ciencias Naturales y Tecnología, así como por el Gobierto español a través del proyecto RTI2018-094653-B-C21, LIVING-LANG.

Bibliografía

Christmann, P., R. Saha Roy, A. Abujabal, J. Singh, y G. Weikum. 2019. Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion. En *28 CIKM*, páginas 729–738.

Fensel, A., Z. Akbar, E. Kärle, C. Blank, P. Pixner, y A. Gruber. 2020a. Knowledge graphs for online marketing and sales of touristic services. *Information*, 11(5):253.

Fensel, D., U. Şimşek, K. Angele, E. Huaman, E. Kärle, O. Panasiuk, I. Toma, J. Umbrich, y A. Wahler. 2020b. Why we need kg: Applications. En *Knowledge Graphs*. Springer, páginas 94–123.

Hogan, A., E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, y others. 2021. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37.

Ranta, A. 2004. Grammatical framework. *Journal of Functional Programming*, 14(2):145–189.

Plataforma de exploración de la Composición Semántica a partir de Modelos de Lenguaje pre-entrenados y embeddings estáticos

Platform for exploring Semantic Composition from pre-trained Language Models and static embeddings

Adrián Ghajari, Víctor Fresno
Enrique Amigó
Universidad Nacional de Educación a Distancia, UNED
{aghajari,vfresno,enrique@lsi.uned.es}

Resumen: El crecimiento de la capacidad de procesamiento y el advenimiento del modelo Transformer han modificado el panorama del PLN. El proceso conocido como Transferencia de Aprendizaje ha facilitado la consecución de resultados cercanos al estado-del-arte a una fracción del coste computacional. En este ámbito, este artículo presenta una aplicación cliente-servidor capaz de obtener vectores contextualizados (o estáticos) de palabras dentro de textos y a partir de una gran cantidad de modelos pre-entrenados, realizar composición semántica para, finalmente, visualizar en un espacio tridimensional las representaciones obtenidas y estimar su similitud semántica; todo esto, explotando los recursos hardware disponibles.
Palabras clave: Composición semántica, Vectores de frases, Transformers.

Abstract: The computing power growth and the advent of the Transformer model have changed the NLP landscape. Transfer Learning has allowed the posibility of achieving state-of-the-art results at a fraction of the computational cost. In this scope, this work presents the development of a server-client application capable of obtaining contextual and static word vectors from a wide variety of models, operate with them to achieve semantic composition to, lastly, visualize them in a 3-dimensional space and obtain semantic similarity; all of this, while exploiting the hardware resources available.
Keywords: Semantic composition, sentence embedding, Transformers.

1 Introducción

La llegada del modelo Transformer (Vaswani et al., 2017) ha revolucionado el área del NLP, fundamentalmente por su capacidad de aplicar patrones aprendidos durante su entrenamiento sobre distintas tareas nuevas, aunque relacionadas, lo que se conoce como Transferencia de Aprendizaje (TA). Este modelo captura información sobre el contexto en el que se encuentra cada palabra dentro de una frase, generando representaciones vectoriales de las mismas como paso intermedio antes de un proceso de ajuste fino (fine tuning). Estas representaciones de palabras se pueden procesar para realizar Composición Semántica. El Principio de Composicionalidad está basado en que el significado del todo es una función del significado de sus partes y de cómo están sintácticamente combinadas; por su parte, el Principio de Contextualidad afirma que el significado de las unidades lingüísticas emerge del contexto en el que se usan. Se conoce como Composición semántica al proceso por el que se generan representaciones vectoriales de frases a partir de los significados individuales de sus palabras constituyentes y de cómo estas se combinan.

En este artículo se presenta una plataforma software[1] que realiza composición semántica a partir de modelos pre-entrenados de los repositorios de HuggingFace[2] (contextuales) y Gensim[3] (estáticos), utilizando textos facilitados por el usuario, siendo capaz de representar los resultados en el espacio tridimensional, permitiendo así la comparación de embeddings en diferentes tareas de similitud semántica (STS), o estudiar el efecto de la contextualidad en este tipo de problemas.

[1] https://github.com/adriangh-ai/AllSpark
[2] https://huggingface.co
[3] https://radimrehurek.com/gensim/

2 Motivación

El éxito de la TA ha traído consigo una sobrecarga de modelos; sólo en el repositorio de HuggingFace hay más de 34 mil almacenados. Asimismo, la composición semántica sobre la salida de los modelos de lenguaje se enfrenta a problemas tales como la degradación de representación (Ethayarajh, 2020), (Li et al., 2020), (Gao et al., 2019), o si realmente capturan la información semántica del texto de entrada. Esto hace abstrusa la evaluación y establecimiento de métricas, más allá de la inspección supervisada del resultado. Los objetivos de este trabajo son el análisis e implementación de mecanismos de composición semántica, con una interfaz que sirva como capa de abstracción para la búsqueda, obtención y almacenamiento de diversos modelos de representación basados en RNA como un marco de trabajo para el control y visualización de los diferentes modelos. Lo anterior provee de un mecanismo para el estudio de la estructura interna de modelos de lenguaje, al permitir observar representaciones provenientes de la salida de capas intermedias con las herramientas de reducción dimensional implementadas para la visualización 3D de embeddings n-dimensionales; así como el estudio comparativo de modelos preentrenados y métodos de composición mediante métricas de similitud semántica.

3 Funcionalidad y Caso de Uso

Su función principal es la obtención de representaciones vectoriales a partir del modelo de lenguaje neuronal descargado desde un repositorio. La selección de capas internas del modelo a partir de las cuales obtener la composición semántica mediante distintos mecanismos (suma, media aritmética, [CLS] token y F_{inf}, F_{joint}, F_{ind} de ICDS (Amigó et al., 2021)) y la posibilidad de procesar de forma concurrente y con paralelismo de datos el conjunto de muestra. Finalmente, la visualización de las frases en una gráfica 3D interactiva. Se trata de una aplicación que puede ejecutarse con independencia de despliegue del cliente y servidor, pudiendo encontrarse y explotar recursos en máquinas locales o estaciones de trabajo remotas y distintos sistemas operativos.

El usuario tendrá conocimiento esencial sobre modelos de lenguaje y composición y se intentan cubrir los siguientes casos de usos diferenciados:

- Inferencia sobre una muestra de datos dada. Obtención de la composición de uno o más conjuntos de frases, bien para su visualización o para su uso en otra tarea.

- Recuperación de sesión anterior. Volver a cargar una o varias sesiones anteriores para su visualización y comparación.

4 Arquitectura general

Modelo cliente-servidor multi-plataforma, con *back-end* escrito en Python y *front-end* en ElectronJS[4] y Plotly DASH[5] con comunicación remota basada en gRPC protobuf[6].

4.1 Servidor

El servidor contiene la lógica relacionada con la gestión de modelos y el procesamiento de la evaluación, así como la composición semántica, pudiendo ejecutarse en una máquina remota. Es quien implementa la definición de la interfaz gRPC para ofrecer servicios a clientes, gestiona el almacenamiento de modelos y mantiene la relación hardware del sistema. Por último, procesa la entrada de texto, inferencia y composición semántica.

- Módulo de sesión Realiza las tareas de inferencia y composición individuales mediante multiprocesamiento, haciendo uso de los módulos de modelos y composición. Se ha implementado paralelismo de datos instanciando el modelo en cada dispositivo con hilo exclusivo y dividiendo la carga a partes iguales entre dispositivos, mostrando mejor rendimiento frente a Pytorch DataParallel. A su vez, la técnica de Uniform Length Batching evita el procesamiento de tokens [PAD] innecesarios mediante ordenación y agrupación en batches de frases según longitud.

- Módulo de modelos y composición Instanciarán un objeto con los métodos necesarios para la inferencia y la composición que se asignarán a *workers*; se ha optado por la eliminación de los tokens que no se corresponden con una palabra o un fragmento de palabra con una función de limpieza que convierte en una máscara los identificadores de los tokens especiales. Asimismo, en caso de haber seleccionado un rango de capas para su procesado, se operará la media aritmética sobre los resultados de composición individuales por capa.

[4] https://www.electronjs.org/
[5] https://plotly.com/
[6] https://grpc.io/

4.2 Cliente

Contiene la interfaz de usuario y los métodos gRPC de comunicación con el servidor. Se ocupa del pre-procesamiento de los datos, así como la lógica que atiende a la reducción dimensional y representación de resultados. Se divide en dos bloques funcionales gobernados por ElectronJS y DASH, que se comunican entre ellos por HTTP mediante un Web Server Gateway Interface, Waitress[7], para proveer la interfaz. Tras el lanzamiento y la conexión a un servidor externo o ejecución y conexión local, se llega a la pantalla de ajuste y selección de parámetros de inferencia. En su inicio la aplicación cliente actualiza la relación de dispositivos y modelos disponibles del servidor, o para su descarga desde los almacenes de HuggingFace y Gensim.

- Pestaña Principal Los modelos disponibles (contextuales y estáticos) se ofrecerán en forma de lista predictiva al introducir texto en el área de búsqueda. Una vez descargados podrá elegirse la salida de una de las capas del mismo, o un rango de ellas. Asimismo, se ofrece para su selección una relación de métodos de composición semántica y la lista de dispositivos de computación encontrados en el servidor para su asignación.

El área de selección de archivo acepta diversos formatos: estructurados, como csv, json o excel, y texto desestructurado en txt. En este último caso, el sistema tratará de reconocer las frases que contiene el bloque de texto del archivo a través de la librería Natural Language Toolkit (NLTK)[8]. Finalmente, se podrán seleccionar columnas, que serán visualizadas en la misma gráfica con colores distintos por columna. Por otro lado, los archivos de sesiones anteriores guardados se pueden volver a cargar y visualizar.

Tras seleccionar la configuración completa, pueden añadirse a la lista de peticiones de inferencia. Es posible añadir y borrar cuantas peticiones se desee; pulsando el botón de lanzamiento de inferencia, serán procesadas en el servidor de forma concurrente en los dispositivos que cada uno tenga asignados.

- Pestañas de inferencia: Tras el proceso de evaluación, el servidor envía los datos al cliente, que los mostrará en una nueva pestaña de inferencia. Esta pestaña contiene elementos para el control de la visualización. Desde la misma es también posible guardar los vectores resultado del proceso de composición. Se ofrecen distintos métodos de reducción dimensional, t-SNE (Van Der Maaten y Hinton, 2008), Principal Component Analysis (Hotelling, 1933) y UMAP (Mcinnes, Healy, y Melville, 2020), seleccionables como subpestañas. Éstas contienen los elementos de ajuste de parámetros de cada uno de estos métodos. El resultado de estos métodos se mostrará en la gráfica, que ofrecerá una representación interactiva y tridimensional por color según columna de procedencia en la muestra original con los datos de cada frase. Seleccionando el método de similitud semántica, como similitud coseno, y un punto en la gráfica de representación, se mostrará una tabla con las 10 frases más cercanas en el conjunto de origen, así como la columna a la que pertenecen.

5 Ejemplo de uso

Es en esta pantalla (ver Figura1) el usuario puede seleccionar los dispositivos de computación a usar, descargar y seleccionar modelos y capas a procesar, ver la estimación de ocupación de memoria, elegir método de composición y cargar el archivo de muestras. Una vez realizada la selección, se procede a la inferencia, añadiendo los resultados a una nueva pestaña. Finalmente (ver Figura2), pueden elegirse distintos métodos de reducción dimensional (esquina superior izquierda), modificar sus parámetros de operación (izquierda), visualizar las frases más cercanas a un punto (tabla de similitud coseno, parte inferior de la imagen con la frase seleccionada marcada por una etiqueta) y guardar los resultados de la sesión. A modo de ejemplo, se ofrecen los puntos correspondientes a las columnas *hipótesis* (azul) y *premisa* (rojo) del conjunto de datos GLUE, subset *mnli*, según la última capa del modelo BERT; puede observarse empíricamente la posición relativa entre frases, distancia y agrupación, según la función de composición semántica, el modelo y capa del mismo elegidos.

6 Conclusiones y trabajos futuros

Este trabajo presenta una aplicación distribuida multiplataforma cuya finalidad es la asistencia a la investigación en el estudio de la composición a partir de modelos de lenguaje neuronales. Se han implementado algoritmos de composición semántica, reducción de di-

[7] https://github.com/Pylons/waitress
[8] https://www.nltk.org/

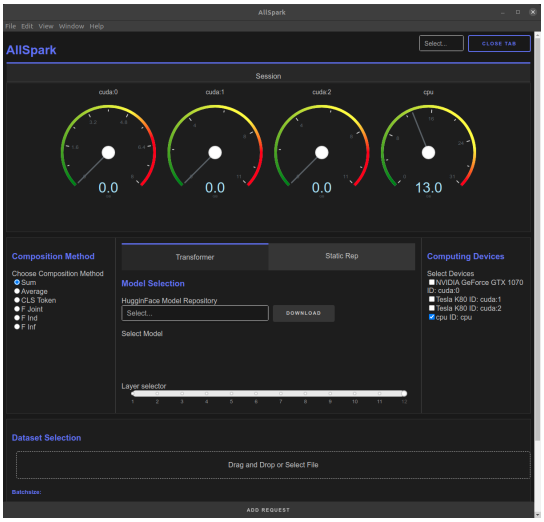


Figura 1: Pantalla principal.

Figura 2: Pestaña de inferencia.

mensionalidad y similitud semántica, además de técnicas de optimización de inferencia.

En lo relativo a futuras funcionalidades, se abordará el paralelismo de un solo modelo en inferencia, dividiendo el modelo en capas y repartiéndolas entre dispositivos. Adicionalmente, se pretenden implementar soluciones propuestas al problema de la isometría semántica de las representaciones, evidenciado en los trabajos de (Amigó et al., 2020) y (Ethayarajh, 2020). Finalmente, algunos estudios apuntan a la posibilidad de que distintas cabezas de auto-atención del modelo Transformer atienda a distintos aspectos semánticos, (Rogers, Kovaleva, y Rumshisky, 2020); aislarlos y generar su representación podría ofrecer una nueva perspectiva.

Agradecimientos

Este trabajo ha sido financiado por el proyecto del Ministerio de Ciencia y Innovación DOTT-HEALTH (PID2019-106942RB-C32), gracias al acuerdo UNED - Ministerio de Economía y Competitividad de España con ref. C039/21-OT, y al proyecto LyrAIcs a través del Consejo Europeo de Investigación (ERC, con Grant agreement Nº [964009]).

Bibliografía

Amigó, E., A. Ariza, V. Fresno, y M. A. Martí. 2021. Information-teoretic compositional distributional semantics (IN PRESS).

Amigó, E., F. Giner, J. Gonzalo, y M. Verdejo. 2020. On the foundations of similarity in information access. *Information Retrieval*, 23, Issue 3:216–254.

Ethayarajh, K. 2020. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMO, and GPT-2 embeddings. *EMNLP-IJCNLP 2019*, páginas 55–65.

Gao, J., D. He, X. Tan, T. Qin, L. Wang, y Y. Liu. 2019. Representation Degeneration Problem in Training Natural Language Generation Models. Informe técnico.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *J. of Educational Psychology*, 24:498–520.

Li, B., H. Zhou, J. He, M. Wang, Y. Yang, y L. Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv*.

Mcinnes, L., J. Healy, y J. Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Rogers, A., O. Kovaleva, y A. Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. Informe técnico.

Van Der Maaten, L. y G. Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Vaswani, A., G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin. 2017. Attention Is All You Need. Informe técnico.

Crossroads 2.0 - Juego educativo sobre el impacto del cambio climático con generación de lenguaje natural

Crossroads 2.0 - Learning game for climatic change awareness with generation of natural language

David Escudero-Mancebo, Adrián Santos-Manzano, Manuel Alda, Yania Crespo, María Robles
Dpto. de Informática
Universidad de Valladolid

Resumen: El cambio climático y la disponibilidad de recursos energéticos son problemas de dimensión planetaria con importantes implicaciones sociológicas y económicas. En este trabajo se presenta una aplicación educativa gamificada para la concienciación sobre la importancia del problema y la trascendencia de adoptar medidas políticas de alcance. El juego permite a los usuarios proponer medidas políticas para la mitigación del problema del cambio climático y genera los resultados de dichas medidas junto con una recomendación textual que describe el escenario alcanzado y propone medidas para mejorarlo. Esta comunicación muestra cómo la incorporación de un módulo de generación automática de lenguaje natural en el juego permite aportar una realimentación a los usuarios de manera eficiente.
Palabras clave: Juegos serios, cambio climático, generación de lenguaje natural

Abstract: Climate change and the availability of energy resources are problems of a planetary dimension with important sociological and economic implications. In this work, we present a learning game for awareness about the importance of the problem. The game allows users to propose political measures to mitigate the problem of climate change and generates the results of such measures together with a written recommendation that describes the scenario reached and proposed measurements for improving it. This project shows how adding a module for automatic natural language, allows the system to provide feedback in an efficient way.
Keywords: Learning games, climate change, natural language generation

1 Introducción

Crossroads 2.0 es la versión electrónica de un juego colaborativo previo cuyo fin es la concienciación sobre la necesidad de adoptar medidas políticas relevantes para frenar el impacto del cambio climático (Capellán-Pérez, Álvarez-Antelo, y Miguel, 2019). Concienciar sobre la transcendencia del problema es una empresa en la que la Unión Europea ha puesto todo su empeño durante los últimos años. La financiación de proyectos de investigación que arrojen luz sobre el problema o de actividades que permitan concienciar a la sociedad han sido líneas de actuación prioritarias durante los últimos años. Uno de esos proyectos es el proyecto H2020 Locomotion `https://www.locomotion-h2020.eu/` en el que se enmarca el trabajo presentado en esta comunicación.

En el proyecto Locomotion se desarrollan modelos sistémicos que relacionan un elevado número de variables de tipo económico, social y de recursos energéticos con el cambio climático y los índices de bienestar a nivel planetario. Como parte de las actividades del proyecto, está el desarrollo de aplicaciones informáticas que permitan poner en valor los modelos desarrollados. Una de estas aplicaciones es el juego Crossroads 2.0 presentado en esta comunicación.

El uso de juegos educativos para concienciar sobre el cambio climático no es una idea original, existen abundantes experiencias que han intentado explotar las capacidades cautivadoras de los juegos para implicar a los usuarios en la causa de la lucha contra el cambio climático (Wu y Lee, 2015; Fernández Galeote y Hamari, 2021). Más original es el uso que hacemos en el juego Crossroads 2.0 de técnicas de generación de lenguaje natural para aportar argumentos que sirvan de realimentación en el juego. Los mensajes, obtenidos

automáticamente a partir del análisis del conocimiento sobre el problema, sirven para dar argumentos informativos a los jugadores, que los utilizan para corregir sus propuestas.

El uso de lenguaje natural está muy extendido en los juegos educativos, pero generalmente se utilizan bases de datos con mensajes pre-elaborados, dependientes del contexto a los que se accede en función de una casuística que depende del juego (Johnson, Bailey, y Buskirk, 2017). En esta comunicación describimos primero la interfaz y la interacción del juego, después describimos la operativa del módulo de generación de lenguaje natural, detallamos la integración de dicho módulo en la arquitectura del juego y presentamos la estrategia de pruebas.

2 *Descripción del juego*

La figura 1 muestra la interfaz web y la dinámica de juego. Crossroads utiliza la metáfora de la sala de juego a la que entran los usuarios para participar en una partida. Se organizan grupos de trabajo que compiten entre ellos por aportar la mejor solución. Una cinemática inicial explica el objetivo del juego. Los jugadores asumen el papel de persona responsable de tomar decisiones para salvar el planeta del problema del cambio climático sin hundir la economía. Las medidas políticas se toman completando un formulario. Deben tomarse de forma colaborativa dentro del grupo para lo cual disponen de un chat y de información sobre las medidas elegidas por sus compañeros de grupo. Una vez llegado a un consenso, el equipo ve los resultados alcanzados en forma de sendas gráficas con la evolución esperada de PIB medio a nivel global y de temperatura global del planeta.

Los grupos de trabajo disponen de varias rondas para conseguir un resultado satisfactorio. Transcurrido un número de rondas pre-establecido, se muestran los resultados de la competición comparando los resultados de cada equipo en un ranking. El ranking tiene en cuenta cuestiones relativas al comportamiento más o menos adecuado de la economía y de la ecología en función de los objetivos propuestos por los equipos.

El juego permite registrarse como moderador para organizar partidas. El moderador puede gestionar varias partidas estableciendo el número de grupos y el número de personas por grupo. Es también el responsable de comenzar y finalizar las partidas pudiendo seguir las actividades de los participantes en un *dashboard*.

Figura 1: Interfaz y fases del juego. Se marca con un rectángulo rojo los puntos en los que aparece la realimentación.

2.1 Generación de lenguaje natural

En la presentación de resultados del grupo y en la presentación de los rankings finales, el sistema aporta realimentación a los usuarios. Esta realimentación tiene como objetivo indicar cómo de bueno o de malo es su rendimiento en el juego, y dar claves sobre cómo mejorar los resultados.

La generación de lenguaje natural reali-

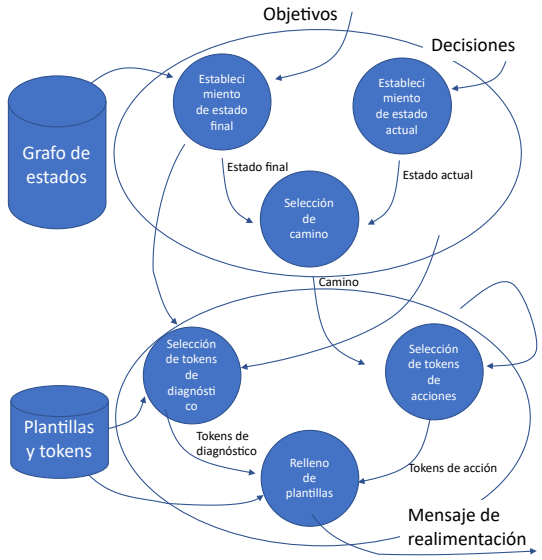


Figura 2: Diagrama funcional que muestra la operativa del módulo de generación de lenguaje natural.

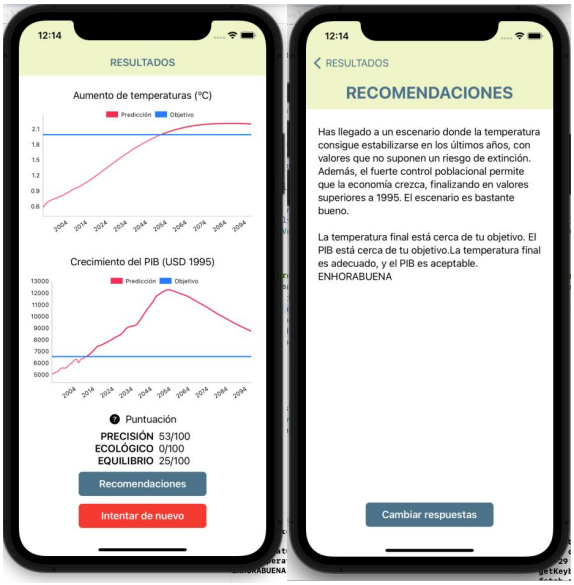


Figura 3: Apartado de realimentación en la interfaz iOS del juego.

zada en este módulo sigue el método descrito en (Escudero-Mancebo y Manzano-Santos, 2022). Los mensajes de realimentación constan de tres partes: una descripción del resultado obtenido, una estimación de cómo de lejos se encuentra el grupo de trabajo de obtener un resultado satisfactorio y, por último, una recomendación de los cambios que deben hacerse para mejorar el resultado.

La figura 2 describe el método de generación de lenguaje natural. Se apoya en un grafo de estados que representa el conocimiento sobre el juego. Los jugadores establecen una serie de objetivos y de medidas políticas a adoptar que determinan el estado actual y final del juego teniendo en cuenta el grafo de estados. Empleando dicho grafo se identifica el camino entre dichos estados.

En una segunda fase, se utiliza la información que caracteriza el estado inicial para generar los tokens que enriquecen las plantillas relativas la descripción del resultado obtenido. La longitud del camino en el grafo se emplea para informar de lo lejos o cerca que está el usuario de llegar a un estado satisfactorio y los pasos del camino se utilizan para generar tokens que permitan informar sobre las acciones que deben realizar los jugadores.

El grafo de estados se obtiene mediante la ejecución iterativa del simulador MEDEAS (Capellán-Pérez et al., 2020), que genera series temporales con prospecciones socio económicas y de temperatura del planeta.

Las posibles entradas de los formularios utilizados para introducir las decisiones políticas han sido previamente traducidas en variables empleadas por el simulador para generar un número de series temporales que componen las prospecciones. Se genera un *dataset* que es convertido en un grafo de estados mediante un algoritmo de clustering multivariante (Manzano-Santos, Escudero-Mancebo, y Miguel-González, en revisión).

2.2 Arquitectura software

El frontend del videojuego Crossroads ha sido desarrollado con Angular, y se comunica con el backend a través de servicios desarrollados con Spring boot. Como principales componentes están los que gestionan el registro de jugadores y su integración en grupos de trabajos; el componente vinculado al chat, que permite dialogar a los jugadores; y el componente de monitorización de las opciones elegidas por cada miembro del grupo. En el backend, se gestiona una base de datos relacional con la información de las partidas y una base de NoSQL con la información de los resultados del simulador (el simulador tarda varios segundos en hacer una simulación por lo que se almacenaron los resultados de las ejecuciones realizadas en lotes de trabajo).

El módulo de generación de lenguaje natural está aislado del resto de la arquitectura para facilitar la modularidad. Accede a una base de conocimiento que guarda el grafo de estados y las plantillas a aplicar. El módu-

lo se ha implementado en Python como un servicio REST externo al juego, desarrollado mediante el framework *Flask*, accesible mediante una petición `HTTP POST`.

2.3 Pruebas realizadas

La aplicación está desplegada online en `http://crossroads.geeds.eu/`. La versión para móvil (Android e iOS) se está desarrollando actualmente en el marco del proyecto FeCYT FCT-20-16138. Se han realizado pruebas de usabilidad en diferentes sesiones de trabajo con diferentes perfiles de usuario: investigadores en energía, economía y cambio climático; expertos en videojuegos; estudiantes de secundaria; profesorado de primaria y secundaria. En cada una de las sesiones de prueba realizadas hasta el momento se han recogido cuestionarios de evaluación que están dirigiendo la mejora continua que se está aplicando a la interfaz del juego.

Las sesiones de pruebas se han orientado a evaluar la robustez, la facilidad de uso, la capacidad de motivar, los capacidad de concienciar sobre el cambio climático y la eficiencia de la aplicación al compararla con la versión manual del juego (Capellán-Pérez, Álvarez-Antelo, y Miguel, 2019). En las pruebas de usabilidad no se ha recogido ningún comentario negativo sobre los mensajes de *feedback*, evidenciando la correcta integración del módulo en el juego.

3 Conclusiones

La aplicación Crossroads 2.0, disponible en línea, muestra cómo es posible generar mensajes automáticos utilizando el conocimiento del dominio disponible sobre el juego. Las pruebas realizadas muestran que, el uso de una base de conocimiento amplia, permite generar mensajes no sólo apropiados para el contexto, sino también lo suficientemente expresivos como para integrarlos de forma natural en el juego. El método propuesto para la generación de mensajes, extensible a otros juegos y aplicaciones que puedan representar el conocimiento en forma de grafo, es especialmente interesante porque utiliza un número pequeño de plantillas, derivando la complejidad de las respuestas a los tokens a integrar en las plantillas que son generados automáticamente mediante comparaciones entre los estados que integran el camino en el grafo.

Agradecimientos

Este proyecto ha sido desarrollado en el marco del proyecto LOCOMOTION , financiado por la Unión Europea en el programa Horizon 2020 número de contrato 821105. Las pruebas del juego se están realizando en el marco del proyecto FeCYT 2020 código 16138 titulado ENCRUCIJADA-MUNDO: ECOHERRAMIENTAS LÚDICAS PARA LA TRANSICIÓN ENERGÉTICA. Han participado en el desarrollo software de la aplicación, además de los autores, Lucas Calderón y María Galindo. Han colaborado en la pruebas ISF País Vasco y Cátedra UNESCO EHU, Instituto Juana I de Castilla, grupo de investigación GEEDs, grupo EcoProfes. Especial agradecimiento a Carmen Duce, José María Enriquez y Luis Javier de Miguel por la búsqueda de financiación.

Bibliografía

Capellán-Pérez, I., D. Álvarez-Antelo, y L. J. Miguel. 2019. Global sustainability crossroads: A participatory simulation game to educate in the energy and sustainability challenges of the 21st century. *Sustainability*, 11(13):3672.

Capellán-Pérez, I., I. de Blas, J. Nieto, C. de Castro, L. J. Miguel, O. Carpintero, M. Mediavilla, L. F. Lobejón, N. Ferreras-Alonso, P. Rodrigo, y others. 2020. Medeas: A new modeling framework integrating global biophysical and socioeconomic constraints. *Energy & environmental science*, 13(3):986–1017.

Escudero-Mancebo, D. y A. Manzano-Santos. 2022. Generación automática de realimentación en lenguaje natural en juegos serios a partir del grafo de estado. *Revista de la Sociedad Española de Procesamiento de Lenguaje Natural*, página en revisión.

Fernández Galeote, D. y J. Hamari. 2021. Game-based climate change engagement: Analyzing the potential of entertainment and serious games. *Proceedings of the ACM on Human-Computer Interaction*, 5(CHI PLAY):1–21.

Johnson, C. I., S. K. Bailey, y W. L. V. Buskirk. 2017. Designing effective feedback messages in serious games and simulations: A research review. *Instructional techniques to facilitate learning and motivation of serious games*, páginas 119–140.

Manzano-Santos, A., D. Escudero-Mancebo, y J. M. Miguel-González. en revisión. A multivariate time series clustering algorithm for the analysis of the cross relation between the constituent univariate time series patterns. *Pattern Recognition*.

Wu, J. S. y J. J. Lee. 2015. Climate change games as tools for education and engagement. *Nature Climate Change*, 5(5):413–418.

ICA2TEXT: Un sistema para la descripción automática en lenguaje natural de series temporales de calidad del aire

ICA2TEXT: A system for the automatic natural language description of air quality time series

Andrea Cascallar-Fuentes,[1] **Javier Gallego-Fernández,**[1] **Alejandro Ramos-Soto,**[1]
Anthony Saunders-Estévez,[2] **Alberto Bugarín-Diz,**[1]
[1]Grupo de Sistemas Intelixentes, Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela
[2]Rede de Calidade do Aire de Galicia, MeteoGalicia, Xunta de Galicia
{andrea.cascallar.fuentes, alberto.bugarin.diz}@usc.es, javier.gallego.fernandez@rai.usc.es, alejandro.ramos@inverbisanalytics.com, calidadedoaire.cma@xunta.gal

Resumen: En este proyecto describimos ICA2TEXT, un sistema data-to-text para generar automáticamente descripciones textuales sobre series temporales de calidad del aire proporcionadas por MeteoGalicia. Los resultados de la evaluación por parte de dos expertos meteorólogos fueron muy satisfactorios, lo que confirma que las descripciones textuales propuestas se ajustan a este tipo de datos y servicios tanto en contenido como en diseño. Actualmente, este sistema se encuentra en una fase final de pruebas y será desplegado como servicio público de la web de MeteoGalicia (MeteoGalicia, 2021).
Palabras clave: términos lingüísticos borrosos, sistemas data-to-text, generación de lenguaje natural

Abstract: In this project we describe ICA2TEXT, a data-to-text system to automatically generate textual descriptions about air quality time series provided by MeteoGalicia. Assessment results by two experts meteorologists were very satisfactory, which confirm that the proposed textual descriptions fit this type of data and service both in content and layout. This system is currently in a final testing phase and will be deployed as a public service on the MeteoGalicia website (MeteoGalicia, 2021).
Keywords: fuzzy linguistic terms, data-to-text systems, natural language generation

1 Introducción

Profundizar en la información realmente relevante que hay detrás de los datos plantea la necesidad de emplear técnicas que se adapten a las necesidades específicas de cada dominio y que puedan escalar a medida que se acumulan los datos.

La Generación de Lenguaje Natural (NLG) es un campo centrado en la generación de texto a partir de varias fuentes de datos. Dentro del NLG, los sistemas data-to-text (D2T) (Reiter, 2007) generan automáticamente textos a partir de grandes conjuntos de datos numéricos o simbólicos, proporcionando información comprensible. Normalmente, el diseño de los sistemas D2T incluye *i)* una etapa de análisis de datos donde se extrae la información relevante y *ii)* una etapa de generación donde se transmite la información en lenguaje natural. Relacionado con esto, desde el campo de la lógica borrosa se ha propuesto varios enfoques para generar descripciones lingüísticas de los datos (LDD) o resúmenes lingüísticos utilizando términos lingüísticos.

En este trabajo describimos ICA2TEXT, un sistema data-to-text basado en la lógica borrosa y la generación de lenguaje natural para describir automáticamente series temporales sobre el índice de calidad del aire (ICA), que es un indicador ampliamente utilizado en todo el mundo de la calidad del aire.

2 Contexto del problema

La presencia de contaminantes en el aire y, por tanto, el deterioro de la calidad del aire puede tener efectos nocivos para la salud de las personas. Hemos trabajado con datos describen el Índice de Calidad del Aire (ICA) en la red de 50 estaciones meteorológicas que envían datos actualizados cada hora en tiempo real en Galicia proporcionados por MeteoGalicia (MeteoGalicia, 2021). Para determinar la calidad del aire, este servicio mide cinco contaminantes diferentes: SO_2, NO_2, $PM25$, $PM10$ and O_3.

Basándose en los criterios de la Agencia Europea de Medio Ambiente (European Environment Agency, 2021), esta variable tiene seis etiquetas con una percepción positiva, neutra o negativa (Tabla 1).

Percepción	Positiva		Neutra	Negativa		
Etiqueta	Muy bueno	Bueno	Regular	Malo	Muy malo	Pésimo
Índice	0	1	2	3	4	5

Tabla 1: Etiquetas del índice de calidad del aire con su percepción e índice numérico.

Debido a la importancia de esta información, los meteorólogos de MeteoGalicia pretenden ofrecerla a los ciudadanos de forma comprensible, hasta ahora en formato gráfico. Por ello, surge la necesidad de dotar a esta información gráfica de una descripción textual que facilite su comprensión. En este contexto, hemos desarrollado el sistema ICA2TEXT en colaboración con los expertos de MeteoGalicia para describir lingüísticamente las series temporales de calidad del aire. El diseño de este sistema ha sido realizado de modo que atiende a las necesidades de este ámbito en cuanto a la flexibilidad de la riqueza lingüística requerida, abordando el manejo de la imprecisión en la descripción de series temporales. En los siguientes apartados se muestra en detalle el diseño del sistema siguiendo los requerimientos de los expertos.

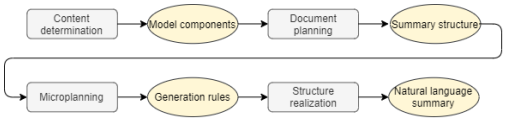


Figura 1: Representación de la arquitectura de nuestra propuesta. Los rectángulos representan las etapas, mientras que las elipses representan los resultados.

3 *Descripciones lingüísticas de las series temporales del ICA*

Este sistema se compone de las siguientes etapas (Figura 1), que componen la arquitectura data-to-text propuesta para describir las series temporales.

3.1 Determinación de contenido

Esta fase se compone de dos sub-etapas: *i)* Análisis de los datos, en el que se identifican los patrones y las tendencias, e *ii)* interpretación de los datos, en la que se identifican los mensajes que representan los patrones y la relación entre ellos.

Hemos diseñado un modelo temporal borroso para abordar el problema de manejar la imprecisión de la información temporal al resumir las series temporales. Este modelo temporal se ha diseñado para agrupar los datos, si es posible, en la referencia temporal más general. Nuestro objetivo es que el discurso sea legible y comprensible, aunque se pierda algo de precisión o exactitud en las descripciones.

3.2 Planificación del documento

Una vez identificados los mensajes y sus relaciones, en esta fase se generan todos los mensajes que se pueden incluir en la descripción final y se da una estructura a la descripción lingüística. La estructura de la descripción lingüística es la siguiente: *i* resumen general, *ii* intensificación (si procede) y *iii* excepción (si procede). Además, el resumen general incluye una descripción general y la descripción de la tendencia si procede, mientras que las secciones de intensificación y excepción contienen valores excepcionales ordenados de forma ascendente por valor o fecha. Realizamos las descripciones lingüísticas en los idiomas español y gallego utilizando SimpleNLG-ES (Ramos-Soto, Gallardo, y Bugarín, 2017) y SimpleNLG-GL (Cascallar-Fuentes, Ramos-Soto, y Bugarín, 2018).

3.3 Microplanificación

A partir de los mensajes generados previamente y de la estructura definida, en esta fase se seleccionan los casos a destacar y, por tanto, los mensajes que se van a mostrar. Las reglas de microplanificación se basan en las máximas griceanas (Grice, 1975).

En cuanto al resumen general se define que *i)* al describir un caso negativo se debe incluir el contaminante causante y *ii)* la tendencia sólo se incluye si las etiquetas de inicio y fin son diferentes.

En cuanto a la intensificación y a la excepción se define que *i)* el contaminante causante de un ICA negativo se omite si ha sido indicado en el resumen general, *ii)* se debe seleccionar la referencia temporal más general posible con un grado de verdad mayor o igual a 0.9 y *iii)* los periodos de tiempo con el mismo valor se agrupan en la descripción.

3.4 Realización de la estructura

Una vez que hemos definido la estructura y los mensajes que compondrán la descripción lingüística, se genera automáticamente asegurándonos de que sea correcta ortográfica, morfológica y sintácticamente. En este escenario, tanto en la intensificación como en la excepción, si el número de casos destacados es superior a 2, se dispon-

drán como una lista. Sin embargo, cuando el número de elementos sea igual o inferior a 2 se incluirán ambos como texto plano.

3.5 Definición de los componentes

En esta sección, presentamos el diseño de los componentes necesarios para generar la descripción lingüística de la serie de índices de calidad del aire.

3.5.1 Cálculo de las etiquetas

En primer lugar, calculamos la etiqueta del índice general de calidad del aire que mejor representa la serie temporal global para incluirla en la descripción general. Esta etiqueta se obtuvo como una media ponderada en la que el valor más reciente es el más relevante para describir la situación general a través de la referencia temporal "En las últimas horas". Además, en descripción de la tendencia, su valor también se calculaba con una media ponderada.

3.5.2 Referencias temporales

En el libro de estilo de MeteoGalicia, se define la franja horaria para las diferentes partes del día {mañana, tarde, noche} en verano e invierno.

Aunque los rangos que definen estos momentos del día se declaran de forma estática (al igual que la definición de un día completo desde las 00:00:00 hasta las 23:59:59), su uso al hablar está condicionado por la imprecisión del lenguaje. De modo que hemos definido de forma difusa las siguientes referencias temporales:

- Día completo: en lugar de una definición estricta desde las 00:00:00 hasta las 23:59:59, agrupamos como día también las dos horas anteriores y posteriores con un peso en el rango [0, 1].
- Mañana, tarde, noche: como se ha mencionado anteriormente, estas referencias temporales están definidas en el libro de estilo de MeteoGalicia. Utilizando esa definición como base, las hemos definido como un conjunto borroso trapezoidal en el que las dos horas anteriores y posteriores a los límites se consideran con un peso en el rango [0, 1].
- Primeras, centrales y últimas horas de la {mañana, tarde y noche}: hemos definido estas tres referencias temporales para describir situaciones más específicas. Estas etiquetas también se definen como conjuntos borrosos trapezoidales.

Código	Pregunta
Q1	La descripción lingüística representa correctamente los datos representados en la figura
Q2	La descripción concuerda con la forma en que describirías los datos
Q3	El vocabulario se usa correctamente
Q4	La organización de la descripción lingüística facilitar su comprensión
Q5	La ortografía, la puntuación y la estructura son correctas

Tabla 2: Preguntas del cuestionario de validación de expertos del índice de calidad del aire.

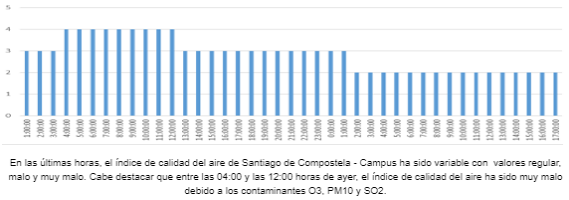


Figura 2: Ejemplo del cuestionario de evolución del ICA diseñado para la validación de expertos.

4 Validación por expertos

Hemos pedido a dos meteorólogos expertos de la Red de Calidad del Aire de MeteoGalicia (MeteoGalicia, 2021) que evaluaran la calidad de las descripciones lingüísticas generadas por ICA2TEXT en este dominio y su adecuación rellenando el cuestionario compuesto por 30 situaciones meteorológicas diferentes utilizando una escala de 5 puntos donde 1 significa "el experto está absolutamente en desacuerdo" y 5 "el experto está absolutamente de acuerdo". Ninguno de estos dos expertos había participado en la definición de ninguna parte del modelo.

Es cuestionario está formado por cinco preguntas, agrupadas en dos categorías: contenido de la descripción lingüística (Q1, Q2) y diseño (Q3, Q4, Q5). Cada caso del cuestionario está formado por una representación gráfica de la serie temporal y la descripción textual generada que describía el caso, pidiéndoles que evaluaran la idoneidad de las descripciones para describir las distintas situaciones. La figura 2 muestra un ejemplo extraído del cuestionario.

En la Tabla 3 presentamos un resumen de las puntuaciones de los expertos para cada una de las preguntas de forma individual y agrupada por dimensión. En general, los resultados muestran que los expertos están de acuerdo con las descripciones lingüísticas, ya que la media de las puntuaciones es de 4,67 y la moda muestra que el ma-

	Media	Desv. Típica	Moda	Mediana	IQR
Q1	4.58	0.87	5	5	1
Q2	4.15	1.01	5	4	1
Q3	4.75	0.70	5	5	0
Q4	4.92	0.28	5	5	0
Q5	4.97	0.18	5	5	0
Contenido	4.37	0.96	5	5	1
Estructura	4.88	0.46	5	5	0
General	4.67	0.75	5	5	0

Tabla 3: Resultado de la evaluación realizada por expertos.

yor valor utilizado es 5, es decir, la puntuación máxima. Por lo tanto, podemos concluir que estas descripciones lingüísticas generadas son muy adecuadas tanto en contenido como en forma para describir series temporales de índices de calidad del aire.

5 Discusión y conclusiones

En este trabajo hemos descrito el desarrollo de ICA2TEXT, un sistema que genera descripciones lingüísticas de datos de calidad del aire en castellano y gallego en colaboración con expertos de MeteoGalicia. Nuestro objetivo era cubrir las necesidades detectadas de acompañar la información gráfica que ofrecen en su web con descripciones textuales que faciliten su comprensión por parte de los usuarios.

Las series temporales para cada estación nunca supera los 150 registros. Nuestra aproximación consume una media de 10s para generar las dos descripciones textuales (una por idioma) para las 50 estaciones de MeteoGalicia. Este tamaño es lo usual por lo que nuestra aproximación puede ser utilizada con datos de cualquier agencia meteorológica realizando las adaptaciones pertinentes.

ICA2TEXT permite incluir un nuevo idioma, incluyendo los elementos necesarios en los archivos de configuración. Para los idiomas para los que ya existe una versión de SimpleNLG se podría adaptar fácilmente teniendo en cuenta las características de cada idioma. En caso de que no exista, habría que crear plantillas o un realizador lingüístico para este idioma.

Con respecto a su reutilización con otro tipo de datos, a la hora de describir series temporales se utiliza un tipo de relato muy habitual, donde se describe una valoración general de una situación incluyendo matices de intensificación y excepción. En el modelo que hemos definido hemos seguido esta estructura, de modo que, para reutilizar ICA2TEXT con otros tipos de datos, debería adaptarse la fase de preprocesado de los datos y las tareas realizadas dentro de la fase de determinación de contenido. Por otro lado, en caso de que los requisitos del lenguaje sean muy diferentes, habría que adaptar todas la fases del diseño en gran medida.

Los resultados de la validación realizada por expertos en la materia han sido muy satisfactorios. Como consecuencia, actualmente está siendo sometido a una fase final de pruebas y se desplegará como servicio público en la web oficial de MeteoGalicia.

Como trabajo actual y futuro, estamos aplicando nuestro modelo al diseño de nuevos sistemas D2T en otros ámbitos, como la notificación automática de series temporales en el ámbito de la sanidad electrónica.

Agradecimientos

Esta investigación ha sido financiada por el Ministerio de Ciencia, Innovación y Universidades (subvenciones TIN2017-84796-C2-1-R, PID2020-112623GB-I00, y PDC2021-121072-C21) y la Consellería de Educación, Universidade e Formación Profesional (subvenciones ED431C2018/29 y ED431G2019/04). Todas las subvenciones han sido cofinanciadas por el Fondo Europeo de Desarrollo Regional (programa FEDER).

Bibliografía

Cascallar-Fuentes, A., A. Ramos-Soto, y A. Bugarín. 2018. Adapting SimpleNLG to Galician language. En *Proceedings of the 11th International Conference on Natural Language Generation*, páginas 67–72. Association for Computational Linguistics.

European Environment Agency. 2021. European Air Quality Index website. [Accessed February 2021].

Grice, H. P. 1975. Logic and conversation. En *Speech acts*. Brill, páginas 41–58.

MeteoGalicia. 2021. MeteoGalicia website. [Accessed February 2021].

Ramos-Soto, A., J. J. Gallardo, y A. Bugarín. 2017. Adapting SimpleNLG to Spanish. En *Proceedings of the 10th International Conference on Natural Language Generation, INLG*, páginas 144–148. Association for Computational Linguistics.

Reiter, E. 2007. An architecture for data-to-text systems. En *Proceedings of the Eleventh European Workshop on Natural Language Generation*, páginas 97–104. Association for Computational Linguistics.

NLP4SM: Natural Language Processing for Social Media

NLP4SM: Procesamiento del Lenguaje Natural para Redes Sociales

Gonzalo Medina Medina,[1] **Jose Camacho Collados,**[2] **Eugenio Martínez Cámara**[1]
[1]Department of Computer Science and Artificial Intelligence
Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI)
University of Granada, Spain
[2]School of Computer Science and Informatics, Cardiff University, United Kingdom
gmedina95@correo.ugr.es, camachocolladosj@cardiff.ac.uk, emcamara@decsai.ugr.es

Abstract: NLP4SM is a website for the execution, analysis and comparison of tweet classification methods based on language models. Currently, NLP4SM supports the text classification tasks considered in TweetEval, but it aims at integrating additional text clasification tasks and to wider the number of language models available with the goal of becoming to a benchmark platform for assessing text classification methods with real data from social media.
Keywords: Language models, text classification, social media.

Resumen: NLP4SM es una plataforma web orientada a la ejecución, análisis y comparación de modelos de clasificación de texto en redes sociales basados en modelos del lenguaje. Actualmente integra las tareas de clasificación incluidas en TweetEval, pero aspira a incluir un mayor número de tareas y de ampliar la cantidad de modelos de lenguaje usados con el fin de convertirse en la plataforma de referencia de evaluación de modelos de clasificación de texto con datos reales de redes sociales.
Palabras clave: Modelos de lenguaje, clasificación de texto, redes sociales.

1 Introduction

The most likely source of the vertiginous progress of Natural Language Processing (NLP) in the recent years is the proposal of the Word2Vec model (Mikolov et al., 2013), which eases the generation of unsupervised linguistic features that are known as word embeddings and they represent the meaning of words in vectors of real numbers. The strong results reached by word embeddings based on Word2Vect enhanced the design of new word embeddings models, such as Glove.[1] These models set an embedding vector to each word regardless of its context, and for this reason the next landmark were starred by the contextual word embeddings models (Pilehvar and Camacho-Collados, 2020). The transformers models stand out as contextual word embeddings, with BERT (Devlin et al., 2019) as outstanding example. These models are known as language models, and their capacity of representing the meaning of words couple with the possibility of using them as pre-trained models have driven the progress of a broad branch of NLP tasks, especially those mostly linked to the classification of the semantic meaning of text, such as the opinion polarity of a review, the offensive meaning or the underlying emotional meaning of a message.

The potential of language models has made them the baseline of a wide range of NLP tasks, and they can even be used for developing learning models in production environments. On the other hand, the ease of tuning these models to specific NLP tasks has led the development and release of a huge amount of pre-trained language models in a large bunch of NLP task, with HuggingFace and especially its Transformers library (Wolf et al., 2020) standing out. This vast variety of language models makes their comparison and analysis really difficult as a previous step of the particular language model to fine-tune to a specific use case.

The certain use of language in social networks makes to adapt the NLP methods to the specific use of language of each social net-

[1] `https://nlp.stanford.edu/projects/glove/`

work, as for instance to Twitter (Martínez-Cámara et al., 2014). Language models also needs this fitting to the use of language of social networks, which makes them to be at the top of most NLP shared-tasks.

The great availability of language models has not been coupled with the release of web platforms for comparing and analysing the different language models in specific NLP tasks. Nevertheless, the issue of the great availability of training corpora and the evaluation of learning models begins to be resolved by the publication of leader boards of learning models trained on gold standards, such as SuperGLUE (Wang et al., 2019) or TweetEval (Barbieri et al., 2020).

Following the example of the NLP classification tasks leader boards, we present the web platform NLP4SM,[2,3] whose demonstrative prototype is described in this paper. NLP4SM is a web application for analysing the performance of Twitter language models fine-tuned to the tasks of (1) sentiment analysis, (2) emotion analysis, (3) offensive language classification, (4) hate speech classification, (5) irony detection and (6) stance classification on abortion, climate change, atheism, feminism and Hillary Clinton. NLP4SM allows on one hand the classification of a free span of text, and on the other hand the classification of the meaning of a bunch of tweets returned by Twitter. Furthermore, the classification results are shown as charts to ease their understanding. NLP4SM can be used by non-NLP experts and NLP scientists that need to compare different language models in one of the mentioned tasks on real data. The design of the system allows the consideration of new language models of the previous NLP tasks, as well as the incorporation of new result visualisation methods.

2 Language Models in NLP4SM

The first version of NLP4SM incorporates learning models that classify the meaning of tweets. The learning models are based on the fine-tuning of Twitter language models to the specific NLP tasks, which we subsequently describe.

[2]Prototype: `https://nlp4sm.on.fleek.co/`

[3]Production (Camacho-Collados et al., 2022): `https://tweetnlp.org/demo/`

2.1 NLP tasks

We select the NLP tasks according to their scientific relevancy, as well as the high social demand to have automatic systems that can identify specific kind of messages. The tasks are also part of TweetEval, and we present them as what follows.

Emotion analysis It identifies the underling emotion of a text. Although it is a multi-label task, we redefined it as a multi-class classification task. The corpus "Affect in Tweets" (Mohammad et al., 2018) was used to fit the model to the most frequent emotions of the corpus: joy, optimism, anger and sadness.[4]

Sentiment analysis It classifies the opinion polarity in positive, negative or neutral. The corpus of the subtask A of "Sentiment Analysis in Twitter" task of SemEval17 (Rosenthal, Farra, and Nakov, 2017) was used to fit the model.[5]

Hate speech It aims at classifying whether a tweet express hate. The corpus of HateEval from SemEval19 was used to fit the model (Basile et al., 2019).[6]

Irony detection The goal is to classify whether a tweet is ironic. The corpus of the Irony Detection task from SemEval18 was used to fit the model (Van Hee, Lefever, and Hoste, 2018).[7]

Offensive language It identifies whether a span of text has an offensive meaning. The corpus of OffensEval from SemEval19 was used to fit the model (Zampieri et al., 2019).[8]

Emoji prediction It aims at predicting the emoji that best represent the meaning of a tweet. The corpus of Emoji Prediction from SemEval18 was used to fit the model (Barbieri et al., 2018).[9]

Stance classification It classifies the author stance according to a topic. The corpus of the task Detectin Stance from SemEval16

[4]`https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion`

[5]`https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment`

[6]`https://huggingface.co/cardiffnlp/twitter-roberta-base-hate`

[7]`https://huggingface.co/cardiffnlp/twitter-roberta-base-irony`

[8]`https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive`

[9]`https://huggingface.co/cardiffnlp/twitter-roberta-base-emoji`

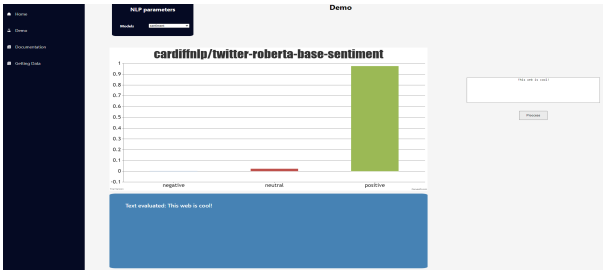


Figure 1: Sentiment analysis, 'text mode' mode.

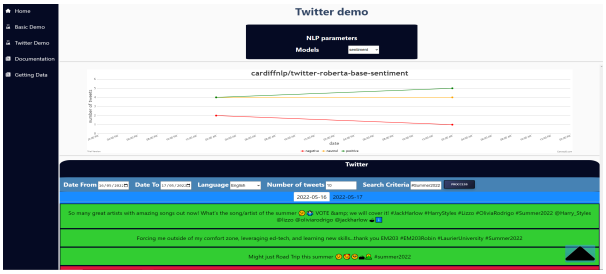


Figure 2: Sentiment analysis, 'Twitter mode'.

was used to fit the model. The topics considered are: abortion,[10] atheism,[11] feminism,[12] climate change[13] and Hillary Clinton.[14]

Multinguality Social networks are multilingual, and for this reason NLP4SM also allows to analyse multilingual language models, namely those ones based on XLM-R (Conneau et al., 2020) that is fitted on a large set of tweets written in more than 50 languages. NLP4SM also provides the XLM-T language model fitted to the sentiment analysis task in eight different languages (Barbieri, Espinosa-Anke, and Camacho-Collados, 2021).

2.2 Language Models

The language models currently included in NLP4SM match with the ones in TweetEval and they are available in HuggingFace. We have used the RoBERTa-base model (Liu et al., 2019) pre-trained on English text from social networks (Barbieri et al., 2020).

The fine-tuning of RoBERTa-base to each NLP task is based on a output layer with the same output units than the number of classes of each task (Liu et al., 2019). The languages models used are described and linked in section 2.1.

3 Description of NLP4SM

We aim at providing an unified and accessible platform for assessing and analysing social network text classification models. Hence, we have developed a web application for the first version of NLP4SM.

NLP4SM is built upon a client-server architecture led by a REST API. Moreover, we have relied on external services for running the language models. NLP4SM uses Huggingface because it is currently the on-cloud service that hosts the language models included in NLP4SM, it is the artificial intelligence service platform most used by the NLP research community and it provides a high quality service.

The server side is developed in Python and it is based on the micro-framework Flask. The server side is responsible of the communication with HuggingFace through using its API. Moreover, the server side queries Twitter according to the user query.

The client side is a web interface based on JavaScript React. It allows two different forms of evaluating the models, namely:

Text mode It evaluates any language model described in section 2 with a span of text written down by the user in a text box. Several charts show the result of the evaluation. Figure 1 depicts and example of the text mode.

Twitter mode It process a set of tweets returned in real-time from Twitter according to the user query. The user can configure his query according to the language, the time and the specific text of the query. NLP4SM retrieves the tweets and shows with different kind of charts the result of running the selected language model. Figure 2 depicts and example of the text mode.

4 Conclusions and future work

In this paper, we presented the prototype demonstration NLP4SM, which aims at easing the access, analysis and comparison of classification models based on language models of different NLP tasks with real data from social networks. NLP4SM allows the evaluation of any span of text, and the evaluation

[10] https://huggingface.co/cardiffnlp/twitter-roberta-base-stance-abortion

[11] https://huggingface.co/cardiffnlp/twitter-roberta-base-stance-atheism

[12] https://huggingface.co/cardiffnlp/twitter-roberta-base-stance-feminist

[13] https://huggingface.co/cardiffnlp/twitter-roberta-base-stance-climate

[14] https://huggingface.co/cardiffnlp/twitter-roberta-base-stance-hillary

of tweets from a user query.

We plan as future work: (1) to integrate more NLP tasks, (2) to extend the number of language models considered, and (3) to add a greater number of visualisation methods of results.

Agradecimientos

This research work is supported by the R&D&I grant PID2020-116118GA-I00 funded by MCIN/ AEI/10.13039/501100011033.

References

Barbieri, F., J. Camacho-Collados, L. Espinosa Anke, and L. Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the ACL: EMNLP 2020*.

Barbieri, F., J. Camacho-Collados, F. Ronzano, ..., and H. Saggion. 2018. SemEval 2018 task 2: Multilingual emoji prediction. In *Proc. of The 12th Int. Workshop on Semantic Evaluation*, pages 24–33.

Barbieri, F., L. Espinosa-Anke, and J. Camacho-Collados. 2021. A Multilingual Language Model Toolkit for Twitter. In *arXiv preprint arXiv:2104.12250*.

Basile, V., C. Bosco, E. Fersini, ..., and M. Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proc. of the 13th Int. Workshop on Semantic Evaluation*, pages 54–63.

Camacho-Collados, J., K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa-Anke, F. Liu, E. Martínez-Cámara, et al. 2022. Tweetnlp: Cutting-edge natural language processing for social media. *arXiv preprint arXiv:2206.14774*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, ..., and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. of the 58th of the ACL*, pages 8440–8451.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conf. of the NAACL, Vol. 1 (Long and Short Papers)*, pages 4171–4186.

Liu, Y., M. Ott, N. Goyal, J. Du, ..., and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Martínez-Cámara, E., M. T. Martín-Valdivia, L. A. Ureña-López, and A. Montejo-Ráez. 2014. Sentiment analysis in twitter. *Natural Language Engineering*, 20(1):1–28.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of Workshop at ICLR*.

Mohammad, S., F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proc. of The 12th Int. Workshop on Semantic Evaluation*, pages 1–17.

Pilehvar, M. T. and J. Camacho-Collados. 2020. Embeddings in natural language processing: theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4):1–175.

Rosenthal, S., N. Farra, and P. Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proc. of the 11th Int. Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Van Hee, C., E. Lefever, and V. Hoste. 2018. SemEval-2018 Task 3: Irony detection in English tweets. In *Proc. of The 12th Int. Workshop on Semantic Evaluation*.

Wang, A., Y. Pruksachatkun, N. Nangia, ..., and S. R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proc. of NeurIPS 2019*.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, ..., and A. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP 2020: system demonstrations*, pages 38–45.

Zampieri, M., S. Malmasi, P. Nakov, ..., and R. Kumar. 2019. SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proc. of the 13th Int. Workshop on Semantic Evaluation*, pages 75–86.

Transcripción de periódicos históricos: aproximación CLARA-HD

Transcription in historical newspapers: the CLARA-HD approach

Antonio Menta, Eva Sánchez-Salido y Ana García-Serrano
ETSI Informática de la UNED
amenta@invi.uned.es, {evasan, agarcia}@lsi.uned.es

Resumen: Analizar periódicos de los siglos XVIII, XIX y principios del XX exige cierta calidad de las fuentes digitalizadas y la utilización de recursos específicos de dominio o de la lengua. Cualquier aproximación utilizando las tecnologías actuales, se encuentra con que la mayoría de los modelos PLN disponibles para la transcripción o el reconocimiento de entidades están entrenados con textos en "lenguajes actuales". Si además el reto consiste en extraer información de periódicos históricos en español, la complejidad aumenta, ya que la normalización del español es relativamente "moderna" y hay que intentar refinar los modelos de PLN o generar nuevos recursos. En esta presentación del corpus construido desde los textos disponibles en la Hemeroteca Digital de la BNE, Diario de Madrid (1788-1825)[1], se mostrarán los pasos seguidos para su transcripción automática generando un modelo (99% de rendimiento) en el marco del proyecto CLARA-HD. Finalmente se incluyen unas conclusiones iniciales.

Palabras clave: Transcripción de textos, modelos del lenguaje, recursos lingüísticos.

Abstract: The analysis of historical newspapers from the 18th, 19th, and early 20th centuries requires a certain quality of digitized sources and the use of specific domain or language resources. Any approach using current technologies finds that most of the NLP models available for transcription or entity recognition are trained with texts in "current languages". If, in addition, the challenge consists of extracting information from historical newspapers in Spanish, the complexity increases since the normalization of Spanish is relatively "modern" and it is necessary to try to refine the NLP models or generate new resources. In this demonstration for the corpus built from the BNE Digital Hemeroteca, Diario de Madrid (1788-1825)[1] the steps followed will be shown for its automatic transcription using a defined model (99% performance), within the framework of the CLARA-HD project. Finally, some initial conclusions are included.

Keywords: Transcription of texts, language models, linguistic resources.

1 Introducción

La utilización de técnicas de Procesamiento de Lenguaje Natural (PLN) en el tratamiento de documentos textuales, en concreto en el ámbito de las Humanidades Digitales (HD), se ha convertido en una práctica referente en muchos de los proyectos actuales (Toscano et al, 2020). En los últimos veinte años se han realizado multitud de procesos de digitalización para la conservación de colecciones culturales tanto a nivel local como nacional y europeo. Estos proyectos han generado millones de imágenes que necesitan ser tratadas para la transcripción del texto que contienen, ya sea de forma manual o mediante la aplicación de procesos de reconocimiento óptico de caracteres, conocido como OCR (del inglés *Optical Character Recognition*).

La elaboración de corpus históricos está sujeta a múltiples factores, entre ellos su finalidad (Torruella, 2017). Por ejemplo, para el estudio de una lengua actual en general se pretende que el corpus sea proporcional, es decir,

[1] http://hemerotecadigital.bne.es/details.vm?q=id:0001510462&lang=es

que la cantidad de palabras o de textos de cada muestra esté en proporción respecto a su distribución en el total de la población. Sin embargo, este requisito es difícil de conseguir en corpus históricos, ya que a menudo no se conservan suficientes documentos representativos de cada tipo, o incluso se desconocen las proporciones en que deberían aparecer. Por otra parte, la creación del corpus también depende del tipo de consulta que se desee realizar sobre los resultados que proporcione su análisis. En función de las posibilidades de consulta, los corpus son etiquetados mediante marcas declarativas que describen los elementos formales del texto (cursiva, tamaño de la fuente), estructurales (capítulos, páginas) y lingüísticos (entidades, cambios de registro).

La comunidad científica se ha dado cuenta de la dificultad de tratar documentos históricos, y en los últimos años se está realizando un esfuerzo en mejorar el acceso y las herramientas disponibles para su consulta (Ehrmann et al., 2020). Aquí es donde entran en juego las técnicas de PLN. Estas son capaces de extraer, procesar y relacionar la información que contienen los documentos para su posterior utilización y que sirvan de ayuda a los humanistas en sus reflexiones y análisis.

En esta presentación del corpus construido desde los textos disponibles en la Hemeroteca Digital de la BNE, Diario de Madrid (1788-1825)[1], se justifica, en el apartado segundo, la necesidad de construir corpus de suficiente calidad para el análisis PLN previo al estudio de historiadores o público en general, se muestran los pasos seguidos para su transcripción en el apartado tercero y finalmente se incluyen algunos comentarios sobre este trabajo.

2 Necesidad de corpus de textos históricos de calidad

Las crecientes facilidades que ofrece la informática propician la confección de corpus que presentan el mismo texto en diversas modalidades de edición: facsímil (reproducción fotográfica del original), paleográfica (transcripción sin correcciones ni interpretaciones), normalizada (transcripción siguiendo la normativa ortográfica, léxica y sintáctica vigente), crítica (transcripción que pretende reconstruir el texto original) o interpretativa (transcripción que sigue los postulados de la edición paleográfica pero permite corregir ciertos errores para poder explicar el sentido del texto). Ejemplos son el corpus burckhardtsource.org y el proyecto CHARTA[2].

En el estudio del impacto de la tarea de reconocimiento de entidades nombradas (NER, por sus siglas en inglés) en el ámbito de las HD, en (van Hooland et al., 2015) se reflexiona sobre las posibilidades de utilizar NER y otros métodos de extracción de información en textos no estructurados y proponen ampliar el debate sobre la forma de utilizar las tecnologías del PLN a la comunidad humanística.

Dentro de las HD, el estudio de las ediciones de periódicos históricos entre el siglo XVIII y principios del siglo XX es un campo idóneo para aplicar estas técnicas debido a la presencia de todo tipo de entidades en ellos y a su evolución temporal a lo largo de los años para recuperar, almacenar y consultar la herencia cultural transmitida. Aun así, su uso directo presenta varios inconvenientes al utilizarlos en textos históricos. La mayoría de los modelos actuales son modelos estadísticos que necesitan un conjunto de datos etiquetados para ser entrenados en el contexto que se quieren utilizar, y estos conjuntos escasean o no son públicos en las HD. Esto repercute en otra dificultad añadida, que es la representación que deben tener los textos para ser utilizados por las técnicas del PLN.

Desde hace años se ha impuesto la utilización de modelos vectoriales de baja dimensión para representar los textos, conocidos como *word embeddings*. Para obtener estos modelos, en la mayoría de las ocasiones es necesario realizar un entrenamiento en una gran cantidad de textos del contexto en el que se quieren utilizar para aprender las relaciones entre las palabras y conceptos. Para obtener una mejor representación final se suele realizar un pre-procesamiento de los textos para eliminar información irrelevante (como código HTML y algunos metadatos). Una vez limpio el texto, se utiliza como entrada para generar los *word embeddings*, ya sean estáticos o contextuales como los modelos basados en *Transformers* (Vaswani et al., 2017).

Últimamente, las redes neuronales basadas en modelos de lenguaje han permitido la continua mejora en la detección de entidades,

[2] https://www.corpuscharta.es

especialmente desde la publicación del modelo BERT (Devlin et al., 2018) en 2018, o los modelos de lenguajes basados en *Transformers*. En (Jiang et al., 2021) se realiza un estudio del impacto de la salida del OCR en el rendimiento de los modelos basados en BERT en un problema de clasificación de extractos de libros que van desde finales del siglo XVIII a finales del siglo XX. En sus conclusiones mencionan una degradación de los resultados y recomiendan realizar un ajuste fino de los modelos en esta tipología de documentos con anterioridad a realizar la clasificación para hacerlos más robustos a los errores ortográficos. Además, el vocabulario utilizado en siglos pasados dista enormemente del usado hoy en día y es un reto y una motivación para hacer hincapié en la utilización de los modelos de lenguaje basados en redes neuronales.

En definitiva, los intentos de análisis de documentos históricos mediante tecnologías de PLN actuales se encuentran con el problema de que la mayoría de los modelos disponibles están entrenados con textos en "lenguas modernas", y aumenta la complejidad al intentar extraer información de documentos históricos en español, ya que la normalización del español es relativamente "moderna" y hay que refinar los modelos de PLN o generar nuevos recursos.

3 Construcción del modelo de transcripción

Como ya se ha indicado, para aplicar tecnología actual de PLN, la dificultad añadida en las HD es el origen de los datos, porque la mayoría de las fuentes están almacenadas en imágenes de mala calidad con tipografías antiguas que necesitan de un OCR específico.

Transkribus[3] es una plataforma para la digitalización, el reconocimiento de texto, la transcripción y la búsqueda en documentos históricos. Es resultado de un proyecto europeo y, una vez finalizado, para su mantenimiento y explotación han decidido hacerla de pago a partir de un cierto límite de uso. Con el registro se obtienen 500 créditos (unas 500 páginas). La herramienta está muy bien documentada[4] y cuenta con funcionalidades de acceso libre desde el navegador[5] o la aplicación.

Para la transcripción dispone de modelos basados en redes neuronales públicos y entrenados en distintos idiomas y grafías[6], lo que facilita encontrar uno que se aproxime al de los documentos a transcribir. De no ser así, la herramienta permite entrenar uno propio y automatizar la transcripción de nuestros documentos. De hecho, ya disponemos de un modelo entrenado a partir de transcripciones manuales en el proyecto CLARA-HD.

Para ello, se comienza creando una colección y cargando los ficheros que contienen los textos en ella (Figura 1).

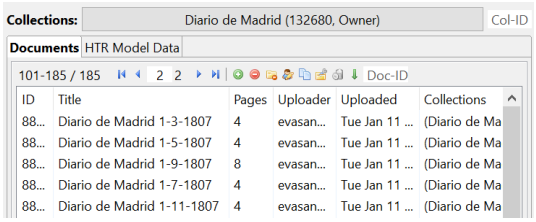


Figura 1. Carga de ficheros.

Para poder transcribir los documentos hay que realizar manualmente el reconocimiento de su estructura (o *layout*), diferenciando las regiones en las que se encuentra el texto (Figura 2). El reconocimiento en general no es perfecto, por lo que en ocasiones habrá que corregir errores o modificar el resultado manualmente.

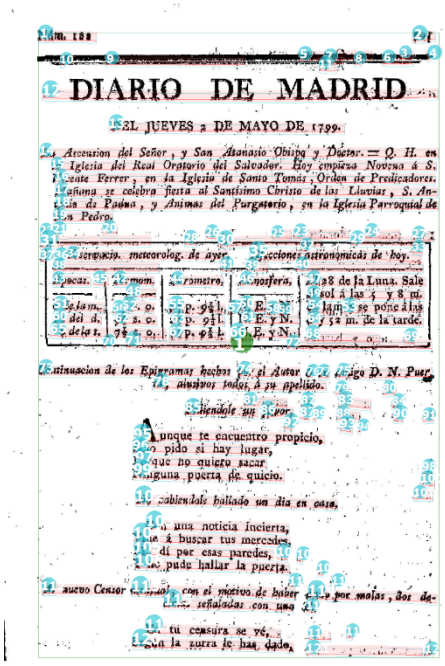


Figura 2. Reconocimiento de la estructura.

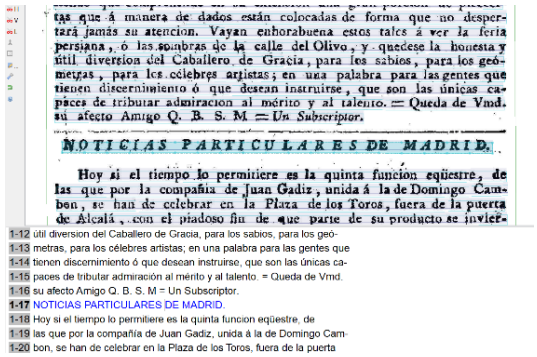
[3] https://readcoop.eu/transkribus/

[4] https://readcoop.eu/transkribus/howto/use-transkribus-in-10-steps/

[5] https://transkribus.eu/lite/

[6] https://readcoop.eu/transkribus/public-models/

Una vez hecho el reconocimiento de las regiones de texto, se transcribe el texto, línea a línea manualmente o con la ayuda de un modelo público seleccionado. De nuevo, la transcripción automática no es perfecta y hay que editarla para corregir los errores (Figura 3).


tas que á manera de dados están colocadas de forma que no despertará jamás su atencion. Vayan enhorabuena estos tales á ver la feria persiana, ó las sombras de la calle del Olivo, y quedese la honesta y útil diversion del Caballero de Gracia, para los sabios, para los geómetras, para los célebres artistas; en una palabra para las gentes que tienen discernimiento ó que desean instruirse, que son las únicas capaces de tributar admiracion al mérito y al talento. = Queda de Vmd. su afecto Amigo Q. B. S. M = *Un Subscriptor.*

NOTICIAS PARTICULARES DE MADRID.

Hoy si el tiempo lo permitiere es la quinta funcion eqüestre, de las que por la compañia de Juan Gadiz, unida á la de Domingo Camben, se han de celebrar en la Plaza de los Toros, fuera de la puerta de Alcalá, con el piadoso fin de que parte de su producto se invier-

1-12 útil diversion del Caballero de Gracia, para los sabios, para los geó-
1-13 metras, para los célebres artistas; en una palabra para las gentes que
1-14 tienen discernimiento ó que desean instruirse, que son las únicas ca-
1-15 paces de tributar admiración al mérito y al talento. = Queda de Vmd.
1-16 su afecto Amigo Q. B. S. M = Un Subscriptor.
1-17 NOTICIAS PARTICULARES DE MADRID.
1-18 Hoy si el tiempo lo permitiere es la quinta funcion eqüestre, de
1-19 las que por la compañía de Juan Gadiz, unida á la de Domingo Cam-
1-20 bon, se han de celebrar en la Plaza de los Toros, fuera de la puerta

Figura 3. Transcripción manual.

Para automatizar este proceso se ha creado un modelo propio de transcripción a partir de un conjunto de entrenamiento junto con una guía de estilo, realizando los pasos mostrados anteriormente: (1) subida de documentos a la herramienta, (2) reconocimiento manual de la estructura de todas las páginas de los documentos, (3) transcripción de un cierto número de páginas manualmente o con la ayuda de un modelo público y (4) revisión manual final de las mismas, para entrenar nuestro modelo de transcripción.

4 Comentarios finales

Se ha presentado cómo construir un corpus con la herramienta Transkribus, entrenando un nuevo modelo de transcripción capaz de reconocer caracteres no vistos por el modelo base, alcanzando una precisión en el reconocimiento de caracteres nuevos del 99%.

En este momento estamos trabajando con historiadores de la UNED interesados en el contenido del Diario de Madrid, para identificar tanto la terminología como los temas de interés para su investigación y evaluar cuánto es soportada por la tecnología PLN utilizada. Una vez identificados los tipos de entidades útiles para los historiadores, se seguirá con la extracción de las menciones de cada tipo, como las localizaciones, las profesiones o palabras complejas de entender.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto coordinado CLARA-NLP (www.clara-nlp.uned.es) con los subproyectos en historia (PID2020-116001RB-C32), biomedicina (Campillos-Llanos, 2022) (PID2020-116001RA-C33) y economía (Moreno-Sandoval, 2020) (PID2020-116001RB-C31). Los autores agradecen especialmente la colaboración de los estudiantes V. Sánchez-Sánchez, R. Garcia-Sánchez y A. Rodriguez-Francés.

Bibliografía

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Underst.", *arXiv preprint 1810.04805,* 2018.

Campillos-Llanos, L., A. Terroba, S. Zakhir, A. Valverde and A. Capllonch, "Building a comparable corpus and a benchmark for Spanish medical text simpli." *Procesamiento del Lenguaje Natural* 69, 2022.

Ehrmann, M., M. Romanello, A. Flückiger, and S. Clematide, "Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers", *CLEF proc. 2020.*

Jiang, M., Y. Hu, G. Worthey, R. C. Dubnicek, and T. Underwood, "Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts", in *CHR 2021: Computational Humanities Research Conference*, pp. 266–279.

Moreno-Sandoval, A., A. Gisbert and H. Montoro: "Fint-esp: a corpus of financial reports in Spanish" en *Multiperspectives in Analysis and Corpus Design*, Granada, Editorial Comares, 2020, pp. 89-102.

Torruella Casañas, J., "Lingüística de corpus: Génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística", Peter Lang Ed., 2017.

Toscano, M., Rabadán, A., Ros, S., and González-Blanco, E. (2020). Digital humanities in Spain: Historical perspective and current scenario. *Profesional De La Información*, *29*(6).

van Hooland, S., M. de Wilde, R. Verborgh, T. Steiner, and R. Van de Walle, "Exploring entity recognition and disambiguation for cultural heritage collections", *Digital Scholarship Humanities*, V30, N2, 2015.

Vaswani, A., *N.* Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin "Attention is all you need", *Advances in neural inf. Proc. Sys.* 30, 2017.

iASSIST: Low-cost, Portable and Embedded Assistants for On-Premise Automated Transcription and Translation Services

iASSIST: Asistentes embebidos, portables y de bajo coste para servicios on-premise de transcripción y traducción

Aitor Álvarez[1], Víctor Ruiz[1], Iván G. Torre[1], Thierry Etchegoyhen[1], Harritxu Gete[1], Joaquín Arellano[1]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián, Spain;
{aalvarez,vruiz,igonzalez,tetchegoyhen,hgete,jarellano}@vicomtech.org

Resumen: Presentamos iASSIST, un sistema embebido, portable y de bajo coste con servicios de transcripción y traducción para inglés, castellano y euskera. El sistema es operativo, embebido en placas Jetson, y accesible mediante interfaz de usuario para tareas de transcripción y traducción en tiempo real con modelos neuronales de alta calidad.
Palabras clave: edge computing, IA embebida, transcripción, traducción neuronal

Abstract: We present iASSIST, a low-cost, portable and embedded solution for on-premise automated neural transcription and translation services, currently for the English, Spanish and Basque languages. The system is fully operational, embedded in Jetson boards, and accessible via a user-friendly interface to perform real-time transcription and translation with high-quality neural models.
Keywords: edge computing, embedded AI, neural transcription, neural translation

1 Introduction

Recent advances in deep neural networks (DNNs) have led to significant improvements in both Automatic Speech Recognition (ASR) and Neural Machine Translation (NMT) (Chiu et al., 2018; Vaswani et al., 2017). However, these advances are mainly achieved with large neural architectures, trained on massive volumes of data and typically deployed on high-end expensive servers in the cloud to provide efficient services, which raises a number of critical issues.

First, privacy is an important concern, since sending personal or confidential data over the Internet makes the information vulnerable to attacks and breaches. The General Data Protection Regulation (GDPR) and similar policies set to protect sensitive data also need to be taken into account.

Secondly, high-quality AI models typically require servers with significant computational capacity and GPU acceleration cards for both training and inference. Acquiring this type of hardware resources for local computing, or renting appropriate infrastructure in the cloud, can represent a significant budget that many companies cannot cover.

Thirdly, deep AI models are significantly impacting energy consumption worldwide, with serious consequences on the increasing climate crisis. Reducing the ecological footprint of current AI technology is a critical part of the current research agenda.

Finally, latency issues and information loss can impact cloud computing services, making it difficult at times to deploy responsive and robust AI solutions.

Edge computing aims to move computational power and data processing closer the originating data (Sarker et al., 2019), with AI algorithms running on local networks or embedded devices to guarantee data privacy and reduce latency, energy consumption and network load. However, integrating high-performance AI models into embedded systems with low computational capabilities requires system and model optimization.

Within this context, we present iASSIST, a low-cost, portable and embedded solution for on-premise automated neural transcription and translation services for the English, Spanish and Basque languages. This solution has been developed within the applied research project iASSIST, partially supported

by the Department of Economic Development of the Basque Government. The project started in September 2019 and finalised in December 2021, and was carried out by the following consortium: SPC[1] (project coordinator), MondragonLingua[2], Serikat[3], Natural Vox[4], Haresi[5] and Vicomtech[6].

2 iASSIST

The core architecture of iASSIST is shown in Figure 1. It consists of the following main components:

- A front-end, composed of a web-based graphical user interface (GUI).
- A REST API, which exposes the functionalities of the back-end.
- A back-end, which orchestrates all the functionalities of the solution, including automatic transcription and translation, client request management, model loading and unloading, and operational modes (batch and streaming).

Among the different options for embedded systems offered by the market (e.g. Raspberry Pi, NVIDIA Jetson, Google Coral or Intel Movidius, among others), we selected the NVIDIA Jetson embedded computing boards for the project. Specifically, we focused on two specific devices with different capabilities: Jetson TX2 and Jetson AGX Xavier. Although these two boards were relatively similar prices at the time, the AGX Xavier (32 TOPS, 512-core GPU, 8-core CPU, 32 GB of shared memory) offered significantly more computational power than the TX2 system (1.3 TOPS, 256-core GPU, dual-core CPU, 8 GB of shared memory), while also being more energy efficient. During the project, we explored the capacities of both boards and evaluated the integration of different AI models depending on their architecture, size, number of parameters and performance in each embedded system.

In the following subsections, each of the main components of the iASSIST solution is presented in more detail.

[1]https://www.spc.es/
[2]https://www.mondragonlingua.com/en
[3]https://www.serikat.es/
[4]https://www.naturalvox.eu/en/home/
[5]https://haresi.es/
[6]https://www.vicomtech.org/en

2.1 Front-end

The iASSIST GUI aims to facilitate the communication between the user and the back-end. It was designed from a usability and user experience perspective, prioritizing simplicity. The GUI provides users with different input options, from text to audio file (batch mode) and audio source (streaming mode), and allows them to select different transcription and translation models to perform the corresponding tasks. Additionally, it integrates two main text-boxes to present the transcription and translation results and a graphical interface to manage model loading and unloading in memory. It is worth noting that the transcription results can be downloaded in different formats (txt, rtf, xml, srt, vtt) that can be used for different applications such as subtitling, keyword spotting and rich transcription. The GUI was developed using the Angular framework[7] and deployed via a Nginx web server[8].

2.2 REST API

The REST API serves as the main interface between the GUI and the back-end. In addition, it provides an alternative way for the user to directly access all the features of the solution via http requests, allowing third party systems to be built on top of iASSIST and thus extend its functionality.

2.3 Back-end

The iASSIST back-end is composed of several modules which encompass the features of the solution. The main modules are described in turn in the next subsections.

2.3.1 Orchestrator

This module encompasses the automated configuration, management, and coordination of the main components and services of the back-end. At its core, it manages user requests, communication between modules and I/O interaction. The module also implements the logic and interfaces for the batch and streaming applications, manages automatic language identification for translation with bilingual ASR models, and controls the input sources, devices and audio streams. The iASSIST solution is able to process audio files, texts or streaming audio coming from any microphone connected to the board or machine where the GUI is launched.

[7]https://angular.io/
[8]https://www.nginx.com/

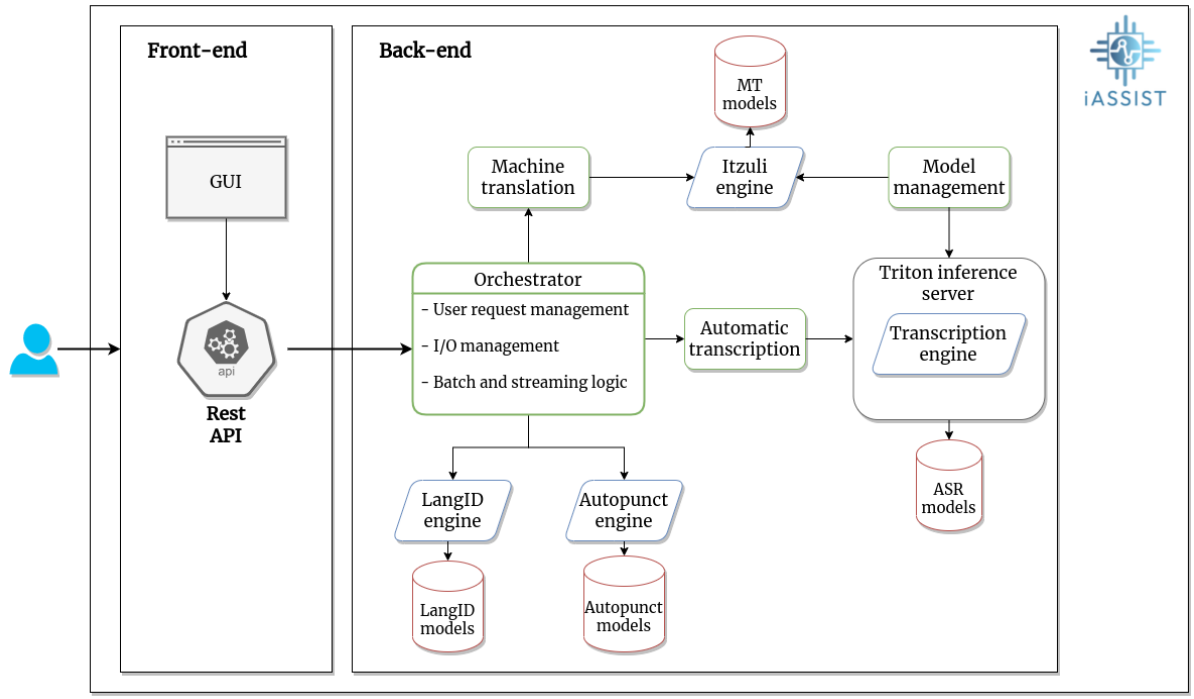


Figure 1: Core architecture of the iASSIST solution.

2.3.2 Model management

Running applications composed of several AI models on embedded systems requires dynamically controlling model activation and memory usage, given common limitations of the supporting boards. The model management module ensures proper model loading and unloading in memory, allowing users to enable or disable the relevant functionality depending on the AI task at hand.

2.3.3 Automatic transcription

The Automatic Transcription module is managed by the Triton Inference Server[9], which is in charge of handling workloads and integrating the three main modules of the transcription pipeline. The first module processes the raw audio input by extracting features as spectrogram chunks, which are sent to an acoustic model for probabilistic classification in a second stage. The final module, composed by the decoder, determines the most likely transcription for that audio using the likelihoods produced by the previous classification with the help of a language model.

For iASSIST, we developed acoustic models based on the NVIDIA's Quartznet E2E architecture (Kriman et al., 2020), designed by the need to reduce the size and complexity of the recognition models, making them lighter, faster and more easily deployed on embedded systems. This architecture is composed of multiple blocks with residual connections in between. Each unique block consists of one or more modules with 1D time-channel separable convolutional layers, batch normalisation, and ReLU layers.

For each of the selected Jetson embedded systems, we experimented with different versions of the Quartznet architecture. After evaluating their performance in terms of latency and quality, we decided to deploy the Quartznet Q15×5 based model on the Jetson AGX Xavier, and the Q10×5 based model on the Jetson TX2 board. The main difference lies in the number of times the Quartznet models repeat the five unique blocks, which modifies the total number of parameters from 18.9M (Q15×5) to 12.8M (Q10×5). To further optimize the performance of the Quartznet acoustic models, quantization and layer fusion techniques were also applied via the TensorRT library (Vanholder, 2016).

Finally, the raw transcriptions are enriched with capitalisation and punctuation marks generated by the BERT-based AutoPunct engine (González-Docasal et al., 2021).

[9]https://developer.nvidia.com/nvidia-triton-inference-server

In addition to enhancing readability, splitting the raw text into correctly punctuated sentences increases the quality of machine translation results.

2.4 Machine Translation

The Machine Translation module is based on Vicomtech's Itzuli Translator engine, a robust and scalable text translation system, which can be deployed under Kubernetes orchestration or as a standalone platform in a dedicated server, and integrates MarianNMT (Junczys-Dowmunt et al., 2018) in its own back-end to perform efficient NMT inference.

To optimise Transformer (Vaswani et al., 2017) NMT models, in terms of size and inference latency, we explored different strategies based on network pruning, quantization and knowledge distillation. Our final optimised models, suitable for the more constrained TX2, were student models trained on the knowledge distilled by large teacher models, with 6 Self-Attention layers for encoding and 2 SSRU layers (Kim et al., 2019) for decoding. The student models halved the memory footprint of teacher models, increased inference speed between 200% and 400% depending on beam size, with minor losses in terms of translation quality ranging between 0.2 and 1.4 BLEU points.

Translation models can be loaded and unloaded in memory on the fly, thus giving users the ability to switch to new translation tasks as needed within the constrained environment. Translation can be performed directly on user-provided source text or on the output of the ASR component to perform real-time speech translation.

3 Conclusions

We described the iASSIST solution, an embedded assistant for on-premise neural transcription and translation services. The application was validated by each of the companies of the consortium within three evaluation campaigns, where they accessed the embedded system externally and tested the solution at operational, usability and quality levels over their own contents and devices.

iASSIST demonstrates the ability to embed neural transcription and translation technology in Jetson boards with hardly any loss in performance, performing both batch and streaming tasks within a secure, portable and low-cost edge device. As future work, we will explore other embedded systems in which iASSIST could be integrated and will continue to improve AI model optimization for less powerful environments, particularly CPU-based client-side computation.

References

Chiu, C.-C., T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *Proc. of ICASSP*, pages 4774–4778.

González-Docasal, A., A. García-Pablos, H. Arzelus, and A. Álvarez. 2021. AutoPunct: A BERT-based automatic punctuation and capitalisation system for Spanish and Basque. *Proces. de Leng. Nat.*, 67:59–68.

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch. 2018. Marian: Fast neural machine translation in C++. In *Proc. of ACL 2018*, pages 116–121.

Kim, Y. J., M. Junczys-Dowmunt, H. Hassan, A. F. Aji, K. Heafield, R. Grundkiewicz, and N. Bogoychev. 2019. From research to production and back: Ludicrously fast neural machine translation. In *Proc. of WNGT*, pages 280–288.

Kriman, S., S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *Proc. of ICASSP 2020*, pages 6124–6128.

Sarker, V. K., J. P. Queralta, T. N. Gia, H. Tenhunen, and T. Westerlund. 2019. A survey on LoRa for IoT: Integrating edge computing. In *Proc of FMEC 2019*, pages 295–300.

Vanholder, H. 2016. Efficient inference with TensorRT. In *GPU Technology Conference*, volume 1, page 2.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

GUAITA: Monitorización y análisis de redes sociales para la ayuda a la toma de decisiones

GUAITA: Monitoring and analysis of social media to help decision making

Ferran Pla, Lluís-F. Hurtado, José-Ángel González, Vicent Ahuir, Encarna Segarra, Emilio Sanchis, María-José Castro, Fernando García
VRAIN: Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València

Resumen: El proyecto GUAITA tiene como objetivo extraer grandes cantidades de información proveniente de las redes sociales y proporcionar herramientas de análisis de dicha información que puedan ser útil para la toma de decisiones de las organizaciones. En ese sentido, instituciones y empresas de pueden beneficiar de los avances logrados. Fruto de este proyecto en la actualidad se dispone de un prototipo software que integra diferentes modelos basados en redes neuronales para la monitorización y análisis multilingüe de Twitter. Aunque ya existen en el mercado herramientas de recogida de datos y de análisis, un factor diferencial de GUAITA es el uso modelos de estado del arte en el análisis del lenguaje natural en redes sociales. Esto permite ampliar la funcionalidad básica de análisis de sentimiento, detectando, por ejemplo, el lenguaje inapropiado, el discurso del odio o el nivel de toxicidad presente en los mensajes. Además, GUAITA tiene en consideración idiomas habitualmente no considerados por este tipo de herramientas como es el caso del catalán.
Palabras clave: Redes Sociales, Procesamiento del Lenguaje Natural, Redes Neuronales

Abstract: The GUAITA project aims to extract large amounts of information from social media and provide tools for the analysis of this information that may be useful for decision-making in organizations. In this sense, institutions and companies can benefit from the progress achieved. As a result of this project, there is currently a software prototype that integrates different models based on neural networks for multilingual monitoring and analysis of Twitter. Although there are already data collection and analysis tools on the market, a differentiating factor of GUAITA is the use of state-of-the-art models in the analysis of natural language in social networks. This allows the analysis to be extended not only to sentiment analysis but also to go further and be able to detect, among others, inappropriate language, hate speech or the level of toxicity present in the messages. In addition, GUAITA takes into account languages that are not usually considered by this type of tool, such as Catalan.
Keywords: Social Media, Natural Language Processing, Neural Networks

1. Introducción

La tecnología actual ha permitido que la información disponible para la toma de decisiones sea cada vez más abundante y oportuna; esto, unido a nuevos desarrollos en ciencia de datos, ha ayudado a mejorar la velocidad de reacción y la calidad de dichas decisiones por parte de las empresas y organizaciones. Uno de los grandes apoyos en esta nueva toma de decisiones es el desarrollo de plataformas, servicios y modelos de analítica avanzada y visualización de datos. Se hace necesario desarrollar herramientas que presenten una interfaz amigable para el usuario y que no impliquen un alto coste de implantación, tanto económico como temporal.

Aunque ya existen en el mercado herramientas de recogida de datos y de análisis, estas herramientas están todavía en un nivel de desarrollo muy inicial cuando la fuente de datos son las redes sociales. El procesamiento automático del lenguaje natural utilizado

en este tipo de redes constituye un problema abierto dentro de la comunidad científica, por lo que la transferencia al mercado de herramientas de las tecnologías del habla logradas dentro del área de investigación del procesamiento del lenguaje natural es de gran interés.

En la actualidad existe una gran cantidad de compañías que ofrecen herramientas destinadas a dar soporte a los Community Manager (CM) en su labor de gestionar la presencia corporativa en redes sociales. Típicamente permiten la agregación en una única plataforma o aplicación de múltiples cuentas de usuario en varias redes sociales.

Además del soporte a la gestión del CM, muchas de las herramientas permiten el análisis de la actividad de usuarios en redes sociales, fundamentalmente en Twitter. Entre algunas de estas herramientas que permiten monitorizar la actividad de usuarios podemos destacar las siguientes: Twitter Analytics, Followerwonk, Twitonomy, Rebold, Brandwatch Consumer Research, Buffer, BuzzSumo, Klear, Union Metrics, Mentionmapp, Foller.me, PressClipping.

La información proporcionada por estas herramientas se basa principalmente en el análisis de los metadatos proporcionados por las redes. Esta información consiste en datos agregados y la distribución temporal de estos: número de seguidores, repercusión de un post a lo largo del tiempo (número de me gusta, número de retweets, número de replicas). En algunos casos se incluye la información geográfica de los tweets, basada en la geolocalización de los usuarios o en su perfil. En general, no se realiza un análisis profundo del contenido textual.

En este trabajo se presenta una demostración de los principales logros obtenidos en el proyecto GUAITA: Monitorización y análisis de redes sociales para la ayuda a la toma de decisiones subvencionado por la Agencia Valenciana de la Innovació (AVI) de la Generalitat Valenciana. Se describe el sistema desarrollado incluyendo su arquitectura y principales funcionalidades.

El sistema integra diferentes modelos basados en redes neuronales para la monitorización y análisis multilingüe de la red social Twitter. Un factor diferencial de GUAITA es el uso modelos de estado del arte en el análisis del lenguaje natural en redes sociales. Esto permite ampliar el análisis no solo al análisis de sentimientos sino ir más allá y ser capaz de detectar, entre otros, el lenguaje inapropiado, el discurso del odio o el nivel de toxicidad presente en los mensajes. Asimismo, GUAITA incluye el catalán entre los idiomas soportados; idioma que no suele estar considerado por otras herramientas de esta índole.

2. Descripcion del sistema

El sistema GUAITA está concebido como una herramienta software que permita el seguimiento de acontecimientos, personas o cualquier tema de interés para el usuario en la red social Twitter.

Las principales funcionalidades del sistema son las siguientes:

- *Seguimiento de redes sociales.* Permite realizar la monitorización de la red social Twitter para la obtención y almacenamiento de la información relacionada con el tema de interés. Para ello, nos permite definir tareas y programarlas en el tiempo. En cada tarea se pueden definir capturas (búquedas de Twitter) siguiendo los criterios que se consideren oportunos.

- *Obtención de modelos específicos de análisis de textos para diferentes lenguas.* El sistema también permite la recolección de textos que sean útiles para aprender modelos específicos en una lengua en concreto o dominio. La herramienta dispone de modelos para el español, inglés y catalán que permiten el procesado y etiquetado de corpus en estas lenguas.

- *Visualización de resultados.* El sistema presenta gráficamente los resultados de los análisis desarrollados mediante una serie de interfaces web. También dispone de un generador de informes, que de forma automática, elabora un dossier de toda la información relacionada con una tarea definida por el usuario. Dichos informes pueden ser de gran utilidad para su análisis con el fin de determinar la reputación de una institución o compañía.

- *API REST.* Permite la comunicación de nuestra aplicación con aplicaciones de terceros.

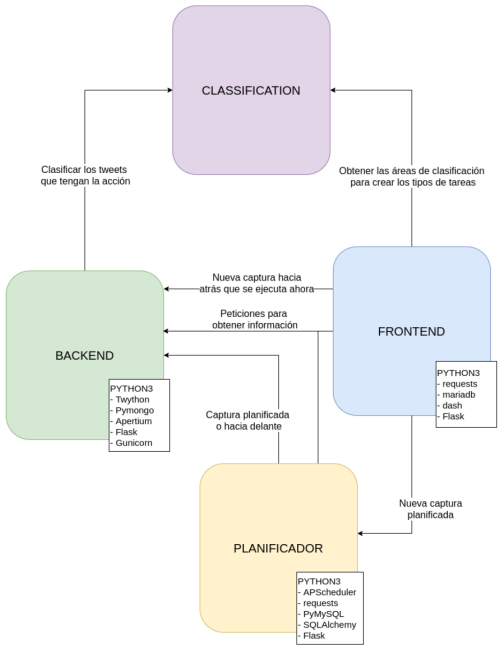


Figura 1: Arquitectura del Sistema GUAITA.

3. Arquitectura del sistema

En la Figura 1 se muestra la arquitectura general de la aplicación y las interconexiones entre los distintos módulos que la componen. Como se puede observar, el sistema GUAITA está compuesto por cuatro módulos principales que se describirán de forma sucinta en esta sección.

3.1. Backend

El Backend es el núcleo del sistema GUAITA. Este módulo es el responsable de gestionar el comportamiento de toda la aplicación. Fundamentalmente realiza la descarga de contenido de redes sociales y las peticiones al motor de clasificación y perfilado de usuario, genera las estadísticas y datos necesarios para la presentación gráfica de la información; información que será accesible mediante la API REST del módulo.

El Backend consta de una capa de persistencia dividida en dos. Por un lado se utiliza una instancia distribuida de una base de datos orientada a documentos MongoDB donde se guarda toda la información referente al análisis de cada captura. Por otro lado, se utiliza MariaDB para la persistencia relacionada con los procesos de captura. Esta capa es la responsable del almacenamiento en memoria secundaria de toda la información necesaria para el funcionamiento global de la aplicacióndesde el contenido descargado de Twitter o la información generada por el motor de clasificación.

3.2. Planificador

Este módulo se encarga de enviar capturas periódicas o futuras al Backend. En el contexto de la monitorización de redes sociales es habitual querer planificar tareas con antelación, por ejemplo para seguir el impacto de una nueva campaña publicitaria. También es habitual querer hacer consultas recurrentes en el tiempo, por ejemplo, para seguir un programa de televisión que se emite siempre a la misma hora un día determinado de la semana.

3.3. Motor de Clasificación

GUAITA permite obtener información del contenido textual de los tweets descargados mediante el uso de modelos de clasificación basados en redes neuronales. Todos estos modelos están aprendidos utilizando arquitecturas neuronales del estado del arte y corpus de múltiples competiciones internacionales. En la Sección 4 se describen los modelos y corpus utilizados.

3.4. Frontend

El sistema GUAITA esta dotado de una interfaz de programación de aplicaciones (API REST) que le permite ser utilizado e integrado en software de terceros. Sin embargo, también existe una versión web que facilita el uso de la herramienta a usuarios humanos. El Frontend es el encargado de gestionar los formularios y demás páginas de la aplicación web y realizar las peticiones al Backend utilizando la API. En la Figura 2 se muestra parte de la salida gráfica proporcionada por la aplicación para la consulta: #SagitarioA OR #SagittariusA OR #BlackHole.

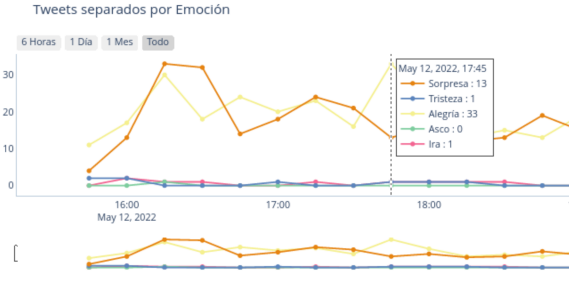


Figura 2: Salida del sistema, detección de emociones.

4. Modelos de clasificación

Para el motor de clasificación se han utilizado tres *encoders* : TWILBERT(González, Hurtado, y Pla, 2021), BETO(Cañete et al., 2020) y XLM-T(Barbieri, Anke, y Camacho-Collados, 2021). Se han utilizado corpus de múltiples competiciones para aprender diversos modelos de clasificación. GUAITA utiliza cada uno de los tres encoders dependiendo de la tarea y en función del rendimiento. Los corpus utilizados han sido los que se enumeran a continuación. TASS 2019(Manuel Carlos Díaz-Galiano, 2019) para los modelos de polaridad, Irosva 2019(Bueno et al., 2019) para los modelos de detección de ironía, EmoEvalEs 2021(Plaza-del Arco et al., 2021) para la detección de emociones y lenguaje ofensivo, HateEval 2019(Basile et al., 2019) para los modelos de detección de lenguaje del odio y agresividad, HaHa 2019(Chiruzzo, Castro, y Rosá, 2020) para la detección de humor presente en los tweets y Detoxis 2021(Taulé Delor et al., 2021-09) para varios modelos: lenguaje impropio, sarcasmo, toxicidad, etc.

5. Conclusiones y trabajos futuros

En este trabajo se ha presentado el sistema desarrollado en el proyecto GUAITA: Monitorización y análisis de redes sociales para la ayuda a la toma de decisiones. Se ha descrito su arquitectura y principales funcionalidades actuales. No obstante GUAITA es una herramienta en constante crecimiento y mejora. Entre las ampliaciones que se pretenden incorporar podemos destacar la detección de aspectos y las alertas automáticas ante mensajes que fomenten el odio durante el seguimiento de algún evento.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por la Agencia Valenciana de la Innovació (AVI) de la Generalitat Valenciana, proyecto GUAITA (INNVA1/2020/61) y el Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-11-21).

Bibliografía

Barbieri, Francesco, Luis Espinosa Anke, y Jose Camacho-Collados. 2021. Xlm-t: A multilingual language model toolkit for twitter.

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, y Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. En *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 54–63.

Bueno, Reynier Ortega, Francisco Manuel Rangel Pardo, D. I. H. Farías, Paolo Rosso, Manuel Montes y Gómez, y José Eladio Medina-Pagola. 2019. Overview of the task on irony detection in spanish variants. En *IberLEF@SEPLN*.

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, y Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. En *PML4DC at ICLR 2020*.

Chiruzzo, Luis, Santiago Castro, y Aiala Rosá. 2020. HAHA 2019 dataset: A corpus for humor analysis in Spanish. En *Proceedings of the 12th Language Resources and Evaluation Conference*, páginas 5106–5112.

González, José Ángel, Lluís-F. Hurtado, y Ferran Pla. 2021. Twilbert: Pre-trained deep bidirectional transformers for spanish twitter. *Neurocomputing*, 426:58–69.

Manuel Carlos Díaz-Galiano, et al. 2019. Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus. En *Proceedings of the Iberian Languages Evaluation, IberLEF@SEPLN 2019, Bilbao, Spain*, volumen 2421 de *CEUR Workshop Proceedings*, páginas 550–560.

Plaza-del Arco, Flor Miriam, Salud M. Jiménez Zafra, Arturo Montejo Ráez, M. Dolores Molina González, Luis Alfonso Ureña López, y María Teresa Martín Valdivia. 2021. Overview of the emoevales task on emotion detection for spanish at iberlef 2021.

Taulé Delor, Mariona, Alejandro Ariza, Montserrat Nofre, Enrique Amigó Cabrera, y Paolo Rosso. 2021-09. Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish.

Plugin para la automatización del análisis fonético-fonológico y la obtención de retroalimentación analítica para estudiantes de español

Plugin for automating phonetic-phonological analysis and obtaining analytical feedback for Spanish learners.

Tamara Couto Fernández[1], Albina Sarymsakova[2], Nelly Condori Fernández[3], Patricia Martín Rodilla[4]

[1 3 4] Facultad de Informática, Universidade da Coruña, Spain
[2] Facultad de Filología, Universidade da Coruña, Spain

albina.sarymsakova@udc.es, tamara.cfernandez@udc.es, n.condori.fernandez@udc.es, patricia.martin.rodilla@udc.es

Resumen: Presentamos en este artículo el *Plugin* para el análisis fonético-fonológico en español (PAFe) que consiste en una serie de *scripts* (un código escrito con un lenguaje de programación (Python) que, a su vez, implementan tres algoritmos diferentes de comparación de la entonación (de un alumno ELE (español como lengua extranjera) y un hablante nativo de español), permitiendo a su vez tres tipos de análisis distintos: global, tendencia tonal e intersilábico. Además, PAFe cuenta con una base de datos para mantener un histórico de diferentes tipos de datos (perfil de usuario, ejercicios de pronunciación y audios) y una interfaz gráfica para incluir reportes sobre la evolución de pronunciación en Praat, una herramienta para el análisis acústico. PAFe es una solución de *software* que ofrece nuevas funcionalidades de Praat y permite lo siguiente: (i) realizar un análisis comparativo entre los patrones entonativos de un estudiante de ELE y un hablante nativo; (ii) reportar la evolución de la adquisición de dichos patrones en español gracias al histórico de los datos almacenados. De esta manera, se brinda una retroalimentación automatizada tanto al alumnado como a los/las docentes.
Palabras clave: Praat, análisis de la entonación, TIC, Python.

Abstract: We present in this article the Plugin for phonetic-phonological analysis in Spanish (PAFe), which consists of a series of scripts (a code written with a programming language (Python) that, implement three different intonation comparison algorithms (of an ELE (Spanish as a foreign language) student and a native speaker of Spanish), allowing in turn three different types of analysis: global, tonal tendency and intersyllabic. In addition, PAFe has a database to keep a history of different types of data (user profile, pronunciation exercises and audios) and a graphical interface to include reports on pronunciation evolution in Praat, a tool for acoustic analysis. PAFe is a software solution that offers new functionalities of Praat and allows the following: (i) to perform a comparative analysis between the intonational patterns of an ELE student and a native speaker; (ii) to report the evolution of the acquisition of such patterns in Spanish thanks to the history of the stored data. In this way, automated feedback is provided to both students and teachers.
Keywords: Praat, intonation analysis, ICT, Python.

1 Introducción

El presente trabajo se enmarca en el área del procesamiento de lenguaje natural, en concreto, en el análisis de la entonación comparativo-contrastivo con los fines didácticos que proporciona nuestra herramienta original PAFe. A pesar de la existencia de algunas herramientas, como, por ejemplo, la de Oplustil y Toledo (2019) o el estudio de Helmer Strik y Khiet

Truong (2009), que ofrecen resultados de similitud fonético-fonológica o detectan errores cometidos en la pronunciación, no existe ninguna que aporte ambas facilidades al mismo tiempo, ni tampoco que ofrezca un seguimiento de la evolución del alumnado.

Por este motivo, hemos decidido desarrollar un sistema que complementa la enseñanza de idiomas, en particular, aquella que se imparte en remoto o para modalidades híbridas.

Nuestra herramienta ofrece la funcionalidad de realizar análisis comparativos instantáneos de la pronunciación de un estudiante, tomando como referencia el habla de un nativo, y observar la evolución de este a través de datos almacenados en un histórico.

Para la elaboración de nuestro *plugin* se han empleado diversas tecnologías para dar soporte al trabajo realizado tales como Praat, Python y PostgreSQL.

2 *Metodología*

Comenzamos a diseñar nuestro trabajo basándonos en los siguientes principios esenciales del análisis de la entonación:

(1) anotamos las sílabas de cada acto de habla en un *textgrid* de Praat (Boersma y Weenink, 2019); identificamos valores del tono de todas las vocales de las sílabas (se miden las consonantes sonoras o sonantes también), empleando el *Script* de Praat elaborado por Mateo (2010a, 2010b), que extrae los valores absolutos en Hz, los relativiza y dibuja el gráfico de la melodía estandarizada;

(2) discriminamos los valores frecuenciales relevantes entre los segmentos tonales de los valores irrelevantes; según Cantero (2002, 2019), Font-Rotchés y Cantero Serena (2008, 2009), menos de 10% de diferencia entre los segmentos se considera imperceptible.

Una vez hemos obtenido los datos relevantes del análisis de la entonación, pasamos a la arquitectura de PAFe.

En nuestro proyecto se desarrolla una extensión a una aplicación de escritorio para análisis acústicos de habla ya existente: Praat. Por lo tanto, se parte de una arquitectura desarrollada a la cual se le acopla un nuevo módulo (PAFe) (Figura 1) constituido por *scripts* de Praat, código de Python y una base de datos en PostgreSQL.

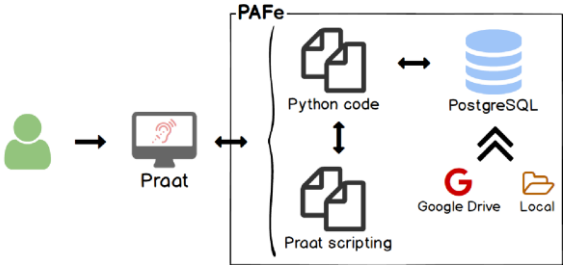


Figura 1: Arquitectura global de PAFe

Praat, mediante su *scripting*, permite hacer llamadas por línea de comandos a otros sistemas, tal y como describe Dragos-PaulPop (2013), siendo posible, de esta forma, la extensión de la aplicación mediante el uso de otros lenguajes y tecnologías externas a Praat. Este nuevo módulo (PAFe) se comunica con el sistema original por medio de nuevos *scripts* de Praat que se asocian a *ítems* del menú de la aplicación (Véase la Figura 2), desde los cuales se ejecutan dichos ficheros. En ocasiones, el nuevo módulo prescinde de llamadas a Praat y genera directamente ventanas con información a partir de ficheros con código en Python. El intermediario entre Praat y los datos que se gestionan en la base de datos es Python.

Empleamos técnicas de procesamiento del lenguaje natural y procesado de audio en nuestra herramienta tomando como fuente principal las grabaciones de voz humana de hablantes nativos y estudiantes. Praat nos permiten extraer información cuantitativa a nivel prosódico de los audios.

Posteriormente, los algoritmos comparativos nativo-estudiante en cuanto a aspectos prosódicos que se presentan y que implementa la herramienta son capaces de ofrecer la información comparativa entre dos audios nativo/estudiante para ofrecer la retroalimentación en el aprendizaje del idioma español. Estos algoritmos son una contribución original implementada en la herramienta dado que no existía una propuesta algorítmica de este tipo para español hasta ahora.

Hemos desarrollado el *Plugin* PAFe siguiendo una metodología iterativa e incremental basada en tecnologías ágiles y metodología de desarrollo *scrum*, apoyándonos en el trabajo de Sutherland (2020).

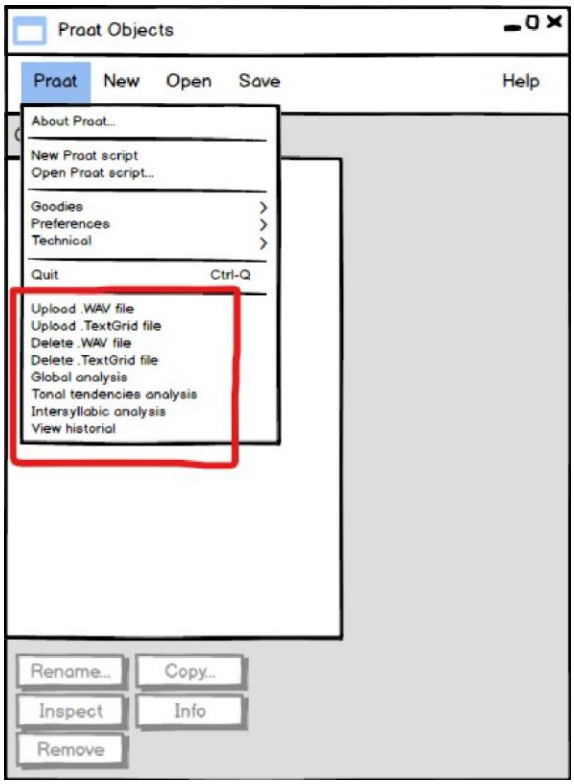


Figura 2: Ejemplo de Interfaz de Usuario que visualiza las nuevas funcionalidades añadidas en Praat

A continuación, describimos el desarrollo de nuestra herramienta.

3 *Solución: Plugin PAFe*

Nuestra herramienta PAFe, en su versión final, permite realizar los análisis comparativos proporcionando resultados de similitud y gráficas de la entonación en base a los valores de *pitch*[1] y a la tendencia tonal en cada segmento definido y, por último, visualización del progreso de un estudiante a lo largo del tiempo. Resaltamos las siguientes operaciones que posibilita ejecutar nuestro *plugin*:

(1) La aplicación permite crear distintos perfiles con el fin de facilitar el proceso de gestión de los datos que suben los usuarios.
   a) En primer lugar, se registra el profesor/la profesora.
   b) A continuación, se asigna un alumno/una alumna al profesor/la profesora registrados previamente. Este paso permite evitar la confusión si hay más de un usuario del mismo ordenador o portátil.
   c) Finalmente, se registra el perfil de un hablante/una hablante nativo/nativa de español para subir los datos que servirán como referenciales para el programa;

(2) PAFe posibilita la gestión de ficheros en formato .WAV y TextGrid [2] : nuestro programa comprende tanto almacenamiento como eliminación de audios y anotaciones;

(3) Asimismo, permite la realización de distintos tipos de análisis acústicos (análisis global, análisis de tendencias tonales y análisis intersilábico): el algoritmo que lleva a cabo el análisis global consiste en dividir los audios previamente guardados de los alumnos y hablantes nativos de español en unos 1000 intervalos (descartando los silencios) para obtener valores comparativos muy precisos. Sin embargo, este tipo de análisis no proporciona una retroalimentación acerca de las posibles desviaciones en el tono, sino que aporta datos genéricos del porcentaje de similitud del audio del hablante nativo y el alumno. En lo que se refiere al análisis de tendencias tonales, el programa trabaja con las anotaciones del formato .TextGrid y los audios en formato .WAV guardados previamente. En este caso, los enunciados se dividen por palabras y, para obtener la similitud de forma local, se indica si se ha reproducido el tono de cada palabra correctamente o no y, en caso de no ser reproducido correctamente, se indica el porcentaje de desviación; asimismo, se obtiene el porcentaje de la similitud del tono y la diferencia media entre dos audios. Por último, el análisis acústico intersilábico es un análisis comparativo, sílaba a sílaba, acerca de la semejanza entre la realización del tono de un/una estudiante y la de un nativo/una nativa; en este caso, para cada sílaba se indica la diferencia de pronunciación respecto al audio de referencia, así como se obtiene el porcentaje de la similitud del tono y la diferencia media entre dos audios. Según los resultados obtenidos a través de este último tipo de análisis, tanto la similitud como la diferencia entre el audio de referencia y el de los estudiantes se muestran con mayor precisión. Finalmente, podemos ver la evolución de los resultados de nuestros

[1] Frecuencia del tono en Hz

[2] Fichero con etiquetas que segmentan el audio asociado

alumnos mediante la opción de ver el historial.

Finalmente, mostramos un diagrama de flujo (Figura 3) que aporta información sobre su comportamiento de nuestro *plugin*, exponiendo las funcionalidades y su interrelación, además de presentar los operadores que interactúan con la aplicación y sus restricciones.

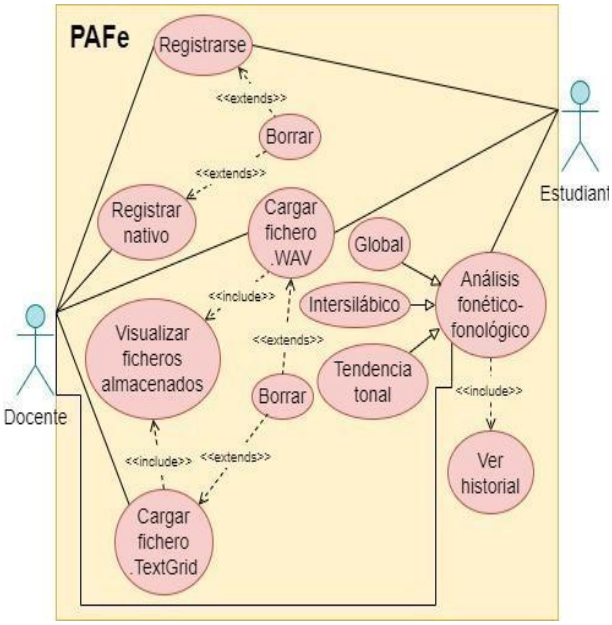
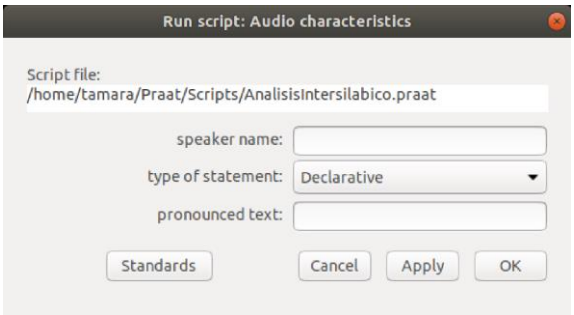


Figura 3: Diagrama de casos de uso (funcionalidades y actores principales de PAFe)

4 *Ejemplo ilustrativo de análisis intersilábico*

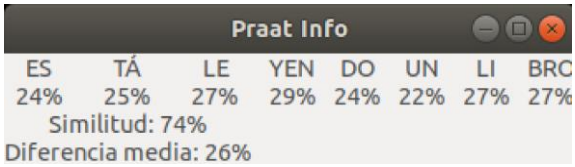
En este apartado mostramos cómo se lleva a cabo un tipo de análisis comparativo. Para ejecutar el análisis intersilábico, es necesario cubrir un formulario (Figura 4) con los datos que caracterizan al audio del estudiante que queremos comparar.



Figuras 4: Formulario para realizar un análisis intersilábico

A continuación, se filtran los audios de ese estudiante que cumplen dichas propiedades y se muestra una ventana con un desplegable para la selección del audio a analizar. Una vez seleccionado el audio, se selecciona el fichero .TextGrid correspondiente de la misma manera.

Cada tipo de análisis devuelve resultados diferentes. Para el análisis intersilábico, se muestra un resultado de similitud por sílabas y el porcentaje de diferencia media (Figura 5). Finalmente, obtenemos un gráfico con las curvas de diferenciación tonal en cada sílaba para cada audio (Figura 6).



Figuras 5: Información del análisis intersilábico

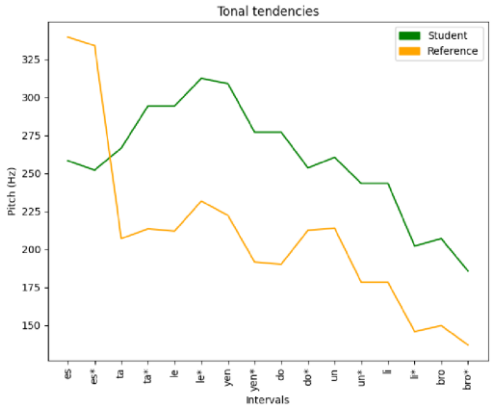


Figura 6: Gráfico con las curvas tonales de cada audio para cada sílaba (el eje X representa la división silábica de un enunciado y el eje Y los valores del pitch)

5 *Conclusiones*

A modo de conclusión, resaltamos los siguientes aspectos clave que hemos abordado en el presente trabajo:

(1) La herramienta PAFe permite realizar distintos tipos de análisis comparativo-contrastivos de la entonación (global, tendencias tonales e intersilábico) de los alumnos de ELE y los hablantes nativos de español; dentro de ellos, consideramos el intersilábico como el más preciso dado que los resultados de diferencia tonal aparecen sílaba a sílaba y muestran las desviaciones tonales de los estudiantes, y el global como el más eficiente respecto a tiempo de

respuesta ya que no requiere la subida de TextGrids, y la segmentación se realiza de manera automatizada, según muestran los datos empíricos del trabajo Couto Fernández (2021).

(2) Dicha aplicación cuenta con varias funciones; aparte de realizar el análisis entonativo, permite almacenar los audios, los archivos .TextGrid y los resultados del análisis (el historial) de cada enunciado según el perfil del hablante (alumno o hablante nativo de español).

(3) PAFe ha sido elaborado para conseguir los siguientes objetivos didácticos: facilitar la labor de los docentes en lo que se refiere a la identificación y corrección de las desviaciones de la entonación (hemos llevado a cabo un análisis empírico con docentes de español como lengua extranjera, donde medimos el grado de satisfacción con PAFe, con resultados positivos, tal y como indica el trabajo Couto Fernández, 2021); almacenar los resultados de los análisis ejecutados para su futuro perfeccionamiento; servir como herramienta de autoevaluación y autocorrección para los alumnos de ELE dado que la propia herramienta les permite subir los archivos .WAV y .TextGrid, ejecutar los análisis y obtener los resultados sin acudir a la ayuda constante de los docentes.

Como futura línea de investigación, destacamos la necesidad de medir ese grado de retroalimentación a los estudiantes de manera empírica.

Hasta donde sabemos es la única solución existente tanto bajo Praat, como fuera del mismo que permite este tipo de analíticas y ofrece la retroalimentación al estudiante en la lengua española. Resaltamos que a modo de retroalimentación y autoevaluación nuestra herramienta ofrece el porcentaje de similitud y diferencia de los valores de *pitch* para que el estudiante pueda corregir su pronunciación. Asimismo, como futuras líneas de trabajo, nos planteamos mejorar el entorno gráfico del *plugin* y abrir al estudiante, como usuario final, la posibilidad de su uso vía web.

Bibliografía

Boersma, P. y Weenink, D. 2019. Praat: doing phonetics by computer [Computer program]. Version 6.0.51. http://www.praat.org/.

Cantero Serena, F. J. 2002. *Teoría y análisis de la entonación*. Edicions Universitat Barcelona, vol. 54.

Cantero Serena, F. J. 2019. Análisis prosódico del habla: más allá de la melodía. *Comunicación Social: Lingüística, Medios Masivos, Arte, Etnología, Folclor y otras ciencias afines*, 2: 485-498.

Couto Fernández , T. 2021. Una herramienta de análisis del habla de audio para proporcionar retroalimentación automática a los estudiantes en la pronunciación en español. UDC. A Coruña

Dragos-PaulPop, A. A. 2013. Designing an mvc model for rapid web application development, [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S187770581400352X

Font-Rotchés, D. y Cantero Serena, F. J. 2008. La melodía del habla: acento, ritmo y entonación. *Eufonía: Didáctica de la música*, 19-39.

Font-Rotchés, D. y Cantero Serena, F. J. 2009. Melodic Analysis of Speech Method applied to Spanish and Catalan. *Phonica*, 5:33-47.

Helmer Strik, F. d. W. y Khiet Truong C. C. 2009. Comparing different approaches for automatic pronunciation error detection, *Speech Communication*, 845–852, [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S0167639309000715

Mateo, M. 2010a. Protocolo para la extracción de los datos tonales y curva estándar en análisis melódico del habla (AMH). *Phonica*, 6:49-90.

Mateo, M. 2010b. Scripts en Praat para la extracción de datos tonales y curva estándar. *Phonica*, 6:91-111.

Oplustil, P., y Toledo, G. 2019. Uso de una herramienta didáctica para la práctica de la entonación en hablantes no nativos de español. *Sintagma: Revista de lingüística*, 31:37–50.

Sutherland, K. S. J. 2020. La guía definitiva de scrum: Las reglas del juego. Disponible en: https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-Spanish-Latin-South-American.pdf

appForum: Una Aplicación para el Procesamiento de Foros

appForum: An application for Forums Processing

Alvaro Rodrigo,[1] José Luis Fernandez,[1] Jorge Pérez-Martín,[1] Ismael Iglesias,[2] Víctor Fresno,[1] Aitor Díaz,[1] Francisco Javier Sánchez,[2] Roberto Centeno[1]
[1]Intelligent Systems for Learning, Grupo de Innovación Docente de la UNED
[2]Intecca, UNED
alvarory@lsi.uned.es,jlvindel@dia.uned.es,jperezmartin@dia.uned.es,
iiglesias@intecca.uned.es,vfresno@lsi.uned.es,adiazm@scc.uned.es,
fjsanchez@intecca.uned.es,rcenteno@lsi.uned.es

Resumen: Los foros siguen siendo la forma predominante de comunicación en algunas comunidades y sobre todo en cursos virtuales. Estos foros no solo recogen las dudas de sus usuarios y las consiguientes respuestas, sino que contienen una gran información relativa a su frecuencia de uso, las dinámicas que se generan entre usuarios, etc. Para facilitar su procesamiento y posterior análisis, hemos desarrollado la aplicación *appForum*. Esta aplicación permite transformar a formato tabular los foros que recibe y ofrece distintas vistas y estadísticas sobre la información contenida en dichos foros. Debe servir también a futuro como plataforma sobre la que aplicar algoritmos inteligentes y basados en tecnologías del lenguaje.
Palabras clave: Foros, Procesamiento Lingüístico

Abstract: Forums are still the predominant form of communication in some communities and especially in virtual courses. These forums not only collect the doubts of their users and the consequent answers, but also contain a great deal of information regarding their frequency of use, the dynamics generated between users, etc. To facilitate their processing and subsequent analysis, we have developed the *appForum* application. This application allows you to transform the forums you receive into a tabular format and offers different views and statistics on the information contained in these forums. It should also serve in the future as a platform on which to apply intelligent algorithms based on language technologies.
Keywords: Forums, Linguistic Processing

1 Introducción

Los foros representan un recurso cada vez más importante para la comunicación entre usuarios en distintas comunidades. Esta importancia se hace todavía más evidente en cursos online, donde los usuarios normalmente no pueden comunicarse en persona y tienen como único recurso de comunicación los foros. Por otro lado, la proliferación de cursos en línea masivos y abiertos (MOOC), así como la enseñanza a distancia y virtual, que se ha manifestado tan importante durante la pandemia, nos han mostrado la importancia de este tipo de comunicaciones.

En los foros educativos, los distintos participantes (principalmente equipo docente y estudiantes) realizan intervenciones que pueden ser relativas o expresar dudas sobre contenidos o procedimientos académicos, respuestas a las dudas de otros usuarios, aportaciones relacionadas con la asignatura, etc. Todas estas comunicaciones se hacen a través de mensajes de texto y quedan registradas dentro de un curso virtual. Como consecuencia, los foros representan una gran fuente de información para conocer las interacciones entre los distintos usuarios, el seguimiento realizado, las dudas más comunes, etc.

Teniendo en cuenta estos factores, nuestro grupo de innovación docente, *ISL: Intelligent Systems for Learning*, ha desarrollado una aplicación centrada en el procesamiento de foros en el dominio educativo. La aplicación

se encarga de procesar foros de una comunidad o asignatura, y generar distintas visualizaciones, estructuradas por campos, lo que permite un mejor análisis de los foros, y con la posibilidad de ser anonimizadas. Además, se realizan distintos procesamientos, como por ejemplo análisis de sentimiento y de emociones, que nos pueden permitir realizar análisis más detallados.

El objetivo de esta aplicación es facilitar a los equipos docentes o a las coordinaciones de grado y másterla posibilidad de ejecutar diferentes tipos de analíticas sobre los mensajes en foros de una universidad. Actualmente funciona sobre los foros de la UNED, pero se pueden crear fácilmente adaptadores de los foros de cualquier plataforma. La aplicación se hará accesible a medida que se vayan desarrollando estos adaptadores a las diferentes herramientas de gestión de aprendizajes existentes. *appForum* está desarrollada en Python y utiliza Django para ofrecer sus funcionalidades a través de un interfaz web. A día de hoy la aplicación está siendo utilizada por distintos equipos docentes para analizar los datos de sus asignaturas. La evaluación de la herramienta será, por tanto, cualitativa; más adelante se agregarán todas las conclusiones extraídas por parte de los diferentes equipos docentes.

2 Flujo de Procesamiento

En la Figura 1 se muestra el flujo global de procesamiento de nuestra aplicación.

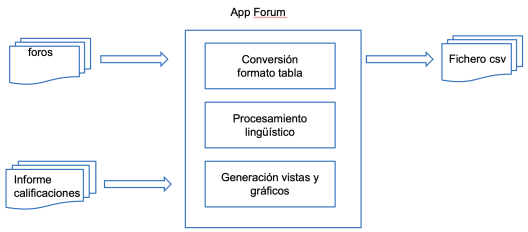


Figura 1: Fjujo de procesamiento de información de la aplicación *appForum*.

Conversión a formato tabla: La aplicación lee los foros en texto plano y los convierte a formato tabular estándar, facilitando después las vistas y los posteriores procesamientos, con campos relativos al NOMBRE DEL FORO, NOMBRE DEL HILO, NÚMERO DE MENSAJE dentro del hilo, MENSAJE AL QUE SE RESPONDE, el AUTOR, la FECHA, el TÍTULO DEL MENSAJE y el propio TEXTO DEL MENSAJE.

Procesamiento lingüístico: A continuación, se analizan todos los textos de los mensajes usando la librería spacy[1]. El objetivo es disponer de información lingüística (lemas, Entidades Nombradas, qué palabras no son palabras vacías, etc) que se pueda utilizar para crear las distintas vistas. Aplicamos análisis de sentimiento usando la librería sentiment-spanish [2] y de emoción siguiendo el modelo de afecto de Russell (Russell, 1980). Creamos también representaciones de cada mensaje dentro del Modelo de Espacio Vectorial y con TF-IDF como función de pesado.

Generación de Vistas y Gráficos: Posteriormente, y como paso siguiente, se generan cuatro vistas complementarias en relación a cómo se agrupa la información a mostrar: "*por mensajes*", "*por foros*", "*por participantes*" y "*por hilos*". Cada vista contiene información distinta (relacionada con dicha agrupación) y permite generar una serie de gráficos sobre la información mostrada en dicha vista.

Unido a lo anterior, la aplicación ofrece también las siguientes funcionalidades:

Informe de calificaciones: Permite cargar las calificaciones de los estudiantes de una asignatura en formato procesable y añadirla a la información de foros.

Exportación a fichero csv: Las distintas vistas e información de gráficos se pueden descargar en ficheros csv para su posterior procesamiento. La aplicación permite crear estos ficheros que pueden mostrar la información de foros desde distintas perspectivas, mezcladas con notas, así como con la inclusión de información lingüística.

En la siguiente sección vamos a describir las distintas vistas que ofrece la aplicación.

3 Vistas y Gráficos

Como hemos comentado en la sección anterior, *appForum* ofrece cuatro vistas principales en función de cómo agrupa la información contenida en los foros:

3.1 Vista de Mensajes

Esta es la vista por defecto, donde se muestran en formato tabla todos los mensajes publicados con su respectiva información. Esta información consta del NOMBRE DEL FORO y NOMBRE DEL HILO, el NÚMERO DE MENSAJE

[1] https://spacy.io/

[2] https://pypi.org/project/sentiment-analysis-spanish/

dentro del hilo, MENSAJE AL SE QUE RESPONDE, el AUTOR DEL MENSAJE, la FECHA y HORA DEL MENSAJE y el TÍTULO DEL MENSAJE. Además, muestra el número de caracteres de cada mensaje. Dado que la inclusión de cada mensaje en la vista podría no ser completa por su longitud, para poder ver cada mensaje hay que seleccionar la fila que lo contiene.

En la Figura 2 podemos ver un ejemplo de la vista de mensajes[3]. Nos permite además crear distintos gráficos sobre la información que se muestra. Algunos de estos gráficos son una nube de palabras generada a partir del peso TF-IDF de los términos dentro de cada mensaje (ver Figura 3, donde el tamaño codifica el peso del término y el color actualmente no está aportando ninguna semántica especial), la evolución temporal de las palabras más utilizadas (ver Figura 4), que se muestra como una animación, la distribución de los mensajes a lo largo de todo el curso o las horas con más mensajes por día de la semana.

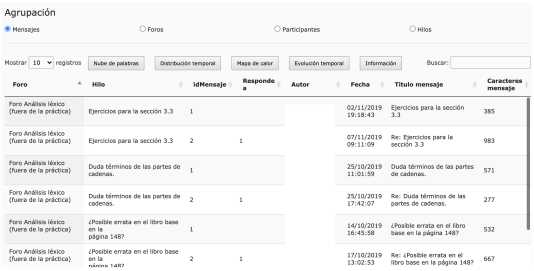


Figura 2: Ejemplo de cómo se muestra la información en la vista de mensajes.

Figura 3: Ejemplo de nube de palabras generada a partir del texto de los mensajes de un determinado foro.

Para hacer más visual la información relativa al sentimiento y emoción de cada mensaje, en lugar de mostrar valores numéricos o etiquetas, se muestra información gráfica.

[3]La información del 'Autor' ha sido anonimizada.

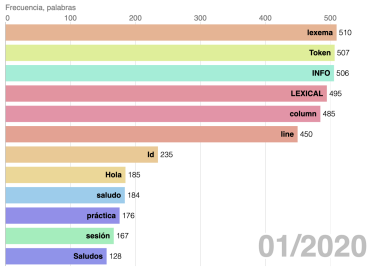


Figura 4: Ejemplo de un gráfico de barras animado que muestra a lo largo del tiempo los términos más relevantes acerca del contenido de un foro.

Se asocian colores a cada sentimiento (rojo a 'negativo', verde a 'positivo' y amarillo a 'neutral'), y para el análisis de emoción se utilizan distintos tipos de caras; por ejemplo, una cara sonriente para un texto con emoción 'alegre'. Actualmente aún se están ajustando los valores de referencia. En la Figura 5 se puede ver una primera versión de cómo quedaría la vista de mensajes incluyendo información gráfica de sentimiento y emoción.

Figura 5: Vista de mensajes incluyendo información de sentimiento y emoción.

3.2 Vista de Foros

Se ofrece agrupada por los foros en los que está dividido el fichero de entrada. Aunque el número de foros depende de cada asignatura y podemos encontrarnos con asignaturas con un solo foro y otras con, por ejemplo, un foro por tema, esta vista permite una visión más general de los distintos mensajes. La información que se ofrece en esta vista es relativa al NÚMERO DE AUTORES POR FORO, NÚMERO DE HILOS, NÚMERO DE MENSAJES y NÚMERO DE CARACTERES. Además, también se puede seleccionar un foro y obtener la nube de tags asociada a dicho foro. En la Figura 6 se muestra un ejemplo de esta vista.

3.3 Vista de Participantes

La tercera vista se centra en los usuarios que escriben mensajes en el foro, ya que no se

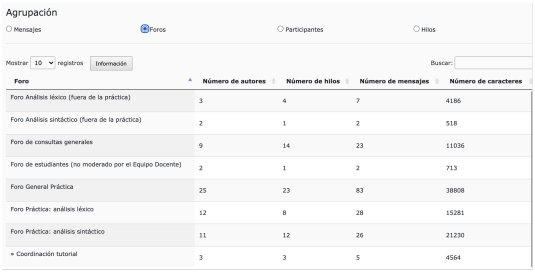


Figura 6: Ejemplo de cómo se muestra la información en la vista de foros.

dispone de información sobre los visitantes que no escriben. La información agrupada de esta vista permite realizar distintos análisis, como por ejemplo aquellos basados en análisis de redes sociales e interconexión de usuarios. Por ejemplo, a partir de esta vista podemos ver la estructura que representa la red de usuarios en función de los mensajes que publican y cómo se responden e interactúan (ver ejemplo en la Figura 8). Actualmente hemos incluido también medidas relativas a autoridades y hubs, pero en los foros analizados hasta la fecha no han ofrecido resultados relevantes. La Figura 7 ofrece una muestra de cómo se ofrece la información con esta vista.

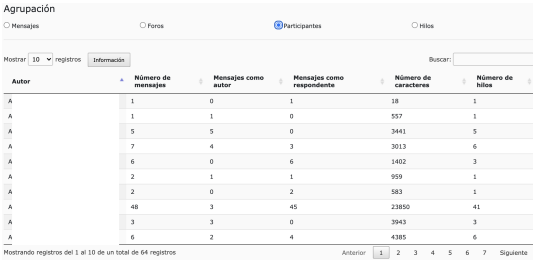


Figura 7: Ejemplo de cómo se muestra la información en la vista de participantes.

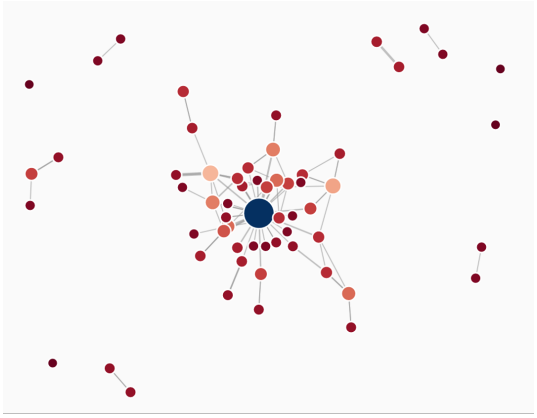


Figura 8: Representación en forma de grafo de las interacciones entre usuarios.

3.4 Vista de Hilos

La última vista que ofrece la aplicación agrupa los mensajes por los hilos donde se producen. La información que se ofrece es relativa al NÚMERO DE AUTORES, MENSAJES y CARACTERES de cada hilo. Permite además seleccionar y mostrar la nube de tags de cada hilo. En la Figura 9 se ve un ejemplo de cómo se muestra la información con esta vista.

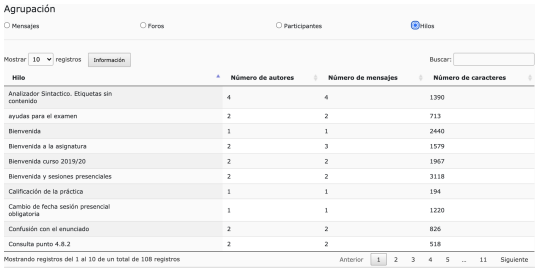


Figura 9: Ejemplo de cómo se muestra la información en la vista de hilos.

4 Conclusiones y trabajos futuros

Este trabajo presenta una aplicación, llamada *appForum*, para el análisis de foros en entornos educativos. Esta aplicación permite transformar a formato tabular el contenido de uno o varios foros de una asignatura, ofreciendo distintas vistas, así como estadísticas sobre la información contenida en los mismos. *appForum* integra actualmente diferentes procesos de análisis lingüístico, como un proceso de POS tagging o de análisis de sentimiento y emoción, que se pretenden ampliar incluyendo funciones de composición semántica, algoritmos de generación automática de resúmenes extractivos, de *Community Question&Answering*, detección de toxicidad en mensajes, lenguaje ofensivo, etc. aplicando algoritmos del estado del arte.

Agradecimientos

Este trabajo ha sido financiado con convocatorias de aplicativos de la UNED, así como por proyectos concedidos al grupo *ISL: Intelligent Systems for Learning* (GID2016-39) en las convocatorias PID 20/21 y 21/22.

Bibliografía

[Russell1980] Russell, J. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178.

A Neural Machine Translation System for Galician from Transliterated Portuguese Text

Un sistema de tradución neuronal para el gallego a partir de texto portugués transliterado

John E. Ortega, Iria de-Dios-Flores, Pablo Gamallo and José Ramom Pichel
Centro de Investigación en Tecnoloxías da Información (CITIUS)
{john.ortega,iria.dedios,pablo.gamallo,jramon.pichel}@usc.gal

Abstract: We present a neural machine translation (NMT) system for translating both Spanish and English to Galician (*ES–GL* and *EN–GL*). Galician is a language closely related to Portuguese, with low to medium resources, spoken in northwestern Spain. Our NMT system is trained on large-scale synthetic $ES \rightarrow PT \rightarrow GL$ and $EN \rightarrow PT \rightarrow GL$ parallel corpora created by the spelling transliteration of Portuguese to Galician from a high-quality Spanish to Portuguese (*ES–PT*) and English to Portuguese (*EN–PT*) translation memories. The NMT system is then made available via a public web interface at `https://demos.citius.usc.es/nos_tradutor`.
Keywords: Galician Language, Neural Machine Translation, Transliteration

Resumen: Presentamos un sistema de traducción automática neuronal (NMT) para traducir desde el castellano y el inglés al gallego (*ES–GL* y *EN–GL*). El gallego es una lengua estrechamente relacionada con el portugués, con recursos entre bajos y medios, que se habla en el noroeste de España. Nuestro sistema NMT se entrena con corpus paralelos sintéticos a gran escala $ES \rightarrow PT \rightarrow GL$ y $EN \rightarrow PT \rightarrow GL$ creados mediante la transliteración ortográfica del portugués al gallego a partir de unas memorias de traducción de alta calidad del español al portugués (*ES–PT*) y del inglés al portugués (*EN–PT*). El sistema NMT está disponible a través de una interfaz web pública en `https://demos.citius.usc.es/nos_tradutor`.
Palabras clave: Gallego, Traducción automática neuronal, Transliteración

1 Introduction

Several systems have been compared and developed to perform machine translation (MT), ranging from rule-based systems to systems based on neural networks (Knowles, Ortega, and Koehn, 2018) Traditionally, rule-based systems like Apertium (Forcada et al., 2011) are used for languages with a small amount of parallel data. That is because MT systems backed by neural networks, or neural machine translation (NMT) systems, require high amounts of data, typically on the order of millions of sentences or more (Bahdanau, Cho, and Bengio, 2014; Koehn and Knowles, 2017). An interesting option for low-resource languages is the use of zero-shot translation techniques, that is, translating in multilingual settings between language pairs for which the NMT system has never been trained. However, as Gu et al. (2019) state, training zero-shot NMT models easily fails as this task is very sensitive to hyper-parameter setting. The performance of zero-shot strategies is usually lower than that of more conventional pivot-based approaches.

We describe and implement an approach inspired by previous work (Campos et al., 2009) that uses the proximity of Portuguese and Galician to overcome the lack of resources problem and produces corpora to build an NMT system, similar to low-resource NMT systems found in previous work (Ortega, Mamani, and Cho, 2020; Ortega, Castro-Mamani, and Montoya Samame, 2020), for translating both Spanish to Galician and English to Galician. Our system first uses high-quality Spanish–Portuguese (ES–PT) and English–Portuguese (EN–PT) parallel corpora to translate the target-sided (Portuguese) sentences (or segments) to Galician using *transliteration*, the conversion of

text in one language to another through spelling. Transliteration between Portuguese and Galician works well due to the orthographic nearness of the two languages found in previous work (Pichel et al., 2021). Second, NMT systems with the transliterated Galician parallel text are created to form a Spanish–Galician (ES–GL) and English–Galician (EN–GL) MT system where both Spanish and English are the source languages and Galician is the target language. Two different neural-based architectures were tested: Long short-term memory (LSTM) and Transformers.

2 Method

Our translation strategy consists of two steps. The first step uses *transliteration* (Knight and Graehl, 1997) to create parallel Galician segments from the Portuguese segments in the aligned corpus, by making using of the transliteration tool port2gal[1], which contains several hundreds of rules on characters and sequences of characters. Both training and validation sets are transliterated leaving a final parallel Galician corpus. Then, in the second step, the Galician (transliterated) corpus is used to train an NMT system with Spanish or English as the source language and Galician as the target language. For the first transliteration step, we also tested a more complex strategy by combining PT→GL Apertium translator (Forcada et al., 2011), which uses a basic bilingual dictionary to translate word by word, with the transliteration tool for those words that are not in the bilingual dictionary.

The NMT system that we use for ES–GL and EN–GL translations was created using OpenNMT (Klein et al., 2017), a generic deep learning framework for creating sequence-to-sequence models in machine translation. In particular, we trained a LSTM (long short term memory) seq2seq model as well as a Transformer model for each language pair.

Concerning LSTM, we used the following default neural network training parameters: two hidden layers, 500 hidden LSTM units per layer, input feeding enabled, 13 epochs, batch size of 64. Alternatively, we modified the default learning step parameters to 100,000 training steps and 10,000 validation steps. Traditional tokenization was performed with Linguakit (Gamallo et al., 2018)

The Transformer implementation, described in Garg et al. (2019), was configured with default training parameters: 6 layers for both encoding and decoding and batch size of 4096 tokens. We also modified the learning step parameters to the same values as the LSTM configuration. In this case, we used sub-word tokenization, performed with SentencePiece (Kudo and Richardson, 2018).

3 Corpora

The main parallel sources we used to train the NMT system come from Opus[2]. In particular we used the *ES–PT* and *EN–PT* partitions of both Europarl[3], with about 2 million sentences per language, and OpenSubtitles[4], containing about 30 million sentences in *ES–PT* and 25 in *EN–PT*. The Portuguese partition was transliterated to Galician so as to build *ES–GL* and *EN–GL* parallel corpora. In addition, we also added the Spanish-Galician partition of CLUVI[5], to the *ES–GL* corpus, containing 144 thousand sentences.

4 Test results

Table 1 show the results of different experiments for *ES–GL* and *EN–GL* combining the system, LSTM or Transformer, with the size of the corpus. We observe that LSTM works very well for close languages (*ES–GL*), but for the pair (*EN–GL*), two distant languages, the results are slightly better with Transformer. In addition, we also observe that the whole OpenSubtitles corpus hurts the performance in *ES–GL*. The best results in *ES–GL* combine Europarl with OpenSubtitles and are comparable to the state-of-the-art (Bayón and Sánchez-Gijón, 2019). Let us note that the Movie and TV subtitles of OpenSubtitles are a highly valuable resource but the quality of the resulting sentence alignments is often lower than for other parallel corpora (Lison, Tiedemann, and Kouylekov, 2018). The results in Table 1 allow us to confirm that using transliteration between two closely aligned languages like Portuguese and Galician, favorable outcomes can be achieved.

[1] https://github.com/gamallo/port2gal

[2] https://opus.nlpl.eu

[3] https://opus.nlpl.eu/Europarl.php

[4] https://opus.nlpl.eu/OpenSubtitles.php

[5] https://repositori.upf.edu/handle/10230/20051

system	pair	source	corpus size	bleu	ter	chrF2
lstm	es-gl	Europarl+CLUVI	2.35M	48.9	34.4	69.3
lstm	es-gl	Europarl+CLUVI+OpenSubt(part)	5M	**51.1**	**32.8**	**70.8**
lstm	es-gl	Europarl+CLUVI+OpenSubt	30M	46.0	37.2	66.5
transformer	es-gl	Europarl+CLUVI	2.35M	17.5	67.4	53.0
transformer	es-gl	Europarl+CLUVI+OpenSubt	30M	13.9	66.7	46.4
lstm	en-gl	Europarl+OpenSubt	27.M	26.6	50.3	45.5
transformer	en-gl	Europarl+OpenSubt	27.M	**29.3**	**49.7**	**51.0**

Table 1: Results obtained for the two language pairs (*ES–GL* and *EN–GL*) evaluated on two different systems, LSTM and Transformer, by making use of three quantitative measures: BLEU, TER and ChrF2. The corpus size is quantified in millions of sentences (M).

Figure 1: A screen capture of the web interface.

5 *Demonstration*

Our demonstration is made up of a public-facing web page[6] that provides Galician translations for both Spanish and English inputs. Users will be able to test the system via an open web interface (see Figure 1) where they could select the language pair (*ES–GL* or *EN–GL*) and translation system (LSTM or Transformer) to then enter text and generate translations.

In our demonstration, we plan to show where our system performs well and where it does not perform well. As an example, the sentence translated from Spanish to Galician using the LSTM system in Table 2 is an excellent translation despite its long length. Additionally, our system translations perform well with syntax and seem to generally translate better than previous systems tested on the same domain. Nonetheless, we have found that when comparing our system's performance for lexical and morphological quality, the Portuguese transliteration affect the performance, found to be better on other rule-based MT systems like Apertium (Forcada et al., 2011) for example.

[6] `https://demos.citius.usc.es/nos_tradutor`

6 *Future work*

We plan to perform further work with a human-in-the-loop to increase the performance based on quality. This is outlined by a continuous improvement plan which insinuates the inclusion of translators for user functionality tests. For example, spelling and lexical issues such as *acidente* instead of *accidente*, formal Galician differences that need to be addressed are first to be solved using newly-developed heuristics as part of our future contingency plan. The aim will be to create the highest-quality system in order expand the language pairs to other languages such as Russian or Chinese.

References

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bayón, M. D. C. and P. Sánchez-Gijón. 2019. Evaluating machine translation in a low-resource language combination: Spanish-galician. In *Machine Translation Summit XVII Vol. 2: Translator, Project and User Tracks*, pages 30–35.

Campos, J. R. P., P. M. Fernández, O. Gomez, P. Gamallo, and A. C. García. 2009. Carvalho: English-galician smt system from europarl english-portuguese parallel corpus. *Procesamiento Del Lenguaje Natural*, pages 379–381.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A.

Spanish	Galician
Debemos imponer el cumplimiento de los reglamentos y velar por que se aplique el principio de que "el que contamina paga" para que se utilicen sanciones y también incentivos financieros a fin de presionar a los propietarios de los buques y las compañías petroleras y lograr que se introduzcan los procedimientos mejores.	Temos de impor o cumpremento dos regulamentos e celar por que o principio do poluidor-pagador sexa aplicado para que sexan utilizadas sancións e tamén incentivos financeiros a fin de exercer presión sobre os proprietarios dos navíos e das compañías petrolíferas e conseguir que os procedementos mellores sexan introducidos.

Table 2: Translation using the best performing machine translation system (LSTM).

Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Gamallo, P., M. Garcia, C. Piñeiro, R. Martinez-Castaño, and J. C. Pichel. 2018. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Garg, S., S. Peitz, U. Nallasamy, and M. Paulik. 2019. Jointly learning to align and translate with transformer models. *CoRR*, abs/1909.02074.

Gu, J., Y. Wang, K. Cho, and V. O. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy, July. Association for Computational Linguistics.

Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations.*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Knight, K. and J. Graehl. 1997. Machine transliteration. *arXiv preprint cmp-lg/9704003*.

Knowles, R., J. Ortega, and P. Koehn. 2018. A comparison of machine translation paradigms for use in black-box fuzzy-match repair. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 249–255.

Koehn, P. and R. Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Kudo, T. and J. Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Lison, P., J. Tiedemann, and M. Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Ortega, J. E., R. A. Castro-Mamani, and J. R. Montoya Samame. 2020. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China, December. Association for Computational Linguistics.

Ortega, J. E., R. C. Mamani, and K. Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Pichel, J. R., P. Gamallo, I. Alegria, and M. Neves. 2021. A methodology to measure the diachronic language distance between three languages based on perplexity. *Journal of Quantitative Linguistics*, 28(4):306–336.