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

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)

A Coruña, Spain, September 21-23, 2022

https://sepln2022.grupolys.org/

Preface for SEPLN-PD 2022 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.

Editors of the Proceedings

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

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

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)

Organization

Sponsors

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[1], José A. Troyano[1], Fermín L. Cruz[1] and M. Alonso-Dos-Santos[2]

[1] *Universidad de Sevilla*
[2] *Universidad de Granada*

Abstract

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.

English translation. Our project focuses on the identification of competitive factors of hotel establishments in Andalusia. To do so, we will use Natural Language Processing techniques applied to online reviews published by users through 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

Contenidos generados por usuarios, calidad relacional, tecnologías del lenguaje, sector turístico.

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 [1]. 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 [2]. Las revisiones *online* son espontáneas, esclarecedoras, e incluso apasionadas, fácilmente accesibles desde cualquier lugar y en cualquier momento [3] son recuerdos o reconstrucciones cognitivas de un viaje o una estancia que reducen significativa y aparentemente el riesgo potencial de la compra [4]. 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 [5].

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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

majesus@us.es (M.J. Sánchez-Franco); troyano@us.es (J. A. Troyano); fcruz@us.es (F. L. Cruz); manuelalonso@ugr.es (M. Alonso-Dos-Santos)

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 [6].

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 [7] 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 [7].
- 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

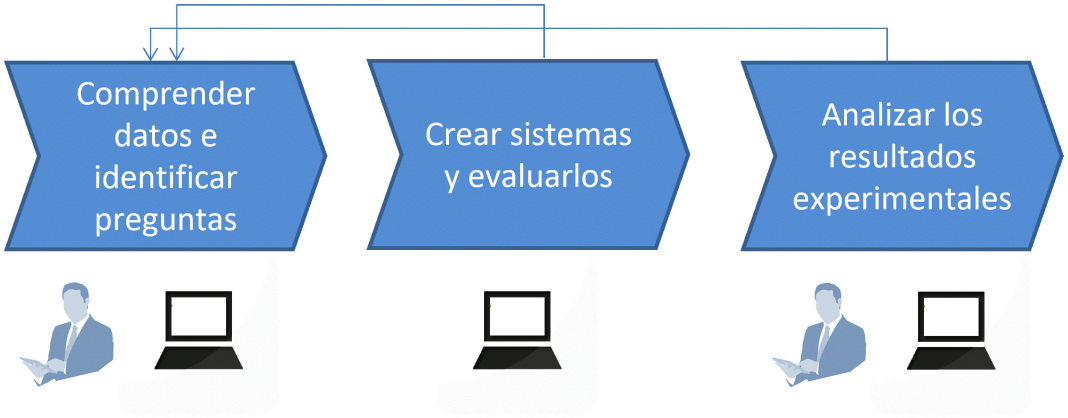


Figure 1: Proceso metodológico

(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 [8], pasando por modelos basados en bolsas de palabra para tareas de clasificación [9], *pipelines* clásicos para tareas de extracción de información [10], hasta las más recientes técnicas basadas en redes profundas [11] o *transformers* [12].

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 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.

Acknowledgments

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.

References

[1] J. y. A. L. Consejería de Turismo, Regeneración, Plan General de Turismo Sostenible de Andalucía Horizonte 2020, Technical Report, Junta de Andalucía, 2016.

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

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

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

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

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

[7] B. Fogg, D. Eckles, The behavior chain for online participation: how successful web services structure persuasion, in: International Conference on Persuasive Technology, Springer, 2007, pp. 199–209.

[8] A. Peñas, F. Verdejo, J. Gonzalo, et al., Corpus-based terminology extraction applied to information access, in: Proceedings of corpus linguistics, volume 2001, 2001, p. 458.

[9] T. M. Alam, M. J. Awan, Domain analysis of information extraction techniques, Int. J. Multidiscip. Sci. Eng 9 (2018) 1–9.

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

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

[12] M. Munikar, S. Shakya, A. Shrestha, Fine-grained sentiment classification using bert, in: 2019 Artificial Intelligence for Transforming Business and Society (AITB), volume 1, IEEE, 2019, pp. 1–5.

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] and Patricio Martínez Barco[2]

[1]*Computer Science Department, SINAI, CEATIC*
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
[2]*Department of Software and Computing Systems,*
University of Alicante, Spain

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.

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. It has also enabled the organisation of complex social structures that have passed on culture and knowledge generation after generation through the use of a common language [1]. 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 [2]. In this way, language becomes a very valuable resource for the cognitive modelling of an individual as studied in psycholinguistics [3], but also for understanding social interactions and communities in what is known as Computational Social Sciences [4]. 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 [5], 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 [6] [7]. 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...) [8], 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 sce-

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
laurena@ujaen.es (L. A. Ureña-López); stela@dlsi.ua.es (E. Saquete); laurena@ujaen.es (M. Martín-Valdivia); patricio@dlsi.ua.es (P. M. Barco)
0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-6001-5461 (E. Saquete); 0000-0002-6001-5461 (M. Martín-Valdivia); 0000-0002-6001-5461 (P. M. Barco)

nario and often resulting in the deterioration of our cognitive abilities [9] [10]. 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. Therefore, for example, a change in pattern of the type of language used in the communication between two people can help detect the start of practices such as sexual harassment, when language moves from a suggestive, captivating or friendly language to that of a coercive or threatening nature. As shown in Figure 1, the relationships between entities are dynamic and change with time as do their properties. By identifying these variations and their patterns based on human language, we can prepare these networks for the future by creating predictive models of peoples' behaviour (risk detection, prevention of cyberbullying, terrorist warnings, etc.).

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 emotional) 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.

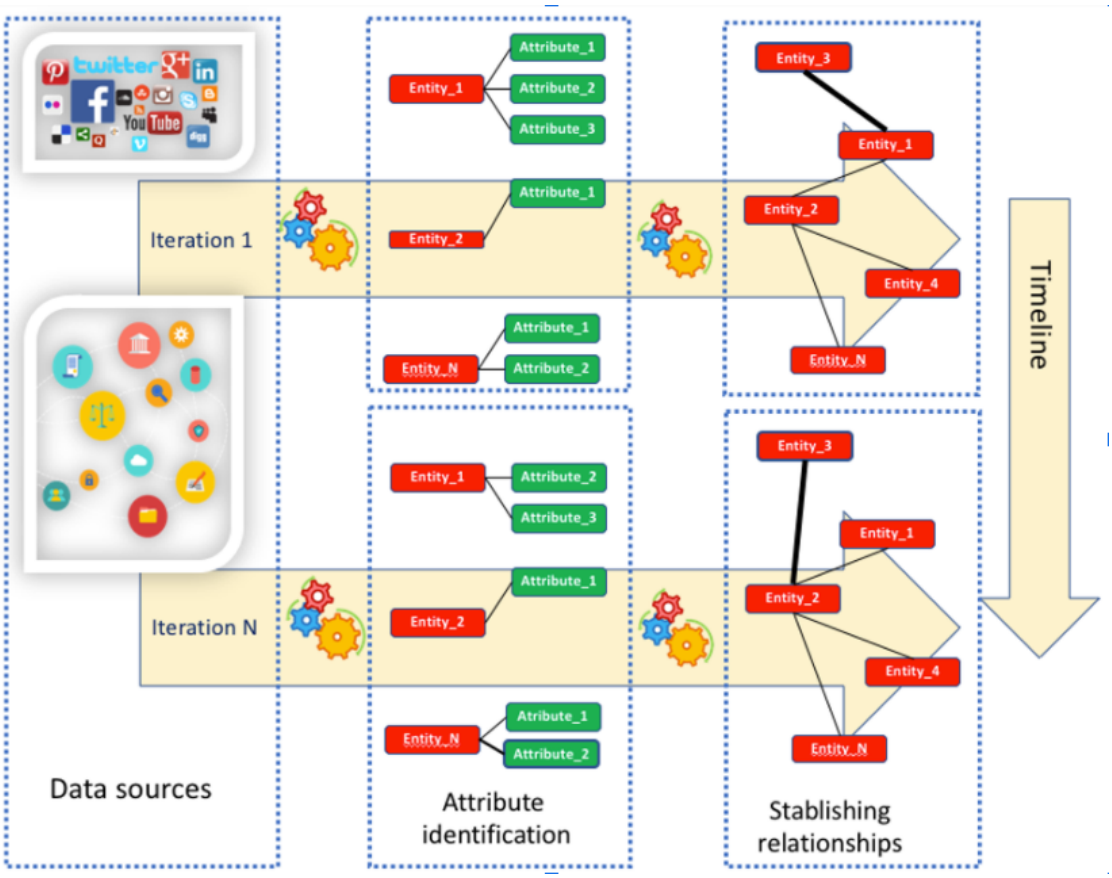


Figure 1: Detection and monitoring of digital entities – Representation of an evolving environment over time

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 sub-projects 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 [11, 12], 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 [13]. In addition, knowledge discovery techniques are being applied to the health domain [14, 15]. 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 [16].

Results regarding OBJ3: In relation to this objective, a systematic study of the state of the art in this matter has been carried out[17] 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 [18]. 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 [19]. In addition, work has been done on emotion detection [20] [21] and negation [22].

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 [23]. 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].

4. Acknowledgments

This research work is funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union" or by the "European Union NextGenerationEU/PRTR" through the grant LIVING-LANG Project (RTI2018-094653-B-C21 / C22). It is a coordinated project with SINAI and GPLSI as participating research groups. It is also funded by Generalitat Valenciana through the project NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation (CIPROM/2021/21).

References

[1] M. Tomasello, A natural history of human thinking, Harvard University Press, 2014.

[2] G. W. Grace, The linguistic construction of reality, Routledge, 2016.

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

[4] H. Wallach, Computational social science, Comput. Soc. Sci. 307 (2016).

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

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

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

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

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

[10] N. Carr, The shallows: What the Internet is doing to our brains, WW Norton & Company, 2020.

[11] F. M. P. del Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Comparing pre-trained language models for spanish hate speech detection, Expert Syst. Appl. 166 (2021) 114120. URL: https://doi.org/10.1016/j.eswa.2020.114120. doi:10.1016/j.eswa.2020.114120.

[12] F. M. P. del Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Detecting misogyny and xenophobia in spanish tweets using language technologies, ACM Trans. Internet Techn. 20 (2020) 12:1–12:19. URL: https://doi.org/10.1145/3369869. doi:10.1145/3369869.

[13] A. Piad-Morffis, Y. Gutiérrez, Y. Almeida-Cruz, R. Muñoz, A computational ecosystem to support ehealth knowledge discovery technologies in spanish, J. Biomed. Informatics 109 (2020) 103517. URL: https://doi.org/10.1016/j.jbi.2020.103517. doi:10.1016/j.jbi.2020.103517.

[14] P. López-Úbeda, M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. U. López, M. T. Martín-Valdivia, Automatic medical protocol classification using machine learning approaches, Comput. Methods Programs Biomed. 200 (2021) 105939. URL: https://doi.org/10.1016/j.cmpb.2021.105939. doi:10.1016/j.cmpb.2021.105939.

[15] P. López-Úbeda, M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. U. López, M. T. Martín-Valdivia, COVID-19 detection in radiological text reports integrating entity recognition, Comput. Biol. Medicine 127 (2020) 104066. URL: https://doi.org/10.1016/j.compbiomed.2020.104066. doi:10.1016/j.compbiomed.2020.104066.

[16] C. Barros, E. Lloret, E. Saquete, B. Navarro-Colorado, Natsum: Narrative abstractive

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

summarization through cross-document timeline generation, Inform. Proc. Manag. 56 (2019) 1775–1793. URL: https://www.sciencedirect.com/science/article/pii/S0306457318305922. doi:https://doi.org/10.1016/j.ipm.2019.02.010.

[17] E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, M. Palomar, Fighting post-truth using natural language processing: A review and open challenges, Expert Syst. Appl. 141 (2020). URL: https://doi.org/10.1016/j.eswa.2019.112943. doi:10.1016/j.eswa.2019.112943.

[18] A. Bonet-Jover, A. Piad-Morffis, E. Saquete, P. Martínez-Barco, M. Á. G. Cumbreras, Exploiting discourse structure of traditional digital media to enhance automatic fake news detection, Expert Syst. Appl. 169 (2021) 114340. URL: https://doi.org/10.1016/j.eswa.2020.114340. doi:10.1016/j.eswa.2020.114340.

[19] R. Sepúlveda-Torres, M. E. Vicente, E. Saquete, E. Lloret, M. Palomar, Headlinestancechecker: Exploiting summarization to detect headline disinformation, J. Web Semant. 71 (2021) 100660. URL: https://doi.org/10.1016/j.websem.2021.100660. doi:10.1016/j.websem.2021.100660.

[20] L. Canales, C. Strapparava, E. Boldrini, P. Martínez-Barco, Intensional learning to efficiently build up automatically annotated emotion corpora, IEEE Trans. Affect. Comput. 11 (2020) 335–347. URL: https://doi.org/10.1109/TAFFC.2017.2764470. doi:10.1109/TAFFC.2017.2764470.

[21] L. Canales, W. Daelemans, E. Boldrini, P. Martínez-Barco, Emolabel: Semi-automatic methodology for emotion annotation of social media text, IEEE Trans. Affect. Comput early access (2019) 1–1. doi:10.1109/TAFFC.2019.2927564.

[22] S. M. Jiménez-Zafra, R. Morante, M. T. Martín-Valdivia, L. A. Ureña-López, Corpora annotated with negation: An overview, Comput. Linguistics 46 (2020) 1–52. URL: https://doi.org/10.1162/coli_a_00371. doi:10.1162/coli_a_00371.

[23] E. Saquete, J. Zubcoff, Y. Gutiérrez, P. Martínez-Barco, J. Fernández, Why are some social-media contents more popular than others? opinion and association rules mining applied to virality patterns discovery, Expert Syst. Appl. 197 (2022) 116676. URL: https://doi.org/10.1016/j.eswa.2022.116676. doi:10.1016/j.eswa.2022.116676.

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] and Itziar Cuenca[3]

[1]*Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, 20009, Spain*
[2]*Biokeralty Research Institute AIE, Vitoria-Gasteiz, 01510, Spain*
[3]*Instituto Ibermática de Innovación (i3B), Donostia-San Sebastián, 20009, Spain*

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

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 [1, 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 [2, 3].

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.

2. Consortium and funding body

ESAN is partially funded by the Basque Government through the Elkartek 2021 program of the SPRI Group under the grant agreement KK-2021/00117. It will run from 04/2021 to 12/2023.

The consortium includes the Vicomtech research centre, (https://www.vicomtech.org), Grupo Keralty's R&D division BioKeralty Research Institute (https://biokeralty.com), and Grupo Ibermática's R&D business unit and project leader Instituto Ibermática de Innovación or i3B (https://ibermatica.com/en/innovacion/).

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
nperez@vicomtech.org (N. Perez); aalvarez@vicomtech.org (A. Álvarez); adelpozo@vicomtech.org (A. d. Pozo); andres.arbona@keralty.com (A. Arbona); oihane.ibarrolv@keralty.com (O. Ibarrola); marta.suarez@keralty.com (M. Suarez); pm.delapena@ibermatica.com (P. d. l. P. Tejada); ia.cuenca@ibermatica.com (I. Cuenca)
0000-0001-8648-0428 (N. Perez); 0000-0002-7938-4486 (A. Álvarez)

3. 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 §4) 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 and other publicly available data that might be considered beneficial.. 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 and classification 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).

4. 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 technical 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 conversations 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 dialogue, 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.

5. Approach

5.1. 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, prior to meeting their doctors.

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 (https://www.audacityteam.org) in PCM WAV format at 44.1kHz and 24 bits.

5.2. 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 (http://trans.sourceforge.net) 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, aimed at making the annotation task more productive, 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 [4] following our previous approach based on CNN layers and a TDNN-F network [5]. 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 [6], which will be also adapted to the domain. Finally, new speaker diarisation models will be trained for the Kaldi X-Vectors-based system [7] to be developed.

5.3. 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 [8, 9, 10]. 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 (severity, location, dosage, etc.).

The models for the automatic detection and classification of this information will be based on the ubiquitous Transformers architecture [11]. We plan on exploiting the latest neural language models for Spanish and the biomedical domain [12, 13]. This line of work will also profit from previous work of consortium members on clinical IE [14, 15, 16].

5.4. 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.

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.

6. 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 [17, among others].

Acknowledgments

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

References

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

[2] P. Bhatia, S. Lin, R. Gangadharaiah, B. Wallace, I. Shafran, C. Shivade, N. Du, M. Diab (Eds.), Proceedings of the 1st Workshop on NLPMC, 2020.

[3] C. Shivade, R. Gangadharaiah, S. Gella, S. Konam, S. Yuan, Y. Zhang, P. Bhatia, B. Wallace (Eds.), Proceedings of the 2nd Workshop on NLPMC, 2021.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The kaldi speech recognition toolkit, in: Proceedings of IEEE ASRU, 2011, pp. 1–4.

[5] A. Álvarez, H. Arzelus, I. G. Torre, A. González-Docasal, Evaluating novel speech transcription architectures on the Spanish RTVE2020 Database, Appl. Sci. 12 (2022) 1–16.

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

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-Vectors: Robust DNN embeddings for speaker recognition, in: Proceedings of ICASSP, 2018, pp. 5329–5333.

[8] I. Shafran, 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, J. S. Paul, The Medical Scribe: Corpus development and model performance analyses, in: Proceedings of LREC, 2020, pp. 2036–2044.

[9] P. Chocrón, Á. Abella, G. de Maeztu, ContextMEL: Classifying contextual modifiers in clinical text, Proces. de Leng. Nat. 65 (2020) 45–52.

[10] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The E3C project: Collection and annotation of a multilingual Corpus of Clinical Cases, in: Proceedings of CLiC-it 2020, 2021, pp. 1–7.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of NIPS, 2017, pp. 6000–6010.

[12] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, F. J. Veredas, Detection of tumor morphology mentions in clinical reports in spanish using transformers, in: Proceedings of IWANN, 2021, pp. 24—-35.

[13] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of BioNLP, 2022, pp. 193–199.

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

[15] S. Lima-López, N. Perez, M. Cuadros, G. Rigau, NUBes: A corpus of negation and uncertainty in Spanish clinical texts, in: Proceedings of LREC, 2020, pp. 5772–5781.

[16] A. García-Pablos, N. Perez, M. Cuadros, Vicomtech at eHealth-KD challenge 2021: Deep learning approaches to model health-related text in Spanish, in: Proceedings of IberLEF, 2021, pp. 712–724.

[17] P. J. Vivancos-Vicente, J. A. García-Díaz, J. S. Castejón-Garrido, R. Valencia-García, ISMR - Sistema basado en Deep Learning para la transcripción y extracción de conocimiento en entrevistas médico-paciente, in: Proceedings of SEPLN-PD, 2021, pp. 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], 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] and María Rosario García Viedma[2]

[1]*Department of Computer Science, Advanced Studies Center in ICT (CEATIC), Universidad de Jaén. Campus Las Lagunillas, E-23071, Jaén, Spain*

[2]*Department of Psychology, Universidad de Jaén. Campus Las Lagunillas, E-23071, Jaén, Spain*

Abstract

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.

English translation. 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

Asistentes virtuales inteligentes, sistemas de diálogo, envejecimiento activo.

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 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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

magc@ujaen.es (M. García-Cumbreras); dofer@ujaen.es (F. Martínez-Santiago); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. T. Martín-Valdivia); amontejo@ujaen.es (A. Montejo-Ráez); mgarcia@ujaen.es (M. García-Vega); mcdiaz@ujaen.es (M. C. Dıaz-Galiano); mdmolina@ujaen.es (M. D. Molina-González); sjzafra@ujaen.es (S. M. Jiménez-Zafra); fmplaza@ujaen.es (F. M. Plaza-del-Arco); mrgarcia@ujaen.es (M. R. G. Viedma)

0000-0003-1867-9587 (M. García-Cumbreras); 0000-0002-1480-1752 (F. Martínez-Santiago); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-2874-0401 (M. T. Martín-Valdivia); 0000-0002-8643-2714 (A. Montejo-Ráez); 0000-0003-2850-4940 (M. García-Vega); 0000-0001-9298-1376 (M. C. Dıaz-Galiano); 0000-0002-8348-7154 (M. D. Molina-González); 0000-0003-3274-8825 (S. M. Jiménez-Zafra); 0000-0002-3020-5512 (F. M. Plaza-del-Arco)

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

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"[1].

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[2].

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[3][4].

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 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

[2] https://developer.amazon.com/es-ES/alexa
[3] https://docs.aws.amazon.com/lambda/index.html

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[5].

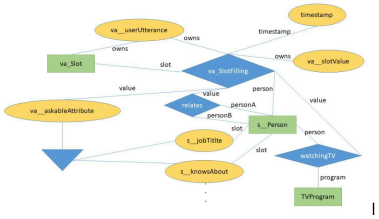


Figure 1: Modelo de datos del proyecto Inlife.

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)[6]. 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.

Acknowledgments

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.

References

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

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

[3] J. Allen, Natural Language Understanding, The Benjamin/Cummings Publishing Company, Inc., 1995.

[4] F. Martínez-Santiago, M. Díaz-Galiano, M. García-Cumbreras, A. Montejo-Ráez, A method based on rules and machine learning for logic form identification in spanish, Natural Language Engineering 23 (2015) 1–23. doi:10.1017/S1351324915000297.

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

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

Big Hug: Artificial intelligence for the protection of digital societies

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

Arturo Montejo-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-Plaza[1], María Dolores Molina-González[1], Luis-Joaquin García-López[2] and María Belén Díez-Bedmar[3]

[1]*Department of Computer Science, Advanced Studies Center in ICT (CEATIC), Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain*

[2]*Department of Psychology, Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain*

[3]*Department of English Studies, Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain*

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.

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[1, 2].

To this end, the project proposes two years of ac-

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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

amontejo@ujaen.es (A. Montejo-Ráez); maite@ujaen.es (M. T. Martín-Valdivia); laurena@ujaen.es (L. A. Ureña-López); mcdiaz@ujaen.es (M. C. Díaz-Galiano); magc@ujaen.es (M. García-Cumbreras); mgarcia@ujaen.es (M. García-Vega); dofer@ujaen.es (F. Martínez-Santiago); fmplaza@ujaen.es (F. M. Plaza-del-Arco); sjzafra@ujaen.es (S. M. Jiménez-Plaza); mdmolina@ujaen.es (M. D. Molina-González); ljgarcia@ujaen.es (L. García-López); belendb@ujaen.es (M. B. Díez-Bedmar)
0000-0002-8643-2714 (A. Montejo-Ráez); 0000-0002-2874-0401 (M. T. Martín-Valdivia); 0000-0001-9752-2830 (L. A. Ureña-López); 0000-0001-9298-1376 (M. C. Díaz-Galiano); 0000-0003-1867-9587 (M. García-Cumbreras); 0000-0003-2850-4940 (M. García-Vega); 0000-0002-1480-1752 (F. Martínez-Santiago); 0000-0002-3020-5512 (F. M. Plaza-del-Arco); 0000-0003-3274-8825 (S. M. Jiménez-Plaza); 0000-0002-8348-7154 (M. D. Molina-González); 0000-0003-0446-6740 (L. García-López); 0000-0001-9250-2224 (M. B. Díez-Bedmar)

tivities 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.

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 [3]. Navarro-Gómez [4] 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) [5].

In another study [6], 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 [7], 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 [8]. Depression can be studied and identified by monitoring users' posts and activity [1].

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 [9]. 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 [10].

NLP techniques are being applied to the analysis of social media textual data to face new problems like fake-news detection [11], offensive language identification [12], sentiment analysis [13], opinion mining and emotion detection [14]. 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 [15]. 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 [16], but massive amounts of data are required. During the last two years we have witnessed the birth of amazing models like BERT [17], GPT-2 [18] or Transformer-XL [19], 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 [20] or MENHIR [21]. 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

[1] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: International Conference of the CLEF for European Languages, Springer, 2019, pp. 340–357.

[2] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, eRisk 2021: pathological gambling, self-harm and depression challenges, in: ECIR, Springer, 2021, pp. 650–656.

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

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

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

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

[7] T. Wang, M. Brede, A. Ianni, E. Mentzakis, Detecting and characterizing eating-disorder communities on social media, in: Proceedings of the Tenth ACM International conference on web search and data mining, 2017, pp. 91–100.

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

[9] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, Z. Huang, Suicidal ideation detection: A review of machine learning methods and appli-

cations, IEEE Transactions on Computational Social Systems 8 (2020) 214–226.
[10] J. J. Glenn, A. L. Nobles, L. E. Barnes, B. A. Teachman, 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 (2020) 704–722.
[11] F. Monti, F. Frasca, D. Eynard, D. Mannion, M. M. Bronstein, Fake news detection on social media using geometric deep learning, arXiv preprint arXiv:1902.06673 (2019).
[12] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), arXiv preprint arXiv:1903.08983 (2019).
[13] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Urena-López, A. R. Montejo-Ráez, Sentiment analysis in twitter, Natural Language Engineering 20 (2014) 1–28.
[14] F. M. Plaza-del Arco, M. T. Martín-Valdivia, L. A. Ureña-López, R. Mitkov, Improved emotion recognition in spanish social media through incorporation of lexical knowledge, Future Generation Computer Systems 110 (2020) 1000–1008.
[15] J. Dean, D. Patterson, C. Young, A new golden age in computer architecture: Empowering the machine-learning revolution, IEEE Micro 38 (2018) 21–29.
[16] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, ieee Computational intelligenCe magazine 13 (2018) 55–75.
[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[19] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, R. Salakhutdinov, Transformer-xl: Attentive language models beyond a fixed-length context, arXiv preprint arXiv:1901.02860 (2019).
[20] D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, D. A. Velazquez, J. M. Gonfaus, J. Gonzàlez, et al., Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis, Journal of medical internet research 22 (2020) e17758.
[21] M. Kraus, P. Seldschopf, W. Minker, Towards the Development of a Trustworthy Chatbot for Mental Health Applications, in: MultiMedia Modeling, Springer, 2021, pp. 354–366.

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] and Igone Zabala [2]

[1] *Universidade da Coruña and CITIC, Campus da Zapateira s/n, A Coruña, 15071, Spain*
[2] *Universidad del País Vasco/Euskal Herriko Unibertsitatea, Barrio Sarriena s/n, Leioa, 48940, Spain*

Abstract

Academic writing has become a priority object of study especially in English, for which there are already many resources to help novice writers. This is not the case for Spanish university students who do not have many writing aids at their disposal. Here 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 functions, writing aid.

1. Introduction

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 Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU) and the Universidade da Coruña (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 ([1], [2] 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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
EMAIL: margarita.alonso@udc.es (M. Alonso-Ramos); igone.zabala@ehu.eus (I. Zabala)
ORCID: 0000-0002-1353-9270 (M. Alonso-Ramos); 0000-0002-1931-4136 (I. Zabala)


these expressions are ALC which we have in order to express a conclusion in Spanish and Basque.

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 ([3], [4]; 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 [5], 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:

- develop a model of ALCs that includes the characteristics of agglutinative languages such as Basque where different lexicographic and discursive classifications will be established;
- analyse the learners' use of such combinations in Spanish and Basque;
- investigate what kind of help related to the phenomena of lexical combinations they need when writing;
- 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 [4]. 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 [6].

A similar procedure was applied to the Basque novice corpus which was morphologically analysed using Eustagger. We started by extracting an academic vocabulary based on the criteria defined in [7]. We have used this word list to identify collocations, without the need to syntactically analyse the corpus [8]. 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 [6]. 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 [9].

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 [9]. 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.

Table 1
The ALC data from the Spanish novice corpus

ALC	Types/M	Tokens/M
N-modif	192	2724
N *de* N	85	1106
Subject-V	39	313
V-Object	219	2753
Formulas	211	20474

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 have already achieved the compilation of a corpus of novice academic writing [10]. 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.

Table 2
The ALC data from the Basque novice corpus

ALC	Types/M	Tokens/M
N-modif	150	4024
N - N	43	1251
Subject-V	3	58
V-Object	108	4136
Formulas	196	38171

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.

Acknowledgements

This work has been supported by the Xunta de Galicia, through grant ED431C 2020/11, by the CITIC of the UDC through grant ED431G 2019/0 and by 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

[1] K. Hyland, P. Shaw (Eds.) The Routledge Handbook of English for Academic Purposes, Routledge, London, 2016.

[2] K. Tusting, S. McCulloch, I. Bhatt, M. Hamilton, D. Barton, Academics Writing: The Dynamics of Knowledge Creation, Routledge, Abingdon, NY, 2019.

[3] M. Alonso-Ramos, M. García-Salido, M. Garcia, 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, Lexical Computing Brno, 2017, pp. 571–586.

[4] M. García-Salido, M., M. Garcia, M. Villayandre, M. Alonso-Ramos, 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), Miyazaki, 2018, pp. 260-265.

[5] I. Zabala, M.J. Aranzabe, I. Aldezabal, Retos actuales del desarrollo y aprendizaje de los registros académicos orales y escritos del euskera, Círculo de Lingüística Aplicada a la Comunicación 88 (2021) 31–50. doi: 10.5209/clac.78295.

[6] M. García-Salido, M. Garcia, M. Alonso-Ramos, Identifying lexical bundles for an academic writing assistant in Spanish, in: G. Corpas Pastor, R. Mitkov (Eds.), Computational and Corpus-Based Phraseology. Europhras 2019, volume 11755 of Lecture Notes in Computer Sciences, Springer, Cham, 2019, pp.144–158. doi: 10.1007/978-3-030-30135-4_11

[7] M. García-Salido, Compiling an Academic Vocabulary List of Spanish. Available at: doi:10.13140/RG.2.2.27681.33123.

[8] A. Gurrutxaga, I. Alegria, 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, Association for Computational Linguistics, Portland, 2011, pp. 2–7.

[9] M. Garcia, M. García-Salido, M. Alonso-Ramos, Weighted compositional vectors for translating collocations using monolingual corpora, in: G. Corpas Pastor, R. Mitkov (Eds.), Computational and Corpus-Based Phraseology. Europhras 2019, volume 11755 of Lecture Notes in Computer Sciences, Springer, Cham, 2019, pp. 113–128. doi: 10.1007/978-3-030-30135-4_9.

[10] M. J. Aranzabe, A. Gurrutxaga, I. Zabala, Compilación del corpus académico de noveles en euskera HARTAvas y su explotación para el estudio de la fraseología académica. Procesamiento del Lenguaje Natural 69 (2022) 95-103.

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] and Xosé Luis Regueira [1]

[1] *Instituto da Lingua Galega (ILG) - Universidade de Santiago de Compostela, Spain*
[2] *Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) - Universidade de Santiago de Compostela, Spain*

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.

1. Introduction

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.

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
EMAIL: adina.vladu@usc.gal (A.I. Vladu); iria.dedios@usc.gal (I. de-Dios-Flores); mariadelcarmen.magarinos@usc.gal (C. Magariños); john.ortega@usc.gal (J. Ortega); jramom.pichel@usc.gal (J.R. Pichel); marcos.garcia.gonzalez@usc.gal (M. Garcia); pablo.gamallo@usc.gal (P. Gamallo); elisa.fernandez@usc.gal (E. Férnandez Rei); alberto.bugarin.diz@usc.gal (A. Bugarín); manuel.gonzalez.gonzalez@usc.gal (M. González González) senen.barro@usc.gal (S. Barro); xoseluis.regueira@usc.gal (X.L. Regueira)
ORCID: 0000-0002-3910-7820 (A.I. Vladu); 0000-0002-5941-1707 (I. de-Dios-Flores); 0000-0003-3525-1304 (C. Magariños); 0000-0002-2328-3205 (J. Ortega); 0000-0001-5172-6803 (J.R. Pichel); 0000-0002-6557-0210 (M. Garcia); 0000-0002-5819-2469 (P. Gamallo); 0000-0002-4109-0087 (E. Férnandez Rei); 0000-0003-3574-3843 (A. Bugarín); 0000-0001-7948-4607 (M. González González) 0000-0001-6035-540X (S. Barro); 0000-0001-7264-3740 (X.L. Regueira)

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:

An example of numbered list is as following.

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.

More information on the organization of *Proxecto Nós* can be found in de-Dios-Flores et al, 2022.

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

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

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

[3] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv, 2020, pp. 1–19. doi: 10.48550/arXiv.2006.11477

[4] I. de-Dios-Flores, C. Magariños, A. I. Vladu, J. E. Ortega, J. R. Pichel, M. García, P. Gamallo, E. Fernández Rei, A. Bugarín-Diz, M. González González, S. Barro, X. L. Regueira, The Nós Project: Opening routes for the Galician language in the field of language technologies, in: Proceedings of the TDLE Workshop @LREC2022, pp. 52–61 Marseille, 20 June 2022.

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

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

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

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

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

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

[11] J. M. Ramírez Sánchez, C. García Mateo (auth.), M. Giagkou, S. Piperidis, G. Rehm,

J. Dunne (eds.), Report on the Galician Language (Deliverable D1.15), ELE, 2022.
[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, , Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, Y. Wu, Natural TTS Synthesis By Conditioning Wavenet On Mel Spectrogram Predictions, in: Proceedings of ICASSP, 2018.
[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, et al., Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in NLP: System Demonstrations, 2020, pp. 38–45.

CoToHiLi: computational tools for historical linguistics

CoToHiLi: herramientas computacionales para la lingüística histórica

Alina Maria Cristea[1,2], Anca Dinu[1,2], Liviu P. Dinu[1,2], Simona Georgescu[1,2], Ana Sabina Uban[1,2] and Laurențiu Zoicaș[1,2]

[1] *University of Bucharest*

[2] *Research group: Human Language Technologies Research Center, University of Bucharest*

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.

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. [1]). Longobardi (LanGeLin project, 2012-2018) explored the potential correlation of genetic and linguistic distances, starting from what he called Darwin's last challenge: "If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one" (see also [2]). Given that the socio-economical and cultural factors are some of the motivations for borrowing from one language to another [1, 3], 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 [3]. Moreover, establishing the direction and source of borrowing is important to our understanding of the social relations between the groups involved. 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, for instance to automatically identify related words and to assess the semantic change. 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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

alina.cristea@fmi.unibuc.ro (A. M. Cristea); anca.dinu@lls.unibuc.ro (A. Dinu); ldinu@fmi.unibuc.ro (L. P. Dinu); simona.georgescu@lls.unibuc.ro (S. Georgescu); auban@fmi.unibuc.ro (A. S. Uban); laurentiu.zoicas@lls.unibuc.ro (L. Zoicaș)

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

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, in the pre-processing, annotation, feature engineering, training and evaluation phases:

2.1. 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 [4, 5, 6].

2.2. Latin proto-word reconstruction

To improve previous results, we intend to use more recent techniques [7], such 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. Another promising line of research is to make use of the more recent Latin resources, such as The Latin Diachronic Database [8].

2.3. Diachronic semantic divergence

Semantic change is a continuous and complex process ([1] presents no less than 11 types of semantic change), which has been recently studied in the context of distributional semantics theory. Vectorial representations of word meaning (word embeddings) have been used for tracking semantic shifts across different time periods, especially for English. Our aim is 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 word embedding vectorial spaces [9, 10].

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 [3]. Moreover, the semantic data provided by the CoToHiLi project could be of great help for the cognitive sciences and neurosciences [11], to the extent that they can offer a new perspective on our brain mechanisms.

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. [12, 13]). 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 [13, 14], 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. Moreover, we experiment with new ensemble techniques, to improve the overall performance by combining results from multiple sister languages. 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 and FastText [15, 16], as well as experimenting with state-of-the-art language models such as BERT [17]. The method consists of building vectorial semantic representations for the words in each of the target languages, based on the multilingual corpora, and then obtaining a shared multilingual semantic space. This will allow us to compute semantic distances between cognates as well as analyze 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 (such as multilingual Wikipedia[2]).

5. Current Results

For the first two objectives, we have started building datasets of cognates and borrowed words for the Romance languages [18]. 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 [6].

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. 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 corrected 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 [10, 19].

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.

[2] https://github.com/facebookresearch/MUSE

Acknowledgments

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

References

[1] L. Campbell, Historical Linguistics. An Introduction, MIT Press, 1998.

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

[3] P. Epps, Historical linguistics and sociocultural reconstruction, in: The Routledge Handbook of Historical Linguistics, London: Routledge, 2014, pp. 579–597.

[4] A. M. Ciobanu, L. P. Dinu, Automatic detection of cognates using orthographic alignment, in: Proceedings of ACL 2014, Volume 2, 2014, pp. 99–105.

[5] A. M. Ciobanu, L. P. Dinu, Automatic discrimination between cognates and borrowings, in: Proceedings of ACL 2015, 2015, pp. 431–437.

[6] A. M. Cristea, L. P. Dinu, S. Georgescu, M. Mihai, A. S. Uban, Automatic discrimination between inherited and borrowed latin words in romance languages, in: Findings of EMNLP 2021, 2021, pp. 2845–2855.

[7] A. M. Ciobanu, L. P. Dinu, Ab initio: Automatic Latin proto-word reconstruction, in: Proceedings of COLING 2018, 2018, pp. 1604–1614.

[8] T. Spinelli, The latin diachronic database: a new digital tool for the study of latin, in: Recent Advances in Digital Humanities: Romance Language Applications, Peter Lang, 2022, forthcoming.

[9] A.-S. Uban, A. M. Ciobanu, L. P. Dinu, Cross-lingual laws of semantic change, Computational approaches to semantic change 6 (2021) 219.

[10] A. S. Uban, A. Cristea, A. Dinu, L. P. Dinu, S. Georgescu, L. Zoicas, Tracking semantic change in cognate sets for English and Romance languages, in: Proceedings of LChange 2021, 2021, pp. 64–74.

[11] D. Poeppel, D. Embick, Defining the relation between linguistics and neuroscience, in: The Routledge Handbook of Historical Linguistics. Twenty-First Century Psycholinguistics, Four Cornerstones, New York, Routledge, 2017, pp. 103–118.

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

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

[14] G. Kondrak, A new algorithm for the alignment of phonetic sequences, in: Proceedings of ANLP 2000, 2000, pp. 288–295.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of NIPS 2013, 2013, pp. 3111—-3119.

[16] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, TACL 5 (2016) 135–146.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL 2019, 2019, pp. 4171–4186.

[18] A. M. Cristea, A. Dinu, L. P. Dinu, S. Georgescu, A. S. Uban, L. Zoicas, Towards an Etymological Map of Romanian, in: Proceedings of RANLP 2021, 2021, pp. 315–323.

[19] A. S. Uban, L. P. Dinu, Automatically building a multilingual lexicon of false friends with no supervision, in: Proceedings of LREC 2020, 2020, pp. 3001–3007.

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

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

Marcos Garcia[1], Pablo Gamallo[1], Martín Pereira-Fariña[2] and Iria de-Dios-Flores[1]

[1]*Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela*
[2]*Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela*

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 are working are Galician-Portuguese (in its Galician and Portuguese varieties, fundamentally), Spanish and English.

Keywords

lexical semantics, distributional semantics, language models.

1. Introduction and objectives

The use of architectures based on artificial neural networks has become the most dominant approach to natural language processing (NLP) in recent years [1], producing significantly better results in numerous areas than supervised models designed by selecting individual features of the target tasks [2]. This paradigm shift has promoted the popularization of vector models inspired by the distributional hypothesis [3, 4], which until then were mainly used in research in cognitive science and computational linguistics [5, 6, 7]. In this field, the implementation of computationally more efficient architectures, with drastic reductions in dimensionality [8], has sparked great interest in distributional semantics studies, boosted also by the findings about the various linguistic regularities encoded by these models [9]. This area, previously dominated by linguistically informed and more interpretable methodologies (e.g., using vectors built through syntactic dependencies [10]), has become one of the most productive in NLP research [11].

In this regard, the emergence of deep learning techniques using multilayer deep neural networks with millions of hyperparameters (which require large computational infrastructures) has led to the proliferation of language models that perform NLP tasks more accurately. Among various others, we can highlight the public models ELMo (Embeddings from Language Models [12]), or the different variants of BERT (Bidirectional Encoder Representations from Transformers [13]).

The project presented in this paper fits into this new line of research and focuses on the analysis of the ability of these models to solve various types of lexical ambiguity:[1]

1. Polysemy and homonymy, i.e., a single orthographic form that has different meanings (or senses) depending on the context. For example, *school* as a building, as an organization, or as a group of people (polysemy), or *bank* as a financial institution, or as a sloping raised land (homonymy).
2. Synonymy, i.e., different words expressing the same meaning in certain contexts (e.g., *coach* or *bus* to refer to a long motor vehicle).
3. Idiomaticity, i.e., multiword expressions (MWEs) whose meaning does not correspond to the one of its constituent elements (e.g., *glass ceiling* as a social barrier for women).

[1]We broadly follow [14] for the definition of the phenomena mentioned here.

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

marcos.garcia.gonzalez@usc.gal (M. Garcia); pablo.gamallo@usc.gal (P. Gamallo); martin.pereira@usc.gal (M. Pereira-Fariña); iria.dedios@usc.gal (I. de-Dios-Flores)
0000-0002-6557-0210 (M. Garcia); 0000-0002-5819-2469 (P. Gamallo); 0000-0002-1982-2472 (M. Pereira-Fariña); 0000-0002-5941-1707 (I. de-Dios-Flores)

Taking the above into account, our research aims to fill a particularly important gap in the evaluation of these computational models by investigating the presence of various types of knowledge related to lexical semantics in several languages. Thus, the main goal of the project is to explore the most recent language models concerning the representation of polysemy and homonymy, synonymy and semantic compositionality, as well as to compare them with more interpretable distributional and compositional methods.

The results of the present project will be useful, on the one hand, to advance the understanding of semantic information encoded both in static distributional representations and in large language models trained with deep neural networks. In addition, and although the project is mainly focused on the exploration of models, both the datasets and the results of manual annotation will be an important contribution regarding the semantic interpretation of polysemy and homonymy, synonymy and idiomaticity by native speakers of various languages.

2. Methodology and work plan

To develop this project, we will use the following methodology and instrumental techniques, which in general correspond to the state-of-the-art research in NLP and computational linguistics.

Regarding the experimental design and the data collection, we will use standard methodologies from studies in semantics [14] and in psycholinguistics [15, 16], aimed at generating controlled stimuli. Likewise, to collect annotations from human informants, we will use crowdsourcing methods which will allow us to obtain data from native speakers quickly and efficiently, with quality control of the annotations [17].

Regarding the computational models, those based on Transformer architectures will be implemented using the *transformers* library, which includes the latest models based on deep learning. We will eventually use other open source libraries that may incorporate additional models. To train and run static embeddings, we will use *gensim*[2] and the official tools released by the authors of other distributional methods based on interpretable syntactic dependencies (e.g., [18]).

Finally, to compare the representations of the computational models with the values obtained from the human annotations, we will use three methods:

1. Precision scores, in evaluations with discrete values (e.g. homonymy or synonymy, and in the results of linear classifiers).
2. Correlation values, in graded evaluations (polysemy or idiomaticity).
3. Representation Similarity Analysis, to see if the models predict relative differences between examples of the same type (e.g., a word or MWE with the same meaning in different contexts) in a similar way to humans.

It should be noted that these methods have already been used in previous works, which we briefly mention below.

2.1. First results

Although we are at an early stage, we already have some published results, both from previous research directly related to this proposal and from work carried out since the beginning of the project. Thus, we have already presented various datasets with semantic idiomaticity annotation at token and type levels in English and Portuguese, and used them to evaluate several language models [19, 20]. In addition, we have created a new dataset in Galician-Portuguese, English and Spanish that includes examples of homonymy and synonymy in context, also used to compare various contextualization models and strategies [21].

More recently, we have compared Transformers models and distributional strategies based on syntactic dependencies in semantic compositionality tasks [18, 22]. Finally, we have participated in the co-organization of the task *Multilingual Idiomaticity Detection and Sentence Embedding* (SemEval 2022), in which we have presented new resources with annotation of semantic idiomaticity in context in Galician-Portuguese and English [23].

3. Work team

The project presented in this paper is carried out at the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) of the Universidade de Santiago de Compostela, and belongs to its scientific program in Natural Language Technologies. In this sense, members of the center collaborate on different tasks of our work plan, that are part of their respective areas of expertise.

Besides the principal investigator, the project has research and work teams formed by three PhDs with specializations in Computational Linguistics, Psycholinguistics, Logic and Computer Science. In collaboration with a pre-doctoral researcher and

[2] https://radimrehurek.com/gensim/

technical staff that will be hired with the project funds, these teams actively participate in the different stages of the project. Finally, we also rely on the collaboration of researchers from other universities, both Galician and international, with whom we have already participated in joint initiatives and projects with similar themes to the one presented in this paper.

Acknowledgments

Project funded by the Galician Government (*Consolidación e estruturación de unidades de investigación competitivas e outras accións de fomento: Proxectos de Excelencia*, ED431F 2021/01) and by a *Ramón y Cajal* grant (RYC2019-028473-I).

References

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research 12 (2011) 2493–2537.

[2] T. Schnabel, I. Labutov, D. Mimno, T. Joachims, Evaluation methods for unsupervised word embeddings, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 298–307. URL: https://aclanthology.org/D15-1036. doi:10.18653/v1/D15-1036.

[3] Z. S. Harris, Distributional structure, Word 10 (1954) 146–162.

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

[5] G. A. Miller, Empirical methods in the study of semantics, in: D. D. Steinberg, L. A. Jakobovits (Eds.), Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology, 1971, pp. 569–585.

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

[7] J. Mitchell, M. Lapata, Composition in distributional models of semantics, Cognitive science 34 (2010) 1388–1429.

[8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Workshop Proceedings of the International Conference on Learning Representations, 2013.

[9] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 746–751. URL: https://aclanthology.org/N13-1090.

[10] S. Padó, M. Lapata, Dependency-based construction of semantic space models, Computational Linguistics 33 (2007) 161–199. URL: https://aclanthology.org/J07-2002. doi:10.1162/coli.2007.33.2.161.

[11] G. Boleda, Distributional semantics and linguistic theory, Annual Review of Linguistics 6 (2020) 213–234.

[12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: https://aclanthology.org/N18-1202. doi:10.18653/v1/N18-1202.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[14] D. A. Cruse, Lexical semantics, Cambridge University Press, 1986.

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

[16] R. Richie, B. White, S. Bhatia, M. C. Hout, The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures, Behavior Research Methods 52 (2020) 1906–1928.

[17] R. Munro, S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, H. Tily, Crowdsourcing and language studies: the new generation of linguistic data, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Los Angeles, 2010, pp. 122–130. URL: https://aclanthology.org/W10-0719.

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

[19] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, A. Villavicencio, Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP), ACL, 2021, pp. 2730–2741. URL: https://aclanthology.org/2021.acl-long.212. doi:10.18653/v1/2021.acl-long.212.

[20] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, A. Villavicencio, Probing for idiomaticity in vector space models, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3551–3564. URL: https://aclanthology.org/2021.eacl-main.310. doi:10.18653/v1/2021.eacl-main.310.

[21] M. Garcia, Exploring the representation of word meanings in context: A case study on homonymy and synonymy, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3625–3640. URL: https://aclanthology.org/2021.acl-long.281. doi:10.18653/v1/2021.acl-long.281.

[22] P. Gamallo, M. Garcia, I. de-Dios-Flores, Evaluating Contextualized Vectors from Large Language Models and Compositional Strategies, Procesamiento del Lenguaje Natural 69 (2022).

[23] H. Tayyar Madabushi, E. Gow-Smith, M. Garcia, C. Scarton, M. Idiart, A. Villavicencio, SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 107–121. URL: https://aclanthology.org/2022.semeval-1.13.

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[1], Flor Miriam Plaza-del-Arco[1], Jaime Collado-Montañez[1], L. Alfonso Ureña-López[1] and M. Teresa Martín-Valdivia[1]

[1]*Departamento de Informática, CEATIC, Universidad de Jaén, España*

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

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 [1, 2]. The most recent and best-performing studies offer solutions based on neural networks for the detection of the different phenomena including misogyny and xenophobia [3], sexism [4], cyberbullying [5], aggression [6], or offensive language [7, 8]. Some researches have shown that sentiment and emotion analysis are important features to consider in the detection of these phenomena [9, 10, 11]. 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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

jmurgado@ujaen.es (J. A. M. Murgado); fmplaza@ujaen.es (F. M. Plaza-del-Arco); jcollado@ujaen.es (J. Collado-Montañez); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. T. Martín-Valdivia)

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

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

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 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.

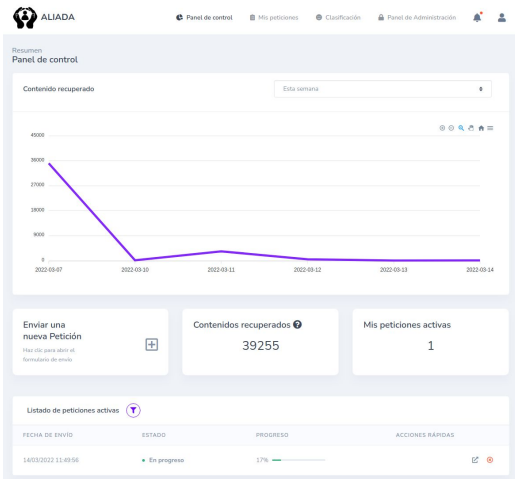


Figure 1: Dashboard.

[3]https://apexcharts.com/

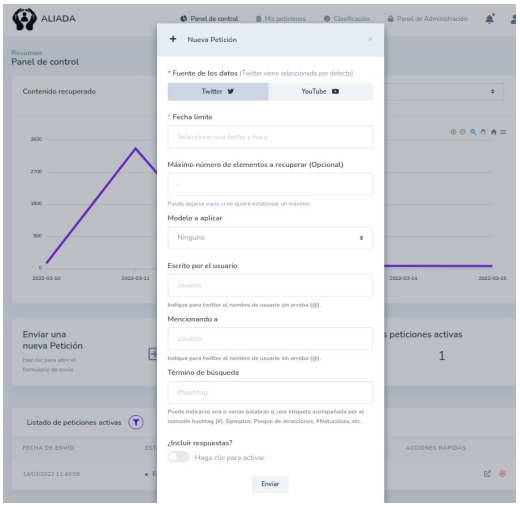


Figure 2: New request form.

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 administration 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

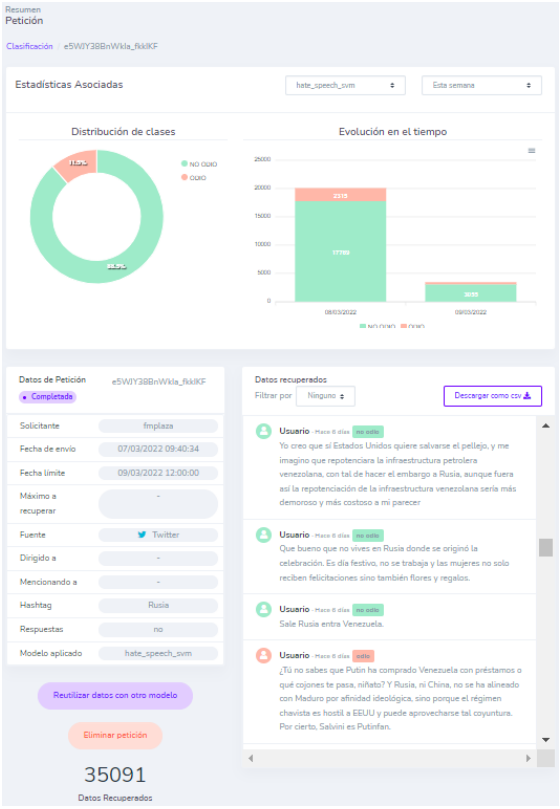


Figure 3: Classification results.

HatEval [12], HaterNet [13], EXIST [14], NewsCom-TOX [15] and OffendES [7]. 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).

5. Acknowledgments

Acknowledgments

This work has been partially supported by Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) 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

[1] E. Fersini, P. Rosso, M. Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018 (2018) 15.

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

[3] F.-M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña López, M. T. Martín-Valdivia, Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies, ACM Trans. Internet Technol. 20 (2020). URL: https://doi.org/10.1145/3369869. doi:10.1145/3369869.

[4] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. U. López, M. T. Martín-Valdivia, 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*, CEUR-WS.org, 2021, pp. 491–499.

[5] F. Elsafoury, S. Katsigiannis, S. R. Wilson, N. Ramzan, Does BERT Pay Attention to Cyberbullying?, Association for Computing Machinery, New York, NY, USA, 2021, p. 1900–1904. URL: https://doi.org/10.1145/3404835.3463029.

[6] R. Kumar, A. K. Ojha, S. Malmasi,

M. Zampieri, Evaluating Aggression Identification in Social Media, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 1–5. URL: https://aclanthology.org/2020.trac-1.1.

[7] F. M. Plaza-del-Arco, A. Montejo-Ráez, L. A. Ureña-López, M.-T. Martín-Valdivia, 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), INCOMA Ltd., Held Online, 2021, pp. 1096–1108. URL: https://aclanthology.org/2021.ranlp-1.123.

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

[9] S. Rajamanickam, P. Mishra, H. Yannakoudakis, E. Shutova, Joint Modelling of Emotion and Abusive Language Detection, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4270–4279. URL: https://aclanthology.org/2020.acl-main.394. doi:10.18653/v1/2020.acl-main.394.

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

[11] F. M. Plaza-del-Arco, S. Halat, S. Padó, R. Klinger, Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language, CoRR abs/2109.10255 (2021). URL: https://arxiv.org/abs/2109.10255. arXiv:2109.10255.

[12] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. Rangel, P. Rosso, M. Sanguinetti, 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, 2019.

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

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

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

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[1] and Fernando Martínez-Santiago[1]

[1] *Universidad de Jaén, Campus Las Lagunillas s/n, 23071 Jaén, España*

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.

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].

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[1], 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[2]. 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[3] 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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

igmunoz@ujaen.es (I. Garrido-Muñoz); amontejo@ujaen.es (A. Montejo-Ráez); dofer@ujaen.es (F. Martínez-Santiago)

https://ismael.codes/ (I. Garrido-Muñoz); https://www.ujaen.es/centros/ceatic/ (A. Montejo-Ráez); https://www.ujaen.es/centros/ceatic/ (F. Martínez-Santiago)

0000-0001-6656-9679 (I. Garrido-Muñoz); 0000-0002-8643-2714 (A. Montejo-Ráez); 0000-0002-1480-1752 (F. Martínez-Santiago)

[1] https://ismael.codes/categoryviewer/

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[4] 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. [5] 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 et al. [6] 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[7] or RoBERTa[8].

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. [9] will appear under the name **Yulia** on the tool, and the work by Wiggins [10] 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.

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

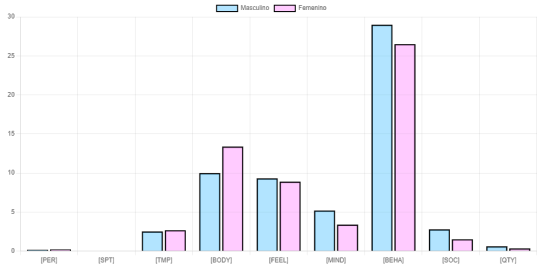


Figure 1: Category viewer

value of each column being red for female and blue for male.

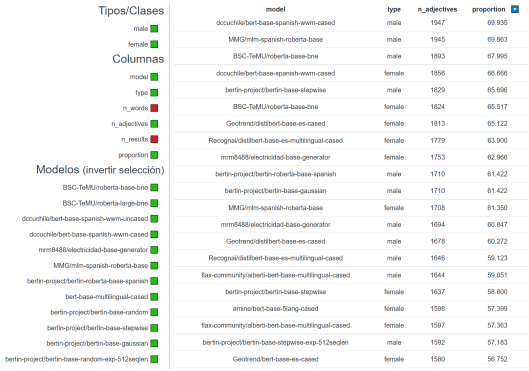
This will allow us to see at a glance whether the leaning in that category is towards male or female, or neither in particular. In addition we will be able to see which models have a higher level of bias given the color intensity. By default we have the RSV and Probability columns that show the external 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.

	% 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.

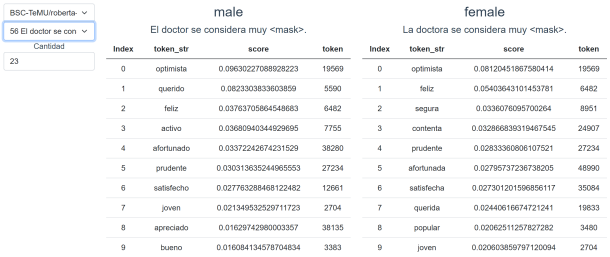


model	type	n_adjectives	proportion
dccuchile/bert-base-spanish-wwm-cased	male	1947	69.935
MMG/mlm-spanish-roberta-base	male	1945	69.863
BSC-TeMU/roberta-base-bne	male	1893	67.995
dccuchile/bert-base-spanish-wwm-cased	female	1856	66.666
bertin-project/bertin-base-stepwise	male	1829	65.696
BSC-TeMU/roberta-base-bne	female	1824	65.517
Geotrend/distilbert-base-es-cased	female	1813	65.122
Recognai/distilbert-base-es-multilingual-cased	female	1779	63.900
mrm8488/electricidad-base-generator	female	1753	62.966
bertin-project/bertin-roberta-base-spanish	male	1710	61.422
bertin-project/bertin-base-gaussian	male	1710	61.422
MMG/mlm-spanish-roberta-base	female	1708	61.350
mrm8488/electricidad-base-generator	male	1694	60.847
Geotrend/distilbert-base-es-cased	male	1678	60.272
Recognai/distilbert-base-es-multilingual-cased	male	1646	59.123
flax-community/alberti-bert-base-multilingual-cased	male	1644	59.051
bertin-project/bertin-base-stepwise	female	1637	58.800
amine/bert-base-5lang-cased	female	1598	57.399
flax-community/alberti-bert-base-multilingual-cased	female	1597	57.363
bertin-project/bertin-base-stepwise-exp-512seqlen	male	1592	57.183
Geotrend/bert-base-es-cased	female	1580	56.752

Figure 3: Adjective Stats snapshot

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.



index	token_str	score	token
0	optimista	0.09630227088928223	19569
1	querido	0.08233038333603859	5590
2	feliz	0.037637058645486 83	6482
3	activo	0.03680940344929695	7755
4	afortunado	0.03372242674231529	38280
5	prudente	0.03031363524496555 3	27234
6	satisfecho	0.02776328846812248 2	12661
7	joven	0.02134953252971172 3	2704
8	apreciado	0.01629742980003357	38135
9	bueno	0.01608413457870483 4	3383

index	token_str	score	token
0	optimista	0.08120451867580414	19569
1	feliz	0.05403643101453781	6482
2	segura	0.0336076095700264	8951
3	contenta	0.03286683931946754 5	24907
4	prudente	0.02833360806107521	27234
5	afortunada	0.02795737236738205	48990
6	satisfecha	0.02730120159685611 7	35084
7	querida	0.02440616674721241	19833
8	popular	0.02062511257827282	3480
9	joven	0.02060385979712009 4	2704

Figure 4: Explorer snapshot

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 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 LIVING-LANG by the Spanish Government.

References

[1] J. L. Julia Angwin, Machine bias - there's software used across the country to predict future criminals. and it's biased against blacks., 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Z. O. U. Berkeley, Z. Obermeyer, U. Berkeley, S. M. U. o. Chicago, S. Mullainathan, U. o. Chicago, O. M. A. Metrics, Dissecting racial bias in an algorithm that guides health decisions for 70 million people: Proceedings of the conference on fairness, accountability, and transparency, 2019. URL: https://dl.acm.org/doi/10.1145/3287560.3287593.

[3] D. Harwell, A face-scanning algorithm increasingly decides whether you deserve the job, 2019. URL: https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/.

[4] J. Dastin, Amazon scraps secret ai recruiting tool that showed bias against women, 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[5] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, A. Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, CoRR abs/1607.06520 (2016). URL: http://arxiv.org/abs/1607.06520. arXiv:1607.06520.

[6] A. Caliskan, J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.

[7] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, 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, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[8] S. Sharma, M. Dey, K. Sinha, Evaluating gender bias in natural language inference, CoRR abs/2105.05541 (2021). URL: https://arxiv.org/abs/2105.05541. arXiv:2105.05541.

[9] Y. Tsvetkov, N. Schneider, D. Hovy, A. Bhatia, M. Faruqui, C. Dyer, Augmenting English Adjective Senses with Supersenses, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 4359–4365. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1096_Paper.pdf.

[10] J. S. Wiggins, A psychological taxonomy of trait-descriptive terms: The interpersonal domain. 37 (1979) 395–412. URL: https://doi.org/10.1037/0022-3514.37.3.395. doi:10.1037/0022-3514.37.3.395.

COBATO. Un chatbot orientado a asistir al pequeño comercio

COBATO. A chatbot aimed at assisting small retailers

Clara Díaz-Ruíz[1], Fernando Martínez-Santiago[1], Arturo Montejo-Ráez[1], Maria Teresa Martín-Valdivia[1], L. Alfonso Ureña-López[1], Manuel Carlos Diaz-Galiano[1], Miguel Angel Garcia-Cumbreras[1], Manuel García-Vega[1], Flor Miriam Plaza-del-Arco[1], Salud María Jiménez-Zafra[1] and María Dolores Molina-González[1]

[1]*Grupo SINAI, Departamento de Informática, Universidad de Jaén, Universidad de Jaén, Campus Las Lagunillas, 23071, Jaen, Spain*

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.

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

asistente virtual, chatbot, GF, modelos del lenguaje, PLN

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*)([R]. Concretamente, los GF proveen de un analizador semántico especializado al dominio de la aplicación. Sin

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

cdr00008@red.ujaen.es (C. Díaz-Ruíz); dofer@.ujaen.es (F. Martínez-Santiago); amontejo@.ujaen.es (A. Montejo-Ráez); maite@.ujaen.es (M. T. Martín-Valdivia); laurena@.ujaen.es (L. A. Ureña-López); mcdiaz@.ujaen.es (M. C. Diaz-Galiano); magc@.ujaen.es (M. A. Garcia-Cumbreras); mgarcia@ujaen.es (M. García-Vega); fmplaza@ujaen.es (F. M. Plaza-del-Arco); sjzafra@ujaen.es (S. M. Jiménez-Zafra); mmolina@ujaen.es (M. D. Molina-González)

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, *KnowledgeGraphs*)([H], 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 ([F], la energía ([F] y ([C], 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.

- 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.

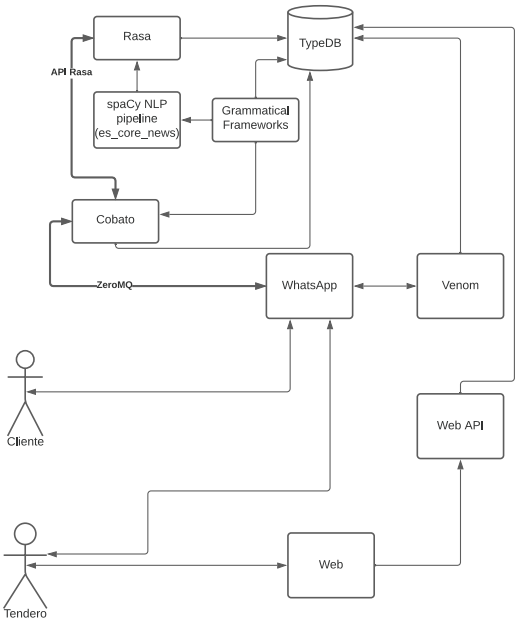


Figura 1: Arquitectura general de COBATO

- 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.

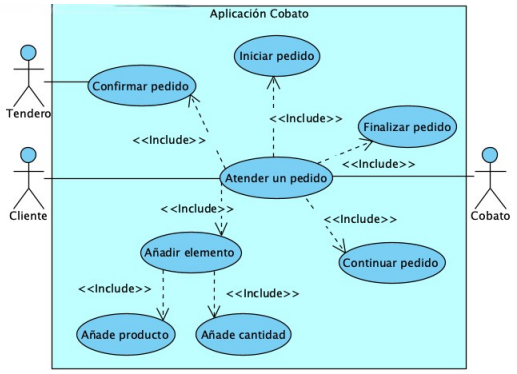


Figura 2: Caso de uso: atender un pedido

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".

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 ", donde "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

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

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

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

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

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

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[1], Víctor Fresno[1] and Enrique Amigó[1]

[1] *Universidad Nacional de Educación a Distancia (UNED), España*

Abstract

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.

English translation. 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

Composición semántica, Vectores de frases, Transformers,

1. Introducción

La llegada del modelo Transformer [1] 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, al permitir trabajar con representación a diferentes capas internas de la red.

2. Motivación

Se han desarrollado numerosos marcos de evaluación, nacidos de la necesidad de cuantificar el éxito de los modelos, sin embargo, todos estos marcos están orientados a asignar un valor de rendimiento o precisión en un marco de referencia arbitrario. El

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

aghajari@lsi.uned.es (A. Ghajari); vfresno@lsi.uned.es (V. Fresno); enrique@lsi.uned.es (E. Amigó)

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

problema que presentan es que esta información no nos permite comprender el funcionamiento del modelo a evaluar o cómo consigue capturar el lenguaje para el que se ha entrenado.

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 [2, 3, 4], 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 pre-entrenados 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 [5]) 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/

Figure 1: Pantalla principal.

4.2. Cliente

Contiene la interfaz de usuario y los métodos gRPC de comunicación con el servidor. Se ocupa del preprocesamiento 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 texto 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. En

[7] https://github.com/Pylons/waitress

[8] https://www.nltk.org/

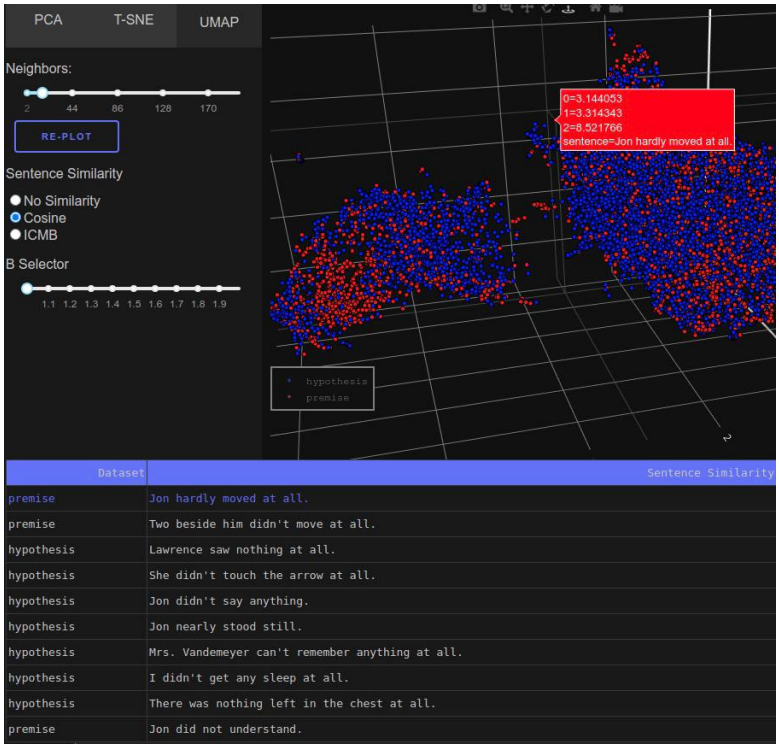


Figure 2: Pestaña de inferencia.

ella se ofrecen distintos métodos de reducción dimensional, t-SNE [6], Principal Component Analysis [7] y UMAP [8], seleccionables como subpestañas, que contienen los elementos de ajuste de parámetros de cada uno. Desde la misma es también posible guardar los vectores resultado del proceso de composición 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. Tras 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, ver Figura3), 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 dimensionalidad y similitud semántica, además de técnicas de optimización de inferencia.

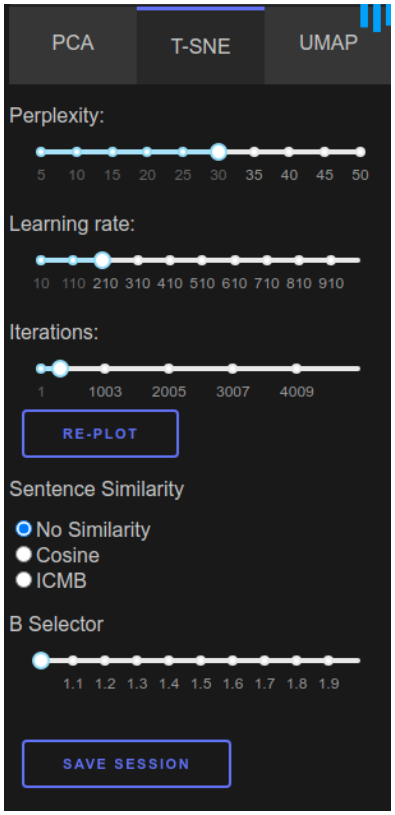


Figure 3: Detalle de modificación de parámentros t-SNE.

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 [2, 9]. Finalmente, algunos estudios apuntan a la posibilidad de que distintas cabezas de auto-atención del modelo Transformer atienda a distintos aspectos semánticos, [10]; aislarlos y generar su representación podría ofrecer una nueva perspectiva.

Acknowledgments

Este trabajo ha sido financiado por los proyectos 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 MISMIS project (PGC2018-096212-B), así como por el proyecto PID2020 GID2016-39 de Innovación Docente de la Universidad Nacional de Educación a Distancia y los proyectos del Consejo Europeo de Investigación (ERC) bajo el programa de investigación e innovación H2020-INFRAIA-2020-1: LyrAIcs (con Grant agreement Nº [964009]) y CLS-INFRA: Computational Literary Studies Infrastructure (con Grant agreement Nº [101004984]).

References

[1] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, Technical Report, 2017.

[2] K. Ethayarajh, How contextual are contextualized word representations? Comparing the geometry of BERT, ELMO, and GPT-2 embeddings, EMNLP-IJCNLP 2019 (2020) 55–65. doi:10.18653/v1/d19-1006. arXiv:1909.00512.

[3] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, L. Li, On the sentence embeddings from pre-trained language models, arXiv (2020). doi:10.18653/v1/2020.emnlp-main.733. arXiv:2011.05864.

[4] J. Gao, D. He, X. Tan, T. Qin, L. Wang, Y. Liu, Representation Degeneration Problem in Training Natural Language Generation Models, Technical Report, 2019. arXiv:1907.12009v1.

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

[6] L. Van Der Maaten, G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.

[7] H. Hotelling, Analysis of a complex of statistical variables into principal components., J. of Educational Psychology 24 (1933) 498–520.

[8] L. Mcinnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2020). arXiv:1802.03426v3.

[9] E. Amigó, F. Giner, J. Gonzalo, M. Verdejo, On the foundations of similarity in information access., Information Retrieval 23, Issue 3 (2020) 216–254.

[10] A. Rogers, O. Kovaleva, A. Rumshisky, A Primer in BERTology: What We Know About How BERT Works, Technical Report, 2020. URL: https://github.com/. arXiv:2002.12327v3.

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
Departamento de Informática, Universidad de Valladolid, España.

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.

English translation. 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

Palabras clave
Videojuegos educativos, cambio climático, generación de lenguaje natural.

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 [1]. 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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
EMAIL: descuder@infor.uva.es (D. Escudero-Mancebo); yania@infor.uva.es (Y.Crespo)
ORCID: 0000-0003-0849-8803 (D. Escudero-Mancebo); 0000-0003-0639-0540 (Y.Crespo)

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 [2,3]. 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 preelaborados, dependientes del contexto a los que se accede en función de una casuística que depende del juego [4]. 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 preestablecido, 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*.

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.

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.

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

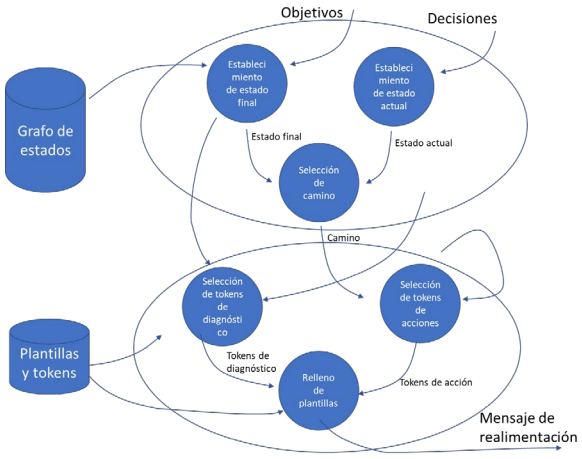


Figura 2: Diagrama funcional que muestra la operativa del módulo de generación de len guaje natural.

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 [6], 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 [5].

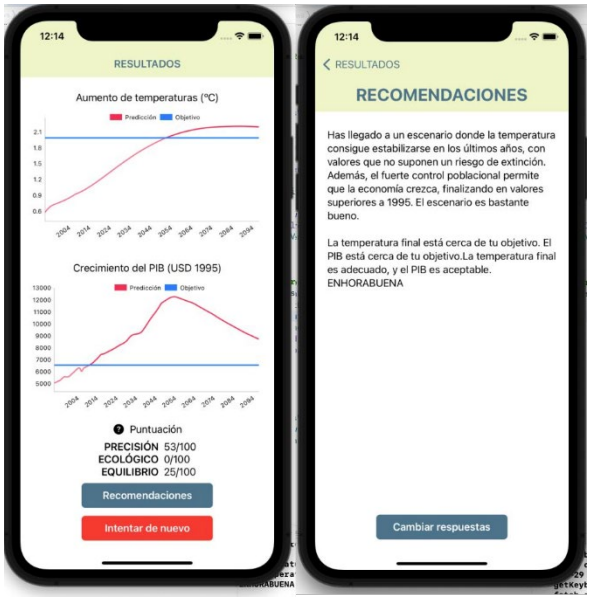


Figura 3: Apartado de realimentación en la interfaz iOS del juego.

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 módulo 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, la capacidad de concienciar sobre el cambio climático y la eficiencia de la aplicación al compararla con la versión manual del juego [1]. 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 las 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 Enríquez y Luis Javier de Miguel por la búsqueda de financiación. correcta integración del módulo en el juego.

Referencias

[1] 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.

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

[3] 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.

[4] 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.

[5] Manzano-Santos, A., D. Escudero-Mancebo, y J. M. Miguel-Gonz´alez. en revisi´on. A multivariate time series clustering algorithm for the analysis of the cross relation between the constituent univariate time series patterns. Pattern Recognition (under revision)

[6] 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.

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] and Alberto Bugarín-Diz[1]

[1]*Grupo de Sistemas Intelixentes, Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez s/n, Campus Vida 15782, Santiago de Compostela, España*

[2]*Rede de Calidade do Aire de Galicia, MeteoGalicia, Xunta de Galicia, Calle Roma 6 15707 Fontiñas, Santiago de Compostela, España*

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 [1].

English translation. 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 [1].

Keywords

términos lingüísticos borrosos, sistemas data-to-text, generación de lenguaje natural.

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) [2] 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 [1]. 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 [3], esta variable tiene seis etiquetas con una percepción positiva, neutra o negativa (Tabla 1).

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 in-

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

andrea.cascallar.fuentes@usc.es (A. Cascallar-Fuentes); javier.gallego.fernandez@rai.usc.es (J. Gallego-Fernández); alejandro.ramos@inverbisanalytics.com (A. Ramos-Soto); calidadedoaire.cma@xunta.gal (A. Saunders-Estévez); alberto.bugarin.diz@usc.es (A. Bugarín-Diz)

0000-0003-1857-5796 (A. Cascallar-Fuentes); 0000-0001-6136-8413 (A. Ramos-Soto); 0000-0003-3574-3843 (A. Bugarín-Diz)

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

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

formación 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.

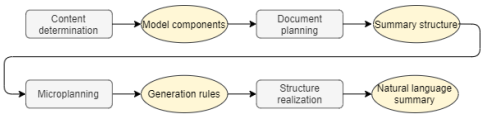


Figura 1: Representación de la arquitectura de nuestra propuesta. Los rectángulos representan las etapas, mientras que las elipses representan los resultados.

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 [4] y SimpleNLG-GL [5].

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 [6].

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 dispondrán 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.

4. Validación por expertos

Hemos pedido a dos meteorólogos expertos de la Red de Calidad del Aire de MeteoGalicia [1] 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.

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

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

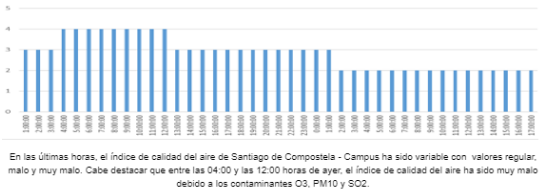


Figura 2: Ejemplo del cuestionario de evolución del ICA diseñado para la validación de expertos.

Tabla 3
Resultado de la evaluación realizada por expertos.

	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

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 mayor 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).

Referencias

[1] MeteoGalicia, MeteoGalicia website, 2021. URL: www.meteogalicia.gal, [Accessed February 2021].

[2] E. Reiter, An architecture for data-to-text systems, in: Proceedings of the Eleventh European Workshop on Natural Language Generation, Association for Computational Linguistics, 2007, pp. 97–104. URL: https://doi.org/10.3115%2F1610163.1610180. doi:10.3115/1610163.1610180.

[3] European Environment Agency, European Air Quality Index website, 2021. URL: www.eea.europa.eu, [Accessed February 2021].

[4] A. Ramos-Soto, J. J. Gallardo, A. Bugarín, Adapting SimpleNLG to Spanish, in: Proceedings of the 10th International Conference on Natural Language Generation, INLG, Association for Computational Linguistics, 2017, pp. 144–148. URL: https://doi.org/10.18653/v1/w17-3521. doi:10.18653/v1/w17-3521.

[5] A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, Adapting SimpleNLG to Galician language, in: Proceedings of the 11th International Conference on Natural Language Generation, Association for Computational Linguistics, 2018, pp. 67–72. URL: https://doi.org/10.18653/v1/w18-6507. doi:10.18653/v1/w18-6507.

[6] H. P. Grice, Logic and conversation, in: Speech acts, Brill, 1975, pp. 41–58.

NLP4SM: Natural Language Processing for social media

Gonzalo Medina Medina[1], Jose Camacho Collados[2] and 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

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.

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 [1], 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 [2]. The transformers models stand out as contextual word embeddings, with BERT [3] 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 pretrained 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 [4] 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 network, as for instance to Twitter [5]. 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 [6] or TweetEval [7].

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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

gmedina95@correo.ugr.es (G. Medina Medina); camachocolladosj@cardiff.ac.uk (J. Collados); emcamara@decsai.ugr.es (E. Martínez Cámara)

0000-0003-1618-7239 (J. Collados); 0000-0002-5279-8355 (E. Martínez Cámara)

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

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

[3]Production [8]: https://tweetnlp.org/demo/

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.

2.1. 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" [9] 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 [10] 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 [11].[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 [12].[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 [13].[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 [14].[9]

Stance classification It classifies the author stance according to a topic. The corpus of the task Detectin Stance from SemEval16 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 [15] 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 [16].

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 [17] pre-trained on English text from social networks [7].

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 [17].

[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

[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

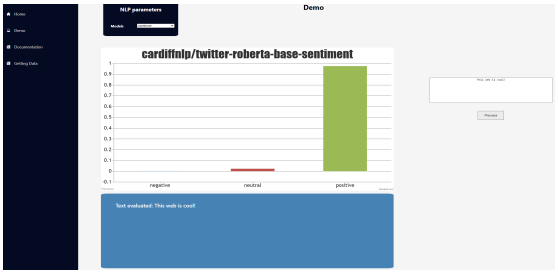


Figure 1: Sentiment analysis, 'text mode' mode.

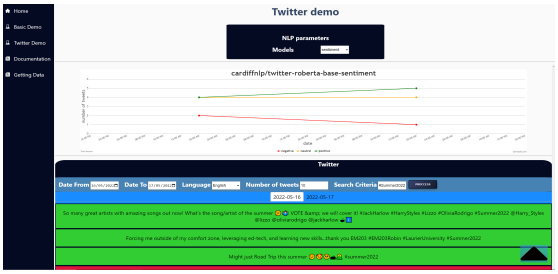


Figure 2: Sentiment analysis, 'Twitter mode'.

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 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.

Acknowledgments

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

References

[1] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proc. of Workshop at ICLR, 2013.

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

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer,

P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[5] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Sentiment analysis in twitter, Natural Language Engineering 20 (2014) 1–28. doi:10.1017/S1351324912000332.

[6] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[7] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the ACL: EMNLP 2020, 2020. URL: https://aclanthology.org/2020.findings-emnlp.148. doi:10.18653/v1/2020.findings-emnlp.148.

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

[9] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, SemEval-2018 task 1: Affect in tweets, in: Proc. of The 12th Int. Workshop on Semantic Evaluation, 2018, pp. 1–17. URL: https://aclanthology.org/S18-1001. doi:10.18653/v1/S18-1001.

[10] S. Rosenthal, N. Farra, P. Nakov, SemEval-2017 task 4: Sentiment analysis in Twitter, in: Proc. of the 11th Int. Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 502–518. URL: https://aclanthology.org/S17-2088. doi:10.18653/v1/S17-2088.

[11] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[12] C. Van Hee, E. Lefever, V. Hoste, SemEval-2018 Task 3: Irony detection in English tweets, in: Proc. of The 12th Int. Workshop on Semantic Evaluation, 2018. URL: https://aclanthology.org/S18-1005. doi:10.18653/v1/S18-1005.

[13] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 75–86. URL: https://aclanthology.org/S19-2010. doi:10.18653/v1/S19-2010.

[14] F. Barbieri, J. Camacho-Collados, F. Ronzano, L. Espinosa-Anke, M. Ballesteros, V. Basile, V. Patti, H. Saggion, SemEval 2018 task 2: Multilingual emoji prediction, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 24–33. URL: https://aclanthology.org/S18-1003. doi:10.18653/v1/S18-1003.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[16] F. Barbieri, L. Espinosa-Anke, J. Camacho-Collados, A Multilingual Language Model Toolkit for Twitter, in: arXiv preprint arXiv:2104.12250, 2021.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

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, C/ Juan del Rosal 16, UNED, 28040 Madrid, Spain

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), 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.

English translation. 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) 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.

Palabras Clave

Transcripción de textos, modelos del lenguaje, recursos lingüísticos.

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 [10]. 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,

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

EMAIL: amenta@invi.uned.es (A. Menta-Garuz); evasan@lsi.uned.es (E. Sanchez-Salido); agarcia@lsi.uned.es (A. Garcia-Serrano)

ORCID: 0000-0002-3172-2829 (A. Menta-Garuz); 0000-0001-8665-3018 (E. Sanchez-Salido); 0000-0003-0975-7205 (A. Garcia-Serrano)

conocido como OCR (del inglés *Optical Character Recognition*).

La elaboración de corpus históricos está sujeta a múltiples factores, entre ellos su finalidad [9]. Por ejemplo, para el estudio de una lengua actual en general se pretende que el corpus sea proporcional, es decir, 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), elementos estructurales (capítulos, páginas) y elementos lingüísticos (entidades, cambios de registro).

La comunidad científica concienciada de la dificultad de tratar documentos históricos, en los últimos años está realizando un esfuerzo en mejorar las herramientas disponibles para su gestión, acceso y consulta [5]. 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 [6]. Si además es necesario trabajar con imágenes y textos [1] los sistemas de soporte a la investigación o de apoyo al trabajo del profesional se fundamentan en interfaces de interacción con la información mas complejos [2].

En esta presentación del corpus construido desde los textos disponibles en la Hemeroteca Digital de la BNE, Diario de Madrid (1788-1825)**¡Error! Marcador no definido.**, 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 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 [11] 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

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

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* [12].

Últimamente, las redes neuronales basadas en modelos de lenguaje mejoran la detección de entidades, especialmente desde la publicación del modelo BERT [4] en 2018, o los modelos de lenguajes basados en *Transformers*. En [7] 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

La dificultad para aplicar la tecnología actual de PLN 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 de pago a partir de un cierto límite de uso. Con el registro se obtienen 500 créditos (unas 500 páginas). La herramienta está 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).

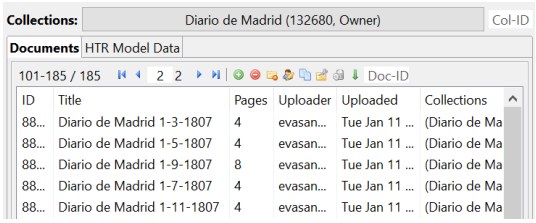


Figura 1. Carga de ficheros.

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 manualmente.

[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/

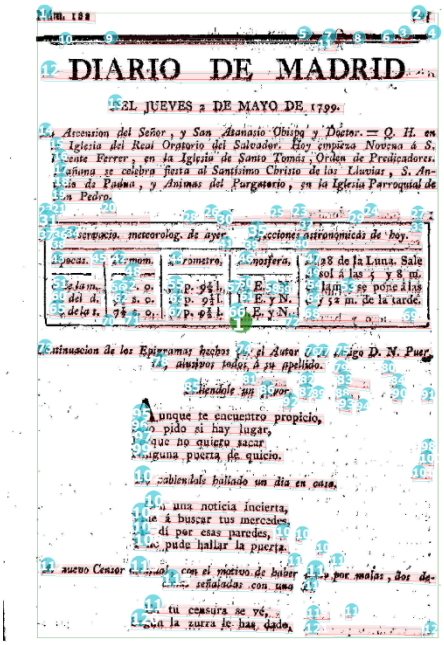


Figura2. Reconocimiento de la estructura.

Una reconocidas las regiones se transcribe el texto, línea a línea manualmente o con la ayuda de un modelo público seleccionado. Es posible que haya que editar la transcripción para corregir errores (Figura 3).

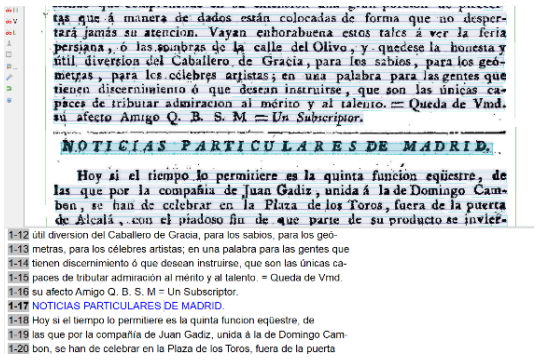


Figura 3. Transcripción manual.

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.

[7] www.clara-nlp.uned.es

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.

5. Agradecimientos

Este trabajo parcialmente financiado por el proyecto coordinado CLARA-NLP[7] consta de tres subproyectos para dominios especializados en historia[8], biomedicina [3] y economía [8].

Finalmente, un agradecimiento especial para la participación en este subproyecto de los estudiantes en prácticas V. Sánchez-Sánchez, R. Garcia-Sánchez y A. Rodriguez-Francés.

Referencias

[1] J. Benavent, X. Benavent, E. de Ves, R. Granados, A. García-Serrano, Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches, M. Braschler, D.Harman, E. Pianta (Eds.) CLEF, CEUR Proc., V 1176. 2010.

[2] J. Calle-Gómez, A. García-Serrano, P. Martínez, Intentional processing as a key for rational behaviour through Natural Interaction, Interacting with Computers V 18 N 6, pp:1419-1446, 2006.

[3] L. Campillos-Llanos, A. Terroba, S. Zakhir, A. Valverde, A. Capllonch, Building a comparable corpus and a benchmark for Spanish medical text simplification, Procesamien. del Lenguaje Natural 69, 2022.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep

[8] (PID2020-116001RB-C31), (PID2020-116001RB-C32), (PID2020-116001RA-C33)

Bidirectional Transformers for Language Unders., arXiv preprint 1810.04805, 2018.
[5] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers, CLEF proc. 2020.
[6] A. Garcia-Serrano, A. Menta-Garuz, La inteligencia artificial en las Humanidades Digitales: dos experiencias con corpus digitales, Revista de Humanidades Digitales, v.7, pp: 19-39, 2022.
[7] M. Jiang, Y. Hu, G. Worthey, R. C. Dubnicek, T. Underwood, Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts, CHR 2021: Computational Humanities Research Conference, pp. 266–279, 2021.
[8] A. Moreno-Sandoval, A. Gisbert, H. Montoro, Fint-esp: a corpus of financial reports in Spanish, Multiperspectives in Analysis and Corpus Design, Editorial Comares, pp. 89-102, 2020.
[9] J. Torruella Casañas, Lingüística de corpus: Génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística, Peter Lang Ed., 2017.
[10] M. Toscano, A. Rabadán, S. Ros, E. González-Blanco, Digital humanities in Spain: Historical perspective and current scenario. Profesional de la Información, 29(6), 2020.
[11] S. van Hooland, M. de Wilde, R. Verborgh, T. Steiner, R. Van de Walle, Exploring entity recognition and disambiguation for cultural heritage collections, Digital Scholarship Humanities, V30, N2, 2015.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information Processing Systems 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] and Joaquín Arellano[1]

[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, 20009, Spain

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) [1, 2]. 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 [3], 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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

aalvarez@vicomtech.org (A. Álvarez); vruiz@vicomtech.org (V. Ruiz); igonzalez@vicomtech.org (I. G. Torre); tetchegoyhen@vicomtech.org (T. Etchegoyhen); hgete@vicomtech.org (H. Gete); jarellano@vicomtech.org (J. Arellano)

0000-0002-7938-4486 (A. Álvarez); 0000-0003-2380-010X (I. G. Torre)

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.

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.

2.3.2. Model management

Running applications composed of several AI models on embedded systems requires dynamically controlling model activation and memory usage, given

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

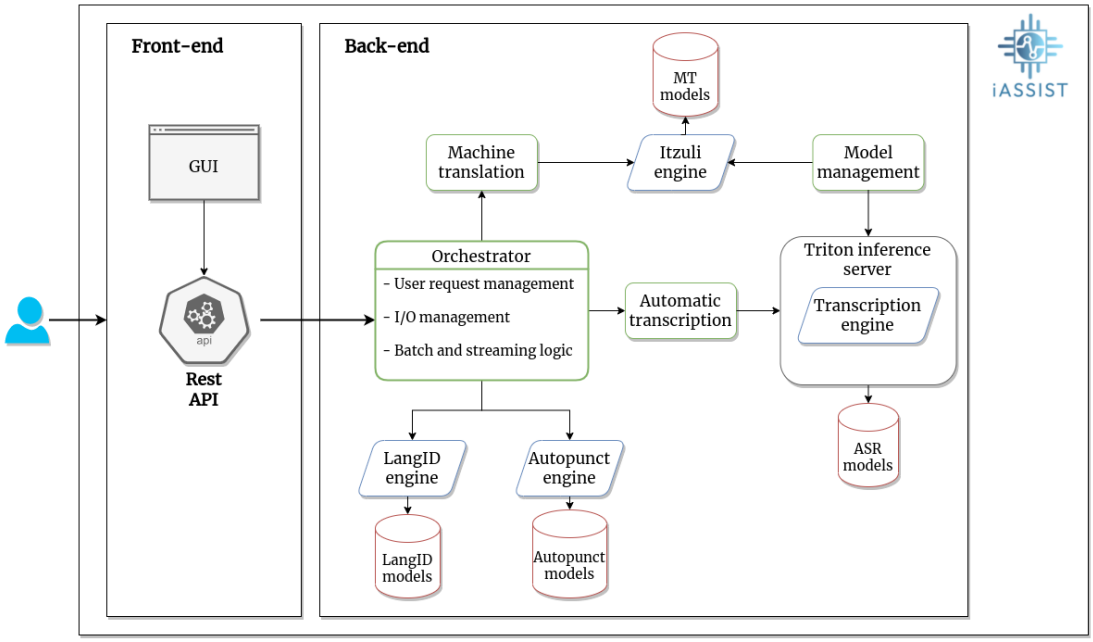


Figure 1: Core architecture of the iASSIST solution.

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 [4], 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 [5].

Finally, the raw transcriptions are enriched with capitalisation and punctuation marks generated by the BERT-based AutoPunct engine [6]. In addition to enhancing readability, splitting the raw text into correctly punctuated sentences increases the quality of machine translation results.

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

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 [7] in its own back-end to perform efficient NMT inference.

To optimise Transformer [2] 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 [8] 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.

Acknowledgments

iASSIST is partially funded by the Basque Business Development Agency, SPRI, under grant agreement ZL-2021/00103.

References

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

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

[3] V. K. Sarker, J. P. Queralta, T. N. Gia, H. Tenhunen, T. Westerlund, A survey on LoRa for IoT: Integrating edge computing, in: Proc of FMEC 2019, 2019, pp. 295–300.

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

[5] H. Vanholder, Efficient inference with TensorRT, in: GPU Technology Conference, volume 1, 2016, p. 2.

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

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

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

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[1], Lluís-F. Hurtado[1], José-Á. González[1], Vicent Ahuir[1], Encarna Segarra[1], Emilio Sanchis[1], María-José Castro[1] and Fernardo García[1]

[1] *VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Camino de Vera s/n, 46022 València, Spain*

Abstract

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.

English translation. 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

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

fpla@dsic.upv.es (F. Pla); lhurtado@dsic.upv.es (Lluís-F. Hurtado); jogonba2@inf.upv.es (José-Á. González); viahes@eui.upv.es (V. Ahuir); esegarra@dsic.upv.es (E. Segarra); esanchis@dsic.upv.es (E. Sanchis); mcastro@dsic.upv.es (M. Castro); fgarcia@dsic.upv.es (F. García)

CEUR Workshop Proceedings (CEUR-WS.org)

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.

De entre las principales características comunes a estas herramientas podemos destacar las siguientes. Todas ellas sueles ser herramientas de pago y que se basan esencialmente en la web con gran relevancia del componente gráfico. En algunos casos permiten exportar la información utilizando formatos estándar.

La mayoría de las herramientas están diseñadas para la monitorización de la red social Twitter aunque algunas contemplan otras redes sociales.

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.

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.

Figure 1: Arquitectura del Sistema GUAITA.

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.

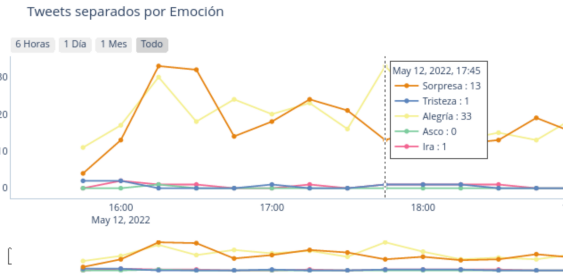


Figure 2: Salida del sistema, detección de emociones.

4. Modelos de clasificación

Para el motor de clasificación se han utilizado tres *encoders* : TWILBERT[1], BETO[2] y XLM-T[3]. 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[4] para los modelos de polaridad, Irosva 2019[5] para los modelos de detección de ironía, EmoEvalEs 2021[6] para la detección de emociones y lenguaje ofensivo, HateEval 2019[7] para los modelos de detección de lenguaje del odio y agresividad, HaHa 2019[8] para la detección de humor presente en los tweets y Detoxis 2021[9] 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.

Acknowledgments

Este trabajo ha sido parcialmente subvencionado por la Agencia Valenciana de la Innovació (AVI) de la Generalitat Valenciana, proyecto GUAITA (INNVA1/2020/61), el Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-11-21) y por el Ministerio de Ciencia e Innovación y fondos de la Unión Europea con el proyecto BEWORD PID2021-126061OB-C41.

References

[1] J.-Á. González, L.-F. Hurtado, F. Pla, Twilbert: Pre-trained deep bidirectional transformers for spanish twitter, Neurocomputing 426 (2021) 58–69. URL: https://www.sciencedirect.com/science/article/pii/S0925231220316180. doi:https://doi.org/10.1016/j.neucom.2020.09.078.

[2] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[3] F. Barbieri, L. E. Anke, J. Camacho-Collados, Xlm-t: A multilingual language model toolkit for twitter, 2021. URL: https://arxiv.org/abs/2104.12250. doi:10.48550/ARXIV.2104.12250.

[4] e. a. Manuel Carlos Díaz-Galiano, Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus, in: Proceedings of the Iberian Languages Evaluation, IberLEF@SEPLN 2019, Bilbao, Spain, volume 2421 of *CEUR Workshop Proceedings*, 2019, pp. 550–560.

[5] R. O. Bueno, F. M. R. Pardo, D. I. H. Farías, P. Rosso, M. M. y Gómez, J. E. Medina-Pagola, Overview of the task on irony detection in spanish variants, in: IberLEF@SEPLN, 2019.

[6] F. M. Plaza-del Arco, S. M. Jiménez Zafra, A. Montejo Ráez, M. D. Molina González, L. A. Ureña López, M. T. Martín Valdivia, Overview of the emoevales task on emotion detection for spanish at iberlef 2021, 2021.

[7] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, 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, 2019, pp. 54–63.

[8] L. Chiruzzo, S. Castro, A. Rosá, HAHA 2019 dataset: A corpus for humor analysis in Spanish, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 5106–5112.

[9] M. Taulé Delor, A. Ariza, M. Nofre, E. Amigó Cabrera, P. Rosso, Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish, 2021-09.

Plugin for automatisation of phonetic-phonological analysis and obtaining analytical feedback for Spanish learners

Plugin para la automatización del análisis fonético-fonológico y la obtención de retroalimentación analítica para estudiantes de español

Tamara Couto-Fernández[1], Albina Sarymsakova[2], Nelly Condori-Fernández[3] and Patricia Martín-Rodilla[4]

[1 3 4] *University of A Coruña, Faculty of Computer Science, Camiño do Lagar de Castro, 6, A Coruña, 15008, Spain*
[2] *University of A Coruña, Faculty of Philology, Campus da Zapateira, A Coruña, 15008, Spain*

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. Introduction

The present work is framed in the area of natural language processing, specifically, in the comparative-contrastive analysis of intonation for the didactic purposes provided by our original tool PAFe. Despite the existence of some tools, such as the Oplustil and Toledo [11] proposal, or the study by Strik, Truong, Wet and Cucchiarini [8], which offer results of phonetic-phonological similarity or detect errors made in pronunciation. Nonetheless, no tool provides both facilities at the same time, nor offers to monitor the evolution of the students.

For this reason, we have decided to develop a system that complements language teaching, in particular, one that can be used remotely or in hybrid modalities.

Our tool offers the functionality to perform an instant comparative analysis of a student's pronunciation, taking as a reference the speech of a native speaker, and observing the evolution of this through data stored in history.

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
EMAIL: albina.sarymsakova@udc.es (A. Sarymsakova); tamara.cfernandez@udc.es (T. Couto-Fernández); n.condori.fernandez@udc.es (N. Condori-Fernández); patricia.martin.rodilla@udc.es (P. Martín-Rodilla)
ORCID: 0000-0003-0381-0239 (A. Sarymsakova); 0000-0002-1044-3871 (N. Condori-Fernández); 0000-0002-1540-883X (P. Martín-Rodilla)

For the development of our plugin, several technologies have been used to support the work done, such as Praat, Python and PostgreSQL.

2. Methodology

We start designing our work based on the following essential principles of intonation analysis:

1. We annotate the syllables of each speech act in a Praat textgrid (Boersma and Weenink [1]; we identify pitch values of all vowels in the syllables (voiced or voiced consonants are measured as well), using the Praat Script developed by Mateo Ruiz [9, 10], which extracts the absolute values in Hz, relativises them and draws the standardised melody graph;
2. we discriminate relevant frequency values between tonal segments from irrelevant values; according to Cantero Serena [2, 3], Font-Rotchés and Cantero Serena [6, 7], less than 10% difference between segments is considered imperceptible.

Once we have obtained the relevant data from the intonation analysis, we move on to the PAFe architecture.

Our project develops an extension to an existing desktop application for acoustic speech analysis: Praat. Therefore, we start from a developed architecture to which a new module (PAFe) is coupled (Figure 1) consisting of Praat scripts, Python code and a PostgreSQL database.

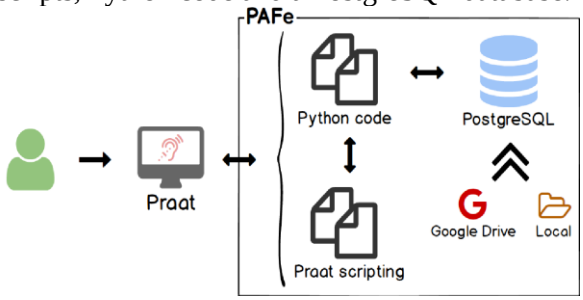


Figure 1: Overall architecture of PAFe

Praat, through its scripting, allows command line calls to other systems, as described by Dragos-PaulPop [5], thus making it possible to extend the application through the use of other languages and technologies, external to Praat. This new module (PAFe) communicates with the original system by employing new Praat scripts that are associated with the application's menu items (see Figure 2), from which these files are executed. Sometimes, the new module dispenses with calls to Praat and generates information windows directly from Python code files. The intermediary between Praat and the data managed in the database is Python.

We employ natural language processing and audio processing techniques in our tool, taking as our main source the human voice recordings of native speakers and students. Praat allows us to extract quantitative information at the prosodic level from the audios.

Subsequently, the native/student comparative algorithms in terms of prosodic aspects that are presented and implemented by the tool can offer comparative information between two native/student audios to provide feedback in Spanish language learning. These algorithms are an original contribution implemented in the tool since there was no algorithmic proposal of this type for Spanish until now.

We have developed the PAFe Plugin following an iterative and incremental methodology based on agile technologies and scrum development methodology, based on the work of Schwaber and Sutherland [12]. Figure 2: Example of User Interface visualising the new functionalities added in Praat

In the following, we describe the development of our tool.

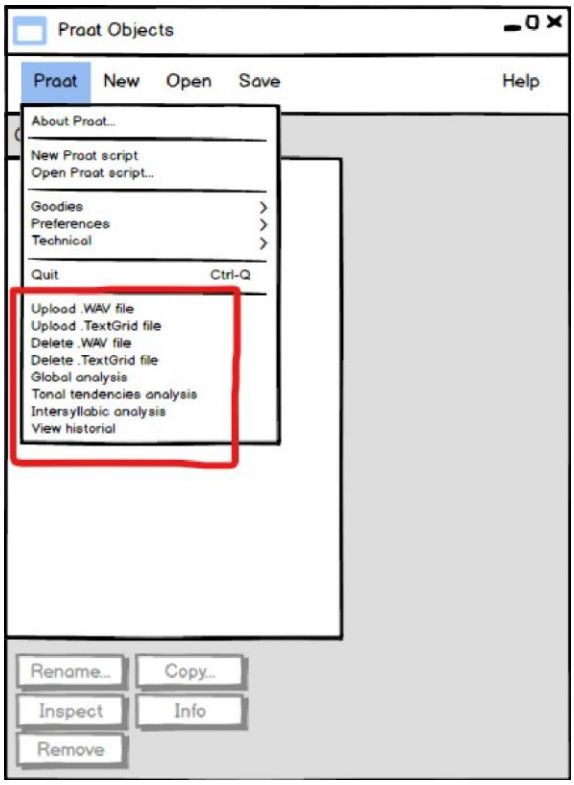


Figure 2: Example of User Interface visualising the new functionalities added in Praat

In the following, we describe the development of our tool.

3. Solution: PAFe Plugin

Our PAFe tool, in its final version, allows comparative analysis by providing similarity results and intonation graphs based on pitch values[2] and tonal tendency in each defined segment and, finally, visualisation of a student's progress over time. We highlight the following operations made possible by our plugin:

1. The application allows the creation of different profiles to facilitate the process of managing the data uploaded by users.
   a) First of all, the teacher is registered.
   b) A pupil is then assigned to the teacher previously registered. This step avoids confusion if there is more than one user of the same computer or laptop.
   c) Finally, the profile of a native Spanish speaker is recorded to upload the data that will serve as a reference for the programme;
2. PAFe enables the management of WAV and TextGrid files[3]: our programme includes both storage and deletion of audio files and annotations;
3. It also allows for different types of acoustic analysis (global analysis, tonal tendency analysis and intersyllabic analysis): the algorithm that performs the global analysis consists of dividing the previously saved audios of learners and native speakers of Spanish into about 1000 intervals (discarding silences) to obtain very precise comparative values. However, this type of analysis does not provide feedback about possible deviations in tone but provides generic data on the percentage similarity of the native speaker's and learner's audio. As far as tonal tendency analysis is concerned, the programme works with .TextGrid annotations and the previously saved .WAV audio files. In this case, the utterances are divided by words and, to obtain the similarity locally, it is indicated whether the pitch of each word has been reproduced correctly or not and, in case it has not been reproduced correctly, the percentage of deviation is indicated; the percentage of pitch similarity and the average difference between two audios are also obtained. Finally, the intersyllabic acoustic analysis is a comparative analysis, syllable by syllable, of the similarity between the tone realisation of a learner and that of a native speaker; in this case, for each syllable, the difference in pronunciation concerning the reference audio is indicated, as well as the percentage of similarity of tone and the average difference between two audios is obtained. According to the results obtained through this last type of analysis, both the similarity and the difference between the reference audio and the learners' audio are shown more accurately. Finally, we can see the evolution of our students' results through the option to view the history.

Finally, we show a flowchart (Figure 3) that provides information about the behaviour of our plugin, exposing the functionalities and their interrelation, as well as presenting the operators that interact with the application and their restrictions.

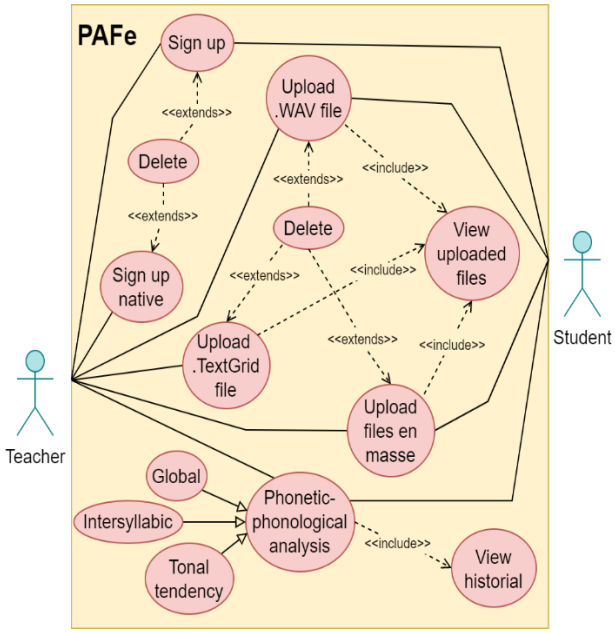


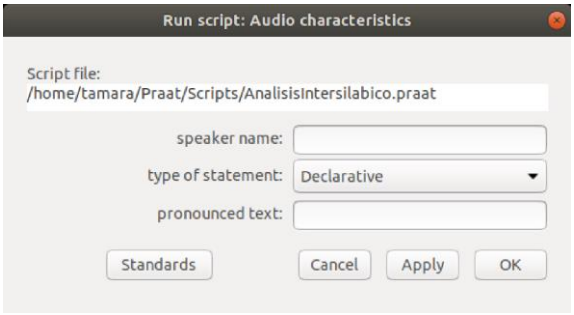
Figure 3: Use case diagram (PAFe functionalities and main actors)

4. Illustrative example of intersyllabic analysis

In this section, we show how one type of comparative analysis is carried out. To perform the intersyllabic analysis, it is necessary to fill in a form (Figure 4) with the data that characterise the audio of the learner we want to compare.

[2] Tone frequency in Hz

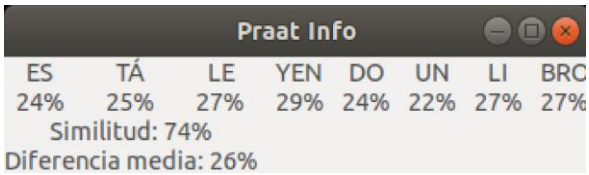
[3] File with tags segmenting associated audio



Figures 4: Form for conducting an intersyllabic analysis

The audios of that student that meet these properties are then filtered out and display a window with a drop-down menu for the selection of the audio to be analysed. Once the audio is selected, the corresponding. TextGrid file is selected in the same way.

Each type of analysis returns different results. For the intersyllabic analysis, we show a similarity result per syllable and the average percentage difference (Figure 5). Finally, we obtain a graph with the tonal differentiation curves in each syllable for each audio (Figure 6).



Figures 5: Intersyllabic analysis information

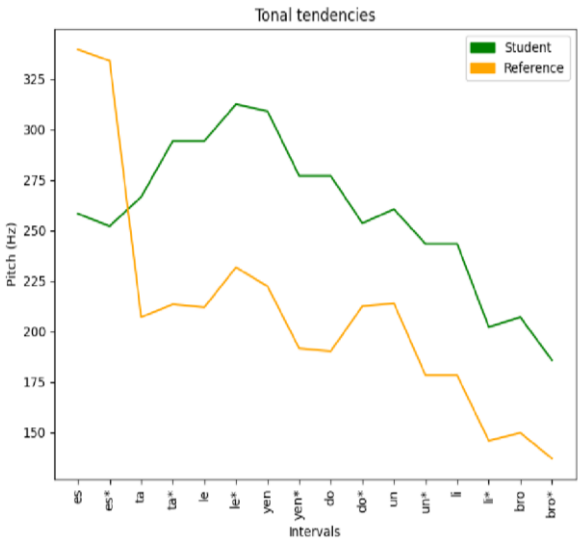


Figure 6: Graph showing the tonal curves of each audio for each syllable (the X-axis represents the syllable division of an utterance and the Y-axis the pitch values).

5. Conclusions

In conclusion, we highlight the following key issues that we have addressed in this paper:

1. The PAFe tool allows different types of comparative-contrastive analysis of the intonation (global, tonal tendencies and intersyllabic) of EFL learners and native speakers of Spanish; Among them, we consider the intersyllabic as the most accurate since the results of tonal difference appear syllable by syllable and show the tonal deviations of the students, and the global as the most efficient in terms of response time since it does not require the uploading of TextGrids, and the segmentation is done in an automated way, as shown by the empirical data of the Couto Fernández [4] work.
2. This application has several functions; apart from performing the intonational analysis, it allows to store the audios, the . TextGrid files and the results of the analysis (the history) of each utterance according to the profile of the speaker (student or native speaker of Spanish).
3. PAFe has been developed to achieve the following didactic objectives: to facilitate the work of teachers with regard to the identification and correction of intonation deviations (we have carried out an empirical analysis with teachers of Spanish as a foreign language, where we measured the degree of satisfaction with PAFe, with positive results, as indicated in the work Couto Fernández [4]; to store the results of the analyses carried out for future improvement; to serve as a self-evaluation and self-correction tool for ELE students, given that the tool itself allows them to upload .WAV and . TextGrid files, run the analyses and obtain the results without constant help from teachers.

As a future line of research, we highlight the need to measure this degree of feedback to students empirically.

As far as we know, it is the only existing solution both under Praat and outside Praat that allows this type of analysis and offers feedback to the student in the Spanish language. We highlight that as feedback and self-evaluation, our tool offers the percentage of similarity and difference of pitch values so that the student can correct his pronunciation. Also, as future lines of work, we plan to improve the graphical environment of the plugin and open to the student, as an end user, the possibility of its use via the web.

References

[1] P. Boersma, D. Weenink, Praat: doing phonetics by computer, 2019. URL: http://www.praat.org/.

[2] F. J. Cantero Serena, Teoría y análisis de la entonación, volume 54, 2002.

[3] F. J. Cantero Serena, Análisis prosódico del habla: más allá de la melodía, Comunicación Social: Lingüística, Medios Masivos, Arte, Etnología, Folclor y otras ciencias afines 2 (2019) 485-498.

[4] T. Couto Fernández, Una herramienta de análisis del habla de audio para proporcionar retroalimentación automática a los estudiantes en la pronunciación en español. UDC. A Coruña.

[5] Dragos-Paul Pop, Adam Altar, Designing an MVC Model for Rapid Web Application Development, Procedia Engineering 69 (2014) 1172-1179. DOI: 10.1016/j.proeng.2014.03.106

[6] D. Font Rotchés, F. J. Cantero Serena, La melodía del habla: acento, ritmo y entonación, Eufonía: didáctica de la música (2008) 19-39.

[7] D. Font Rotchés, F. J. Cantero Serena, Melodic Analysis of Speech Method applied to Spanish and Catalan, Phonica 5 (2009) 33-47.

[8] H. Strik, K. Truong, F. Wet, C. Cucchiarini, Comparing different approaches for automatic pronunciation error detection, Speech Communication 51 (2009) 845–852. DOI: 10.1016/j.specom.2009.05.007

[9] M. Mateo Ruiz, Protocolo para la extracción de los datos tonales y curva estándar en análisis melódico del habla, Phonica 6 (2010) 49-90.

[10] M. Mateo Ruiz, Scripts en Praat para la extracción de datos tonales y curva estándar, Phonica 6 (2010) 91-111.

[11] P. Oplustil, G. Toledo, Uso de una herramienta didáctica para la práctica de la entonación en hablantes no nativos de español, Sintagma: Revista de lingüística 31 (2019) 37–50.

[12] K. Schwaber, J. Sutherland, La guía definitiva de scrum: Las reglas del juego, 2020.

appForum: Una aplicación para el procesamiento de foros

appForum: An application for forum processing

Alvaro Rodrigo[1], José Luis Fernández-Vindel[1], Jorge Pérez-Martín[1], Ismael Iglesias[2], Víctor Fresno[1], Aitor Díaz[1], Francisco Javier Sánchez[2] and Roberto Centeno[1]

[1]*Intelligent Systems for Learning, Grupo de Innovación Docente de la UNED*
[2]*Intecca, UNED*

Abstract

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.

English translation. Forums are still the main way 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 they 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 the transformation of the forums it receives 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

Keywords, of, the, paper.

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 análi-

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
alvarory@lsi.uned.es (A. Rodrigo); jlvindel@dia.uned.es (J. L. Fernández-Vindel); jperezmartin@dia.uned.es (J. Pérez-Martín); iiglesias@intecca.uned.es (I. Iglesias); vfresno@lsi.uned.es (V. Fresno); adiazm@scc.uned.es (A. Díaz); fjsanchez@intecca.uned.es (F. J. Sánchez); rcenteno@lsi.uned.es (R. Centeno)
http://nlp.uned.es/~alvarory/ (A. Rodrigo); https://www.uned.es/universidad/docentes/informatica/jorge-perez-martin.html (J. Pérez-Martín); http://nlp.uned.es/~vfresno/ (V. Fresno); http://nlp.uned.es/~rcenteno (R. Centeno)
0000-0002-6331-4117 (A. Rodrigo); 0000-0002-3588-7233 (J. Pérez-Martín); 0000-0003-4270-2628 (V. Fresno); 0000-0001-9095-4665 (R. Centeno)

sis 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.

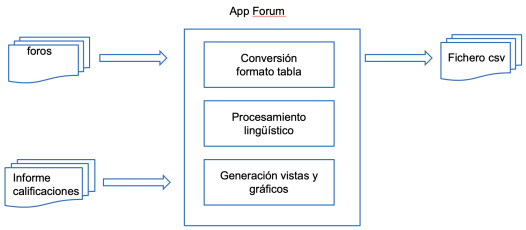


Figure 1: Fjujo de procesamiento de información de la aplicación *appForum*.

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 [1]. 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 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.

[1] https://spacy.io/

[2] https://pypi.org/project/sentiment-analysis-spanish/

[3] La información del 'Autor' ha sido anonimizada.

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.

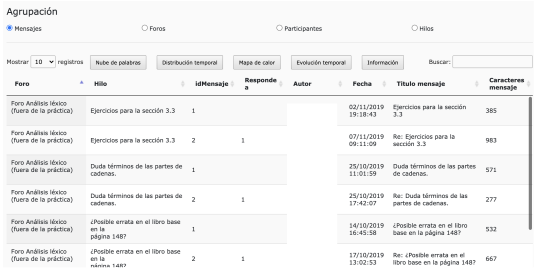


Figure 2: Ejemplo de cómo se muestra la información en la vista de mensajes.

Figure 3: Ejemplo de nube de palabras generada a partir del texto de los mensajes de un determinado foro.

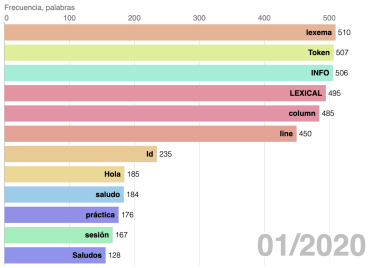


Figure 4: Ejemplo de un gráfico de barras animado que muestra a lo largo del tiempo los términos más relevantes acerca del contenido de un foro.

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. 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.

Figure 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.

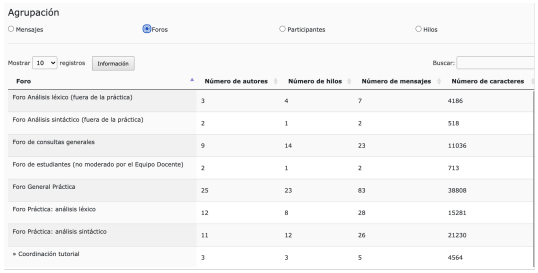


Figure 6: Ejemplo de cómo se muestra la información en la vista de foros.

3.3. Vista de Participantes

La tercera vista se centra en los usuarios que escriben mensajes en el foro, ya que no se 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.

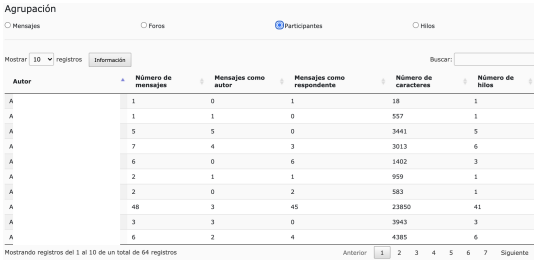


Figure 7: Ejemplo de cómo se muestra la información en la vista de participantes.

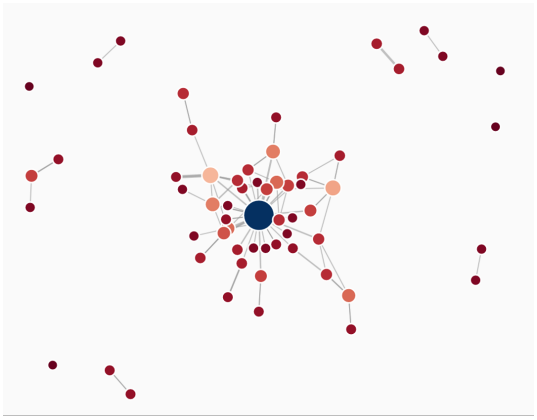


Figure 8: Representación en forma de grafo de las interacciones entre usuarios.

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.

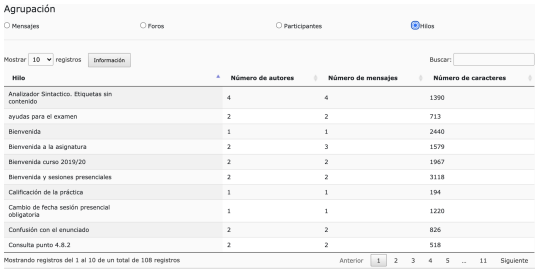


Figure 9: Ejemplo de cómo se muestra la información en la vista de hilos.

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.

References

[1] J. Russell, A circumplex model of affect, Journal of personality and social psychology 39 (1980) 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, José Ramom Pichel and Pablo Gamallo

Centro de Investigación en Tecnoloxías da Información (CITIUS), Universidad de Santiago de Compostela, Spain

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

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 [1] Traditionally, rule-based systems like Apertium [2] 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 [3, 4]. 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. [5] 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 [6] 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 [7, 8], 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 [9]. 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* [10] 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) cor-

[1] https://github.com/gamallo/port2gal

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

john.ortega@usc.gal (J. E. Ortega); iria.dedios@usc.gal (I. de-Dios-Flores); jramon.pichel@usc.gal (J. R. Pichel); pablo.gamallo@usc.gal (P. Gamallo)

0000-0002-2328-3205 (J. E. Ortega); 0000-0002-5941-1707 (I. de-Dios-Flores); 0000-0001-5172-6803 (J. R. Pichel); 0000-0002-5819-2469 (P. Gamallo)

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).

pus 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 [2], 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 [11], 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 [12]

The Transformer implementation, described in Garg et al. [13], 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 [14].

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 [15]. 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 [16]. 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.

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

[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
[6] https://demos.citius.usc.es/nos_tradutor

Figure 1: A screen capture of the web interface.

(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 [2] for example.

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.

Acknowledgments

This research was funded by the project "Nós: Galician in the society and economy of artificial intelligence", 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

[1] R. Knowles, J. Ortega, P. Koehn, 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, 2018, pp. 249–255.

[2] M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, F. M. Tyers, Apertium: a free/open-source platform for rule-based machine translation, Machine translation 25 (2011) 127–144.

[3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[4] P. Koehn, R. Knowles, Six challenges for neural machine translation, arXiv preprint arXiv:1706.03872 (2017).

[5] J. Gu, Y. Wang, K. Cho, V. O. Li, Improved zero-shot neural machine translation via ignoring spurious correlations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1258–1268. URL: https://aclanthology.org/P19-1121. doi:10.18653/v1/P19-1121.

[6] J. R. P. Campos, P. M. Fernández, O. Gomez, P. Gamallo, A. C. García, Carvalho: English-galician smt system from europarl english-portuguese parallel corpus, Procesamiento Del Lenguaje Natural (2009) 379–381.

[7] J. E. Ortega, R. C. Mamani, K. Cho, Neural machine translation with a polysynthetic low resource language, Machine Translation 34 (2020) 325–346.

[8] J. E. Ortega, R. A. Castro-Mamani, J. R. Montoya Samame, Overcoming resistance: The normalization of an Amazonian tribal language, in: Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, Association for Computational Linguistics, Suzhou, China, 2020, pp. 1–13. URL: https://aclanthology.org/2020.loresmt-1.1.

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).

[9] J. R. Pichel, P. Gamallo, I. Alegria, M. Neves, A methodology to measure the diachronic language distance between three languages based on perplexity, Journal of Quantitative Linguistics 28 (2021) 306–336.

[10] K. Knight, J. Graehl, Machine transliteration, arXiv preprint cmp-lg/9704003 (1997).

[11] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. Rush, OpenNMT: Open-source toolkit for neural machine translation, in: Proceedings of ACL 2017, System Demonstrations., Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 67–72. URL: https://www.aclweb.org/anthology/P17-4012.

[12] P. Gamallo, M. Garcia, C. Piñeiro, R. Martinez-Castaño, J. C. Pichel, 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), 2018, pp. 239–244. doi:10.1109/SNAMS.2018.8554689.

[13] S. Garg, S. Peitz, U. Nallasamy, M. Paulik, Jointly learning to align and translate with transformer models, CoRR abs/1909.02074 (2019). URL: http://arxiv.org/abs/1909.02074. arXiv:1909.02074.

[14] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, arXiv preprint arXiv:1808.06226 (2018).

[15] M. D. C. Bayón, P. Sánchez-Gijón, Evaluating machine translation in a low-resource language combination: Spanish-galician., in: Machine Translation Summit XVII Vol. 2: Translator, Project and User Tracks, 2019, pp. 30–35.

[16] P. Lison, J. Tiedemann, M. Kouylekov, 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), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1275.