CONVERSA: Effective and Efficient Resources and Models for Transformative Conversational AI in Spanish and Co-official Languages

David Griol[1], Ksenia Kharitonova[1], David Pérez-Férnandez[2], Asier Gutiérrez-Fandiño[3] and Zoraida Callejas[1,4]

[1]*Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Granada, Granada, Spain*

[2]*Universidad Autónoma de Madrid, Madrid, Spain*

[3]*LHF Labs, Bilbao, Spain*

[4]*Centro de Investigación en TIC de la Universidad de Granada (CITIC-UGR), Granada, Spain*

Abstract

Access to information is increasingly conversational. However, there is a lack of conversational AI training material for Spanish and the co-official languages, in general and for specific key tasks and domains. Additional barriers include steep computational costs for training conversational agents and challenging inference times, and a lack of guarantees for the safety and transparency of conversational systems. The CONVERSA project (TED2021-132470B-I00) constitutes a step forward to democratize access to conversational AI through computation and data efficient development and testing of innovative, open and safe resources in Spanish and co-official languages. The duration of the project is from December 2022 to November 2024.

Keywords

Conversational AI, conversational systems, dialogue systems, corpus, datasets, language models, open access

1. Introduction

The term conversational artificial intelligence, coined recently in academic research, refers to Natural Language Processing (NLP) technologies, such as dialog systems, chatbots or intelligent virtual assistants, with which users can engage in a conversation in natural language and artificial intelligence techniques are extensively used [1, 2]. These systems provide a low barrier entry for users, enabling them to access information, interact in an intuitive way with services, resources, and data on the Internet, as well as with their surrounding environment [3, 4].

The development of conversational AI has seen rapid progress in recent years, both text- and voice-based, enabled by large pretrained language models and a number of new consumer-facing applications and intelligent devices, such as personal mobile assistants [5, 6], social networks, messaging applications, and intelligent speakers. Application domains have increased dramatically ranging from retail, telecommunication, finance, health and e-government [7].

Recent market reports [8, 9], as well as relevant surveys [7] and both academic and industrial vision statements on the future of conversationally enabled applications [10] highlight how the use of natural language (and speech interaction) is changing the way people connect to the information that they need [11].

This has been accentuated during the Covid-19 pandemic with the highly successful increased use of Conversational AI for e-government [12, 13, 14] and customer purchases [15, 16]. There are predictions by 2024 that consumer retail spending via chatbots will reach $142 billion worldwide, up from just $2.8 billion in 2019 [9]. Revenue in NLP is estimated to increase from $3.2 billion in 2017 to more than $43 billion in 2025 [17].

Since the 2000s the emphasis in the scientific spoken dialogue systems community has moved from handcrafted systems (symbolic and logic-based AI) to data-driven systems using machine learning. Machine-learning techniques avoid specifying dialogue state machines and make it possible to address the inconveniences derived from unexpected patterns that slot filling approaches cannot anticipate. Key to the development of effective chat-based conversational AI technology using this paradigm is the availability of a large volume of training material [1, 18].

These new systems rely on high-quality datasets/corpora for the training of deep-learning algorithms to develop precise models. The preparation of a high-quality gold standard corpora for natural

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations

dgriol@ugr.es (D. Griol); ksenia@ugr.es (K. Kharitonova); david.perez@inv.uam.es (D. Pérez-Férnandez); asier@lhf.ai (A. Gutiérrez-Fandiño); zoraida@ugr.es (Z. Callejas)
0000-0001-6266-5321 (D. Griol); 0000-0002-9139-2622 (K. Kharitonova); 0000-0002-2214-0245 (D. Pérez-Férnandez); 0000-0002-7368-6950 (A. Gutiérrez-Fandiño); 0000-0001-8891-5237 (Z. Callejas)

language processing on a large scale is a challenging task due to the need of data cleaning, accurate language identification models, and precise content parsing tools. Transformer-based models and self-supervised learning mechanisms have shown promising results in key NLP tasks and academia and industry are currently developing large transformer-based linguistic models [1, 19].

Due to their size, the training and adjustment for the implementation of these conversational services currently requires large computational resources and expert knowledge for their optimization [20, 21]. Current models have a large number of parameters and are trained with huge collections of training examples. GPT-1 had 117 million parameters to work with, GPT-2 had 1.5 billion. GPT-3 used for its training, among others, 410 billion crawled data tokens, 67 billion book tokens, and 3 billion Wikipedia tokens. The full version of OpenAI GPT-3 has around 175 billion trainable parameters. GPT-3.5 is significantly larger, with a staggering 355 billion parameters. GPT-4 has been trained on a large amount of internet content and it is able to handle more complex instructions and produce higher-quality long-form writing. Bard, powered by Google's Language Model for Dialogue Applications (LaMDA), was released in 2021 with 137 billion parameters trained using 1.56 trillion words of public dialog data and web text.

Regarding open-source alternatives, the LLaMA project has encompassed a set of foundational language models that vary in size from 7 billion to 65 billion parameters and were trained on millions of tokens extracted exclusively from publicly available datasets. Stanford Alpaca claims that it can compete with ChatGPT and the training can be completed with less than 600$. Vicuna is finetuned from the LLaMA model on user-shared conversations collected from ShareGPT. GPT4ALL is a community-driven project and was trained on a massive curated corpus of assistant interactions, including code, stories, depictions, and multi-turn dialogue.

In addition, the black box nature of the neural networks that these techniques require, makes it difficult to explain the reasons behind the behavior of a conversational agent at different points of the conversation, undermining user-system trust and making it difficult to tackle changes and evolve the system reliably. In addition, topic detection and switching is challenging in particular for end-to-end neural dialogue systems, as it is difficult to incorporate long-distance context information.

There are still remaining open issues that concern flowing/navigating through dialogues within a conversational AI system. One example of it is the monolithic structure that dialogues exhibit and their single-block granularity (i.e., without dialogue sub-components), so that it is not possible to architect dialogues with several entrance points. This is particularly desirable when data about users does not change frequently over time and are required by one or more tasks. Being able to exploit the history/record for several dialogues would make the system more effective and enhance the user experience. Hence, monolithic dialogues negatively affect the quality of interactions and users' satisfaction, since the same repetitive questions/answers pairs are followed for every user.

A more elaborated case of this kind of situation -that we have experienced in past projects- occurs when it can be inferred (possibly following a machine-learning process) that particular responses to a question at some point allow some subcomponents of a dialogue to be skipped. However, once the first model is constructed and deployed, the experience gained over time cannot be exploited, without constructing a new complete model. Again, data and knowledge sharing between different conversational AI applications could also help boost learning processes from past experiences and training data.

A related challenge is to ensure system proactivity, that is, the ability of the system to start the conversation at adequate moments or to show unsolicited behaviours and contents that may be helpful for the users. This requires having accurate knowledge about the user and the context of the interaction, which involves being able to identify the user, to identify whether the user is actively involved in the conversation (and not for example someone just present in the environment), manage turn-taking and distinguishing between multiple topics.

2. Description and Objectives of the Project

A key aspect of digital transformation is to be able to engage, serve, and empower the users it is directed to. This implies that digital services must be accessible, provide open access to reliable information and open interfaces for businesses and citizens, make full use of existing online services with more agile discovery and personalization, ensure ease of use, and guarantee security and privacy.

Access to information is increasingly conversational and task-oriented chat systems help users achieve their goals efficiently using natural language, ensuring accessibility and personalized omnichannel interaction, fostering user-system trust through explainable social dialogues. However, there are three main barriers to digital transformation in Spain through the democratic adoption of conversational AI technology by a wide spectrum of technological and societal stakeholders:

- A lack of training material for Spanish and the co-official languages, in general and for specific key tasks and domains.

- Steep computational costs for training conversational agents and challenging inference times.
- A lack of guarantees for the safety and transparency of conversational systems.

The CONVERSA project (TED2021-132470B-I00) addresses these challenges by addressing the following research questions:

- How can we automatically generate, simulate and transfer training data to produce engaging chat-based conversations in Spanish and the coofficial languages? By using data-driven technology to create better performing chat-based technology, public institutions and companies can outsource more citizen/ customer interaction to neural network-based agents.
- How can we develop and adapt chat-based conversational systems in a computationally and data-efficient manner? By further developing and improving neural network-based models for conversational systems for less resource-intensive scenarios, we will directly contribute to the way computers can communicate with humans.
- How can we do all of this securely and transparently, without violating privacy, and with provisions for explainability and data provenance? Accountability also implies security transparency and explainability. Moreover, explainability is a fundamental aspect of knowledge systems, which is also enforced in European regulations.

The Multilingual Technology Alliance and the META-NET Network of Excellence Europe's Languages in the Digital Age, highlight that the lack of natural language processing resources for European languages are the most significant impediment that must be overcome for further conversational AI adoption. Although Spanish is currently the second language in the world by number of native speakers, it is not a well-resourced language in terms of the core technologies and datasets needed to build state-of-the-art language-based solutions. This problem is even more accentuated in the case of co-official languages. Spain has a clear opportunity to lead the development of openly accessible digital services in Spanish and the co-official languages that can be exploited not only in Spain but also in the multiple countries with Spanish-speaking population.

Through CONVERSA, the generation of high-quality and accessible labeled data and pretrained models will drive the development, optimization and deployment of conversational AI for Spanish and co-official languages, fueling the generation and adoption of transformative conversational AI solutions in an environmentally responsible way.

Not undertaking this endeavor would only generate a dependence from the solutions offered by the major tech companies. At present, state-of-the-art conversational AI technology, resources and development tools are mostly owned by big tech companies (e.g., Google DialogFlow, Google Assistant, Amazon Alexa Skills Kit, IBM Watson Assistant, or Microsoft Bot Framework). However, many public and private stakeholders are keen to take a more autonomous position concerning the technology to interact with citizens and customers for strategic (competitiveness) reasons, economic reasons, and legal reasons. In particular, accessibility and non-disclosure of customer/citizen data to third parties is key.

To be engaging and effective, modern conversational systems are computationally intensive and data hungry, requiring large-scale, language-specific, domain-specific, and task-specific training data. CONVERSA will address the development of effective methods and tools for domain adaptation and data augmentation for conversational AI, in order to be fast and deployable in lightweight computing environments, and overcome the current limited transferability of task knowledge between tasks.

Another important aspect is to enhance trust and ensure the reliability of the processes. Conversational interfaces to digital services pose the risk of harming users by conversing with them inappropriately, revealing sensitive private information about them or incurring in bias (gender and racial bias implicitly learned from automatically crawled language resources). A particular challenge for conversational systems is that they often directly confront users and so the impact can be immediate. CONVERSA specifically targets the development of explainability techniques to prevent harmful utterances, both safety and privacy-wise, and design trustworthy conversational systems more resilient against malicious manipulations.

CONVERSA has the general objective of democratizing access to conversational AI through computation and data efficient development and testing of innovative, open and safe resources in Spanish and co-official languages. The specific objectives are:

- To create and annotate open-access multi domain dialog corpora in Spanish and the co-official languages to train the specific models required to develop neural-bases AI conversational systems.
- To develop compute- and data- efficient neural architectures and open-access models for conversational AI, including methods and tools for domain adaptation and data augmentation.
- To develop open source tools for corpus creation, data debugging, bias analysis, and privacy analysis.
- To provide a replicable evaluation of the project

outcomes with particular attention to replicability, bias avoidance and privacy preservation.
- To showcase the use of the resources and models generated through the development and deployment of an initial pilot demonstrator of conversationally enabled innovative e-government services.

3. Scientific and Technical Impact

CONVERSA is meant to facilitate the following changes in the way that public and private industrial end users develop and deploy conversationally enabled applications:

- Change 1: Faster prototyping and retraining of conversationally enabled applications
- Change 2: Shift development of conversationally enabled applications from a supervised learning paradigm to a transfer learning-based paradigm;
- Change 3: Reduced development times and ecologically pernicious effects (waste or computation resources) of conversationally enabled applications for novel domains, tasks, and scenarios;
- Change 4: Improve customer trust by means of explanations of actions taken by a chat-based conversational assistant

In particular, the outcomes that will facilitate these changes are:

- Outcome 1: Efficient implementations and adaptation of conversational technology for new domains
- Outcome 2: Enriched domain-specific language models for the Spanish and co-official languages retail and service domains
- Outcome 3: Better coverage of stakeholders' product bases in conversational agents
- Outcome 4: Open data and open-source conversational technology
- Outcome 5: Improved user satisfaction and user trust in conversational interaction
- Outcome 6: Development of R&D technology in new contexts

The algorithmic methods and insights developed will be published as open-source software, and the data and corpora will be also openly available. Validated versions of the packages will be integrated in the open-source technology of Rasa Technologies, that is available for any organization to use and implement in their own service and retail environments.

The project outcomes therefore impact society directly and indirectly: directly because the resulting models and implementations will be available as open-source software and open corpora, and indirectly because the developed methods are in principle technology independent.

CONVERSA results have wide implications for technology areas (human-computer interaction, artificial intelligence, natural language processing, conversational assistants, user modelling, service orchestration) and can revolutionize the development of conversational systems for Spanish and co-official languages. Impact will be pursued at multiple levels:

- Information Retrieval (IR): Conversational search is an important research direction in information retrieval and question answering applications. By addressing the domain-specific challenges of conversational intelligence for e-goverment applications, sales and service-oriented chatbots, we hope to give a boost to the Spanish and cooficial languages conversational IR community with novel methods, models, datasets, and evaluation, especially for less-resourced environments.
- Natural Language Processing (NLP): by utilizing and adapting neural-based architectures and generating tools to this domain, we expect to develop new transformational models, along several dimensions (output quality, memory efficiency, and inference time). Designing novel methods for knowledge-based, multilingual, cross-domain adaptation for conversational AI is expected to transfer learning from transformer-based linguistic modeling, due to the specific challenges of this application.
- Trusted AI (i.e., human-centric AI) and ML security/privacy: Our goal is to develop techniques to avoid harmful utterances, non-disclosure of private data and through the prevention of bias in corpus and generator models (such as gender, race or ideology).

The international impact of CONVERSA is based, not only on the profound implications of its results, but also on the solid international production of the members of the research team, their well-established international connections with ongoing H2020 projects and the dissemination and communication activities envisaged.

Acknowledgments

This publication is part of the "CONVERSA: Effective and efficient resources and models for transformative conversational AI in Spanish and co-official languages" project with reference TED2021-132470B-I00, funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU"/PRTR".

References

[1] M. McTear, Conversational AI. Dialogue systems, Conversational Agents, and Chatbots, Morgan and Claypool Publishers, 2020. doi:10.1007/978-3-031-02176-3.

[2] M. McTear, Z. Callejas, D. Griol, The Conversational Interface: Talking to Smart Devices, Springer, 2016. doi:10.1007/978-3-319-32967-3.

[3] P. Cañas, D. Griol, Z. Callejas, Towards versatile conversations with data-driven dialog management and its integration in commercial platforms, Journal of Computational Science 55 (2021) 101443. doi:10.1016/j.jocs.2021.101443.

[4] T. Fu, S. Gao, X. Zhao, J. rong Wen, R. Yan, Learning towards conversational AI: A survey, AI Open 3 (2022) 14–28. doi:10.1016/j.aiopen.2022.02.001.

[5] S. Lee, R. Jha, Zero-shot adaptive transfer for conversational language understanding, in: Proc. of AAAI'19 Conference on Artificial Intelligence, Honolulu, Hawaii, USA, 2019, pp. 6642–6649. doi:10.1609/aaai.v33i01.33016642.

[6] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, P. Khaitan, Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset, in: Proc. of AAAI'20, New York, NY, USA, 2020, pp. 8689–8696. doi:10.1609/aaai.v34i05.6394.

[7] C. Gao, W. Lei, X. He, M. de Rijke, T.-S. Chua, Advances and Challenges in Conversational Recommender Systems: A Survey, AI Open 2 (2021) 100–126. doi:10.1016/j.aiopen.2021.06.002.

[8] Gartner, Making sense of the chatbot and conversational AI platform market, https://www.gartner.com/en/documents/3993709/making-sense-of-the-chatbot-and-conversational-aiplatfo, 2020. Acceded: June 2023.

[9] BusinessInsider, Chatbot market in 2021: Stats, trends, and companies in the growing AI chatbot industry, https://www.businessinsider.com/chatbot-market-stats-trends, 2021. Acceded: June 2023.

[10] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, et alt., Opinion Paper: So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy, International Journal of Information Management 71 (2023) 102642. doi:10.1016/j.ijinfomgt.2023.102642.

[11] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, Transactions on Machine Learning Research (2022).

[12] A. Androutsopoulou, N. Karacapilidis, E. Loukis, Y. Charalabidis, Transforming the communication between citizens and government through AI-guided chatbots, Government Information Quarterly 36 (2019) 358–367. doi:10.1016/j.giq.2018.10.001.

[13] J. Cabot, Chatbots y asistentes de voz, una oportunidad en la gestión de crisis sanitarias, The Conversation (2020).

[14] D. Griol, D. Pérez Fernández, Z. Callejas, Hispabot-Covid19: the official Spanish conversational system about Covid-19, in: Proc. of IberSPEECH 2021, Valladolid, Spain, 2021, pp. 139–142. doi:10.21437/IberSPEECH.2021-30.

[15] L. Gkinko, A. Elbanna, Designing trust: The formation of employees' trust in conversational AI in the digital workplace, Journal of Business Research 158 (2023) 113707. doi:10.1016/j.jbusres.2023.113707.

[16] I. U. Jan, S. Ji, C. Kim, What (de) motivates customers to use AI-powered conversational agents for shopping? The extended behavioral reasoning perspective, Journal of Retailing and Consumer Services 75 (2023) 103440. doi:10.1016/j.jretconser.2023.103440.

[17] Statista, Revenues from the natural language processing (NLP) market worldwide from 2017 to 2025, https://es.statista.com/estadisticas/1130045/mercado-global-de-procesamiento-de-lenguaje-natural/, 2022. Acceded: June 2023.

[18] P. Su, N. Mrksic, I. Casanueva, I. Vulic, Deep learning for conversational AI, in: Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 27–32.

[19] S. Young, Hey Cyba. The Inner Workings of a Virtual Personal Assistant, Cambridge University Press, 2021.

[20] A. Gutiérrez-Fandiño, D. Pérez-Fernández, J. Armengol-Estapé, D. Griol, Z. Callejas, esCorpius: A Massive Spanish Crawling Corpus, in: Proc. IberSPEECH 2022, 2022, pp. 126–130. doi:10.21437/IberSPEECH.2022-26.

[21] J. Ni, T. Young, V. Pandelea, X. F., E. Cambria, Recent advances in deep learning based dialogue systems: a systematic survey, Artificial Intelligence Review (2023) 3055–3155. doi:10.1007/s10462-022-10248-8.