Project presentation at CONVERSATIONS 2022 – the 6th International Workshop on Chatbot Research, Applications and Design, hosted by the University of Amsterdam, Netherlands, November 22-23.

ENRICH4ALL Project

Dimitra Anastasiou[+], Radu Ion*, Anders Ruge[&], Patrick Gratz[+],
Svetlana Segărceanu[#], George Suciu[#]
+Luxembourg Institute of Science and Technology,
*Romanian Academy Institute for AI, [#]R&D Department, BEIA, Romania
[&]SupWiz, Denmark

{dimitra.anastasiou, patrick.gratz}@list.lu, radu@racai.ro,
{svetlana.segarceanu,george}@beia.ro, a.ruge@supwiz.com

Abstract. This paper describes the project ENRICH4ALL, the aim of which is to develop a Machine Translation (MT)-empowered AI chatbot to be used in public administration in three European countries: Luxembourg, Romania, and Denmark. It is based on Natural Language Understanding (NLU) and Machine Learning language models and covers low resource languages, such as Luxembourgish. During the project, the partners have been developing question-answering datasets in Romanian, English, French, German, Luxembourgish, and Danish in three domains: COVID-19 (Romanian), construction permits (Romanian), and public administration (English, French, German, Luxembourgish, and Danish).

Keywords: government chatbots, multilingual chatbots, public administration.

1 Introduction

Multilingual communication between citizens and public organizations should be a requirement for a digital single market today and part of Europe's digital programs. It is essential in today's globalized economy not only to protect, but also to sustain the EU's linguistic diversity through multilingualism. "Digital by default" is already the new principle for many governments in the EU. Noteworthy is the resolution "Language equality in the digital age", which was passed by the European Parliament in September 2018. In this paper, we describe a project called ENRICH4ALL [1] which targets at lowering language barriers for online services. ENRICH4ALL is about developing a Machine Translation (MT)-powered chatbot and deploying it in public administration in three European countries: Luxembourg, Romania, and Denmark.

Conversational agents are gaining attention and are applied today in many fields, such as e-commerce, education, health, entertainment, and public services to name just a few. According to Gao et al. [2], conversational systems can be grouped into three categories: (1) question answering agents, (2) task-oriented dialogue agents, and (3) chatbots. The history, essential concepts, and classification of chatbots can be found at Adamopoulou & Moussiades [3]. Chatbots can be classified using different parameters: among others, the knowledge domain (open/closed), the service provided (interpersonal/intrapersonal/inter-agent), and the input processing and response generation

method (rule-based/retrieval-based/generative). The chatbot developed in ENRICH4ALL is a closed domain, interpersonal chatbot following a retrieval-based model. Today chatbots have evolved into "virtual personal assistants" and are mainly developed by Google, Amazon, Facebook, Apple, and Microsoft (GAFAM). In the next subsection, we will have a look on e-government chatbot infrastructure currently existing in Europe.

As far as e-government chatbots is concerned, the European Commission has a strategy on e-government in the digital single market concerning the electronic exchange of social security information, electronic payments & invoicing, etc. E-government chatbots are an essential AI application in advancing e-government and facilitating communication between citizens and public services. There are indeed many challenges that prevent governments of deploying chatbots, such as the large number of relevant services, the complexity of administrative services, the context-dependent relevance of user questions, the differences in expert-language and user-language as well as the necessity of providing highly reliable answers for all questions [4]. In the EU and CEF (Connecting Europe Facility) Associated Countries, it is in its infancy. However, in 2019 the Directorate-General for Informatics (DIGIS) has published a document containing the components of a high-level architecture for public service chatbots. A multilingual chatbot, enabling citizens to ask their questions in their preferred language, is a much-needed AI application in the e-government infrastructure. With a more positive glimpse, the COVID crisis has accelerated digitization and there are available chatbot services to combat COVID (and not only); to name just a few examples: HealthBuddy+[1] (a multilingual chatbot developed by UNICEF ECARO and WHO/Europe), WienBot[2] (Austria), Suve[3] (Estonia). Hoehn and Bongard-Blanchy [5] evaluated the usability of 24 COVID-19 chatbots to answer the research questions: what types of COVID-19 chatbots exist and how usable they are. They suggest that conversational e-health applications would be more attractive to users if they invest in UX from the beginning.

2 ENRICH4ALL Project

The ENRICH4ALL project is funded by Connecting Europe Facility CEF-TC-2020-1: Automated Translation and its duration is 2 years (June 2021 – May 2023). The project is coordinated by LIST (Luxembourg Institute of Science and technology); the other three partners are: i) BEIA Consult International [6], ii) Research Institute for Artificial Intelligence, Romanian Academy [7], iii) SupWiz Aps [8]. In the ENRICH4ALL project, we develop a multilingual chatbot using MT, which to our knowledge is the first multilingual bot in the domain of public administration. In ENRICH4ALL, we are using the AI-powered chatbot named *BotStudio*, developed by the Danish company and project partner SupWiz, which now integrates with the *eTranslation* API [9]. *BotStudio*

[1] http://healthbuddy.plus/index#webchat

[2] https://www.wien.gv.at/english/bot/index.html

[3] https://investinestonia.com/estonia-created-suve-an-automated-chatbot-to-provide-trustworthy-information-during-the-covid-19-situation/

can use Natural Language Understanding (NLU), fine-tuned, BERT[4]-like models to appropriately map user intents to developed chat nodes in specific domains. The infrastructure of the *eTranslation* integration and the workflow including both the front-end and back-end can be seen in Figure 1 below.

2.1 Infrastructure

The infrastructure of our chatbot service can be seen in Figure 1 below. The chatbot user selects the language of the query in the chatbot window, then *eTranslation* translates the query into the language of the datasets that we have developed. If we have a dedicated monolingual chatbot for this language, e.g. in Romanian, then there is no need to go through *eTranslation*, but for all other languages, the *eTranslation*-integrated bot will be used. The chatbot accesses the query in our predefined QA datasets with alternative questions, which have been trained with the NLU models respectively, then *eTranslation* translates the answers as well and provides these as an output to the user. We have been preparing monolingual chatbots in Romanian, Danish, German, French, English, and Luxembourgish. Danish is the official language of Denmark, which has approx. 5.831 million inhabitants. Approx. 90% have Danish as their mother tongue[5]. Romanian is the mother tongue for 85% of the population in Romania, followed by Hungarian (6%) and Romani (1%) according to the 2011 census[6]. Luxembourg is a highly multilingual country with Luxembourgish as the national language, French as the legislative language, and French, German and Luxembourgish as the three administrative and judicial languages. Luxembourgish has received an official status only since 1984, and moreover, is still not an official language of the EU.

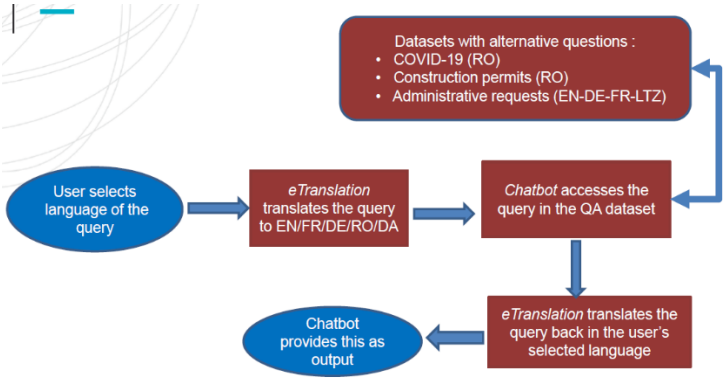

Fig.1. Integration of eTranslation in a chatbot infrastructure

[4] Bidirectional Encoder Representations from Transformers

[5]https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_9__Language_Report_Danish_.pdf

[6] https://www.recensamantromania.ro/rpl-2011/rezultate-2011/

eTranslation. *eTranslation* is the neural MT tool provided by the European Commission to all EU bodies, public services, and public administrations across EU, Iceland and Norway, as well as European SMEs and startups. It currently covers not only the 24 official languages of the EU, but also Russian and simplified Chinese, Turkish, Arabic, and Ukrainian. It should be noted that Luxembourgish is currently not supported in *eTranslation*, i.e. a Luxembourgish monolingual chatbot will be deployed for our use case in ENRICH4ALL.

eTranslation is available both as a stand-alone web service and as an API that can be integrated into other online services. One significant benefit of *eTranslation* over other MT solutions is data privacy preservation. Personal data security is an essential requirement for the deployment and viability of e-government chatbots.

In ENRICH4ALL, *BotStudio* and the live chat solution *SupChat* have been integrated with *eTranslation* via the available API with a particular focus on ensuring real time communication with real time translation. The chatbot developed in the project ENRICH4ALL just uses *eTranslation*, and therefore the ENRICH4ALL team is not responsible for the quality of the translation. The users of the chatbot are informed through a disclaimer that the chatbot is MT-enabled. In addition, in the chatbot widget, the user can report an incorrect translation and also see the original English version.

2.2 Matching

Central to all chatbots is the ability for a node to "match on" what the user writes. Matching can be done in one of two ways: i) providing literal examples of queries the user writes or ii) using a NLU model. An NLU model is trained on real sample-data from users' queries, and as such typically has a better understanding of what the user means. In *BotStudio*, one can upload a fine-tuned BERT language model and use it to label input questions so that the label maps onto the desired dialog node. Users can add labels and training questions for each label and *BotStudio* uses the fine-tuned BERT model to learn a sequence classifier to the label set.

To enable such functionality in *BotStudio*, we must train and/or fine-tune BERT models for the datasets of interest. Luxembourgish did not have any BERT models and thus, we have created one from scratch[7] [10]. In ENRICH4ALL we need targeted datasets, so that we can fine-tune BERT models for the project's languages and domains of interest. We chose three domains of interest to develop and test our multilingual chatbot: COVID-19 (in Romanian), construction permits (in Romanian) and administrative questions (in English, French, German, Luxembourgish, and Danish).

Each QA dataset is organized into question groups, each group having a unique ID and containing multiple formulations of the same question. Each question group contains a single answer that is valid for any question formulation in the group. Table 1 lists the average number of formulations per question group, the number of groups in the QA dataset, and the number of all questions in the QA dataset.

[7] https://huggingface.co/raduion/bert-medium-luxembourgish

Table 1. QA datasets statistics

	Average alternatives	QA groups	Total questions
COVID-19	16.5	168	1123
Construction permits	78.5	28	2200
Administrative issues	16.5	93	2079

Based on our datasets described in Table 1, we have run experiments on question similarity accuracy as well as extractive QA [10, 11]. We fine-tuned the multilingual BERT as well as Romanian BERT [12, 13] and Luxembourgish BERT [10] for the Masked Language Model (MLM) task using the corpora mentioned in Table 1. We measured the accuracy of labeling a question with the correct label from the QA dataset label set and the accuracy of correctly retrieving the ID of the question group (with at least two formulations), out of which one formulation is taken as the test input question. To evaluate question similarity, given an input question from a question group that has at least two formulations, we aimed at recovering the ID of the parent question group. To achieve this, we fed the BERT model the input question and used the last hidden state tensor output to calculate a cosine similarity between the input question and all other questions in the QA dataset. As far as extractive QA is concerned, Ion et al. [11] developed an open-domain QA system which executes the following operations, in sequence, for an input question: question processing, query generation, answer mining. The chatbot only takes the input question and automatically searches for the relevant documents on the web but, for the answer selection, it employs a fine-tuned BERT model for Extractive QA that, using the input question together with the snippet that the web search engine produces for each relevant document, highlights the answer to the input question.

3 Future Prospects

One of our main future goals is to deploy our chatbot in public administration in Luxembourg, Romania, and Denmark. After we identified the public partners, we have to organize a collaboration workflow between the project partners and the public partners, including, among other agreements about data privacy and security, data sharing, evaluation of user requests, and dissemination. After deploying the chatbot, the user questions will be added to our datasets, so that we can further fine-tune our BERT models. This will improve the performance of the chatbot because it will be trained with more domain-specific data.

As far as research-oriented future prospects is concerned, we are interested in comparing various language models and train them with our specific datasets. On a

more general note, building language technologies for low resource languages, such as Luxembourgish, based on sociolonguistic insights [14] is one of our future prospects.

4 Acknowledgments

The Action 2020-EU-IA-0088 has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278547.

References

1. ENRICH4ALL: https://www.enrich4all.eu/, last accessed 2022/09/30.
2. Gao, J., Galley, M., & Li, L. Neural approaches to conversational AI. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1371-1374, (2018).
3. Adamopoulou, E. & Moussiades, L. An overview of chatbot technology. *IFIP International Conference on Artificial Intelligence Applications and Innovations*, Springer, Cham, 373-383, (2020).
4. Lommatzsch, A. A next generation chatbot-framework for the public administration. *International Conference on Innovations for Community Services*, Springer, Cham, 127-141, (2018).
5. Höhn, S., & Bongard-Blanchy, K. Heuristic evaluation of COVID-19 chatbots. In *International Workshop on Chatbot Research and Design*, Springer, Cham, 131-144, (2020).
6. https://www.beiaro.eu/, last accessed 2022/09/30.
7. https://www.racai.ro, last accessed 2022/09/30.
8. https://www.supwiz.com/, last accessed 2022/09/30.
9. https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation, last accessed 2022/09/30.
10. Anastasiou, D., Ion, R., Badea, V., Pedretti, O., Gratz, P., Afkari, H., Maquil, V. & Ruge, A. ENRICH4ALL: A first Luxembourgish BERT Model for a Multilingual Chatbot. *Proceedings of the ELRA/ISCA Special Interest Group on Under-Resourced Languages* (SIGUL 2022), (2022).
11. Ion, R., Avram, A. M., Păiș, V., Mitrofan, M., Mititelu, V. B., Irimia, E., & Badea, V. An Open-Domain QA System for e-Governance. *Proceedings of the Fifth International Conference on Computational Linguistics in Bulgaria (CLIB)*, (2022).
12. Avram, A.M., Catrina, D., Cercel D.C., Dascălu, M., Rebedea, T., Păiș, V., & Tufiș, D. *Distilling the Knowledge of Romanian BERTs Using Multiple Teachers*. arXiv:2112.12650v1, (2021).
13. Dumitrescu, S.D., Avram, A.-M., & Pyysalo, S. *The birth of the Romanian BERT*. arXiv:2009.08712v1, (2020).
14. Doğruöz, A.S., & Sitaram, S. (2022). Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights. *Language Resources and Evaluation Conference (LREC)*, 92-97, (2022).