ENRICH4ALL: A first Luxembourgish BERT Model for a Multilingual Chatbot

Dimitra Anastasiou[+], Radu Ion*, Valentin Badea*, Olivier Pedretti[+],
Patrick Gratz[+], Hoorieh Afkari[+], Valerie Maquil[+], Anders Ruge[&]

[+]Luxembourg Institute of Science and Technology,
*Romanian Academy Institute for AI, [&]SupWiz
[+]5 avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, *13 Calea 13 Septembrie, Bucharest 050711,
[&]Vesterbrogade 35, 1620 Copenhagen
{dimitra.anastasiou, olivier.pedretti, patrick.gratz, hoorieh.afkari, valerie.maquil}@list.lu,
{radu, valentin.badea}@racai.ro, a.ruge@supwiz.com

Abstract

Machine Translation (MT)-powered chatbots are not established yet, however, we see an amazing future breaking language barriers and enabling conversation in multiple languages without time-consuming language model building and training, particularly for under-resourced languages. In this paper we focus on the under-resourced Luxembourgish language. This article describes an experiment we have done with a dataset containing administrative questions that we have manually created to offer BERT QA capabilities to a multilingual chatbot. The chatbot supports visual dialog flow diagram creation (through an interface called *BotStudio*) in which a dialog node manages the user question at a specific step. Dialog nodes can be matched to the user's question by using a BERT classification model which labels the question with a dialog node label.

Keywords: administrative questions, BERT, chatbot, eTranslation, CEF, QA dataset, Luxembourgish

1. Introduction

This paper discusses our own solution of an AI chatbot powered with *eTranslation[1]*, the Machine Translation (MT) system of the European Commission. Since we are developing a conversational chatbot answering user open-ended questions, Natural Language Understanding (NLU) plays an indispensable role in chatbot dialogue management (see Namazifar et al., 2020). In this paper, we describe our work-in-progress in creating a new BERT model for Luxembourgish to drive question classification and answering.

This work has been done within ENRICH4ALL, a CEF-funded project aiming at a Digital Single Market strategy, which is linked with lowering language barriers for online services and public administration procedures. Our chatbot is an AI-based, MT-powered, fully digital and secure service, which automatically simplifies procedures by providing readily available information to citizens 24/7 and reduces the administrative burden from public authorities. One of the goals of ENRICH4ALL is to deploy the chatbot in public services in the Consortium member countries, Luxembourg, Romania, and Denmark.

The goal of this paper is twofold: fine-tune a BERT model for Luxembourgish for i) question labeling and ii) question similarity. The paper is laid out as follows: In Section 2 we provide a short literature review with subsections on the evolution of chatbots, e-government chatbots as well as on multilingual aspects of chatbots. Section 3 describes BotStudio, our AI-based chatbot and its integration with *eTranslation* as well as one of the challenges of MT-enabled chatbots, which is language detection. Section 4 describes briefly the Luxembourgish language and the multilingual setting in Luxembourg. In section 5 we present our Luxembourgish dataset on administrative questions. The dataset is submitted as resource in LREC repository

and is also freely available at the project's website[2]. In Section 6 we describe our training process of BERT models and in Section 7 we present our results on the aforementioned dataset. We conclude this paper in Section 8 with a few future prospects.

2. Literature Review

We begin this literature review on the history and evolution of chatbots (2.1) from simply answering questions to enabling human-like conversations, narrowing down the available infrastructure of chatbots in general to e-government chatbots (2.2) and multilingual chatbots (2.3).

2.1 Evolution of Chatbots

A chatterbot, chatbot, or simply bot is a software application that conducts an online chat conversation with human beings via text or voice through a messaging interface. The term "Chatterbot" was originally coined to describe conversational programs (Mauldin, 1994). However, the first known chatbot dates back to 1966 and its name was Eliza, whose purpose was to act as a psychotherapist returning the user utterances in a question form (Weizenbaum, 1966).

Today chatbots have evolved into "virtual personal assistants" and are mainly developed by Google, Amazon, Facebook, Apple, and Microsoft (GAFAM). 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. (2018), 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 (2020).

---

[1] Links of services or products are included in Section 10.

[2] https://www.enrich4all.eu/language-resources

The advancement of Machine Learning (ML), and particularly transfer learning has shown huge improvements in Natural Language Processing (NLP). Low code-free or open-source development platforms in combination with limited design efforts for a chatbot interface make chatbot development an easy task for developers. *Chatbot.org* is a comparison resource for chatbot buyers by providing user reviews, and research on thousands of chatbot platforms and solutions.

2.2 E-government Chatbots

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. However, there are certain challenges, 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 (Lommatzsch, 2018). While in the USA and India, government agencies use chatbots, in the EU and CEF (Connecting Europe Facility) Associated Countries, it is in its infancy. Currently, there are a few EU countries, where many e-government chatbots are deployed, whereas in other countries, such as Romania or Luxembourg, there are not. 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.[3].

2.3 Multilingual Chatbots

By multilingual chatbot, we mean that a user can choose to ask their question in their preferred language and the chatbot answers respectively in this language. Multilingual communication between citizens and public administration is a major priority of the EU, as it provides customized services for citizens to facilitate their right to speak and write in their native language. Particularly for administrative procedures, there are many requests from citizens who enter a new country. Application for residence, importing a car, start-up a new business, family allowances, etc. are some of such requests. Multilingual bots and guides on how to create them are coming up increasingly in the last few years (Janarthanam, 2017; Boonstra, 2021), but also mainly by the industry and their business solutions.

Many multilingual bots are used for foreign language learning, such as Mondly (supporting 41 languages). Lothritz et al. (2021) tested two strategies for implementing a multilingual chatbot: (S1) For *n* languages, employ *n* chatbots, each of which is trained to handle requests in a single language. (S2) For *n* languages, employ one chatbot which is trained using data written in *n* languages. They compared these two strategies for chatbots in a multilingual environment on two tasks that represent Intent Classification and Slot Filling. They found that in the case of two languages, the combination of a language selector and two monolingual chatbots (S1) usually outperforms chatbots that are directly trained on bilingual datasets (S2).

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. This chatbot is called *BotStudio* and is described in Section 3 below.

3. BotStudio

In ENRICH4ALL, we are using the AI-powered chatbot named *BotStudio*, developed by the Danish company SupWiz, which now integrates the eTranslation API. The BotStudio chatbot has the ability for a node to "match on" what the user writes. This matching can be done either by providing examples of possible user queries or through the usage of an NLU model which is trained on real sample-data from users' queries. BotStudio can use fine-tuned, BERT-like models to appropriately map user intents to developed chat nodes in specific domains.

eTranslation is the neural Machine Translation (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, simplified Chinese, Turkish, and Arabic. eTranslation is a CEF building block that can be integrated into digital services to add translation capabilities.

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, for a government chatbot, 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 multilingual BotStudio chatbot uses eTranslation to automatically translate incoming questions into the language of the QA model and outgoing answers into the language of the user. However, the eTranslation API has not been used for the experiment described in this paper, so it is outside of its scope.

In order to automatically select the translation engine, a language identifier algorithm is needed and we adapted[4] Python's `langdetect`[5] package to the needs of our project. We built a custom Docker container[6] that serves language identification services to the caller, for the languages of the project. Luxembourgish was not supported by the latest distribution of `langdetect` (1.0.9) and thus, we have added it by training `langdetect` on a Luxembourgish Web-based corpus (Leipzig Corpora Collection) containing 1M sentences and more than 16M

words; the used text material was taken from randomly chosen Web sites.

4. Luxembourgish Language

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. The vocabulary of Luxembourgish has many loan words from French and German, the morpho-syntax follows Germanic patterns. According to the STATEC (as of May 2019), French is the most spoken language at work (78%), followed by English (51%) and Luxembourgish (48%). Luxembourgish is the most widely spoken language at home (53%), followed by French (32%) and Portuguese (19%). Luxembourgish is a low-resourced language when it comes to the availability of language resources or tools.

The latest version of the official Luxembourgish orthography can be found at the Zenter fir d'Lëtzebuerger Sprooch (ZLS)/Centre for the Luxembourgish Language and also downloaded as a PDF file[7]. The Luxembourgish orthography officially regulates the spelling of the Luxembourgish language for the areas for which the Luxembourg State is responsible (administrations, schools). The codification and subsequent implementation of orthography in Luxembourgish can be found at Gilles (2015). More information on the languages spoken in Luxembourg can be found at Lulling et al. (2010).

However, the focus on the Luxembourgish language has increased during the last few years both from the governmental side with its long-term strategy and the research side, as a consequence. On the one hand, the government aims at increasing the importance of Luxembourgish by advancing the standardization, use and study of Luxembourgish, promoting learning Luxembourgish and the Luxembourg culture, and promoting culture in the Luxembourgish language. The ZLS contributes to the realization of the government policy on the Luxembourgish language. On the other hand, we see that in the last few years, many research projects focus on Luxembourgish (Lingscape[8], Schnëssen[9]); both of these projects are based on crowdsourcing. This is an excellent example about creating large spoken, image, or written corpora quickly and by diverse users, which can contribute to developing language technology applications.

5. Luxembourgish Dataset on Administrative Questions

As in many countries in the EU, e-government and digitalization are managed by dedicated institutions. In Luxembourg, the Ministry for Digitalization was created on December 11th, 2018 and in Romania, this is the newly established Authority for the Digitalization of Romania. These authorities have helped the ENRICH4ALL project to become a reality and the language resources output of ENRICH4ALL will be fed into the European Language Grid project (Rehm et al., 2020), in which the Luxembourg Institute of Science and Technology and Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăgănescu" are also partners.

In ENRICH4ALL we need targeted datasets, so that we can fine-tune BERT(-like) 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 Luxembourgish). In this paper, we focus only on Luxembourgish.

Concerning the citizen´s online services with the State, Guichet.lu is the information portal in Luxembourg that simplifies citizen's exchanges with the State and offers them quick and user-friendly access to all the information, procedures and services offered by Luxembourg public bodies. The website of Guichet.lu is available in German, English, and French, but not in Luxembourgish.

We have manually created a set of 135 questions with their corresponding answers in Luxembourgish based on Guichet.lu. The questions cover questions about passport, asylum, or certificate requests (see Q1 example in Table 1, below), but also questions that a newly arrived person in Luxembourg might ask, e.g., about the minimum wage (Q2), unemployment rates, school enrollment, etc. Since there are many commuters working in Luxembourg, but living in neighboring countries, we also collected questions relevant to paying taxes in Luxembourg, while living in France or Germany.

Most questions (83%) are *wh*-questions, i.e. starting with Where/When/Whom/How, while 6% are in statement form (see Q3, Table 1). 11% of the questions include both a statement and a *wh*-question. The size of the questions varies from 4 to 15 words.

Q1	Wou muss ech d'Gebuert vu méngem Bebé umellen?	Where should I declare the birth of my baby?
Q2	Wat ass de soziale Mindestloun zu Lëtzebuerg?	Which is the minimum wage in Luxembourg?
Q3	Ech wëll fir e Pass rembourséiert ze kréien.	I want to be reimbursed for a passport.

Table 1: Examples of questions in Luxembourgish and their English translation[10]

This corpus is multilingual (LTZ-EN-DE-FR); we plan further experimentation in future months (see Section 8).

---

6. Training/Fine-tuning of BERT Models

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 for the label set.

To enable such functionality in BotStudio, we must train and/or fine-tune BERT models for the datasets of interest. An additional reason for using BERT models for such a task is to save many hours of manual work creating alternative/synonym sentences and manual labeling of these sentences. Luxembourgish did not have any BERT models and thus, we have created one from scratch. In the next subsection, we detail the training and fine-tuning of Luxembourgish models. In our experiment, we made a comparison between the fine-tuned version of the `bert-base-multilingual-cased` BERT model (Devlin, 2018) as the baseline model and a language-specific, fine-tuned BERT model called `luxmed`.

6.1 The Luxembourgish BERT model for administrative questions

Luxembourgish is a low-resourced language and it is a big challenge to train a standard BERT model for it. According to Wu & Dredze (2020), the multilingual BERT model covers 104 languages and the 30% of languages with the least pretraining resources perform worse than those using no pretrained language model at all.

To bring Luxembourgish among the languages with at least a BERT model and to benefit from language-specific fine-tuning for our evaluated tasks, we proceeded to train a Luxembourgish BERT medium model from scratch, using the 16M Luxembourgish Web-based corpus (Leipzig Corpora Collection) containing 1M sentences and more than 16M words; the used text material was taken from randomly chosen Web sites. This model is available on HuggingFace[11] and can be readily used with the transformers Python API. It was trained for 3 epochs and it reached a final perplexity of 58.76 on the validation set. It has 8 encoder blocks, the size of the hidden layer is 512 and it uses 16 attention heads. The vocabulary has 70K word pieces.

The fine-tuning for administrative questions labeling (see Table 3 below) was done by varying the epochs number (10, 50, 100, 200, 400), the batch size (8, 16, 32), learning rate (5 to 1e-5), and learning rate decay rate (polynomial decay with a learning rate decreasing with `step = size(trainset) * epochs`). The data used for training was 80% of the dataset and the rest for validation. The best results were obtained with 200 epochs, a batch size of 16, and starting learning rate of 1e-5. The whole training process was done using the Tensorflow version of HuggingFace. In what follows, we will refer to the Luxembourgish BERT medium model as `luxmed`.

7. Results

In this section we will evaluate the Luxembourgish fine-tuned BERT models' ability to label input questions with the appropriate label and the ability to find the most similar question to the input question from the train set.

7.1 QA datasets statistics

The dataset has been transformed into JSON objects which are available on GitHub[12]. 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 2 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.

	Average alternatives	QA groups	Total questions
Administrative questions	1.5	93	135

Table 2: QA datasets statistics

7.2 Task evaluation

For our QA dataset, we will provide the following accuracy figures:

a. The accuracy of labeling a question with the correct label from the QA dataset label set;

b. 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, as explained next.

Figure 1 shows the label frequency and distribution in our dataset. 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. The ID of the group in which the most similar QA dataset question is found is the ID we are looking for. If this ID matches the question group ID from which the input question was extracted, we get one accuracy point.
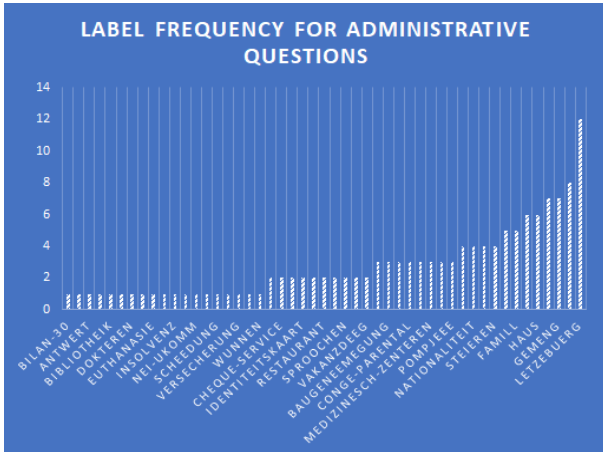

Figure 1: Label distribution for the Administrative questions dataset. The most frequent appears 12 times while there are 19 labels with a count of 1. There are 49 distinct labels in total.

To compute the performance of question similarity, we had to trim the QA datasets and remove all question groups in which a single formulation existed. We ended up with 22.2% of the administrative questions QA dataset.

To optimize the computation time, we introduced an early stop condition: if cosine similarity is over 95%, we assume a very similar question and stop the search for a better one. With this optimization, the whole accuracy calculation time was reduced from 12h to 6h, using an i5-10400 CPU.

We then took the fine-tuned version of the `bert-base-multilingual-cased` BERT model (Devlin, 2018) as the baseline model and we compared accuracy figures against `luxmed` which is a language-specific, fine-tuned BERT model.

Table 3 shows that the multilingual BERT model (`mling`) and the Luxembourgish BERT model (`luxmed`) gave the same accuracy when it came to question labeling. This can be justified by the small size of the administrative questions dataset (135 questions), coupled with the high number of labels (49). In Figure 1 we can see that 19 labels appear only once in our Administrative questions dataset.

When it comes to question similarity accuracy, using the language-specific `luxmed` BERT model is a better choice than using the generic multilingual BERT model (`mling`).

	mling	luxmed
Question labeling accuracy	40.7%	40.7%
Question similarity accuracy	23.3%	26.6%

Table 3: Question labeling and question similarity accuracy with `mling` vs. `luxmed` BERT models

8. Conclusion and Future Prospects

Multilingual communication between citizens and public services should be a requirement for a digital single market. Chatbots are completely missing in the public

administration in Luxembourg, a highly multilingual country. A multilingual chatbot, enabling citizens to ask their questions in their preferred language, is a much-needed AI application in the e-government infrastructure.

The project ENRICH4ALL aims at deploying an MT and AI-enabled chatbot in public services in Luxembourg. Luxembourgish is an under-resourced language, and in addition, is not supported in eTranslation. Within the project ENRICH4ALL, we can overcome these limitations by using the new BERT models we trained for it.

We tested pre-trained and fine-tuned BERT models for question labeling and question similarity. The main limitation of this work was the small size and label imbalance of the QA dataset. In the meantime, we have been adding additional alternative questions under each label.

In the last weeks, we have been extending our Luxembourgish corpus with additional 1,700,000 sentences. We plan to train and validate another medium BERT size model from scratch using this extended corpus data in the coming weeks. Testing with data of similar languages is also among our future prospects. We expect that a subsequent fine-tuning with the improved QA dataset will mitigate the current limitations and yield improved results.

In the coming months we plan to deploy our chatbot in public administration in Luxembourg. Having user interaction logged will result in real user questions that will be added to our existing QA datasets. This will improve the performance of the chatbot, and we will have more data to fine-tune our BERT models. After chatbot deployment, we will analyze user feedback, which will be collected at the end of each conversation. We will calculate the amount of user questions, most used questions as well as the success rate per question.

9. Acknowledgements

10. Bibliographical References

Adamopoulou, E. and Moussiades, L. (2020). An overview of chatbot technology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, Springer, Cham, pp. 373-383.

Boonstra, L. (2021). Creating a multilingual chatbot. In *The Definitive Guide to Conversational AI with Dialogflow and Google Cloud,* pp. 187-194.

Devlin, J. (2018). Multilingual bert readme document.

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

European Language Grid: https://www.european-language-grid.eu/, 14.01.22

ENRICH4ALL: https://www.enrich4all.eu/, 13.01.22

eTranslation: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation, 13.01.22

eTranslation Web Service: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/How+to+submit+a+translation+request+via+the+CEF+eTranslation+webservice, 13.01.22

Janarthanam, S. (2017). Hands-on chatbots and conversational UI development: build chatbots and voice user interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills. Packt Publishing Ltd.

Gao, J., Galley, M., and Li, L. (2018). Neural approaches to conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval,* pp. 1371-1374.

Gilles, P. (2015). From status to corpus: Codification and implementation of spelling norms in Luxembourgish. In *Language planning and microlinguistics*, pp. 128-149.

Guichet.lu: www.guichet.lu, 13.01.22

Leipzig Corpora Collection: Luxembourgish Wikipedia corpus based on material from 2021.

Leipzig Corpora Collection. Dataset. https://corpora.uni-leipzig.de?corpusId=ltz_wikipedia_2021.

Lommatzsch, A. (2018). A next generation chatbot-framework for the public administration. In *International Conference on Innovations for Community Services,* Springer, Cham, pp. 127-141.

Lothritz, C., Allix, K., Lebichot, B., Veiber, L., Bissyandé, T. F., & Klein, J. (2021). Comparing multilingual and multiple monolingual models for intent classification and slot filling. In *International Conference on Applications of Natural Language to Information Systems,* Springer, Cham, pp. 367-375.

Lulling, F. S. (2010). Lëtzebuergesch: la langue nationale du Grand-Duché de Luxembourg. *Lengas. Revue de sociolinguistique*, (60).

Mauldin, M. (1994), ChatterBots, TinyMuds, and the Turing Test: Entering the Loebner Prize Competition. In *Proceedings of the 11th National Conference on Artificial Intelligence*. https://www.reginamaria.ro

Mondly: https://www.mondly.com/

Namazifar, M., Papangelis, A., Tur, G., & Hakkani-Tür, D. (2021). Language model is all you need: Natural language understanding as question answering. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 7803-7807.

SupWiz: https://www.supwiz.com/, 14.01.22

Tutor Mike: https://www.rong-chang.com/tutor_mike.htm, 14.01.22

Weizenbaum, J. (1966). ELIZA–a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, pp. 36-45.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual BERT? arXiv preprint arXiv:2005.09093Language Resource References.