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[&]

www.enrich4all.eu/

+Luxembourg Institute of Science and Technology / 5 avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette,
*Romanian Academy Institute for AI / 13 Calea 13 Septembrie, Bucharest 050711,
& SupWiz / 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

ENRICH4ALL

CEF Funded project Automated Translation (CEF-TC-2020-1: Automated Translation)

Partners

- Luxembourg Institute of Science and Technology
- Romanian Academy Institute for AI
- BEIA Consulting Romania
- SupWiz

Duration

01.06.2021 – 30.05.2023

Objectives

- Development of a multilingual **e-government chatbot** which lowers the language barriers in EU.

- Deployment in public administration in
  - **Luxembourg**
  - **Romania**
  - **Denmark**

Benefits of e-Government Chatbots

- Available 24/7
- Can reach large amounts of people
- Irrespective of
  - age
  - gender
  - educational background
  - geographical barrier
  - LANGUAGE

Languages in Luxembourg

Luxembourgish is the national language, French the legislative language, and French, German and Luxembourgish as the three administrative and judicial languages.

At work:
- French is (78%)
- English (51%)
- Luxembourgish (48%)

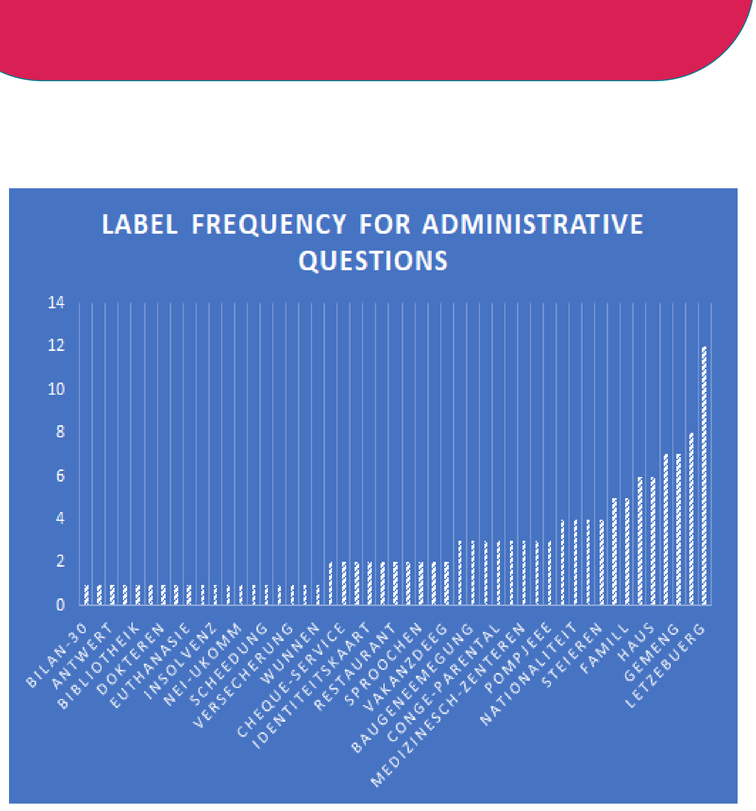
At home:
- Luxembourgish (53%)
- French (32%)
- Portuguese (19%)

Language Resources at ENRICH4ALL

- Three QA datasets
  1. COVID-19 (RO)
  2. Construction permits (RO)
  3. Administrative questions (LTZ–FR–DE–EN) under CC-BY-SA-4.0 license.

- Luxembourgish BERT baseline model for question labelling and similarity

- Language identification



BERT medium size model for LTZ (luxmed), created from a dataset with 1M Luxembourgish sentences from Wikipedia. Parameters L=8 and H=52, vocabulary has 80K word pieces.

Train loss	4.230
Train perplexity	68.726
Validation loss	4.074
Validation perplexity	58.765

Task evaluation

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.

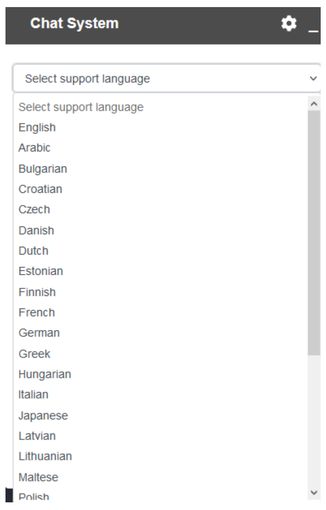
	Average alternatives	QA groups	Total questions
Administrative questions	1.5	93	135

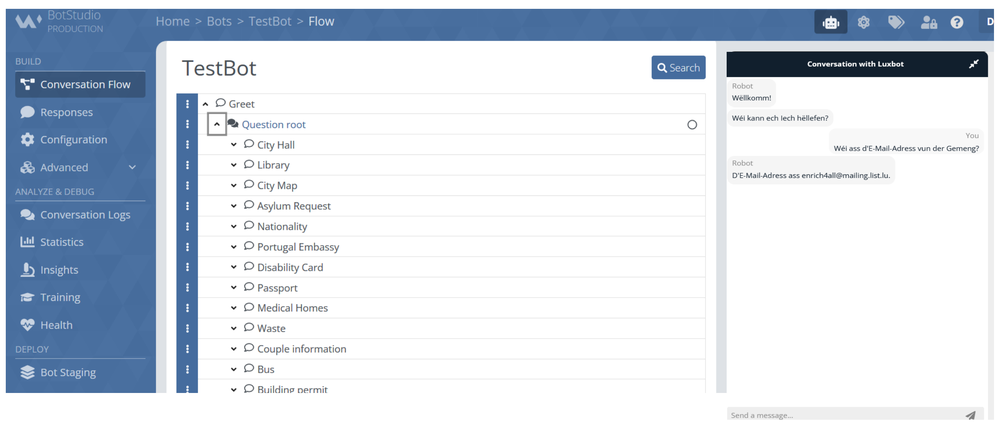
	mling (Multilingual BERT)	luxmed
Question labeling accuracy	40.7%	40.7%
Question similarity accuracy	23.3%	26.6%

Machine Translation system of CEF eTranslation

- Neural MT tool by the EC to EU bodies, public services, public administrations and SMEs for 24 official languages of the EU, plus Chinese, Russian, Turkish, Arabic, but no support of Luxembourgish.

- Building Block you can integrate into your digital systems, if you need to embed them with translation capabilities.





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.

Anastasiou, D., Ruge, A., Ion, R., Segărceanu, S., Suciu, G., Pedretti, O., Gratz, P. & Afkari, H. (2022, June). A Machine Translation-Powered Chatbot for Public Administration. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 327-328).

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

Luxembourg Institute of Science and Technology | 5, avenue des Hauts-Fourneaux, L-4362 Esch/Alzette | phone: (+352) 275 888 - 1, fax: (+352) 275 885 | dimitra.anastasiou@list.lu, **LIST.lu**

Co-financed by the Connecting Europe Facility of the European Union