Computers learn Maltese

Researchers have trained a language model on Maltese textual data and taught it to identify the sentiment and tag names in a text.

DIGITAL ECONOMY

© graphicwithart/Shutterstock.com

AI-based language technologies are opening new paths for digital communication in all European languages. However, language technology tools and resources are lacking for Maltese. The EU-funded LT-BRIDGE project has been working to bridge this gap since its launch in 2021.

In natural language processing, language models are trained to associate words with others in a particular context using neural network approaches. Researchers from LT-BRIDGE project coordinator University of Malta (UM) have trained such a model – BERTu – on Maltese textual data.

Filling in the blanks

However, what exactly are language models? UM PhD student Kurt Micallef describes them in a recent article posted on the 'Times of Malta' website: "Language models are an abstract understanding of a language. You can think of this as an 'intuition' of what a language is. For example, if you had to fill in the blank in the sentence 'Jien _____ il-gazzetta' (I ____ the newspaper), you might come up with 'qrajt' (read) or 'xtrajt' (bought), but you are less likely to suggest 'kilt' (ate) or 'karozza' (car)."

One way to train such language models is using masked language modelling. Words in a text are randomly masked, or covered, and the model has to predict the missing word. "So given the example above, the model should ideally predict 'qrajt'," explains Micallef. This is repeated for many sentences so that the language model can learn Maltese. The neural network is updated with every sentence using machine learning

algorithms, and a probability is assigned to possible words that can fit in the sentence.

Other tasks

Two other tasks on which BERTu was trained are sentiment analysis and named-entity recognition. “Sentiment Analysis is the process of identifying the sentiment of a given text,” notes the researcher in another ‘Times of Malta’ article. “The simplest form is classifying whether a piece of text conveys a positive or negative sentiment with respect to some topic or concept. For example, given Malta’s budget announcements, is this comment supportive or against the announcements made? This type of task is called a classification problem, because for the text we get as input we output a classification label (positive or negative in this example).”

Micallef further describes the second task: “Named-Entity Recognition is a tagging task, where we output a label for each word in the input text. Given an input text, the task is to classify which labels are referring to named entities and what type of entity they are. Compared to sentiment analysis, this task is quite low-level, and would typically be used to complement other language systems. For example, we could use the classification data to identify person names and anonymise them, to abide by data protection laws.”

The research team fine-tuned the pre-trained BERTu model on these tasks by adding an extra layer on top of the model for each task and then running machine learning algorithms on the data set to learn the parameters of the extra layer. BERTu was found to outperform other language models, occasionally by more than 20 %. It is now making it possible to explore more complex language understanding tasks in Maltese. The LT-BRIDGE (“Bridging the technology gap: Integrating Malta into European Research and Innovation efforts for AI-based language technologies”) project ends in December 2023.

For more information, please see:
LT-BRIDGE project website

Keywords

LT-BRIDGE | AI | language | Maltese | language model | BERTu | text | textual data | word

Related projects

PROJECT

“Bridging the technology gap: Integrating Malta into European Research and Innovation efforts for AI-based language technologies”

LT-BRIDGE

7 April 2025

Related articles

NEWS

SCIENTIFIC ADVANCES

Paving the way for AI we can trust

2 November 2022

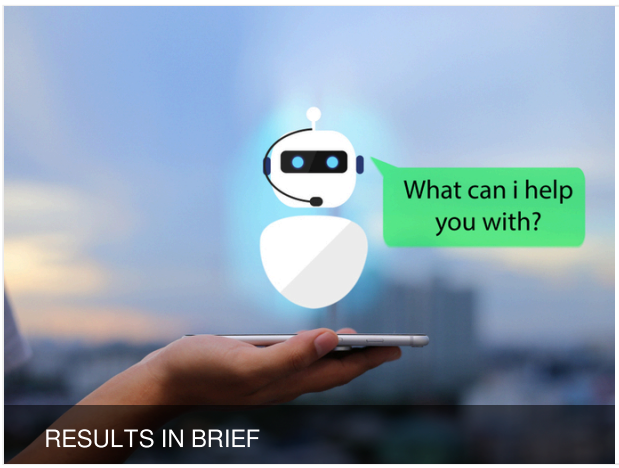
RESULTS IN BRIEF

Personalised AI app teaches foreign languages conversationally

20 March 2020

How machines interpret human language

18 March 2022



Pioneering voice and conversational AI platform boosts business productivity

18 March 2019

Last update: 2 December 2022

Permalink: https://cordis.europa.eu/article/id/442665-computers-learn-maltese