ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

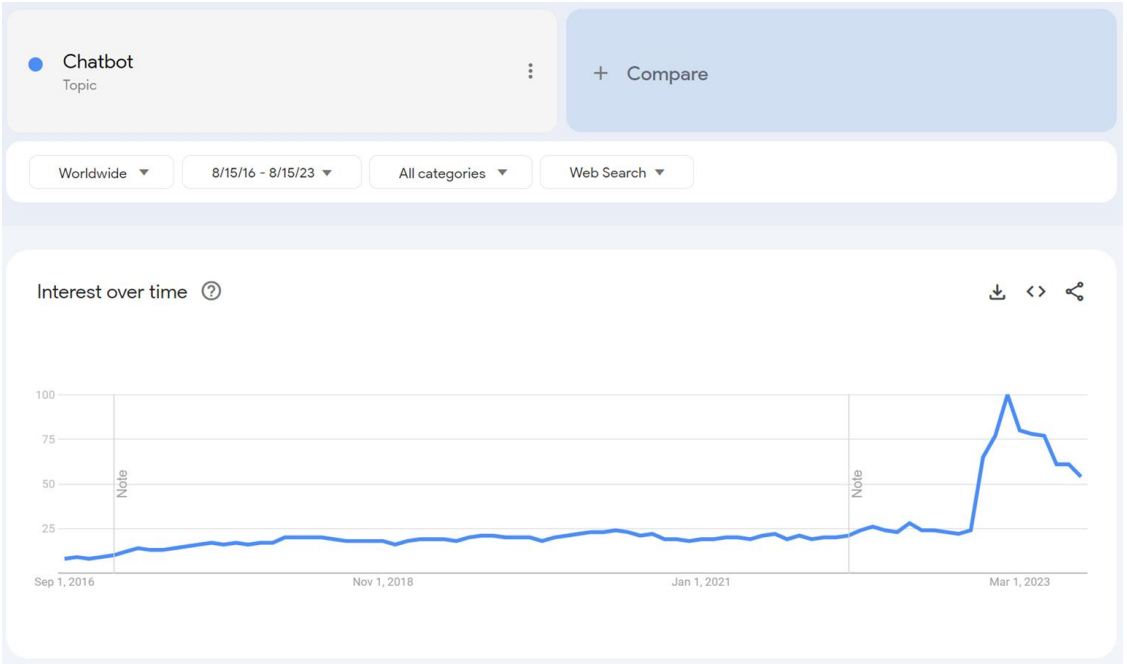
A Review of Greek NLP Technologies for Chatbot Development

Nikolaos Giarelis, Charalampos Mastrokostas, Ilias Siachos, Nikos Karacapilidis

PCI 2023
27TH PAN-HELLENIC CONFERENCE ON PROGRESS IN COMPUTING AND INFORMATICS
WITH INTERNATIONAL PARTICIPATION
November 24-26, 2023, Lamia, Greece

Motivation

The recent advent of OpenAI's **ChatGPT**, a state-of-the-art artificial intelligence (AI) chatbot, has leveraged the public interest and advanced further research in chatbot technology.



Source: https://trends.google.com/trends/explore?date=2016-08-15%202023-08-15&q=%2Fm%2F01305y&hl=en

Major remarks

- Existing surveys mention many successful chatbot applications both in the private (Brachten et al., 2021) and public sector (Androutsopoulou et al., 2019) across many fields.
- The majority of these applications use a **neural classification** architecture to predict the user intent and answer based on pre-determined information (Adamopoulou and Moussiades, 2020).
- However, this leads to the development of chatbots that **fail to understand the nuances** of human communication.
- As reported in (Jain et al., 2018), users prefer chatbots with **'human-like' conversation skills** that offer an engaging experience through a familiar turn-based messaging interface.

From English to Greek NLP

- The integration of recent advancements in Deep Learning (DL), Machine Learning (ML) and Natural Language Processing (NLP) will significantly benefit the current chatbot technology.
- Of particular importance is the utilization of the **Transformer** architecture (Vaswani et al., 2017), and particularly **BERT** (Devlin et al., 2019), which rendered a significant increase in the accuracy achieved in various English NLP tasks.
- With respect to the Greek language, early research work focusing on **Greek NLP** has been thoroughly presented in (Papantoniou and Tzitzikas, 2020);
- Admittedly, compared to widely spoken languages such as English, there exist **much fewer NLP resources** and limited research on the incorporation of DL techniques and models for the Greek language (Evdaimon et al., 2023).

Greek NLP Resources - classical embeddings

- Classical NLP methodologies do not encapsulate textual semantic context. This context can be captured by word embedding models, such as **Word2Vec** (Mikolov et al., 2013) and **FastText** (Bojanowski et al., 2017).
- **spaCy** is an open-source library, which supports more than 73 languages and various NLP tasks. Greek language support was introduced in 2018, through the development of a pre-trained model (el_core_news_lg), which included **FastText** word embeddings, trained on the greek part of Common Crawl and Wikipedia.
- Outsios et al., (2020) build five Greek word embedding models, using **FastText** and **Word2Vec**. Some of these models were trained on a large scale Greek webpages corpus (Outsios et al., 2018), containing 118 million unique sentences, while other ones where trained on the greek parts of Wikipedia and/or Common Crawl.

Greek NLP Resources - transformer-based embeddings

- The **Sentence-BERT** embeddings model (Reimers and Gurevych, 2019), builds on top of **BERT** to create state-of-the-art sentence embeddings, which captures more context than classical embeddings.

Source: https://sbert.net/

- The authors develop this model in an open-source library called **Sentence-Transformers**. In the case of Greek, this library offers a list of multilingual pre-trained embedding models.

- **Sentence-Transformers** can also utilize other Greek transformer-based models to infer sentence embeddings.

Greek NLP Resources - Deep Learning Models

- **GREEK-BERT** (Koutsikakis et al., 2020) is a pre-trained Greek DL model based on **BERT**. This model employs two pre-training learning tasks (Masked Language Modelling and Next Sequence Prediction). This model was trained using a vocabulary of 35,000 tokens and a 29 GB Greek dataset compiled from three sources including:
  - articles from Greek Wikipedia
  - Greek documents collected from European Parliament Proceedings
  - the Greek part of the OSCAR dataset
- **GREEK-LEGAL-BERT** (Athinaios et al., 2020) is a pre-trained Greek DL model created for the NER task on legal texts. For pre-training, a 5 GB dataset containing all legal documents from Greek Legislation was used. This model also utilizes **GREEK-BERT's** training setup.
- Both models achieved similar results, namely a weighted F1 average score of 75% for all predicted NER classes.

Greek NLP Resources - Deep Learning Models

- **GREEK-BART** (Evdaimon et al., 2023) is the first pretrained sequence-to-sequence Greek model based on BART (Lewis et al., 2019). This model is pre-trained on a large 87.6 GB Greek corpus, with a vocabulary of 50,000 sub-words. This corpus comprises the same datasets as **GREEK-BERT** plus the Greek web corpus dataset (Outsios et al., 2018).

- Apart from discriminative tasks (e.g., classification) this model can also be used for **text generation** tasks (e.g., **abstractive** summarization) thanks to the decoder layer of the BART architecture. This is not possible for earlier **BERT-based models**, which do not have this layer.

- The authors fine-tuned this model for abstractive news summarization on a dataset they created called **GreekSUM**, which is a collection of 151,000 news articles from News24/7.gr. The authors report that **GreekBART** achieves better accuracy in most NLP tasks than previous transformer-based models for the Greek language.

Greek NLP Applications - General Language

- **PENELOPIE** - Parallel EN-EL Open Information Extraction - **OIE** (Papadopoulos et al., 2021) is an approach that aims to improve **OIE** for Greek text corpora by employing a multi-step process involving translation, NLP modeling (e.g., coreference resolution, summarization), and triplet extraction.

- Regarding **Greek extractive text summarization** (TS) two open-source libraries exist, (i) **pyTextRank** (Nathan, 2016) and (ii) **sumy** (Belica, 2021). Both libraries implement various extractive TS approaches, that rely on various statistical or graph-based algorithms. These extractive approaches can also be used for the sibling task of **Keyphrase Extraction (KE)**.

- Newer KE approaches build on transformer-based embedding models and mostly support English. **LMRank** (Giarelis et al., 2023) is a transformer-based KE approach that currently supports 14 languages, including Greek.

Greek NLP Applications - Law

- Papaloukas et al., (2021) collect published Greek legislation texts from the Official Government Gazette of the Hellenic Parliament. They use these data to build a legal document classification dataset called **“Greek Legal Codes”**, which contains 47,563 categorized Greek legislation resources that can be classified into 47 legislative volumes, 389 chapters, 2285 subject categories.
- In addition, they use **“Greek Legal Codes”** to evaluate several classification models ranging from traditional ML and RNN-based to Transformer-based ones (i.e., **GREEK-BERT** and **GREEK-LEGAL-BERT**). The results of their experimental evaluation show that the two **BERT** models outperform the other considered ones.
- **GreekLegalSum** (Koniaris et al., 2023) is a Greek legal document summarization dataset. This dataset contains 8395 Court decisions from the Criminal and Supreme Civil Court of Greece along with their summaries. This dataset was collected by web-scraping the **website of Areios Pagos**. Overall, the authors focus on building a dataset that covers a wide variety of legal domains. The authors also use this dataset to fine-tune **GREEK-BERT** for extractive summarization.

Greek NLP Applications - Offensive Language Detection

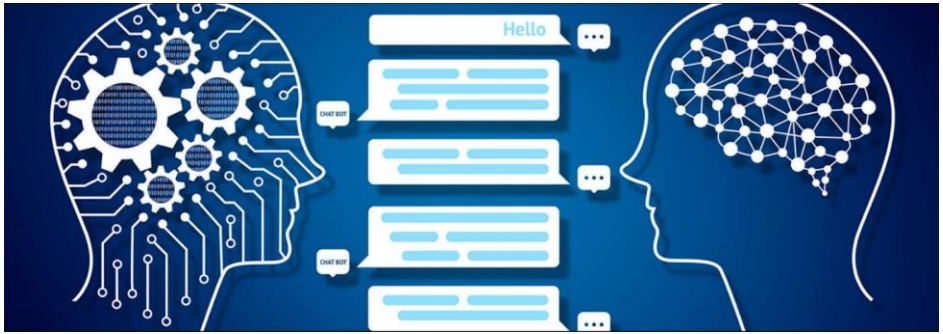
- The **Offensive Greek Tweet Dataset – OGTD** (Pitenis et al., 2020) is a manually created dataset containing 4779 Greek Twitter posts classified as offensive and non-offensive. The dataset was manually annotated by three volunteers, who were given explicit instructions and a proposed definition of offensive language.

- Perifanos and Goutsos (2012) present a hate speech detection approach, combining Computer Vision (image classification networks) and Greek NLP models. Their research focuses on xenophobic and racist Twitter posts directed towards immigrants and refugees. Their proposed approach uses pre-trained embeddings from **GREEK-BERT**, and a dataset of ~23 million Greek tweets generated over a 10-year time period (2008-2018) by 5000 users, to develop a new language model called BERTaTweetGR.

Greek Chatbot Applications - Overview

As mentioned in a recent chatbot survey (Adamopoulou and Moussiades, 2020), the most common techniques incorporated in the development of chatbots are:

- **Rule-based**: chatbots designed around specific rules and constraints.
- **NN-based**: neural network (NN) classification architectures to predict user intention.
- **Knowledge-based:** chatbots that use a knowledge base to infer facts based on user queries.
- **Semantic-based:** chatbots that build ontologies to capture entities and relations (e.g., the Core Public Service Vocabulary – CPSV standard).
- **Deep learning:** chatbots that utilize transformer embeddings and deep learning models for complex NLP tasks.



Source: https://www.analyticsinsight.net/career-insights-all-you-need-to-know-about-nlp-engineers/

Greek Chatbot Applications (1/2)

- Government-Citizen Communication (Androutsopoulou et al., 2019):
  - A chatbot aiming on enhancing digital citizen-government communication channels using chatbots with advanced capabilities.
  - This approach builds on NLP, ML, and data science techniques.
  - Overall, it is elaborated and validated in close cooperation with three Greek governmental bodies (i.e., the Ministry of Finance, a social security agency, and a large local government organization).
- Public Sector Services Chatbot (Karamitsos, 2019):
  - A chatbot aiming to help citizens find public sector services.
  - This approach identifies shortcomings and missing features in existing public sector chatbots.
  - Overall, it incorporates features like real-time data retrieval, multilingual support, and CPSV standard use.
- ERMIS Greek e-Government Chatbot (Stamatis et al., 2020):
  - Four-layered architecture comprising (i) the graphical user interface, (ii) the chatbot engine, (iii) the application programming interface, and (iv) linked data repositories.
  - This approach integrates the method of Life Events for citizen personalization required at different life stages.
  - Overall, this work also demonstrates two usage scenarios for the pilot application along with its evaluation.

Greek Chatbot Applications (2/2)

- PassBot (Antoniadis and Tambouris, 2021):
  - The authors examine several chatbot platforms, as a means of providing citizens with accurate, accessible and personalized information about public services
  - Thus they develop Passbot; a chatbot that provides personalized information regarding the public service of “obtaining a Greek Passport”.
  - Passbot is developed using the CPSV-AP standard for describing public services.
- Chatbot-Knowledge Graph Integration (Patsoulis et al., 2022):
  - The authors focus on integrating chatbot and knowledge graph (KG) technologies.
  - Thus they develop a proof-of-concept chatbot-KG developed for the public service of “Getting a Passport”.
  - Overall, their work achieved positive evaluation results for ease of operation, usefulness, and usability.

Concluding Remarks

- Since the introduction of the transformer and BERT architectures, there is a paradigm shift towards DL in recently published Greek NLP works.
- Admittedly, language resources are, especially in the case of domain-specific applications, quintessential for training robust Greek models that can be integrated into chatbots, and advance their communication skills.
- When a new DL architecture is introduced, offering significant improvement over previous one, researchers should either pre-train it for Greek or fine-tune it for various domains and tasks.
- Most Greek NLP chatbots have been developed to assist citizens accessing services offered by the Greek public sector, while they also do not integrate state-of-the-art DL Greek NLP techniques.

- Future research is required to:
    - Provide more language resources for the Greek language.
    - Fine-tune existing Greek DL models for other domains and tasks.
    - Build chatbots that utilize recent advancements in deep learning for various tasks.

GreekT5

We recently published a pre-print of our recent work on arxiv.org, regarding a series of models for abstractive news summarization, we trained on the GreekSum dataset, which perform better than GreekBART on most evaluation metrics.

https://arxiv.org/abs/2311.07767
https://huggingface.co/IMISLab

Approach	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
TextRank	18.10	5.76	13.84	68.39
GreekT5 (mt5-small)	14.84	1.68	12.39	72.96
GreekT5 (umt5-small)	25.49	12.03	21.32	72.86
GreekT5 (umt5-base)	**26.67**	**13.00**	**22.42**	73.41
GreekBART	17.43	2.44	15.08	**75.89**

Inference API

Summarization

Politics

Αποδεκτή έκανε ο ΣΥΡΙΖΑ, δια στόματος Πόπης Τσαπανίδου την πρόταση του MEGA για διεξαγωγή ντιμπέιτ μεταξύ του Αλέξη Τσίπρα και του Κυριάκου Μητσοτάκη. ""Με χαρά αποδεχόμαστε την πρόσκληση του τηλεοπτικού σταθμού MEGA σε ανοιχτό διάλογο ντιμπέιτ μεταξύ του Αλέξη Τσίπρα και του Κυριάκου Μητσοτάκη. Είμαστε στη διάθεση του καναλιού, όποια μέρα επιθυμεί τροποποιώντας το πρόγραμμα του προέδρου για να γίνει ο δημόσιος διάλογος που θα συμβάλει στην ενημέρωση των πολιτών"", αναφέρει η δήλωση της Πόπης Τσαπανίδου. Υπενθυμίζεται ότι η σχετική πρόταση έγινε από τον Γενικό Διευθυντή Ειδήσεων του καναλιού Σταμάτη Μαλέλη, καθώς ο τηλεοπτικός σταθμός MEGA δεν θα μπορέσει να φιλοξενήσει την τηλεμαχία με όλους τους αρχηγούς, λόγω της προβολής ποδοσφαιρικού αγώνα. Στην πρόσκλησή του ο κ. Μαλέλης ανέφερε χαρακτηριστικά: ""Θεωρούμε ότι η ανταλλαγή απόψεων Μητσοτάκη - Τσίπρα θα συμβάλει στην πληρέστερη ενημέρωση των πολιτών ενόψει των κρίσιμων εκλογών της 21ης Μαΐου"". “Οχι” σε debate μεταξύ του Κυριάκου Μητσοτάκη και του Αλέξη Τσίπρα στο Mega λέει το Μαξίμου, απαντώντας στο ΣΥΡΙΖΑ. Ειδικότερα με δήλωση του ο υπουργός Επικρατείας και κυβερνητικός εκπρόσωπος, Άκης Σκέρτσος, απάντησε στην εκπρόσωπο Τύπου του ΣΥΡΙΖΑ Πόπη Τσαπανίδου: “Η διακομματική επιτροπή, δηλαδή ο συναινετικός θεσμός που εδώ και δεκαετίες ορίζει τις διαδικασίες και τους όρους διεξαγωγής του προεκλογικού αγώνα, αποφάσισε για ένα debate μεταξύ των έξι επικεφαλής των κοινοβουλευτικών κομμάτων την Τετάρτη 10 Μαΐου. Αυτό αποφασίσθηκε ομόφωνα και αυτό θα γίνει. Ο ΣΥΡΙΖΑ μπορεί να προσκαλεί ή να αυτοπροσκαλείται, να επιβεβαιώνει ότι είναι

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 4.682 s

Την πρόταση του MEGA για διεξαγωγή ντιμπέιτ μεταξύ του Αλέξη Τσίπρα και του Κυριάκου Μητσοτάκη σχολίασε η Πόπη Τσαπανίδου, τονίζοντας πως "η δυσανεξία του στους θεσμούς δεν μας εκπλήσσει".

Questions?

Thank you for your attention.

giarelis@ceid.upatras.gr