OPEN WEB SEARCH FOR AI AND NLP IN EUROPE

Jelena Mitrović, Chair of Data Science, University of Passau
Tomáš Mikolov, CIIRC, Czech Technical University
Michael Granitzer, Chair of Data Science, University of Passau

INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) are becoming an increasingly important part of today's businesses. Web search, spam filters and ads recommendation systems rely heavily on AI and ML. However, this "AI Industry" is currently mainly located in Silicon Valley, which creates an imbalanced situation in the research community: while funding, computational resources, and data are all plentiful and available for the researchers located in US and Canada, the situation in Europe is less stellar.

Natural Language Processing (NLP) is an important, and some might argue, the most popular area of AI research and industry nowadays. The goal of NLP is to computationally analyze natural language through simpler, almost accomplished tasks such as part of speech tagging, to complex tasks such as rhetorical figure detection, discourse analysis, or implicit hate speech detection. Most of these tasks are heavily dependant on the presence of well curated data sets, i.e. corpora.

The arguably biggest trend in NLP from the last decade is the adoption of models based on various artificial neural network architectures. Statistical language models (LMs) in particular were dominated for decades by n-gram techniques, to the point that the research community stopped believing it would be possible to significantly overcome these models on large datasets. This changed a decade ago when several important discoveries related to neural language models have been made – especially through open-sourcing these technologies via RNNLM [1] and word2vec [2] projects which were used by many researchers and engineers in companies such as Google, Facebook, and Microsoft as the starting point in using neural models for NLP tasks. While this paradigm shift has been started mainly by researchers from Europe, the neural language models were massively popularized by US companies in recent years.

FUTURE OF NEURAL LANGUAGE MODELS

Recently developed large LMs trained on publicly available data show impressive results. Interestingly, the pretraining dataset for LaMDA [4], e.g., consists of 2.97B documents, 1.12B dialogs, and 13.39B dialog utterances, for a total of 1.56T words. There is an imbalance in research potential at academic institutions because at the moment, access to large amounts of structured, curated data is a reality only for the big companies. When it comes to domain-specific LMs, e.g. LegalBERT [5] or HateBERT [6], it is possible to re-train them successfully on well-curated, comparably smaller corpora. Still, if we wish to bring the performance of these models on par with LaMDA, and have truly valuable specialized models, we should be able to employ frequent and proactive updates using the current Internet data, incorporate the relevant knowledge from social media, and be able to update the performance of our models accordingly - all of which requires a lot of computational power and resources. Furthermore, a proper ethical framework for using the new data for re-training should be at the centre of such efforts. It is not clear how these issues are handled with large LMs at the moment. If we want to build domain-specific language models that can deal with non-literal language, rhetorical figures, negation, humor, we need access to a vast amount of publicly available data, along with literary works, political debates, social media data. Current language models are only partly capable of truly dealing with these phenomena.

CONCLUSION

The chase for building bigger and better AI models is not and should not be seen as the only way forward in science, but the fact that European institutions are lagging behind due to the lack of resources, should not be a defining factor behind our success. Basic research is crucial for progress in AI and NLP, but we also need access to vast amounts of data to test our hypotheses and come to scientific breakthroughs.

It could be argued that to create balance in the current situation, Europe needs to become a leader in areas that are currently being dominated by US monopolies. An example is web search and social networks: while Google and Facebook generate hundreds of billions of dollars in revenue every year, the current business landscape seem to imply that most of the investments are aimed towards the US. This is harmful for the research progress - the scientists no longer compete over having the most innovative ideas, but rather win various benchmarks by using more computational resources to train larger and larger models. This is why having an Open web search paradigm and ultimately an EU-based organisation based on its principles would be an important step towards demonopolization of the AI industry and in creating fairer opportunities in science.

ACKNOWLEDGEMENTS

SPONSORED BY THE
Federal Ministry
of Education
and Research

The project on which this report is based was partly funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01|S20049. The author is responsible for the content of this publication.

REFERENCES

[1] Mikolov et al., Recurrent neural network based language model, INTERSPEECH, 2010.

[2] Mikolov et al., Distributed Representations of Words and Phrases and their Compositionality, NIPS, 2013.

[3] Brown et al., Language Models are Few-Shot Learners, https://arxiv.org/abs/2005.14165

[4] Thoppilan et al., LaMDA: Language Models for Dialog Applications, https://arxiv.org/abs/2201.08239

[5] Chalkidis, Ilias et al., LEGAL-BERT: The Muppets straight out of Law School, ACL, https://aclanthology.org/2020.findings-emnlp.261

[6] Caselli et al., HateBERT: Retraining BERT for Abusive Language Detection in English,Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021).