Dakota Chang

Dr. Zufelt

CSC630

9th November 2021

Bilingual AI: Multilingualism in NLP and How It Relates to Accessibility

Humans have always prided themselves in their invention of language and ways to document stories. However, with the emergence of stronger and faster computers, complex machines are also now closer to achieving the same feat than ever before.

NLP, also known as Natural Language Processing, is a subfield in A.I. that is concerned with the interactions between computers and human language. While it sounds complex, these algorithms are everywhere, Google Translate, chatbots, Amazon Echo, autocorrect, etc. all utilize technologies from this subfield. Through many techniques such as stemming, lemmatization, masking, and word vectorization, computer scientists can now develop algorithms to accurately understand and recreate many aspects of a text, such as sentiment, meaning, tone, etc. However, similar to any other emerging technology, there are many ethical concerns and systemic issues surrounding it, with the lack of support for languages other than English being amongst the most prevalent ones.

Natural Language Processing papers often refer to English as the natural language, which is not only inaccurate but also not inclusive for people whose first language is not English. Computer science has always been Western-centric, especially with A.I. training data often focusing and based on the average upper-middle-class white men. The lack of progress made in NLP for other languages will prove to be problematic and further the class divide in the long run.

Moreover, many machine learning techniques are developed exclusively for English. For example, stemming would only be effective in a language like English as it makes use of the different inflected forms of the language. However, in character-based languages such as Chinese, these practices do not work. In the current technology, we also heavily rely on spaces to separate each word, but languages such as Japanese don't divide their phrases with a certain symbol, adding to the complexity of processing the language with current techniques.

However, there are ways for us to close this gap. For example, when a group of Slovakian researchers tried to develop a chatbot based on Slovak, they ran into the problem of lack of data collected on the language. They devised a typing game that rewards the users for typing faster, collecting common misspellings of certain words. While the game yielded great results and vastly improved the database available for Slovak NLP developers, the English corpora are far more extensive. Thanks to economic inequality and lack of government support, there is a lack of data for many languages, and while that is not an issue with the technology itself, it will prove to be problematic if used on a bigger scale. Researchers have also attempted to use open-sourced databases such as Wikipedia, however, the support for most languages is limited, with Vietnamese ranking second with just one-fifth the size of the English corpora.

While this can be argued to be the most efficient commercial decision made by massive A.I. research companies, we must understand that these technology are implemented into our daily lives, and not at all exclusive to the services these companies offer. For example, in the legal field, many A.I. services are created based on English, disproportionately affecting non-native or less fluent English speakers. The complex nature of NLP also leads to a "black box" environment, allowing the biases in the algorithm to go unchecked. Many feedback sorting algorithms also utilize the current NLP technologies, and the lack of support for foreign

languages would also unintentionally diminish the voices and opinions of those who did not receive an education in English. Overall, the lack of support in NLP for other languages diminishes the voices of those not fluent in English, eliminating a huge demographic of people and perpetuating the gap between different classes.

Although this is an issue that requires systemic changes and time, it is important for us who are benefitting or a part of building the system to acknowledge it and push for changes. With recent advancements in multilingual algorithms such as XLM-RoBERTa, we are closing the gap between different languages. Moving forward, we must attempt to be more inclusive and cultivate an environment where technology will equally benefit everyone and not just the most privileged demographic.

Sources

https://www.tableau.com/learn/articles/natural-language-processing-examples
Hrkút, Patrik & Toth, Štefan & Ďuračík, Michal & Meško, Matej & Krsak, Emil & Mikušová, Miroslava. (2020). Data Collection for Natural Language Processing Systems. 10.1007/978-981-15-3380-8_6.
ULMFit for non-English Languages (NLP Video 10)
NLP for Developers: Non-English Pipeline Components | Rasa
XLM-RoBERTa: The multilingual alternative for non-english NLP
Interesting Novels Written By Artificial Intelligence | by Editorial | The Research Nest | The Research Nest
Unsupervised Cross-lingual Representation Learning at Scale (XML_roBERTa)
XLNet: Generalized Autoregressive Pretraining for Language Understanding
Ethical by Design: Ethics Best Practices for Natural Language Processing
The Staggering Cost of Training SOTA AI Models
XLM-RoBERTa — transformers 4.12.2 documentation

terms

- **autoregressive** (**AR**) **model** is a representation of a type of random process

- **lemmatisation** in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.
- **stemming** just removes or stems the last few characters of a word, often leading to incorrect meanings and spelling.
- **word embeddings** or **word vectorization** is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics.