Why Historical Language Is a Challenge for Artificial Intelligence

Originally published at https://www.unite.ai/why-historical-language-is-a-challenge-for-artificial-intelligence/

One of the central challenges of Natural Language Processing (NLP) systems is to derive essential insights from a wide variety of written materials. Contributing sources for a training dataset for a new NLP algorithm could be as linguistically diverse as Twitter, broadsheet newspapers, and scientific journals, with all the appellant eccentricities unique to each of just those three sources.

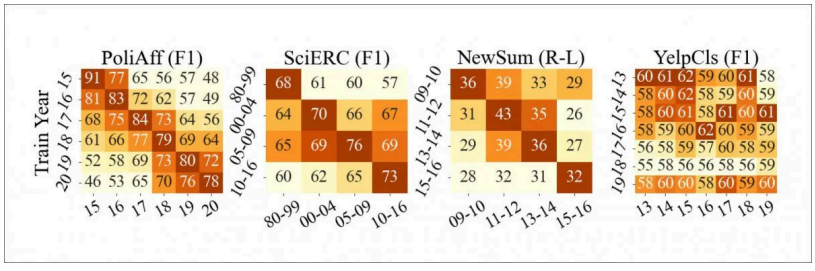
In most cases, that's just for English; and that's just for current or recent text sources. When an NLP algorithm has to consider material that comes from multiple eras, it typically struggles to reconcile the very different ways that people speak or write across national and sub-national communities, and especially across different periods in history.

Yet, using text data (such as historical treatises and venerable scientific works) that straddles epochs is a potentially useful method of generating a historical oversight of a topic, and of formulating statistical timeline reconstructions that predate the adoption and maintenance of metrics for a domain.

For example, weather information contributing to climate change predictive AI models was not adequately recorded around the world until 1880, while data-mining of classical texts offers older records of major meteorological events that may be useful in providing pre-Victorian weather data.

Temporal Misalignment

A new paper from the University of Washington and the Allen Institute for AI has found that even as short an interval as five years can cause *temporal misalignment* which can derail the usefulness of a pre-trained NLP model.



In all cases, higher scores are better. Here we see a heatmap of temporal degradation across four corpora of text material spanning a five year period. Such mism evaluation data, according to the authors of the new paper, can cause a 'massive performance drop'. Source: https://arxiv.org/pdf/2111.07408.pdf

The paper states:

> *'We find that temporal misalignment affects both language model generalization and task performance. We find considerable variation in degradation across text domains and tasks. Over 5 years, classifiers' F1 score can deteriorate as much as 40 points (political affiliation in Twitter) or as little as 1 point (Yelp review ratings). Two distinct tasks defined on the same domain can show different levels of degradation over time.'*

Uneven Splits

The core problem is that training datasets are generally split into two groups, sometimes at a fairly unbalanced 80/20 ratio, due to limited data availability. The larger group of data is trained on a neural network, while the remaining data is used as a control group to test the accuracy of the resulting algorithm.

In mixed datasets containing material that spans a number of years, an uneven distribution of data from various periods could mean that the evaluation data is inordinately composed of material from one particular era.

This will cause it to be a poor testing ground for a model trained on a more diverse mix of eras (i.e. on more of the entire available data). In effect, depending on whether the minority evaluation data over-represents newer or older material, it's like asking your grandfather to rate the latest K-Pop idols.

The long workaround would be to train multiple models on much more time-restricted datasets, and attempt to collate compatible features from the results of each model. However, random model initialization practices alone means that this approach faces its own set of problems in achieving cross-model parity and equity – even before considering whether the multiple contributing datasets were adequately similar to each other to make the experiment meaningful.

Data and Training

To evaluate temporal misalignment, the authors trained four text corpora across four domains:

Twitter
…where they collected unlabeled data by extracting a random selection of 12 million tweets uniformly spread between 2015-2020, where the authors studied named entities (i.e. people and organizations) and political affiliations.

Scientific Articles
…where the authors obtained unlabeled data from the Semantic Scholar corpus, constituting 650,000 documents spanning a 30-year period, and on which they studied mention type classification (SciERC) and AI venue classification (AIC, which distinguishes if a paper was published in AAAI or ICML).

News Articles
…where the authors used nine million articles from the Newsroom Dataset spanning a period 2009-2016, on which they performed three tasks: newsroom summarization, publisher classification and Media frames classification (MFC), which latter task examines the perceived prioritization of various topics across news output.

Food Reviews
…where the researchers used the Yelp Open Dataset on a single task: review rating classification (YELPCLS), a traditional sentiment analysis challenge typical of much NLP research in this sector.

Results

The models were evaluated on GPT-2, with a range of resulting F1 scores. The authors found that performance loss from temporal misalignment is bi-directional, meaning that models trained on recent data can be adversely affected by the influence of older data, and vice versa (see image at start of article for graphs). The authors note that this has particular implications for social science applications.

In general, the results show that temporal misalignment degrades performance loss 'substantially', and has a broad effect on most tasks. Datasets that cover very long periods, such as decades, naturally exacerbate the problem.

The authors further observe that temporal misalignment also affects labeled as well as unlabeled pretraining data. Additionally, their attempts to mitigate the effects via domain adaptation (see below) did not substantially improve the situation, though they assert that fine-tuning the data information in the dataset can help to a certain extent.

Conclusion

The researchers confirm previous findings that earlier-suggested remedies involving *domain adaptation* (DAPT, where allowance is crafted for the data disparity) and *temporal adaptation* (where the data is selected by time period) do little to alleviate the problem.

The paper concludes*:

> *'Our experiments revealed considerable variation in temporal degradation across tasks, more so than found in previous studies. These findings motivate continued study of temporal misalignment across applications of NLP, its consideration in benchmark evaluations, and vigilance on the part of practitioners able to monitor live system performance over time.*
>
> *'Notably, we observed that continued training of LMs on temporally aligned data does not have much effect, motivating further research to find effective temporal adaptation methods that are less costly than ongoing collection of annotated/labeled datasets over time.'*

The authors suggest that further investigation into continual learning, where the data is constantly updated, may be of use in this respect, and that concept drift, and other methods of detecting shifts in tasks could be a useful aid to updating datasets.

** My conversion of inline citations to hyperlinks.*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai