Future Language Modeling from Temporal Document History

Changmao Li, Jeffrey Flanigan
University of California, Santa Cruz
{changmao.li,jmflanig}@ucsc.edu

Abstract

Predicting the future is of great interest across many aspects of human activity. Businesses are interested in future trends, traders are interested in future stock prices, and companies are highly interested in future technological breakthroughs. While there are many automated systems for predicting future numerical data, such as weather, stock prices, and demand for products, there is relatively little work in automatically predicting *textual data*. Humans are interested in textual data predictions because it is a natural format for our consumption, and experts routinely make predictions in a textual format (Christensen et al., 2004; Tetlock & Gardner, 2015; Frick, 2015). However, there has been relatively little formalization of this general problem in the machine learning or natural language processing communities. To address this gap, we introduce the task of **future language modeling**: probabilistic modeling of texts in the future based on a temporal history of texts. To our knowledge, our work is the first work to formalize the task of predicting the future in this way. We show that it is indeed possible to build future language models that improve upon strong non-temporal language model baselines, opening the door to working on this important, and widely applicable problem.[1]

1 Introduction

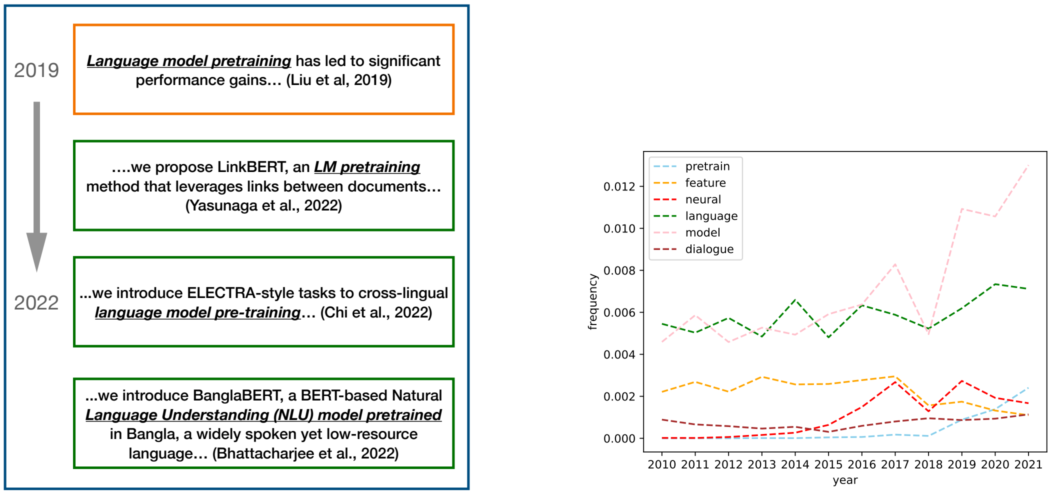
Predicting the future is a standard practice across numerous domains of human life and businesses (Christensen et al., 2004; Tetlock & Gardner, 2015; Frick, 2015). Public and private organizations constantly anticipate future trends, shifts in stock values, or forthcoming technological advancements. The pressure to predict the future has fueled developments in the automated prediction of future numeric data, encompassing areas such as weather forecasting, stock market trends, and demand for goods.

However, it is striking to note the scarcity of work developed towards the automation of predicting textual data. Textual data holds unique significance, given that it is a natural and rich format for human consumption. Moreover, experts frequently offer predictions in a textual format, evident in an array of books, magazines, and academic publications. Despite this, predicting future text data is rarely studied within the machine learning or natural language processing communities.

Our work aims to address this gap by introducing a novel task – future language modeling. The future language modeling task is to construct a generative language model for future text given a temporal history of documents. To the best of our knowledge, this is the first attempt to systematize and advance the task of predicting the future in this specific manner. Beyond formalizing this important task, we also create and develop future language models designed for this task. We evaluate these future language models against strong non-temporal baseline language models using both automatic metrics and human evaluations, and demonstrate their effectiveness at generating future textual content.

A word of caution: predicting the future is a bold claim. We do not wish to argue that all future text can be predicted. There are random events, new named entities, serendipitous discoveries, etc, in text that cannot be predicted. But we hypothesize that there are some important aspects of the future that *can* be predicted given enough historical text. Only by working on this future language modeling task

[1] Our code is available at https://github.com/jlab-nlp/Future-Language-Modeling



(a) (b)

Figure 1: (a) represents an example showing how abstracts in recent history are related to the future. In this example, the text of the abstract of the RoBERTa paper (Liu et al., 2019) anticipates the rise of papers about “language model pretraining” (Du et al., 2022; Bhattacharjee et al., 2022; Chi et al., 2022). (b) shows the word frequencies by year in NLP abstracts for some representative words, which reflects topic/approach changes over the years, i.e., “pretrain” started to dramatically go up after 2018 because of BERT, and “neural” became popular after 2013 because of deep learning.

can this hypothesis be verified. We show, by construction, that future language models can be built that perform better, across various automatic and manual evaluations, than non-temporal language models trained on the most up-to-date text, thereby verifying this hypothesis. While humans can sometimes predict the future, experts are often wrong (Frick, 2015), and we do not know the machine upper-bound on this task. We hope to push the boundaries of predicting future trends by proposing the task of and developing methods for future language modeling.

Our contributions are the following:

- We introduce the future language modeling task (§2) of modeling future textual content using a history of documents, as well as evaluation methods for this task (§4.4 & §4.6).
- We develop a series of future language models (§3) for this task that incorporate temporal information into pre-trained language models, which dynamically adjusts the generation probability to generate content that follows the predicted future trend.
- As a concrete example, we evaluate our model to model future abstracts for ACL conference papers, and our proposed approach outperforms the baseline model on both automatic metrics and human evaluation.

The paper is organized as follows. In §5, we present related work. In §2, we provide a task overview to introduce the proposed future text generation task based on texts in previous time spans. In §3, we present the details of our proposed approaches. §4 presents our experiments and results analysis.

2 Task Overview

We begin by defining some terms. **Without loss of generality, we call the times when we update our language model *years***, but they could be other time spans such as days or hours. Each year has a collection of texts for that year. For simplicity, we call these texts *documents*.[2]

Our proposed **future language modeling** task is to model future texts using documents from previous years. Let i denote the year index, and document $d_{ij} = \langle x_{ij1}, ..., x_{ijk} \rangle$ be jth document from the

[2] In our experiments in §4, the texts (“documents”) are abstracts.

ith year, where x_{ijk} is the kth token from the jth document in the ith year. Let $D_i = \{d_{i1}, ..., dij\}$ represent all documents from year i. The task is to generate D_i based on D_1 to D_{i-1}, which means during generation, the probability of each generated token x_{ijk} is computed not only from a standard language modeling perspective but also considering the content evolution from D_1 to D_i. The conditional probability for each token x_{ijk} is conditioned not only on the previously generated words in the sentence (as usual), but also on all the previous years' documents:

$$P(x_{ijk}|x_{ij1}...x_{ij(k-1)}, D_1 \dots D_{i-1}) \tag{1}$$

We call the model for the above task a **future language model**, formally defined as a statistical language model designed to assign high probability to future texts based on the temporal history of texts.

3 Approach

3.1 Overview of Models

We develop three methods for future language models: a word frequency model (§3.2), a contextual temporal model (§3.3) and a doubly contextualized temporal model (§3.4). In this section, we give some background notation common to all these models.

All our methods modify the language model probabilities to account for the temporal evolution. A language model usually calculates the probabilities with a softmax equation:

$$P(x_k|x_1...x_{k-1}) = \frac{E_{x_k}^T H_k}{\sum_{w'} E_{w'}^T H_k} \tag{2}$$

In this equation, $E_w \in \mathbb{R}^d$ is the learned output embedding vector for the wth word in the vocabulary, and $H_k \in \mathbb{R}^d$ is the contextualized embedding at position k. We use a transformer language model, and H_k is the vector of the last layer of the transformer decoder in position k. This is our baseline to compare with.

Our first two methods (§3.2 & §3.3) compute a temporal bias $B_{iw} \in \mathbb{R}$ for the wth word in the ith year that is calculated from the previous years. The bias term up-weights or down-weights vocabulary items to account for changes across years. The bias is added into the softmax equation to modify the probabilities:

$$P(x_k|x_1...x_{k-1}, D_1 \dots D_{i-1}) = \frac{E_{x_k}^T H_k + B_{ix_k}}{\sum_{w'} \left(E_{w'}^T H_k + B_{iw'}\right)} \tag{3}$$

We describe how B_{iw} is calculated in the following sections.

Our third method (§3.4) is more expressive, and calculates a contextualized bias term that depends on the previous words $x_1...x_{k-1}$ that have been generated. This allows the bias term to be contextualized in the output that is being generated. In our notation, the bias term $B_{ikw} \in \mathbb{R}$ is the bias for the wth word in the kth position in the generated sentence for the ith year. The softmax probability equation becomes:

$$P(x_k|x_1...x_{k-1}, D_1 \dots D_{i-1}) = \frac{E_{x_k}^T H_k + B_{ikx_k}}{\sum_{w'} \left(E_{w'}^T H_k + B_{ikw'}\right)} \tag{4}$$

For training, all our future language models are trained with standard cross-entropy loss:

$$L = -\sum_{k=1}^{|\mathbf{x}|} \log p(x_k|x_1 \dots x_{k-1}; \theta) \tag{5}$$

where θ represents the model parameters.

3.2 The Word Frequency Model

Our simplest method models the change over time of the frequency of the words without using any context from historical documents. It only uses the raw counts of the word over time to compute a