Classification of Direct- and Quasi-identifiers with Transformer Models

Adam Childs and Jack Galvin

{adamchilds, jgalvin}@berkeley.edu

Abstract

Classification of direct- and quasi-identifiers in unstructured text is a task of increasing importance to privacy preservation and responsible data set creation. Although intersecting with well-studied Named Entity Recognition (NER) tasks (e.g. person identification, coreference resolution, location identification), classification of identifiers presents challenges due to the broad definition of identifiers and the length of content over which identifier context must be considered. Entities within legal and medical documents are often named at the beginning of the document with coreferences appearing thousands of tokens later. We reproduce a state-of-the-art baseline for the identification of direct- and quasi-identifiers achieved with Longformer and present two improvements of token classification performance, as measured by AUC_{pr}. Additionally, we share results from a DeBERTA and BERT implementation to demonstrate Longformer advantage and test the benefit of NER-specific pretrained embeddings. We conclude by discussing the strengths and weaknesses of each model.

Introduction

Everyone has a fundamental right to privacy (Art. 12 of the Universal Declaration of Human Rights)[11]. In support of this right, several legal frameworks are established to regulate the protection of private information generated by individuals (GDPR[12], CCPA[2], PIPL). Respect of this right, and legal compliance, demands that individual-identifying data be redacted or otherwise transformed prior to scientific research, long term storage, or commercial use[8].

With this comes the mandate to balance two crucial priorities - the preservation of privacy and the utility of the data set. Redact or change *too little* and the data set violates privacy laws. Redact or change *too much* and the data is less useful for researchers. Additionally, manual annotations are costly and are subject to disagreements among annotators, particularly for quasi-identifiable data. As a result, recent work has focused on model-based approaches to identify direct- (i.e., full name, phone, address) and quasi-identifiable (i.e., data revealing membership in a class or data disclosing an identifying attribute) spans within text.

This is especially important within large documents. Identifiers are often introduced during initial sections of legal and medical documents with coreferences appearing throughout. Model-based methods for the identification of direct- and quasi-identifiers can assist in unlocking datasets that are otherwise inaccessible, supporting responsible academic and commercial use.

Background

The Longformer is a transformer pretrained with masked language modeling on long documents and introducing a novel attention mechanism capable of efficiently applying attention over much longer sequences of text than the transformers which preceded it[1]. Its improved attention operation combines a local windowed attention calculation with a task-motivated global attention mechanism that scales linearly with input length, as opposed to quadratically like the transformers before it. As a result, the Longformer outperforms models like RoBERTa on several long document natural language tasks, as it can process up to 4,096 tokens compared to the 512 token maximum. Like other transformers, the Longformer outputs raw hidden states, which can be used for a variety of downstream language-related tasks.

Pilán et. al. show that a Longformer fine-tuned on 1,268 court documents from the European Court of Human Rights (ECHR) outperforms a RoBERTa model fine-tuned on Ontonotes in the classification of direct- and quasi-identifiers within 553 Wikipedia articles [9]. This suggests that the modified attention mechanism within the Longformer benefits identification of direct- and quasi-identifiers within an out-of-domain dataset, even when sequence length is within the limits imposed by RoBERTa models.

Pilán et. al. feed the raw hidden states from the Longformer into a single linear inference layer. Additional layers, however, allow a network to discover latent representations. Instead of feeding raw hidden states from the Longformer into a linear inference layer, they can be fed into a feedforward neural network (FNN). It is reasonable to expect that the addition of optimized parameters to the classification head of a Longformer-based model improves identifier classification.

SPECTER is a transformer pretrained to produce document-level embeddings on scientific documents and outperforms SciBERT and several other competitive models on document classification and citation prediction tasks[4]. Given its success, feeding a concatenation of document-level embeddings from SPECTER with the hidden states produced by the Longformer into a FNN may yield improvements relative to the Longformer baseline on the classification of direct- and quasi-identifiers. While an ensemble approach to NLP tasks is not new, we are not aware of any other research that combines SPECTER document-level embeddings with Longformer hidden states on a token-level classification task.

The poorly specified problem of identifying quasi-identifiers recommends large language models (LLMs) with high scores across natural language understanding (NLU) tasks. After reviewing GLUE[13] and SQuAD2.0[10] leaderboards, we choose to select two BERT-architecture based models - DeBERTAv3[6] and a BERT model pretrained for NER with the NER CoNLL 2003 dataset[5]. DeBERTAv3 is recommended for use by strong showings on NLU leaderboards[14]. We select the pretrained BERT model to test efficacy of NER specific pretrained embeddings as input for fine-tuning, and evaluate performance on our task.

We formulate this problem as a token-level classification task. Each model we present is fine-tuned to generate token-level predictions indicating whether or not that token should be masked (i.e. a direct- or quasi-identifier). Pilán et. al. report several custom derivatives of precision and recall on this task. Since precision and recall are inversely related, and can change based on the classification threshold, we identify Area Under the Precision-Recall Curve AUC_{pr}(AUC) as the most important performance metric for our models.

Methods

Data

We use the TAB corpus presented by Pilán et al[9]. The dataset consists of 1268 court documents from the ECHR and is stored in json format. Each document has been annotated to identify the spans of text which contain direct- or quasi-identifiers. For evaluation of our models, we use (1) the TAB test set and (2) Wikipedia articles presented by Pilán et. al., which have been annotated the same way as the ECHR documents. Example quasi- and direct-identifiers are shown in Table 1.

Table 1: Example Direct- and Quasi-identifiers

Entity Type	(Q)uasi- or (D)irect Identifier Example
Code	(D) *. . . constitutional complaint (no. 1 BvR* ***170/06****) with the Federal. . .*
Person	(D) *. . . represented by* ***Mr Tyge Trier****, a lawyer. . .*
Datetime	(Q) *. . . On* ***9 December 1990*** *the applicant. . .*
Quantity	(Q) *. . . batting average (****758-for-2433****) with . . .*
Organization	(Q) *. . . news for* ***ITV Breakfast*** *show Daybreak now . . .*
Demographic	(Q) *. . . an artist and* ***runs a small business*** *in France which. . .*
Misc	(Q) *. . . had been an* ***intentional and secret collaborator*** *with the. . .*
Misc	(Q) *. . . involvement in* ***Project Blue Book****, a formal. . .*

Metrics

We use standard formulas for precision, recall, and F_1 scores. Since each of our predictions is a vector of token-level predictions, we report average AUC over the entire batch for ease of comparison between models.

Baseline Longformer

We use the Huggingface implementation for our baseline Longformer. Our data is tokenized using the "Fast" version of the Longformer tokenizer since it includes a mapping between each token and the span of text with which it corresponds, which simplifies labeling at the token-level. Any documents exceeding 4096 tokens are truncated at the maximum length.

Our Longformer model is trained on the training set of the TAB corpus for 2 epochs with a learning rate of 2 x 10-5, Adam optimizer, and a batch size of 1. All Longformer-based models are trained on a Virtual Machine with 96 CPU and 624 GB of memory, with each requiring 8 hours to train.

Longformer with FNN Head

Our baseline Longformer has a linear inference layer which accepts the final hidden states from the Longformer backbone. We remove the linear inference layer and feed the raw hidden states from the Longformer into a FNN with 2 hidden layers. The first hidden layer has 512 neurons and the second has 256. Each dense layer is followed by a dropout layer (0.3) to prevent overfitting. We train and evaluate on the same data as our baseline model.

Ensemble: Longformer and SPECTER with FNN Head

We build on the Longformer FNN with an ensemble of Longformer and SPECTER models by concatenating the hidden states from each model and feeding the concatenated tensor into the same FNN in the above experiment. As with Longformer, we use the Huggingface implementation of SPECTER and its tokenizer.

In order to concatenate the document-level embeddings from SPECTER (512 x 768) with the final hidden states from Longformer (4096 x 768), we use the identity matrix to project the SPECTER document-level embeddings into the Longformer embedding space. Note that the concatenated tensor's first 512 entries combine context between SPECTER and Longformer, but anything after is purely the hidden states generated by Longformer.

DeBERTaV3 Large

We select DeBERTaV3 Large[7] as a representative pretrained language model expanding on training methodologies advanced by ELECTRA[3] and DeBERTA. We apply DeBERTaV3 Large, a 24 layer transformer model pretrained with masked language modeling (MLM) and replaced token detection (RTD), pretrained with gradient-disentangled embedding sharing. As this model's input is limited to 1024 tokens, we train and evaluate within that token count bound.

Results and Discussion

Baseline Longformer

Figures for precision, recall, F_1, and AUC can be found in Table 2. Our Longformer is more precise but has lower recall compared to the fine-tuned Longformer from Pilán et. al. This difference is likely due to two factors. First, Pilán et. al. alter the attention window of the Longformer to 4,096 tokens while we use the default configuration of 512. Secondly, they implement a weighting scheme for masked and unmasked tokens while we evaluate precision and recall with equal weighting. Despite these differences, our baseline model delivers an F_1 score of 0.787 compared to their published F_1 of 0.812 on the Wikipedia dataset.

Longformer with FNN

Replacing the linear inference layer in our baseline Longformer with a FNN improves AUC by over 0.03 points on the Wikipedia dataset and 0.01 points on the TAB test set. As we expected, adding more parameters that can be optimized during training improves overall performance on the identification of direct- and quasi-identifiers. The performance difference relative to baseline is larger on the Wikipedia dataset, attributable to the shorter length of the wikipedia data set text articles, median length of 51 words, compared to the TAB test set with a median article word count of 915.

Ensemble: Longformer and SPECTER with a FNN

Our ensemble of a Longformer with SPECTER combines the power of a modified attention mechanism specialized for long sequences with document-level embeddings from SPECTER. The ensemble model, which feeds concatenated embeddings into the same FNN from the previous experiment, improves AUC by roughly 0.06 points relative to baseline and 0.03 points compared to the same model without SPECTER on the Wikipedia dataset.

As with the prior experiment, this model achieves no appreciable increase in AUC relative to baseline on the TAB test set. We conclude that not only does the difference in average sequence length between the two datasets affect performance, but also that the document-level embeddings generated by SPECTER have more utility on shorter sequences compared to longer ones.

DeBERTaV3 and NER pretraining

The fine-tuned DeBERTaV3 Large pretrained language model performed poorly across all of our metrics. Experiments to improve model performance through adding a FNN head and hyperparameter tuning yielded minimal model performance improvement. The similar performance of these variations, in the best cases, indicate DeBERTaV3 is a poor fit as a base model for fine tuning on our task.

The similarity of our task to NER, specifically NER people and location identification, suggests that adopting embeddings trained specifically for NER may offer an advantage. Experimenting with a CoNLL-2003 pretrained BERT model, fine-tuned the same as our DeBERTAV3 Large, resulted in improved precision and recall, across both sets, but degraded AUC compared to DeBERTaV3 Large.

Table 2: Model Performance

Model	Wikipedia	Data	Results		TAB	Test	Set	Results
	Precision	Recall	F_1	AUC	Precision	Recall	F_1	AUC
Pilán et. al.	0.708	0.952	0.812	—	0.836	0.919	0.876	—
Longformer	0.930	0.682	0.787	0.852	0.966	0.720	0.825	0.875
Longformer$_{FNN}$	0.903	0.763	0.827	0.888	0.946	0.740	0.830	**0.885**
Longformer$_{FNN,SPECTER}$	0.895	0.800	0.845	**0.914**	0.956	0.724	0.824	0.877
DeBERTa-v3$_{Large}$	0.518	0.261	0.347	0.570	0.348	0.312	0.329	0.480
BERT-base-NER	0.580	0.288	0.384	0.552	0.370	0.378	0.374	0.473

Ethical Considerations

The removal of identifiers and quasi-identifiers from documents, in some cases, can be of real critical importance. We recommend any implementation to incorporate multiple layers of result review as part of any production system, and a specialist in de-identification be consulted before deployment. Regarding inherent risks and biases of language models, we recommend review of *Ethical and social risks of harm from Language Models*[16] prior to adopting any methods described in this paper.

Conclusions and Future Work

Based on our experimental analysis, we conclude that context over large spans of text is important for the identification of direct- and quasi-identifiers. Even the Longformer, which is pretrained for masked language modeling on large documents, better identifies direct- and quasi-identifiers within shorter documents across our test sets. We also confirm that additional layers in the classification head and ensembling pre-trained models improve performance over baseline AUC on the direct- and quasi-identifier classification task. In particular, we recognize that pretrained language embeddings for NER tasks can benefit model performance, either through ensembling or as part of a base model.

With more time, memory, and compute power, we recommend next step exploration of embedding concatenation search strategies, e.g. ACE[15], and combining those techniques with the dilated- and sliding-window attention mechanisms introduced by Longformer.

References

[1] Beltagy, I. et al. 2020. Longformer: The long-document transformer. *CoRR*. abs/2004.05150, (2020).

[2] BUKATY, P. 2019. *The california consumer privacy act (CCPA): An implementation guide.* IT Governance Publishing.

[3] Clark, K. et al. 2020. Electra: Pre-training text encoders as discriminators rather than generators. (2020).

[4] Cohan, A. et al. 2020. SPECTER: Document-level representation learning using citation-informed transformers. *CoRR*. abs/2004.07180, (2020).

[5] Devlin, J. et al. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv.

[6] He, P. et al. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv.

[7] He, P. et al. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing.

[8] Ning, Y. et al. 2021. Deep learning based privacy information identification approach for unstructured text. *Journal of physics. Conference series*. 1848, 1 (2021), 12032–.

[9] Pilán, I. et al. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. (2022).

[10] Rajpurkar, P. et al. 2018. Know what you don't know: Unanswerable questions for SQuAD. arXiv.

[11] United Nations 1948. *Universal declaration of human rights*.

[12] Voigt, P. and Bussche, A. von dem 2017. *The EU general data protection regulation (GDPR): A practical guide*. Springer Publishing Company, Incorporated.

[13] Wang, A. et al. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. (2019).

[14] Wang, A. et al. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. arXiv.

[15] Wang, X. et al. 2020. Automated concatenation of embeddings for structured prediction. *CoRR*. abs/2010.05006, (2020).

[16] Weidinger, L. et al. 2021. Ethical and social risks of harm from language models. (2021).