Enhancing Mask Predictions for Text Anonymization

266 Natural Language Processing Final Project

Derrick Chan-Sew, Jason Dong

Agenda

01 **Introduction**
Background

02 **Experimental Design**
Baseline
Attention Mask
Concatenated Embeddings

03 **Results**
SeqEval vs. TAB Metrics
Experiments

04 **Conclusion**

01
IntroDuction

INTRO

- The right to privacy is defined in the **Universal Declaration of Human Rights** (Art. 12) and is further articulated in multiple national and international legal instruments.
- The Health Insurance Portability and Accountability Act of 1996 (**HIPAA**), General Data Protection Regulation (**GDPR**) of 2016 and California Consumer Privacy Act (**CCPA**) of 2018 aim to protect individual privacy as our we increase our digital footprint
- Previous work has mainly focused on de-identification on direct identifiers such as name, address and phone numbers
- The combination of quasi-identifiers 'gender', 'birth date', and 'postal code' reidentify between **63** and **87%** of the U.S. population.
- True anonymization requires protecting quasi-identifiers such as physical appearance, current profession, or political opinions or other geographical or temporal markers based on sensitivity.

ABSTRACT

- Build on novel techniques that evaluate efficacy of anonymization
- Employ and enhance modern architectures such as transformers and attention to improve performance upon traditional NER approaches
- Enhance models using richer, public datasets that include direct and quasi identifiers
- Build framework for improved anonymization in preparation for downstream tasks such as summarization

BACKGROUND

- Dataset
  - Public corpus of 1,268 court cases (Text Anonymization Benchmark, or TAB) including PII text spans, masking decision, confidentiality level, entity and co-reference information
- Baseline
  - 1. spaCy - a generic neural model trained for named entity recognition (NER)
  - 2. Microsoft Presidio - A privacy-oriented NER-based text de-identification system
  - 3. Longformer (BERT) - sequence labeling models based on large, pre-trained language models fine-tuned on NER
- Evaluation Metrics
  - Baseline include inter-annotator agreement across quasi and direct identifiers with emphasis on recall

BACKGROUND

Direct, Quasi, No Mask

The case originated in an application (no . 36110/97) against the Republic of Turkey lodged with the European Commission of Human Rights ("the Commission") under former Article 25 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by four Turkish nationals, Mr Galip Yalman, Mr Bahattin Sarısoy, Mr Osman Çağlayan and Mr Yusuf Çamca ("the applicants"), on 29 November 1996.

The applicants were represented by Mr S. Esmer, a lawyer practising in Ankara. The Turkish Government ("the Government") did not designate an Agent for the purposes of the proceedings before the Convention institutions.

The applicants alleged that their case, which commenced in 1989 and terminated in 1996, was not heard within a reasonable time as required by the Convention.

The application was transmitted to the Court on 1 November 1998, when Protocol No. 11 to the Convention came into force (Article 5 § 2 of Protocol No. 11).

PROCEDURE

The case originated in an application (no. 36110/97 CARDINAL) against the Republic of Turkey GPE lodged with the European Commission of Human Rights ORG ("the Commission") under former Article 25 of the Convention for the Protection of Human Rights and Fundamental Freedoms LAW (" the Convention WORK_OF_ART ") by four CARDINAL Turkish NORP nationals, Mr Galip Yalman PERSON , Mr Bahattin Sarısoy PERSON , Mr Osman Çağlayan PERSON and Mr Yusuf Çamca PERSON ("the applicants"), on 29 November 1996 DATE .

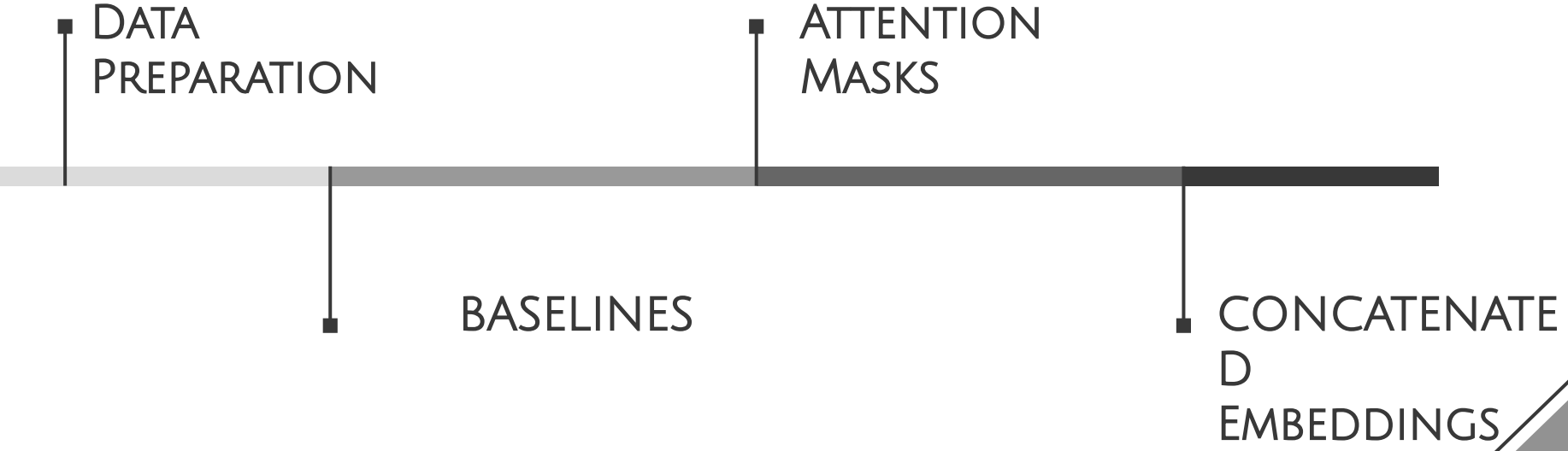
The applicants were represented by Mr S. Esmer PERSON , a lawyer practising in Ankara GPE . The Turkish Government ORG ("the Government") did not designate an Agent for the purposes of the proceedings before the Convention ORG institutions.

The applicants alleged that their case, which commenced in 1989 DATE and terminated in 1996 DATE , was not heard within a reasonable time as required by the Convention.

The application was transmitted to the Court ORG on 1 November 1998 DATE , when Protocol No. 11 PRODUCT to the Convention came into force (Article 5 § 2 of Protocol LAW No. 11).

02
Experimental Design

EXPERIMENTAL DESIGN



Dataset Preparation

1. Removed 274 court cases with annotation disagreements
2. Aligned text span labels to word tokens and IOB labels
3. Wordpiece tokenizer and sequence padding for Longformer

944 court cases:
- Full dataset: 795/101/98
- Mini dataset: 400/50/50

[Mr Galip Yalman] [DIRECT]
[Mr, Galip, Yalman] [B-DIRECT, I-DIRECT, I-DIRECT]
[ĠMr, ĠGal, ip, ĠY, al, man] [B-DIRECT, I-DIRECT…]

Entity Type	Mentions	Direct	Quasi	No Mask
DATETIME	29502 (35%)	7 (0%)	26005 (88%)	3490 (12%)
ORG	24048 (28%)	10 (0%)	8691 (36%)	15347 (64%)
PERSON	13145 (16%)	2123 (16%)	8323 (63%)	2699 (21%)
LOC	5251 (6%)	1 (0%)	3833 (73%)	1417 (27%)
DEM	4499 (5%)	1 (0%)	2128 (47%)	2370 (53%)
MISC	3612 (4%)	22 (1%)	2275 (63%)	1315 (36%)
CODE	2138 (3%)	1186 (55%)	758 (35%)	194 (7%)
QUANTITY	2218 (3%)	0 (0%)	1843 (83%)	375 (17%)
Total	84413	3350 (4%)	53856 (64%)	27207 (32%)

EXPERIMENTS

Baselines:
- spaCy
- Microsoft Presidio
- Longformer (BERT)

Binary Classification: convert dataset to MASK vs. NO MASK

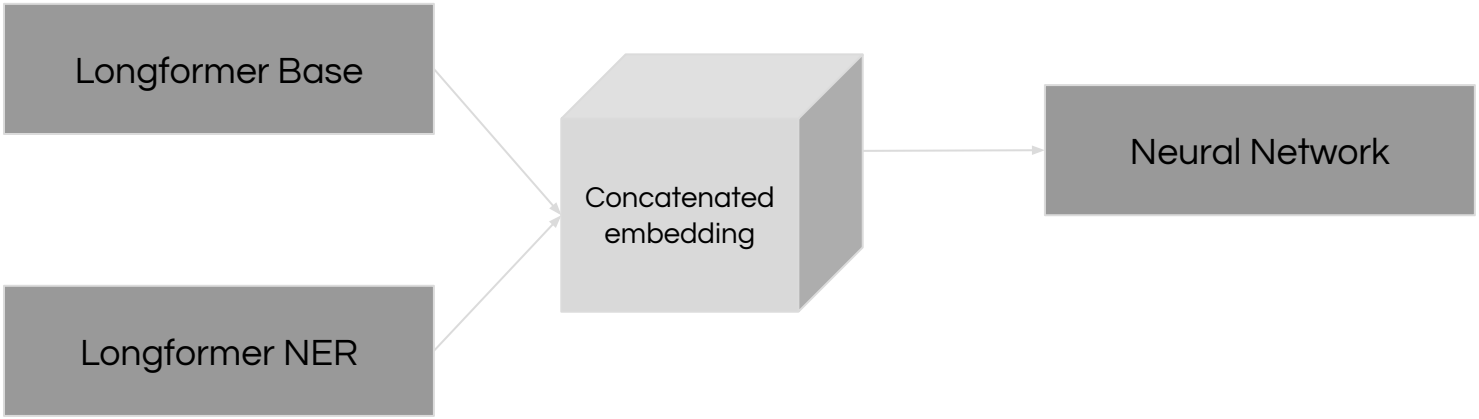
Attention Mask*:
- Overweighting
- Underweighting

Experiment	Attention Mask
Original Text	[...a lawyer practising in Ankara... <PAD>]
Original Mask	[..., 1, 1, 1, 1, 1, ..., 0, 0, 0]
Overweighting	[..., 1, 1, 1, 1, 1.5, ..., 0, 0, 0]
Underweighting	[..., .5, .5, .5, .5, 1, ..., 0, 0, 0]

* Peter Shaw et al. *Self-Attention with Relative Position Representations.* 2018

EXPERIMENTS

Concatenated Embeddings*:



* Yanru Dong et al. *A Fusion Model-Based Label Embedding and Self Interaction Attention for Text Classification.* 2020
* Xinyu Wang et al. *More Embeddings, Better Sequence Labelers?*. 2021

Technical Configuration

Reducing Memory Size:
- Floating point precision 16
- Gradient checkpoint

Hyperparameters:

Parameters	Values
Learning rate scheduler	Linear, Linear w/ warmup, Cosine w/ warmup
Learning rates	5e-4, 1e-4, 2.5-5,
Warmup ratio	10%
Batch size	16, 8
Epochs	20
Early Stopping	3

03
RESULTS

RESULTS

Model	SeqEval						TAB			
	Train			Test			Train		Test	
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	Recall	Precision
spaCy									**0.88**	0.56
Presidio									**0.70**	0.69
Longformer (TAB paper)									**0.95**	0.71
Longformer Baseline	**0.79**	0.77	0.75	**0.72**	0.71	0.71	**0.96**		**0.94**	1.00
Longformer Binary	**0.77**	0.74	0.75	**0.74**	0.74	0.74	**0.96**	0.99	**0.94**	1.00
Attention (over {1, 1.5})	**0.89**	0.04	0.08	**0.80**	0.03	0.07	**0.96**	1.00	**0.94**	1.00
Attention (under {0.75, 1})	**0.83**	0.80	0.81	**0.83**	0.77	0.81	**0.96**	0.99	**0.94**	1.00
Concatenated*	**0.02**	0.74	0.04	**0.04**	0.98	0.07	**0.97**	0.99	**0.91**	0.99

Figures exclude 'O' tokens. Experiments were trained on a mini dataset of 400/50/50 court cases

✶ Concatenated model trained on 64 court cases

DISCUSSION: SEQEVAL VS. TAB

1. Class Distinction vs. Final Masking Decision [MASK, NO MASK]

Source Text: …Mr Ahmet Kenan Er ("the applicant"), on 30 April 2004…
Longformer: …Mr Ahmet Kenan Er ("the applicant"), on 30 April 2004…

2. Token specificity vs. Text Span

Source Text: Hearings were subsequently held on 23 February, 11 August, 22 September, 10 November and 9 December 2005.
Longformer: Hearings were subsequently held on 23 February, 11 August, 22 September, 10 November and 9 December 2005 .

Direct, Quasi, No Mask

DISCUSSION: ATTENTION MASKS

Experiment	Mask Adjustments {O, DQN}	DIRECT			QUASI			OVERALL		
		Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Longformer Base	{1, 1}	**0.70**	0.89	0.79	**0.82**	0.75	0.78	**0.72**	0.71	0.71
Attention (overweight)	{1, 1.5}	**0.73**	0.80	0.76	**0.57**	0.01	0.01	**0.80**	0.03	0.07
Attention (underweight)	{0.75, 1}	**0.70**	0.69	0.69	**0.91**	0.78	0.84	**0.79**	0.77	0.76
Attention (underweight)	{0.5, 1}	**0.68**	0.83	0.75	**0.50**	0.76	0.61	**0.80**	0.77	0.79
Attention (underweight)	{0.25, 1}	**0.59**	0.87	0.70	**0.92**	0.76	0.83	**0.78**	0.77	0.76

Example 1:

Source Text: …to Article 403 of the former Criminal Code (Law no . 765)

Longformer: …to Article 403 of the former Criminal Code (Law no . 765)

Attention: …to Article 403 of the former Criminal Code (Law no . 765)

Direct, Quasi, No Mask

DISCUSSION: CONCATENATED EMBEDDINGS

Experiment	DIRECT			QUASI			OVERALL		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Longformer Base	**0.70**	0.89	0.79	**0.82**	0.75	0.78	**0.72**	0.71	0.71
Concatenated*	**0.46**	0.98	0.63	**0.00**	1.00	0.00	**0.04**	0.98	0.07

✶ Concatenated model trained on 64 court cases

1. High Precision; Low Recall - RNNs or CNNs may be better
2. Concatenated Embeddings (4096 x 1536) - 9.8 GB
   - Reduce dimensions prior to extraction from Longformer
   - Average only NER embeddings
   - Additive embeddings

04
Conclusions

Conclusion

- Next steps
- Anonymization techniques
  - Replace
  - Redact
  - Hash
  - Encyrpt
  - Mask
- Summarization

Q&A