LREC 2022
Marseille

https://www.compriseh2020.eu

Paper ID: 1192

Adapting Language Models When Training on Privacy-Transformed Data

Tugtekin Turan , Dietrich Klakow , Emmanuel Vincent and Denis Jouvet

Introduction

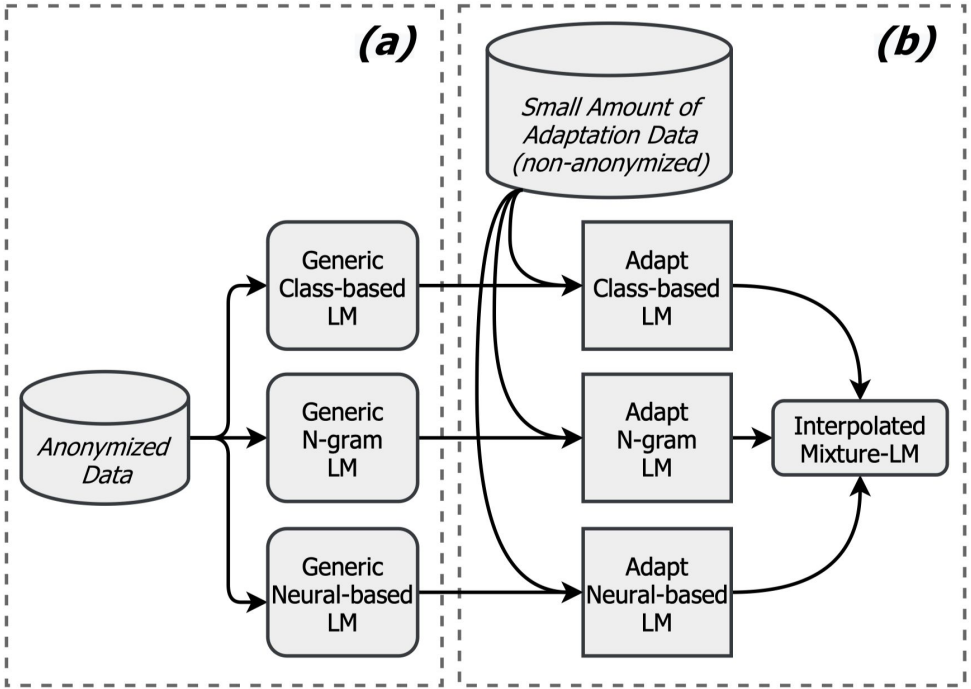
- *Privacy-preserving* models gained too much importance in recent years
    - General Data Protection Regulation (GDPR): storing user data is strongly regulated
- Spoken messages having sensitive information about the user characteristics should not be centralized in a single place
- It is necessary to follow *sanitization* approach that removes private information relating to *persons*, *locations*, *and organization names*
- Yet, hiding information gives *less accurate* automatic speech recognition (ASR)

Paper Idea

- This paper proposes a method to recover the performance loss when training language models (LMs) on *sanitized data*
- We employ a mixture of neural and word-based LMs with *class-based LMs*, where each *named entity* category corresponds to one class
- Our data sanitization process relies on recognizing and replacing named entities by other words from the *same class*
- Applying the *class-based* idea, we were able to represent *anonymous* data better via *named entities*

Methodology

- Our methodology consists of a *two-stage* adaptation scheme
- Following the CoNLL'03, labels are, *PER*, *ORG*, *LOC*, and *MISC*
- The sanitization is performed using the *word-by-word*
- Class-based LMs are inserted over *finite-state transducers*



Language Model Adaptation

- After training the individual LMs, we apply two adaptation strategies
  - For word- and class-based LMs, we use unigram LMs to adapt trigrams where this approach combines unigram and trigram info inspired by the *marginal adaptation*
  - For the neural LSTM-LM, we follow *fine-tuning* idea where we train a background LM on sanitized set and adaptation is performed by fine-tuning final softmax layer
- During the final decoding, we use an *interpolated mixture* LM
  - Also n-gram *approximation* is evaluated for neural-LM into the single-pass decoding
- At the final stage, we smooth the distribution over an *n-gram ARPA* LM

Experimental Setup

Set	Dur. (min.)	Uttr.
Train	4,880	108,221
Test	580	13,059
Adapt.	531	12,612
All	*5,991*	*133,892*

- We use *Augmented Multiparty Interaction (AMI)* corpus for experiments (overall duration: *100 h*)
  - AMI also provides *annotated* entity tags
- The *adaptation* set represents around *12%* of the training data
- As the acoustic model, we use *Kaldi*'s chain model based on the *time-delay neural network*

WER and PPL Results of the Generic LMs

- *[M1]* 3-gram word-based LM in single-pass decoding
- *[M2]* 3-gram class-based LM in single-pass decoding
- *[M3]* 3-gram approximation of the LSTM-based model
- *[M4]* rescoring of the lattice hypotheses with LSTM-LM

Model	Sanitized Data		Original Data		Size: 12%	
	WER [%]	PPL	WER [%]	PPL	WER	PPL
[M1]	32.3	121	28.8	82	31.5	109
[M2]	**30.2**	103	29.3	74	**29.9**	94
[M3]	32.9	137	29.1	88	30.8	101
[M4]	30.5	103	**27.6**	73	30.1	95

Results of Interpolating the LMs

Model	Description	WER	PPL
[M1 + M2]	word-& class-based	29.7	95
[M1 + M2 + M3]	+LSTM (3g app.)	29.6	92
[M1 + M2 + M4]	+LSTM (2nd-pass)	**29.4**	86

- We utilize the default adaptation split (corresponding to 12% of the training size) for our interpolation experiments
- Linear interpolation of the previously adapted LMs is employed with the best weight combinations

Conclusions

- This paper fills the gap by focusing on the *LM adaptation* of initially trained privacy-transformed data using *a few amount* of original *untransformed* data
- We combined *class-based LMs* which overcome data sparsity inside the n-grams and *neural LMs* which handle longer contexts and better predictions
- Experiments show that training an LM on privacy-transformed data result in a relative 11% WER *increase* (compared to the original untransformed data)
- Adapting LM over a limited amount of original untransformed data leads to a relative 8% WER *improvement*

Questions ?