DIFFERENCE MASKING:
Choose What to Mask in Continued Pretraining

CSCI-535: Multimodal Probabilistic Learning of Human Communication

Presenter: Jingmin Wei. Feb 28, 2024

Wilf, Alex, et al. "Difference-Masking: Choosing What to Mask in Continued Pretraining." *arXiv preprint arXiv:2305.14577* (2023).

Outline

- Motivations and Related Work
- Method: Difference Masking
- Qualitative Results
- Further Discussions

Motivations and Related Work

Masking

Masking has led to promising performance gains on a variety of downstream tasks. Masking in adapting pretrained models is effective when the target domain differs from the pretraining domain. Some key approaches include:

EntityBERT: masks tokens identified as entities by a domain-specific named-entity recognizer, applying a uniform strategy across different domains.

Salient Span Masking: employs named-entity recognition to mask entities for improving open-domain QA performance, also requiring a domain-specific entity tagger.

Selective Masking: Focuses on masking tokens based on their contribution to downstream task performance, leveraging supervised task labels.

MST & AttnMask: Utilize attention maps to mask "non-essential" or highly attended regions, without considering domain-specific information.

Motivations

EntityBERT and Salient Span Masking require a domain specific pretrained entity-tagger, and the masking strategy they determine is the same for any domain to which that tagger is applied.

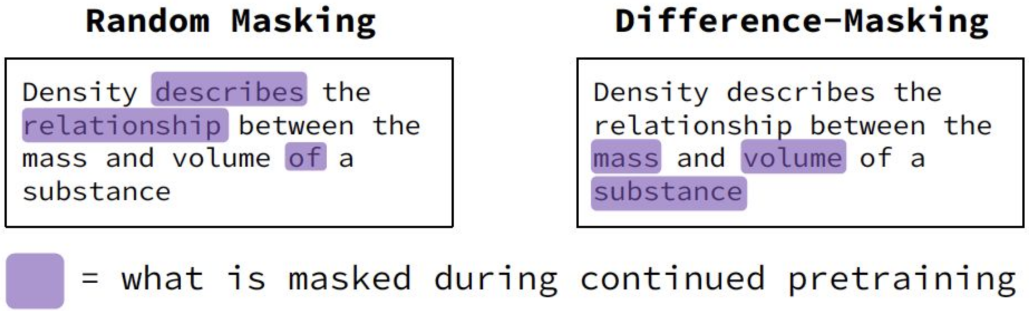
Selective Masking needs supervised downstream task labels.

MST & AttnMask do not take into account domain-specific information.

Develop a Masking method, which can automatically determines what to mask without pretrained entity-taggers, take into account domain-specific information to better keep important information in a given input sequence, and constructed in a self-supervised learning framework.

Method: Difference Masking

Comparison



Random Masking

Density describes the relationship between the mass and volume of a substance

Difference-Masking

Density describes the relationship between the mass and volume of a substance

▨ = what is masked during continued pretraining

Compared with Random Masking, the method should automatically chooses what to mask during continued pretraining by considering what makes a target domain different from the pretraining domain, enhancing model learning on the end task.

Terminology

X_{PT}: data of domain distribution for pretrained model.

X_T: target domain for adapting the pretrained model.

Y : task labels.

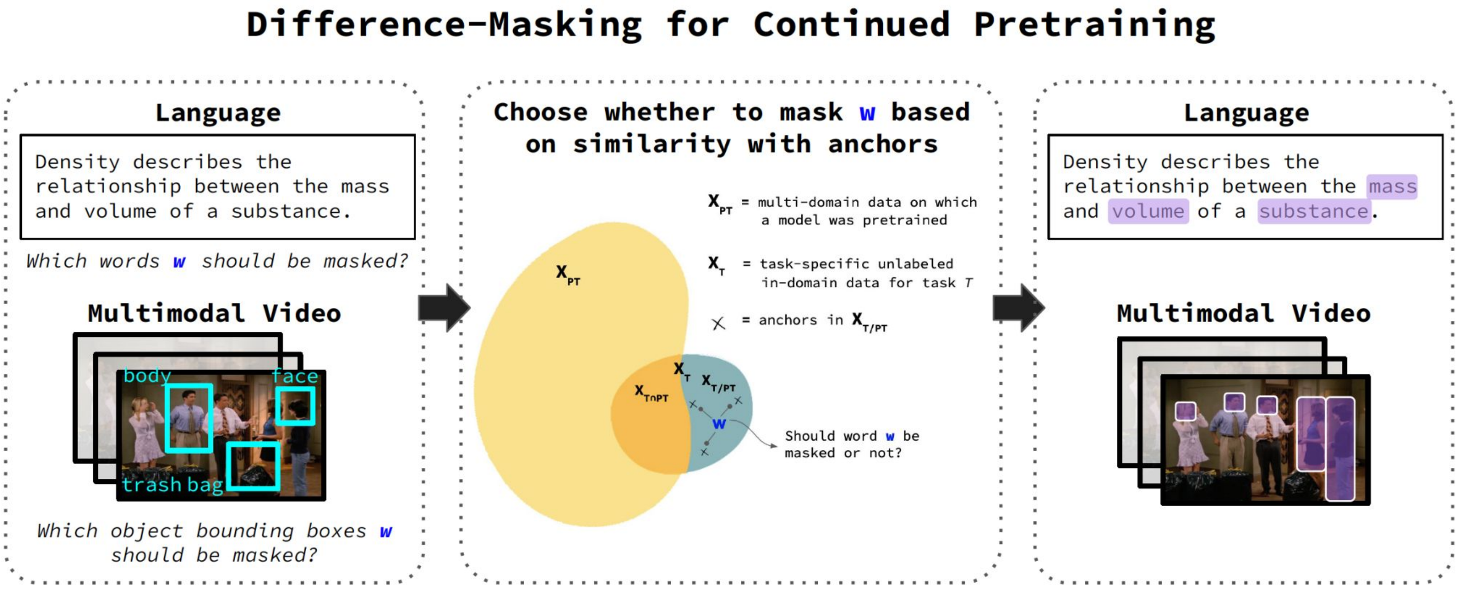
$X_{T \cap PT}$: concepts to appear both in X_T and X_{PT} .

$X_{T/PT}$: domain that make X_T different from X_{PT} .

Concepts common in X_T but uncommon in X_{PT} share higher mutual information with the target task label than concepts found in both domains.

Intuitively, the goal of the new masking method is to learn representation that capture the information unique to the domain ($X_{T/PT}$), which is more relevant for the downstream task.

Structure


Difference-Masking for Continued Pretraining
Language
Density describes the relationship between the mass and volume of a substance.
Which words w should be masked?
Multimodal Video
body
face
trash bag
Which object bounding boxes w should be masked?
Choose whether to mask w based on similarity with anchors
X_{PT} = multi-domain data on which a model was pretrained
X_T = task-specific unlabeled in-domain data for task T
× = anchors in $X_{T/PT}$
X_{PT}
X_T
$X_{T∩PT}$
$X_{T/PT}$
w
Should word w be masked or not?
Language
Density describes the relationship between the mass and volume of a substance.
Multimodal Video

Finding different anchors: TF-ICF to determine words that commonly found in domain X_T and not commonly found in general domains X_{PT} .

Masking based on difference: determine the likelihood that each word should be masked based on its similarity to these difference anchors.

Details

To find words that make a target corpus X_T different from general corpora X_{PT}, the score of a word is highest when it appears frequently in X_T and infrequently in X_{PT} .

$$\text{TF-ICF}(w_i) = \frac{freq(w_i, X_T)}{freq(w_i, X_{PT})}$$

Utilizing TF-ICF, we choose the top k scores as anchors A to represents the target domain, the model then masks the words based on the similarity to these anchors:

$$\text{sim}(w, A_k) = \cos(\text{BERT}(w), \text{BERT}(A_k))$$

Then we need generate probability distribution α over words in the sentence to represent the probability that each word to be masked.

$$\alpha(w_i) = \frac{\max_{k \in K} \text{sim}(w_i, A_k)}{\sum_{j=1}^{N} \max_{k \in K} \text{sim}(w_j, A_k)}$$

Difference-Masking then masks terms by sampling from this distribution α without replacement.

Qualitative Results

Dataset

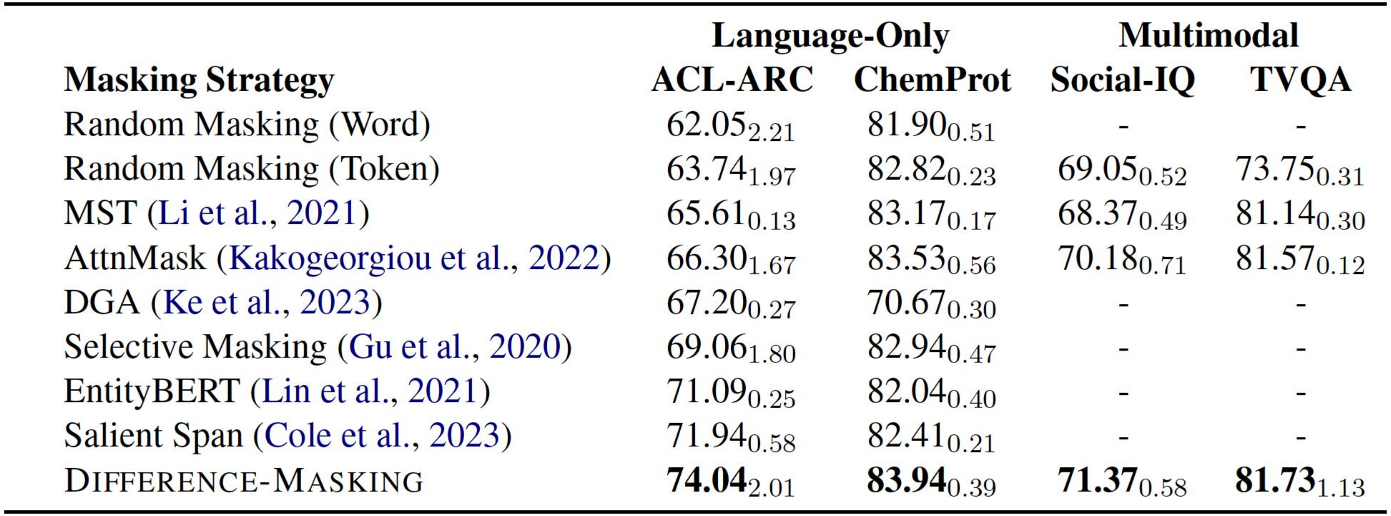
ChemProt: Focuses on chemical document-based relation classification. It's a low-resource task with abundant unlabeled in-domain data, making it suitable for Self-Supervised Learning (SSL) to enhance model performance.

ACL-ARC: Used for citation intent classification within the academic domain, leveraging the ACL Anthology Reference Corpus. It supports structured pretraining and evaluation with defined data splits.

TVQA: Contains 21,792 videos from six TV shows, paired with questions and multiple-choice answers. It integrates video, audio, and subtitles for comprehensive video understanding tasks.

Social-IQ: Features 1,250 videos of social situations with related questions and answers, requiring analysis of video, audio, and subtitle data for understanding social interactions.

Qualitative Results

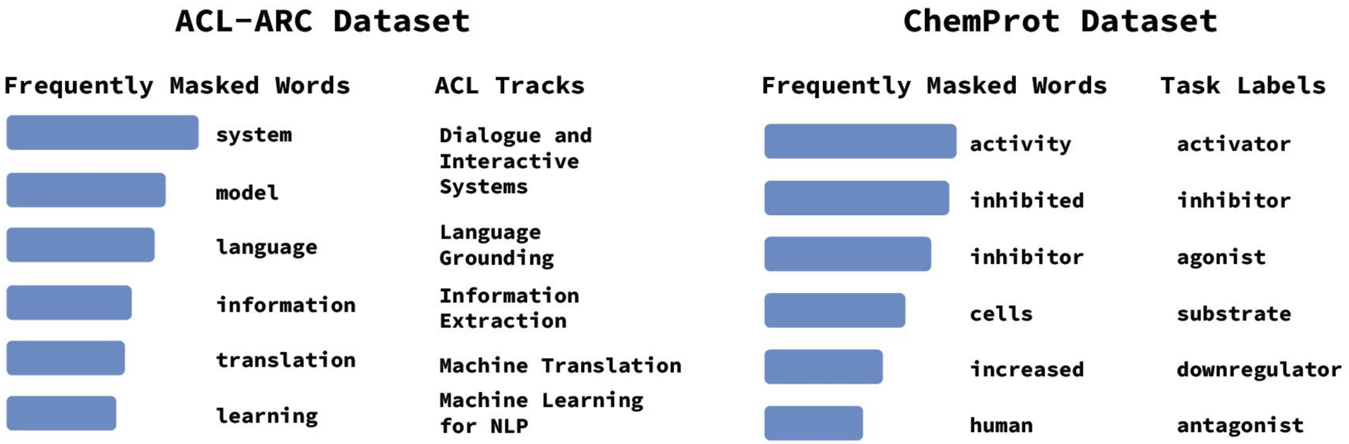


Masking Strategy	Language-Only ACL-ARC	 ChemProt	Multimodal Social-IQ	 TVQA
Random Masking (Word)	$62.05_{2.21}$	$81.90_{0.51}$	-	-
Random Masking (Token)	$63.74_{1.97}$	$82.82_{0.23}$	$69.05_{0.52}$	$73.75_{0.31}$
MST (Li et al., 2021)	$65.61_{0.13}$	$83.17_{0.17}$	$68.37_{0.49}$	$81.14_{0.30}$
AttnMask (Kakogeorgiou et al., 2022)	$66.30_{1.67}$	$83.53_{0.56}$	$70.18_{0.71}$	$81.57_{0.12}$
DGA (Ke et al., 2023)	$67.20_{0.27}$	$70.67_{0.30}$	-	-
Selective Masking (Gu et al., 2020)	$69.06_{1.80}$	$82.94_{0.47}$	-	-
EntityBERT (Lin et al., 2021)	$71.09_{0.25}$	$82.04_{0.40}$	-	-
Salient Span (Cole et al., 2023)	$71.94_{0.58}$	$82.41_{0.21}$	-	-
DIFFERENCE-MASKING	$\mathbf{74.04}_{2.01}$	$\mathbf{83.94}_{0.39}$	$\mathbf{71.37}_{0.58}$	$\mathbf{81.73}_{1.13}$

Difference-Masking outperforms strong baselines in both the language and multimodal experimental settings. The entirely self-supervised method also outperforms Selective Masking, which uses labelled data to inform its masking strategy. Values are average results over five trials, subscripts are standard deviations.

Further Discussions

What is Masked?



ACL-ARC Dataset

Frequently Masked Words	ACL Tracks
system	Dialogue and Interactive Systems
model	
language	Language Grounding
information	Information Extraction
translation	Machine Translation
learning	Machine Learning for NLP

ChemProt Dataset

Frequently Masked Words	Task Labels
activity	activator
inhibited	inhibitor
inhibitor	agonist
cells	substrate
increased	downregulator
human	antagonist

In ACL-ARC dataset, the frequently masked words had an interesting grounding in human intuition. This proves that the words masked by Difference-Masking align with what makes the target domain different.

In ChemProt dataset, the frequently masked words are the same words as the labels for downstream tasks. This means that Difference-Masking is determine a SSL objective that is highly similar to the downstream task without accessing those downstream labels.

Thanks for Listening

CSCI-535: Multimodal Probabilistic Learning of Human Communication

Difference Masking: Choose What to Mask in Continued Pretraining

Presenter: Jingmin Wei. Feb 28, 2024