JAMDEC: Unsupervised Authorship Obfuscation using Constrained Decoding over Small Language Models

Jillian Fisher[1], Ximing Lu[2,3], Jaehun Jung[2], Liwei Jiang[2,3], Zaid Harchaoui[1], Yejin Choi[2,3]

[1]Department of Statistics, University of Washington, [2]Paul G. Allen School of Computer Science & Engineering, University of Washington, [3]Allen Institute for Artificial Intelligence

UNIVERSITY *of* WASHINGTON

Authorship Obfuscation

Task of altering a text to reduce the discovery of the author's identity.

Many settings require obscuring authorship…

- Blind Review for Scientific Papers
- Interaction on Mental Health Forums
- Anonymous Online Review

Current authorship obfuscation methods have challenges….

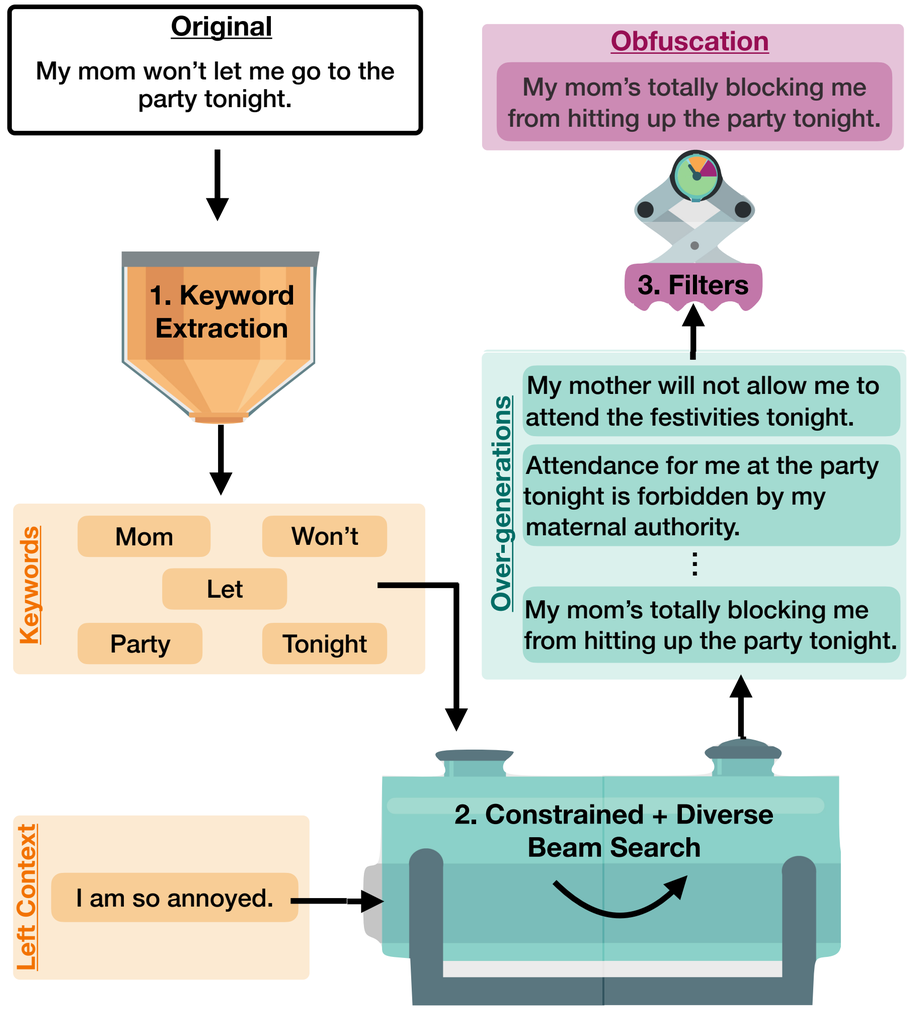
- Require Large Computing
- Involve Proprietary LLMs
- Require an Extra Authorship Corpus

JAMDEC Decoding

JAMDEC Decoding

- user-controlled, inference-time algorithm for authorship obfuscation that can be applied to any text and authorship without a separate authorship corpus

- **3 Stage Approach:**
  1. *Keyword Extraction*: Extract keywords to maintain original content
  2. *Constrained + Diverse Beam Search*: Augmented decoding strategy which encourages diverse but constrained generations
  3. *Filters*: Maintain fluency and content preservation, +any user-specified control



JAMDEC Decoding: Innovation

Keyword Extraction: Likelihood-based Method

- we select the top-k tokens with the lowest conditional probabilities, as measured by a specific language model, as keywords for a given sentence

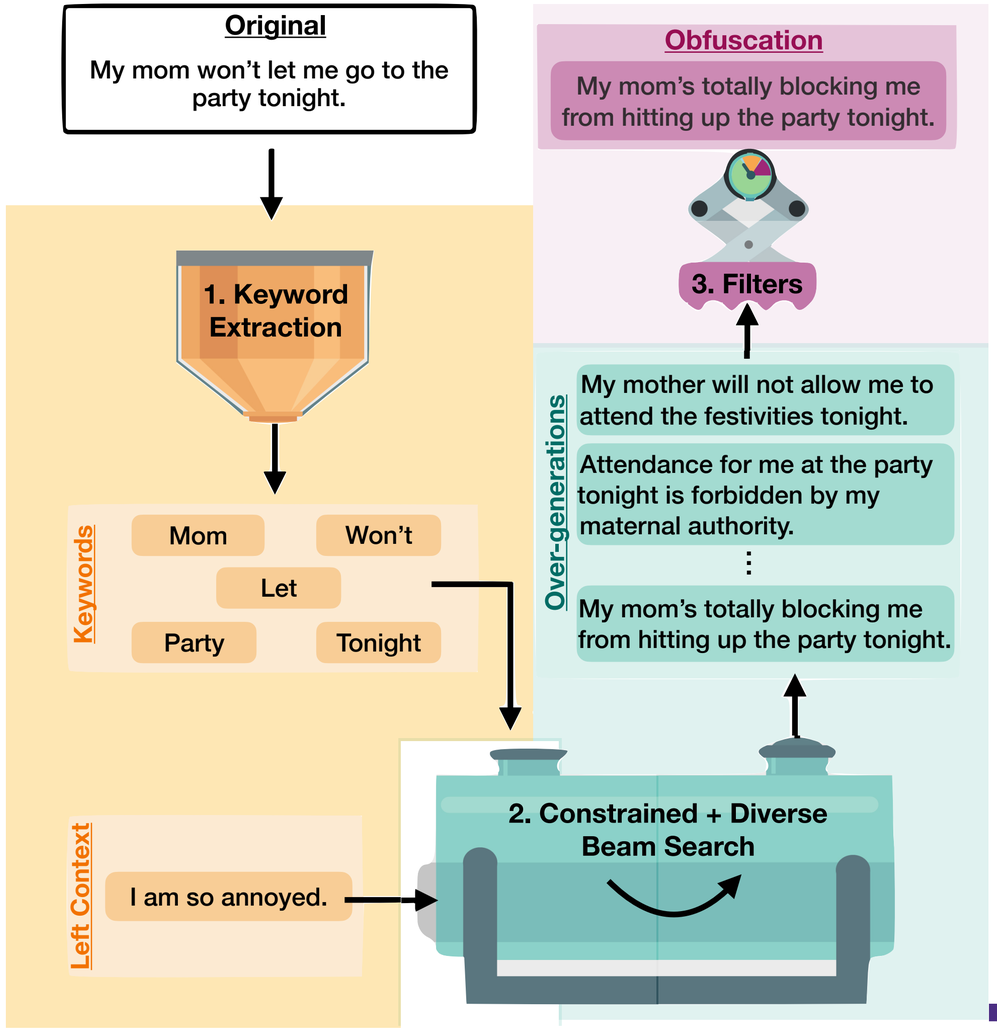
Constrained + Diverse Beam Search (over-generation)

- Constrained Beam Search is base algorithm, but uses the scoring function from Diverse Beam Search instead of likelihoods when iteratively selecting the top-k candidates from each bank.

$$\arg\max_{w \in W} P_w(y \mid x) + \lambda_1 D(y, Y) + \lambda_2 C(y)$$

Filtering

- Reduce pool and allow personalization of user



JAMDEC Decoding: Results

JAMDEC Decoding Experiments

- Two Datasets
  - AMT: formal scholarly passages
  - BLOG: diary-style entries
  - Number of Authors: 3,5, or 10
- Baselines
  - Mutant-X: Iteratively re-writes and combines randomly
  - Paraphrase
  - Round Trip Machine Translation
  - Stylometric: simple changes such as synonyms, number of words, punctuation, etc.
- Metrics
  - Obfuscation (Drop Rate: drop in accuracy compared to original)
  - Fluency: METEOR and CoLA
  - Content Preservation: NLI
  - Overall Metric: Task Score (average of Drop Rate, NLI, and CoLA)

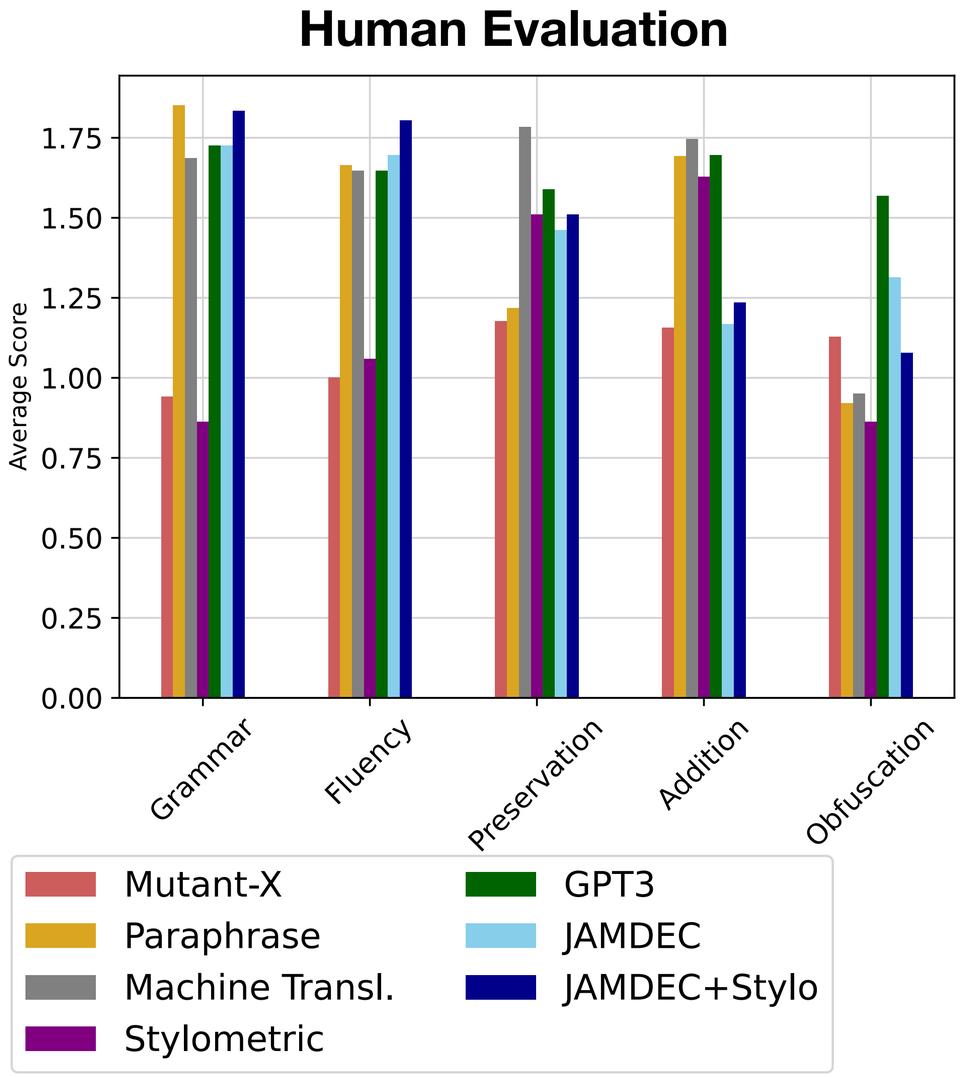
	Method	Mutant-X		Paraphrase	Machine Transl.	Stylometric	JAMDEC	
Dataset	**Metric**	*ENS*	*RFC*				*W/O Stylo*	*W/ Stylo*
AMT-3	Drop Rate (ENS)	⋆	-0.04	0.04	0.04	-0.03	**0.11**	**0.11**
	Drop Rate (BertAA)	0.10	0.04	0.04	0.08	**0.12**	0.04	0.04
	METEOR	0.80	**0.81**	0.55	0.69	0.80	0.62	0.62
	NLI	0.60	0.61	0.62	0.75	0.50	0.75	**0.81**
	CoLA	0.50	0.51	0.78	0.69	0.46	**0.85**	0.79
	Task Score (ENS)	⋆	0.36	0.48	0.49	0.31	**0.57**	**0.57**
	Task Score (BertAA)	0.40	0.39	0.48	0.51	0.36	**0.55**	**0.55**
AMT-5	Drop Rate (ENS)	⋆	0.08	0.20	0.20	**0.23**	0.10	0.13
	Drop Rate (BertAA)	0.07	0.00	-0.06	0.07	0.04	**0.14**	**0.14**
	METEOR	0.74	0.72	0.57	0.68	**0.79**	0.61	0.61
	NLI	0.56	0.57	0.62	0.74	0.48	0.76	**0.82**
	CoLA	0.51	0.55	0.77	0.69	0.46	**0.85**	0.79
	Task Score (ENS)	⋆	0.40	0.53	0.54	0.39	**0.57**	0.58
	Task Score (BertAA)	0.38	0.37	0.44	0.50	0.33	**0.58**	**0.58**
AMT-10	Drop Rate (ENS)	⋆	0.10	0.07	0.19	0.11	**0.44**	0.41
	Drop Rate (BertAA)	0.03	0.04	-0.04	**0.06**	0.00	-0.03	-0.02
	METEOR	0.84	**0.86**	0.54	0.66	0.81	0.60	0.61
	NLI	0.61	0.64	0.61	0.73	0.45	**0.79**	**0.79**
	CoLA	0.53	0.57	0.77	0.68	0.46	**0.78**	**0.78**
	Task Score (ENS)	⋆	0.44	0.48	0.53	0.34	**0.67**	0.66
	Task Score (BertAA)	0.39	0.42	0.45	0.49	0.30	0.51	**0.52**

* Similar results for the BLOG dataset (see paper for all results)

JAMDEC Decoding: Results

JAMDEC Decoding Experiments

- Two Datasets
  - AMT: formal scholarly passages
  - BLOG: diary-style entries
  - Number of Authors: 3,5, or 10
- Baselines
  - Mutant-X: Iteratively re-writes and combines randomly
  - Paraphrase
  - Round Trip Machine Translation
  - Stylometric: simple changes such as synonyms, number of words, punctuation, etc.
- Metrics
  - Human Evaluation



JAMDEC Decoding: Results

Performs similar to much larger models!

Method	GPT3.5		MASQUERADE	
Metric	*Sentence*	*Paragraph*	*W/O Stylo*	*W/ Stylo*
Drop Rate (ENS)	**0.23**	**0.23**	0.11	0.11
Drop Rate (BertAA)	**0.13**	0.09	0.04	0.04
METEOR	0.33	0.41	**0.62**	**0.62**
NLI	0.77	0.73	0.75	**0.81**
CoLA	0.76	0.80	**0.85**	0.79
Task Score (ENS)	**0.59**	**0.59**	0.57	0.57
Task Score (BertAA)	**0.55**	0.54	**0.55**	**0.55**

JAMDEC Decoding: Results

Method	Generation	
Original	The Ex. An ex holding a grudge can do a lot of damage in a short amount of time. He knows enough to open accounts in your name, and he has the motive to hurt you.	
Mutant-X	The Ex. An ex holding a **bitterness able ought** a lot of damage in a **length quantity** of time. He knows enough to **ascend** accounts in **Your prefix**, and he has the **justifiable to impair You.**	Ungrammatical
Paraphrase	**A lot of damage can be done In a short period of time.** He knows **how to** open accounts In your name and he **wants** to hurt you.	Incorrect Content
Machine Translation	**The former.** An **old man who holds a knife** can make a lot of damage in a short time. He knows enough to open accounts in your name, and he has the **reason** to hurt you.	Incorrect Content
Stylometric	An ex **holding, a** grudge can do a lot **inside damage** in a **brief** amount in time, **yet** he knows enough to open accounts in your name, and he has the motive to hurt you.	Missing Meaning
JAMDEC	The Ex. **When the ex is holding his grudge against the person who caused him lot of damage to his life, he is short sighted and will do anything in his power to get back at that person, no matter how much it will hurt the person he is trying to get revenge against.** He knows enough to open accounts in your name, and he has the motive to hurt you.	✓
JAMDEC + Stylo	The Ex. **When the ex is holding his grudge against the person who caused him lot of damage to his life, he is short sighted and will do anything in his power to get back at that person, no matter how much it will hurt the person he is trying to get revenge against.** He **believes** enough to open accounts in your name, and he has the **reason** to hurt you.	✓

More in the Paper

- **Comparison of trade-off between obfuscation, content-preservation, and grammaticality**
- **Ablation of JAMDEC Method (different beam width, with/without diversity, different filters, etc.)**
- **Comparison of “Style Transfer” methods**
- **Evaluation using “Adversarial Threat Models”**
- **Time Consumption Analysis (it’s competitive!)**
- **Discussion of similarity to other tasks (paraphrasing, style transfer, authorship attribution, etc.)**
- ***And MORE*!**

Thank You!

Paper

Code