Natural Language Processing: Inference, Privacy and Controllable Generation

Mark Draś
(with Shervin Malmasi, Chakaveh Saedi, Pedro Faustini, Natasha Fernandes, Annabelle McIver, Mark Johnson, Stephen Wan, . . .)

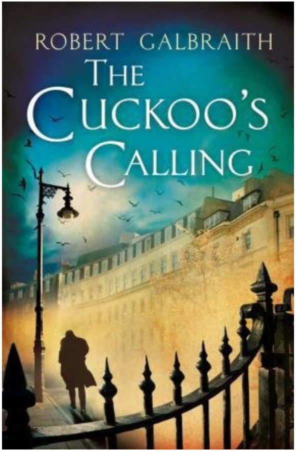
DataX
July 2023

Outline

Author Identification

Goal

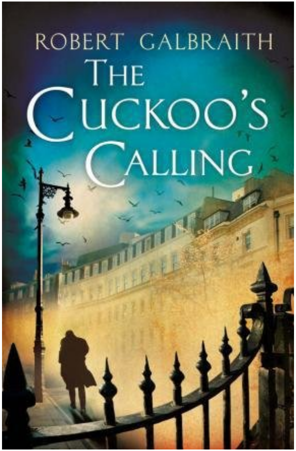
To identify the writer of a text, from among a set of candidates.



Author Identification

Goal

To identify the writer of a text, from among a set of candidates.



67 Nov 03, 2017 5:33:30 PM EDT
Anonymous ID: GVUvg1M7 No. 147816901

Where is John PODESTA?
Where is Tony PODESTA?
Did one or both escape the country and was let out?
WHERE IS BO?
WHERE WAS BO YESTERDAY?
What is the difference between commercial and private re: security clearance for departure?

4chan pol

Paul Furber @paul_furber · Mar 15, 2019
Questions the media should be asking:
- What is Tarrant's real background?
- Why did he travel to both NK and Pakistan recently?
- How did he get hold of so many restricted weapons?
- Why does his "manifesto" not ring true for a real 8chan denizen?

Author Profiling: Native Language Identification

Goal

To identify the native language (L1) of the writer of a given text written in a second language (L2).

Examples

1. *The development of* **country park** *can directly alleviate overcrowdedness*
2. **The English** *is a difficult language*
3. *We located* **the accomodation** *near the sea*

NLI: Use Cases

Security

```
[Title] Message to SONY
We have already given our clear demand to the management team of SONY,however,
they have refused to accept.
It seems that you think everything will be well,if you Bind out the attacker,while no
reacting to our demand.
We are sending you our warning again.
Do carry out our demand if you want to escape us.
And,Stop immediately showing the movie of terrorism which can break the regional
peace and cause the War!
You,SONY & FBI,cannot Bind us.
We are perfect as much.
The destiny of SONY is totally up to the wise reaction & measure of SONY.
```

Language Acquisition and Learning

- Research in Second Language Acquisition: "What is the role played by first language in L2 development, *vis-à-vis* the role of other universal development forces?" (Ortega, 2009)
- Research in Second Language Learning: What phenomena might be problematic for speakers of a particular L2?
  - e.g. Missing determiner errors, Korean speakers (1.1 / 10 sentences) vs German speakers (0.4 / 10 sentences).

NLI: Background

Major Datasets

- International Corpus of Learner English (ICLE)
- Test of English as a Foreign Language (TOEFL11)
  - 11 languages, 1100 essays per language, balanced across topic
- Reddit-derived dataset

Analysis of Shared Tasks

- 2013: TOEFL11 essays; 2017: essays + speech.
- For 2013:
  - Highest accuracy: 83.0%.
  - Most popular learners: logistic regression, SVMs.
  - Consistently useful: ensembles of learners (although use was ad hoc).

Ensemble Architectures

Goal

Systematically investigate ensemble architectures for combining learners.

- Draw on other areas of machine learning.

[Shervin Malmasi and Mark Dras (2018). Native Language Identification using Classifier Ensembles and Stacking. *Computational Linguistics*, 44(3):403-446].

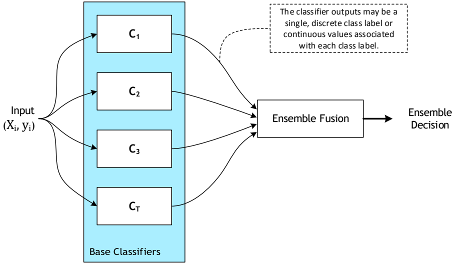


Figure: An example of a parallel ensemble classifier architecture where T independent classifiers provide predictions which are then fused using a rule-based ensemble combination method. The class labels for the input, y_i, are only available during training or cross-validation.

Ensemble Architectures

Components

- Base learners:
  - Logistic regression and linear SVMs.
  - Single feature type.
- Ensemble architecture types:
  - Static ensemble fusion methods (e.g. plurality voting).
  - Meta-classifiers / classifier stacking (e.g. linear SVMs).
  - Meta-classifier ensemble (decision tree ensembles, bagging).

Results

- Cf. 9.1% accuracy random guess, shared task winner 83.0% ...
  - Mean probability was the best static ensemble fusion method (83.3%).
  - Metaclassifiers were better, with LDA the best meta-learner (86.8%).
  - Metaclassifier ensembles were the best approach, again with LDA best for bagging (87.1%).
- Results state-of-the-art, and useful for applications.

Authorship Identification

Goal

To identify the writer of a text, from among a set of candidates.

Examples

Text with author to be determined:

1. gaze. Now was his opportunity. “You go to the Marquee,” he said. “I need to see to something,” Luc

Texts from candidate authors:

1. uggested in reality or outside the story. In the Great Britain wizarding world, Concubines have been
2. of clear water as they reached a tree line, his sodden hair plastered to his forehead, tufts behind
3. don’t you?” “Don’t I, what?” “Have fun?” “Oh yes, I will clear up some time in my busy schedule of n

Authorship Identification

Goal

To identify the writer of a text, from among a set of candidates.

Examples

Text with author to be determined:

1. gaze. Now was his opportunity. “You go to the Marquee,” he said. “I need to see to something,” Luc

Texts from candidate authors:

1. uggested in reality or outside the story. In the Great Britain wizarding world, Concubines have been
2. of clear water as they reached a tree line, his sodden hair plastered to his forehead, tufts behind
3. don’t you?” “Don’t I, what?” “Have fun?” “Oh yes, I will clear up some time in my busy schedule of n

Authorship Identification: Background

Datasets

- Several available from PAN shared tasks.
- We constructed one from online fanfiction.

Approaches

- Conventionally divided into classification-based, similarity-based.
  - Similarity-based better for large number of candidate authors.
  - Also better when candidate set is open, and where authors have not been seen before.
- Similarity-based had fixed notion of similarity.

Siamese Architectures

Goal

To learn a notion of similarity for authorship identification, using deep Siamese architectures.

[Chakaveh Saedi and Mark Dras (2021). Siamese Networks for Large-Scale Author Identification. *Computer Speech and Language*, 70:101241.]

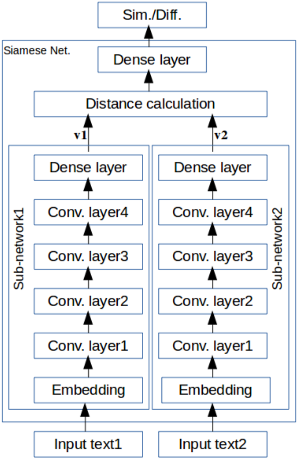


Figure: Siamese network architecture with CNN subnetworks.

Siamese Architectures

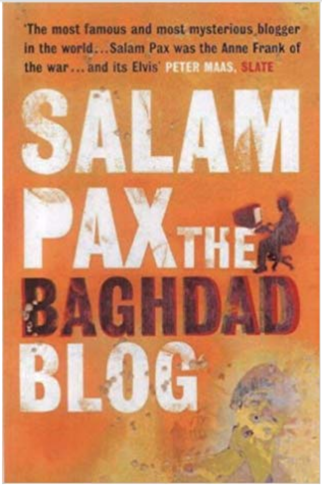
Experimental Setup

- Learn similarity on training set with 1000 authors.
- From test set, pick author of unknown text from among N candidate authors' texts.
- Authors in test set *have not been seen before*.

Results

- For $N = 10$, cf. 10% accuracy random guess, 44% similarity baseline:
  - Best Siamese model gets 94.3%.
  - Cosine similarity better than other metrics (e.g. L_1).
  - CNN better than LSTM for subnetworks.
- Results useful for applications.

Privacy Risks



Goal

Prevent inference methods from inferring authorship.

Outline

Outline

1. Making Inferences about People via Text

2. **Privacy in NLP**
   - Privacy Risks in NLP with Solutions: Three Examples
   - Differential Privacy and NLP Applications
   - Differential Privacy with Metrics

3. Text Generation

4. Wrap-Up

Author Attributes in Representations

Younger

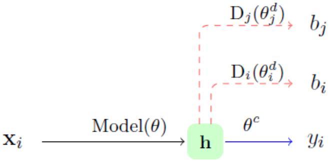
It was air-conditioned, and drove pretty well, once I'd sussed out the bite point (which I had to do after stalling on a hill — a bit scary!)

Older

I've relied on Vodafone since mobile phones were the size and weight of bricks . . .

Summary

- **Data:** TrustPilot reviews.
- **Privacy:** Identify age / gender / location from learned representations.
- **Utility:** PoS, sentiment.
- **Solution:** Adversarial training of classifier.



[Yi, Baldwin, Cohn (2018). Towards Robust and Privacy-preserving Text Representations. *Proc. ACL*.]

Author Attributes in Text

Teen

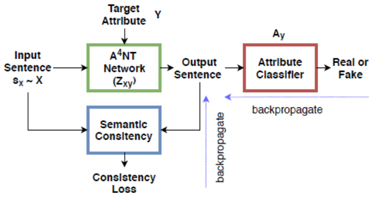
i don't know y i even went into dis relationship

Adult

my first day going out to see clients after vacation.

Summary

- **Data:** Blog posts, speech.
- **Privacy:** Identify age / gender, identity from source text.
- **Utility:** Semantic consistency.
- **Solution:** Adversarial training via NMT seq2seq.



[Shetty, Schiele, Fritz (2018). A[4]NT: Author Attribute Anonymity by Adversarial Training of Neural Machine Translation. *Proc. USENIX Security Symposium*.]

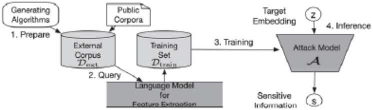
Personally Identifying Information in LMs

Flight Review

I was very unhappy with your flight Delta XXX from *Frankfurt* to *Bangkok* on ...

Summary

- **Data:** Various (e.g. airline reviews).
- **Privacy:** Identify sensitive contents from e.g. BERT representations, such as location, destination.
- **Solution:** Various (tentative).



[Pan, Zhang, Ji, Yang (2020). Privacy Risks of General-Purpose Language Models. *Proc. IEEE Security & Privacy*.]

In General . . .

Idea

- All of these are evaluated empirically.
- No guarantees.

⇒ Consider an approach with guarantees . . .

Outline

Intuition

Idea from Social Science: Randomised Response

Developed to collect statistical information about embarrassing or illegal behavior, captured by having a property P. Study participants are told to report whether or not they have property P as follows:

1. Flip a coin.
2. If tails, then respond truthfully.
3. If heads, then flip a second coin and respond "Yes" if heads and "No" if tails.

Privacy comes from the plausible deniability of any outcome.
Utility comes from being able to recover accuracy, by understanding of the noise generation procedure.

Differential Privacy

Characteristics

- Applied to statistical databases.
- Works by considering *adjacent* databases (ones that differ by a single row/entry).
- **View of privacy**: Differential privacy promises to protect individuals from any *additional* harm that they might face due to their data being in the private database x that they would not have faced had their data not been part of x.
- Achieves privacy by a mechanism that adds noise.

[Dwork et al (2014). *The algorithmic foundations of differential privacy*.]

Example

- Query: What's the mean salary of employees in the database?
- Add noise to reported average salary, as in randomised response.

Differential Privacy

Definition

A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ϵ, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$Pr[\mathcal{M}(x) \in S] \leq \exp(\epsilon) Pr[\mathcal{M}(y) \in S] + \delta,$$

where databases x, y are collections of records from a universe $\mathcal{X}$, and the probability space is over the coin flips of the mechanism $\mathcal{M}$.

Example Mechanism

Add Laplace noise.

Application #1: DP Training of Neural Nets

Motivating Problem

Training neural networks.

- Want to prevent e.g. membership inference or reconstruction attacks.

DP Definition for Problem

Privacy for single examples in training set.

Method for Applying DP

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.
Initialise θ_0 randomly
for $t \in T$ **do**
  $L_t \leftarrow$ random sample of L indices from $1 \ldots N$ ▷ Take a random batch
  for $i \in L_t$ **do**
    $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$ ▷ Compute gradient vector
    $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i)/\max(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C})$ ▷ Clip gradient vector
  end for
  $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\sum_i(\bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, C^2\sigma^2))$ ▷ Add noise
  $\theta_{t+1} \leftarrow \theta_t - \eta_t\tilde{\mathbf{g}}_t$ ▷ Descent
end for
Output θ_T

[Abadi et al (2016). Deep Learning with Differential Privacy. *Proc. ACM CCS*.]

Application #2: DP Language Models

Motivating Problem

Training LSTM models for next-word prediction in a mobile keyboard.

- Sensitive information might include passwords, text messages, and search queries; may also identify a speaker.

DP Definition for Problem

User-level privacy, rather than privacy for single examples.

- User-adjacent datasets differ by all examples by given user.

Method for Applying DP

Key components:

- federated training on user-partitioned data;
- moments accounting method of Abadi et al (2016) to provide tight composition guarantees on noise from Gaussian DP mechanism.

[McMahan, Ramage, Talwar, Zhang (2018). Learning Differentially Private Recurrent Language Models. *Proc. ICLR*.]

Why You Might Want to Extend DP

General Reason

"In some situations the distinguishability level between [datasets] x and x' should depend not only on the *number* of different values between x and x', but also on the *values themselves*."

- Implications for effect on utility.

Example

- Consider an employee database containing salaries.
- In standard DP, the distinguishability between two employees with salaries \$1, \$1m will be treated the same as two other employees with salaries \$20K, \$20.001K.

$\Rightarrow$ Add worst-case noise.

[Chatzikokolakis et al (2013). Broadening the Scope of Differential Privacy Using Metrics. *Proc. PETS*.]

For NLP

Can consider our (high-dimensional) representations as these 'values'.

Outline

1 Making Inferences about People via Text

2 Privacy in NLP
- Privacy Risks in NLP with Solutions: Three Examples
- Differential Privacy and NLP Applications
- Differential Privacy with Metrics

3 Text Generation

4 Wrap-Up

Main Ideas

DP with Metrics: Characterisation

- Standard DP implicitly uses Hamming distance.
- Indistinguishability requirement can be generalised to an arbitrary notion of distance.

[Chatzikokolakis et al (2013). Broadening the Scope of Differential Privacy Using Metrics. *Proc. PETS*.]

Geoindistinguishability

- Add noise to user (Euclidean) location.
- Aim at *geoindistinguishability*: quasi-indistinguishability within a certain area.
- From attacker's PoV, user is almost equally likely to be anywhere within a certain radius r from actual location.

[Andrés et al (2013). Geo-indistinguishability: differential privacy for location-based systems. *Proc. ACM CCS*.]

Geoindistinguishability

Definition

Let $\mathcal{X}$ be a set of points of interest and $\mathcal{Z}$ be a set of possible reported values; and let $x, x' \in \mathcal{X}$ be locations such that $d(x, x') \leq r$, where $d(\cdot, \cdot)$ denotes Euclidean distance.

A mechanism $\mathcal{K}$ satisfies ϵ-geo-indistinguishability iff for all $x, x' \in \mathcal{X}$ and $Z \subseteq \mathcal{Z}$:

$$\mathcal{K}(x)(Z) \leq e^{\epsilon d(x,x')} \mathcal{K}(x')(Z)$$

Intuition

The attacker assigns similar probabilities to the user being located in x or x' after observing reported points Z.

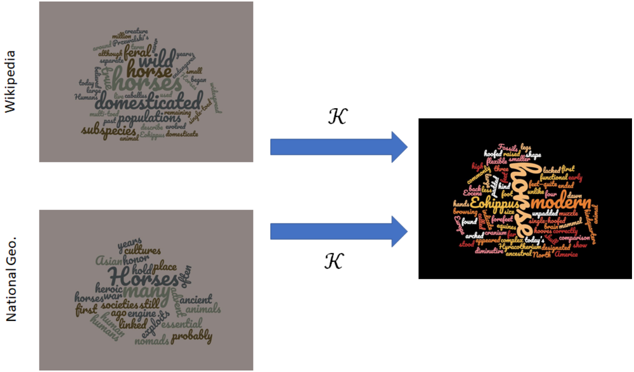
Privacy Mechanism

Extend Laplace mechanism to two dimensions.

NLP Application: Authorship Privacy

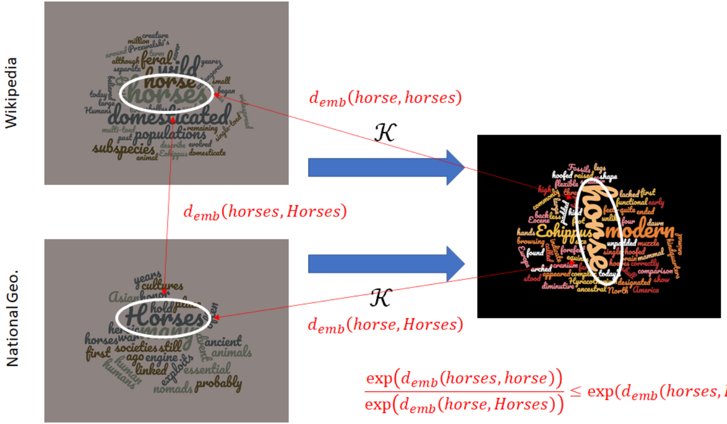
Goal

We want (ultimately) to create a mechanism $\mathcal{K}$ which paraphrases a document i.e. preserving much of its meaning, while giving the actual writer of the document some kind of "plausible deniability".



[Fernandes, Dras, McIver (2019). Generalised Differential Privacy for Text Document Processing. *Proc. POST*.]

Document Privacy: Private Bags of Words



$$\frac{\exp(d_{emb}(horses, horse))}{\exp(d_{emb}(horse, Horses))} \leq \exp(d_{emb}(horses, Horses))$$

Word Mover's Distance

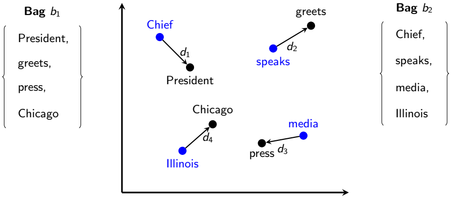
Definition

The Word Mover's Distance is the minimum cost of moving all of the words in the source document to all of the words in the destination document.

Example

$x = \{$ Obama, speaks, Illinois, media $\}$
$y = \{$ President, greets, press, Chicago $\}$

Kantorovich distance:
$|x, y|_T := d_1 + d_2 + d_3 + d_4$



[Kusner, Sun, Kolkin, Weinberger (2015). From word embeddings to document distances. *Proc. ICML*.]

A Mechanism for Private Bags of Words

Mechanism

- For each word w in document d, use word embedding distance and the Laplace distribution to replace w with a randomly chosen word w' so that the probability of choosing w is proportional to

$$e^{d_{\mathrm{emb}}(w,w')}$$

- Apply this *independently* to each word in d.
- Using the *triangle inequality on metrics* and, provided that d and d' have the same number of words, we can prove:

$$\mathcal{K}(d)(Z) \leq e^{\epsilon |d,d'|_T} \mathcal{K}(d')(Z)$$

Experiments

- Fanfiction dataset, 'topic' $\equiv$ fandom (e.g. Harry Potter, Twilight).
- Maintained utility while preserving privacy.

Follow-Ons (1)

GetWeather

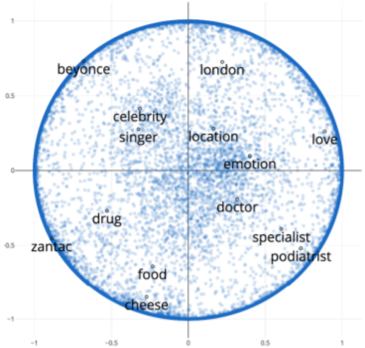
will it be colder in ohio

PlayMusic

play techno on lastfm

Summary

- **Data:** User queries.
- **Privacy:** Identify user identity, location.
- **Utility:** Semantic consistency, query type.
- **Solution:** DP with *hyperbolic* metric.



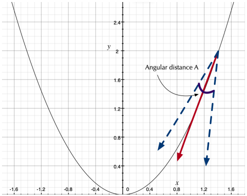
[Feyisetan, Diethe, Drake (2019). Leveraging Hierarchical Representations for Preserving Privacy and Utility in Text. *Proc. ICDM*.]

Follow-Ons (2)

Idea

Metric DP for training neural nets.

- Use direction (angle) as metric.



DP with Von-Mises Fisher Noise

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$.
Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound $C = 1$.
Initialise θ_0 randomly
for $t \in T$ **do**
  $L_t \leftarrow$ random sample of L indices from $1 \ldots N$ ▷ Take a random batch
  for $i \in L_t$ **do**
    $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$ ▷ Compute gradient vector
    $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \frac{\|\mathbf{g}_t(x_i)\|_2}{C}$ ▷ Scale gradient vector
  end for
  $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\sum_i \mathcal{V}(\sigma, \bar{\mathbf{g}}_t(x_i))$ ▷ Add noise
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$ ▷ Descent
end for
Output θ_T

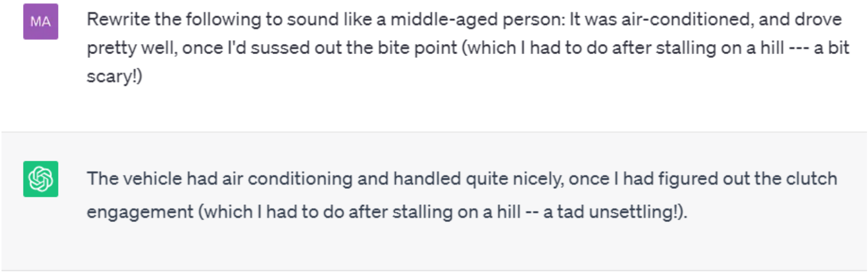
[Faustini, Fernandes, Tonni, McIver, Dras (2023). Directional Privacy for Deep Learning. Under submission.]

Outline

Rewriting for Style (1)

Using Large Language Models (LLMs)

Example using ChatGPT.



Rewriting for Style (2a)

MA

Rewrite the following in the style of a research scientist with citations, including hyperlinks to articles: Have you heard about smart computer programs? Generative Artificial Intelligence (AI) is one of them! It can create pictures, music, and stories all on its own.

It's like having a super creative friend who can make amazing things that nobody has ever seen or heard before. It uses its knowledge and imagination to create brand-new and original stuff.

Rewriting for Style (2b)

Certainly! Here's the revised version with scientific language, citations, and hyperlinks to relevant articles:

"Are you familiar with the paradigm of intelligent computer programs? Generative Artificial Intelligence (AI), a notable exemplar in this domain, has emerged as a powerful tool capable of autonomously producing a diverse array of creative artifacts, encompassing imagery, music, and narratives (Smith et al., 2020; Johnson, 2019). This technology, analogous to an extraordinarily imaginative comrade, possesses the prowess to craft awe-inspiring creations that defy prior human experiences (Gupta et al., 2021). Leveraging its robust knowledge base and imaginative faculties, generative AI generates entirely novel and pioneering content that pushes the boundaries of artistic expression and innovation" (see Smith et al., 2020; Johnson, 2019; Gupta et al., 2021).

For more information on the subject, you may find the following articles helpful:

Smith, A., Davis, B., & Chen, C. (2020). Advances in Generative Artificial Intelligence. Journal of Computational Creativity, 14(2), 205-218. Read here

Johnson, R. W. (2019). The Role of Generative AI in Artistic Expression. International Journal of Creative Computing, 7(1), 45-62. Read here

Problem

Correctness of Generated Text

- Hallucination.
  - One definition: "confident statements that are not true".
- Missing important information.

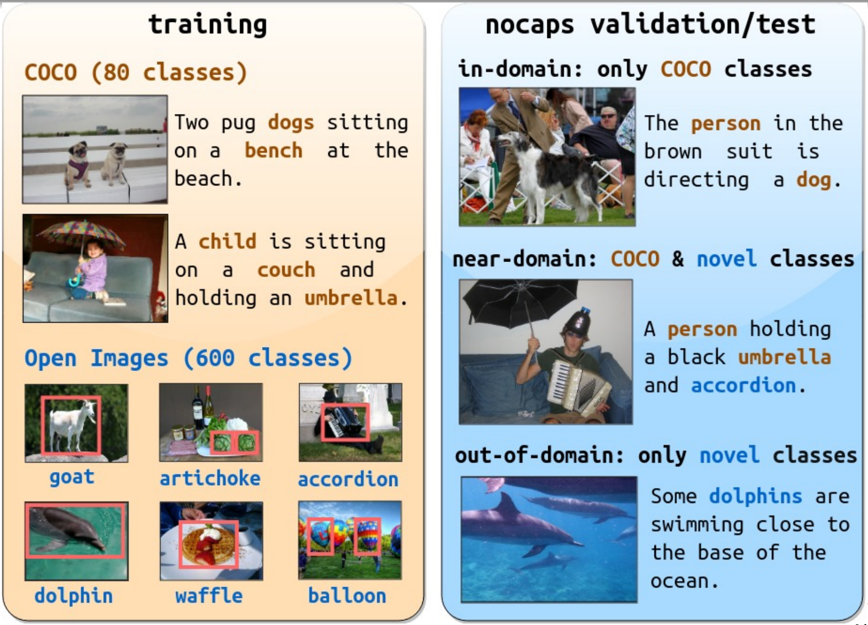
Idea

Force model to produce true things

- Introduce 'Mention Flags' into models.

[Wang, Wood, Wan, Dras, Johnson (2021). Mention Flags (MF): Constraining Transformer-based Text Generators. *Proc. ACL*.]

Data and Tasks: nocaps



Mention Flags Architecture (1)

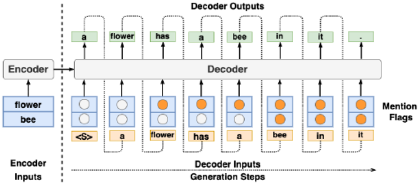


Figure 1: An overview of the Mention Flag mechanism for Transformer-based S2S models. Here, the tokens *flower* and *bee* are required to appear in the generated outputs. Each generated token has a corresponding set of Mention Flags which informs the decoder whether each lexical constraint has been satisfied in the current decoder input sequence. For example, the Mention Flag for *flower* is set (indicated by orange dots) from the third token because it is generated at the second step. Both token and Mention Flag embeddings are the input to the decoder, but Mention Flags are injected into the decoder in a different way to the tokens (see Fig. 3). Note that task specific encoder inputs have been omitted for brevity.

Mention Flags Architecture (2)

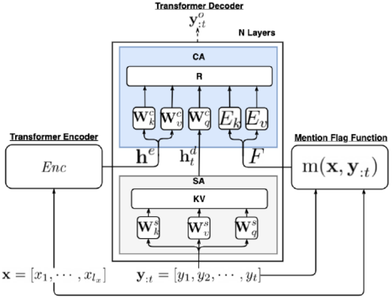


Figure 3: In each decoder layer, the Cross-Attention (CA) module (light blue) integrates Mention Flags as additional inputs describing relationship between encoder contents and decoder input tokens. There are separated representations for Mention Flags in different decoder layers.

Experimental Results

Got state-of-the-art results against strong baselines.

Outline

Wrap-Up

Summary

- You can find out a lot from text that people write.
- Some approaches to privacy can help protect against this kind of inference.
  - Still very much an unsolved problem.
- New LLMs can change stylistic clues, but changes can introduce errors.
  - Relatively early stages of fixing these.