Carnegie Mellon University
Language Technologies Institute

Google AI

A Survey of Data Augmentation Approaches for NLP

Steven Y. Feng*[1], **Varun Gangal*[1]**, Jason Wei[2], Sarath Chandar[3], Soroush Vosoughi[4], Teruko Mitamura[1], Eduard Hovy[1]

[1]Language Technologies Institute, Carnegie Mellon University

[2]Google Research, [3]Mila – Quebec AI Institute, [4]Dartmouth College

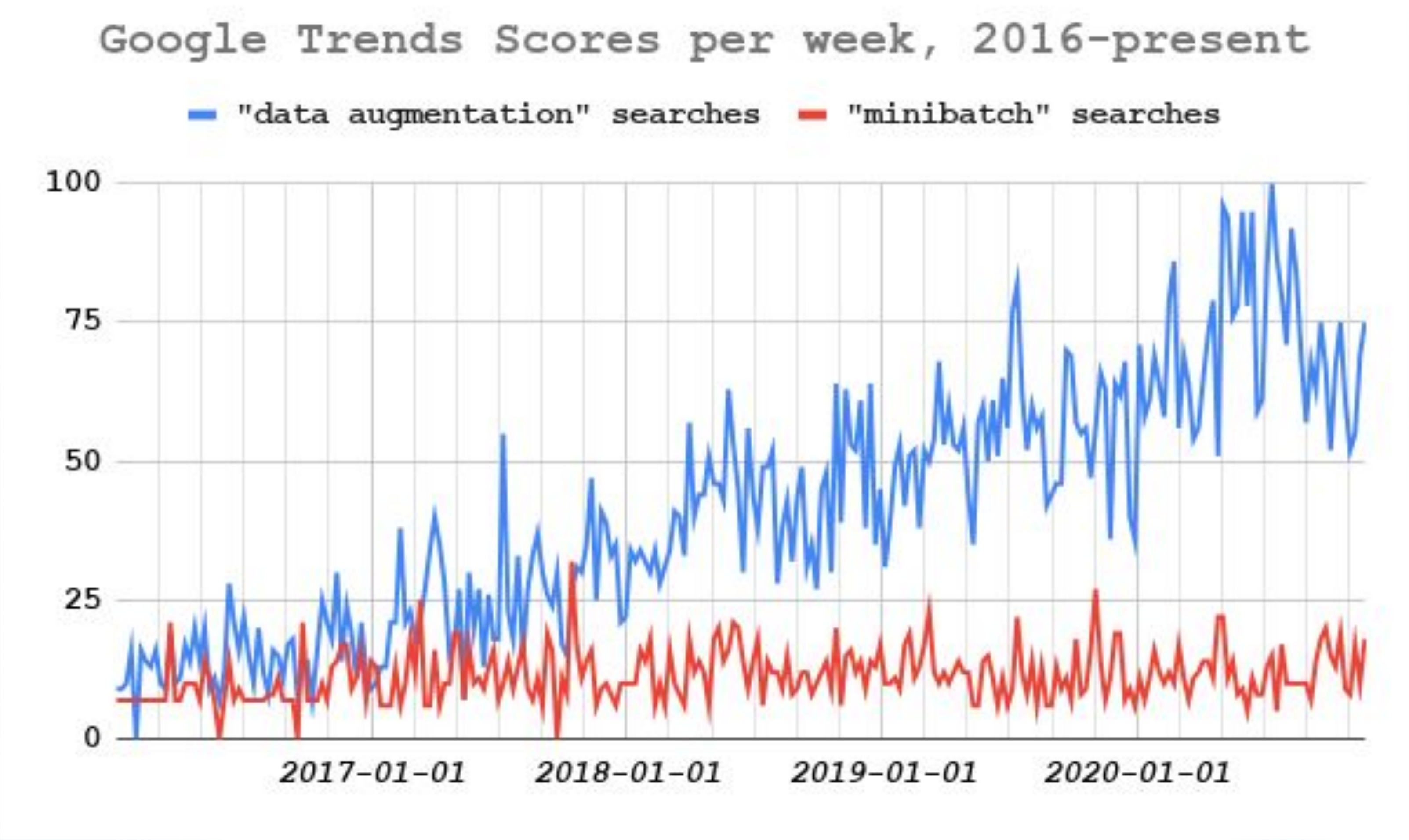
ACL 2021 Findings

Mila

https://github.com/styfeng/DataAug4NLP

DARTMOUTH

Motivation



Google Trends Scores per week, 2016-present
"data augmentation" searches
"minibatch" searches
100
75
50
25
0
2017-01-01
2018-01-01
2019-01-01
2020-01-01

Paper Structure

- Background on Data Augmentation (DA)
- Methodologically Representative DA Techniques
- Useful NLP Applications for DA
- DA Methods for Common NLP Tasks
- Challenges and Future Directions

What is Data Augmentation?

- Methods of increasing training data diversity without directly collecting more data
- Pervasive for Computer Vision, more difficult for NLP where the input space is discrete

Why Challenging for NLP?

- Harder to maintain desired invariances
- Desired invariances are less obvious
- Hard to encode invariances directly into the model or as a lightweight module to apply during training itself
- For NLP, usually generate data and store offline
- Desired invariances can differ substantially across tasks

What Makes a Good DA Technique?

- Ideally, both easy-to-implement and improves performance
  - Rule-based techniques easy but lower gains
  - Model-based techniques more difficult but higher gains
- Balanced distribution of augmented data
  - Not too similar and not too different from original data

Rule-Based DA Techniques

- Uses easy-to-compute and predetermined transforms
- Examples:
  - Easy Data Augmentation (EDA)[1]
  - Unsupervised Data Augmentation (UDA)[2]
  - Dependency Tree Morphing[3]

Dependency Tree Morphing

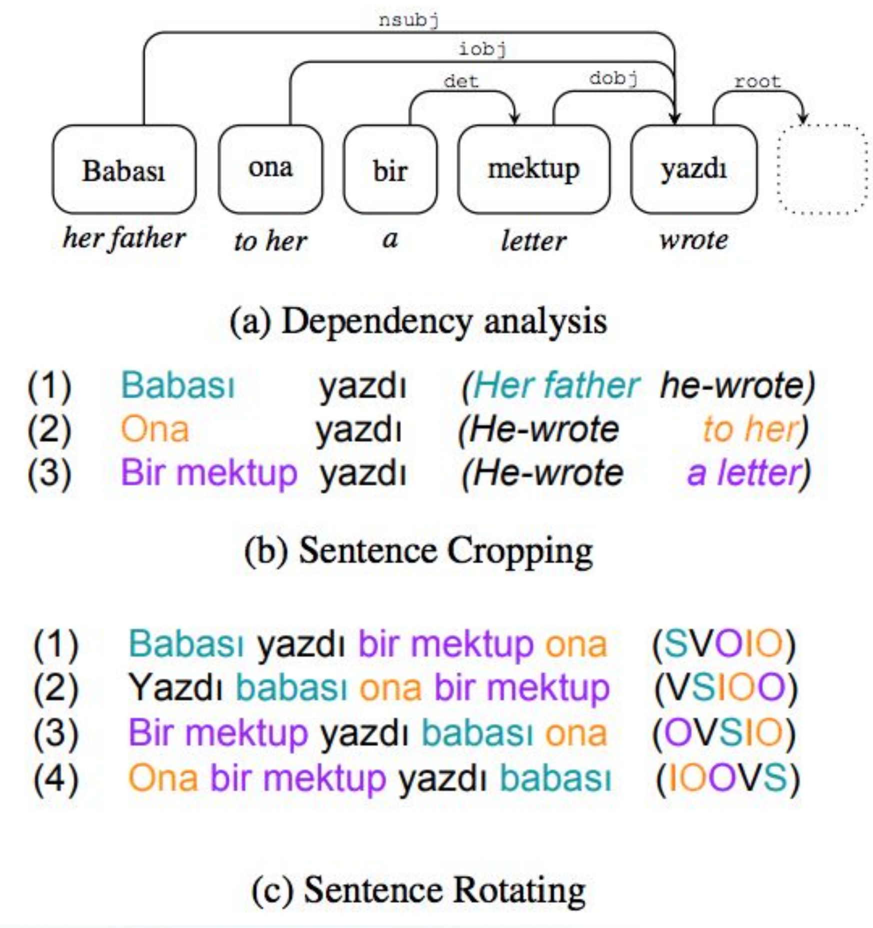
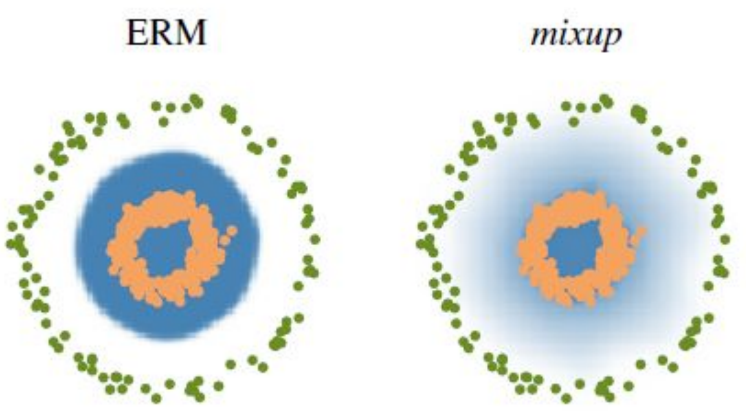


Figure 2: *Dependency tree morphing* DA applied to a Turkish sentence, Şahin and Steedman (2018)

- Augments dependency annotated sentences
- Rotation: children of the same parent are swapped
- Cropping: some children of the same parent are deleted
- Most beneficial for rich case marking system languages (e.g. Baltic, Slavic, Turkic)

Example Interpolation DA Techniques

- Interpolates the inputs and labels of two or more examples
- AKA Mixed Sample Data Augmentation (MSDA)
- Pioneered by MixUp[4]



(b) Effect of *mixup* ($\alpha = 1$) on a toy problem. Green: Class 0. Orange: Class 1. Blue shading indicates $p(y = 1|x)$. From Zhang et al. (2018)

Model-Based DA Techniques

- DA techniques relying on seq2seq and language models
  - E.g. Backtranslation[5]
    - Source language → target language → source language
  - E.g. Contextual Augmentation[6]
  - E.g. Semantic Text Exchange (STE)[7]

Contextual Augmentation

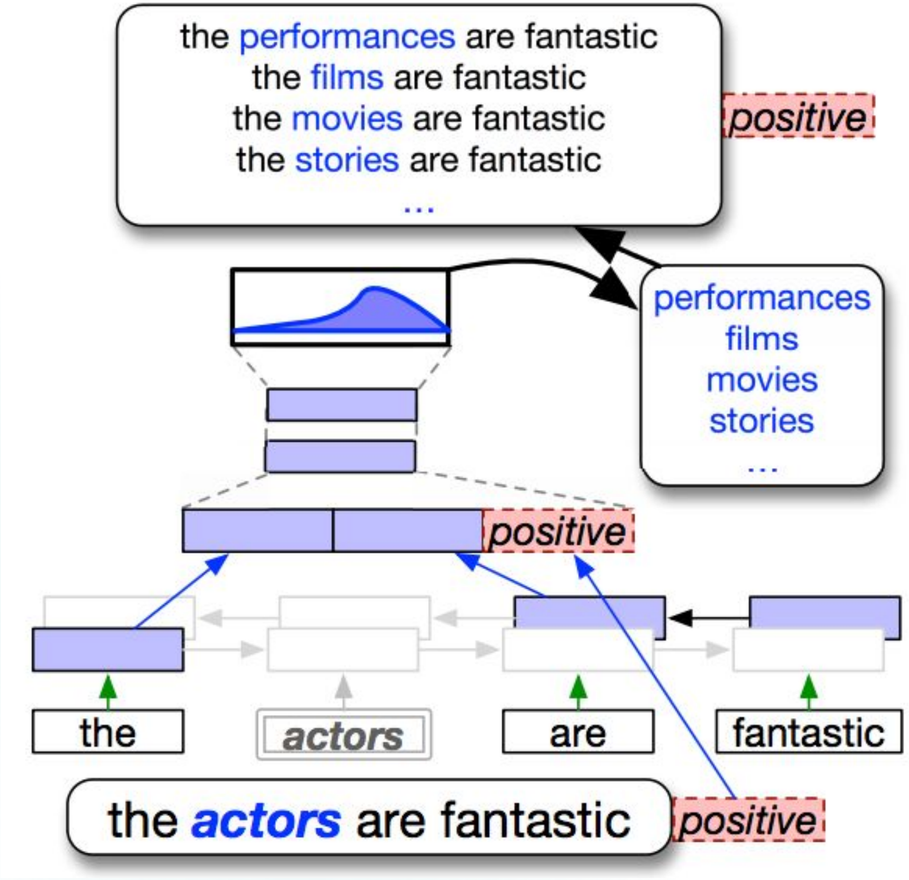
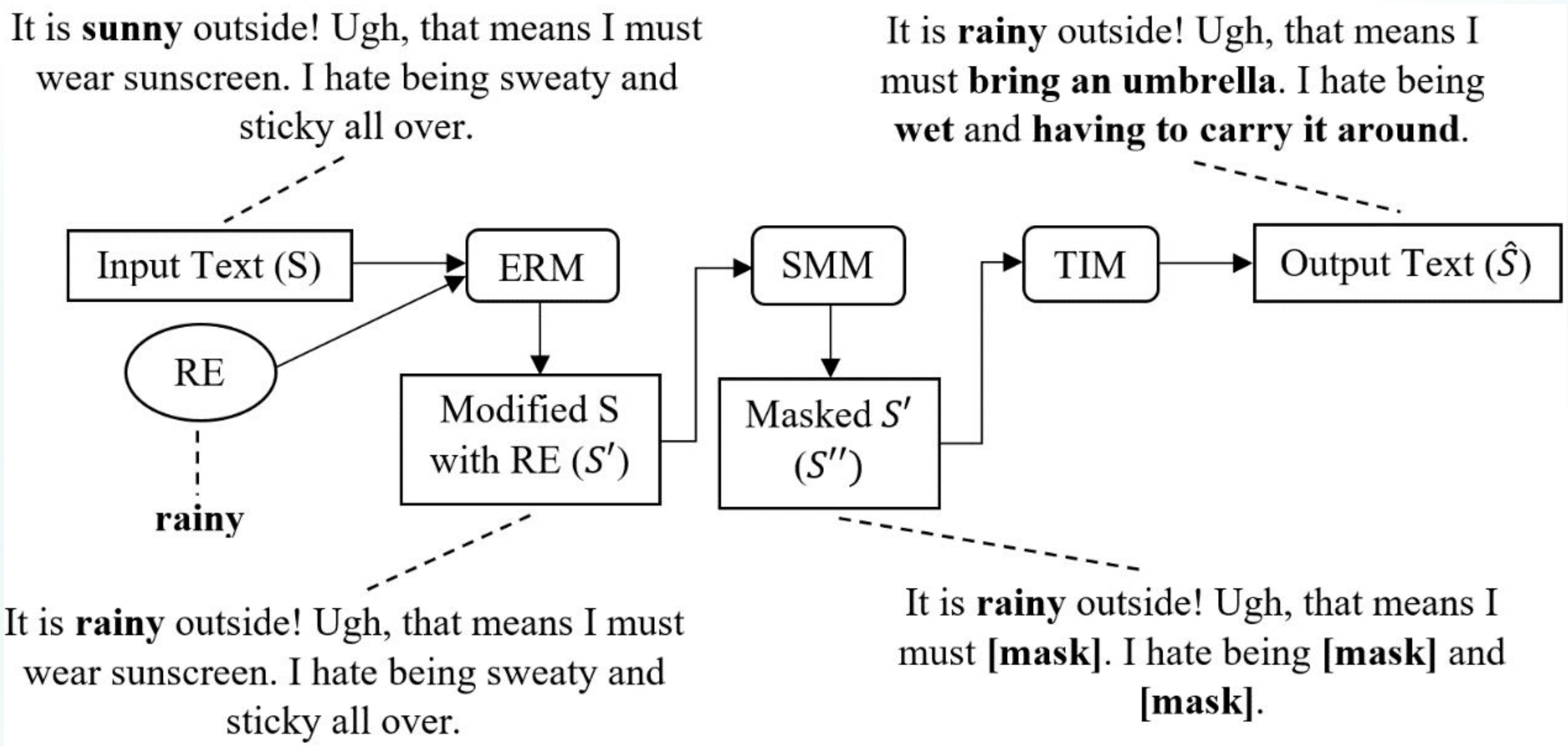


Figure 3: *Contextual Augmentation*, Kobayashi (2018)

- Replace words with randomly drawn other words
- Drawn from the recurrent language model's distribution
- This distribution is based on the current context of the word

Semantic Text Exchange (STE)

- New task proposed in *Keep Calm and Switch On! Preserving Sentiment and Fluency in Semantic Text Exchange (Feng et al., EMNLP '19)*



- Uses SMERTI pipeline: entity replacement, similarity masking, text infilling

Semantic Text Exchange (STE)

- For DA: can replace noun keywords/phrases
- Entity that replaces (RE): another noun keyword/phrase
- Intuition: alters semantics of the entire text with respect to a particular topic

DA for NLP Applications

- Low-Resource Languages
- Mitigating Bias
- Fixing Class Imbalance
- Few-Shot Learning
- Adversarial Examples

CoSDA-ML: Multi-Lingual Code-Switching DA

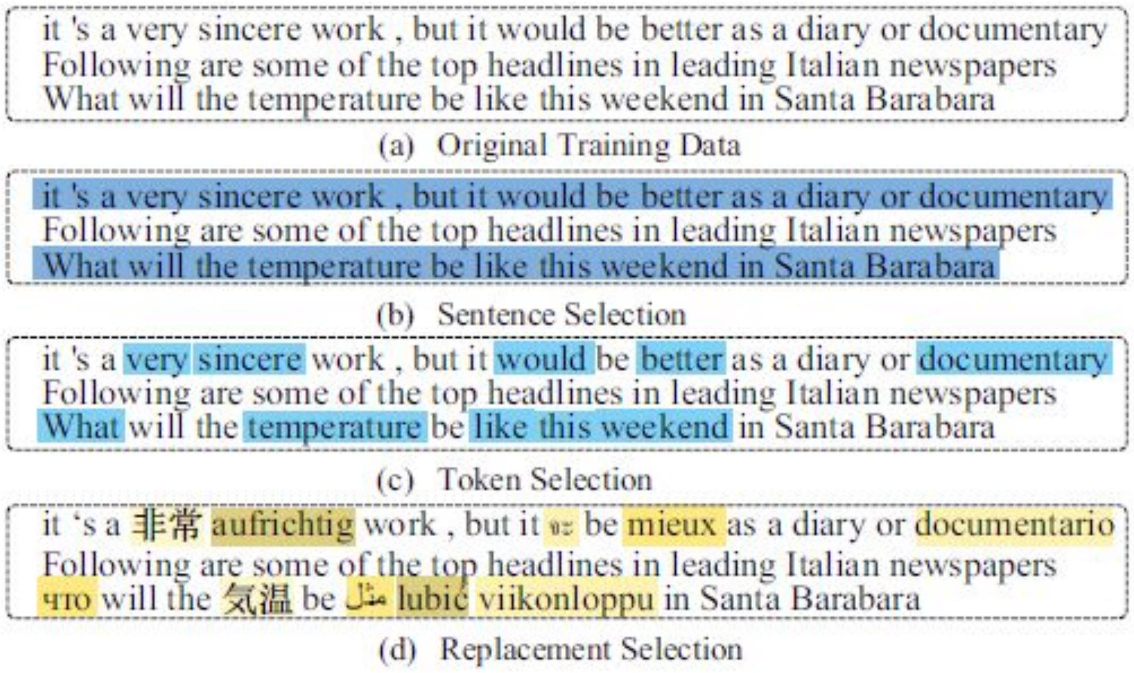


Figure 2: Augmentation process. The source language sentences (a), the sentence selection step (b), the token selection step (c) and the replacement selection step (d) (different shades yellow colors in (d) represent different languages translation).

- Generate multilingual code-switching data
- Purpose: finetune multilingual BERT
- Encourage alignment of representations from source and multiple target languages
- How? By mixing their context information
- Obtain improved performance across 5 tasks with 19 languages

CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP (Qin et al., IJCAI 2020)[8]

DA for Common NLP Tasks

- Summarization
- Question Answering (QA)
- Sequence Tagging Tasks
- Parsing Tasks
- Grammatical Error Correction
- Neural Machine Translation
- Data-to-Text NLG
- Open-Ended Text Generation
- Dialogue
- Multimodal Tasks

DA Method	Ext.Know	Pretrained	Preprocess	Level	Task-Agnostic
SYNONYM REPLACEMENT (Zhang et al., 2015)	✓	×	tok	Input	✓
RANDOM DELETION (Wei and Zou, 2019)	×	×	tok	Input	✓
RANDOM SWAP (Wei and Zou, 2019)	×	×	tok	Input	✓
BACKTRANSLATION (Sennrich et al., 2016)	×	✓	Depends	Input	✓
SCPN (Wieting and Gimpel, 2017)	×	✓	const	Input	✓
SEMANTIC TEXT EXCHANGE (Feng et al., 2019)	×	✓	const	Input	✓
CONTEXTUALAUG (Kobayashi, 2018)	×	✓	-	Input	✓
LAMBADA (Anaby-Tavor et al., 2020)	×	✓	-	Input	×
GECA (Andreas, 2020)	×	×	tok	Input	×
SEQMIXUP (Guo et al., 2020)	×	×	tok	Input	×
SWITCHOUT (Wang et al., 2018b)	×	×	tok	Input	×
EMIX (Jindal et al., 2020a)	×	×	-	Emb/Hidden	✓
SPEECHMIX (Jindal et al., 2020b)	×	×	-	Emb/Hidden	Speech/Audio
MIXTEXT (Chen et al., 2020c)	×	×	-	Emb/Hidden	✓
SIGNEDGRAPH (Chen et al., 2020b)	×	×	-	Input	×
DTREEMORPH (Şahin and Steedman, 2018)	×	×	dep	Input	✓
Sub^2 (Shi et al., 2021)	×	×	dep	Input	Substructural
DAGA (Ding et al., 2020)	×	×	tok	Input+Label	×
WN-HYPERS (Feng et al., 2020)	✓	×	const+KWE	Input	✓
SYNTHETIC NOISE (Feng et al., 2020)	×	×	tok	Input	✓
UEDIN-MS (DA part) (Grundkiewicz et al., 2019)	✓	×	tok	Input	✓
NONCE (Gulordava et al., 2018)	✓	×	const	Input	✓
XLDA (Singh et al., 2019)	×	✓	Depends	Input	✓
SEQMIX (Zhang et al., 2020)	×	✓	tok	Input+Label	×
SLOT-SUB-LM (Louvan and Magnini, 2020)	×	✓	tok	Input	✓
UBT & TBT (Vaibhav et al., 2019)	×	✓	Depends	Input	✓
SOFT CONTEXTUAL DA (Gao et al., 2019)	×	✓	tok	Emb/Hidden	✓
DATA DIVERSIFICATION (Nguyen et al., 2020)	×	✓	Depends	Input	✓
DIPS (Kumar et al., 2019a)	×	✓	tok	Input	✓
AUGMENTED SBERT (Thakur et al., 2021)	×	✓	-	Input+Label	Sentence Pairs

Table 1: Comparing a selection of DA methods by various aspects relating to their applicability, dependencies, and requirements. *Ext.Know*, *KWE*, *tok*, *const*, and *dep* stand for External Knowledge, keyword extraction, tokenization, constituency parsing, and dependency parsing, respectively. *Ext.Know* refers to whether the DA method requires external knowledge (e.g. WordNet) and *Pretrained* if it requires a pretrained model (e.g. BERT). *Preprocess* denotes preprocessing required, *Level* denotes the depth at which data is modified by the DA, and *Task-Agnostic* refers to whether the DA method can be applied to different tasks. See Appendix B for further explanation.

Data Augmentation for Text Generation

- Large pretrained generators like GPT-2 → Possibility to perform generation in many new domains and settings
- GPT-2 still needs to be finetuned to the specific domain!
- Without this, it can't pick up:
  - Length characteristics
  - Stylistic variables (e.g. formality, sentiment)
  - Domain-specific word choices
- Apart from specific tasks like MT, most augmentation methods in NLP have been focused on classification

GenAug: Data Augmentation for Finetuning Text Generators *(Feng et al., DeeLIO Workshop @ EMNLP '20)*[9]

- Suite of perturbation operations to generate augmented examples
- Synthetic Noise: character-level
- Synonym: word choice
- Hypernym/Hyponym: word granularity
- Semantic Text Exchange: topic-level semantics
- Motivated by intuition, greater focus on modestly meaning-altering perturbations, toggle specific aspects

Method	Text
Original Review	got sick from the food . overpriced and the only decent thing was the bread pudding . wouldn't go back even if i was paid a million dollars to do so .
Synthetic Noise (10%)	got **seick** from the **fotod** . **overhpriced** and the only decent **ting** was the bread pudding . wouldn't go back even if i was paid a million dollars to do so .
Synonym Replacement (3 keywords)	got sick from the food . overpriced and the only decent thing was the **scratch pud** . wouldn't go back even if i was paid a **one thousand thousand** dollars to do so .
Hyponym Replacement (3 keywords)	got sick from the food . overpriced and the only decent thing was the **crescent roll corn pudding** . wouldn't go back even if i was paid a million **kiribati dollar** to do so .
Hypernym Replacement (3 keywords)	got sick from the food . overpriced and the only decent thing was the **baked goods dish** . wouldn't go back even if i was paid a **large integer** dollars to do so .
Random Insertion (10%)	got sick from the food **nauseous** . overpriced and the only decent thing was the bread pudding . wouldn't go back even if i was paid a million dollars **boodle** to do so .
Semantic Text Exchange (60% MRT)	got sick from the **coffee** . overpriced and the **food was good** . wouldn't **come back** if i was **in a long hand washing machine** .

GenAug: Data Augmentation for Finetuning Text Generators

- Evaluate various qualities of the generated text: fluency, diversity, content and sentiment preservation
- Two methods: **Synthetic Noise** and **Keyword Replacement with Hypernyms** outperformed a random augmentation baseline and the no-augmentation case
- Augmentations improve quality of the generated text up to **3x** the amount of original training data

Compositionality for Data Augmentation

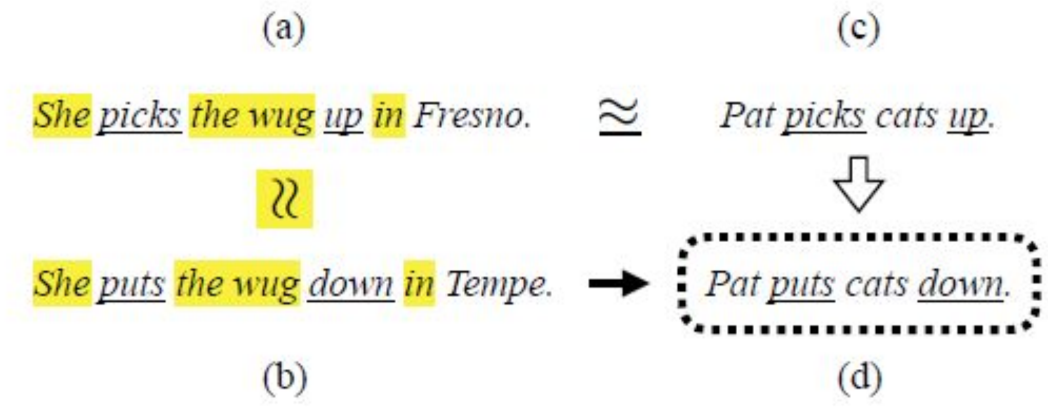


Figure 1: Visualization of the proposed approach: two discontinuous sentence fragments (a–b, underlined) which appear in similar environments (a–b, highlighted) are identified. Additional sentences in which the first fragment appears (c) are used to synthesize new examples (d) by substituting in the second fragment.

Good-Enough Compositional Data Augmentation (Jacob Andreas, ACL 2020)[10]

- Concept of compositionality of meaning
  - Wheels + seat + handle → bike
  - Subwords + morphemes → words
- Constructs synthetic examples for downstream tasks
  - E.g. semantic parsing
- Fragments of original examples are replaced with fragments from other examples in similar contexts

Challenges and Future Directions for DA

- Empirical vs. Theoretical
- Multimodal Challenges
- Span-Based Tasks
- Specialized Domains
- Low-Resource Languages
- More structural and document-level info
- Inspiration from Vision
- Self-Supervised Learning
- Offline vs. Online DA
- Lack of Unification

Empirical vs Theoretical

- Empirical novelties vs theoretical narrative
- What do we mean?
  - Typical "new DA method" paper
    - A task-specific intuition / motivation / invariance
    - Formalized as method, empirically proved better on the task/task family benchmarks
    - End of story
  - Little discussion on
    - What are the factors underlying the success of this method? [What is the space of factors to look at? Is there a common way of coming up with these factors for a set of target tasks?]
    - How does it differ from earlier DA methods on these factors of success?
    - How do the hyperparam variants / ablations of the full DA method do along these factors?

Span-Based Tasks

- Tasks where output labels correspond to multiple tokens or points in the input text, a.k.a spans. Inputs themselves can be quite complex

- No singular label at the global input level, like in generation or classification. Some examples:
  - NER - One label at each token
  - Coreference Detection
    - One label at each entity span
    - Label space = All previous entity spans
  - Event Arg Detection
    - One label at each event trigger
    - Label space = All previous spans

Span-Based Tasks

- Why are they a challenge for data augmentation?
  - Can't rely on easily devised input-level invariances !
  - Most <u>randomized</u> (*token shuffle*) and <u>paraphrasing</u> (*backtranslation*) transforms fiddle with span-level correspondences → can't use them !

Good Data Augmentation Practices

- Unified benchmark tasks, datasets, and frameworks/libraries
- Making code and augmented datasets publicly available
- Reporting variations among results (e.g. across seeds)
- More standardized evaluation procedures
- Transparent hyperparameter analysis
- Explicitly stating failure cases of proposed techniques
- Discussion of the intuition and theory behind DA techniques

Peep@Future#1 - The DataAug4NLP repo

- We maintain a live git repo: https://github.com/styfeng/DataAug4NLP

README.md

Data Augmentation Techniques for NLP

If you'd like to add your paper, do not email us. Instead, read the protocol for adding a new entry and send a pull request.

We group the papers by text classification, translation, summarization, question-answering, sequence tagging, parsing, grammatical-error-correction, generation, dialogue, multimodal, mitigating bias, mitigating class imbalance, adversarial examples, compositionality, and automated augmentation.

This repository is based on our paper, "A survey of data augmentation approaches in NLP (Findings of ACL '21)". You can cite it as follows:

```
@article{feng2021survey,
  title={A Survey of Data Augmentation Approaches for NLP},
  author={Feng, Steven Y and Gangal, Varun and Wei, Jason and Chandar, Sarath and Vosoughi, Soroush a
  journal={Findings of ACL},
  year={2021}
}
```

Authors: Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, Eduard Hovy

Sequence Tagging

Paper	Datasets
Data Augmentation via Dependency Tree Morphing for Low-Resource Languages (EMNLP '18) code	universal dependencies project
DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks (EMNLP '20) code	CoNLL2002/2003
An Analysis of Simple Data Augmentation for Named Entity Recognition (COLING '20)	MaSciP, i2b2- 2010
SeqMix: Augmenting Active Sequence Labeling via Sequence Mixup (EMNLP '20) code	CoNLL-03, ACE05, Webpage

Parsing

Paper	Datasets
Data Recombination for Neural Semantic Parsing (ACL '16) code	GeoQuery, ATIS, Overnight
A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages (EMNLP '19)	Universal Dependencies treebanks version 2.2
Named Entity Recognition for Social Media Texts with Semantic Augmentation (EMNLP '20)code	WNUT16, WNUT17, Weibo
Good-Enough Compositional Data Augmentation (ACL '20) code	SCAN
GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing (ICLR '21)	SPIDER, WIKISQL, WIKITABLEQUESTIONS

- New methods can request inclusion via a PR in specified form
- We also update our arXiv in tandem with the live repo

Peep@Future#2 - NL-Augmenter 🦎→🐍

- Unified benchmark tasks, datasets, and frameworks/libraries
- Making code and augmented datasets publicly available
- ~~Reporting variations among results (e.g. across seeds)~~
- More standardized evaluation procedures
- ~~Transparent hyperparameter analysis~~
- ~~Explicitly stating failure cases of proposed techniques~~
- ~~Discussion of the intuition and theory behind DA techniques~~

🦎→🐍 & the *Transformations* concept

- What's NL-Augmenter? → Participative repo to help NL community define, code, curate large suite of *Transformations*
- What's a *Transformation*? Converts a valid task example → New, distinct [valid] task example → specific to a task (family)
- "Task example": tuple of input sentence, label and whichever other task-specific input + output components get transformed

🦎→🐍 & the *Transformations* concept

- Transformation generalizes the notion of *paraphrase* to be:
  - *Task-specific* in its notion of *invariance*
  - Consider *multiple input components* rather than just *single sentence* → *single sentence* functions
- New transformation → New DA strategy for corresponding task
- Why make process participative?
  - Wisdom [and scale] of the crowds →
    Ensures diverse group of functions, task coverage

🦎→🐍 & *Transformations* - Example

- Task: Sentiment analysis with input sentence x and binary labels y.
- Let 0 = negative sentiment, 1= positive sentiment
- *Add-A-**Not** transformation* for sentiment analysis : x, y → **Not**(x), 1-y
- What's **Not**(x)?
  - Introduces a "not" after the be auxiliary.
  - **Not**(This zombie flick was worth the ticket) → This zombie flick was not worth the ticket
  - **Not** negates meaning of x → not a valid paraphrase!
- However, *Add-A-**Not*** : x, y → **Not**(x), 1-y constitutes a valid *transformation* for sentiment analysis.

Additional Purposes for NL-Augmenter

- NL-Augmenter also helps address additional issues:
  - LR language phenomena and domains not receiving **attention**! E.g. Rare language phenomena, endangered languages, underrepresented groups
  - Can help perform **robustness testing** of models. Specific transformations can help gauge + repair specific capabilities.

We invite you to contribute transformations to 🦎→🐍

- **All submitters of accepted implementations will be included as co-authors on a paper announcing this framework.**
    - Fork the repository @ https://github.com/GEM-benchmark/NL-Augmenter
    - Add your creative transformation
    - Create a Pull request!

⚠️ **Last Date: August 31, 2021**

- **Most Creative Implementations** 🏆 🏆 🏆
    - After all pull-requests have been merged, 3 of the most creative implementations would be selected and featured on the README page and on the NL-Augmenter webpage.

~ 43 implementations (merged) + ~ 80 (under review)

- back_translation
- butter_fingers_perturbation
- change_char_case
- change_date_format
- change_person_named_entities
- change_two_way_ne
- close_homophones_swap
- contraction_expansions
- discourse_marker_substitution
- english_inflectional_variation
- formality_change
- gender_culture_diverse_name
- gender_culture_diverse_name_two_way
- gender_swap
- geonames_transformation
- leet_letters
- lexical_counterfactual_generator

- longer_names_ner
- mixed_language_perturbation
- multilingual_lexicon_perturbation
- negate_strengthen
- p1_noun_transformation
- punctuation
- quora_trained_t5_for_qa
- random_deletion
- random_upper_transformation
- redundant_context_for_qa
- replace_numerical_values
- sentence_reordering
- suspecting_paraphraser
- synonym_substitution
- word_noise

Outputs of Some of the Transformations (Randomly chosen)!

🦎→🐍 **CloseHomophonesSwap**: Kaizhao Liang, (University of Illinois at Urbana-Champaign)

Original: Andrew finally returned **the** French book to Chris that I **bought** last week

Transformed: Andrew finally returned **thee** French book **too** Chris that I **Bot** Lass week

🦎→🐍 **ChangeDateFormat:** Nivranshu Pasricha, (National University of Ireland Galway)

Original: Roger Federer (born 8 August 1981) is a Swiss professional tennis player.

Transformed: Roger Federer (born **8/08/81**) is a Swiss professional tennis player.

🦎→🐍 **Discourse Marker Substitution:** Damien Sileo (KU Leuven)

Original: and platinum behaved independently, first falling and then **later** rising.

Transformed: and platinum behaved independently, first falling and then **subsequently** rising.”

Outputs of Some of the Transformations (Randomly chosen)!

🦎→🐍 **StyleTransfer**: Rishabh Gupta, IITD

Formal2Casual: Original: “This car looks fascinating” → Paraphrase: “This car looks cool!”

Casual2Formal: Original: “who gives a crap?” → Paraphrase: “Who cares about that?”

🦎→🐍 **Increasing the cultural diversity of names**: Xudong Shen, NUS

This transformation changes a name with another, considering gender and cultural diversity. Example: Rachel --> Salome, Phoebe --> Rihab, Joey --> Clarinda, Chandler --> Deon, Monica --> Lamya

🦎→🐍 **DecontextualizedSentenceReordering**: Zijian Wang, Stanford University

Original: John is a great person. He resides in Australia. Peter is also a great person. He resides in India.

Paraphrase: Peter is also a great person. John resides in Australia. Peter resides in India. John is a great person.

Outputs of Some of the Transformations (Randomly chosen)!

🦎→🐍 **Adding Noun Definitions**: Pawan Kumar Rajpoot, Rajpoot

Original: Barack Obama gave a book to me

Paraphrase: Barack obama (44th president of the united states) gave a book (a medium of writing) to me.

🦎→🐍 **GeoNames Transformation**: Vasile Pais, Romanian Academy

Original: Egypt has many pyramids.

Paraphrase: Egypt, a country in Africa, has many pyramids → Egypt, whose capital city is Cairo, has many pyramids!

…. And many more

So, please visit

github.com/GEM-benchmark/NL-Augmenter/tree/main/transformations

to take a look at all the other transformations (& filters)!

⚠️ **Last Date: August 31, 2021**

Organizers & Reviewers

- Kaustubh Dhole (Amelia R&D)
- Sebastian Gehrmann (Google Research)
- Jascha Sohl-Dickstein (Google Brain)
- Varun Gangal (LTI, Carnegie Mellon University)
- Tongshuang Wu (University of Washington)
- Simon Mille (Universitat Pompeu Fabra)
- Zhenhao Li (Imperial College, London)
- Aadesh Gupta (Amelia R&D)
- Samson Tan (NUS & Salesforce Research)
- Saad Mahmood (Trivago R&D)
- Ashish Shrivastava (Amelia R&D)
- Ondrej Dusek (Charles University)
- Abinaya Mahendran (Mphasis Technology)
- Jinho D. Choi (Emory University)
- Steven Y. Feng (LTI, Carnegie Mellon University)

For any questions or to use NL-Augmenter in your projects or to team up with, email us at nl-augmenter@googlegroups.com

Please also read: *Automatic Construction of Evaluation Suites for Natural Language Generation Datasets,* Simon Mille, Kaustubh Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, Sebastian Gehrmann, NeurIPS 2021

CtrlGen Workshop at NeurIPS 2021 (Dec. 13)

Controllable Generative Modeling in Language and Vision

Website: https://ctrlgenworkshop.github.io/   Contact: ctrlgenworkshop@gmail.com

- Aims to explore disentanglement, controllability, and manipulation for the generative vision and language modalities.
- We feature an exciting lineup of speakers, a live QA and panel session, interactive activities, and networking opportunities.

CtrlGen Workshop at NeurIPS 2021 (Dec. 13)

Controllable Generative Modeling in Language and Vision

Website: https://ctrlgenworkshop.github.io/ Contact: ctrlgenworkshop@gmail.com

Invited Speakers and Panelists

Yejin Choi
University of Washington

Jason Weston
Facebook AI

He He
New York University

Alex Tamkin
Stanford University

Yulia Tsvetkov
Carnegie Mellon University

Irina Higgins
DeepMind

Or Patashnik
Tel-Aviv University

Additional Panelists

Sebastian Gehrmann
Google AI

Angela Fan
LORIA and Facebook AI

CtrlGen Workshop at NeurIPS 2021 (Dec. 13)

Controllable Generative Modeling in Language and Vision

Website: https://ctrlgenworkshop.github.io/ Contact: ctrlgenworkshop@gmail.com

Important Dates

- Paper Submission Deadline: **September 27, 2021**
- Paper Acceptance Notification: October 22, 2021
- Paper Camera-Ready Deadline: November 1, 2021
- Demo Submission Deadline: **October 29, 2021**
- Demo Acceptance Notification: November 19, 2021
- Workshop Date: **December 13, 2021**

CtrlGen Workshop at NeurIPS 2021 (Dec. 13)

Controllable Generative Modeling in Language and Vision

Call for Papers: https://ctrlgenworkshop.github.io/CFP.html

Paper submission deadline: **September 27, 2021**. Topics of interest:

Methodology and Algorithms:

- New methods and algorithms for controllability.
- Improvements of language and vision model architectures for controllability.
- Novel loss functions, decoding methods, and prompt design methods for controllability.

Applications and Ethics:

- Applications of controllability including creative AI, machine co-creativity, entertainment, data augmentation (for text and vision), ethics (e.g. bias and toxicity reduction), enhanced training for self-driving vehicles, and improving conversational agents.
- Ethical issues and challenges related to controllable generation including the risks and dangers of deepfake and fake news.

CtrlGen Workshop at NeurIPS 2021 (Dec. 13)

Controllable Generative Modeling in Language and Vision

Call for Papers: https://ctrlgenworkshop.github.io/CFP.html

Submission deadline: **September 27, 2021**.

Tasks:

- Semantic text exchange
- Syntactically-controlled paraphrase generation
- Persona-based text generation
- Style-sensitive generation or style transfer (for text and vision)
- Image synthesis and scene representation in both 2D and 3D
- Cross-modal tasks such as controllable image or video captioning and generation from text

Evaluation and Benchmarks (standard and unified metrics and benchmark tasks)

Cross-Domain and Other Areas (interpretability, disentanglement, robustness, representation learning)

Position and Survey Papers (problems and lacunae in current controllability formulations, neglected areas in controllability, and the unclear and non-standardized definition of controllability)

CtrlGen Workshop at NeurIPS 2021 (Dec. 13)

Controllable Generative Modeling in Language and Vision

Call for Demonstrations: https://ctrlgenworkshop.github.io/demos.html

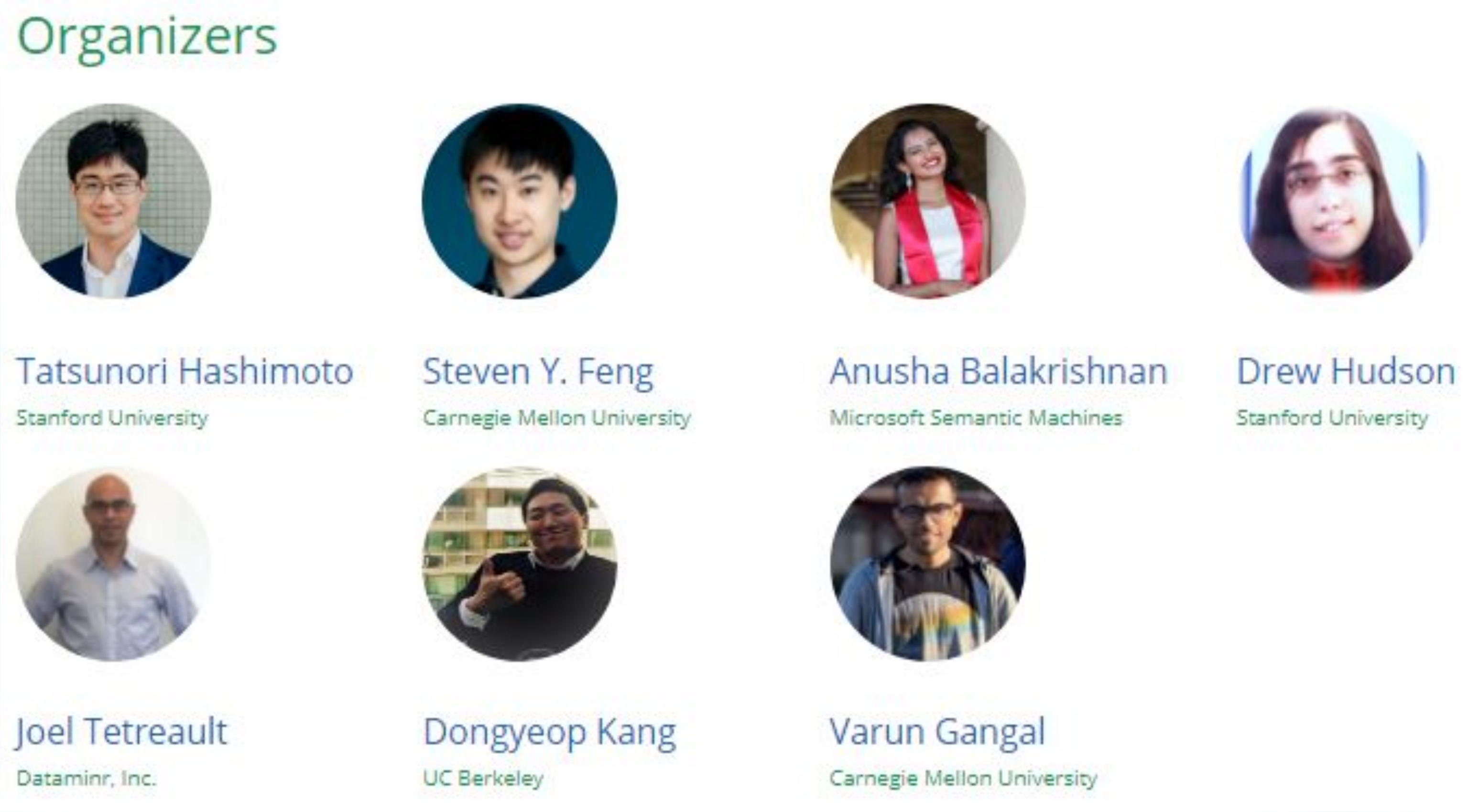
Submission deadline: **October 29, 2021**. Demos of all forms: research-related, demos of products, interesting and creative projects, etc. Creative, well-presented, attention-grabbing. Examples:

- Creative AI such as controllable poetry, music, image, and video generation models.
- Style transfer for both text and vision.
- Interactive chatbots and assistants that involve controllability.
- Controllable language generation systems, e.g. using GPT-2 or GPT-3.
- Controllable multimodal systems such as image and video captioning or generation from text.
- Controllable image and video/graphics enhancement systems.
- Systems for controlling scenes/environments and applications for self-driving vehicles.
- Controllability in the form of deepfake and fake news, specifically methods to combat them.
- And much, much more…

CtrlGen Workshop at NeurIPS 2021 (Dec. 13)

Controllable Generative Modeling in Language and Vision

Website: https://ctrlgenworkshop.github.io/ Contact: ctrlgenworkshop@gmail.com



Podcast - Steven Feng & Eduard Hovy

- **Steven Feng**, **Eduard Hovy**, and **Ben Lorica** discuss data augmentation for NLP (inspired by this survey paper) and general trends and challenges in NLP and machine learning research in a more Joe-Rogan-esque session.
- Video version: https://www.youtube.com/watch?v=qmqyT_97Poc&ab_channel=GradientFlow
- Audio and notes: https://thedataexchange.media/data-augmentation-in-natural-language-processing/

Thanks for Listening!

https://aclanthology.org/2021.findings-acl.84/

https://github.com/styfeng/DataAug4NLP

- **Steven Y. Feng**: syfeng@cs.cmu.edu
  Website: https://styfeng.github.io/
  Twitter: @stevenyfeng

- **Varun Gangal**: vgangal@cs.cmu.edu
  Website: https://vgtomahawk.github.io/
  Twitter: @VarunGangal

- Jason Wei: jasonwei@google.com
  Website: https://jasonwei20.github.io/
  Twitter: @_jasonwei

- Sarath Chandar: sarath.chandar@mila.quebec
  Website: http://sarathchandar.in/
  Twitter: @apsarathchandar

- Soroush Vosoughi: soroush@dartmouth.edu
  Website: https://www.cs.dartmouth.edu/~soroush/
  Twitter: @CrashTheMod3

- Teruko Mitamura: teruko@cs.cmu.edu

- Eduard Hovy: hovy@cs.cmu.edu
  Website: https://www.cs.cmu.edu/~hovy/

References

1. Wei and Zou, EMNLP 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. https://www.aclweb.org/anthology/D19-1670/
2. Xie et al., NeurIPS 2020. Unsupervised Data Augmentation for Consistency Training. https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html
3. Sahin and Steedman, EMNLP 2018. Data Augmentation via Dependency Tree Morphing for Low-Resource Languages. https://www.aclweb.org/anthology/D18-1545/
4. Zhang et al., ICLR 2018. mixup: Beyond Empirical Risk Minimization. https://arxiv.org/abs/1710.09412
5. Sennrich et al., ACL 2016. Improving Neural Machine Translation Models with Monolingual Data.
6. Kobayashi, NAACL 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. https://www.aclweb.org/anthology/N18-2072/
7. Feng et al., EMNLP 2019. Keep Calm and Switch On! Preserving Sentiment and Fluency in Semantic Text Exchange. https://www.aclweb.org/anthology/D19-1272/
8. Qin et al., IJCAI 2020. CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP. https://www.ijcai.org/proceedings/2020/0533.pdf
9. Feng et al., DeeLIO WS @ EMNLP 2020. GenAug: Data Augmentation for Finetuning Text Generators. https://aclanthology.org/2020.deelio-1.4/
10. Andreas, EMNLP 2019. Good-Enough Compositional Data Augmentation. https://aclanthology.org/2020.acl-main.676/