Motivations

- Extend prototypical networks
- Incorporate unlabeled data

Motivations

- Extend prototypical networks
- Incorporate unlabeled data
- Use data augmentation for more robustness

Motivations

- Extend prototypical networks
- Incorporate unlabeled data
- Use data augmentation for more robustness
- Data-augmented texts should be as diverse as possible

Data Augmentation by Paraphrase Generation

Back-translation [7] is a poor data augmentation strategy

Data Augmentation by Paraphrase Generation

Back-translation [7] is a poor data augmentation strategy

→ We use BART [13] for Natural Language Generation

Data Augmentation by Paraphrase Generation

Back-translation [7] is a poor data augmentation strategy

→ We use BART [13] for Natural Language Generation

We train the pre-trained BART model on a paraphrasing task, using the following datasets:

- Quora [21]
- Google's PAWS [31]
- MSR Paraphrase corpus [32]

Constrained generation for more diversity

Diverse Beam Search (**DBS** [27]), an extension of the regular beam search decoding strategy, generates diverse sentences

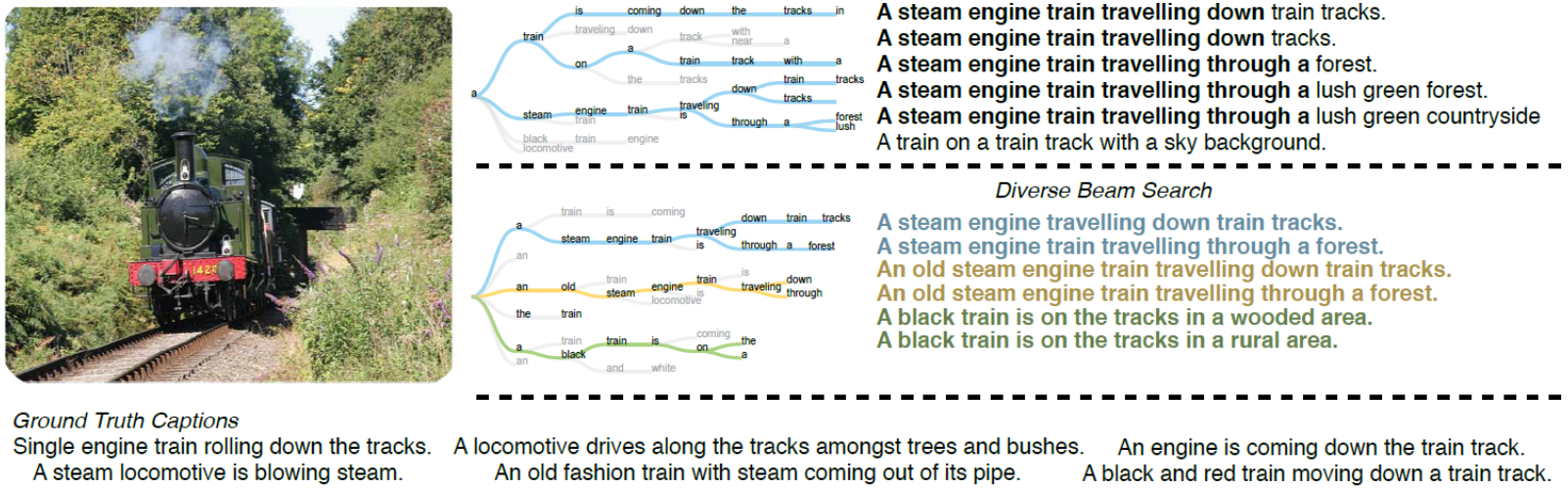

Figure 1: Comparing image captioning outputs decoded by BS (top) and our method, Diverse Beam Search (middle) – we notice that BS captions are near-duplicates with similar shared paths in the search tree and minor variations in the end. In contrast, DBS captions are significantly diverse and similar to the variability in human-generated ground truth captions (bottom).

Constrained generation for more diversity

We further constraint the paraphrase generation using

- **uni-gram masking**: some tokens (according to a probability p_{mask}) of the input sentence cannot appear in the generated paraphrase
- **bi-gram masking**: all bi-grams in the input sentence cannot appear in the generated paraphrase

Paraphrase examples

orig: How long will my transfer be pending for? *back*: How long will my transfer be on hold? *dbs*: How long will my transfer be pending? I am in first year. *dbs_unigram*: When are all transfers coming up and how many days are they expected? *dbs_bigram*: If I have a transfer for a while, how long should I wait for it?
orig: I am not sure where my phone is. *back*: I don't know where my phone is. *dbs*: I am not really sure where my phone is located *dbs_unigram*: How can I find the location of any Android mobile *dbs_bigram*: I don't know where is my cell phone
orig: can you play m3 file *back*: can you read m3 file *dbs*: M3 files: can I play the entire M3 file? *dbs_unigram*: Is there any way to play 3M files on Earth without downloading it *dbs_bigram*: Is there any way to play M3 files on Windows?

Table 5: Examples of sentences paraphrased using back translation, vanilla diverse beam search, DBS with unigram masking, and DBS with bigram masking

Unsupervised Loss

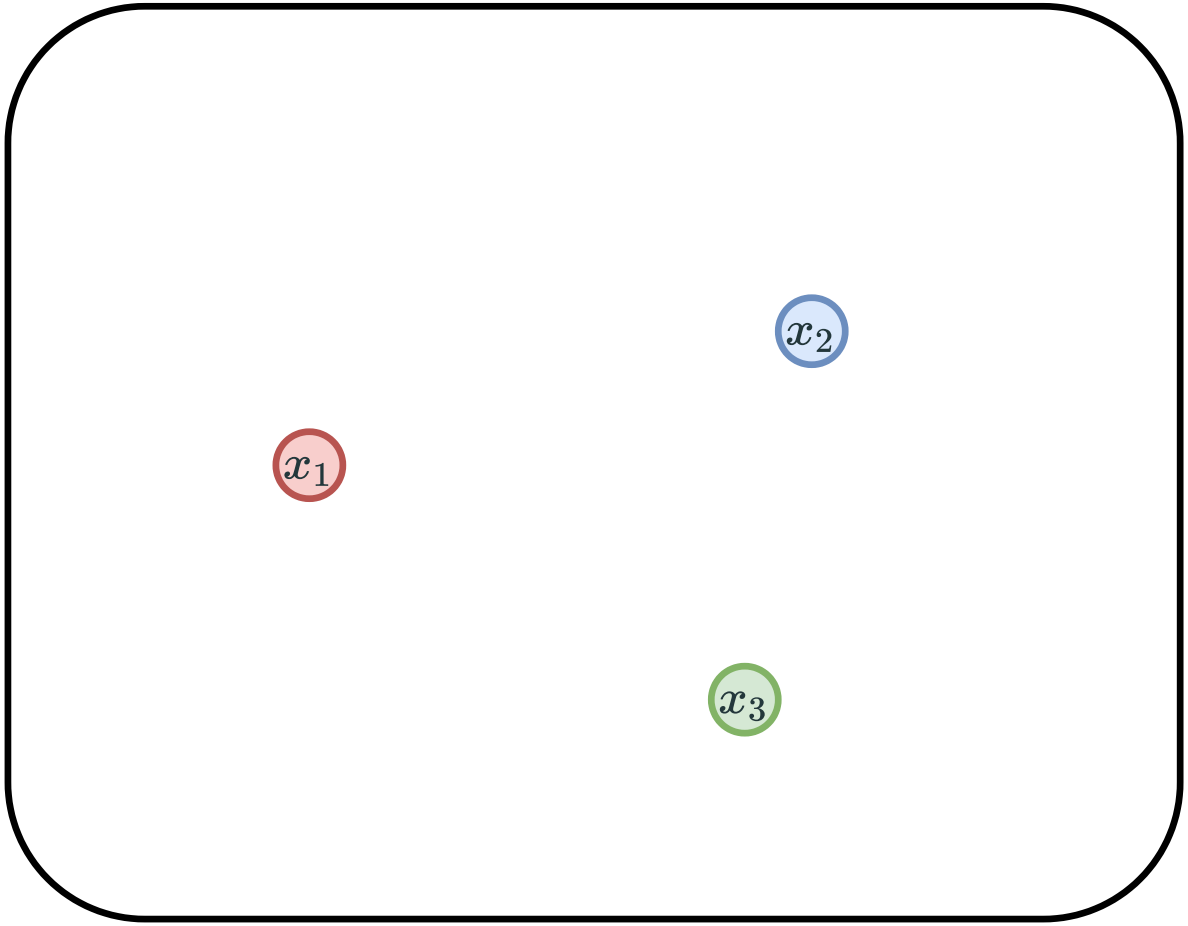

Figure 12: Step 0: Pick unlabeled data

Unsupervised Loss

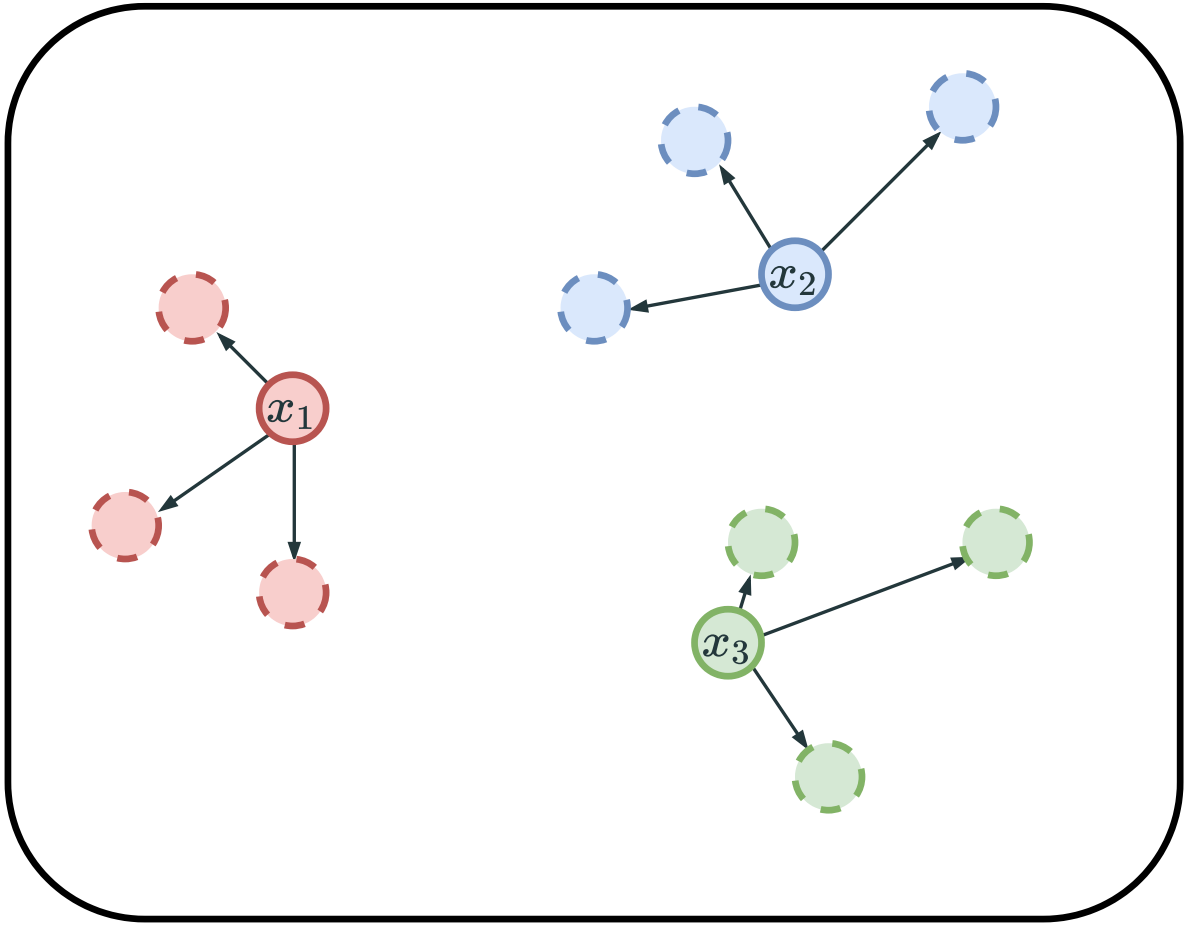

Figure 13: Step 1: Generate paraphrases

Unsupervised Loss

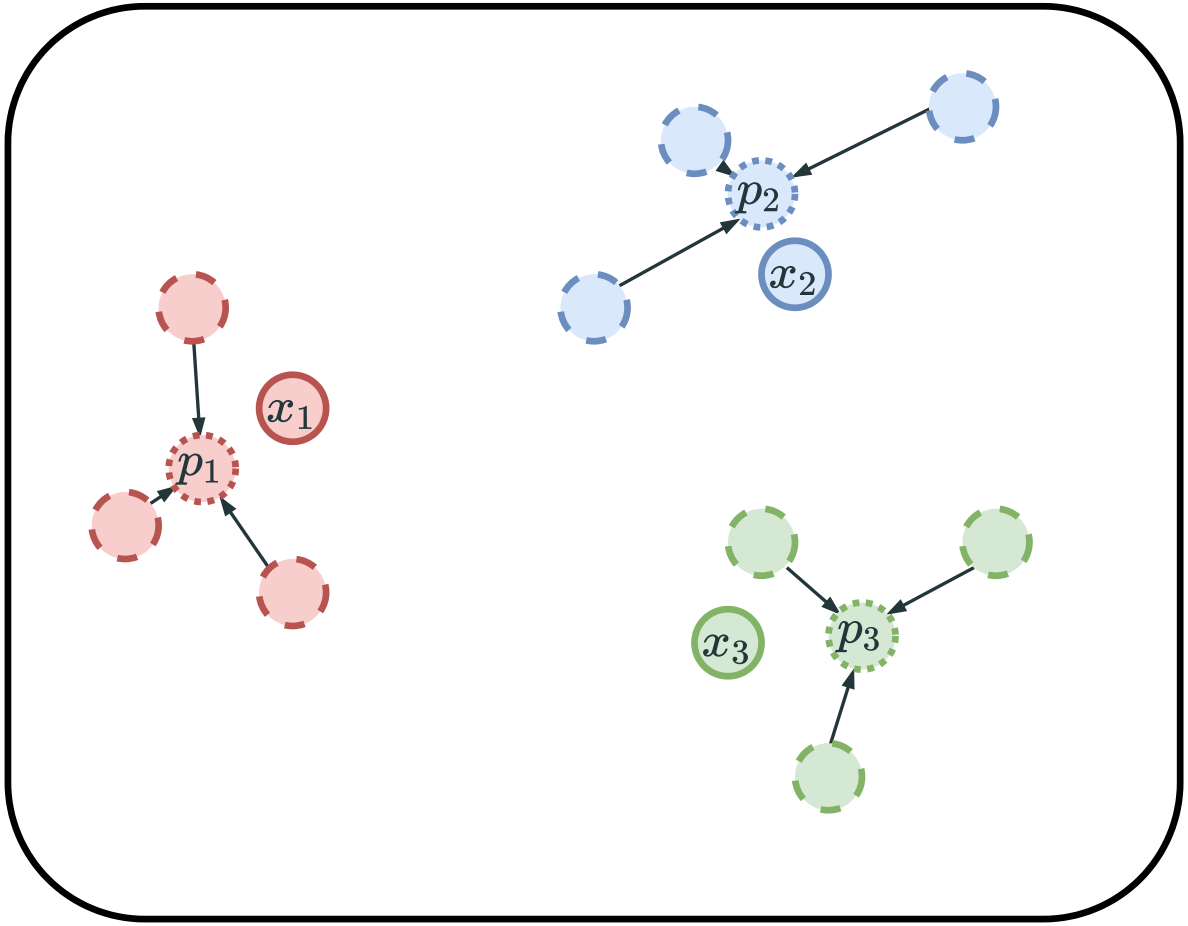

Figure 14: Step 2: Compute the prototypes, as average of paraphrases

Unsupervised Loss

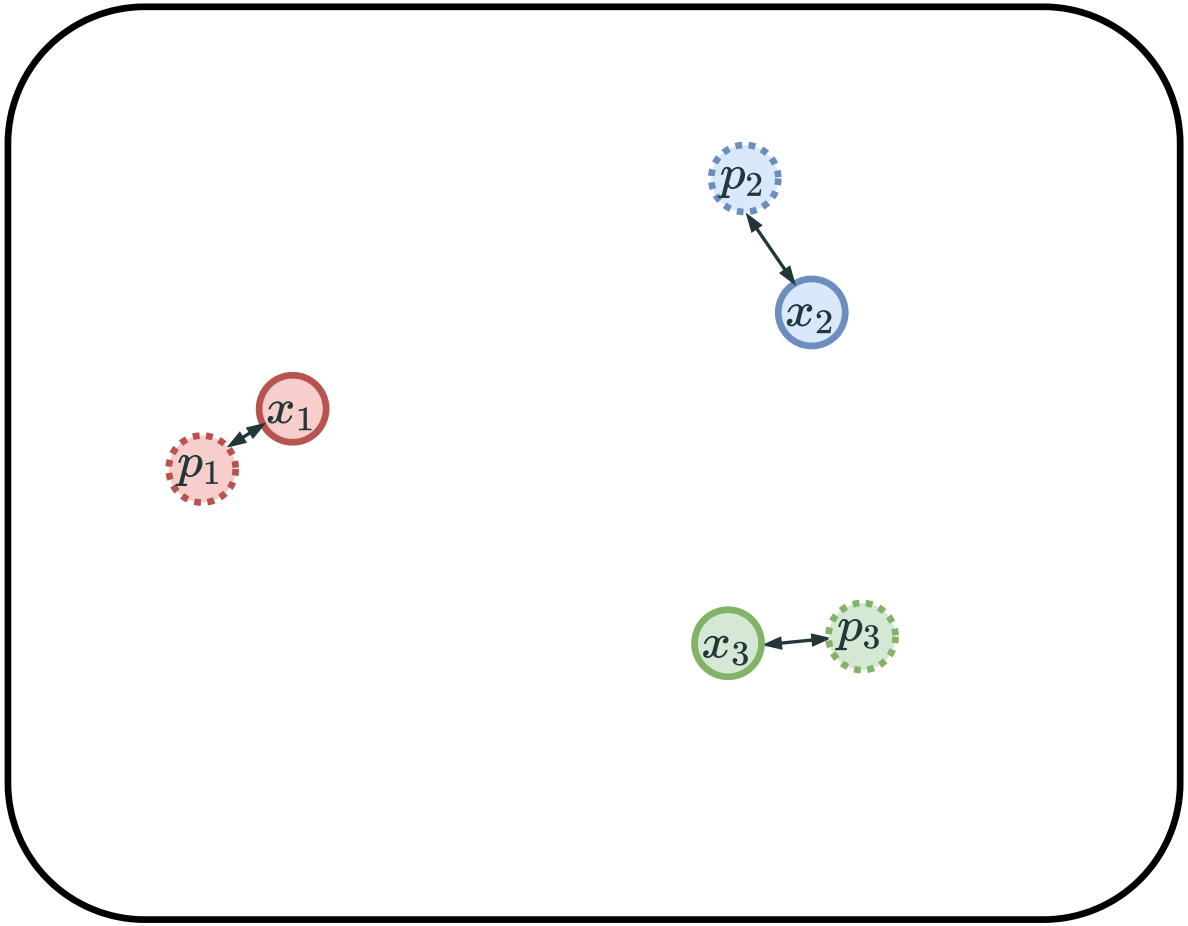

Figure 15: Step 3: Compare unlabeled data to prototypes

ProtAugment's framework

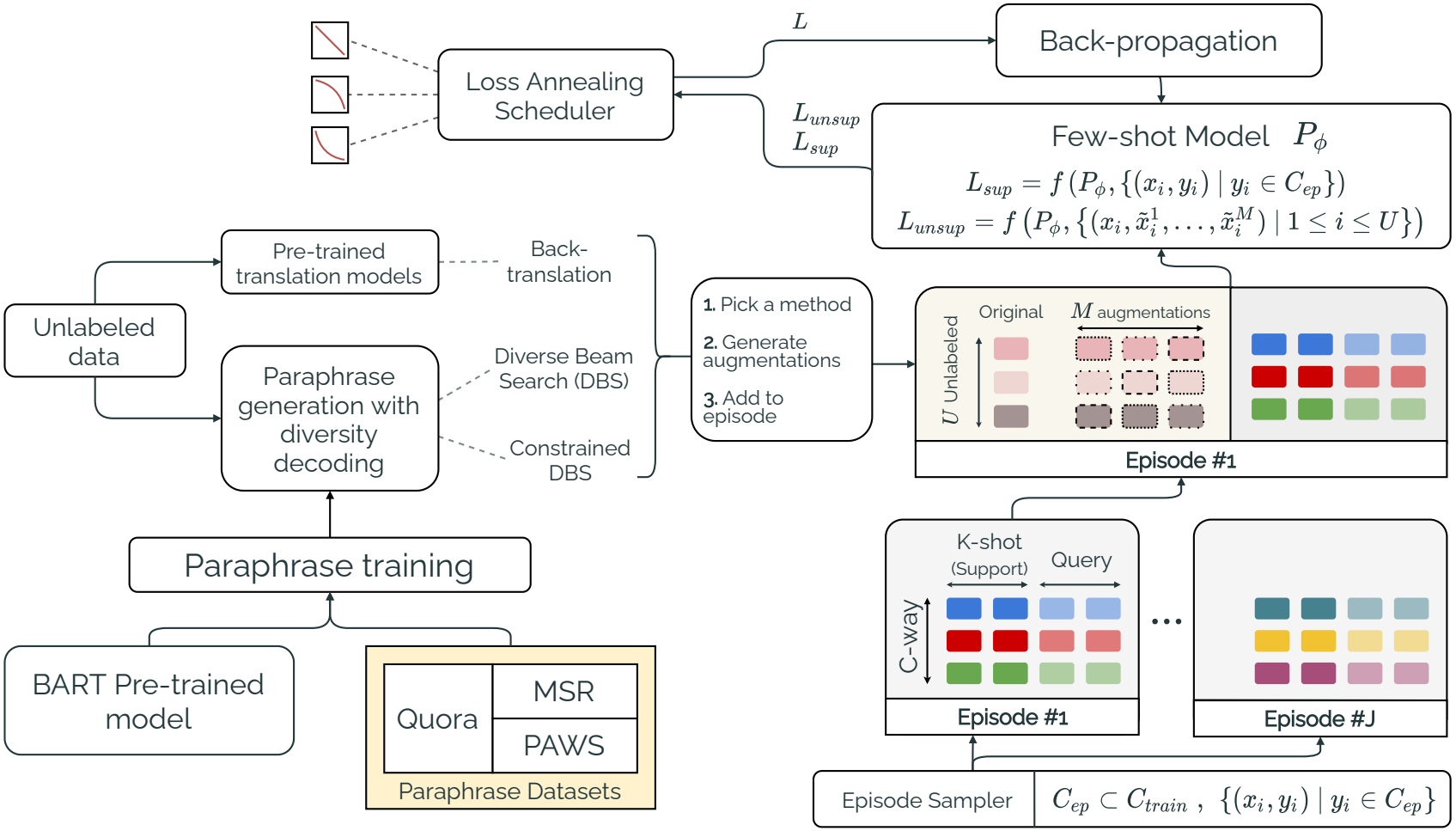

Figure 16: The ProtAugment framework

Results

Data Profile	Method	Datasets								Accuracy stats	
		Banking		HWU		Liu		Clinic		($AVG \pm STD$)	
		$K=1$	$K=5$	$K=1$	$K=5$	$K=1$	$K=5$	$K=1$	$K=5$	$K=1$	$K=5$
low profile	Prototypical Network	82.20	91.57	74.37	86.48	80.06	89.62	94.29	98.10	82.73 ± 2.32	91.44 ± 1.92
	ours w/ BT	83.83	92.16	78.70	89.36	80.84	90.87	94.06	97.62	84.36 ± 1.15	92.50 ± 0.94
	ours w/ DBS	83.10	92.56	80.06	90.21	82.31	91.64	93.70	97.83	84.80 ± 1.26	93.06 ± 0.99
	ours w/ DBS+bigram	86.04	93.55	82.09	**91.57**	83.60	92.71	95.11	98.23	86.71 ± 1.14	94.01 ± 1.05
	ours w/ DBS+unigram	**87.23**	**94.29**	**83.70**	91.29	**85.16**	**93.00**	**95.92**	**98.56**	$\mathbf{88.00 \pm 1.22}$	$\mathbf{94.29 \pm 0.76}$
full profile	Prototypical Network	86.28	93.94	77.09	89.02	82.76	91.37	96.05	98.61	85.55 ± 2.20	93.24 ± 1.22
	ours w/ BT	87.46	94.47	81.31	91.44	84.14	92.67	95.19	98.36	87.02 ± 1.36	94.23 ± 0.82
	ours w/ DBS	86.94	94.50	82.35	91.68	84.42	92.62	94.85	98.41	87.14 ± 1.36	94.30 ± 0.60
	ours w/ DBS+bigram	88.14	94.70	84.05	92.14	85.29	93.23	95.77	98.50	88.31 ± 1.43	94.64 ± 0.59
	ours w/ DBS+unigram	**89.56**	**94.71**	**84.34**	**92.55**	**86.11**	**93.70**	**96.49**	**98.74**	$\mathbf{89.13 \pm 1.13}$	$\mathbf{94.92 \pm 0.57}$

Table 6: 5-way 1-shots and 5-way 5-shots accuracy on the test sets for each dataset. For each data profile, we highlight the best method in **bold**. We underline the methods on the **low** profile which perform better than the Prototypical Networks on the **full** profile.

Masking: Less or More?

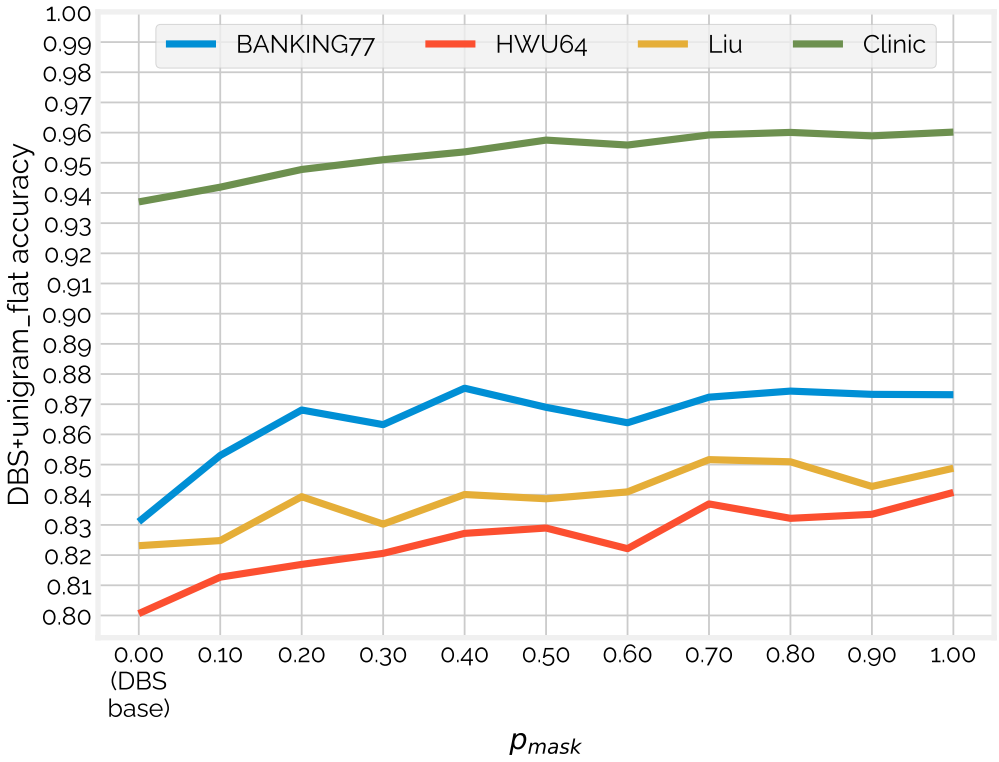

Figure 17: Performance when changing the p_{mask} value

- In terms of masking, **more** is better than **less**
- The optimal value for p_{mask} is around 0.7

Key Findings

- ProtAugment vastly out-performs prototypical networks

Key Findings

- ProtAugment vastly out-performs prototypical networks
- Using constrained paraphrase generation is way better than using back-translated sentences

Key Findings

- ProtAugment vastly out-performs prototypical networks
- Using constrained paraphrase generation is way better than using back-translated sentences
- ProtAugment also yields large increments even using the **full** data profile

Key Findings

- PROTAUGMENT vastly out-performs prototypical networks
- Using constrained paraphrase generation is way better than using back-translated sentences
- PROTAUGMENT also yields large increments even using the **full** data profile

This word was (very) recently accepted as long paper at ACL2021: **PROTAUGMENT: Intent Detection Meta-Learning through Unsupervised Diverse Paraphrasing** [5]

Future Work

Future work

Test this unsupervised loss on other meta-learning models

Extend paraphrase generation in the multilingual case

Feel free to ask questions!

contact: thomas.dopierre@protonmail.com

code: https://github.com/tdopierre

References

[1] W. Chen, Y. Liu, Z. Kira, Y. Wang, and Jia-Bin Huang. A closer look at few-shot classification. *ArXiv*, abs/1904.04232, 2019.

[2] Alice Coucke, Alaa Saade, Adrien Ball, and Bluche et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 2018.

[3] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

References ii

[4] Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. A neural few-shot text classification reality check. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 935–943, Online, April 2021. Association for Computational Linguistics.

[5] Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. Protaugment: Intent detection meta-learning through unsupervised diverse paraphrasing. ACL, 2021.

[6] Thomas Dopierre, Christophe Gravier, Julien Subercaze, and Wilfried Logerais. Few-shot pseudo-labeling for intent detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4993–5003, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

References iii

[7] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[8] Ruiying Geng, Binhua Li, Yongbin Li, Xiao-Dan Zhu, Ping Jian, and Jian Sun. Induction networks for few-shot text classification. In *EMNLP/IJCNLP*, 2019.

[9] J. Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *ACL*, 2018.

[10] Armand Joulin, E. Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *EACL*, 2017.

[11] Yoon Kim. Convolutional neural networks for sentence classification. In *EMNLP*, 2014.

References iv

[12] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*, 2019.

[13] M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, abs/1910.13461, 2020.

[14] Gang Liu and Jiabao Guo. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338, 2019.

References v

[15] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. *arXiv preprint arXiv:1903.05566*, 2019.

[16] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and P. Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

[17] Jeffrey Pennington, R. Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL-HLT*, 2018.

[19] David Pinto and Paolo Rosso. On the relative hardness of clustering corpora. In *International Conference on Text, Speech and Dialogue*, pages 155–161. Springer, 2007.

References vi

[20] Victor Garcia Satorras and Joan Bruna. Few-shot learning with graph neural networks. *ArXiv*, abs/1711.04043, 2018.

[21] Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. Natural language understanding with the quora question pairs dataset, 2019.

[22] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[23] C. Sun, Xipeng Qiu, Yige Xu, and X. Huang. How to fine-tune bert for text classification? *ArXiv*, abs/1905.05583, 2019.

[24] Flood Sung, Yongxin Yang, L. Zhang, Tao Xiang, P. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

References vii

[25] Julien Tissier, Christophe Gravier, and Amaury Habrard. Dict2vec : Learning word embeddings using lexical dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[26] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[27] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of*

References viii

Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 7371–7379. AAAI Press, 2018.

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

[29] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, S. Chang, Saloni Potdar, Yu Cheng, G. Tesauro, Haoyu Wang, and Bowen Zhou. Diverse few-shot text classification with multiple metrics. *ArXiv*, abs/1805.07513, 2018.

[30] S. Yu, Danlei Liu, Wenfeng Zhu, Y. Zhang, and Sheng mei Zhao. Attention-based lstm, gru and cnn for short text classification. *J. Intell. Fuzzy Syst.*, 39:333–340, 2020.

References ix

[31] Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[32] Shiqi Zhao and Haifeng Wang. Paraphrases and applications. In *Coling 2010: Paraphrases and Applications–Tutorial notes*, pages 1–87, Beijing, China, August 2010. Coling 2010 Organizing Committee.