The ability to understand natural language is a foundation of artificial intelligence (AI). My long-term goal is to build **machine learning** models to advance artificial agents' ability to not only **perceive human language**, but also **acquire knowledge, reason, and communicate through it**. While the advent of large-scale deep learning models offers a seemingly clear path forward, challenges come along. The increasing computational overhead heightens the barriers to entry to cutting-edge natural language processing (NLP) research; the sophisticated deep learning models often generalize poorly to real-world settings, and it has become increasingly difficult to explain their decisions. My research aims to design **efficient, robust, generalizable, and interpretable representation learners for NLP**. During my Ph.D. career, I have taken several steps towards my long-term goal:

- **Efficient methods for NLP (§1):** I have developed and empirically validated efficient algorithms and learning paradigms for state-of-the-art NLP models [1, 2, 3, 4, 5].
- **Learning with structural inductive biases (§2):** my research has proposed a joint learning framework for meaning representations, and devised an algorithm to train NLP pipelines end-to-end. This linguistic knowledge in turn helps learn representations that are robust, generalizable, and sample-efficient [6, 7, 8, 9, 10, 11, 12].
- **Formal analysis of deep learning (§3):** I have formally characterized the connections between several modern neural architectures and weighted finite-state automata. The insights inspire accurate, efficient, and interpretable neural models [13, 14].

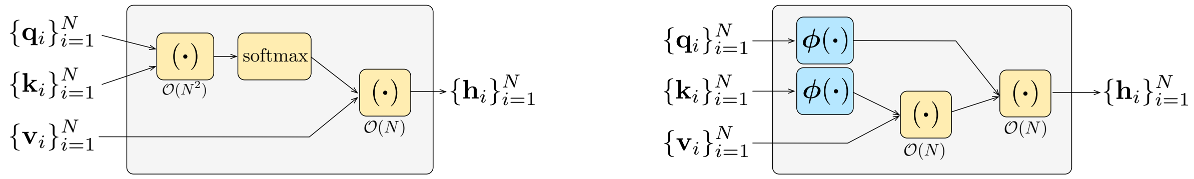
1 Efficient Methods for NLP

Artificial intelligence applications have seen unprecedented progress, which can be primarily attributed to the development of computationally-intensive deep learning models. However, their growing computational requirements have heightened the barriers to entry to state-of-the-art research and negatively impacted the environment. For example, searching for optimal architectures for certain NLP tasks [15] is estimated to cost more than 1 million USD, hardly affordable to less-resourced research groups. It emits a comparable amount of carbon dioxide as an average car does in 5 years [16]. My research aims to **improve the efficiency of state-of-the-art NLP models**, thereby **promoting the accessibility of cutting-edge NLP research, especially for less-funded institutions, and mitigate its environmental concerns.**

Attention with linear complexity. As one step towards this goal, I develop algorithms to improve the efficiency of the transformer architecture [17]. It is a crucial ingredient of recent advance in NLP and the backbone of many foundation models [18] such as BERT [19] and GPT-3 [20]. Transformers use the softmax attention to contextualize the input. Attention comes with a quadratic complexity in the input length and accounts for most of the overhead, especially for long sequences. Therefore, devising accurate and more efficient alternative contextualizations is the key to improving transformers' efficiency, which has become a focal point of the community [21]. I explore two directions to achieve this.

We derive a holistic view of several recent efficient transformer models [2]. Although seemingly disparate, they can be subsumed into one unified abstraction, "compressing" the context with various strategies, primarily using hand-crafted heuristics. This perspective not only connects several efficient attention variants that would otherwise appear apart but also gives fresh insights into established approaches, broadening their impact. Besides, it inspires a new algorithm that learns to compress the context in a context-dependent manner. On a variety of real-world NLP tasks, our model outperforms other efficient transformers.

Drawing inspiration from the classical kernel methods, I devise linear-complexity alternatives to softmax attention [1, 4]. Canonical attention calculates pairwise similarities between input tokens using a softmax normalization. Rfa [1] devises a linear-complexity approximation to it using random feature techniques and a kernel trick (Figure 1). Rfa significantly


$\{\mathbf{q}_i\}_{i=1}^N$
$\{\mathbf{k}_i\}_{i=1}^N$
$\{\mathbf{v}_i\}_{i=1}^N$
$(\cdot)$
$\mathcal{O}(N^2)$
softmax
$(\cdot)$
$\mathcal{O}(N)$
$\{\mathbf{h}_i\}_{i=1}^N$
$\{\mathbf{q}_i\}_{i=1}^N$
$\{\mathbf{k}_i\}_{i=1}^N$
$\{\mathbf{v}_i\}_{i=1}^N$
$\phi(\cdot)$
$\phi(\cdot)$
$(\cdot)$
$\mathcal{O}(N)$
$(\cdot)$
$\mathcal{O}(N)$
$\{\mathbf{h}_i\}_{i=1}^N$

(a) Softmax attention. (b) Random feature attention.

Figure 1: Computation graphs for softmax attention (left) and random feature attention (right), over inputs of length N. $\mathbf{q}_i$, $\mathbf{k}_i$, and $\mathbf{v}_i$ denote query, key, and value vectors, and $\mathbf{h}_i$ the outputs.

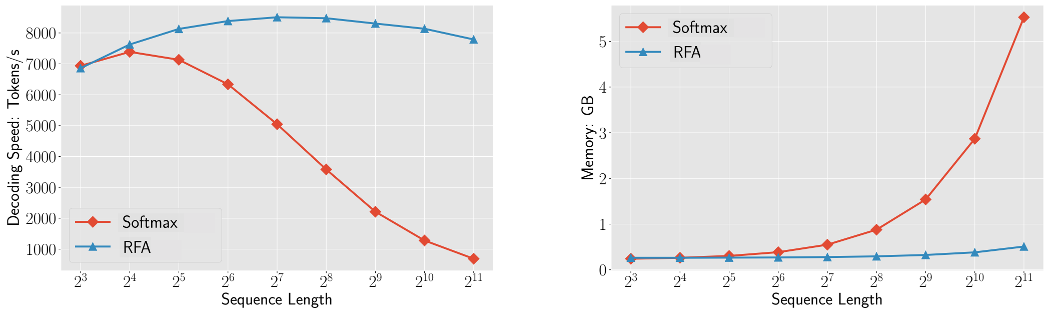
improves time and memory efficiency when applied in transformers, *without* any accuracy loss (e.g., 12× decoding speedup with less than 10% of the memory for 2,048-length sequences; Figure 2). Besides, it connects the attention to recurrent neural networks (RNNs), another important model family in NLP. Such connections allow borrowing existing wisdom from RNNs to enhance attention, which proves especially useful in document-level machine translation [5] and some applications in computer vision [22]. The empirical success of Rfa opens new possibilities: a broader class of kernelized attention offers efficient alternatives to softmax attention. In a follow-up project [4], we observe that *learning* the kernel functions from data improves the accuracy-efficiency tradeoff. Rfa was featured as a spotlight at ICLR 2021 and receives continuing interest from other research groups [23, 24, 22, 25].

Finetuning with efficient attention. Due to their sizes and expensive computation, large pretrained transformers are typically sub-optimal in settings with long inputs or limited computational resources. We propose a swap-then-finetune procedure [2, 4]: it replaces the softmax attention in pretrained transformers with its efficient counterparts (e.g., Rfa) and then finetunes. In practice, this is an appealing approach since it avoids reinvesting the vast amounts of resources already put into pretraining. In an ongoing project, we explore training an efficient attention student from a transformer teacher using knowledge distillation.

Future directions. A long-term goal of my research is to develop efficient methods for NLP. It helps promote the inclusiveness of cutting-edge NLP research and mitigate its negative environmental impact. From an algorithmic perspective, it requires meticulously diving into the mathematical details, rigorous implementations, and familiarity with modern hardware, which I have achieved in the past and will be excited to keep working on. Besides, I believe that getting insights from the tasks and datasets is crucial. For example, we found that simply pairing a shallow decoder with a deep encoder yields significantly faster machine translation decoding *without* accuracy loss [3]. Efficient machine learning has received increasing interest, and it is vital to draw broad and systematic connections among different lines of research to allow for the exchange of progress. I have done so in my previous works [2, 13] and am excited to expand such efforts.

2 Learning with Structural Inductive Biases

Language is structured: meanings are articulated and perceived through hierarchical compositions of words. However, NLP has witnessed a shift away from linguistic structures. In the past, they played a central role. For example, through an NLP pipeline, a sentence is first tagged with parts of speech, then parsed into a syntactic tree, then semantically analyzed, before being fed into a question answering system. Today, general-purpose representations pretrained on massive sequential data are finetuned on task-specific datasets. Although this new paradigm has brought unmatched empirical gains, its sample efficiency and robustness can be enhanced by explicitly encoding structures [9]. **Linguistic structures provide valuable inductive biases** that help models generalize to unseen scenarios; why has the community's



(a) Speed vs. lengths. (b) Memory vs. lengths.

Figure 2: Decoding speed (left) and memory overhead (right) varying the output lengths.

interest in them been fading? Two primary reasons are: (1) for a long time, structured predictors were not accurate enough to produce useful features; (2) traditional NLP pipelines are less flexible than end-to-end training and prone to cascading errors. Answering to these challenges, my research has built **more accurate structured predictors for meaning representations** [6, 7], and proposed **algorithms to train NLP pipelines end-to-end** [8].

Jointly learning multiple meaning representations. Various theories of natural language semantics have been developed, reflecting different linguistic phenomena. They heavily rely on structures to represent meanings. The annotation of semantic structures, already expensive, is fragmented across competing theories. Although these representations may be structurally incompatible, they can be similar in spirit. For example (Figure 3), the phrase "*Only a few books*" fills a semantic role of the frame triggered by "*fell*" (bottom), and at the same time forms a subgraph semantically descending from it (top). We hypothesize that the overlap among different theories can be exploited using multitask learning, allowing us to learn from complementary resources jointly.

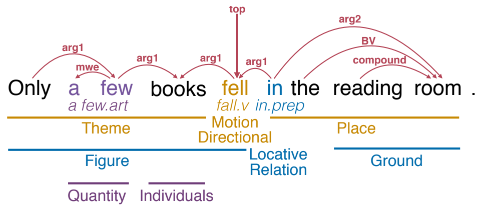


Figure 3: Structural similarities between two meaning representations: bilexical dependencies (top) and phrase-based frame semantics.

Our ACL 2017 work [6] utilizes the invaluable parallel annotations of three semantic formalisms [26]. We propose a joint decoding algorithm on top of other established multitask learning approaches (e.g., parameter sharing). Our results confirm the benefits of symbiosis among multiple semantic representations. This effort is extended to learning from disjoint (i.e., *no* parallel annotations) structurally more divergent resources [7]. These two works have received continuing interest from the community and are broadened by the Cross-Framework Meaning Representation Parsing shared tasks at CoNLL 2019 and 2020 [27, 28].[1] More broadly, they preceded a major shift towards multitask and transfer learning in NLP.

Learning NLP pipelines end-to-end. NLP pipelines run staged structured prediction modules to process text. They are prone to cascading errors and less amenable to neural learning—pipelines make discrete decisions at each stage, incompatible with end-to-end training using backpropagation. Our ACL 2018 work proposes SPIGOT [8], a stable and efficient algorithm to approximate the gradients for discrete structured argmax. SPIGOT calculates a proxy for the gradients **respecting the structured constraints**. As in Figure 4, SPIGOT introduces a projection step to keep the "updated" structures within the relaxed feasible set. SPIGOT allows backpropagating through structured prediction and using it as intermediate layers in neural networks, facilitating training NLP pipelines end-to-end. Our empirical results verify that using SPIGOT to incorporate "learnable" structural inductive biases yields more accurate and

[1] http://mrp.nlpl.eu/2020/index.php

interpretable decisions in downstream tasks. It can also be used in semi-supervised learning and unsupervised structure induction [29]. Spigot received a nomination for the best paper award at ACL 2018.

Future directions. I am passionate about designing NLP models imbued with structural inductive biases. An important aspect is to build neural architectures reflecting linguistic structures, which I have done in my past research [10, 11, 30]. Looking forward, I want to focus on how linguistic structures interact with state-of-the-art attentive models like transformers. In an ongoing project, I am exploring training transformers with a structural prior over the attention distributions. Additionally, improving the efficiency and scalability of structured models is crucial to building more robust, generalizable, and interpretable NLP models in real-world settings. Recent advances in prompt learning provide new opportunities. For example, we bake semantic structural information into a pretrained language model through prompting [12], *without* the efficiency challenges of explicitly encoding structures.

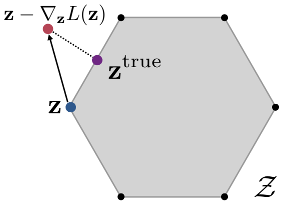

$\mathbf{z} - \nabla_{\mathbf{z}} L(\mathbf{z})$
$\mathbf{z}^{\text{true}}$
$\mathbf{z}$
$\mathcal{Z}$

Figure 4: Illustration of Spigot. The discrete feasible set is relaxed into a convex polytope $\mathcal{Z}$. $\nabla_{\mathbf{z}} L$, the gradients of loss with respect to predicted structures, are computed using backpropagation. If updating $\mathbf{z}$ makes it outside the polytope, it is projected back to $\mathcal{Z}$, resulting in $\mathbf{z}^{\text{true}}$.

3 Formal Analysis of Deep Learning Models

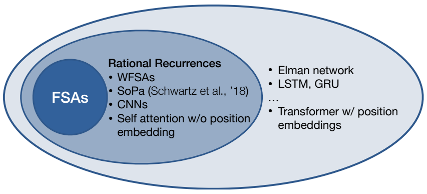

FSAs
Rational Recurrences
• WFSAs
• SoPa (Schwartz et al., '18)
• CNNs
• Self attention w/o position embedding
• Elman network
• LSTM, GRU
...
• Transformer w/ position embeddings

Figure 5: A formal hierarchy of established NLP models in terms of modeling capacities, induced by rational recurrences [31, 13, 32].

Despite their tremendous success in NLP, our understanding of deep learning models lags behind. An important part of my research **grounds modern neural architectures to formal principles in NLP** [13], and uses the insights to develop **better-performing, more efficient, and more interpretable models** [14, 33, 34].

Our EMNLP 2018 work [13] formally studies the connections between modern recurrent neural architectures and weighted finite-state automata (WFSAs). WFSAs naturally relate to pattern-matching methods, offering an intuitive way to explain their decisions (Figure 6); together with their siblings hidden Markov models, they are fundamental tools for NLP. Besides, WFSAs are well-studied, interpretable, efficient, and flexible to design. Our theoretical findings prove that a family of modern recurrent and convolutional neural networks (CNNs), at least in their single-layer cases, are WFSAs parameterized with neural networks, which we dub **rational recurrences**. The significance is both theoretical and empirical: rational recurrences provide a unifying framework for existing approaches and a new way to characterize their capacities (Figure 5); they lead to an algorithm that "translates" WFSAs into neural networks, offering a flexible way to devise new neural architectures imbued with desired inductive biases. Rational models are effective representation learners, as our empirical results show. In a follow-up work [14], we empirically show that rational neural models' decisions are interpretable (Figure 6), and that they are more parameter-efficient and well-suited for low-resource settings. For example, a rationally recurrent model trained with structured sparse regularization performs within 2% of the full model with fewer than 10% of the parameters.

Future directions. Looking ahead, I am passionate about improving our understanding of state-of-the-art deep learning models. Formal methods have been successfully applied to improving past generations of neural architectures; I am excited about extending this effort to attention-based models such as transformers. It will strengthen our understanding of attention, and how it connects to other established neural architectures such as RNNs and CNNs. Practically, it provides practitioners with a flexible way to design neural architectures imbued with desired inductive biases. As recent evidence suggests, a rationally recurrent model equipped with attention can be more accurate and efficient than state-of-the-art transformers [35]. I hope formal analysis and a shared framework for existing approaches could shed some light on future-generation neural networks, which, inevitably, will replace transformers.

Additionally, I aim to bridge the gap between formal principles and practice, and explore better learning paradigms based on the implications of linguistic theories. For example, we are working on a project to provide empirical evidence for the debate whether or not language models can learn meanings without direct supervision [36, 37]. Specifically, we craft grammars satisfying different levels of assumptions, and train language models on synthetic corpora sampled from those grammars and study to what extend they pick up meanings Preliminary results suggest interesting augmentations to the language modeling pretraining, which I am excited to explore.

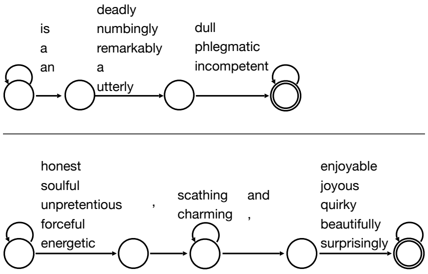


Figure 6: Patterns learned by a rational model, matching negative (top) and positive (bottom) sentiments.

4 Future Research Agenda

Large language models pretrained on massive data have shifted the paradigm of NLP. The empirical gains are genuinely remarkable. However, simply scaling up will hardly lead us to the ultimate goal of building AI with general linguistic capability. Perhaps a more important lesson is that the tasks, data, and evaluation that we thought to be difficult for machines are actually *not*. There has never been a better time for the community to start exploring machine learning models' capability to perform complex reasoning. I am excited to contribute to the reexamination of the tasks, annotation procedures, and evaluation protocols. Meanwhile, algorithmic progress is crucial. I believe that a better understanding of neural models through formal analysis can pave the way towards future neural architectures imbued with desired inductive biases, which are the key to building models that generalize better, are more robust and explainable, and are more sample-efficient to train. Equally importantly, devising efficient methods can help make cutting-edge research more accessible to less-funded institutions, promoting the openness and diversity of the AI community, the very reason it thrives.

Collaborations. Building next-generation NLP models requires joint forces across many fields. For example, my research on formal analysis of neural models complements theoretical approaches to explaining machine learning models' decisions. Learning more robust models relies on insights and inductive biases from linguistics and cognitive science. I am excited to collaborate with experts in these fields to build AI with better linguistic capabilities.

Efficient NLP is largely empowered by advances in hardware and machine learning systems. I plan to work with researchers in systems and architecture to devise algorithms that fully utilize the capacities of modern hardware, and build machine learning systems customized for NLP applications. I expect that the algorithmic aspects of my research expand beyond NLP. For example, I plan to expand my works on efficient transformers to modeling, e.g., DNA sequences or image data, which also require working with long sequences. More broadly, I look forward to working with experts in speech, computer vision, computational biology, reinforcement learning, and robotics to develop machine learning models that can better address the efficiency challenges in various applications.

References

[1] **Hao Peng**, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *Proc. of ICLR*, 2021.

[2] **Hao Peng**, Jungo Kasai, Nikolaos Pappas, Dani Yogatama, Zhaofeng Wu, Lingpeng Kong, Roy Schwartz, and Noah A. Smith. ABC: Attention with bounded-memory control. *arXiv:2110.02488*, 2021. under review.

[3] Jungo Kasai, Nikolaos Pappas, **Hao Peng**, James Cross, and Noah Smith. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *Proc. of ICLR*, 2021.

[4] Jungo Kasai, **Hao Peng**, Yizhe Zhang, Dani Yogatama, Gabriel Ilharco, Nikolaos Pappas, Yi Mao, Weizhu Chen, and Noah A. Smith. Finetuning pretrained transformers into rnns. In *Proc. of EMNLP*, 2021.

[5] Zhaofeng Wu, **Hao Peng**, Nikolaos Pappas, and Noah A. Smith. Modeling context with linear attention for scalable document-level translation. 2021. Under review.

[6] **Hao Peng**, Sam Thomson, and Noah A. Smith. Deep multitask learning for semantic dependency parsing. In *Proc. of ACL*, 2017.

[7] **Hao Peng**, Sam Thomson, Swabha Swayamdipta, and Noah A. Smith. Learning joint semantic parsers from disjoint data. In *Proc. of NAACL*, 2018.

[8] **Hao Peng**, Sam Thomson, and Noah A. Smith. Backpropagating through structured argmax using a spigot. In *Proc. of ACL*, 2018.

[9] Zhaofeng Wu, **Hao Peng**, and Noah A. Smith. Infusing Finetuning with Semantic Dependencies. *TACL*, 9:226–242, 03 2021.

[10] Lili Mou, **Hao Peng**, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. Discriminative neural sentence modeling by tree-based convolution. In *Proc. of EMNLP*, 2015.

[11] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, **Hao Peng**, and Zhi Jin. Classifying relations via long short term memory networks along shortest dependency paths. In *Proc. of EMNLP*, 2015.

[12] Alexis Ross, Tongshuang Wu, **Hao Peng**, Matthew E. Peters, and Matt Gardner. Tailor: Generating and perturbing text with semantic controls. *arXiv:2107.07150*, 2021. Under review.

[13] **Hao Peng**, Roy Schwartz, Sam Thomson, and Noah A. Smith. Rational recurrences. In *Proc. of EMNLP*, 2018.

[14] Jesse Dodge, Roy Schwartz, **Hao Peng**, , and Noah A. Smith. RNN architecture learning with sparse regularization. In *Proc. of EMNLP*, 2019.

[15] David R. So, Quoc V. Le, and Chen Liang. The evolved transformer. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proc. of ICML*, 2019.

[16] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proc. of ACL*, 2019.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of NeurIPS*, 2017.

[18] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 2019.

[20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, P. Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, Tom Henighan, R. Child, Aditya Ramesh, D. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, E. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, A. Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv:2005.14165*, 2020.

[21] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv: 2009.06732*, 2020.

[22] Lin Zheng, Huijie Pan, and Lingpeng Kong. Ripple attention for visual perception with sub-quadratic complexity. *arXiv:2110.02453*, 2021.

[23] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proc. of ICML*, 2021.

[24] Shengjie Luo, Shanda Li, Tianle Cai, Di He, Dinglan Peng, Shuxin Zheng, Guolin Ke, Liwei Wang, and Tie-Yan Liu. Stable, fast and accurate: Kernelized attention with relative positional encoding. In *Proc. of NeurIPS*, 2021.

[25] Sankalan Pal Chowdhury, Adamos Solomou, Avinava Dubey, and Mrinmaya Sachan. On learning the transformer kernel, 2021.

[26] Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, and Zdeňka Urešová. SemEval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proc. of SemEval*, 2015.

[27] Stephan Oepen, Omri Abend, Jan Hajic, Daniel Hershcovich, Marco Kuhlmann, Tim O'Gorman, Nianwen Xue, Jayeol Chun, Milan Straka, and Zdenka Uresova. MRP 2019: Cross-framework meaning representation parsing. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, 2019.

[28] Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, and Daniel Zeman. MRP 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, 2020.

[29] Tsvetomila Mihaylova, Vlad Niculae, and André F. T. Martins. Understanding the mechanics of SPIGOT: Surrogate gradients for latent structure learning. In *Proc. of EMNLP*, 2020.

[30] **Hao Peng**, Roy Schwartz, and Noah A. Smith. PaLM: A hybrid parser and language model. In *Proc. of EMNLP*, 2019.

[31] Roy Schwartz, Sam Thomson, and Noah A. Smith. SoPa: Bridging CNNs, RNNs, and weighted finite-state machines. In *Proc. of ACL*, 2018.

[32] William Merrill, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A. Smith, and Eran Yahav. A formal hierarchy of RNN architectures. In *Proc. of ACL*, 2020.

[33] **Hao Peng**, Ankur P. Parikh, Manaal Faruqui, Bhuwan Dhingra, and Das Dipanjan. Text generation with exemplar-based adaptive decoding. In *Proc. of NAACL*, 2019.

[34] **Hao Peng**, Roy Schwartz, Dianqi Li, and Noah A. Smith. A mixture of h - 1 heads is better than h heads. In *Proc. of ACL*, 2020.

[35] Tao Lei. When attention meets fast recurrence: Training language models with reduced compute. In *Proc. of EMNLP*, 2021.

[36] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proc. of ACL*, 2020.

[37] William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *TACL*, 9:1047–1060, 2021.