Natural Language, Dialog and Speech Symposium (NDS2020)

November 13, 2020 | www.nyas.org/NDS2020

POSTER ABSTRACTS

1. *A Good Summary is Hard to Find! Laying the Groundwork for Advances in Clinical Summarization*
Griffin Adams, PhD Candidate, Columbia University, Mert Ketenci, Noemie Elhadad

2. *Multimodal Emotion Detection via Transfer Learning*
Amith Ananthram, MSc, Columbia University, Amith Ananthram, MSc1, Kailash Karthik1, MSU, Jessica Huynh1, BS, Homayoon Beigi12, PhD, 1 Columbia University, New York, New York, United States 2 Recognition Technologies, Inc., South Salem, New York, United States

3. *On the Semantic Inconsistencies of BERT for Targeted Aspect Based Sentiment Analysis*
Vikas Patidar, M.S student, New York University, Ananth Balashankar, Lakshminarayan Subramanian, New York University

4. *Learning Representations of Causal Evidence Graphs*
Ananth Balashankar, Ph.D Student, New York University, Lakshminarayanan Subramanian, New York University

5. *Active Imitation Learning with Noisy Guidance* [Selected for STAR Talk]
Kianté Brantley, PhD, The University of Maryland College Park, Amr Sharaf, Hal Daumé III, Microsoft Research, New York University

6. *O(n) Connections are Expressive Enough: Universal Approximability of Sparse Transformers*
Yin-Wen Chang, Google Research, Chulhee Yun[1], Yin-Wen Chang[2], Srinadh Bhojanapalli[2], Ankit Singh Rawat[2], Sashank J. Reddi[2], Sanjiv Kumar[2], [1]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA [2]Google Research, New York, New York, USA

7. *Entities as Experts: Sparse Memory Access with Entity Supervision*
Nicholas FitzGerald, PhD, Google Research, Thibault Févry, 1, Livio Baldini Soares, PhD1, Eunsol Choi, PhD2, Tom Kwiatkowski, PhD1, 1Google Reseach, 2University of Texas at Austin

8. *Counterfactually-Augmented SNLI Training Data Does Not Yield Better Generalization Than Unaugmented Data* [Selected for STAR Talk]
William Huang, BSc, New York University, William Huang, BSc1, Haokun Liu, MSc1, and Samuel R. Bowman, PhD1, 1New York University, New York, New York, United States

9. *Adversarial Gaps in Modern Chat Bots*
Josh Kalin, MSc, Department of Computer Science and Software Engineering, Auburn University, Auburn, Alabama, David Noever, PhD2, Matt Ciolino, BSc2, and Gerry Dozier, PhD1, 1Department of Computer Science and Software Engineering, Auburn University, Auburn, Alabama 2PeopleTec, Inc. Huntsville, Alabama

10. *"You Should Probably Read This": Hedge Detection in Text*
Denys Katerenchuk, The Graduate Center, CUNY, Rivka Levitan, PhD, The Graduate Center, CUNY, Brooklyn College, CUNY

11. *Controllable Text Generation from Meaning Representations: Linearization and Data Augmentation Strategies*
Chris Kedzie, Columbia University, Kathleen McKeown

12. *Discourse Coherence, Reference Grounding and Goal Oriented Dialogue*
Baber Khalid, Rutgers University, Malihe Alikhani, Mike Fellner, Brian McMahan, Matthew Stone, University of Pittsburgh, Rutgers Center for Cognitive Science

13. *Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations*
Haau-Sing Li, MSc, New York University Center for Data Science, New York, New York, United States, Alex Warstadt, PhD1, Yian Zhang, BSc2, Haokun Liu, MSc3, Samuel R. Bowman, PhD123, 1New York University Departmant of Linguistics, New York, New York, United States, 2New York University Departmant of Computer Science, New York, New York, United States, 3New York University Center for Data Science, New York, New York, United States.

14. *Semantic Label Smoothing for Sequence to Sequence Problems*
MICHAL LUKASIK, PhD, Google Research, Himanshu Jain, Aditya Krishna Menon, Seungyeon Kim,Srinadh Bhojanapalli, Felix Yu, Sanjiv Kumar

15. *Deception Detection in a Human-Machine Visual Dialogue Task* [Selected for STAR Talk]
Tristan Maidment, PhD, University of Pittsburgh, Patrick Healy, Anthony Sicilia, Dmitriy Babichenko, Malihe Alikhani

16. *Knowledge Graph Based Natural Language Understanding in a Rapid Spoken Dialogue Map-Game*
Deepthi Karkada, MS, Intel Corp, Ramesh Manuvinakurike, Maike Paetzel, Kallirroi Georgila, Intel Corp, Uppsala University, University of Southern California

17. *Parameter Norm Growth During Training of Transformers*
William Merrill, , Allen Institute for AI, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, Noah A. Smith, University of Washington, Bar Ilan University, Hebrew University of Jerusalem

18. *CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models* [Selected for STAR Talk]
Nikita Nangia, New York University, New York City, New York, USA, Clara Vania, Rasika Bhalerao, Samuel R. Bowman

19. *First Impression is the Last Impression? Acoustic-Prosodic Cues to Persuasiveness in Competitive Debate Speeches* [Selected for STAR Talk]
Huyen Nguyen, PhD, Erasmus University Rotterdam, Rotterdam, Netherlands, Sarah Ita Levitan, PhD1, David Lupea, BSc2, Julia Hirschberg, PhD3, 1Hunter College Department of Computer Science, City University of New York, New York, New York, United States. 2New York University, New York, New York, United States. 3Columbia University Department of Computer Science, New York, New York, United States

20. *Query-Key Normalization for Transformers*
Alex Henry, MBA, MA, Cyndx Technologies, New York, New York, United States, Prudhvi Raj Dachapally, MS, Shubham Pawar, MS, Yuxuan Chen, MS

21. *Unsupervised Question Decomposition for Question Answering*
Ethan Perez, Computer Science, Ph.D. Student, New York University, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, Douwe Kiela, Facebook AI Research, University College London, CIFAR Associate Fellow

22. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* [Selected for STAR Talk]
Ethan Perez, Computer Science, Ph.D. Student, New York University, Patrick Lewis, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela, Facebook AI Research, University College London

23. *Probing Saliency in Short Answer Scoring Models for Science Explanations*
Brian Riordan, PhD, ETS, Sarah Bichler, PhD2, Allison Bradford, MSc2, and Marcia C. Linn, PhD2, 2University of California-Berkeley, Berkeley, California, United States

24. *Automatic Fact-Guided Sentence Modification* [Selected for STAR Talk]
Tal Schuster, PhD student, Massachusetts Institute of Technology, Darsh J Shah, Regina Barzilay

25. *Beyond The Text: Analysis of Privacy Statements through Syntactic and Semantic Role Labeling*
Yan Shvartzshnaider, PhD, NYU, Yan Shvartzshnaider, Ananth Balashankar, Vikas Patidar, Thomas Wies, Lakshminarayanan Subramanian, New York University, Courant Institute of Mathematical Sciences

26. *Goal-Oriented Multitask Dialogue Modeling of Supreme Court Oral Arguments* [Selected for STAR Talk]
Ana Smith, MS, Cornell University, Ithaca, New York, United States, Lillian Lee, PhD1, Karen Zhou, 1Cornell University, Ithaca, New York, United States

27. *Asking and Answering Questions to Evaluate the Factual Consistency of Summaries*
Alex Wang, New York University, Kyunghyun Cho, Mike Lewis, Facebook AI

28. *"Talk to Me with Left, Right, and Angles": Lexical Entrainment in Spoken Hebrew Dialogue*
Andreas Weise, MSc, The Graduate Center, CUNY, New York, New York, USA, Vered Silber-Varod, PhD1, Anat Lerner, PhD1,2, Julia Hirschberg, PhD3, Rivka Levitan, PhD4,5, 1Open Media and Information Lab (OMILab), The Open University of Israel, Raanana, Israel, 2Mathematics and Computer Science Department, The Open University of Israel, Raanana, Israel 3Department of Computer Science, Columbia University, New York, New York

1. *A Good Summary is Hard to Find! Laying the Groundwork for Advances in Clinical Summarization*
Griffin Adams, PhD Candidate, Columbia University, Mert Ketenci, Noemie Elhadad

Summary evaluation research has largely focused on salience, fluency, and more recently, factuality. Global evaluation metrics, such as BERTScore, seek to map summaries and references onto a shared semantic space. Local semantic measures, such as FEQA, seek to align single QA pairs between a reference, or source itself, and a model generated summary. Importantly, these metrics do not explicitly assess internal consistency and summary content organization. While effective for most general domain corpora, i.e. news, Wikipedia, they may not be ideal for summarization of clinical narratives. Research has shown that effective synthesis of clinical narratives puts an equally large onus on content organization as it does on salience identification and redundancy minimization. In other words, summary coverage is insufficient; the way the facts are presented, how each fact relates to each other, and the relative ordering, matters. Synthesis of patient information requires linking problem, symptoms, and treatments in a temporally consistent and disorder-specific fashion. We study a particular kind of synthesis: the Brief Hospital Course section of the Discharge Summary, and evaluate it both quantitatively and qualitatively. We demonstrate the shortcomings of existing evaluation metrics at capturing intra-summary coherence, as well as motivate the need for problem-oriented methods of evaluation and content organization.

2. *Multimodal Emotion Detection via Transfer Learning*
Amith Ananthram, MSc, Columbia University, Amith Ananthram, MSc1, Kailash Karthik1, MSU, Jessica Huynh1, BS, Homayoon Beigi12, PhD, 1 Columbia University, New York, New York, United States 2 Recognition Technologies, Inc., South Salem, New York, United States

Automated emotion recognition is a challenging task: emotion is abstract, its expression varies across different modalities, the affect of words is context-dependent and it lacks large datasets. To address these issues, we present an emotion detection system that transfers learning for both speech and text. We begin by training a multilayer TDNN on the task of speaker identification using the VoxCeleb corpus and then fine-tune its final few layers on the task of emotion identification using the CremaD corpus. Using this network, we extract speech embeddings for CremaD from each of its layers, generate and concatenate text embeddings for the accompanying transcripts using a pretrained BERT model and then train an LDA - pLDA model on the resulting dense representations. To understand the merits of each component, we exhaustively evaluate the predictive power of every permutation: the TDNN alone, speech embeddings from each of its layers alone, text embeddings alone and every combination thereof. Our best variant, trained on only VoxCeleb and CremaD and evaluated on IEMOCAP, achieves an EER of 38.05%. Including a portion of IEMOCAP during training improves the 5-fold averaged EER to 25.72% (For comparison, 44.71% of the gold-label annotations include at least one annotator who disagrees). Building on our prior work that shows ASR transferring well to emotion detection, we are currently pursuing a combination of speaker and speech bases for even better accuracies.

3. *On the Semantic Inconsistencies of BERT for Targeted Aspect Based Sentiment Analysis*
Vikas Patidar, M.S student, New York University, Ananth Balashankar, Lakshminarayan Subramanian, New York University

Transformers like BERT have been fine-tuned for supervised tasks with state of the art results. In this paper, we analyze if providing more contextual information leads to better accuracy in BERT fine-tuned sentence-pair classification models for the targeted aspect based sentiment analysis task (TABSA). Our results on the SentiHood and SemEval 2014 Task-4 datasets show that the BERT fine-tuned models do not perform consistently with increase in context size and have a range of up to 6% and 3% in F1 accuracy variations respectively. Further, our analysis shows severe inconsistencies with more than 25% of test samples predictions getting inverted with increase in context size.

4. *Learning Representations of Causal Evidence Graphs*
Ananth Balashankar, Ph.D Student, New York University, Lakshminarayanan Subramanian, New York University

Identifying causal graphs are critical to perform complex natural language reasoning tasks. In this paper, we propose a framework to identify causal graphs based on Granger causality and domain knowledge cause-effect mentions in the text. We learn asymmetric word representations of the underlying causal graph which can be used to perform complex causal reasoning task. We show that using these representations improve F1 accuracy on the causal question answering task and causal link prediction between emergent terms in news streams.

5. *Active Imitation Learning with Noisy Guidance*
Kianté Brantley, PhD, The University of Maryland College Park, Amr Sharaf, Hal Daumé III, Microsoft Research, New York University

Imitation learning algorithms provide state-of-the-art results on many structured prediction tasks by learning near-optimal search policies. Such algorithms assume training-time access to an expert that can provide the optimal action at any queried state; unfortunately, the number of such queries is often prohibitive, frequently rendering these approaches impractical. To combat this query complexity, we consider an active learning setting in which the learning algorithm has additional access to a much cheaper noisy heuristic that provides noisy guidance. Our algorithm, LEAQI, learns a difference classifier that predicts when the expert is likely to disagree with the heuristic, and queries the expert only when necessary. We apply LEAQI to three sequence labeling tasks, demonstrating significantly fewer queries to the expert and comparable (or better) accuracies over a passive approach.

6. *O(n) Connections are Expressive Enough: Universal Approximability of Sparse Transformers*
Yin-Wen Chang, Google Research, Chulhee Yun[1], Yin-Wen Chang[2], Srinadh Bhojanapalli[2], Ankit Singh Rawat[2], Sashank J. Reddi[2], Sanjiv Kumar[2], [1]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA [2]Google Research, New York, New York, USA

Transformer networks use pairwise attention to compute contextual embeddings of inputs, and have redefined the state of the art in many NLP tasks. However, these models suffer from quadratic computational cost in the input sequence length n to compute attention in each layer. This has prompted recent research into faster attention models, with a predominant approach involving sparsifying the connections in the attention layers. While empirically promising for long sequences, fundamental questions remain unanswered: Can sparse transformers approximate any arbitrary sequence-to-sequence function, similar to their dense counterparts? How does the sparsity pattern and the sparsity level affect their performance? In this paper, we address these questions and provide a unifying framework that captures existing sparse attention models. Our analysis proposes sufficient conditions under which we prove that a sparse attention model can universally approximate any sequence-to-sequence function. Surprisingly, our results show the existence of models with only O(n) connections per attention layer that can approximate the same function class as the dense model with n^2 connections. Lastly, we present experiments comparing different patterns/levels of sparsity on standard NLP tasks.

7. *Entities as Experts: Sparse Memory Access with Entity Supervision*
Nicholas FitzGerald, PhD, Google Research, Thibault Févry, 1, Livio Baldini Soares, PhD1, Eunsol Choi, PhD2, Tom Kwiatkowski, PhD1, 1Google Reseach, 2University of Texas at Austin

We focus on the problem of capturing declarative knowledge about entities in the learned parameters of a language model. We introduce a new model - Entities as Experts (EAE) - that can access distinct memories of the entities mentioned in a piece of text. Unlike previous efforts to integrate entity knowledge into sequence models, EAE's entity representations are learned directly from text. We show that EAE's learned representations capture sufficient knowledge to answer TriviaQA questions such as "Which Dr. Who villain has been played by Roger Delgado, Anthony Ainley, Eric Roberts?", outperforming an encoder-generator Transformer model with 10x the parameters. According to the LAMA knowledge probes, EAE contains more factual knowledge than a similarly sized BERT, as well as previous approaches that integrate external sources of entity knowledge. Because EAE associates parameters with specific entities, it only needs to access a fraction of its parameters at inference time, and we show that the correct identification and representation of entities is essential to EAE's performance.

8. *Counterfactually-Augmented SNLI Training Data Does Not Yield Better Generalization Than Unaugmented Data*

William Huang, BSc, New York University, William Huang, BSc1, Haokun Liu, MSc1, and Samuel R. Bowman, PhD1, 1New York University, New York, New York, United States

A growing body of work shows that models exploit annotation artifacts to achieve state-of-the-art performance on standard crowdsourced benchmarks—datasets collected from crowdworkers to create an evaluation task—while still failing on minimally perturbed examples. Recent work has explored the use of counterfactually-augmented data—data built by minimally editing a set of seed examples to yield counterfactual labels—to augment training data associated with these benchmarks and build more robust classifiers that generalize better. We use English natural language inference data to test model generalization and robustness and find that models trained on counterfactually-augmented Stanford Natural Language Inference (SNLI) data do not generalize better compared to similarly large unaugmented datasets, yielding slightly worse out-of-domain performance with a difference of 0.4 points. Further, we find that the data augmentation hurts performance by 6.7 points on one of our evaluation sets with the counterfactually-augmented training set yielding worse results than the seed examples. We thus argue that careful consideration should be given to the trade-offs between seed examples and augmented data in counterfactually-augmented datasets and encourage researchers to explore this line of work before using such data for training.

9. *Adversarial Gaps in Modern Chat Bots*

Josh Kalin, MSc, Department of Computer Science and Software Engineering, Auburn University, Auburn, Alabama, David Noever, PhD2, Matt Ciolino, BSc2, and Gerry Dozier, PhD1, 1Department of Computer Science and Software Engineering, Auburn University, Auburn, Alabama 2PeopleTec, Inc. Huntsville, Alabama

Popular chat bot systems like ParlAI use transformers to supply the backend machine learning model for a given task. Adversarial attacks can be used to manipulate or trick the output of transformers. Recent workshops with ParlAI have focused on safety and security within these conservational systems. In this work, the focus is to demonstrate the gaps in transformer derived systems and suggest fixes to each identified issue. Adversarial actors can exploit weaknesses within these bots to have gibberish like conversations. The exploit works by using words within the language models that have similar embeddings. Likewise, it is also possible to demonstrate manipulation of the positive, negative, and neutral sentiments that can dominate the responses of a chat bot. In retail settings using chat bots with customers, this can provide favorable outcomes to these actors. Finally, a demonstration of production style protections is shown for each attack. The ultimate goal is to demonstrate issues inherent in these systems and provide suggested fixes to each problem. All attacks on these systems are detectable and correctable in a production setting.

10. *"You Should Probably Read This": Hedge Detection in Text*
Denys Katerenchuk, The Graduate Center, CUNY, Rivka Levitan, PhD, The Graduate Center, CUNY, Brooklyn College, CUNY

Humans express ideas, beliefs, and statements through language. The manner of expression can carry information indicating the author's degree of confidence in their statement. Understanding the certainty level of a claim is crucial in areas such as medicine, finance, engineering, etc. where errors can lead to disastrous results. In this work, we apply a joint model that leverages word sentences and part-of-speech tags to improve uncertainty detection in text. This approach achieves an F1 score of 70.24 on the CoNLL-2010 Wikipedia corpus and outperforms current best result. In current work, we explore several approaches and different neural network architectures to improve current top results on CoNLL-2010 Wikipedia corpus. The main contributions of this work are: 1) a comprehensive analysis of various neural network architectures and their performance, 2) a model formulation for including part-of-speech information in the input, 3) a new top score on the CoNLL-2010 Wikipedia dataset. From our results we find that the joint model approach achieves high scores across all three language models. This shows that adding POS-based information can improve performance. However, to our surprise, the GRU model with a custom pre-trained word embedding achieved the highest result.

11. *Controllable Text Generation from Meaning Representations: Linearization and Data Augmentation Strategies*
Chris Kedzie, Columbia University, Kathleen McKeown

We study the degree to which neural sequence-to-sequence (S2S) models exhibit fine-grained controllability when performing natural language generation (NLG) from a meaning representation (MR). In this setting, the NLG model is expected to produce an utterance that faithfully communicates the MR. In the S2S paradigm, the MR must first be "linearized" (i.e. represented as a linear sequence of tokens) before being presented to the S2S encoder. Using two dialogue generation benchmarks, we systematically compare the effect of four MR linearization strategies on controllability and faithfulness. Additionally, we evaluate how a phrase-based data augmentation method can improve performance. We find that properly aligning input sequences during training leads to highly controllable generation on several popular neural S2S models. Such models can follow an "utterance plan" at test time which determines the order in which they realize different parts of the MR. We test the limits of this behavior and evaluate models on their ability to follow the realization order specified by the utterance plan. We compare performance when following plans produced by either a planner model, a human, or difficult, randomly generated plans. We show that the alignment training produces highly controllable NLG models, especially when following a planner model. Furthermore, we demonstrate that phrase-based data augmentation improves the robustness of the control even on more difficult utterance plans.

12. *Discourse Coherence, Reference Grounding and Goal Oriented Dialogue*
Baber Khalid, Rutgers University, Malihe Alikhani, Mike Fellner, Brian McMahan, Matthew Stone, University of Pittsburgh, Rutgers Center for Cognitive Science

Prior approaches to realizing mixed-initiative human--computer referential communication have adopted information-state or collaborative problem-solving approaches. In this paper, we argue for a new approach, inspired by coherence-based models of discourse such as SDRT, in which utterances attach to an evolving discourse structure and the associated knowledge graph of speaker commitments serves as an interface to real-world reasoning and conversational strategy. This helps in reference resolution of anaphora and tracking discourse contributions which is difficult to handle for current dialogue architectures. We show that it is practical to use data-driven dialogue modules to approximate discourse meaning representations, ground these using this graph-like discourse structure and couple it with planning frameworks like RL to learn effective communication strategies. We document our findings on the use of this approach on different goal-oriented dialogue domanis and argue for the adoption of this approach as a practical and scalable way of building dialogue systems.

13. *Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations*
Haau-Sing Li, MSc, New York University Center for Data Science, New York, New York, United States, Alex Warstadt, PhD1, Yian Zhang, BSc2, Haokun Liu, MSc3, Samuel R. Bowman, PhD123, 1New York University Departmant of Linguistics, New York, New York, United States, 2New York University Departmant of Computer Science, New York, New York, United States, 3New York University Center for Data Science, New York, New York, United States.

One reason for the effectiveness in pretraining self-supervised RoBERTa models is that it teaches models linguistic features useful for language understanding. However, we want pretrained models not only to learn these features, but also to use these features preferentially when fine-tuned on target tasks. With this goal in mind, we introduce a new English-language diagnostic set called MSGS (the Mixed Signals Generalization Set), which consists of 20 ambiguous binary classification tasks, designed by pairing 4 linguistic and 5 surface features, to test model preferences for either linguistic or surface generalizations during fine-tuning. MSGS also consists of 9 control tasks for the features, to test whether pretrained models learn features in an unambiguous setting during fine-tuning. We also pretrain RoBERTa models from scratch on data from 1M to 1B words, to compare their performances with RoBERTa-base, which is pretrained on about 30B words. We find that models can learn to represent linguistic features with little pretraining data, but require far more data to learn preferences for linguistic generalizations over surface ones. Eventually, compared with our pretrained models, RoBERTa-base demonstrates a strong linguistic bias. We conclude that while self-supervised pretraining is an effective way to learn helpful inductive biases, there is likely room to improve the rate at which models learn which features matter.

14. *Semantic Label Smoothing for Sequence to Sequence Problems*
Michal Lukasik, PhD, Google Research, Himanshu Jain, Aditya Krishna Menon, Seungyeon Kim,Srinadh Bhojanapalli, Felix Yu, Sanjiv Kumar,

Label smoothing has been shown to be an ef-fective regularization strategy in classification,that prevents overfitting and helps in label de-noising. However, extending such methodsdirectly to seq2seq settings, such as MachineTranslation, is challenging: the large targetoutput space of such problems makes it in-tractable to apply label smoothing over all pos-sible outputs. Most existing approaches forseq2seq settings either do token level smooth-ing, or smooth over sequences generated byrandomly substituting tokens in the target se-quence. Unlike these works, in this paper, wepropose a technique that smooths overwellformedrelevant sequences that not only havesufficient n-gram overlap with the target se-quence, but are alsosemantically similar. Ourmethod shows a consistent and significant im-provement over the state-of-the-art techniqueson different datasets.

15. *Deception Detection in a Human-Machine Visual Dialogue Task*
Tristan Maidment, PhD, University of Pittsburgh, Patrick Healy, Anthony Sicilia, Dmitriy Babichenko, Malihe Alikhani,

When humans attempt to detect deception, they perform two actions: recognize signs of deception, and ask questions to attempt to unveil a deceptive conversational partner. We focus on the latter, constructing a dialogue system that asks questions to attempt to catch a potentially deceptive conversation partner. To explore these complexities in a non-stationary environment, we appeal to an eye-spy style visual dialogue game where a questioner and oracle communicate, achieving common ground to identify a pre-specified object within an image. The questioner interrogates the oracle via yes or no questions, in an attempt to identify some predetermined target entity. To investigate deception, we instruct humans to interact with this autonomous questioner and act in any way they believe would cause the questioner to fail. We use this dialogue to ground an autonomous oracle with human deceptive strategies. We then introduce the questioner to a modified game where it is randomly paired with a cooperative or deceptive oracle with a new goal to either identify the pre-specified object or identify if it is paired with the deceptive oracle. Using reinforcement learning, we train the questioner to succeed in this modified game setting. Our work explores the design of conversational systems which exhibit resilience to human deception in non-stationary environments and establishes a test-bed for investigation of human-machine deception and misinformation.

16. *Knowledge Graph Based Natural Language Understanding in a Rapid Spoken Dialogue Map-Game*
Deepthi Karkada, MS, Intel Corp, Ramesh Manuvinakurike, Maike Paetzel, Kallirroi Georgila, Intel Corp, Uppsala University, University of Southern California

Knowledge graphs (KGs) provide an elegant way to encode domain knowledge for natural language understanding (NLU) tasks in spoken dialogue systems. In this work we present a model for utilizing KG

embeddings for the task of visual reference resolution in a rapid spoken dialogue map-game. The map-game is a two player country identification game between a human describer and an agent selector. The agent has to guess the target country using natural language descriptions from the user (e.g., it is north-west of India, it looks like a rectangle). Our dialogue corpus has been collected in a Wizard of Oz setting and consists of 80 human-Wizard conversations. The agent utilizes KG-based spatial and shapes reasoning to identify the target country. The KG is constructed using structured data (e.g., Wikidata, Wikipedia tables) available from existing resources. Such KG-based NLU helps expand the information available to the system, and we show that this approach outperforms alternate classifier-based approaches, but it still falls short of human performance. We discuss the causes for lower performance of the model compared to humans in the task, which are complex spatial reasoning, utilizing context, and creative shapes descriptions. We also discuss a spatial attention approach that is used by humans to solve the task in the domain.

17. *Parameter Norm Growth During Training of Transformers*
William Merrill, , Allen Institute for AI, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, Noah A. Smith, University of Washington, Bar Ilan University, Hebrew University of Jerusalem

The capacity of neural networks like the widely adopted transformer is known to be very high. Evidence is emerging that they learn successfully due to inductive bias in the training routine, typically some variant of gradient descent (GD). To better understand this bias, we study the tendency of transformer parameters to grow in magnitude during training. We find both theoretically and empirically, that, in certain contexts, GD increases the parameter L2 norm up to a threshold that itself increases with training-set accuracy. This means increasing training accuracy over time enables the norm to increase. Empirically, we show that the norm grows continuously over pretraining for T5 (Raffel, 2019). We show that pretrained T5 approximates a discretized infinite-norm network. Such "saturated'' networks are known to have a reduced capacity compared to the original network family that can be described in automata-theoretic terms. This suggests a new characterization of the inductive bias of GD that is of particular interest for NLP. While our experiments focus on transformers, our theoretical analysis extends to other architectures with similar formal properties, such as feedforward ReLU networks.

18. *CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models*
Nikita Nangia, New York University, New York City, New York, USA, Clara Vania, Rasika Bhalerao, Samuel R. Bowman

Pretrained language models, especially masked language models (MLMs) have seen success across many NLP tasks. However, there is ample evidence that they use the cultural biases that are undoubtedly present in the corpora they are trained on, implicitly creating harm with biased representations. To measure some forms of social bias in language models against protected demographic groups in the US,

we introduce the Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs). CrowS-Pairs has 1508 examples that cover stereotypes dealing with nine types of bias, like race, religion, and age. In CrowS-Pairs a model is presented with two sentences: one that is more stereotyping and another that is less stereotyping. The data focuses on stereotypes about historically disadvantaged groups and contrasts them with advantaged groups. We find that all three of the widely-used MLMs we evaluate substantially favor sentences that express stereotypes in every category in CrowS-Pairs. As work on building less biased models advances, this dataset can be used as a benchmark to evaluate progress.

19. *First Impression is the Last Impression? Acoustic-Prosodic Cues to Persuasiveness in Competitive Debate Speeches*

Huyen Nguyen, PhD, Erasmus University Rotterdam, Rotterdam, Netherlands, Sarah Ita Levitan, PhD1, David Lupea, BSc2, Julia Hirschberg, PhD3, 1Hunter College Department of Computer Science, City University of New York, New York, New York, United States. 2New York University, New York, New York, United States. 3Columbia University Department of Computer Science, New York, New York, United States

Do men and women persuade differently? Are they evaluated differently? Using a data set of over 1800 audio segments of first and last minutes of tournament speeches, their evaluation scores and demographic data, we investigate gender disparity in acoustic-prosodic features and any ensuing impacts on evaluations of persuasiveness. Debate tournaments provide a useful means of systematically answering these questions because they include these four components: (i) a diverse pool of intrinsically motivated professional debaters; (ii) exogenously assigned speaking position, topic and opponents; (iii) transparent scoring criteria, based purely on comparative argumentation strength; and (iv) accountable, selective panels of judges. With this data set, we analyze the acoustic-prosodic correlates of persuasiveness (i.e.: pitch, intensity, harmonic-to-noise ratio (HNR), jitter, shimmer and speaking rate), taking into account individual traits (e.g: gender, native language, institution ranking, study major), to explore the existence and magnitude of discriminatory evaluation standards across social groups. This work contributes a large-scale analysis of acoustic-prosodic cues in a strategically relevant context, and discusses how demographic characteristics of speakers influence judges' perception of persuasive argumentative speeches.

20. *Query-Key Normalization for Transformers*

Alex Henry, MBA, MA, Cyndx Technologies, New York, New York, United States, Prudhvi Raj Dachapally, MS, Shubham Pawar, MS, Yuxuan Chen, MS,

Low-resource language translation is a challenging but socially valuable NLP task. Building on recent work adapting the Transformer's normalization to this setting, we propose QKNorm, a normalization technique that modifies the attention mechanism to make the softmax function less prone to arbitrary

saturation without sacrificing expressivity. Specifically, we apply ℓ2 normalization along the head dimension of each query and key matrix prior to multiplying them and then scale up by a learnable parameter instead of dividing by the square root of the embedding dimension. We show improvements averaging 0.928 BLEU over state-of-the-art bilingual benchmarks for 5 low-resource translation pairs from the TED Talks corpus and IWSLT'15.

21. *Unsupervised Question Decomposition for Question Answering*
Ethan Perez, Computer Science, Ph.D. Student, New York University, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, Douwe Kiela, Facebook AI Research, University College London, CIFAR Associate Fellow

We aim to improve question answering (QA) by decomposing hard questions into simpler sub-questions that existing QA systems are capable of answering. Since labeling questions with decompositions is cumbersome, we take an unsupervised approach to produce sub-questions, also enabling us to leverage millions of questions from the internet. Specifically, we propose an algorithm for One-to-N Unsupervised Sequence transduction (ONUS) that learns to map one hard, multi-hop question to many simpler, single-hop sub-questions. We answer sub-questions with an off-the-shelf QA model and give the resulting answers to a recomposition model that combines them into a final answer. We show large QA improvements on HotpotQA over a strong baseline on the original, out-of-domain, and multi-hop dev sets. ONUS automatically learns to decompose different kinds of questions, while matching the utility of supervised and heuristic decomposition methods for QA and exceeding those methods in fluency. Qualitatively, we find that using sub-questions is promising for shedding light on why a QA system makes a prediction.

22. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*
Ethan Perez, Computer Science, Ph.D. Student, New York University, Patrick Lewis, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela, Facebook AI Research, University College London

Large pretrained language models have been shown to store factual knowledge in their parameters and achieve state of the art when finetuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, so on knowledge-intensive tasks, their performance lags behind task-specific architectures. Also, explaining their decisions and updating their world knowledge remain open research problems. Pretrained models with a differentiable access mechanism to explicit non-parametric memory can overcome this issue but have so far been only investigated for extractive downstream tasks. We explore a general-purpose finetuning recipe for retrieval-augmented generation (RAG) - models that combine parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pretrained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pretrained retriever.

We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, the other can use different passages per token. We finetune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set state of the art on three open-domain QA tasks. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

23. *Probing Saliency in Short Answer Scoring Models for Science Explanations*
Brian Riordan, PhD, ETS, Sarah Bichler, PhD2, Allison Bradford, MSc2, and Marcia C. Linn, PhD2, 2University of California-Berkeley, Berkeley, California, United States

Recent work on automated scoring of student responses in educational applications has shown gains in human-machine agreement from neural models, particularly recurrent neural networks (RNNs) and pre-trained transformer (PT) models. However, prior research has neglected investigating the reasons for improvement – in particular, whether models achieve gains for the "right" reasons. Through expert analysis of saliency maps, we analyze the extent to which models attribute importance to words and phrases in student responses that align with question rubrics. We focus on responses to questions that are embedded in science units for middle school students accessed via an online classroom system. RNN and PT models were trained to predict an ordinal score from each response's text, and experts analyzed generated saliency maps for each response. Our analysis shows that RNN and PT-based models can produce substantially different saliency profiles while often predicting the same scores for the same student responses. While there is some indication that PT models are better able to avoid spurious correlations of high frequency words with scores, results indicate that both models focus on learning statistical correlations between scores and words and do not demonstrate an ability to learn key phrases or longer linguistic units corresponding to ideas, which are targeted by question rubrics. These results point to a need for models to better capture student ideas in educational applications.

24. *Automatic Fact-Guided Sentence Modification*
Tal Schuster, PhD student, Massachusetts Institute of Technology, Darsh J Shah, Regina Barzilay,

Online encyclopediae like Wikipedia contain large amounts of text that need frequent corrections and updates. The new information may contradict existing content in encyclopediae. In this paper, we focus on rewriting such dynamically changing articles. This is a challenging constrained generation task, as the output must be consistent with the new information and fit into the rest of the existing document. To this end, we propose a two-step solution: (1) We identify and remove the contradicting components in a target text for a given claim, using a neutralizing stance model; (2) We expand the remaining text to be consistent with the given claim, using a novel two-encoder sequence-to-sequence model with copy attention. Applied to a Wikipedia fact update dataset, our method successfully generates updated sentences for new claims, achieving the highest SARI score. Furthermore, we demonstrate that

generating synthetic data through such rewritten sentences can successfully augment the FEVER fact-checking training dataset, leading to a relative error reduction of 13%.

25. *Beyond The Text: Analysis of Privacy Statements through Syntactic and Semantic Role Labeling*
Yan Shvartzshnaider, PhD, NYU, Yan Shvartzshnaider, Ananth Balashankar, Vikas Patidar, Thomas Wies, Lakshminarayanan Subramanian, New York University, Courant Institute of Mathematical Sciences

This paper formulates a new task of extracting privacy parameters from a privacy policy, through the lens of Contextual Integrity, an established social theory framework for reasoning about privacy norms. Privacy policies, written by lawyers, are lengthy and often comprise incomplete and vague statements. In this paper, we show that traditional NLP tasks, including the recently proposed Question-Answering based solutions, are insufficient to address the privacy parameter extraction problem and provide poor precision and recall. We describe 4 different types of conventional methods that can be partially adapted to address the parameter extraction task with varying degrees of success: Hidden Markov Models, BERT fine-tuned models, Dependency Type Parsing (DP) and Semantic Role Labeling (SRL). Based on a detailed evaluation across 36 real-world privacy policies of major enterprises, we demonstrate that a solution combining syntactic DP coupled with type-specific SRL tasks provides the highest accuracy for retrieving contextual privacy parameters from privacy statements. We also observe that incorporating domain-specific knowledge is critical to achieving high precision and recall, thus inspiring new NLP research to address this important problem in the privacy domain.

26. *Goal-Oriented Multitask Dialogue Modeling of Supreme Court Oral Arguments*
Ana Smith, MS, Cornell University, Ithaca, New York, United States, Lillian Lee, PhD1, Karen Zhou, 1Cornell University, Ithaca, New York, United States

Dialogue modeling has advanced in a number of domains, but many complex domains such as the Supreme Court of the United States (SCOTUS) remain understudied. Goal-oriented dialogues like those found in SCOTUS have more strictly defined outcomes than open-domain dialogues, while having less strictly scripted language than task-oriented dialogues. This work evaluates the effect of balancing turn-level goal prediction tasks -- (a) deciding an outcome of a conversation and (b) addressing a topic -- with a next turn ranking task in a multitask setting using the ParlAI dialogue framework and SCOTUS oral arguments. While most meeting corpora such as AMI and ICSI involve multiple roles and multifarious meeting objectives, SCOTUS dialogues have relatively simple outcomes (vote for or against petitioner), two sides arguing opposing points (petitioner/respondent), and strict power distinctions (Justice/Counsel). This work experiments with modeling Justice turns in a Counsel-Justice dyadic exchange. Justices' turns are targeted as they participate in multiple conversations and exert more control over the dialogue than Counsel. This enables us to control for power, initiative, and speaker traits. We jointly model a next turn ranking task and a turn-level goal prediction task, the latter of which may use a vote objective for (a) or a topic objective for (b). The effects of different objectives are analyzed in aggregate, in ideological aggregate, and for each speaker.

27. *Asking and Answering Questions to Evaluate the Factual Consistency of Summaries*
Alex Wang, New York University, Kyunghyun Cho, Mike Lewis, Facebook AI

Practical applications of abstractive summarization models are limited by frequent factual inconsistencies with respect to their input. Existing automatic evaluation metrics for summarization are largely insensitive to such errors. We propose Question Answering and Generation for Summarization (QAGS, pronounced ``kags''), an automatic evaluation protocol that is designed to identify factual inconsistencies in a generated summary. QAGS is based on the intuition that if we ask questions about a summary and its source, we will receive similar answers if the summary is factually consistent with the source. To evaluate QAGS, we collect human judgments of factual consistency on model-generated summaries for the CNN/DailyMail and XSUM summarization datasets. QAGS has substantially higher correlations with these judgments than other automatic evaluation metrics. Also, QAGS offers a natural form of interpretability: The answers and questions generated while computing QAGS indicate which tokens of a summary are inconsistent and why.

28. *"Talk to Me with Left, Right, and Angles": Lexical Entrainment in Spoken Hebrew Dialogue*
Andreas Weise, MSc, The Graduate Center, CUNY, New York, New York, USA, Vered Silber-Varod, PhD1, Anat Lerner, PhD1,2, Julia Hirschberg, PhD3, Rivka Levitan, PhD4,5, 1Open Media and Information Lab (OMILab), The Open University of Israel, Raanana, Israel, 2Mathematics and Computer Science Department, The Open University of Israel, Raanana, Israel 3Department of Computer Science, Columbia University, New York, New York

It has been well-documented for several languages that human interlocutors tend to adapt their linguistic productions to become more similar to each other. This behavior, known as entrainment, affects lexical choice as well, both with regard to specific words, such as referring expressions, and overall style. We offer what we believe to be the first investigation of such lexical entrainment in Hebrew. Using two existing measures, we analyze Hebrew speakers interacting in a Map Task, a popular experimental setup, and find rich evidence of lexical entrainment. Analyzing speaker pairs by the combination of their genders as well as speakers by their individual gender, we find no clear pattern of differences. We do, however, find that speakers in a position of less power entrain more than those with greater power, which matches theoretical accounts. Overall, our results mostly accord with those for American English, with a lack of entrainment on hedge words being the main difference.