Welcome to the Tutorials Session of ACL 2020.

The ACL tutorials session is organized to give conference attendees a comprehensive introduction by expert researchers to some topics of importance drawn from our rapidly growing and changing research field.

This year, as has been the tradition over the past few years, the call, submission, reviewing and selection of tutorials were coordinated jointly for multiple conferences: ACL, AACL-IJCNLP, COLING and EMNLP. We formed a review committee of 19 members, including the ACL tutorial chairs (Agata Savary and Yue Zhang), the EMNLP tutorial chairs (Banjamin van Durme and Aline Villavicencio), the COLING tutorial chairs (Daniel Beck and Lucia Specia), the AACL-IJCNMP tutorial chairs (Timothy Baldwin and Fei Xia) and 11 external reviewers (Emily Bender, Erik Cambria, Gaël Dias, Stefan Evert, Yang Liu, João Sedoc, Xu Sun, Yulia Tsvetkov, Taro Watanabe, Aaron Steven White and Meishan Zhang). A reviewing process was organised so that each proposal receives 3 reviews. The selection criteria included clarity, preparedness, novelty, timeliness, instructors’ experience, likely audience, open access to the teaching materials, diversity (multilingualism, gender, age and geolocation) and the compatibility of preferred venues. A total of 43 tutorial submissions were received, of which 8 were selected for presentation at ACL.

We solicited two types of tutorials, including cutting-edge themes and introductory themes. The 8 tutorials for ACL include of 3 introductory tutorials and 5 cutting-edge tutorials. The introductory tutorials are dedicated to reviewing, ethics and commonsense reasoning in NLP. The cutting-edge discussions address interpretability of neural NLP, multi-modal information extraction and dialogue, stylized text generation, and open-domain question answering.

We would like to thank the tutorial authors for their contributions and flexibility while organising the conference virtually. We are also grateful to the 11 external reviewers for their generous help in the decision process. Finally, our thanks go to the conference organizers for effective collaboration, and in particular to the general chair Dan Jurafsky, the website chairs Sudha Rao and Yizhe Zhang, the publicity chair Emily Bender, the ACL anthology director Matt Post.

We hope you enjoy the tutorials.

ACL 2020 Tutorial Co-chairs
Agata Savary
Yue Zhang
General Chair
   Dan Jurafsky, Stanford University

Program Chairs
   Joyce Chai, University of Michigan
   Natalie Schluter, IT University of Copenhagen, Denmark
   Joel Tetreault, Dataminr, USA

Tutorial Chairs
   Agata Savary, University of Tours, France
   Yue Zhang, Westlake University, China
# Table of Contents

*Interpretability and Analysis in Neural NLP*
  Yonatan Belinkov, Sebastian Gehrmann and Ellie Pavlick ........................................... 1

*Integrating Ethics into the NLP Curriculum*
  Emily M. Bender, Dirk Hovy and Alexandra Schofield .................................................... 6

*Achieving Common Ground in Multi-modal Dialogue*
  Malihe Alikhani and Matthew Stone .................................................................................. 10

*Reviewing Natural Language Processing Research*
  Kevin Cohen, Karën Fort, Margot Mieskes and Aurélie Névéol ........................................ 16

*Stylized Text Generation: Approaches and Applications*
  Lili Mou and Olga Vechtomova ......................................................................................... 19

*Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web*
  Xin Luna Dong, Hannaneh Hajishirzi, Colin Lockard and Prashant Shiralkar ..................... 23

*Commonsense Reasoning for Natural Language Processing*
  Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi and Dan Roth ......................... 27

*Open-Domain Question Answering*
  Danqi Chen and Wen-tau Yih ............................................................................................. 34
Conference Program

Sunday, July 5, 2020

6:00–9:30  *Interpretability and Analysis in Neural NLP*
(10:00–13:30)
Yonatan Belinkov, Sebastian Gehrmann and Ellie Pavlick

6:00–9:30  *Integrating Ethics into the NLP Curriculum*
(10:00–13:30)
Emily M. Bender, Dirk Hovy and Alexandra Schofield

6:00–9:30  *Achieving Common Ground in Multi-modal Dialogue*
(15:00–18:30)
Malihe Alikhani and Matthew Stone

10:00–13:30  *Reviewing Natural Language Processing Research*
Kevin Cohen, Karën Fort, Margot Mieskes and Aurélie Névéol

10:00–13:30  *Stylized Text Generation: Approaches and Applications*
Lili Mou and Olga Vechtomova

15:00–18:30  *Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web*
Xin Luna Dong, Hannaneh Hajishirzi, Colin Lockard and Prashant Shiralkar

15:00–18:30  *Commonsense Reasoning for Natural Language Processing*
Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi and Dan Roth

15:00–18:30  *Open-Domain Question Answering*
Danqi Chen and Wen-tau Yih
Abstract

While deep learning has transformed the natural language processing (NLP) field and impacted the larger computational linguistics community, the rise of neural networks is stained by their opaque nature: It is challenging to interpret the inner workings of neural network models, and explicate their behavior. Therefore, in the last few years, an increasingly large body of work has been devoted to the analysis and interpretation of neural network models in NLP.

This body of work is so far lacking a common framework and methodology. Moreover, approaching the analysis of modern neural networks can be difficult for newcomers to the field. This tutorial aims to fill this gap and introduce the nascent field of interpretability and analysis of neural networks in NLP.

The tutorial will cover the main lines of analysis work, such as structural analyses using probing classifiers, behavioral studies and test suites, and interactive visualizations. We will highlight not only the most commonly applied analysis methods, but also the specific limitations and shortcomings of current approaches, in order to inform participants where to focus future efforts.

1 Tutorial Description

Deep learning has transformed the NLP field and impacted the larger computational linguistics community. Neural networks have become the preferred modeling approach for various tasks, from language modeling, through morphological inflection and syntactic parsing, to machine translation, summarization, and reading comprehension.

The rise of neural networks is, however, stained by their opaque nature. In contrast to earlier approaches that made use of manually crafted features, it is more challenging to interpret the inner workings of neural network models, and explicate their behavior. Therefore, in the last few years, an increasingly large body of work has been devoted to the analysis and interpretation of neural network models in NLP.

The topic has so far been represented in two dedicated workshops (Blackbox 2018 and 2019) and was recently established as a track in the main *CL conferences. Due to these recent developments, methods for the analysis and interpretability of neural networks in NLP are so far lacking a common framework and methodology. Moreover, approaching the analysis of modern neural networks can be difficult for newcomers to the field, since it requires both a familiarity with recent work in neural NLP and with analysis methods which are not yet standardized. This tutorial aims to fill this gap and introduce the nascent field of interpretability and analysis of neural networks in NLP.

The tutorial will cover the main lines of analysis work, mostly drawing on the recent TACL survey by Belinkov and Glass (2019). In particular, we will devote a large portion to work aiming to find linguistic information that is captured by neural networks, such as probing classifiers (Hupkes et al., 2018; Adi et al., 2017; Conneau et al., 2018a,b; Tenney et al., 2019b, inter alia), controlled behavior studies on language modelling (Gulordava et al., 2018; Linzen et al., 2016a; Marvin and Linzen, 2018) or inference tasks (Poliak et al., 2018a,b; White et al., 2017; Kim et al., 2019; McCoy et al., 2019; Ross and Pavlick, 2019), psycholinguistic methods (Ettinger et al., 2018; Chrupala and Alishahi, 2019), layerwise analyses (Peters et al., 2018; Tenney et al., 2019a), among other methods (Hewitt and Manning, 2019; Zhang

---

1 A comprehensive bibliography is found in the accompanying website of the survey: https://boknilev.github.io/nlp-analysis-methods/.
and Bowman, 2018; Shi et al., 2016). We will also present various interactive visualization methods such as neuron activations (Karpathy et al., 2015; Dalvi et al., 2019), attention mechanisms (Bahdanau et al., 2014; Strobelt et al., 2018), and saliency measures (Li et al., 2016; Murdoch et al., 2018; Arras et al., 2017), including a walkthrough on how to build a simple attention visualization. Next, we will discuss the construction and use of challenge sets for fine-grained evaluation in the context of different tasks (Conneau and Kiela, 2018; Wang et al., 2018; Isabelle and Kuhn, 2018; Sennrich, 2017, inter alia). Finally, we will review work on generating adversarial examples in NLP, focusing on the challenges brought upon by the discrete nature of textual input (Papernot et al., 2016b; Ebrahimi et al., 2018; Jia and Liang, 2017; Belinkov and Bisk, 2018, inter alia). A detailed outline is provided in Section 3.

Throughout the tutorial, we will highlight not only the most commonly applied analysis methods, but also the specific limitations and shortcomings of current approaches. By the end of the tutorial, participants will be better informed where to focus future research efforts.

2 Tutorial Type

This tutorial will cover cutting-edge research in interpretability and analysis of modern neural NLP models. The topic has not been previously covered in *CL tutorials.

3 Outline

1. Introduction
2. Structural Analyses
   (a) Methodology: Analysis by Probing Classifiers
   (b) Example Studies: Different Components and Linguistic Phenomena
   (c) Limitations
3. Behavioral Studies
   (a) Background on Test Suites and Challenge Sets
   (b) Types of Probing Tasks
   (c) Experimental Designs
   (d) Construction Methods
   (e) Languages
4. Interaction and Visualization
   (a) How Interaction can help and its limitations
   (b) Classification and Review of Related Efforts
   (c) Demo Walk-through: Simple Attention Visualization
   (d) Broader Perspectives and Opportunities
5. Other Methods
   (a) Generating Explanations
   (b) Psycholinguistic Methods
   (c) Testing on Formal Languages
6. Conclusion

4 Prerequisites

We would assume acquaintance with core linguistic concepts and basic knowledge of machine learning and neural networks, such as covered in most introductory NLP courses.

5 Reading List

In addition to the papers mentioned in this proposal, a comprehensive bibliography can be found in the following website: https://boknilev.github.io/nlp-analysis-methods/.

For trainees interested in reading important studies before the tutorial, we recommend the following: Belinkov and Glass (2019); Hupkes et al. (2018); Tenney et al. (2019b); Linzen et al. (2016b); Ettinger et al. (2018); Bahdanau et al. (2014); Li et al. (2016); Sennrich (2017); Papernot et al. (2016a); Ebrahimi et al. (2018).

6 Names and Affiliations

Yonatan Belinkov, Postdoctoral Fellow, Harvard University and MIT
email: belinkov@seas.harvard.edu
website: http://people.csail.mit.edu/belinkov

Yonatan Belinkov is a Postdoctoral Fellow at the Harvard School of Engineering and Applied Sciences (SEAS) and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). His research interests are in interpretability and robustness of neural models of language. He has done previous work in machine translation, speech recognition, community question answering, and syntactic parsing. His research has been
published at ACL, EMNLP, NAACL, CL, TACL, ICLR, and NeurIPS. His PhD dissertation at MIT analyzed internal language representations in deep learning models. He co-organized or co-organizes BlackboxNLP 2019, BlackboxNLP 2020, and the WMT 2019 machine translation robustness task, and serves as an area chair for the analysis and interpretability track at ACL and EMNLP 2020.

Sebastian Gehrmann, Research Scientist, Google AI
e-mail: gehrmann@google.com
website: http://sebastiangehrmann.com

Sebastian is research scientist at Google AI. He received his PhD in 2020 from Harvard University. His research focuses on the development and evaluation of controllable and interpretable models for language generation. By applying methods from human-computer interaction and visualization to problems in NLP, he develops interactive interfaces that help with the interpretation and explanation of neural networks. His research has been published at ACL, NAAACL, EMNLP, CHI, and IEEE VIS. He received an honorable mention at VAST 2018 and was nominated for ACL best demo 2019 for his work on interactive visualization tools. He co-organized INLG 2019 and served as an area chair in summarization for ACL 2020.

Ellie Pavlick, Assistant Professor of Computer Science, Brown University
e-mail: ellie.pavlick@brown.edu
website: http://cs.brown.edu/people/epavlick

Ellie Pavlick is an Assistant Professor at Brown University and a Research Scientist at Google. She received her PhD in 2017 with her thesis on modeling compositional lexical semantics. Her current work focuses on computational models of semantics and pragmatics, with a focus on building cognitively-plausible representations. Her recent work has focused on “probing” distributional models in order to better understand the linguistic phenomena that are and are not encoded “for free” via language modelling. Her work has been published at ACL, NAACL, EMNLP, TACL, *SEM, and ICLR, including two best paper awards at *SEM 2016 and 2019. Ellie co-organized the 2018 JSALT summer workshop on building and evaluating general-purpose sentence representations. She also served as area chair for ACL’s sentence-level semantics track.

References


Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018b. What you can cram into a single $&!* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th Annual Meeting of


Integrating Ethics into the NLP Curriculum

Emily M. Bender
University of Washington
ebender@uw.edu

Dirk Hovy
Bocconi University
dirk.hovy@unibocconi.it

Xanda Schofield
Harvey Mudd College
xanda@cs.hmc.edu

1 Description

As NLP technology becomes more ubiquitous, it has ever more impact on the lives of people all around the world. As a field, we have become increasingly aware that we have a responsibility to evaluate the effects of our research and mitigate harmful outcomes. This is true for both researchers and developers in universities, government labs, and industry. However, without experience of how to productively engage with the many ethical conundrums in NLP, it is easy to become overwhelmed and remain inactive. To raise awareness among future NLP practitioners and prevent inertia in the field, we need to place ethics in the curriculum for all NLP students—not as an elective, but as a core part of their education. Though ethical considerations are achieving new currency in NLP, similar issues have been under consideration for decades, if not centuries, in other fields, and there are robust existing practices for approaching these problems. The difference is that there is no agreed-upon way to engage with them in our field.

Our goal in this tutorial is to empower NLP researchers and practitioners with tools and resources to teach others about how to ethically apply NLP techniques. Our tutorial will present both high-level strategies for developing an ethics-oriented curriculum, based on experience and best practices, as well as specific sample exercises that can be brought to a classroom.

We consider three primary topics with our session that frequently underlie ethical issues in NLP research:

1. Dual Use: Learning how to anticipate how a developed technology could be repurposed for harmful or negative results, and designing systems so that they do not inadvertently cause harm.

2. Bias: Understanding the different ways in which bias interacts with language data, including over- and under-sampling of different populations as well as the effects of human bias expressed in language; building less biased datasets and debiasing trained models; strategies for matching appropriate training data to a given use case.

3. Privacy: Protecting the privacy of speakers/writers of text used in the construction or evaluation of a new NLP technology.

In this setting, a key lesson is that there is no single approach to ethical NLP: each project requires thoughtful consideration about what steps can be taken to best support people affected by that project. However, we can learn (and teach) what kinds of issues to be aware of and what kinds of strategies are available for mitigating harm. To teach this process, we apply and promote interactive exercises that provide an opportunity to ideate, discuss, and reflect. We plan to facilitate this in a way that encourages positive discussion, emphasizing the creation of ideas for the future instead of negative opinions of previous work.

2 Type of tutorial

Introductory. Though this is a topic of importance to the NLP community internally, it relies
on existing expertise from both pedagogical and philosophical work, and it is not meant to depend on any particular research area of NLP. However, we do believe the content of this workshop also explores questions not fully answered in our field about concrete best practices in the specific context of NLP courses.

A note on interactivity: The proposed format of this tutorial is different from many past introductory tutorials, in the sense that it relies heavily on participation as part of the instruction. However, we believe this is a necessary part of the format of this tutorial for several reasons:

- Because our tutorial is focused on pedagogy, it makes sense to use effective and equitable pedagogical classroom techniques in it. Interactivity through active or cooperative learning (Slavin, 1980; Johnson and Johnson, 2008) and guided discovery-based learning (Alfieri et al., 2011) are proven to enable students to learn more effectively across diverse classrooms, and our design models this.

- The outcome of this tutorial is one focused on training and professional development, which comes with practice. In the same way one might encourage developing a sample neural network in a tutorial on deep learning, we encourage performing steps of educational practice to develop skills to then use in our lives as instructors.

- While there exists literature in ethics pedagogy and ethics in NLP, there do not exist large pools of resources and papers to refer when designing a course, but instead only a small collection of syllabi for ethics in machine learning/NLP courses. An interactive tutorial format allows us to use the learning experiences of our participants as a starting point to construct a more centralized pool of resources from which faculty and educators in NLP can draw.

3 Outline

1. Introduction, background, motivation [10m]

2. Core concepts and terminology, and warm up exercises. [50m] We will have the participants discuss what motivates them and core concepts of ethics and pedagogy that might be useful in the subsequent ideation.

3. Big class exercise I [55m] (5 minutes intro, 35 minutes doing the exercise with the group, 10 minutes talking about how to teach it). The exercises in this set are centered around thinking through how systems behave in the world. There will be a separate exercise for each of the three groups: dual use, bias, and privacy.

   **Dual Use** A student approaches you because they want to explore gendered language in the LGBTQ community. They are very engaged in the community themselves and have access to data. Their plan is to write a text classification tool that distinguishes LGBTQ from heterosexual language. What do you tell the student?

   **Bias** Pick an application of speech/language technology, determine what kind of training data is typically used for it (whose language? recorded when/where/how?). Next, imagine real world use cases for this technology. What speaker groups would come in contact with the system? If their language differs from substantially from the training data, what would the failure mode of the system be and what would the real-world impacts of that failure be? How could systems, their training data or documentation be designed to be robust to this kind of problem?

   **Privacy** Consider a simple Naive Bayes classifier trained on a subset of 20 News-groups using word frequencies as features. For five sample messages, could you tell whether or not they were included in the subset? How would you check? How certain could you be?

4. Big class exercise II [55m] (5 minutes intro, 25 minutes refining the exercise, 25 minutes talking about how to teach it). The exercises in this set involve building a system and observing its behavior.

   **Dual Use** (1) An ACL submission claims to be able to undo ciphers used by dissenters on social media. Who benefits from this? Is it better to release it in a peer-reviewed venue than to not know it? (2) You develop a tool that can detect depression with high accuracy.
Why, or why not, should you release it as an app?

**Bias** Taking inspiration from Speer (2017b), build a sentiment analysis system over restaurant reviews using different sources of training data for word embeddings. What kind of biases can be observed in system behavior for different types of cuisine? What patterns in language use in the underlying training data are responsible? What kinds of analogous problems can arise in other systems that use word embeddings as input?

**Privacy** Design a small search engine around an inverted index that uses random integer noise from a two-sided geometric distribution (Ghosh et al., 2012) to shape which queries are retrieved. Analyze how much this changes the search results with different noise levels. Are there systematic changes?

5. Wrap up [20m]: big points, reflections from people, where to find resources and keep talking

4 **Prerequisites**

This tutorial is meant to be accessible to anyone actively working with NLP and either currently teaching, interested in teaching, or interested in informal instruction outside of university contexts.

5 **Reading List**

We recommend the following short readings to get a sense of the kinds of issues we will be approaching:

- Dual Use: Ehni 2008
- Bias: Speer 2017a
- Privacy: Coavoux et al. 2018

In addition, we recommend the following papers for a sense of what can be learned from other fields:

- Value scenarios, a technique from value-sensitive design: Nathan et al. 2007
- A history of notions of fairness in education and hiring: Hutchinson and Mitchell 2019
- Disparate impact: Feldman et al. 2015

Participants are encouraged to have read at least some of these papers ahead of time, but familiarity with all of them will not be assumed.

6 **Instructors**

Emily M. Bender
University of Washington
ebender@uw.edu
faculty.washington.edu/ebender

Emily M. Bender is a Professor of Linguistics and Adjunct Professor of Computer Science and Engineering at the University of Washington. Her research interests include computational semantics, grammar engineering, computational linguistic typology, and ethics in NLP. She is the Faculty Director of UW’s Professional Masters in Computational Linguistics (CLMS) and has been engaged with integrating ethics into the CLMS curriculum since 2016. She co-organized the first EthNLP workshop. Her first publication in this area is the TACL paper “Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science” (Bender and Friedman, 2018) and she has been an invited speaker at workshops and panels related to ethics and NLP (or AI more broadly) at the Taskar Memorial Event (UW, March 2018), The Future of Artificial Intelligence: Language, Ethics, Technology (Cambridge, March 2019), West Coast NLP (Facebook, September 2019), Machine Learning Competitions for All (NeurIPS, December 2019) and AAAS (Seattle, February 2020).

Xanda Schofield
Harvey Mudd College
xanda@cs.hmc.edu
www.cs.hmc.edu/~xanda

Xanda Schofield is an Assistant Professor of Computer Science at Harvey Mudd College. Her work focuses on the practical aspects of using distributional semantic models for analysis of real-world datasets, with problems ranging from understanding the consequences of data pre-processing on model inference (Schofield and Mimno, 2016; Schofield et al., 2017) to enforcing text privacy for these models (Schein et al., 2018). She also is interested in pedagogy at this intersection, having co-developed a Text Mining for History and Literature course at Cornell University with David Mimno. She is currently focusing pedagogical ef-
forts on how to introduce considerations of ethics and bias into other courses such as Algorithms.

**Dirk Hovy**
Bocconi University
dirk.hovy@unibocconi.it
www.dirkhovy.com

Dirk Hovy is an Associate Professor of Computer Science in the Department of Marketing at Bocconi University in Milan, Italy. His research focuses on how social dimensions influence language and in turn NLP models, as well as on questions of bias and fairness. He strives to integrate sociolinguistic knowledge into NLP models to counteract demographic bias. Dirk has written on ethics and bias in NLP (Hovy and Spruit, 2016), co-organized two editions of the EthNLP workshops and one of the abusive language workshop, and was an invited speaker on panels on ethics at NAACL 2018 and SLT 2018. He is teaching a related tutorial (on ethics and biases) at CLiC-IT in November 2019.

**References**


Achieving Common Ground in Multi-modal Dialogue

Malihe Alikhani
Computer Science
Rutgers University
malihe.alikhani@rutgers.edu

Matthew Stone
Computer Science
Rutgers University
matthew.stone@rutgers.edu

1 Description

All communication aims at achieving common ground (grounding): interlocutors can work together effectively only with mutual beliefs about what the state of the world is, about what their goals are, and about how they plan to make their goals a reality (Clark et al., 1991). Computational dialogue research, in particular, has a long history of influential work on how implemented systems can achieve common ground with human users, from formal results on grounding actions in conversation (Traum, 1994) to machine learning results on how best to fold confirmation actions into dialogue flow (Levin et al., 1998; Walker, 2000). Such classic results, however, offer scant guidance to the design of grounding modules and behaviors in cutting-edge systems, which increasingly combine multiple communication modalities, address complex tasks, and include the possibility for lightweight practical action interleaved with communication. This tutorial is premised on the idea that it’s time to revisit work on grounding in human–human conversation, particularly Brennan’s general and important characterization of grounding as seeking and providing evidence of mutual understanding (Brennan, 1990), in light of the opportunities and challenges of multi-modal settings such as human–robot interaction.

In this tutorial, we focus on three main topic areas: 1) grounding in human-human communication; 2) grounding in dialogue systems; and 3) grounding in multi-modal interactive systems, including image-oriented conversations and human–robot interactions. We highlight a number of achievements of recent computational research in coordinating complex content, show how these results lead to rich and challenging opportunities for doing grounding in more flexible and powerful ways, and canvass relevant insights from the literature on human–human conversation. We expect that the tutorial will be of interest to researchers in dialogue systems, computational semantics and cognitive modeling, and hope that it will catalyze research and system building that more directly explores the creative, strategic ways conversational agents might be able to seek and offer evidence about their understanding of their interlocutors.

Grounding in human-human communication. Clark et al. (1991) argued that communication is accomplished in two phases. In the presentation phase, the speaker presents signals intended to specify the content of the contributions. In the second phase, the participants work together to establish mutual beliefs that serve the purposes of the conversation. The two phases together constitute a unit of communication—contributions. Clark and Krych (2004) show how this model applies to coordinated action, while Stone and Stojnić (2015) applies the model to text-and-video presentations.

Figure 1: Examples of the generation and interpretation of grounded referring expressions in multimodal interactive settings. Grounding is making sure that the listener understands what the speaker said.

A: A green bike with tan handlebars. B: Got it (Manuvinakurike et al., 2017)

A: The green cup is called Bill. B: Ok, the green cup is Bill. [point to the inferred object] (Liu and Chai, 2015)
Coherence is key.

**Grounding in dialogue systems.** Computer systems achieve grounding mechanistically by ensuring they get attention and feedback from their users, tracking user state, and planning actions with reinforcement learning to resolve problematic situations. We will review techniques for maintaining engagement (Sidner et al., 2005; Bohus and Horvitz, 2014; Foster et al., 2017) and problems that arises in joint attention (Kontogiorgos et al., 2018) and turn taking such as incremental interpretation ( DeVault and Stone, 2004; DeVault et al., 2011), ambiguity resolution (DeVault and Stone, 2009) and learning flexible dialogue management policies (Henderson et al., 2005). Similar questions have been studied in the context of instruction games (Perera et al., 2018; Thomason et al., 2019; Suhr and Artzi, 2018), and interactive tutoring systems (Yu et al., 2016; Wiggins et al., 2019).

**Grounding in multi-modal systems.** Multi-modal systems offer the ability to use signals such as nodding, certain hand gestures and gazing at a speaker to communicate meaning and contribute to establishing common ground (Mavridis, 2015). However, multi-modal grounding is more than just using pointing to clarify. Multi-modal systems have diverse opportunities to demonstrate understanding. For example, recent work has aimed to bridge vision, interactive learning, and natural language understanding through language learning tasks based on natural images (Zhang et al., 2018; Kazemzadeh et al., 2014; De Vries et al., 2017a; Kim et al., 2020). The work on visual dialogue games (Geman et al., 2015) brings new resources and models for generating referring expression for referents in images (Suhr et al., 2019; Shekhar et al., 2018), visually grounded spoken language communication (Roy, 2002; Gkatzia et al., 2015), and captioning (Levinboim et al., 2019; Alikhani and Stone, 2019), which can be used creatively to demonstrate how a system understand a user. Figure 1 shows two examples of models that understand and generate referring expressions in multi-modal settings.

Similarly, robots can demonstrate how they understand a task by carrying it out—in research on interactive task learning in human-robot interaction (Zarrieß and Schlangen, 2018; Carlmeyer et al., 2018) as well as embodied agents performing interactive tasks (Gordon et al., 2018; Das et al., 2018) in physically simulated environments (Anderson et al., 2018; Tan and Bansal, 2018) often drawing on the successes of deep learning and reinforcement learning (Branavan et al., 2009; Liu and Chai, 2015). A lesson that can be learned from this line of research is that one main factor that affects grounding is the choice of medium of communication. Thus, researchers have developed different techniques and methods for data collection and modeling of multimodal communication (Alikhani et al., 2019; Novikova et al., 2016). Figure 2 shows two example resources that were put together using crowdsourcing and virtual reality systems. We will discuss the strengths and shortcomings of these methods.

We pay special attention to non-verbal grounding in languages beyond English, including German (Han and Schlangen, 2018), Swedish (Kontogiorgos, 2017), Japanese (Endrass et al., 2013; Nakano et al., 2003), French (Lemaignan and Alami, 2013; Steels, 2001), Italian (Borghi and Cangelosi, 2014; Taylor et al., 1986), Spanish (Kery et al., 2019), Russian (Janda, 1988), and American sign language (Emmorey and Casey, 1995). These investigations often describe important language-dependent characteristics and cultural differences in studying non-verbal grounding.

**Grounding in end-to-end language & vision systems.** With current advances in neural mod-
elling and the availability of large pretrained models in language and vision, multi-modal interaction often is enabled by neural end-to-end architectures with multimodal encodings, e.g. by answering questions about visual scenes (Antol et al., 2015; Das et al., 2017). It is argued that these shared representations help to ground word meanings. In this tutorial, we will discuss how this type of lexical grounding relates to grounding in dialogue from a theoretical perspective (Larsson, 2018), as well as within different interactive application scenarios – ranging from interactively identifying an object (De Vries et al., 2017b) to dialogue-based learning of word meanings (Yu et al., 2016). We then critically review existing datasets and shared tasks and showcase some of the shortcomings of current vision and language models, e.g. (Agarwal et al., 2018). In contrast to previous ACL tutorials on Multimodal Learning and Reasoning, we will concentrate on identifying different grounding phenomena as identified in the first part of this tutorial.

2 Outline

We begin by discussing grounding in human-human communication (~20 min). After that, we discuss the role of grounding in spoken dialogue systems (~30 min) and visually grounded interactions including grounding visual explanations in images and multimodal language grounding for human-robot collaboration (~90 min). We then survey methods for developing and testing multimodal systems to study non-verbal grounding (~20 min). We follow this by describing common solution concepts and barrier problems that cross application domains and interaction types (~20 min).

3 Prerequisites and reading list

The tutorial will be self-contained. For further readings, we recommend the following publications that are central to the non-verbal grounding framework as of late 2019:

1. Grounding in communication, Herb Clark and Susan Brennan. (Clark et al., 1991)
2. Meaning and demonstration by Una Stojnic and Matthew Stone (Stone and Stojnić, 2015)
4. Language to Action: Towards Interactive Task Learning with Physical Agents, Joyce Y. Chai by Joyce Y. Chai et al.(Chai et al., 2018)
7. A survey of nonverbal signaling methods for non-humanoid robots by Elizabeth Cha et al. (Cha et al., 2018)
8. The Devil is in the Details: A Magnifying Glass for the GuessWhich Visual Dialogue Game by Alberto Testoni et al. (Testoni et al., 2019)

4 Authors

Malihe Alikhani is a 5th year Ph.D. student in the department of Computer Science at Rutgers University mail95@cs.rutgers.edu, advised by Prof. Matthew Stone. She is pursuing a certificate in cognitive science through the Rutgers Center for Cognitive Science and holds a BA and MA in Mathematics. Her research aims at teaching machines to understand and generate multimodal communication. She is the recipient of the fellowship award for excellence in computation and data sciences from Rutgers Discovery Informatics Institute in 2018 and the Anita Berg student fellowship in 2019. Before joining Rutgers, she was a lecturer and an adjunct professor of Mathematics and Statistics for a year at San Diego State University and San Diego Mesa College. She has served as the program committee of ACL, NAACL, EMNLP, AAAI, ICRL, ICMI, and INLG and is currently the associate editor of the Mental Note Journal. email: mail95@cs.rutgers.edu, webpage: www.malihealikhani.com

Matthew Stone is professor and chair in the Department of Computer Science at Rutgers University; he holds a joint appointment in the Rutgers Center for Cognitive Science. His research focuses on discourse, dialogue and natural language
generation; he is particularly interested in leveraging semantics to make interactive systems easier to build and more human-like in their behavior. He was program co-chair for NAACL 2007, general co-chair for SIGDIAL 2014. He has also served as program co-chair for INLG and IWCS, as an information officer for SIGSEM, and on the editorial board for Computational Linguistics.

email: mdstone@cs.rutgers.edu, website: www.cs.rutgers.edu/~mdstone/

Acknowledgments

Preparation of this tutorial was supported in part by the DATA-INSPIRE Institute at Rutgers University, website: http://robotics.cs.rutgers.edu/data-inspire/ under NSF HDR TRIPODS award CCF-1934924. We gratefully acknowledge the effort of Professor Verena Rieser of Heriot-Watt University, who discussed the tutorial with us extensively but was ultimately unable to participate due to the disruption of COVID-19.

References


Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In Proceedings of the 16th international conference on multimodal interaction. ACM.


Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017b. Guesswhat!! visual object discovery through multi-modal dialogue. In CVPR.


Reviewing Natural Language Processing Research

Kevin B. Cohen
Computational Bioscience Program
University of Colorado, USA
kevin.cohen@gmail.com

Karen Fort
Sorbonne Université, EA STIH, Paris
LORIA, Nancy
France
karen.fort@sorbonne-universite.fr

Margot Mieskes
University of Applied Sciences, Darmstadt
Germany
margot.mieskes@h-da.de

Aurélie Névéol
LIMSI, CNRS
Université Paris-Saclay
France
neveol@limsi.fr

1 Tutorial Content

This tutorial will cover the goals, processes, and evaluation of reviewing research in natural language processing. As has been pointed out for years by leading figures in our community (Webber, 2007), researchers in the ACL community face a heavy—and growing—reviewing burden. Initiatives to lower this burden have been discussed at the recent ACL general assembly in Florence (ACL 2019)\textsuperscript{1}. Simultaneously, notable “false negatives”—rejection by our conferences of work that was later shown to be tremendously important after acceptance by other conferences (Church, 2005)—has raised awareness of the fact that our reviewing practices leave something to be desired...and we do not often talk about “false positives” with respect to conference papers, but conversations in the hallways at *ACL meetings suggest that we have a publication bias towards papers that report high performance, with perhaps not much else of interest in them (Manning, 2015).

It need not be this way. There is good reason to think that reviewing is a learnable (and teachable) skill (Basford, 1990; Paice, 2001; Benos et al., 2003; Koike et al., 2009; Shukla, 2010; Tandon, 2014; Spyns and Vidal, 2015; Stahel and Moore, 2016; Kohnen, 2017; McFadden et al., 2017; Hill, 2018). To address the issues raised above, we propose this tutorial on reviewing natural language processing research, focusing on conference submissions and various review forms used in the NLP community. The extended part also covers journal submissions.

As the demand for reviewing grows, so must the pool of reviewers. As the survey presented by Graham Neubig at the 2019 ACL showed, a considerable number of reviewers are junior researchers, who might lack the experience and expertise necessary for high-quality reviews. A tutorial on this topic might increase reviewers’ confidence, as well as the quality of the reviews. Given the importance of conferences in NLP, the reviewing standards should be as high as with journals in other fields.

2 Timetable

Table 1 shows an outline of the content discussed during the tutorial. Apart from a general introduction to the topic of peer reviewing and its role in the publishing circle, we will go into details on reviewing for *ACL-venues. All sections will include exercises and practical examples to get a better grasp for individual elements mentioned during the theoretical input. We will also take a look at problems with respect to peer reviewing and specific peer reviewing models, such as double-blind reviewing, which is the primary mode in *ACL-publication venues vs. single-blind and open reviewing. The case study will look at an actual example paper including reviews for that example.

3 Suggested Reading List


\textsuperscript{1}http://www.livecongress.it/aol/indexSA.php?id=E2EAED7D&ticket=
• Mark Peplow. 2014. *Peer review reviewed*. *Nature*

4 **Presenters (in alphabetical order)**

**Kevin Bretonnel Cohen** has written, overseen, and received hundreds of reviews in his capacity as deputy editor-in-chief of a biomedical informatics journal, associate editor of five natural language processing or bioinformatics journals, special issue editor, workshop organizer, and author of 100+ publications in computational linguistics and natural language processing. His forthcoming book *Writing about data science research: With examples from machine and natural language processing* includes coverage of a number of aspects of the reviewing process. His current research focuses on issues of reproducibility.

**Karén Fort** is an associate professor at Sorbonne Université. Besides being a reviewer for most major NLP conferences, she has been editor in chief for a *Traitement automatique des langues* journal special issue on ethics and acted as Area Chair for ACL in 2017 and 2018 (as senior AC). Her main research interests are ethics, and the construction of language resources for natural language processing. She co-authored the report on the EMNLP reviewer survey (*Névél et al., 2017*).

**Aurélie Névéol** is a permanent researcher at LIMSI CNRS and Université Paris Saclay. She has been involved in reviewing natural language processing papers at many stages of the reviewing process, including: reviewer, associate editor for three journals, area chair for *ACL* and bioinformatics conferences, workshop organizer. Her research focuses on biomedical natural language processing as well as ethics issues in NLP research. She co-authored the report on EMNLP reviewer survey (*Névéol et al., 2017*).
References
Stylized Text Generation: Approaches and Applications

Lili Mou
University of Alberta; Amii
doublepower.mou@gmail.com

Olga Vechtomova
University of Waterloo
ovechtom@uwaterloo.ca

Type of Tutorial: Cutting-edge.

1 Tutorial Introduction

Text generation has played an important role in various applications of natural language processing (NLP), such as paraphrasing, summarization, and dialogue systems. With the development of modern deep learning techniques, text generation is usually accomplished by a neural decoder (e.g., a recurrent neural network or a Transformer), which generates a word at a time conditioned on previous generated words. The decoder can be further conditioned on some source information, such as a source language sentence in machine translation, or a previous utterance in dialogue systems. In recent studies, researchers are paying increasing attention to modeling and manipulating the style of the generation text, which we call stylized text generation in this tutorial. The goal is to not only model the content of text (in traditional text generation), but also control some “style” of the text, for example, the persona of a speaker in a dialogue (Li et al., 2016), or the sentiment of product reviews (Hu et al., 2017).

Stylized text generation is related to various machine learning techniques, for example, embedding learning techniques to represent style (Fu et al., 2018), adversarial learning and reinforcement learning with cycle consistency to match “content” but to distinguish different styles (Hu et al., 2017; Xu et al., 2018; John et al., 2019); very recent work is even able to disentangle latent features in an unsupervised way (Xu et al., 2019).

In this tutorial, we will provide a comprehensive literature review on stylized text generation. We start from the definition of style and different settings of stylized text generation, illustrated with various applications.

In the second part, we will describe style-conditioned text generation. In this category, style serves as a certain type of source information, which the decoder is conditioned on. We describe three types of approaches: (1) embedding-based techniques that capture the style information by real-valued vectors, which can be used to condition a language model (Tikhonov and Yamshchikov, 2018) or concatenated with the input to a decoder (Li et al., 2016; Vechtomova et al., 2018) (2) approaches that encode both style and content in the latent space (Shi et al., 2019a; Yang et al., 2017; Li et al., 2020). We will discuss techniques that structure latent space to encode both style and content, and include Gaussian Mixture Model Variational Autoencoders (GMM-VAE) (Shi et al., 2019a; Wang et al., 2019a; Shi et al., 2019b), Conditional Variational Autoencoders (CVAE) (Yang et al., 2017), and Adversarially Regularized Autoencoders (ARAE) (Li et al., 2020). (3) approaches with multiple style-specific decoders (Syed et al., 2019; Chen et al., 2019). We highlight several applications including persona-based dialogue generation (Li et al., 2016) and creative writing (Yang et al., 2017; Tikhonov and Yamshchikov, 2018; Vechtomova et al., 2018).

Next, we will introduce evaluation methods for style-conditioned text generation. We will present the current practice in the literature, involving both human evaluation and automatic metrics. A few important evaluation aspects include the success of being in the target style, the preservation of content information, as well as language fluency in general.

In the third part, we will focus on style-transfer text generation. Given an input sentence of a certain style, the goal of style transfer is to synthesize a new sentence that has the same content but with different styles. Particularly, style-transfer text generation can be categorized into three settings: (1) Parallel-supervised style transfer, where a par-
allel corpus is available (Xu et al., 2012; Rao and Tetreault, 2018). This is similar to machine translation, but semi-supervised learning is adopted to address small-data training (Wang et al., 2019b). (2) Non-parallel style transfer, where each sentence is annotated by a style label (e.g., positive or negative sentiment). This setting is the most explored setting in previous style transfer literature. We will discuss classification losses to distinguish different styles (John et al., 2019), and adversarial losses/cycle consistency to match content information (Shen et al., 2017). We will also present an editing-based approach that edits style-specific words and phrases into the desired style (Li et al., 2018). (3) Unsupervised style transfer, where the entire corpus is unlabeled (no parallel pairs or style labels). In recent studies, researchers have applied auxiliary losses (such as orthogonality penalty) to detect the most prevalent variation of text in a corpus, and are sometimes able to accomplish style transfer in a purely unsupervised fashion. Since unsupervised style transfer is new to NLP and less explored, we will also introduce several studies in the computer vision domain, bringing future opportunities to text generation in this setting (Gatys et al., 2016; Chen et al., 2016).

Next, we will discuss style adversarial text generation (Zhang et al., 2019). The setting of adversarial attacks is similar to style transfer in that it aims to change the style classifier’s prediction. However, the synthesized sentence in this setting should in fact keep the actual style as humans perceive, but “fool” the style classifier. Thus, it is known as the adversarial attack. We will discuss style adversarial generation in the character level, the word level, as well as the sentence level. Techniques include discrete word manipulation and continuous latent space manipulation.

Finally, we will conclude our tutorial by presenting the challenges of stylized text generation and discussing future directions, such as small-data training, non-categorical style modeling, and a generalized scope of style transfer (e.g., controlling the syntax as a style (Bao et al., 2019)).

By the end of the tutorial, the audience will have a systematic view of different settings of stylized text generation, understand common techniques to model and manipulate the style of text, and be able to apply existing approaches to new scenarios that require stylized text generation. Our tutorial also investigates stylized generative models in non-NLP domains, and thus would inspire future NLP studies in this direction.

2 Tutorial Outline

PART I: Introduction (20 min)

• Definition of style
• Settings and Problem formulations
• Examples of style (e.g., sentiment, artistic style, grammatical style)

PART II: Style-Conditioned Text Generation (50 min)

• Techniques
  – Encoding style in embeddings: sequence-to-sequence models with style embeddings, style conditioned language models, Variational Autoencoder (VAE) with style embeddings;
  – Encoding style and content in latent space: Conditional Variational Autoencoder (CVAE) Gaussian Mixture Variational Autoencoder (GMM-VAE), Adversarially Regularized Autoencoder (ARAE).
  – Models with multiple style-specific decoders
• Applications
  – Creative text generation (e.g., poetry composition)
  – Persona and emotion conditioned dialogue models
  – Stylized image caption generation
• Evaluation measures
  – Stylistic adherence
  – Content preservation
  – Language fluency
  – Novelty and diversity

PART III: Style-Transfer Text Generation (60 min)

• Parallel supervised style transfer
  – Sequence-to-sequence learning
  – Semi-supervised training with limited parallel data
– Applications: Shakespearean–modern English transfer, formality style transfer

• Non-parallel supervised style transfer
  – Auxiliary classification for style modeling
  – Adversarial learning for matching content
  – Cycle consistency for content matching
  – Edit-based style transfer
  – Applications: Sentiment, genre and grammatical style transfer

• Unsupervised style transfer
  – Approaches: Mutual information penalties and correlation penalties for automatic style detection
  – A brief introduction of unsupervised style transfer in image domain (e.g., color, shape, angle)

PART IV: Style-Adversarial Text Generation
(30 minutes)

• Style adversarial vs. style transfer

• Approaches
  – Character-level attack
  – Word-level attack
  – Sentence-level attack

PART IV: Conclusion, Future Work, and Q&A
(20 min)

• Challenges: non-categorical style, small-data training

• A broader view of “style”: text summarization/simplification as style transfer, syntax-semantic disentanglement

3 Instructors

Lili Mou
doublepower.mou@gmail.com
https://lili-mou.github.io

Dr. Lili Mou is an Assistant Professor at the Department of Computing Science, University of Alberta. He is also an Amii Fellow and a Canadian CIFAR AI Chair. Lili received his BS and PhD degrees in 2012 and 2017, respectively, from School of EECS, Peking University. After that, he worked as a postdoctoral fellow at the University of Waterloo and a research scientist at Adeptmind (a startup in Toronto, Canada). His research interests include deep learning applied to natural language processing as well as programming language processing. Recently, he has been focusing more on text generation, from both continuous latent space and discrete word space. He has more than 30 papers published at top-tier conferences and journals, including AAAI, ACL, CIKM, COLING, EMNLP, ICASSP, ICLR, ICML, IJCAI, INTERSPEECH, NAACL-HLT, and TACL. He presented a tutorial “Discreteness in Neural Natural Language Processing” at EMNLP-IJCNLP’19.

Olga Vechtomova
ovechtomova@uwaterloo.ca
https://ov-research.uwaterloo.ca

Dr. Olga Vechtomova is an Associate Professor in the Department of Management Sciences, Faculty of Engineering, cross-appointed in the School of Computer Science at the University of Waterloo. Olga leads the Natural Language Processing Lab, affiliated with the Waterloo.AI Institute. Her research has been supported by a number of industry and government grants, including Amazon Research Award and Natural Sciences and Engineering Research Council (NSERC). The research in her Lab is mainly focused on designing deep neural networks for natural language generation tasks. Her current and recent projects include controlled text generation, text style transfer, and designing text generative models for creative applications. She has over 50 publications in NLP and Information Retrieval conferences and journals, including NAACL-HLT, COLING, ACL, ACM SIGIR, and CIKM. She and her colleagues recently received the ACM SIGIR 2019 Test of Time Award.

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), under grant Nos. RGPIN-2019-04897 and RGPIN-2020-04465. Lili Mou is also supported by AltaML, the Amii Fellow Program, and the Canadian CIFAR AI Chair Program.

References


Yuan Li, Chunyuan Li, Yizhe Zhang, Xiujun Li, Guoqing Zheng, Lawrence Carin, and Jianfeng Gao. 2020. Complementary auxiliary classifiers for label-conditional text generation. In AAAI.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GY AFC dataset: Corpus, benchmarks and metrics for formality style transfer. In NAACL-HLT, pages 129–140.


Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web

Xin Luna Dong  
Amazon  
lunadong@amazon.com

Hannaneh Hajishirzi  
University of Washington  
Allen Institute for AI  
hannaneh@washington.edu

Colin Lockard  
University of Washington  
lockardc@cs.washington.edu

Prashant Shiralkar  
Amazon  
shiralp@amazon.com

Abstract

How do we surface the large amount of information present in HTML documents on the Web, from news articles to Rotten Tomatoes pages to tables of sports scores? Such information can enable a variety of applications including knowledge base construction, question answering, recommendation, and more. In this tutorial, we present approaches for information extraction (IE) from Web data that can be differentiated along two key dimensions: 1) the diversity in data modality that is leveraged, e.g., text, visual, XML/HTML, and 2) the thrust to develop scalable approaches with zero to limited human supervision.

1 Description

Motivation: The World Wide Web contains vast quantities of textual information in several forms: unstructured text, template-based semi-structured webpages (which present data in key-value pairs and lists), and tables. Methods for extracting information from these sources and converting it to a structured form have been a target of research from the natural language processing (NLP), data mining, and database communities. While these researchers have largely separated extraction from web data into different problems based on the modality of the data, they have faced similar problems such as learning with limited labeled data, defining (or avoiding defining) ontologies, making use of prior knowledge, and scaling solutions to deal with the size of the Web.

In this tutorial we take a holistic view toward information extraction, exploring the commonalities in the challenges and solutions developed to address these different forms of text. We will explore the approaches targeted at unstructured text that largely rely on learning syntactic or semantic textual patterns, approaches targeted at semi-structured documents that learn to identify structural patterns in the template, and approaches targeting web tables which rely heavily on entity linking and type information.

While these different data modalities have largely been considered separately in the past, recent research has started taking a more inclusive approach toward textual extraction, in which the multiple signals offered by textual, layout, and visual clues are combined into a single extraction model made possible by new deep learning approaches. At the same time, trends within purely textual extraction have shifted toward full-document understanding rather than considering sentences as independent units. With this in mind, it is worth considering the information extraction problem as a whole to motivate solutions that harness textual semantics along with visual and semi-structured layout information. We will discuss these approaches and suggest avenues for future work.

Tutorial Content: We will start by defining unstructured, semi-structured, and tabular text, and discussing the challenges and opportunities that differentiate these data sources, as well as those they have in common. We will then provide introductions to the basic models and learning algorithms used in extraction from unstructured, semi-structured, and tabular text. We will pay special attention to methods that enable extraction to be expanded to the scope of entity and relation types found on the web, such as the distant supervision and data programming paradigms of creating training data, and schema-less “OpenIE” extraction. After introducing the separate approaches targeting these data modalities, we will then explore research that combines signals from textual, visual, and layout information to consider all aspects of a document.

Throughout the tutorial, we will bring together lessons learned from the different communities involved in information extraction research and will
provide insights from industry experiences building a production knowledge graph leveraging both unstructured and semi-structured text. Section 3 contains a full outline of planned content.

Tutorial slides are available at https://sites.google.com/view/acl-2020-multi-modal-ie

Relevance to ACL: Information Extraction is a core task in natural language processing, with the web serving as a rich source of information for constructing knowledge bases (KBs). A 2018 NAACL tutorial, “Scalable Construction and Reasoning of Massive Knowledge Bases” (Ren et al., 2018), provided an overview of recent IE and KB research. However, like most NLP research, that tutorial focused on methods that treat text as a simple string of natural language sentences in a txt file, while many real-world documents convey information via visual and layout relationships. A separate line of information extraction work has focused on learning to extract from these template-based documents. As interest in multi-modal NLP techniques has grown in recent years, we think the community will be interested in a tutorial that compares and contrasts these approaches and examines recent research that brings together textual, visual, and layout features of documents.

2 Type of the tutorial:

The tutorial will cover cutting-edge work in both unstructured and semi-structured information extraction, including visual and GCN-based approaches. However, our coverage of semi-structured and tabular IE will cover introductory material since it is likely new to much of the NLP community.

3 Outline

1. (30 mins) Introduction and Applications
   - Knowledge Base Population
     - Intro to knowledge graphs
     - Applications
     - Industry examples
     - Importance of the long tail
   - Unstructured, Semi-structured, and Tabular text
     - Unstructured Text
     - HTML and DOM trees
     - Webtables
     - Template learning vs. generalization
   - Schema-aligned extraction vs. OpenIE
   - Common challenges, opportunities, and key intuitions

2. (45 mins) IE from unstructured text:
   - Tasks
     - Named Entity Recognition
     - Co-reference Resolution
     - Relation Extraction
     - Event Extraction
   - Featurization and Modeling
     - OpenTag (Zheng et al., 2018)
     - DyGIE (Luan et al., 2019)
   - Limited Training Data
     - Distant Supervision (Mintz et al., 2009)
     - Data Programming (Ratner et al., 2017)
   - OpenIE

3. (45 mins) IE from semi-structured documents
   - Supervised Wrapper Induction
     - Vertex (Gulhane et al., 2011)
   - Distantly Supervised approaches
     - LODIE (Ciravegna et al., 2012)
     - DIADEM (Furche et al., 2012)
     - Ceres (Lockard et al., 2018)
   - OpenIE / Schema-less approaches
     - WEIR (Bronzi et al., 2013)
     - OpenCeres (Lockard et al., 2019)

4. (15 mins) IE from tables
   - WebTables (Cafarella et al., 2018)
   - Subject detection (Venetis et al., 2011)
   - Joint approaches (LimayeGirija et al., 2010)

5. (30 mins) Multi-modal extraction
   - Benefits of multi-modal extraction
     - Connecting tables and text (Ibrahim et al., 2019)
   - Visual signals for keyphrase extraction (Xiong et al., 2019)
   - Documents as images (Katti et al., 2018)
   - GCN-based encoders (Qian et al., 2019; Liu et al., 2019)
   - Multi-modal signals for creating training data (Wu et al., 2018)
Multi-modal OpenIE

6. (15 mins) Conclusion and Open Directions

4 Prerequisites
The tutorial should be accessible to anyone with a background in natural language processing. It would be helpful to have a basic understanding of classification algorithms, preferably with some knowledge of neural network approaches, as well as unsupervised clustering algorithms.

5 Reading list
- “Web-Scale Information Extraction With Vertex”, Gulhane et al. (2011)
- “Ten Years of WebTables”, Cafarella et al. (2018)
- “Fonduer: Knowledge Base Construction from Richly Formatted Data”, Wu et al. (2018)
- “Extraction and Integration of Partially Overlapping Web Sources” Bronzi et al. (2013)
- “Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion”, Dong et al. (2014)
- “GraphIE: A Graph-Based Framework for Information Extraction”, Qian et al. (2019)

6 Presenters
In alphabetical order,

**Xin Luna Dong** is a Principal Scientist at Amazon, leading the efforts of constructing Amazon Product Knowledge Graph. She was one of the major contributors to the Google Knowledge Vault project, and has led the Knowledge-based Trust project, which is called the “Google Truth Machine” by the Washington Post. She co-authored the book “Big Data Integration”, was awarded ACM Distinguished Member, VLDB Early Career Research Contribution Award for “advancing the state of the art of knowledge fusion”, and Best Demo award in Sigmod 2005. She serves on the VLDB endowment and PVLDB advisory committee, and was a PC co-chair for VLDB 2021, ICDE Industry 2019, VLDB Tutorial 2019, Sigmod 2018 and WAIM 2015. She has given multiple tutorials on data integration, graph mining, and knowledge management.
Email: lunadong@amazon.com
Homepage: http://lunadong.com/.

**Hannaneh Hajishirzi** is an Assistant Professor at the Paul G. Allen School of Computer Science & Engineering at the University of Washington. She works on NLP, AI, and machine learning, particularly designing algorithms for semantic understanding, reasoning, question answering, and information extraction from multimodal data. She has earned numerous awards for her research, including an Allen Distinguished Investigator Award, a Google Faculty Research Award, a Bloomberg Data Science Award, an Amazon Research Award, and a SIGDIAL Best Paper Award.
Email: hannaneh@washington.edu
Homepage: https://homes.cs.washington.edu/~hannaneh/

**Colin Lockard** is a PhD student at the Paul G. Allen School of Computer Science & Engineering at the University of Washington, where he has published papers on knowledge extraction from both unstructured and semi-structured text.
Email: lockardc@cs.washington.edu
Homepage: https://homes.cs.washington.edu/~lockardc/

**Prashant Shiralkar** is an Applied Scientist in the Product Graph team at Amazon. He currently works on knowledge extraction from semi-structured data. Previously, he received a Ph.D. from Indiana University Bloomington where his dissertation work focused on devising computational approaches for fact checking by mining knowledge graphs. His research interests include machine learning, data mining, information extraction and NLP, and Semantic Web technologies.
Email: shiralp@amazon.com
Homepage: https://sites.google.com/site/shiralkarprashant/

**References**


Michael Cafarella, Alon Halevy, Hongrae Lee, Jayant Madhavan, Cong Yu, Daisy Zhe Wang, and Eu-


Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. In *NAACL-HLT*.


Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL/IJCNLP*.


Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. OpenTag: Open attribute value extraction from product profiles. In *KDD*, pages 1049–1058. ACM.
1 Introduction

Commonsense knowledge, such as knowing that “bumping into people annoys them” or “rain makes the road slippery”, helps humans navigate everyday situations seamlessly (Apperly, 2010). Yet, endowing machines with such human-like commonsense reasoning capabilities has remained an elusive goal of artificial intelligence research for decades (Gunning, 2018).

Commonsense knowledge and reasoning have received renewed attention from the natural language processing (NLP) community in recent years, yielding multiple exploratory research directions into automated commonsense understanding. Recent efforts to acquire and represent common knowledge resulted in large knowledge graphs, acquired through extractive methods (Speer et al., 2017) or crowdsourcing (Sap et al., 2019a). Simultaneously, a large body of work in integrating reasoning capabilities into downstream tasks has emerged, allowing the development of smarter dialogue (Zhou et al., 2018) and question answering agents (Xiong et al., 2019).

Recent advances in large pretrained language models (e.g., Devlin et al., 2019; Liu et al., 2019b), however, have pushed machines closer to human-like understanding capabilities, calling into question whether machines should directly model commonsense through symbolic integrations. Despite these impressive performance improvements in a variety of NLP tasks, it remains unclear whether these models are performing complex reasoning, or if they are merely learning complex surface correlation patterns (Davis and Marcus, 2015; Marcus, 2018). This difficulty in measuring the progress in commonsense reasoning using downstream tasks has yielded increased efforts at developing robust benchmarks for directly measuring commonsense capabilities in multiple settings, such as social interactions (Sap et al., 2019b; Rashkin et al., 2018a) and physical situations (Zellers et al., 2019; Talmor et al., 2019).

We hope that in the future, machines develop the kind of intelligence required to, for example, properly assist humans in everyday situations (e.g., a chatbot that anticipates the needs of an elderly person; Pollack, 2005). Current methods, however, are still not powerful or robust enough to be deployed in open-domain production settings, despite the clear improvements provided by large-scale pretrained language models. This shortcoming is partially due to inadequacy in acquiring, understanding and reasoning about commonsense knowledge, topics which remain understudied by the larger NLP, AI, and Vision communities relative to its importance in building AI agents. We organize this tutorial to provide researchers with information about the critical foundations and recent advances in commonsense, in the hopes of casting a brighter light on this promising area of future research.

In our tutorial, we will (1) outline the various types of commonsense (e.g., physical, social), and (2) discuss techniques to gather and represent commonsense knowledge, while highlighting the challenges specific to this type of knowledge (e.g., reporting bias). We will also (3) discuss the types of commonsense knowledge captured by modern NLP systems (e.g., large pretrained language models), (4) review ways to incorporate commonsense knowledge into downstream task models, and (5) present various benchmarks used to measure systems’ commonsense reasoning abilities.

2 Description

What is commonsense? The tutorial will start with a brief overview of what commonsense is, how it is defined in the literature, and how hu-
mans acquire it (Moore, 2013; Baron-Cohen et al., 1985). We will discuss notions of social commonsense (Burke, 1969; Goldman, 2015) and physical commonsense (Hayes, 1978; McRae et al., 2005). We will cover the differences between taxonomic and inferential knowledge (Davis and Marcus, 2015; Pearl and Mackenzie, 2018), and differentiate commonsense knowledge from related concepts (e.g., script learning; Schank and Abelson, 1975; Chambers and Jurafsky, 2008).

**How to represent commonsense?** We will review existing methods for representing commonsense, most of which focus solely on English. At first, symbolic logic approaches were the main representation type (Forbus, 1989; Lenat, 1995). While still in use today (Davis, 2017; Gordon and Hobbs, 2017), computational advances have allowed for more data-driven knowledge collection and representation (e.g., automatic extraction; Etzioni et al., 2008; Zhang et al., 2016; Elazar et al., 2019). We will cover recent approaches that use natural language to represent commonsense (Speer et al., 2017; Sap et al., 2019a), and while noting the challenges that come with using data-driven methods (Gordon and Van Durme, 2013; Jastrzebski et al., 2018).

**What do machines know?** Pretrained language models (LMs) have recently been described as “rediscovering the NLP pipeline” (Tenney et al., 2019a), i.e. replacing previous dedicated components of the traditional NLP pipeline, starting from low- and mid-level syntactic and semantic tasks (POS tagging, parsing, verb agreement, e.g., Peters et al., 2018; Jawahar et al., 2019; Schwartz and Dagan, 2019, *inter alia*), to high-level semantic tasks such as named entity recognition, coreference resolution and semantic role labeling (Tenney et al., 2019b; Liu et al., 2019a). We will discuss recent investigations into pretrained LMs’ ability to capture world knowledge (Petroni et al., 2019; Logan et al., 2019) and learn or reason about commonsense (Feldman et al., 2019).

**How to incorporate commonsense knowledge into downstream models?** Given that large number of NLP applications are designed to require commonsense reasoning, we will review efforts to integrate such knowledge into NLP tasks. Various works have looked at directly encoding commonsense knowledge from structured KBs as additional inputs to a neural network in generation (Guan et al., 2018), dialogue (Zhou et al., 2018), QA (Mihaylov and Frank, 2018; Bauer et al., 2018; Lin et al., 2019; Weissenborn et al., 2017; Musa et al., 2019), and classification (Chen et al., 2018; Paul and Frank, 2019; Wang et al., 2018) tasks. For applications without available structured knowledge bases, researchers have relied on commonsense aggregated from corpus statistics pulled from unstructured text (Tandon et al., 2018; Lin et al., 2017; Li et al., 2018; Banerjee et al., 2019). More recently, rather than providing relevant commonsense as an additional input to neural networks, researchers have looked into indirectly encoding commonsense knowledge into the parameters of neural networks through pretraining on commonsense knowledge bases (Zhong et al., 2018) or explanations (Rajani et al., 2019), or by using multi-task objectives with commonsense relation prediction (Xia et al., 2019).

**How to measure machines’ ability of commonsense reasoning?** We will explain that, despite their design, many natural language understanding (NLU) tasks hardly require machines to reason about commonsense (Lo Bue and Yates, 2011; Schwartz et al., 2017). This prompted efforts in creating benchmarks carefully designed to be impossible to solve without commonsense knowledge (Roemmele et al., 2011; Levesque, 2011).

In response, recent work has focused on using crowdsourcing and automatic filtering to design large-scale benchmarks while maintaining negative examples that are adversarial to machines (Zellers et al., 2018). We will review recent benchmarks that have emerged to assess whether machines have acquired physical (e.g., Talmor et al., 2019; Zellers et al., 2019), social (e.g., Sap et al., 2019b), or temporal commonsense reasoning capabilities (e.g., Zhou et al., 2019), as well as benchmarks that combine commonsense abilities with other tasks (e.g., reading comprehension; Ostermann et al., 2018; Zhang et al., 2018; Huang et al., 2019).

3 Outline

3.1 Schedule

**Talk 1 (15 min.)** will introduce and motivate this tutorial and discuss long term vision for NLP commonsense research.

**Talk 2 (20 min.)** will focus on the question “Do pre-trained language models capture com-
monsense knowledge?” and review recent work that studies what such models already capture due to their pre-training, what they can be fine-tuned to capture, and what types of knowledge are not captured.

**Talk 3 (20 min.)** will discuss ways of defining and representing commonsense, covering established symbolic methods and recent efforts for natural language representations.

**Talk 4 (20 min.)** will discuss neural and symbolic models of commonsense reasoning, focusing on models based on external knowledge integration for downstream tasks.

If time permits, we will end the first half with an interactive session and a preview to the second half.

**Break (30 min.)**

**Talk 5 (20 min.)** will continue the discussion on neural and symbolic models of commonsense knowledge representation, focusing on COMET (Bosselut et al., 2019), a language model trained on commonsense knowledge graphs. We will present its utility in a zero-shot model for a downstream commonsense question answering task.

**Talk 6 (25 min.)** will focus on temporal commonsense: how to represent it, how to incorporate it into downstream models, and how to test it.

**Talk 7 (20 min.)** will discuss ways to assess machine commonsense abilities, and challenges in developing benchmarks for such evaluations.

**Concluding discussion (10 min.)** will summarize the remaining challenges of commonsense research, and wrap up the tutorial.

### 3.2 Breadth

Due to the research interests and output of the presenters, we estimate that approximately 30% of the tutorial will center around work done by the presenters (Rashkin et al., 2018b; Sap et al., 2019a; Bosselut et al., 2019; Rashkin et al., 2018a; Sap et al., 2019b; Zellers et al., 2018, 2019; Sakaguchi et al., 2019; Bosselut and Choi, 2019; Shwartz et al., 2020).

### 4 Prerequisites

We will not expect attendees to be familiar with previous research on commonsense knowledge and reasoning, but participants should be familiar with:

- Knowledge of machine learning and deep learning – recent neural network architectures (e.g., RNN, CNN, Transformers), as well as large pre-trained language models models (e.g., BERT, GPT, GPT2).
- Familiarity with natural language processing tasks – understanding the basic problem to solve in tasks such as question answering (QA), natural language generation (NLG), textual entailment/natural language inference (NLI), etc.

### 5 Reading List

- Storks et al. (2019) – a survey on commonsense
- Speer et al. (2017) – A description of a prototypical commonsense knowledge base, its structure, and its curation
- Gordon and Van Durme (2013) – Overview of issues surrounding reporting bias, making automatic commonsense acquisition difficult
- Mostafazadeh et al. (2016) – A dataset that appears often in recent commonsense research
- Talmor et al. (2019) – One approach for leveraging crowdsourcing to construct a commonsense evaluation benchmark

### 6 Instructor information

**Maarten Sap** is a PhD student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. His research focuses primarily on social applications of NLP, specifically on endowing machines with social intelligence, social commonsense, or theory of mind.

**Vered Shwartz** is a postdoctoral researcher at the Allen Institute for Artificial Intelligence (AI2) and the Paul G. Allen School of Computer Science & Engineering at the University of Washington, working on lexical semantics, multiword expressions, and commonsense reasoning. She co-organized the ACL 2018 Student Research Workshop, the SemEval 2018 shared task on hypernymy discovery, and the AAAI 2020 Workshop on Reasoning for Complex Question Answering, Special Edition on Commonsense Reasoning.
Antoine Bosselut is a PhD student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington and a student researcher at the Allen Institute for Artificial Intelligence (AI2). His research interests are in integrating commonsense knowledge and reasoning into downstream applications for text generation, summarization, and conversational dialogue. He organized the West Coast NLP (WeCNLP) in 2018 and 2019 and the NeuralGen workshop at NAACL 2019.

Yejin Choi is an associate professor at the Paul G. Allen School of Computer Science & Engineering at the University of Washington and also a senior research manager at AI2 overseeing the project Mosaic. Her research interests include language grounding with vision, physical and social commonsense knowledge, language generation with long-term coherence, conversational AI, and AI for social good. She was a recipient of Borg Early Career Award (BECA) in 2018, among the IEEEs AI Top 10 to Watch in 2015, a co-recipient of the Marr Prize at ICCV 2013, and a faculty advisor for the Sounding Board team that won the inaugural Alexa Prize Challenge in 2017. She was on the steering committee of the NeuralGen workshop at NAACL 2019.

Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of Computer and Information Science, University of Pennsylvania, and a Fellow of the AAAS, the ACM, AAAI, and the ACL. In 2017 Roth was awarded the John McCarthy Award, the highest award the AI community gives to mid-career AI researchers. He was the Editor-in-Chief of the Journal of Artificial Intelligence Research (JAIR) and a program co-chair of AAAI, ACL and CoNLL. Dan has presented several tutorials in conferences including at ACL, on entity linking, temporal reasoning, transferable representation learning, and more.

References


Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. In IJCAI.

Jian Guan, Yansen Wang, and Minlie Huang. 2018. Story ending generation with incremental encoding and commonsense knowledge. In AAAI.


Qian Li, Ziwei Li, Jin-Mao Wei, Yanhui Gu, Adam Jatowt, and Zhenglu Yang. 2018. A multi-attention based neural network with external knowledge for story ending predicting task. In COLING.


Peter Lo Bue and Alexander Yates. 2011. Types of Common-Sense knowledge needed for recognizing textual entailment. In ACL.


Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Mark Johnson, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In HLT-NAACL.


Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In IJCAI.
1 Description

Open-domain question answering (QA), the task of answering questions using a large collection of documents of diversified topics, has been a longstanding problem in NLP, information retrieval (IR) and related fields (Voorhees et al., 1999; Moldovan et al., 2000; Brill et al., 2002; Ferrucci et al., 2010). Traditional QA systems were usually constructed as a pipeline, consisting of many different components such as question processing, document/passage retrieval, and answer processing. With the rapid development of neural reading comprehension (Chen, 2018), modern open-domain QA systems have been restructured by combining traditional IR techniques and neural reading comprehension models (Chen et al., 2017; Yang et al., 2019; Min et al., 2019a) or even implemented in a fully end-to-end fashion (Lee et al., 2019; Seo et al., 2019; Guu et al., 2020; Roberts et al., 2020). In this tutorial, we aim to provide a comprehensive and coherent overview of cutting-edge research in this direction. ¹

We will start by first giving a brief background of open-domain question answering, discussing the basic setup and core technical challenges of the research problem. We aim to give the audience a historical view of how the field has advanced in the past several decades, from highly-modulated pipeline systems in the early days, to modern end-to-end training of deep neural networks in the present.

We will then discuss modern datasets proposed for open-domain QA (Voorhees et al., 1999; Berant et al., 2013; Rajpurkar et al., 2016; Joshi et al., 2017; Dhingra et al., 2017; Dunn et al., 2017; Kwiatkowski et al., 2019), as well as common evaluation metrics and benchmarks. We plan to provide a detailed discussion on available datasets — their collection methodology and properties — as well as insights on how these datasets should be viewed in the context of open-domain QA.

Next, the focus will shift to cutting-edge models proposed for open-domain QA, which is also the central part of this tutorial. We divide existing models into three main categories: Two-stage retriever-reader approaches, Dense retriever and end-to-end training, and Retriever-free approaches. We will present the logical elements behind different sorts of models and discuss their pros and cons.

Two-stage retriever-reader approaches. We will start by discussing two-stage retriever-reader frameworks for open-domain QA, pioneered by Chen et al. (2017): a retriever component finding documents that (might) contain an answer from a large collection of documents, followed by a reader component finding the answer in a given paragraph or a document. In this category, the retriever component is usually implemented by traditional sparse vector space methods, such as TF-IDF or BM25 and the reader is implemented by neural reading comprehension models. We will further discuss several challenges and techniques arising in this area, including multi-passage training (Clark and Gardner, 2018; Wang et al., 2019), passage reranking (Wang et al., 2018; Nogueira and Cho, 2019), and denoising distantly-supervised data (Lin et al., 2018).

Dense retriever and end-to-end training. The first category mainly employs a non-machine learning model for the retrieval stage. The second category will focus on how to learn the retriever component by replacing traditional IR methods with dense representations, as well as joint training of both components. Learning and searching in dense vector space is challenging, as it usually involves...
an enormous search space (easily ranging from millions to billions of documents). We will discuss in depth how this was achieved by existing models, including novel pre-training methods (Lee et al., 2019; Guu et al., 2020), carefully-designed learning algorithms (Karpukhin et al., 2020) or a hybrid approach using both dense and sparse representations (Seo et al., 2019).

Retriever-free approaches. The third category, which is a recent emerging trend, only relies on large-scale pre-trained models (Radford et al., 2018; Devlin et al., 2018; Liu et al., 2019) as implicit knowledge bases and doesn’t require access to text data during inference time. These pre-trained models will be used directly to answer questions, in a zero-shot manner (Radford et al., 2019; Raffel et al., 2019) or fine-tuned using question-answer pairs. As these methods don’t need a retriever component, we call them Retriever-free approaches.

Up to this point, our tutorial has mainly focused on textual question answering. At the end, we also plan to discuss some hybrid approaches for answering open-domain questions using both text and large knowledge bases, such as Freebase (Bollacker et al., 2008) and Wikidata (Vrandečić and Krötzsch, 2014), and give a critical review on how structured data complements the information from unstructured text. The approaches include (1) how to leverage structured data to guide the retriever or reader stage of existing textual QA systems (Asai et al., 2020; Min et al., 2019b), or (2) how to synthesize information from these two heterogeneous sources and build effective QA models on the combined information (Sun et al., 2018, 2019; Xiong et al., 2019).

Finally, we will discuss some important questions, including (1) How much progress have we made compared to the QA systems developed in the last decade? (2) What are the main challenges and limitations of current approaches? (3) How to trade off the efficiency (computational time and memory requirements) and accuracy in the deep learning era? We hope our tutorial will not only serve as a useful resource for the audience to efficiently acquire up-to-date knowledge, but also provide new perspectives to stimulate the advances of open-domain QA research in the next phase.

Prerequisites The tutorial will be accessible to anyone who has the basic knowledge of machine learning and natural language processing. The tutorial will target both NLP researchers/students in academia and NLP practitioners in industry.

2 Tutorial Outline

The intended duration of this tutorial is 3.5 hours, including a half an hour break.

1. Introduction
2. Problem definition & motivation
3. A history of open-domain (textual) QA
   (a) Early QA systems
   (b) TREC QA competitions
   (c) IBM’s DeepQA project
   (d) More recent developments: 2017-2020
4. Datasets & evaluation
   (a) Reading comprehension vs QA datasets
   (b) Categorization of QA datasets
   (c) Evaluation metrics
5. Two-stage retriever-reader approaches
   (a) General framework
   (b) Multi-passage training
   (c) Passage reranking
   (d) Denoising distantly supervised data
6. Dense retriever and end-to-end training
   (a) Dense passage retrieval
   (b) Joint training of retriever and reader
   (c) Dense-sparse phrase indexing
7. Retriever-free approaches
8. Open-domain QA using KBs and text
   (a) Improving retriever and reader using structured KBs
   (b) Answering questions over combined KBs and text
9. Open problems and future directions

3 Presenters

Danqi Chen Danqi Chen is an Assistant Professor of Computer Science at Princeton University and co-directs the Princeton NLP Group. Danqi’s research interests lie within deep learning for natural language processing, with an emphasis on the intersection between text understanding and knowledge representation/reasoning and applications. 
such as question answering and information extraction. Before joining Princeton University, Danqi worked as a visiting scientist at Facebook AI Research (FAIR). She received her PhD from Stanford University (advised by Christopher Manning) in 2018 and B.Eng from Tsinghua University in 2012. Website: https://www.cs.princeton.edu/~danqic/.

Scott Wen-tau Yih  
Scott Wen-tau Yih is a Research Scientist at Facebook AI Research (FAIR), and his recent research focuses on continuous representations and neural network models, with applications in knowledge base embedding, semantic parsing and question answering. Yih received the best paper award from CoNLL’11, an outstanding paper award from ACL’15 and has served as an area co-chair and a program co-chair for several top conferences. He is also a co-presenter for several popular tutorials on topics including Semantic Role Labeling, Deep Learning for NLP, Question Answering with Knowledge Base, Web and Beyond and NLP for Precision Medicine. Website: http://scottyih.org/.

References


36


Author Index

Alikhani, Malihe, 10
Belinkov, Yonatan, 1
Bender, Emily M., 6
Bosselut, Antoine, 27
Chen, Danqi, 34
Choi, Yejin, 27
Cohen, Kevin, 16
Dong, Xin Luna, 23
Fort, Karën, 16
Gehrmann, Sebastian, 1
Hajishirzi, Hannaneh, 23
Hovy, Dirk, 6
Lockard, Colin, 23
Mieskes, Margot, 16
Mou, Lili, 19
Névéol, Aurélie, 16
Pavlick, Ellie, 1
Roth, Dan, 27
Sap, Maarten, 27
Schofield, Alexandra, 6
Shiralkar, Prashant, 23
Shwartz, Vered, 27
Stone, Matthew, 10
Vechtomova, Olga, 19
Yih, Wen-tau, 34