NLP for Precision Medicine

Hoifung Poon[1] **Chris Quirk**[1] **Kristina Toutanova**[2] **Wen-tau Yih**[1]
[1] Microsoft Research, Redmond, WA, USA
[2] Google Research, Seattle, WA, USA
{hoifung,chrisq,scottyih}@microsoft.com
kristout@google.com

1 Tutorial Overview

We will introduce precision medicine and showcase the vast opportunities for NLP in this burgeoning field with great societal impact. We will review pressing NLP problems, state-of-the art methods, and important applications, as well as datasets, medical resources, and practical issues. The tutorial will provide an accessible overview of biomedicine, and does not presume knowledge in biology or healthcare. The ultimate goal is to reduce the entry barrier for NLP researchers to contribute to this exciting domain. Our motivation stems from the shocking inefficiency of medicine today. For the top 20 prescription drugs in the US, 80% of patients are non-responders. The result is ineffective care delivery, which leads to missed opportunities for treatment and constitutes a large part of the estimated trillion-billion waste in the US health system each year. Recent technological disruptions such as $1000 human genome have enabled more personalized and effective treatments, with great potential to improve patient health and save lives.

A major bottleneck to advancing precision medicine is access to structured information encoded in free text. In cancer, for example, it takes hours for a molecular tumor board of many specialists to review one patient's genomics data and make treatment decisions. With 1.7 million new cancer patients in the US alone each year, this is clearly not scalable. Most relevant knowledge resides in published literature, whereas rich patient information is scattered in clinical notes in electronic medical records (EMRs). NLP holds the key to unlock such structured information for supporting predictive analytics and medical decision making. Compared to the newswire and web domains, healthcare also exhibits important differences and offers a fertile ground of novel research challenges.

In this tutorial, we will first present an overview of precision medicine, and highlight key research challenges and opportunities for NLP. We will then dive into main research areas and review problem formulations and cutting-edge methods. To illustrate the potential impact of NLP, we will present several real-world applications with promising results. To facilitate new entry to the filed, we will provide a systematic review of relevant resources and conclude with a list of exciting open problems.

2 Outline

1. Precision Medicine (20 minutes)
   - Motivation: imprecise medicine, disruptions, what successes look like
   - Challenges: knowledge, reasoning
   - Opportunities for NLP
2. Annotation Bottleneck (25 minutes)
   - Example tasks: entity linking, relation extraction
   - Distant supervision
   - Learning with indirect supervision
3. Extract complex structured information (25 minutes)
   - Example task: complex event extraction
   - Grounded semantic parsing
4. Beyond the sentence boundary (25 minutes)
   - Motivation: knowledge extraction for molecular tumor board
   - Cross-sentence relation extraction
   - Graph LSTM
5. Reasoning (25 minutes)
   - Standard approaches and challenges

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-Tutorial Abstracts, pages 1–2
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-5001

- Neural embeddings of structured knowledge
- Example application: Knowledge base completion

6. Applications in Precision Medicine (30 minutes)
   - Decision support for molecular tumor board
   - Rational design of cancer drug combinations
   - Clinical machine reading
7. Resources (20 minutes)
   - Text, ontologies, and knowledge bases
   - Shared tasks
   - Practical issues: publishers, privacy, regulations
8. Open Problems (10 minutes)

3 Instructors

Hoifung Poon is a Researcher at Microsoft Research Redmond. His research interests lie in advancing machine learning and natural language processing (NLP) to help automate discovery and decision support in precision medicine. He received his Ph.D. in computer science & engineering at the University of Washington. His past work has been recognized with Best Paper Awards from premier NLP and machine learning venues such as NAACL-09 (unsupervised morphological segmentation), EMNLP-09 (unsupervised semantic parsing), and UAI-11 (sum-product networks).

Chris Quirk is a Principal Researcher at Microsoft Research Redmond. Since joining Microsoft Research in 2001, his research has focused on effective computational systems for aiding human communication, understanding, and task completion. His primary focus is in machine translation, building practical and widely-used system implementations and authoring a number of influential papers. He has also worked in paraphrase, information extraction, and most recently biological applications of natural language processing and machine learning. He has served on numerous program committees, acted Area Chair (ACL 2010, EMNLP 2012), and is currently an action editor of the TACL journal.

Kristina Toutanova is a Staff Research Scientist at Google Research Seattle and affiliate faculty member at the University of Washington. In 2005, she obtained her Ph.D. from the Computer Science Department at Stanford University, where she was advised by Christopher Manning. She focuses on modeling the structure of natural language using machine learning, in the areas of semantic parsing, knowledge extraction, information retrieval, and text-to-text generation. She has co-authored more than 50 publications at refereed conferences and journals, including four papers that have won awards at conferences (EMNLP, NAACL, CoNLL, ECML). She served as a program co-chair for CoNLL 2008 and ACL 2014 and is currently serving as a co-editor-in-chief of the TACL journal.

Wen-tau Yih is a Senior Researcher at Microsoft Research Redmond. His research interests include natural language processing, machine learning and information retrieval. Yih received his Ph.D. in computer science at the University of Illinois at Urbana-Champaign. His work on joint inference using integer linear programming (ILP) helped the UIUC team win the CoNLL-05 shared task on semantic role labeling, and the approach has been widely adopted in the NLP community since then. After joining MSR in 2005, he has worked on email spam filtering, keyword extraction and search & ad relevance. His recent work focuses on continuous semantic representations using neural networks and matrix/tensor decomposition methods, with applications in lexical semantics, knowledge base embedding and question answering. Yih received the best paper award from CoNLL-2011, an outstanding paper award from ACL-2015 and has served as area chairs (HLT-NAACL-12, ACL-14, EMNLP-16,17), program co-chairs (CEAS-09, CoNLL-14) and action/associated editors (TACL, JAIR) in recent years.