Advanced NLP
Lecture 1: Introduction + Intrinsic Tasks

Dr. Gabriel Stanovsky

March 11, 2023

Suggested reading: Speech and Language Processing: an Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. Daniel Jurafsky & James H Martin

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM
الجامعة العبرية في اورشليم القدس

Target Audience

- **Background in machine learning**
  - That's all we'll do
- **Background in NLP to varying degrees**
  - From students who just did first NLP course to PhD students
- **Want to learn more about current tech & research trends**

Course Agenda

- We aim to bring you up to speed with **latest NLP innovations**
- **NLP is fast moving field**
  - Likely to change as we're giving the course
- We built the syllabus around **long-standing challenges and themes**
- **Ask questions, challenge assumptions**
  - Mid-size, diverse course to prompt discussion
  - Some (many?) questions which we don't know the answer to
  - We'll be happy for feedback

This isn't a seminar

- Lectures are **self-contained**
- But we provide a bibliography for additional reading
- and you'll benefit from reading into what you find interesting
- Lots of room for **creativity**

Today

1. Course Structure
2. Pretraining, Intrinsic, and Extrinsic Tasks
3. Intrinsic NLP Applications
   - Textual Entailment (or NLI)
   - Coreference
   - Entity Linking
4. Discussion
5. Conclusion

In previous chapters you learned . . .

- **Q: How do we define NLP?**
  - Models taking natural language as part of their input or output
  - **Q: Can you give example of tasks?**
- **Automatically understanding language is hard**
  - **Q: Why?**
- **Linguistic background**
  - Lexical (POS), syntactic (dep trees), semantic (SRL)
- **Machine learning is ubiquitous**
- **Word embedding is a powerful technique**
  - As word features (e.g., Word2vec)
  - With finetuning (e.g., ELMo, BERT)

Where we're going

- **Better understanding of our tasks & data**
- **Finetuning and zero-shot**
- **Interpreting model performance**
- **Efficient models**
- **Real-world tasks**

Disclaimer: We're biased towards our research topics

- We don't aim to give an exhaustive overview of NLP
- **Many courses from other researchers**
  - Self-supervised Statistical Models (Daniel Khashabi, JHU)
  - Local Explanations for Deep Learning Models (Ana Marasovic, Utah)
  - Exploration on Language, Knowledge, and Reasoning (Yejin Choi, UW)
  - Computational Ethics in NLP (Emma Strubell, Maarten Sap, CMU)
- **Interesting to contrast & compare**

Course Objectives

- Familiarize with **topics at the forefront** of NLP today
- Exprience phrasing a **research question**
- **Hands-on exprience** with state-of-the-art NLP
- Read relevant **literature**
- Present your work in **scientific writing**

Course Requirements (w/o the gritty details)

- Two relatively small **coding & evaluation excercises**
- **An Open-Ended Research Project**
  - **You formulate your idea**
  - Define goals and intended outcomes
  - Describe your work in a final report
  - Work in groups

Start thinking about your project today

- Today we'll talk about **longstanding NLP tasks**
- **We won't discuss modelling**
- Focus on understanding **importance & challenges**

Pretraining vs. Intrinsic vs. Extrinsic

- **Extrinsic tasks** (aka *downstream*)
  - Tasks which have applicable value for external users
  - Machine translation, information extraction, summarization...
- **Intrinsic tasks** (aka *intermediate*)
  - Inherently required across extrinsic tasks
  - But are not directly useful on their own
  - Often correspond to much-studied linguistic phenomena
  - You've seen: SRL, grammar (dependency trees)
- **Pretraining tasks**
  - Do not fall neatly into any of the above
  - But we have order of magnitudes more data for them
  - and they transfer well to other tasks

Synthetic vs. Real-World Data

- **Synthetic data** is constructed specifically for training the model
  - E.g., asking humans to write text according to guidelines
- **Real-world data** is written independently from model development
  - E.g., news outlets, books, or financial reports
- **Orthogonal to the type of task**
- We'll come back to this later in the course

Task vs. Format

- **Task:** The human skill required by the model
  - E.g., translation, commonsense, information extraction
- **Format:** How it's encoded or collected
  - E.g., sequence labeling, seq2seq, span selection
- **Q: Is QA a task or a format?** [1]
  - We'll see how it's used to collect data for different tasks

Recognizing Textual Entailment (Or NLI)

The task of deciding whether the meaning of one text (the **Hypothesis***) is entailed, or can be inferred, from another text (the* **Premise***)*[2]

- Typically consisting of **three labels**
  - **Premise:** "Yoko Ono unveiled a bronze statue for her late husband, John Lennon."
- **Entailment**
  "Yoko Ono is John Lennon's widow"
- **Contradiction**
  "John Lennon is Yoko Ono's widow"
- **Neutral**
  "John Lennon and Yoko Ono married in 1969"

Why is NLI Important?

- Traditionally considered a **facet of many NLP tasks**
- Consider QA model answering the question
  Who is John Lennon's widow?
- Would require understanding it is **entailed** from hypothesis above

Why is NLI challenging?

- Requires a combination of **world knowledge and common sense**
  - **Q: How did we infer the contradiction above?**
- Language if often **ambiguous and evades logic operators**
  - Reconsider:
    "Yoko Ono unveiled a bronze statue for her late husband, John Lennon."
- **Q: Maybe John was late to the event so Yoko unveiled for him?**
  - Thus changing many of the labels we gave before
- NLI (and NLP in general) goes for the **most reasonable reading**
  - Compare with:
    "Yoko Ono ordered a sandwich for her late husband, John Lennon."

NLI Datasets

- **Recognizing Textual Entailment** (RTE) [3]
  - Yearly challenges
  - 1600 annotated pairs in each
  - Directly coupled with a downstream application
- **Stanford Natural Language Inference** (SNLI) [4]
  - 500K training pairs, 10K for test
  - Annotators write hypotheses on image caption premises
- **Multi-Genre NLI** (MNLI) [5]
  - 433K pairs from multiple generes (chat, literature, ...)
  - Collected similarly to SNLI
- **We'll discuss SNLI & MNLI in future lectures**

Grounding

- The next tasks we'll discuss today revolve around **Grounding**
- Mapping from **text** (or form) to a world (**ontology, or meaning**)
  - Form: *"Quick call John!"*
  - Grounding: Identify the correct John, find his number, call, etc.
- In fierce debate around LLMs
  - **Q: How is grounding and meaning defined?**
  - **Q: Are LLMs exposed only to form?**
  - **Q: If so, can they still learn meaning?** [6]

Coreference: Task Definition

An important component of language processing is knowing who is being talked about in a text.[7] [Chap. 26]

Victoria Chen, CFO of Megabucks Banking, saw *her* pay jump to $2.3 million, as *the 38-year-old* became the company's president. It is widely known that *she* came to Megabucks from rival Lotsabucks.

- **mentions** (or *coreferring expression*)
  Refer to the same entity in an extra-textual world
- **evoking mention** (or *antecedent*)
  The first mention in which the entity is identified
- *anaphoras*
  Other mentions which accesses an entity evoked elsewhere
- Other variants include **event coreference**

Coreference: Task Definition [7] (Chap. 26)

An important component of language processing is knowing who is being talked about in a text.

Victoria Chen, CFO of Megabucks Banking, saw *her* pay jump to $2.3 million, as *the 38-year-old* became the company's president. It is widely known that *she* came to Megabucks from rival Lotsabucks.

- **Output:** Coreference chains (or *clusters*)
  {Victoria Chen, her, the 38-year-old, She}
  {Megabucks Banking, the company, Megabucks}
  {her pay}
  {Lotsabucks}
- The last two are termed **singletons**
- **Note that coreference doesn't link to an ontology** (only form)
  - But assumes an ontology exists?

Coreference Tasks

- **Mention detection**:
  Identifying spans of texts referring to external entities
- **Finding coreference links**:
  Forming coreference clusters from mentions

Why is coreference challenging?

- Often requires long-range dependencies
- How are *external entities* defined?
- Often ambiguous
  - The trophy didn't fit in the suitcase because **it** was too **big**
  - The trophy didn't fit in the suitcase because **it** was too **small**
- This format is known as the **Winograd Schema** [8]

Evaluating coreference is hard

- **Requires comparing & aligning two groups of clusters**
- **Many types of errors**
  - Missing entities, missing mentions, different spans
- Compare gold:
  {Victoria Chen; her; *the* 38-year-old; She}
  {<u>Megabucks Banking</u>; *the* company; Megabucks}
- ... with predicted:
  {Victoria Chen, *CFO of Megabucks Banking*; her; 38-year-old; She}
  {company; Megabucks}

Coreference Evaluation Metrics

- **Many metrics measuring different aspects of coreference**
  - Most popular are MUC, B^3, CEAF
  - Common practice is to report their average
- **MUC: Definition**
  - Let R - reference clusters, H - predicted hypothesis
  - MUC precision: $\frac{\#\text{common links}}{\#\text{links in } H}$
  - MUC recall: $\frac{\#\text{common links}}{\#\text{links in } R}$
  - A *link* is any (unordered) pair of mentions in the same cluster
- **Q: What does MUC miss?**
  - No reward for slight errors in spans
  - Doesn't reward (or punish) singletons

Coreference Datasets

- **Ontonotes** [9]
  - About 1M words in English and 1M words in Chinese
  - Newswire, web data and conversational speech
- **Quoref** [10]
  - Annotates coreference through QA
  - 24K questions over Wikipedia

Byzantines were avid players of tavli (Byzantine Greek: τάβλη), a game known in English as backgammon, which is still popular in former Byzantine realms, and still known by the name tavli in Greece. Byzantine nobles were devoted to horsemanship, particularly ***tzykanion***, now known as *polo*. The game came from Sassanid Persia in the early period and a Tzykanisterion (stadium for playing *the game*) was built by Theodosius II (r. 408–450) inside the Great Palace of Constantinople. Emperor Basil I (r. 867–886) excelled at **it**; Emperor Alexander (r. 912–913) died from exhaustion while playing, Emperor Alexios I Komnenos (r. 1081–1118) was injured while playing with Tatikios, and John I of Trebizond (r. 1235–1238) died from a fatal injury during a game. Aside from Constantinople and Trebizond, other Byzantine cities also featured tzykanisteria, most notably Sparta, Ephesus, and Athens, an indication of a thriving urban aristocracy.

Q1. What is the Byzantine name of the game that Emperor Basil I excelled at? **it → tzykanion**
Q2. What are the names of the sport that is played in a Tzykanisterion? *the game → tzykanion; polo*
Q3. What cities had tzykanisteria? cities → Constantinople; Trebizond; Sparta; Ephesus; Athens

Coreference: Annotation

- Traditionally requires **highly-experienced annotators**
  - Dozens of pages of annotation guidelines
- QA annotation (e.g., Quoref) **eases annotation difficulty**
  - At the cost of exhaustiveness, inter-annotator agreement

Coreference Variants

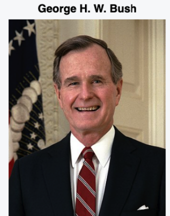
- **Multi-document coreference**
  - Annotates entity mentions across different documents
  - E.g., news reports of the same events
- **Event coreference**
  - Annotates mentions of events rather than entities
  - E.g., *The Great War*; *World War I*; *WWI*

Entity Linking: Task Definition

Entity linking is the task of associating a mention in text with the representation of some real-world entity in an ontology.[11, 7] [Chap. 14.3]

George Bush reveals how he repeatedly turned to **his father** for advice as he contemplated following him into war against Saddam Hussein.

- Requires a representation of an extra-textual world (often Wikipedia)
  - Assigns a Wikipedia page to each entity mention
  - This is sometimes called **Wikification** [12]



Why is entity linking challenging?

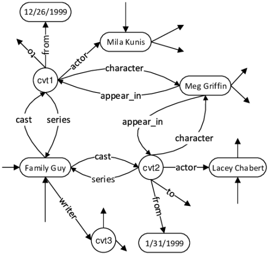
- A mention span may be ambiguous with respect to the ontology
  - Consider the George Bush example
- Requires world knowledge corresponding to the reference ontology

Other Ontologies

- **Medical**
    - E.g., SNOMED [13], NCBI
    - Ontology of e.g., drugs, symptoms, adverse reactions
- **Knowledge graphs**
    - E.g., Freebase [14], Wikidata [15], YAGO [16]
    - Represent entities in the world as well as events and relations

Other Ontologies

- **Medical**
  - E.g., SNOMED [13], NCBI
  - Ontology of e.g., drugs, symptoms, adverse reactions
- **Knowledge graphs**
  - E.g., Freebase [14], Wikidata [15], YAGO [16]
  - Represent entities in the world as well as events and relations



Entity Linking: Datasets & Annotation

- **Entity linking as Question Answering** [17]
- **AIDA CoNLL-YAGO** [18]
  - News texts mapped to YAGO & DBPedia
- **CADEC** [19]
  - Blog posts mapped to a medical ontology (SNOMED)

Grounding: Connecting Coreference and Entity Linking

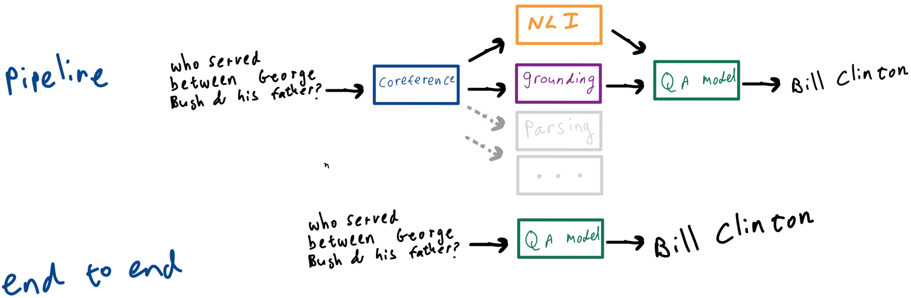
- **Coreference + entity linking** assigns an entity per cluster
- **Facilitates agents interacting in the world**

Grounding: Connecting Coreference and Entity Linking

- **Coreference + entity linking** assigns an entity per cluster
- **Facilitates agents interacting in the world**
- {Victoria Song, her, the 38-year-old, She}

Pipeline vs. End-to-End Approaches

- **E2E models do not require intermediate task labels**
- **NLP has shifted almost entirely to E2E approaches**
  - Trained on input and outputs w/o intermediate labels



How does E2E learning intrinsic tasks?

- They observe intrinsic phenomena from **downstream examples**
  - E.g., coreference in enough relevant parallel data
- The data is rich enough $\Rightarrow$ they'll learn the required intrinsic tasks
- This is often tested with **probing**
  - More on this later in the course

- **Q: Do large language models observe only form?**
  - An open research question

- **For example, in entity linking, they may observe**
  - George Bush Senior *link-to-wiki-page*

  - **Effectively training LLM with entity linking labels**

So we don't need intrinsic tasks?

- The set of intrinsic tasks is **arbitrary and incomplete**
- **Q: Can we enumerate all subtasks required for e.g. MT?**
- E2E models improve **without additional supervision**

Then why should we still study intrinsic tasks?

- **Integrating intrinsic signal into models can alleviate biases**
  - E2E models often sidestep intrinsic tasks with shortcuts
  - E.g., gender bias in coreference resolution [20]
  - More on this later
- **Using intrinsic tasks to evaluate E2E model performance**
  - To understand their boundaries and bottlenecks
- **E2E learning of complex phenomena may not be data efficient**

  $\Rightarrow$ Not applicable for low-resource domains and languages

  $\Rightarrow$ Wasteful in terms of compute

Conclusion

- **Extrinsic tasks are readily useful for end users**
  - E.g., Machine translation, summarization, information extraction
- **Intrinsic tasks are needed for many extrinsic tasks**
  - Aren't useful on their own
- We discussed **grounding**
  - Maps text (form) and extra-textual entities (ontology)
  - E.g., database entries such as Wikipedia
- **QA is an intuitive annotation paradigm**
  - Project idea: extend to other tasks
- Originally motivated by a **pipeline approach**
- Recently for **interpretability & mitigating biases**

Next Week

Extrinsic tasks!

Which you can also think about for inspiration for your project!

References I

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min.
Question answering is a format; when is it useful?
arXiv preprint arXiv:1909.11291, 2019.

Ido Dagan, Oren Glickman, and Bernardo Magnini.
The pascal recognising textual entailment challenge.
In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer, 2006.

References II

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan.
The third pascal recognizing textual entailment challenge.
In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9, 2007.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning.
A large annotated corpus for learning natural language inference.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

References III

Adina Williams, Nikita Nangia, and Samuel Bowman.
A broad-coverage challenge corpus for sentence understanding through inference.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller.
Climbing towards NLU: On meaning, form, and understanding in the age of data.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics.

James H Martin.
Speech and language processing: an introduction to speech recognition, computational linguistics and natural language processing. daniel jurafsky & 4 n-grams.

Hector J Levesque, Ernest Davis, and Leora Morgenstern.
The winograd schema challenge.
KR, 2012:13th, 2012.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue.
CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes.
In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner.
Quoref: A reading comprehension dataset with questions requiring coreferential reasoning.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5925–5932, Hong Kong, China, November 2019. Association for Computational Linguistics.

Heng Ji and Ralph Grishman.
Knowledge base population: Successful approaches and challenges.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Rada Mihalcea and Andras Csomai.
Wikify! linking documents to encyclopedic knowledge.
In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, 2007.

Tim Benson.
Principles of health interoperability HL7 and SNOMED.
Springer Science & Business Media, 2012.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor.
Freebase: a collaboratively created graph database for structuring human knowledge.
In *SIGMOD Conference*, 2008.

Denny Vrandeić and Markus Krötzsch.
Wikidata: a free collaborative knowledgebase.
Commun. ACM, 57:78–85, 2014.

References VII

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum.
Yago: a core of semantic knowledge.
In *The Web Conference*, 2007.

Wenzheng Zhang, Wenyue Hua, and Karl Stratos.
Entqa: Entity linking as question answering, 2021.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum.
Robust disambiguation of named entities in text.
In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang.
Cadec: A corpus of adverse drug event annotations.
Journal of biomedical informatics, 55:73–81, 2015.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme.
Gender bias in coreference resolution.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.