Natural Language Processing for Law and Social Science

12. Global Semantics

Week 11 Group Activity Second Presentation

Methods Review: Text Data in Social Science

- What has been done:
  - **Dictionary pattern matching**: e.g., Baker, Bloom, and Davis (2017), measuring economic policy uncertainty from newspaper text.
  - **Supervised learning to scale a dimension in text**: e.g. Osabruegge, Ash, and Morelli (2021) doing cross-domain topic classification in political text, Widmer, Galletta, and Ash (2022) measuring cable-news slant.

Methods Review: Text Data in Social Science

- What has been done:
  - **Dictionary pattern matching**: e.g., Baker, Bloom, and Davis (2017), measuring economic policy uncertainty from newspaper text.
  - **Supervised learning to scale a dimension in text**: e.g. Osabruegge, Ash, and Morelli (2021) doing cross-domain topic classification in political text, Widmer, Galletta, and Ash (2022) measuring cable-news slant.
  - **Topic models**: e.g., Hansen, McMahon, and Prat (2018), measuring allocation of attention in Central Bank discussion transcripts.
  - **Word embeddings**: e.g. Ash, Chen, and Ornaghi (2021), measuring gender stereotypes in judicial opinions; Gennaro and Ash (2021), measuring emotionality in congressional speeches.

Methods Review: Text Data in Social Science

- What has been done:
  - **Dictionary pattern matching**: e.g., Baker, Bloom, and Davis (2017), measuring economic policy uncertainty from newspaper text.
  - **Supervised learning to scale a dimension in text**: e.g. Osabruegge, Ash, and Morelli (2021) doing cross-domain topic classification in political text, Widmer, Galletta, and Ash (2022) measuring cable-news slant.
  - **Topic models**: e.g., Hansen, McMahon, and Prat (2018), measuring allocation of attention in Central Bank discussion transcripts.
  - **Word embeddings**: e.g. Ash, Chen, and Ornaghi (2021), measuring gender stereotypes in judicial opinions; Gennaro and Ash (2021), measuring emotionality in congressional speeches.
  - **Syntactic parsing**: e.g. Ash et al (2020) extracting modal verb structures ("workers shall have X").
  - **Semantic role labeling**, extracting agents, actions, patients: Ash, Gauthier, and Widmer (2021) on narratives in Congressional speeches.

Methods Review: Text Data in Social Science

- What has been done:
  - **Dictionary pattern matching**: e.g., Baker, Bloom, and Davis (2017), measuring economic policy uncertainty from newspaper text.
  - **Supervised learning to scale a dimension in text**: e.g. Osabruegge, Ash, and Morelli (2021) doing cross-domain topic classification in political text, Widmer, Galletta, and Ash (2022) measuring cable-news slant.
  - **Topic models**: e.g., Hansen, McMahon, and Prat (2018), measuring allocation of attention in Central Bank discussion transcripts.
  - **Word embeddings**: e.g. Ash, Chen, and Ornaghi (2021), measuring gender stereotypes in judicial opinions; Gennaro and Ash (2021), measuring emotionality in congressional speeches.
  - **Syntactic parsing**: e.g. Ash et al (2020) extracting modal verb structures ("workers shall have X").
  - **Semantic role labeling**, extracting agents, actions, patients: Ash, Gauthier, and Widmer (2021) on narratives in Congressional speeches.
- What has not (at least not substantially):
  - local semantics and global semantics

Note on Narratives

- "Narratives" are of major interest in social science, e.g. political economy:

"Higher taxes will hurt the economy."

"Immigrants steal our jobs."

"Murderers deserve the death penalty."

`https://sites.google.com/view/trump-narratives`

- Long-range narrative comprehension and storytelling is a key endpoint of strong computational natural language understanding.
- as these papers show, it is a difficult task
  - also, defining "narrative" is already difficult.
- explains why it is not well-researched compared to some other tasks.

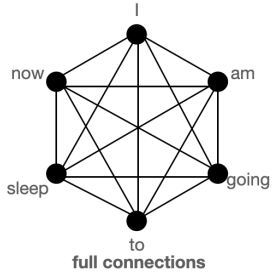
Outline

The Problem with BERT/GPT

- The 2018/2019 generation transformers like BERT have a computational constraint on the length of sequences they can consider (usually limited to $n = 512$).

The Problem with BERT/GPT

- The 2018/2019 generation transformers like BERT have a computational constraint on the length of sequences they can consider (usually limited to $n = 512$).
- BERT's attention heads take as input the embeddings for each pair-wise interaction between tokens.

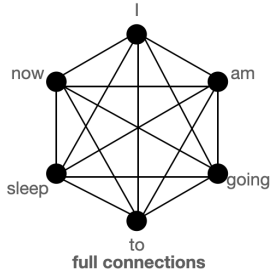


$$h_i = \sum_{j=1}^{n_L} a(x_i, x_j) x_j, \forall i \in \{1, ..., n_L\}$$

- n^2 computations are needed at each step, so computation time is convex in sequence length.

The Problem with BERT/GPT

- The 2018/2019 generation transformers like BERT have a computational constraint on the length of sequences they can consider (usually limited to $n = 512$).
- BERT's attention heads take as input the embeddings for each pair-wise interaction between tokens.



$$h_i = \sum_{j=1}^{n_L} a(x_i, x_j) x_j, \forall i \in \{1, ..., n_L\}$$

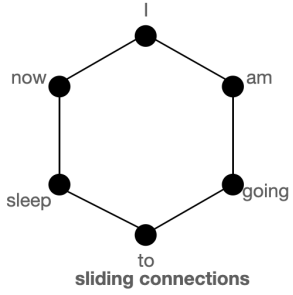
- n^2 computations are needed at each step, so computation time is convex in sequence length.
- Long-document transformers like BigBird try to approximate fully connected attention while enforcing sparsity between some/most tokens.

Three alternatives to full pairwise attention

$$h_i = \sum_{j=1}^{n_L} a(x_i, x_j) x_j, \forall i \in \{1, ..., n_L\}$$

Three alternatives to full pairwise attention

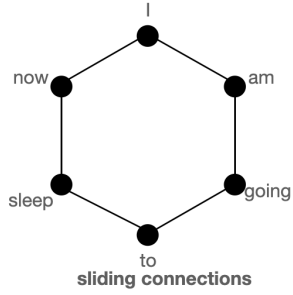
$$h_i = \sum_{j=1}^{n_L} a(x_i, x_j) x_j, \forall i \in \{1, ..., n_L\}$$



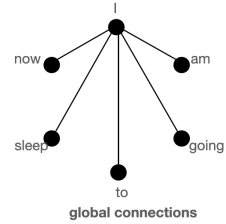
(1) $a(\cdot)$ always includes the tokens j before and after i.

Three alternatives to full pairwise attention

$$h_i = \sum_{j=1}^{n_L} a(x_i, x_j) x_j, \forall i \in \{1, ..., n_L\}$$



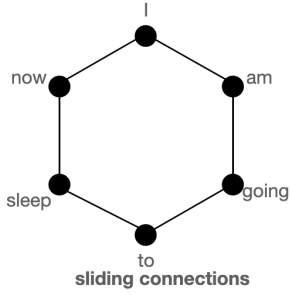
(1) $a(\cdot)$ always includes the tokens j before and after i.



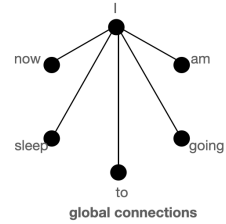
(2) Pick some important tokens (eg the first and last), and always include them in $a(\cdot)$.

Three alternatives to full pairwise attention

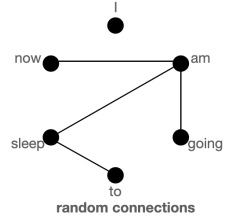
$$h_i = \sum_{j=1}^{n_L} a(x_i, x_j) x_j, \forall i \in \{1, ..., n_L\}$$



(1) $a(\cdot)$ always includes the tokens j before and after i.



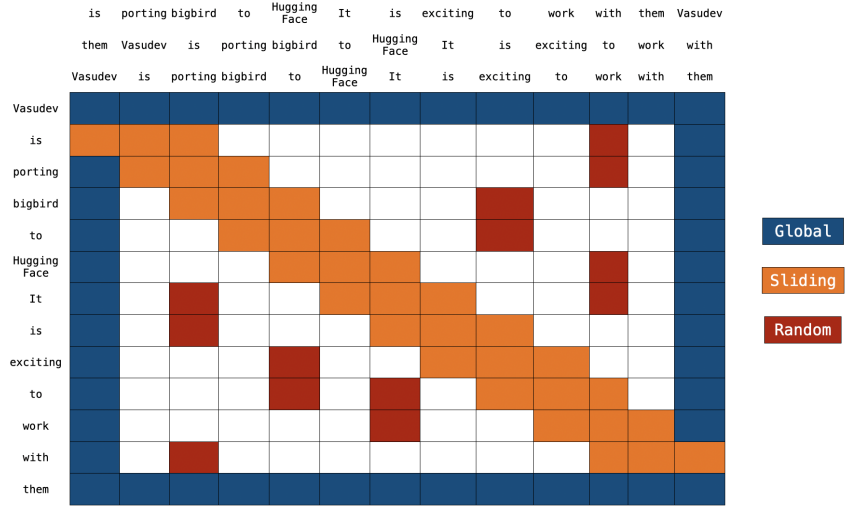
(2) Pick some important tokens (eg the first and last), and always include them in $a(\cdot)$.



(3) Pick some random tokens j to be included in $a(\cdot)$.

BigBird's Block Sparse Attention

BigBird's Block Sparse Attention



- Block sparse attention is an efficient implementation of these three alternative attention mechanisms: Each token attends to sliding tokens, some global tokens, & some random tokens.
- Extends the context window from 512 tokens to 4096 tokens, superior on long-document tasks.

Long-context transformers are still local

- **local semantics** get at linguistic information/relations at the level of sentences or paragraphs.
- **global semantics** are at the level of documents longer than paragraphs, or at the level of whole corpora.

- 4000 tokens $\approx$ 3000 words.
  - that's great but still shorter than many interesting/important documents
  - eg long news articles, political speeches, judicial opinions.
- further, can't solve problems that need information from multiple documents in a large corpus.

Applying local-semantics methods

Group Activity

What is the endpoint of NLP?

What is the endpoint of NLP?

Machine understanding of text **discourse** across long documents and corpora**.**

- good summaries of long texts: extraction of relevant information, discarding of irrelevant information.
- question answering: retrieving evidence and answers from large corpora
- what else?

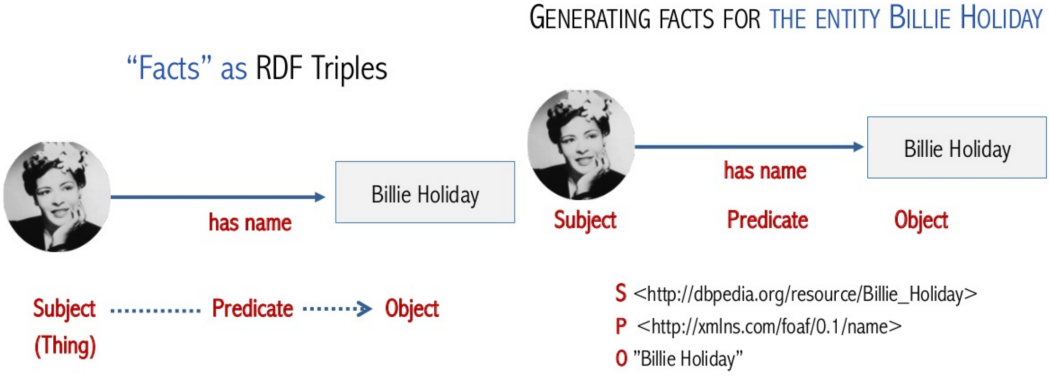
Outline

Knowledge Graphs

- A structured graph representing facts (and assertions?) as tuples.
- Entities are nodes, relations are edges:
  - (head entity, relation, tail entity)

Knowledge Graphs

- A structured graph representing facts (and assertions?) as tuples.
- Entities are nodes, relations are edges:
  - (head entity, relation, tail entity)
- E.g., DBPedia: crowd-sourced effort to extract structured information from Wikipedia and make it available as linked open data.



Soares et al (2019): Contrastive Learning for Relation Extraction

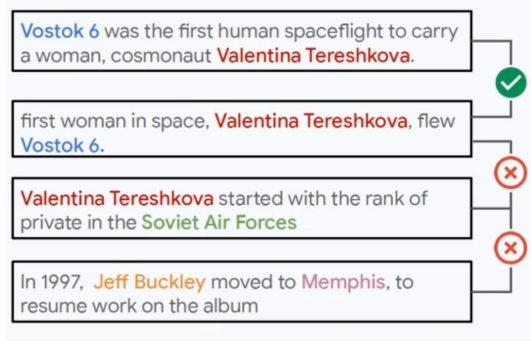
- English Wikipedia:
  - use named entity recognition and co-reference resolution to assign unique ID's.
  - take sentences with at least 2 entities.

Soares et al (2019): Contrastive Learning for Relation Extraction

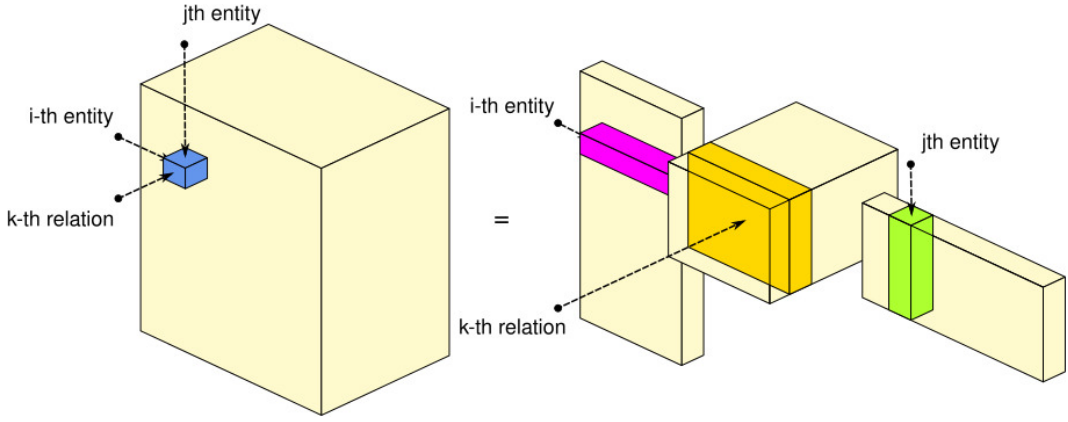
- English Wikipedia:
  - use named entity recognition and co-reference resolution to assign unique ID's.
  - take sentences with at least 2 entities.

Negative sampling objective:

- Positive samples:
  - statements with the same two entities
- Negative samples:
  - same entity, one other entity
  - two different entities



Knowledge Graph Embeddings



- Many proposed approaches for embedding KG entities, for example by predicting the presence of a bilateral relation.
  - seems to be an unsolved problem.

Learning from Knowledge Graph Relations

- Knowledge graph embeddings can be used to fill in missing relations.

e.g., parents of a person are often married, so

(John, parent of, Anne) + (Mary, parent of, Anne) $\rightarrow$ (John, married to, Mary)

Learning from Knowledge Graph Relations

- Knowledge graph embeddings can be used to fill in missing relations.

e.g., parents of a person are often married, so

(John, parent of, Anne) + (Mary, parent of, Anne) $\rightarrow$ (John, married to, Mary)

EXAMPLES OF PATHS LEARNED BY PRA ON FREEBASE TO PREDICT WHICH COLLEGE A PERSON ATTENDED

Relation Path	F1	Prec	Rec	Weight
(draftedBy, school)	0.03	1.0	0.01	2.62
(sibling(s), sibling, education, institution)	0.05	0.55	0.02	1.88
(spouse(s), spouse, education, institution)	0.06	0.41	0.02	1.87
(parents, education, institution)	0.04	0.29	0.02	1.37
(children, education, institution)	0.05	0.21	0.02	1.85
(placeOfBirth, peopleBornHere, education)	0.13	0.1	0.38	6.4
(type, instance, education, institution)	0.05	0.04	0.34	1.74
(profession, peopleWithProf., edu., inst.)	0.04	0.03	0.33	2.19

Hayashi et al (2019): Generating with KG inputs.

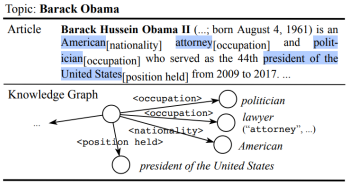


Figure 1: Overview of our task of language modeling conditioned on structured knowledge. For a given topic, we want to learn an LM that leverages the knowledge graph through relations when modeling the text.

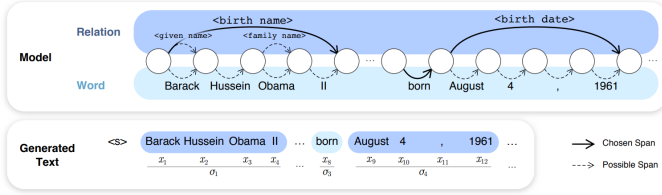


Figure 2: While generating, our model switches between the two sources, namely "Relation" and "Word". Nodes represent hidden states up to each token, and edges represent possible span matches, *i.e.*, choice of latent variables. In this example, we show one choice of latent variables with solid lines, and other options as dashed lines. We also show an "annotation" of the token sequence by the spans and sources we choose.

Language models and knowledge graphs

Hayashi et al (2019): Generating with KG inputs.

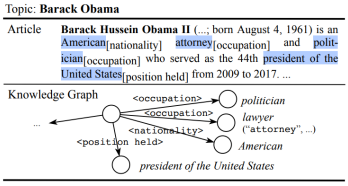


Figure 1: Overview of our task of language modeling conditioned on structured knowledge. For a given topic, we want to learn an LM that leverages the knowledge graph through relations when modeling the text.

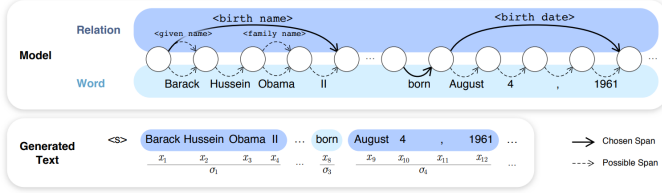


Figure 2: While generating, our model switches between the two sources, namely "Relation" and "Word". Nodes represent hidden states up to each token, and edges represent possible span matches, *i.e.*, choice of latent variables. In this example, we show one choice of latent variables with solid lines, and other options as dashed lines. We also show an "annotation" of the token sequence by the spans and sources we choose.

Wang et al (2020): Language models *are* knowledge graphs.

- iterate over entity pairs, search LM attention weights for highest-probability phrase that connect them.
- construct / complete knowledge graph with these LM outputs.

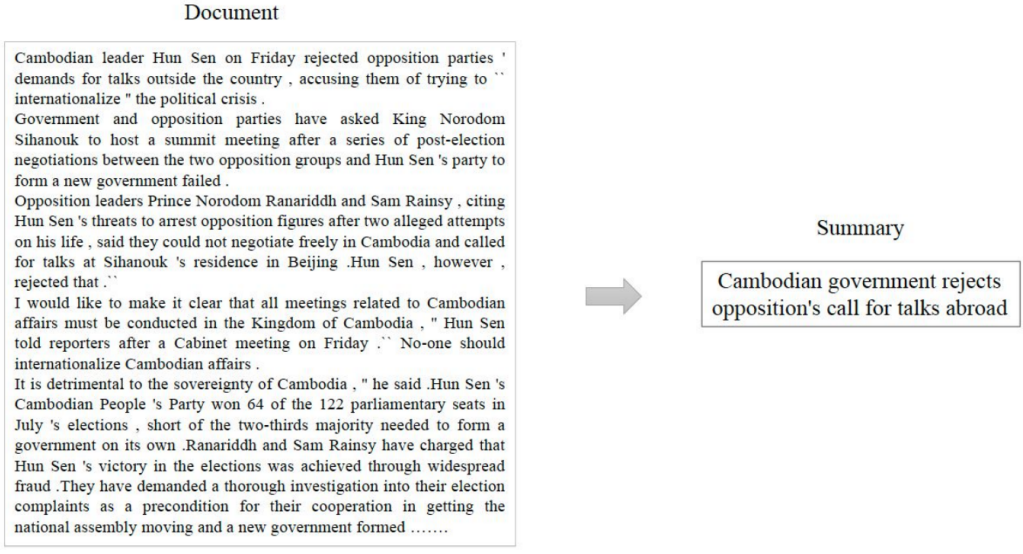
Outline

Text Summarization

Goal: produce a shorter version of a text that contains the most relevant or important information.

- obvious applications in law / legal practice.
- not proven: dimension reduction or information extraction for social science measurement.

Single Document Summarization



Source: Chan Young Park, Dan Jurafsky Slides

- **Extractive summarization:**
  - create the summary from phrases or sentences in the source document(s)
  - e.g. MemSum (Gu et al, ACL 2022) is a light-weight reinforcement-learning model that scores sentences and then stops summarizing based on the extraction history.

- **Extractive summarization:**
  - create the summary from phrases or sentences in the source document(s)
  - e.g. MemSum (Gu et al, ACL 2022) is a light-weight reinforcement-learning model that scores sentences and then stops summarizing based on the extraction history.
- **Abstractive summarization:**
  - express the ideas in the source documents using (at least in part) different words
  - e.g., fine-tune **Big Bird Pegasus** to reconstruct provided summaries.

Summarization with Human Feedback (Stiennon et al 2020)

Collect a large, high-quality dataset of human comparisons between summaries, train a model to predict the human-preferred summary, and use that model as a reward function to fine-tune a summarization policy using reinforcement learning

Summarization with Human Feedback (Stiennon et al 2020)

Collect a large, high-quality dataset of human comparisons between summaries, train a model to predict the human-preferred summary, and use that model as a reward function to fine-tune a summarization policy using reinforcement learning

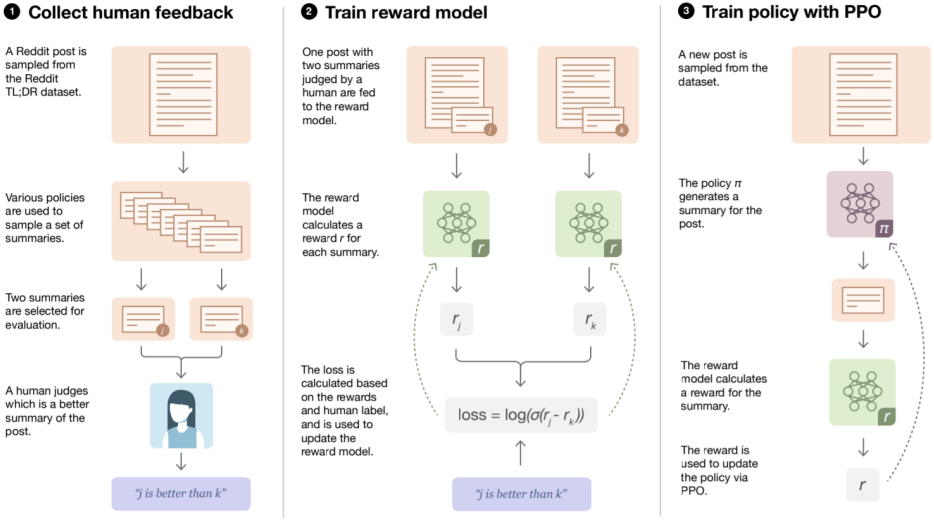
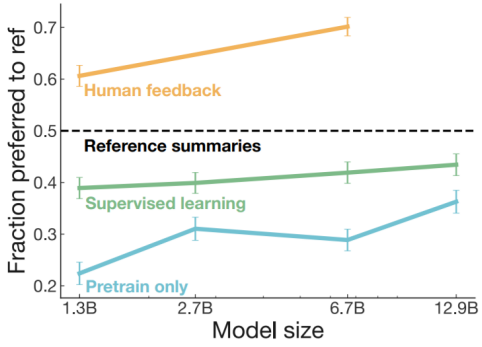


Figure 2: Diagram of our human feedback, reward model training, and policy training procedure.



[r/dating_advice] **First date ever, going to the beach. Would like some tips**
Hey Reddit! I (20M) would like some tips, because I have my first ever date tomorrow (although I've had a gf for 3 years, but no actual dating happened), and we're going to the beach.

I met this girl, we have mutual friends, at a festival a few days ago. We didn't kiss, but we talked, held hands, danced a bit. I asked her to go on a date with me, which was super hard as it is the first time I've asked this to anybody. What I mean to say is, it's not like a standard *first* date because we already spent some time together.

I'm really nervous and excited. I'm going to pick her up tomorrow, we're cycling to the beach which will take 30 minutes, and then what? I'm a bit scared. Should I bring something (the weather, although no rain and sunny, is not super so no swimming), should we do something. I'd like all the tips I can get. Thanks!

Human written reference TL;DR	**6.7B supervised model**	**6.7B human feedback model**
First date after 3 years in a relationship, going to the beach, terrified. What to bring with me, what to do?	Going on a date with a girl I met a few days ago, going to the beach. What should I bring, what should we do?	Going on my first ever date tomorrow, cycling to the beach. Would like some tips on what to do and bring. I'm a bit nervous and excited. Thanks!

Table 1: Example of post and samples on the TL;DR dataset, chosen to be particularly short. For random samples (along with posts), see Appendix H and our website.

Outline

Open Question Answering and Claim Verification

Perhaps the most difficult global semantics tasks:

- Open question answering:
  - Answer any question.
  - "What are the responsibilities of the mayor of Zurich?"
- Open claim verification:
  - Check whether a plain-text claim is true or false.
  - "Zurich has the second-highest per-capita income of any city in Europe."

Open Question Answering and Claim Verification

Perhaps the most difficult global semantics tasks:

- Open question answering:
  - Answer any question.
  - "What are the responsibilities of the mayor of Zurich?"
- Open claim verification:
  - Check whether a plain-text claim is true or false.
  - "Zurich has the second-highest per-capita income of any city in Europe."
- Both problems are solved using information retrieval pipelines:
  - search large corpora or knowledge graphs for evidence
  - use evidence to answer the question or check the claim

Information Retrieval for Question Answering

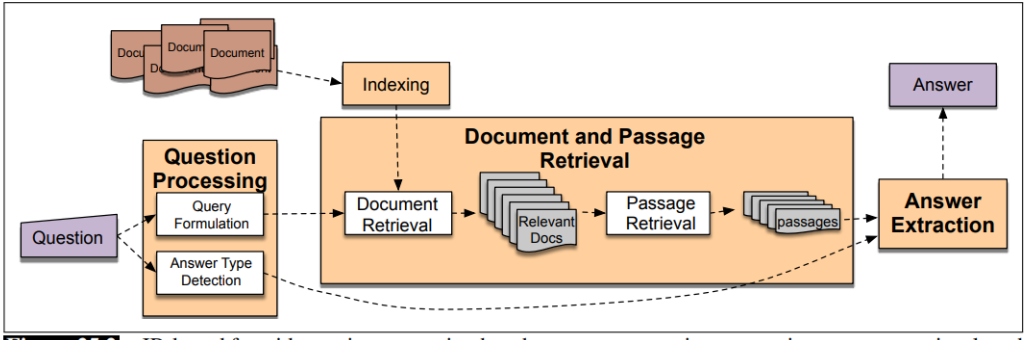


Figure 25.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

- e.g., IBM Watson is a fast search engine over a knowledge base.

Automated Claim Verification

Claim *(by Minister Shailesh Vara)*
"The average criminal bar barrister working full-time is earning some £84,000."
Verdict: FALSE *(by Channel 4 Fact Check)*
The figures the Ministry of Justice have stressed this week seem decidedly dodgy. Even if you do want to use the figures, once you take away the many overheads self-employed advocates have to pay you are left with a middling sum of money.

1. Claim spotting (what to fact check – facts vs opinions, etc)
2. Evidence retrieval
3. Evidence filtering
4. Fact-check claim given evidence (textual entailment)

Information Retrieval

- Input is a plain text query (a question or a claim)

Information Retrieval

- Input is a plain text query (a question or a claim)
- The standard approach is **BM25:**
  - roughly, rank all documents by their TF-IDF cosine similarity to the query.
  - index every document by which words it contains, to quickly filter out irrelevant documents.

Information Retrieval

- Input is a plain text query (a question or a claim)
- The standard approach is **BM25:**
  - roughly, rank all documents by their TF-IDF cosine similarity to the query.
  - index every document by which words it contains, to quickly filter out irrelevant documents.
- Problem: requires exact word overlap (not synonyms).
  - alternatives use embeddings, e.g. S-BERT (but separate problem of how to encode long documents).
  - then can do fast approximate search over dense vectors (e.g. Faiss, Johnson et al 2017)
  - BM25 still usually does better, but this is being actively researched

Inference Step

- Question answering:
  - take retrieved evidence as the context passage, and do local question answering (as in Week 11 lecture)

Inference Step

- Question answering:
  - take retrieved evidence as the context passage, and do local question answering (as in Week 11 lecture)
- Claim verification:
  - take retrieved evidence as the premise, and the claim as the hypothesis, and do textual entailment (also like in Week 11).

Choice of KB in Automated Fact-Checking (Stammbach, Zhang, and Ash 2021)

Claim	Wikipedia Evidence	Evidence from Scientific Abstracts
Global warming is driving polar bears toward extinction. **True label: SUPPORTED**	Habitat destruction: Rising global temperatures, caused by the greenhouse effect, contribute to habitat destruction, endangering various species, such as the polar bear. Polar bear: "Bear hunting caught in global warming debate". Global warming: Rising temperatures push bees to their physiological limits, and could cause the extinction of bee populations. Extinction risk from global warming: "Recent Research Shows Human Activity Driving Earth Towards Global Extinction Event". **predicted label: NOT ENOUGH INFO**	Polar bears will largely disappear from the southern portions of their range by mid-century (Stirling and Derocher, 2012). While the polar bear is the most well-known species imperiled by global warming, and the first to be listed under the ESA solely due to this factor, it was not the first species protected under the statute in which global warming played a significant role. This highly publicized milestone firmly cemented the polar bear as the iconic example of the devastating impacts of global warming on the planet's biodiversity (Cummings and Siegel, 2009). **predicted label: SUPPORTED**

Table 2: Label Accuracy Across Claim Verification Tasks Using Different Knowledge Bases

	Claim Verification Task						
Knowledge Base	FEVER	SciFact	Climate	Presidential	Real-World	Fool Me Twice	Avg.
Wikipedia	**74**	39	43	1	38	24	37
Science Abstracts	47	**60**	**45**	1	31	9	32
NYTimes	55	39	43	10	36	16	33
Google API	72	50	36	**26**	**61**	**40**	47
None	38	38	34	0	14	0	21
All	75	59	47	21	54	41	50
Best Evidence	74	60	43	26	61	40	**51**

Recap: Machine Learning vs Causal Inference

Recap: Machine Learning vs Causal Inference

Machine Learning:

- in ML, we already know the truth from the dataset.
- we take the labels as given, we just want to predict them.
- we can always verify our model works using the test set.

Recap: Machine Learning vs Causal Inference

Machine Learning:

- in ML, we already know the truth from the dataset.
- we take the labels as given, we just want to predict them.
- we can always verify our model works using the test set.

Causal Inference:

- Causal inference is about *what we don't know yet*.
- how do we know if a new policy will work?
  - for example, effect of central bank announcements on inflation.

Recap: Machine Learning vs Causal Inference

Machine Learning:

- in ML, we already know the truth from the dataset.
- we take the labels as given, we just want to predict them.
- we can always verify our model works using the test set.

Causal Inference:

- Causal inference is about *what we don't know yet*.
- how do we know if a new policy will work?
  - for example, effect of central bank announcements on inflation.
- There isn't a machine learning dataset to train a model on.
  - central bank can't experiment with monetary policy.
- this is where causal inference methods are needed.

Causal Inference: Examples

- Tons of A/B testing done by tech companies all the time.
  - eg, how YouTube gets people to watch more videos.
- Hansen, McMahon, and Prat (2018) paper on topic models and central bank meetings:
  - natural experiment where there was an unexpected increase in transparency of the meetings content.
- Widmer, Galletta, and Ash (2022) on cable-news slant:
  - instrumental-variables design where local Fox News channel position makes people watch more Fox News.

Local Semantics Idea: Adding in Causality

- In your slide template from earlier, at the end you will see "Extra Slide for Follow-Up Task".
- Task: Brainstorm an idea for how causal inference could be used in your project.
- e.g.:
  - A **causal question/task** with a text-based treatment
  - A **causal question/task** with a text-based outcome
  - A **causal question/task** using a dialogue system
- For social science, this may be observational data. for an app, this may be an impact evaluation or HCI study.
- 10 minutes:
  - brainstorm some ideas with your group and add to this last slide.
  - explain how the extended project uses both causal inference and machine learning, and the distinction between them.