Natural Language Processing for Law and Social Science

10. Semantics and Information Retrieval

In-Class Presentations

- Antoniak, Mimno, and Levy, "Narrative Paths and Negotiation of Power in Birth Stories"
- Widmer, Galletta, Ash,"Media Slant is Contagious"

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?

Four modes for NLP

- **Local:** get at linguistic information/relations from local context, e.g. sentences, paragraphs:
  - computing local sentiment
  - textual entailment
  - co-reference resolution
  - closed question answering
- **Long document**: linguistic information from long documents:
  - TF-IDF and CBOW representations $\rightarrow$ supervised learning
  - cosine distance between vectors
- **Global / knowledge base**: corpus level tasks:
  - information retrieval / search
  - open question answering / claim checking
  - knowledge graphs
- **Generative/Creative:** generate text for some purpose.
  - compose a sonnet
  - draft a legal brief to attack the opponent's brief

Is BERT obsolete?

- Can RLHF-autoregressive models like GPT do everything?

Is BERT obsolete?

- Can RLHF-autoregressive models like GPT do everything?
- Not quite:
  - GPT does not generate document encodings
  - GPT cannot do search/retrieval

Outline

Local Semantics

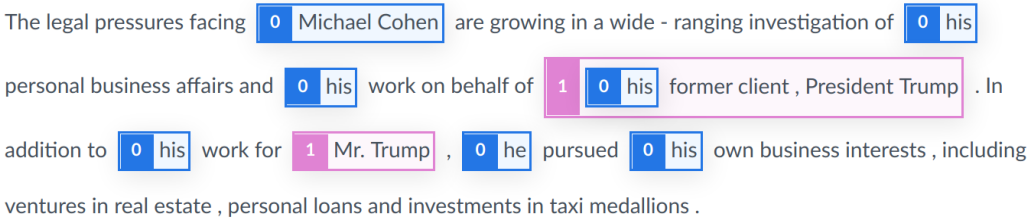
Global / Retrieval

Local Semantics is Basically Solved

- Sentiment:
  - subjective vs objective statements
  - positive, negative, neutral
  - emotions (angry, happy, sad, etc)
- Stance (whether entity A favors or disfavors entity B)

Local Semantics is Basically Solved

- Sentiment:
  - subjective vs objective statements
  - positive, negative, neutral
  - emotions (angry, happy, sad, etc)
- Stance (whether entity A favors or disfavors entity B)
- Co-Reference Resolution



Semantic Role Labeling (PropBank Labels)

Ex1: [$_{\text{Arg0}}$ The group] *agreed* [$_{\text{Arg1}}$ it wouldn't make an offer].

Ex2: [$_{\text{ArgM-TMP}}$ Usually] [$_{\text{Arg0}}$ John] *agrees* [$_{\text{Arg2}}$ with Mary] [$_{\text{Arg1}}$ on everything].

ARG0	agent	ARG3	starting point, benefactive, attribute
ARG1	patient	ARG4	ending point
ARG2	instrument, benefactive, attribute	ARGM	modifier

Table 1.1: List of arguments in PropBank

- Agent (ARG0)
  - Volitional/sentient involvement in event or state
  - Causes an event or change of state in another participant
- Patient (ARG1)
  - Causally affected by an agent/action
  - Undergoes change of state
- ARG2 has three functions:
  - instrument for an action ("Pat opened the door with a crowbar.")
  - attribute assigned to a patient ("Pat is an agent").
  - benefactive: the dative/indirect object ("Sasha gave the crowbar to Pat.")

ARG-M: Modifiers

ArgM-TMP	when?	yesterday evening, now
LOC	where?	at the museum, in San Francisco
DIR	where to/from?	down, to Bangkok
MNR	how?	clearly, with much enthusiasm
PRP/CAU	why?	because ... , in response to the ruling
REC		themselves, each other
ADV	miscellaneous	
PRD	secondary predication	...ate the meat raw

- AllenNLP semantic role labeling demo:

  `https://demo.allennlp.org/semantic-role-labeling`

Reading Comprehension $\leftrightarrow$ Local Question Answering

Answering questions about a passage of text to show that the system understands the passage.

Reading Comprehension ↔ Local Question Answering

Answering questions about a passage of text to show that the system understands the passage.

`https://demo.allennlp.org/reading-comprehension`

Passage Context

The institutional framework of Navarre was preserved following the 1512 invasion. Once Ferdinand II of Aragon died in January, the Parliament of Navarre gathered in Pamplona, urging Charles V to attend a coronation ceremony in the town following tradition, but the envoys of the Parliament were met with the Emperor's utter indifference if not contempt. He refused to attend any ceremony and responded with a brief "let's say I am happy and pleases me." Eventually the Parliament met in 1517 without Charles V, represented instead by the **Duke of Najera** pronouncing an array of promises of little certitude, while the acting Parliament kept piling up grievances and demands for damages due to the Emperor, totalling 67—the 2nd Viceroy of Navarre Fadrique de Acuña was deposed in 1515 probably for acceding to send grievances. Contradictions inherent to the documents accounting for the Emperor's non-existent oath pledge in 1516 point to a contemporary manipulation of the records.

Question

Who represented the Charles V at Parliament?

Local Question Answering with BERT

Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".
Q: "In what city and state did Beyoncé grow up?" A: "Houston, Texas"
Q: "What areas did Beyoncé compete in when she was growing up?" A: "singing and dancing"
Q: "When did Beyoncé release Dangerously in Love?" A: "2003"

Figure 23.11 A (Wikipedia) passage from the SQuAD 2.0 dataset (Rajpurkar et al., 2018) with 3 sample questions and the labeled answer spans.

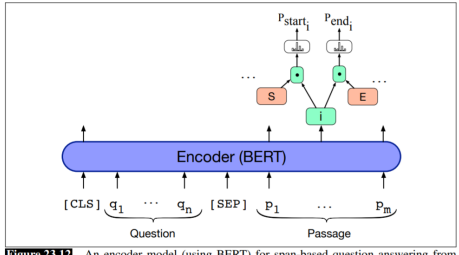


Figure 23.12 An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.

- BERT-based models learn to solve these problems by inputting {question + [SEP] + passage} and predicting the token indexes for the start and end of the answer.

Reading Comprehension with GPT-3

Stammbach, Antoniak, and Ash (2022)

Who is the villain in the following text?

Text: In questioning the use of hydraulic fracturing in New York to help produce natural gas, you do not note that the technology has been employed and continuously improved for more than 50 years and that studies by the Environmental Protection Agency and the Ground Water Protection Council have not identified a single instance of groundwater contamination. Wells where fracturing is used are specially constructed to protect drinking water sources. Regulatory oversight is extensive. The fluids mostly water that are forced into a well to create pressure to fracture rock are pushed back out by the oil and gas flowing upward for safe processing. Protecting our water supplies is important, as are reductions in greenhouse gas emissions through use of clean-burning natural gas. Banning hydraulic fracturing would be unwarranted and shortsighted, preventing production of large amounts of natural gas that could directly benefit New York consumers for decades and create thousands of good jobs.

Villain: **The villain in this text is the person who is questioning the use of hydraulic fracturing in New York.**

Textual Entailment $\leftrightarrow$ Natural Language Inference

- TE is the task of predicting whether, for a pair of sentences, the facts in the first sentence necessarily imply the facts in the second.

Sentence A (Premise)	Sentence B (Hypothesis)	Label
A soccer game with multiple males playing.	Some men are playing a sport.	entailment
An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.	neutral
A man inspects the uniform of a figure in some East Asian country.	The man is sleeping.	contradiction

- The SNLI (Stanford Natural Language Inference) dataset contains 570k human-written English sentence pairs manually labeled (by Amazon Mechanical Turk Workers) for balanced classification with the labels: entailment, contradiction, neutral.

https://demo.allennlp.org/textual-entailment

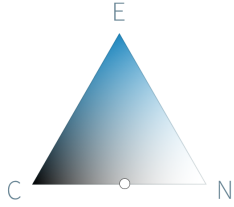
Premise

A handmade djembe was on display at the Smithsonian.

Hypothesis

Visitors could not hear the djembe.

It is **somewhat likely** that there is **no correlation** between the premise and hypothesis.



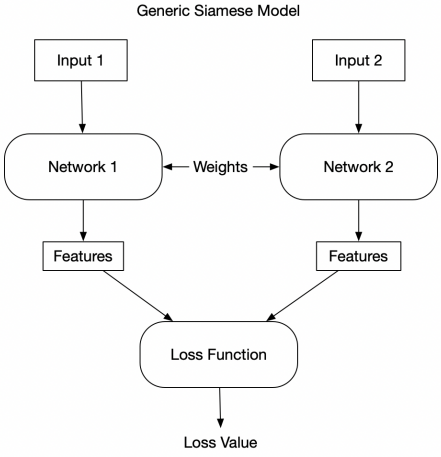
Judgement	Probability
Entailment	0.7%
Contradiction	46.4%
Neutral	52.9%

Sentence-BERT

- S-BERT (Reimers and Gurevych 2019):
  - fine-tune BERT embeddings to classify sentence pairs in textual entailment task.
  - significantly improves performance of sentence embeddings on standard tasks.

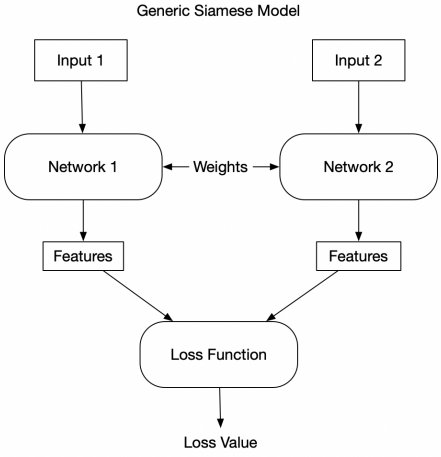
Siamese Networks

- A Siamese Neural Network (SNN) is a class of neural network architectures that contain two or more identical subnetworks (shared architecture and parameters).



Siamese Networks

- A Siamese Neural Network (SNN) is a class of neural network architectures that contain two or more identical subnetworks (shared architecture and parameters).



1. start with an "anchor" document
2. take one positive sample (document from the same class) and k negative samples (documents from randomly chosen class)
3. send all of them through the same set of hidden layers to produce embeddings
4. compute **contrastive loss** that rewards high similarity between anchor and positive sample, and rewards low similarity between the anchor and the negative samples

S-BERT Training Objective

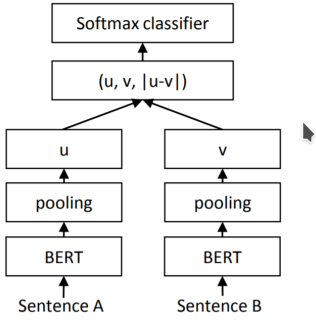


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

S-BERT SNN:

1. send sentences through two siamese BERT networks (with shared weights)
2. produce document embeddings
3. use each embedding, and their distance, as features to predict labels in the Stanford Natural Language Inference dataset (entail, contradict, neutral)

- Then at inference time, form embeddings and then compute cosine similarity between sentences.

SentenceTransformers

- SentenceTransformers (sbert.net) is an amazing python package for embedding texts or short documents.
- Initially based on S-BERT but expanded to many additional models, including embeddings trained on other tasks besides entailment:
  - paraphrase identification
  - semantic textual similarity
  - duplicate question detection
  - question-answer retrieval
- monolingual and multilingual models (for over 100 languages)

Sentence-Grams (Not Done Yet)

- With sentence embedding, we can think of documents as collections of sentences, rather than of words or phrases.

Sentence-Grams (Not Done Yet)

- With sentence embedding, we can think of documents as collections of sentences, rather than of words or phrases.
- e.g., take the average across sentence embeddings for a document, or produce sentence clusters:
  - run k-means clustering on the dataset of sentences embeddings, then represent documents as counts/frequencies over "sentence-grams".
  - show example sentences to interpret the clusters

Sentence-Grams (Not Done Yet)

- With sentence embedding, we can think of documents as collections of sentences, rather than of words or phrases.
- e.g., take the average across sentence embeddings for a document, or produce sentence clusters:
  - run k-means clustering on the dataset of sentences embeddings, then represent documents as counts/frequencies over "sentence-grams".
  - show example sentences to interpret the clusters
- Multilingual encoders (e.g. MUSE, LASER)
  - trained on multilingual machine translatio tasks
  - sentences with similar meanings in different languages should have similar vectors

Text Summarization

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

- obvious applications in law / legal practice.

Text Summarization

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

- obvious applications in law / legal practice.
- **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.
  - or use GPT

Text Summarization

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

- obvious applications in law / legal practice.
- **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.
  - or use GPT
- **Abstractive summarization:**
  - express the ideas in the source documents using different words
  - e.g., fine-tune **Longformer LED** to reconstruct provided summaries.
  - or use GPT

Not attempted yet: use summarization as pre-processing before social science measurement.

In-Class Presentations

- Licht, "Cross-lingual classification of political texts using multilingual sentence embeddings"
- Bana et al, "Work2Vec"

Outline

Open Question Answering and Claim Verification

- 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

- 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

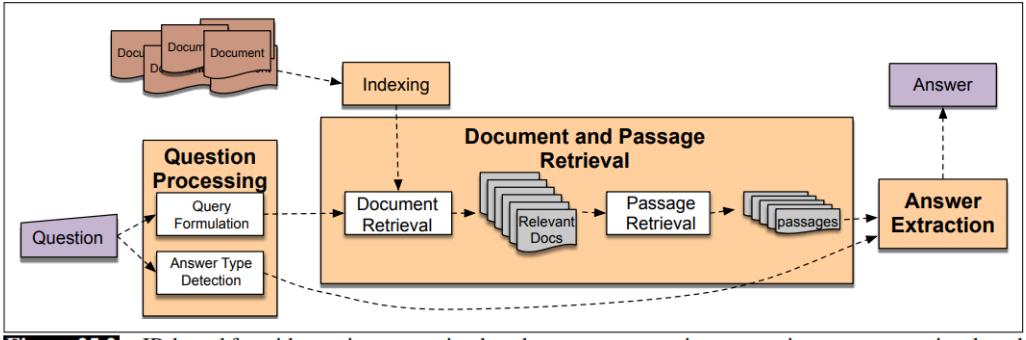


Figure 25.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

- 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)

(1) Information Retrieval Step

For both open question answering and automated claim verification:

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

(1) Information Retrieval Step

For both open question answering and automated claim verification:

- 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.

(1) Information Retrieval Step

For both open question answering and automated claim verification:

- 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

(2) Inference Step

- Question answering:
  - take retrieved evidence as the context passage, and do local question answering

(2) Inference Step

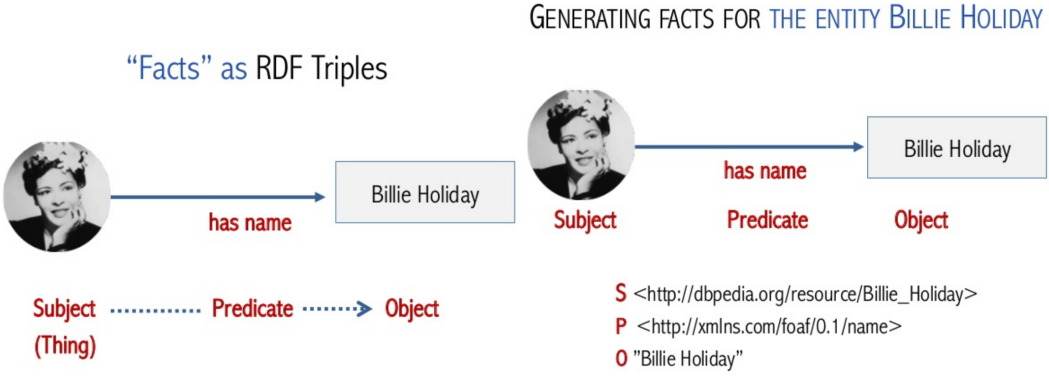
- Question answering:
  - take retrieved evidence as the context passage, and do local question answering
- Claim verification:
  - take retrieved evidence as the premise, and the claim as the hypothesis, and do textual entailment.

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.



In-Class Presentation

Ash, Gauthier, Widmer

"Relatio: Text Semantics Capture Political and Economic Narraitves"