IN-SIGHT.it

INDE lab
INtelligent Data Engineering

Large Language Models in Standards Discourse Analysis

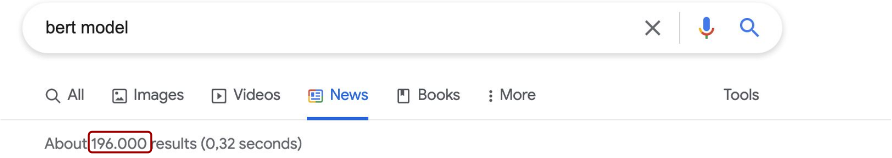
Effy Xue Li
x.li3@uva.nl
University of Amsterdam

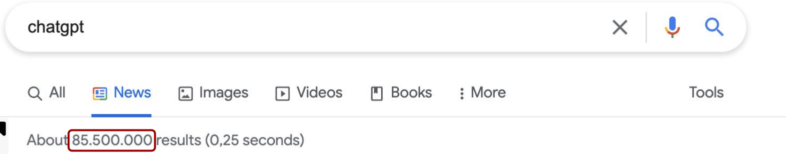
March 2023

Table of contents

Large Language Models

ChatGPT



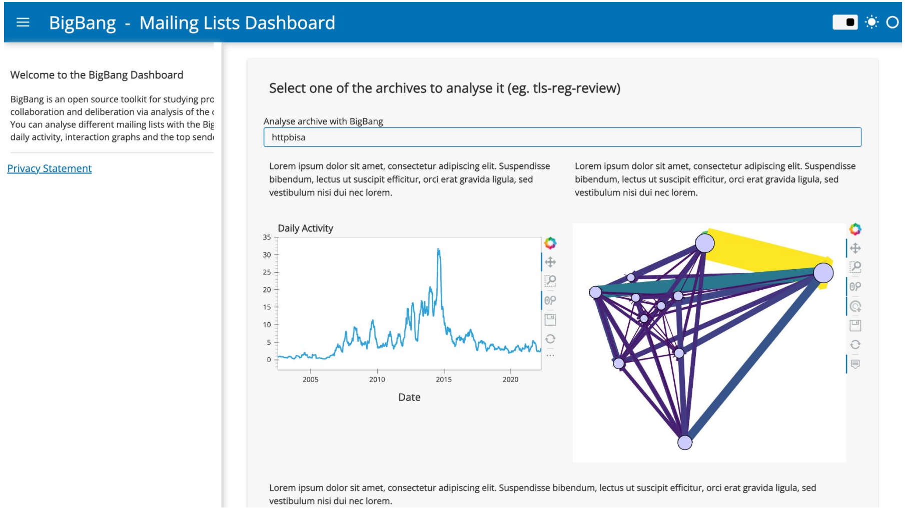


RQ: How can we use it for standards discourse analysis?

Source: https://openai.com/blog/chatgpt

BigBang Package

Toolkit for studying communications data from collaborative projects.



Bigbang dashboard

NER in Emails

- Bert-base model
- Fine-tuned with CEREC [1]

I would say the latter of the two I started linalg months ago and **Travis O** PER put a lot of effort into over the last several weeks I am not really familiar with we are really focusing on **ATLAS** ORG because it is so dang fast on most platforms It does not provide a full **LAPACK** ALLCAPS though so you have to merge it with another **LAPACK** ALLCAPS to get everything If you can figure out how to write a generic interface not to hard but only partially documented in linalgdocsmore _ notes then have at it The actual fpy interfaces are generated from a python script The more interfaces the merrier but the compatibility issue has to be addressed On **Unix** MISC we could use nm to check if the function is there On windows it are not so easy Maybe it should just be an optional function for now ie defaults to being commented out for the widest compatibility eric

[1] CEREC: A Corpus for Entity Resolution in Email Conversations

Top 10 frequent entities

- We quantitatively extract the Top 10 frequent entities for each type.
- Sample mailing list: 3gv6

```
Top 10 occurence for type:  LOC
```

	entity	counts
0	San Francisco	9
1	USA	3
2	Shanghai	2
3	China	2
4	Anaheim	1
5	Tower Hui Hui Deng denghuigmailcom	1
6	Vista level	1
7	Vista Room at the Hilton San Francisco The Vis...	1
8	Vista level of Tower	1
9	the Vista Room at the Hilton San Francisco The...	1

```
Top 10 occurence (pronouns excluded) for type:  PER
```

	entity	counts
0	Teemu	20
1	Cameron	15
2	Jari	11
3	Dan	10
4	Jouni	9
5	Cameron Byrne	9
6	David Crowe	9
7	Brian	8
8	Julien	7
9	Dan Wing	6

<= Extracted person entities align with sender-receiver analysis from meta data.

Top 10 frequent entities

```
Top 10 occurence for type:  MISC
```

	entity	counts
0	Internet	4
1	Windows	2
2	RFC	1
3	Internet Protocol	1
4	MacOS	1
5	Windows OS	1
6	IGI	1

```
Top 10 occurence for type:  ORG
```

	entity	counts
0	UE	34
1	IETF	24
2	GPP	20
3	UEs	17
4	IPvonly	16
5	DS	9
6	PDN	8
7	RFC	8
8	IMHO	7
9	GPP EPC	7

```
Top 10 occurence for type:  DIG
```

	entity	counts
0	IPv	3
1	DHCPv	1
2	teemusavolainennokiacom	1
3	PGW	1
4	IHdpdGggREhDUFYIHNIcnZlciwgdGhlbiBaGVzZSBdgREh...	1
5	ba sis	1
6	withIETFDocs	1
7	listA	1
8	STUNTURN	1
9	PNAT	1

Pros & Cons

Pros:

- Great quantitative tool for analyzing **email bodies** from large scale mailing lists.
- Extract information with types that users define.

Cons:

- Fine-tuning with labelled data makes results much better. But we don’t have …
- Fixed sets of types.
- Limited information.

Pros & Cons

Pros:

- Great quantitative tool for analyzing **email bodies** from large scale mailing lists.
- Extract information with types that users define.

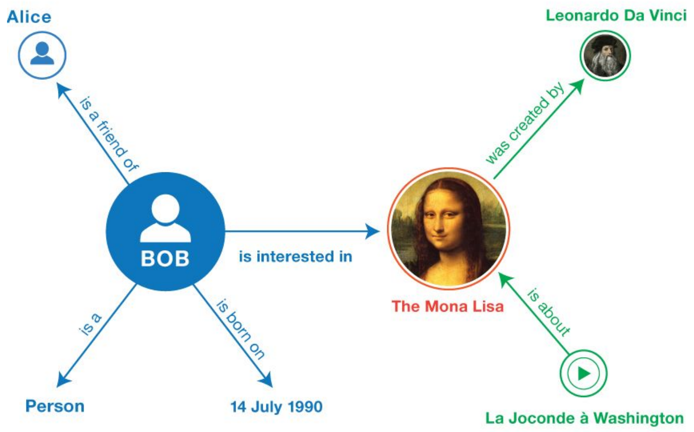
Cons:

- Fine-tuning with labelled data makes results much better. But we don't have ...
- Fixed sets of types.
- Limited information.

One step further...

Knowledge Graphs

- Definition: Network of real-world entities and their relations.
  - Entity extraction; relation extraction.
  - Multiple tasks needed.
- Challenges: Specialized domains.
  - Standards in different domains.
  - No unified schema.
- Applications.
  - Structured data.
  - Connected data.
  - Can be intervened on.

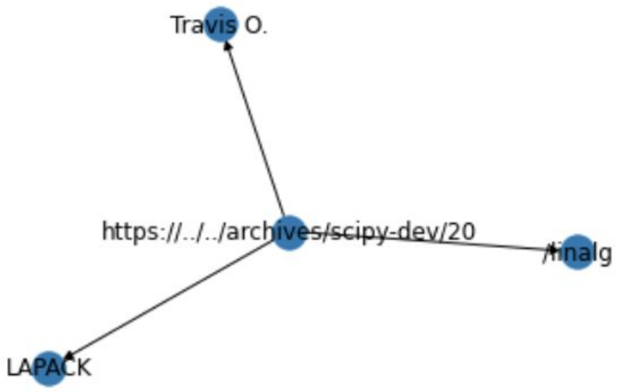


Source: https://www.w3.org/TR/rdf11-primer/

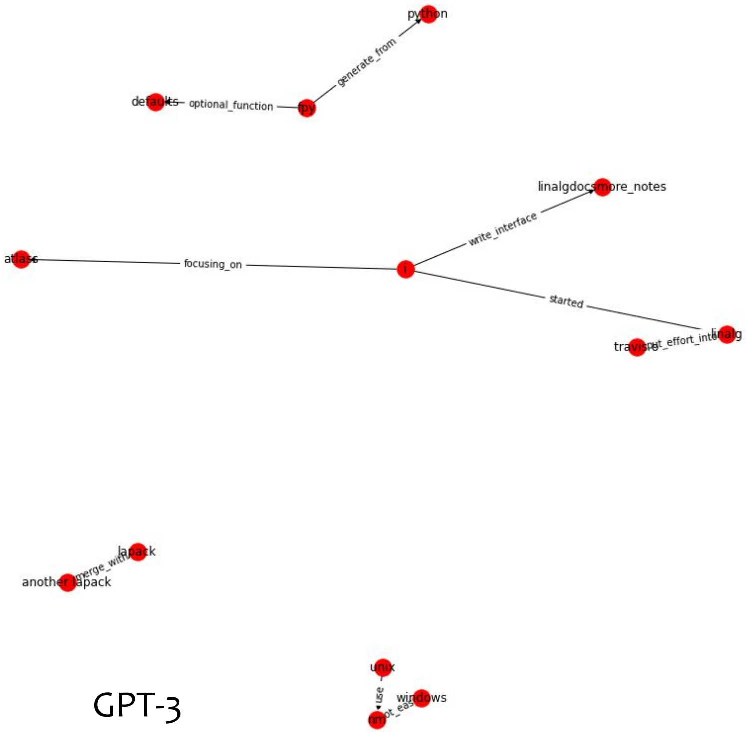
GPT-3

- Generative Pre-trained Transformers 3
- It is HUGE!
  - GPT-3 has 175 billion parameters. Bert has 110 million parameters.
  - 1,591 x larger than Bert! 100x larger than GPT-2.
- Prompt-engineering with OpenAI's APIs.
  - Task-agnostic.
  - No access to the underlying trained weights.
  - It costs money.

Knowledge Graph extraction with KGcreator and GPT-3



KGcreator[1]



GPT-3

[1] https://pypi.org/project/kgcreator/

Knowledge Graph extraction with KGcreator and GPT-3

	Entity	Type
0	Travis O.	PERSON
1	LAPACK	ORG
2	LAPACK	ORG
3	/linalg	GPE

KGcreator[1]

	source	source_attr	target	target_attr	edge
0	i	person	linalg	software	started
1	travis o	person	linalg	software	put_effort_into
2	i	person	atlass	software	focusing_on
3	lapack	software	another lapack	software	merge_with
4	i	person	linalgdocsmore_notes	document	write_interface
5	fpy	software	python	programming_language	generate_from
6	unix	operating_system	nm	software	use
7	windows	operating_system	nm	software	not_easy
8	fpy	software	defaults	software	optional_function

GPT-3

[1] https://pypi.org/project/kgcreator/

Natural Language Prompt with One-shot Example

Extract all entities with types and their relations from texts:

John Doe works at Google.

Apple is located in Cupertino.

Results:

Entities:

Entity 1: John Doe Type: Person

Entity 2: Google Type: Company

Entity 3: Apple Type: Company

Entity 4: Cupertino Type: City

Relations:

works_at(person:john doe,company:google)

located_in(company:apple, city:cupertino)

Extract all entities with types and their relations from texts:

{Email body}

Results:

Limitation & Concerns

- Potential privacy and ethical issues.
  - We would like not to send our data to another company.
- It costs more when the amount of emails goes up.
  - For 2 million emails, it will cost ~17,900 USD.
  - It takes ~1 min for processing one API call.
- No control over the model.
  - The results are not deterministic.
  - No access to the underlying weights. No way to debug the model.

Future Directions

- Denoising results given constraints.
- Prompt optimisation.
- Local models that can achieve comparable performance with GPT-3.
    - GPT-3 as a labeler.
    - Hierarchical information extraction.
    - …

Thank you!