Industry Use Case: Knowledge Search and Discovery

Bloomberg

Engineering

ECIR 2023
6 April, 2023

Edgar Meij, Ph.D.
Head of AI Search
@edgarmeij

TechAtBloomberg.com

Who uses KGs?

- Bloomberg
- IBM
- Amazon
- Walmart
- eBay
- LinkedIn
- Yahoo
- Facebook
- Microsoft/Bing
- Google
- Uber
- Airbnb
- Siemens
- Zalando
- Elsevier

	Data model	Size of the graph	Development stage
Microsoft	The types of entities, relations, and attributes in the graph are defined in an ontology.	~2 billion primary entities, ~55 billion facts	Actively used in products
Google	Strongly typed entities, relations with domain and range inference	1 billion entities, 70 billion assertions	Actively used in products
Facebook	All of the attributes and relations are structured and strongly typed, and optionally indexed to enable efficient retrieval, search, and traversal.	~50 million primary entities, ~500 million assertions	Actively used in products
eBay	Entities and relation, well-structured and strongly typed	Expect around 100 million products, >1 billion triples	Early stages of development and deployment
IBM	Entities and relations with evidence information associated with them.	Various sizes. Proven on scales documents >100 million, relationships >5 billion, entities >100 million	Actively used in products and by clients

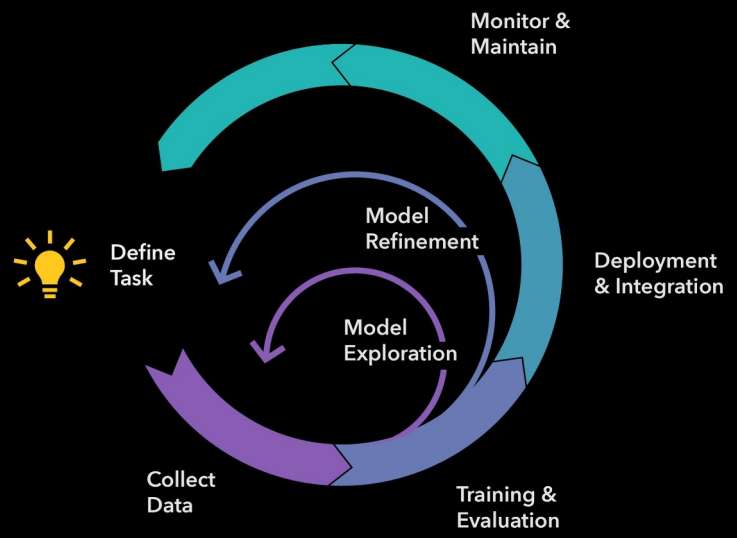
From https://cacm.acm.org/magazines/2019/8/238342-industry-scale-knowledge-graphs/

TechAtBloomberg.com

What is symbolic AI typically used for?

- Discovery
- Querying for and reasoning over instances or topology/structure
- Search, (stateful) QA, autocomplete suggestions, recommender systems
- Joining data: schema mapping, constraint validation, deduplication, and merging
- Entity linking and information extraction
- Link/Type/Attribute prediction

Use Case: Graph Analytics

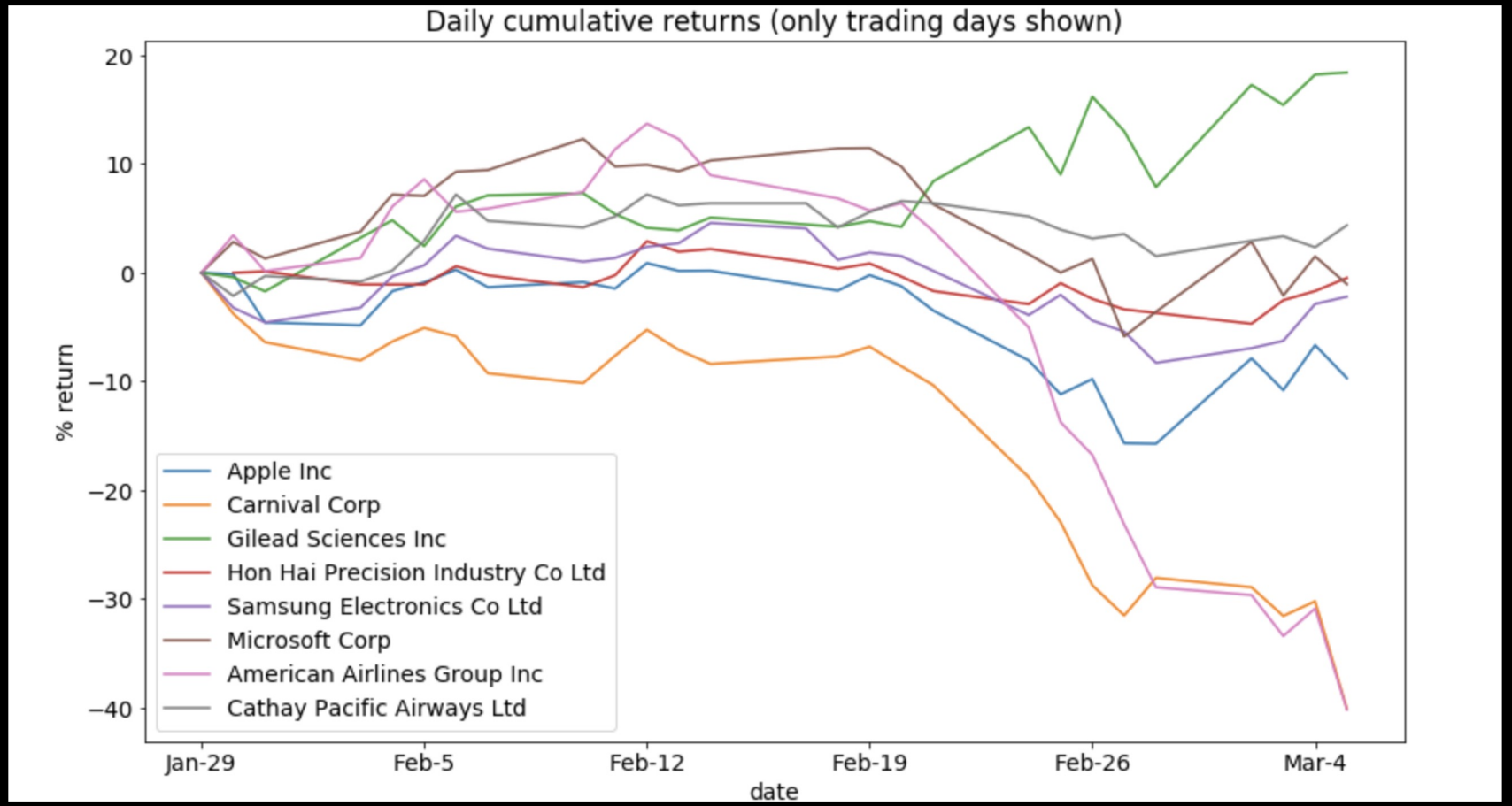


TechAtBloomberg.com

Financial Graph Analytics: COVID-19's Impact

- Combining bbKG, news, and market data
- Using entities in news as starting point
  - *Which companies are most affected directly by COVID-19?*
- Expanding the seed universe using bbKG to obtain a broader context
  - *Which companies in the supply chain are also or indirectly affected?*
  - *Which industries are most affected directly by COVID-19?*
  - *Which industries in the supply chain are also or indirectly affected?*
- Aggregate analysis on the broader context

Financial Graph Analytics: COVID-19's Impact



TechAtBloomberg.com

Bloomberg

Engineering

Enabling Financial Analytics: Factor Investing

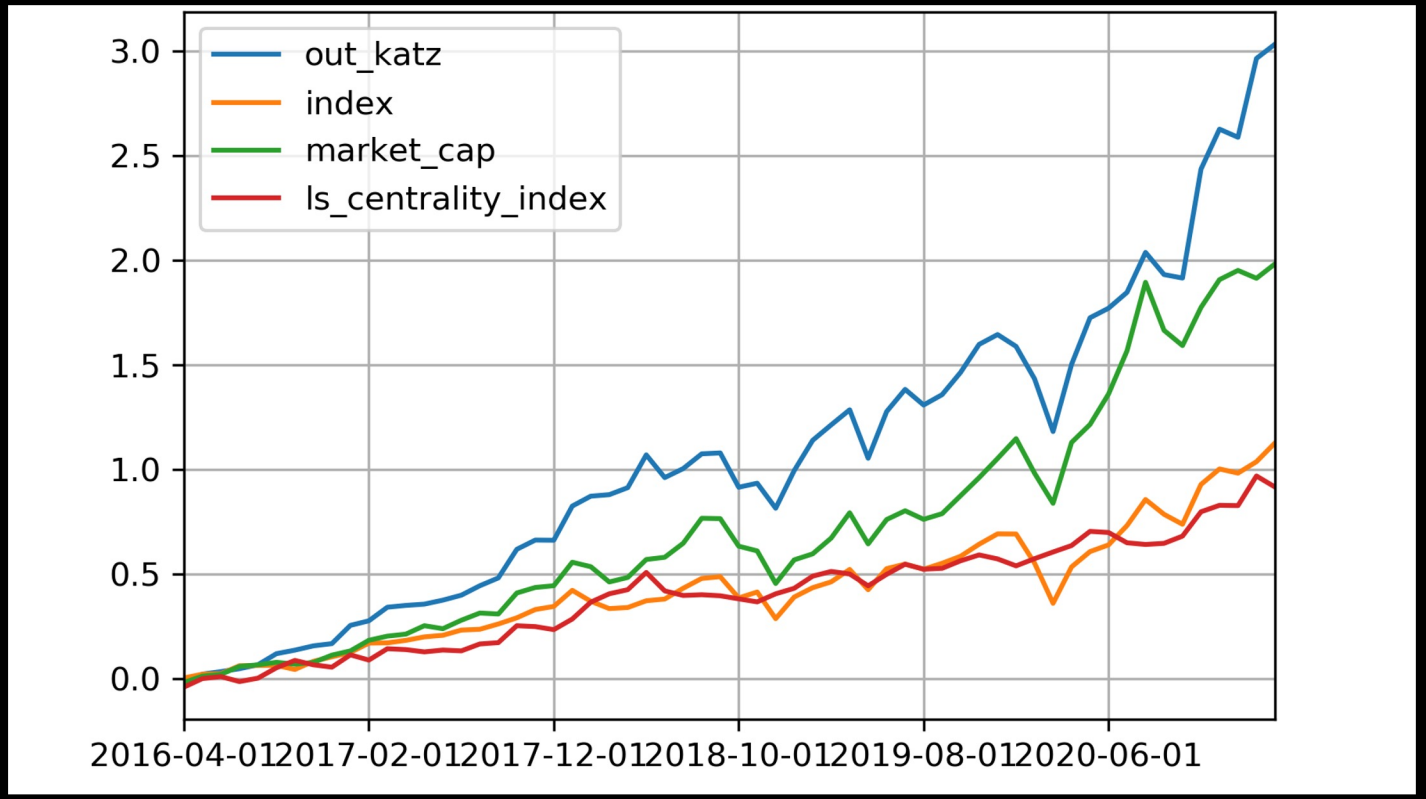
- Taking the analysis in the previous study one step further
- **Factor investing**: investment approach that targets quantifiable company characteristics (i.e., “factors”) that can explain differences in stock returns
- An alternative way to generate those characteristics: graph analytics

Enabling Financial Analytics: Factor Investing

- How?
  - Compute centrality of companies from the supply chain graph
  - Use the centrality scores to construct investment portfolios

- Experimental setup
  - Use historical supply chain data
  - Backtesting: use historical stock returns to assess the investment strategy
  - Baselines for comparison
    - Indices
    - Market cap

Enabling Financial Analytics: Factor Investing



Bloomberg Whitepaper: “Unlocking the Alpha of Supply Chains using Centrality Measures”

Symbolic Approaches: Challenges & Opportunities

TechAtBloomberg.com

Symbolic Approaches

- Maintenance
  - Ensuring data quality, timeliness, coverage, and consistency
  - Disambiguate across sources and languages, maintain identity, handle multi-linguality

- Applications and access
  - Merge global, company/domain-specific, and product-specific knowledge
  - Supporting point-in-time queries for companies, products, brands, people, etc.
  - Representation learning

- Do all of the above, *at scale*

Ensuring Quality & Consistency

- Human-in-the-loop: curation and augmentation by human experts
- Machine-in-the-loop: have a system tee up items needing curation
- Automated checks
  - Continuous monitoring
  - Cross-domain consistency
  - Downstream integration, regression, and smoke tests

Entities Evolve on a Regular Basis...

- Entities emerge and dissolve
  - Mergers and acquisitions
  - IPO
- Properties and relationships change
  - Stock ticker changes
  - Index membership
  - Supply chain

Dartpoints Holding to Buy Immedion

By Bloomberg Automation

(Bloomberg) -- Dartpoints Holding to Buy Immedion, according to a press release.

Key excerpts:

- DartPoints, an owner and operator of edge colocation data centers, announces that it has signed definitive agreements to acquire Immedion, a provider of colocation, cloud, and managed services with eight data centers in seven markets throughout South Carolina, North Carolina, Ohio and Indiana.
- The acquisition is expected to close in the second quarter of 2021.

To view the source of this information, click here

Related tickers:
1349252D US (Immedion LLC)
1389230D US (Astra Capital Management LLC)

Supporting Point-in-Time Queries

- Financial analytics require awareness of these changes, e.g.,
  - Changes in supply chain must be factored into COVID-19 impact analyses
  - Historical company names need to be considered when running text analytics on historical documents

Challenge: storing this data and support these queries efficiently.

Subsymbolic Approaches: Challenges & Opportunities

TechAtBloomberg.com

State of the Art

- Scaling out deep learning architectures
- Speeding up inference using multi-GPU
  - Exploit parallelism
- Make models smaller and faster
  - Reduce dimensionality, e.g., using teacher-student distillation or by training a smaller model to "distill" a larger model
  - Quantization

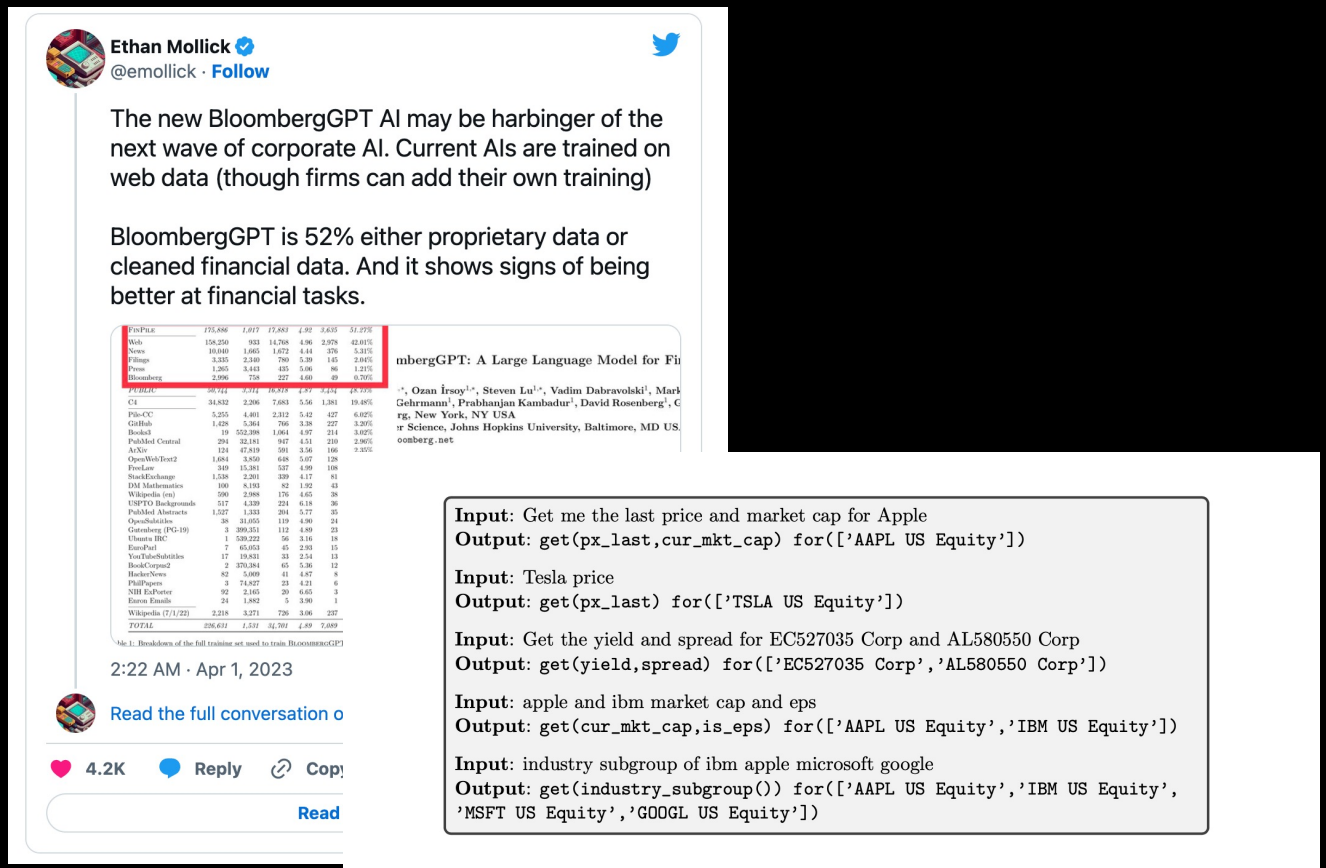
State of the Art

- Representation learning and reusing knowledge across tasks
  - Pretrain general language models to predict next words (BERT, GPT-*k*, …)
  - Use it to initialize a task-specific model
  - Achieve higher accuracy with fewer examples
  - Zero or few-shot learning

- Learning more from less annotations
  - Distant supervision
  - Transfer learning
  - Active learning

BloombergGPT

- https://arxiv.org/pdf/2303.17564.pdf



Input: Get me the last price and market cap for Apple
Output: `get(px_last,cur_mkt_cap) for(['AAPL US Equity'])`

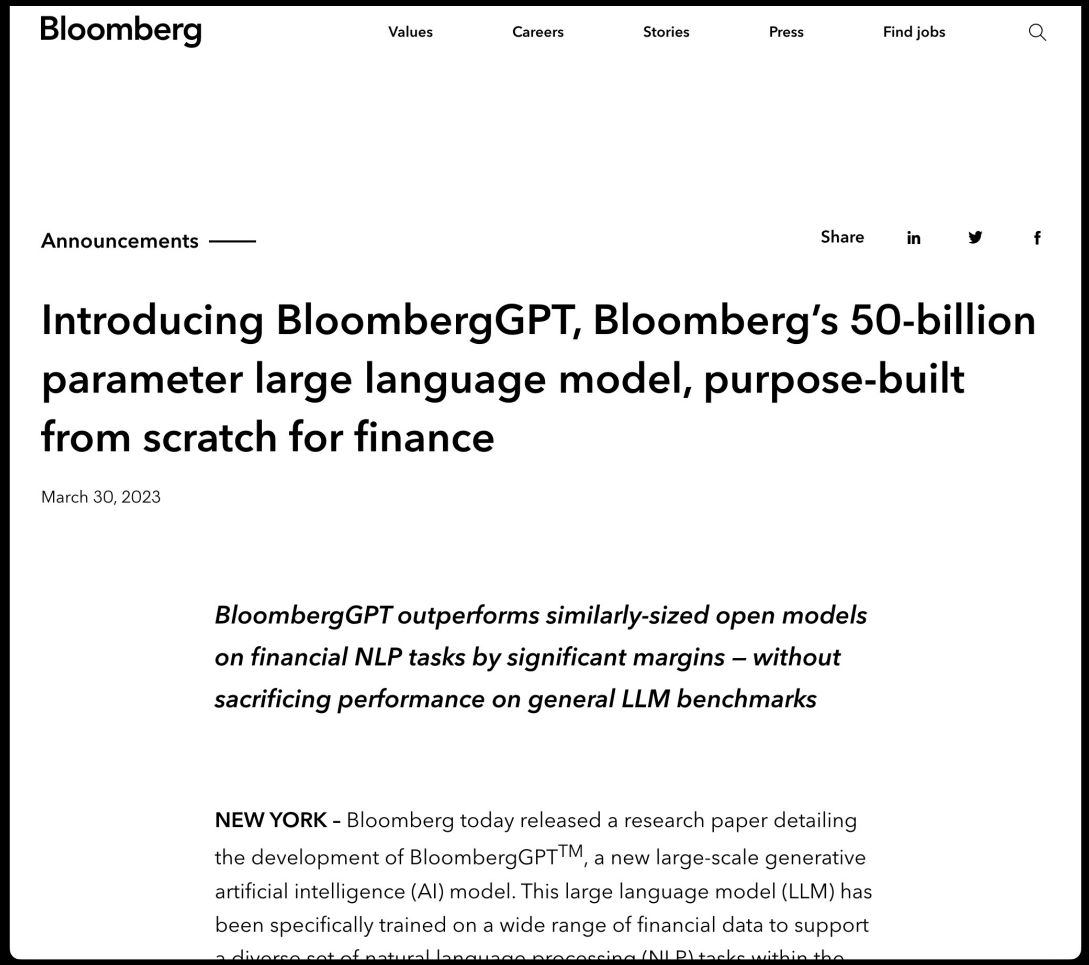
Input: Tesla price
Output: `get(px_last) for(['TSLA US Equity'])`

Input: Get the yield and spread for EC527035 Corp and AL580550 Corp
Output: `get(yield,spread) for(['EC527035 Corp','AL580550 Corp'])`

Input: apple and ibm market cap and eps
Output: `get(cur_mkt_cap,is_eps) for(['AAPL US Equity','IBM US Equity'])`

Input: industry subgroup of ibm apple microsoft google
Output: `get(industry_subgroup()) for(['AAPL US Equity','IBM US Equity', 'MSFT US Equity','GOOGL US Equity'])`

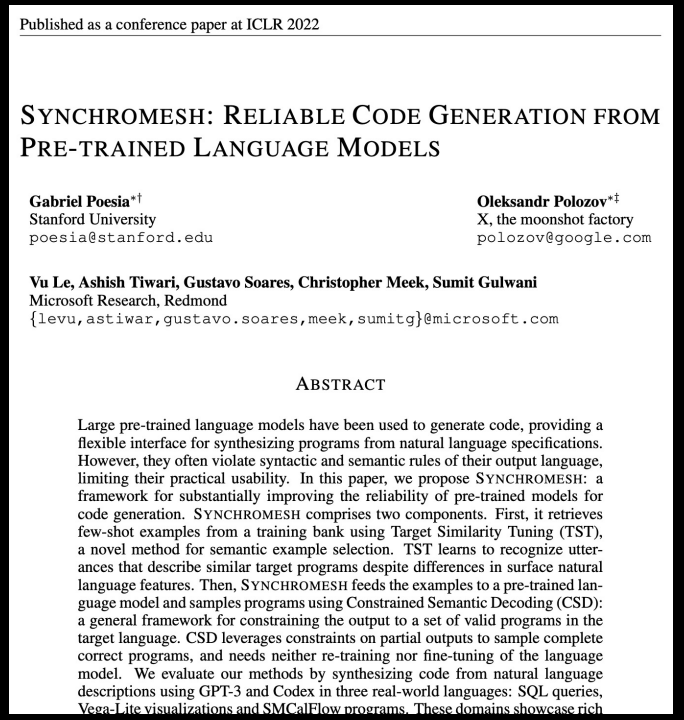
Figure 4: Using BLOOMBERGGPT to generate valid Bloomberg Query Language. Using only a few examples in a few-shot setting, the model can utilize its knowledge about stock tickers and financial terms to compose valid queries to retrieve the data, given a request in natural language. In each case, the model is given 3



Bloomberg
Engineering

State of the Art – neurosymbolic

- Add domain, task, and contextual knowledge to LLMs
- Weave in “common (sense)” knowledge into LLMs
  - “Humans eat food” / “Cake is a type of food” / “Humans eat cake”
  - Use declarative constraints to filter LLM in/outputs
- Generative semi-supervised learning: use generative/large models to generate examples to train a smaller model, or even a more traditional ML pipeline



Published as a conference paper at ICLR 2022

SYNCHROMESH: RELIABLE CODE GENERATION FROM PRE-TRAINED LANGUAGE MODELS

Gabriel Poesia[*†]
Stanford University
poesia@stanford.edu

Oleksandr Polozov[*‡]
X, the moonshot factory
polozov@google.com

Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, Sumit Gulwani
Microsoft Research, Redmond
{levu,astiwar,gustavo.soares,meek,sumitg}@microsoft.com

ABSTRACT

Large pre-trained language models have been used to generate code, providing a flexible interface for synthesizing programs from natural language specifications. However, they often violate syntactic and semantic rules of their output language, limiting their practical usability. In this paper, we propose SYNCHROMESH: a framework for substantially improving the reliability of pre-trained models for code generation. SYNCHROMESH comprises two components. First, it retrieves few-shot examples from a training bank using Target Similarity Tuning (TST), a novel method for semantic example selection. TST learns to recognize utterances that describe similar target programs despite differences in surface natural language features. Then, SYNCHROMESH feeds the examples to a pre-trained language model and samples programs using Constrained Semantic Decoding (CSD): a general framework for constraining the output to a set of valid programs in the target language. CSD leverages constraints on partial outputs to sample complete correct programs, and needs neither re-training nor fine-tuning of the language model. We evaluate our methods by synthesizing code from natural language descriptions using GPT-3 and Codex in three real-world languages: SQL queries, Vega-Lite visualizations and SMCalFlow programs. These domains showcase rich

More Practical Considerations

- PII / GDPR / Right to be Forgotten
- Bias, perspectives, and explainability
  - Bias is a capital (and moral!) risk
  - Many (financial) models need to be explainable for compliance and trust
- Regulatory / compliance
  - Data permissioning: per-role/per-person, store data on- or off-prem
  - Encrypting data at-rest and in-transfer, private data flows, separated networks
- Legacy systems and patchwork processes
  - Allocate time to disentangle and move to modern platforms and architectures

Lessons Learned

- Simpler = better
- Mistakes will be made! Don’t try and plan everything in advance
  - Get products out in front of people and iterate
- Representation learnings
- Buy vs. Build
  - Invest and build from scratch, or partner with vendor(s)?
  - Assemble the right team, with the right mix of skill sets
- Make data veracity a cornerstone
  - Confidence, curation, provenance, human-in-the-loop

TechAtBloomberg.com

Bloomberg
Engineering

Thank You!

Part of https://github.com/laura-dietz/neurosymbolic-representations-for-IR/

Contact me: emeij@bloomberg.net

https://TechAtBloomberg.com/ai
https://TechAtBloomberg.com/data-science-research-grant-program

https://www.bloomberg.com/careers

Bloomberg
Engineering

TechAtBloomberg.com

Agenda

Part 1: Symbolic AI representations and tasks
- (Sub)symbolic AI, and representations
- Foundations for this tutorial
- Question Answering on Knowledge Graphs

Part 2: Text-to-symbols and Ranking
- Neural Text and Graph Representations
- Text-Symbol Alignment and Semantic Annotations
- Entity Representations and Entity Ranking

Part 3: Neuro-symbolic representations for Reasoning
- Reasoning about Relevance
- Neuro Pseudo-Relevance Feedback with Explainability

Part 4: Applications for Neuro-symbolic approaches
- Industry Use Case: Knowledge Search and Discovery
- Panel & Discussion

TechAtBloomberg.com