Developing Selective Question Answering @ Bloomberg

Search Solutions 2022
November 23, 2022

Mohamed Yahya
Software Engineer & Team Lead, Question Answering
AI Engineering Group

TechAtBloomberg.com
What is the Bloomberg Terminal?

350,000+ subscribers

Software that delivers a diverse array of information, news and analytics to facilitate financial decision-making
Bloomberg by the numbers

● 7K+ software engineers
● 200+ data scientists & engineers working on AI/IR/NLP & related problems
● Increased presence in academic research
Outline

- Context: Search & QA @ Bloomberg
- Why QA now?
- Selective Question Answering
- Takeaways and future work
Outline

● Context: Search & QA @ Bloomberg

● Why QA now?

● Selective Question Answering

● Takeaways and future work
Search @ Bloomberg
Search @ Bloomberg

**Table:**

<table>
<thead>
<tr>
<th>Issuer Name</th>
<th>Ticker</th>
<th>Coupon</th>
<th>Maturity Date</th>
<th>Currency</th>
<th>Amt Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>R.1 Sutuzker International Fin.</td>
<td>SUGR</td>
<td>5.125</td>
<td>10/31/2027</td>
<td>EUR</td>
<td>400MM</td>
</tr>
<tr>
<td>R.1 Nidda Healthcare Holding G.</td>
<td>SUGR</td>
<td>3.500</td>
<td>09/30/2024</td>
<td>EUR</td>
<td>479.2MM</td>
</tr>
<tr>
<td>R.1 Landskab Baden-Wurtte...</td>
<td>LBBX</td>
<td>2.750</td>
<td>10/18/2024</td>
<td>EUR</td>
<td>188MM</td>
</tr>
<tr>
<td>R.1 Talanx AG</td>
<td>TALAX</td>
<td>4.000</td>
<td>10/25/2029</td>
<td>EUR</td>
<td>500MM</td>
</tr>
<tr>
<td>R.1 thyssenkrupp AG</td>
<td>TKAGR</td>
<td>2.875</td>
<td>02/22/2024</td>
<td>EUR</td>
<td>1.5MM</td>
</tr>
<tr>
<td>R.1 State of North Rhine-Westph...</td>
<td>NRW</td>
<td>3.000</td>
<td>01/27/2028</td>
<td>EUR</td>
<td>2.5MM</td>
</tr>
</tbody>
</table>

**Selected News Articles:***

- Four Big Market Things to Know That Aren’t the Fed: John Authors
- ECB Needs a Formidable Anti-Fragmentation Tool: Marcus Ashworth
- ECB Delay Presages Bigger Rate Rises Ahead: Marcus Ashworth
- Fitch Revises Techem’s Outlook to Stable, Affirms IDR at ‘B’
- Fitch Assigns PCF GmbH ‘B+’ Final IDR; Outlook Stable
- Fitch Assigns PCF GmbH First-Time ‘B+(EXP)’ IDR; Outlook Stable

**FAQ:**

**In DDIS, why do bonds maturing in other years appear in the securities?**

The list of securities for any maturity year in the Debt Distribution (DDIS) function includes any securities with principal cash flows in that year. For example, a sinkable bond with a 2025 maturity that has principal cash flows in 2022 is included in the list of securities for the 2022 year.

**How do I find the constant maturity yield (CMY) curve for a US Treasury?**

There are two ways to find the CMY curve for US Treasury bonds. To find the CMY for US Treasury bonds in the All Index Sector Display (AIXS) function, from the command line, enter ALXX HIST GO. Note: You can click an index to load it, then run HP QG to view the CMY curve.

**YAS - Workflows - Analyzing Yield/Spread - Changing the Benchmark: Function Help**

YAS launches with a default benchmark. You can change the benchmark to re-price the loaded security and recalculate spreads to a custom bond, swap, treasury curve, or to a local sovereign curve for emerging markets bonds. Steps: select the benchmark.

**YAS - Reference - Calculations - Calculations by Security Type - Bond: Function Help**

This section describes the calculations used in yield analysis for bonds in default. A bond is said to be in default when it fails to fulfill its contractual obligations, such as the delay or failure to make timely payments of interest or principal. Projected...
which drones did russia buy from iran?

Shahed
been found after being shot down. What’s the advantage of using the Shahed drones? As Russia runs down its stockpile of precision-guided missiles and struggles to buy or make replacements while under sanctions, Iran’s drones offer an accurate, mid-range alternative.

What Are Iranian Drones Russia Is Using in Ukraine?: QuickTake
Published: 17/10/2022 | Bloomberg News | NSN>>

Selected News Articles | More »
Al Jazeera: Analysis: What’s behind Iran’s alleged drone deal with Russia? AZ 10/31
Ukraine Latest: Putin Demands Safety Guarantee for Russian Ships BN 10/31
Al Jazeera: Analysis: What’s behind Iran’s alleged drone deal with Russia? AZ 10/31
Ukraine Latest: Russia Says UK Involved in Naval Vessel Strike BN 10/29
US Says Will Make It Harder for Iran to Sell Weapons to Russia BN 10/20
Ukraine Latest: Kyiv Energy Use Limited; New Threat From North BN 10/20

Where can I find information on Russian repurchase agreements? FAQ
You can find information on Russian repurchase agreements for bonds and equities on the Moscow Exchange REPO Monitor (MICR) function. For more on using MICR: MICR Help Page.
Why Productionize QA?

- Make Bloomberg textual sources more easily discoverable
  - News; Financial research; Company filings; Earning call transcripts; Reference material & documentation

- Expressive search to uncover the wealth of knowledge in the Terminal
  - 2.7K+ journalists and analysts writing news and in-depth investment research
  - 2M news articles/day from 125K+ curated third-party news sources

- Testbed for Transformer-based LLMs at Bloomberg
Outline

● Context: Search & QA @ Bloomberg

● Why QA now?

● Selective Question Answering

● Takeaways and future work
Outline

- Context: Search & QA @ Bloomberg
- Why QA now?
- Selective Question Answering
- Takeaways and future work
Why Productionize QA now?
Why Productionize QA now?

QA =

Modeling +

Engineering & Infrastructure → PaaS
Outline

● Context: Search @ Bloomberg

● Why QA now?

● Selective Question Answering

● Takeaways and future work
A Good Answer

Correct at the time presented & explainable

always high precision

recall can wait
Core QA pipeline

- Document Stores
- Retriever
- Re-ranker
- (Extractive) Reader
- Contextualized Answer
Answer Correctness: What Can Go Wrong?

1. Corpus doesn’t contain answer (OOD question or just a “limited” corpus)

2. Imperfect answering pipeline, that’s always trying to answer
Selective Answering

- Document Stores
- Retriever
- Re-ranker
- (Extractive) Reader
- Selective QA Model

Confidence > θ
High confidence

Confidence < θ
Low confidence

Abstain from answering

Iteration = release

Precision

Recall

On the Foundations of Noise-free Selective Classification
Ran El-Yaniv
Yaze Wiener

Selective Question Answering under Domain Shift
Anita Kamath Robin Jia Percy Liang
Computer Science Department, Stanford University
{anitaka, robinjia, pliang}@cs.stanford.edu
Selective Answering

- Core QA pipeline produces **uncalibrated** scores, in general
- Iterate by experimenting with **instantiations of pipeline components**
- **Inject features** not necessarily captured by the core pipeline
Selective Answering

Core QA pipeline

- Document Stores
  - Retriever
  - Re-ranker
  - (Extractive) Reader
  - Selective QA Classifier

Selective QA Classifier

Binary classifier: show/no-show
- Show: query in-domain AND answer correct
- No-show: query out-of-domain OR answer incorrect

Confidence > θ
High confidence

Confidence < θ
Low confidence

Abstain from answering
Let’s talk about data & annotation first…
Evaluation Data Needs - Setting

QA system over a specific corpus deployed in an aggregated search setting

- Need in-domain questions
- Representative of what users will ask
- QA pipeline can still go wrong on in-domain questions
- OOD questions are the norm
Evaluation Data Needs - Setting

SQuAD Setting

{(Question, Reference Answer)}

Core QA pipeline

{(Question, Predicted Answer)}

Selective QA Model Training/Eval

{(0, 1)}
Evaluation Data Needs - Data Sources

(OOD)

In-domain Cold start

{(Question, Reference Answer)}
Evaluation Data Needs - Data Sources

In-domain

{(Question, Reference Answer)}

OOD

Cold start

1

Buildings Decarbonization Funding Lags Other Climate Tech: BNEF

Commercial buildings accounted for 11% of global emissions in 2019, making them a key target for decarbonization. However, the sector has not yet been met with much enthusiasm by the venture capital and private equity (VCPE) community.

Factoid:

11%

Factoid question:

What share of global emissions comes from commercial buildings?
Evaluation Data Needs - Data Sources

{(Question, Reference Answer)}

In-domain (&OOD)

six months

in a refrigerator. AstraZeneca and J&J’s are even easier to handle: they can be transported and stored at normal refrigerator temperatures until their expiry, which is up to six months for AstraZeneca. The Sinovac, Sinopharm and Covaxin vaccines can be stored in refrigerators for up to three years.
Evaluation Setting

- Retriever
- Re-ranker
- (Extractive) Reader
- Selective QA Model

Core QA pipeline

50% Exact Match

Evaluating quality of show/no-show decisions made by selective QA model.

Abstain from answering
<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy (TP+TN)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Iteration 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Lexical features</td>
<td>Crowdsourced in-domain</td>
<td>0.89</td>
<td>0.70</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Logged OOD</td>
<td>N/A</td>
<td>N/A</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Naturally asked (not in training)</td>
<td>0.82</td>
<td>0.23</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Evaluation of Retriever, Re-ranker, (Extractive) Reader, and Selective QA Model.

- **Model Dataset Precision Recall Accuracy (TP+TN)**
  - **Crowdsourced in-domain**: 0.89, 0.70, 0.74
  - **Logged OOD**: N/A, N/A, 0.98
  - **Naturally asked (not in training)**: 0.82, 0.23, 0.51

- **Iteration 2**
  - Features based on Reader embeddings
  - **Crowdsourced in-domain**: 0.83, 0.85, 0.78
  - **Logged OOD**: N/A, N/A, 0.99
  - **Naturally asked (not in training)**: 0.83, 0.54, 0.66
## Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy (TP+TN)</th>
</tr>
</thead>
</table>
| - Lexical features  
(iterative independent of pipeline) | Crowdsourced in-domain | 0.89 | 0.70 | 0.74 |
| | Logged OOD | N/A | N/A | 0.98 |
| | Naturally asked (not in training) | 0.82 | 0.23 | 0.51 |
| - Features based on Reader embeddings | Crowdsourced in-domain | 0.83 | 0.85 | 0.78 |
| | Logged OOD | N/A | N/A | 0.99 |
| | Naturally asked (not in training) | 0.83 | 0.54 | 0.66 |
Outline

- Context: Search @ Bloomberg
- Why QA now?
- Selective Question Answering
- Takeaways and future work
Takeaways

- QA enabled by LLMs + PaaS + MLOps
- Selective QA enables early productionization and rapid iteration
- Most search settings are NOT Web-scale
  a. Rely on crowdsourcing for data
  b. Bootstrap a QA system
  c. Use actual logged data
Ongoing & Future Work

- Temporal dimension of QA
- Discovery for QA
  How do you tell users that you do QA, and what you can/cannot answer
- Snippeting & explanations for QA
- Self-service QA (QAaaS)
We Are Hiring!

Senior IR Research Engineer - Artificial Intelligence Group
London
Posted Nov 8, 2022 - Requisition No. 111378

Senior IR Research Engineer - Artificial Intelligence Group
New York, NY
Posted Nov 8, 2022 - Requisition No. 111377

Senior NLP Research Engineer - Artificial Intelligence Group
London
Posted Nov 8, 2022 - Requisition No. 111383

Senior MLOps Engineer - Artificial Intelligence Group
London
Posted Nov 8, 2022 - Requisition No. 111384

Data Engineer - Artificial Intelligence Group
New York, NY
Posted Oct 5, 2022 - Requisition No. 110069

https://careers.bloomberg.com/
Thank you!

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

https://www.bloomberg.com/careers

Contact me: myahya6@bloomberg.net