Towards Knowledge-Intensive Text-to-SQL Semantic Parsing with Formulaic Knowledge

EMNLP 2022 Main Conference

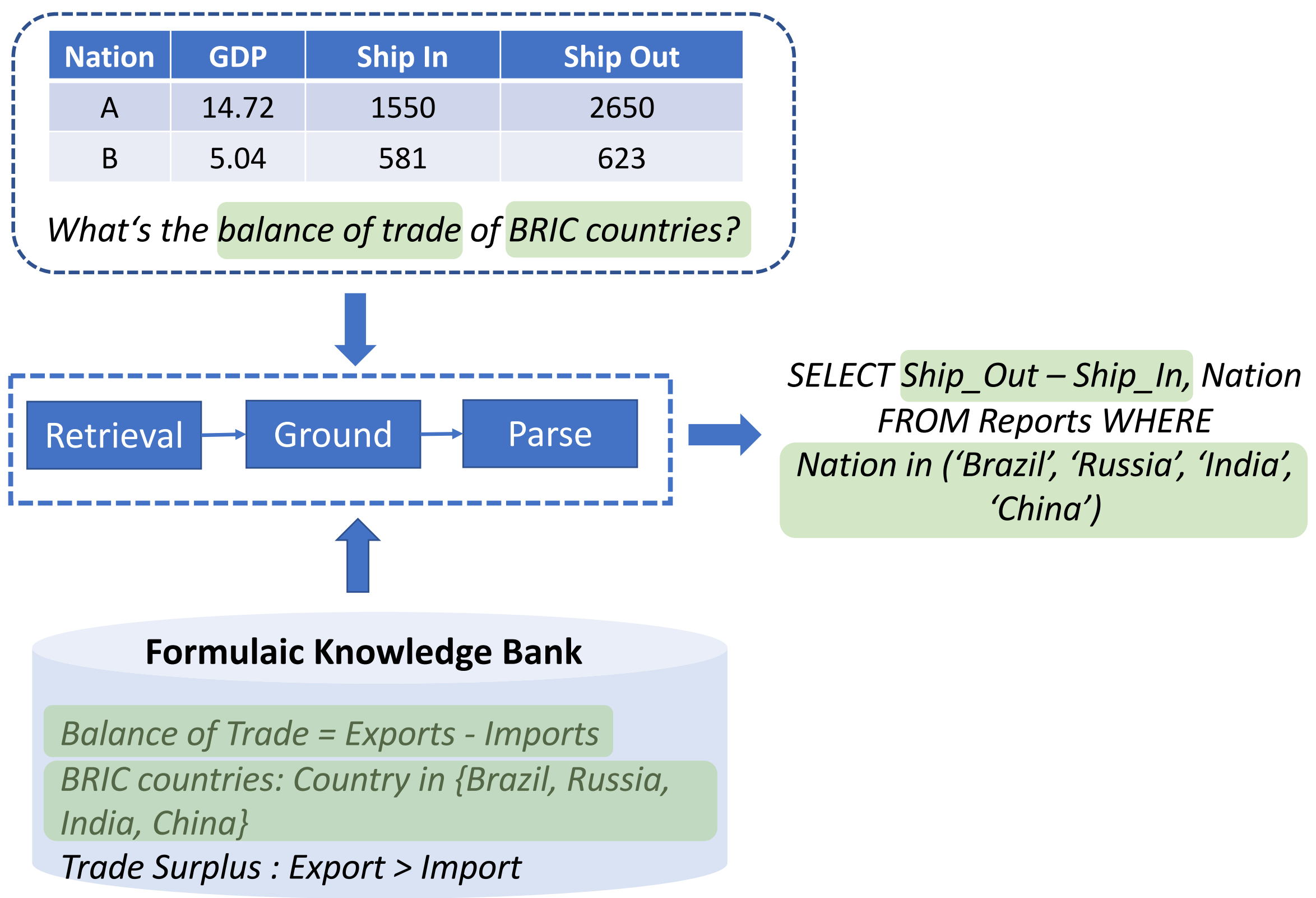
Longxu Dou[1], Yan Gao[2], Xuqi Liu[1], Mingyang Pan[1], Dingzirui Wang[1], Wanxiang Che[1], Min-Yen Kan[3], Dechen Zhan[1], Jian-Guang Lou[2]

[1]Harbin Institute of Technology [2]Microsoft Research Asia [3]National University of Singapore

❖ Introduction: Problem Definition & Traditional Solution

- **Knowledge-Intensive Text-to-SQL**
- In the ***professional*** application of text-to-SQL, such as in the ***data analysis of financial reports***, models require external knowledge to map the expert query with the domain-specific database.

- **Traditional Solution (Data-Centric)**
- Annotating more data pairs on a target domain. Then such mappings are induced during the training process.
- However, it is **fragile** and **expertise-heavy**. Such knowledge does not port across domains and requires expert knowledge to craft.



Nation	GDP	Ship In	Ship Out
A	14.72	1550	2650
B	5.04	581	623

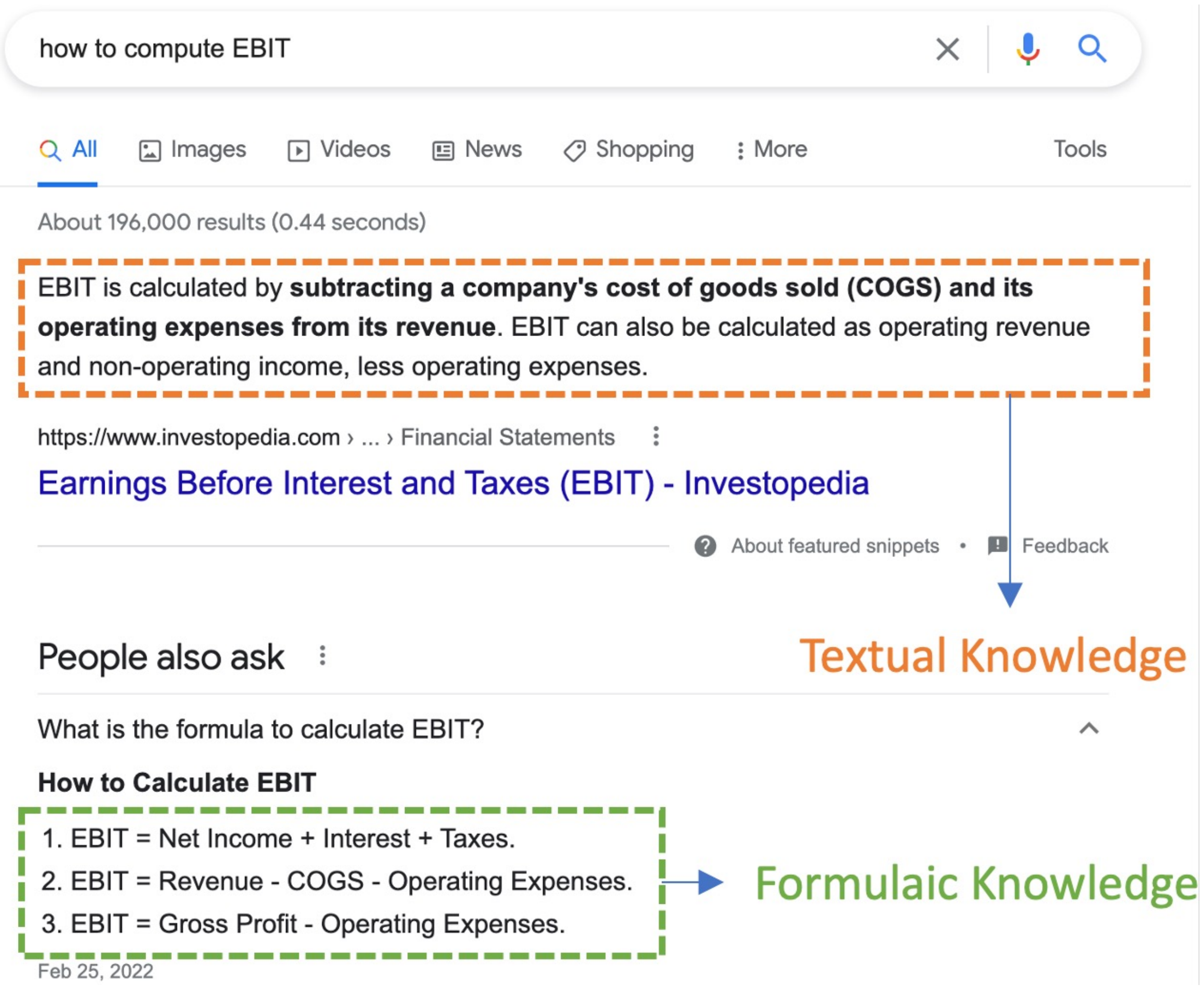
What's the balance of trade of BRIC countries?

SELECT Ship_Out – Ship_In, Nation FROM Reports WHERE Nation in ('Brazil', 'Russia', 'India', 'China')

Formulaic Knowledge Bank

Balance of Trade = Exports - Imports
BRIC countries: Country in {Brazil, Russia, India, China}
Trade Surplus : Export > Import

❖ Approach: Formulaic Knowledge & ReGrouP Architecture



how to compute EBIT

About 196,000 results (0.44 seconds)

EBIT is calculated by **subtracting a company's cost of goods sold (COGS) and its operating expenses from its revenue**. EBIT can also be calculated as operating revenue and non-operating income, less operating expenses.

https://www.investopedia.com › ... › Financial Statements

Earnings Before Interest and Taxes (EBIT) - Investopedia

Textual Knowledge

People also ask

What is the formula to calculate EBIT?

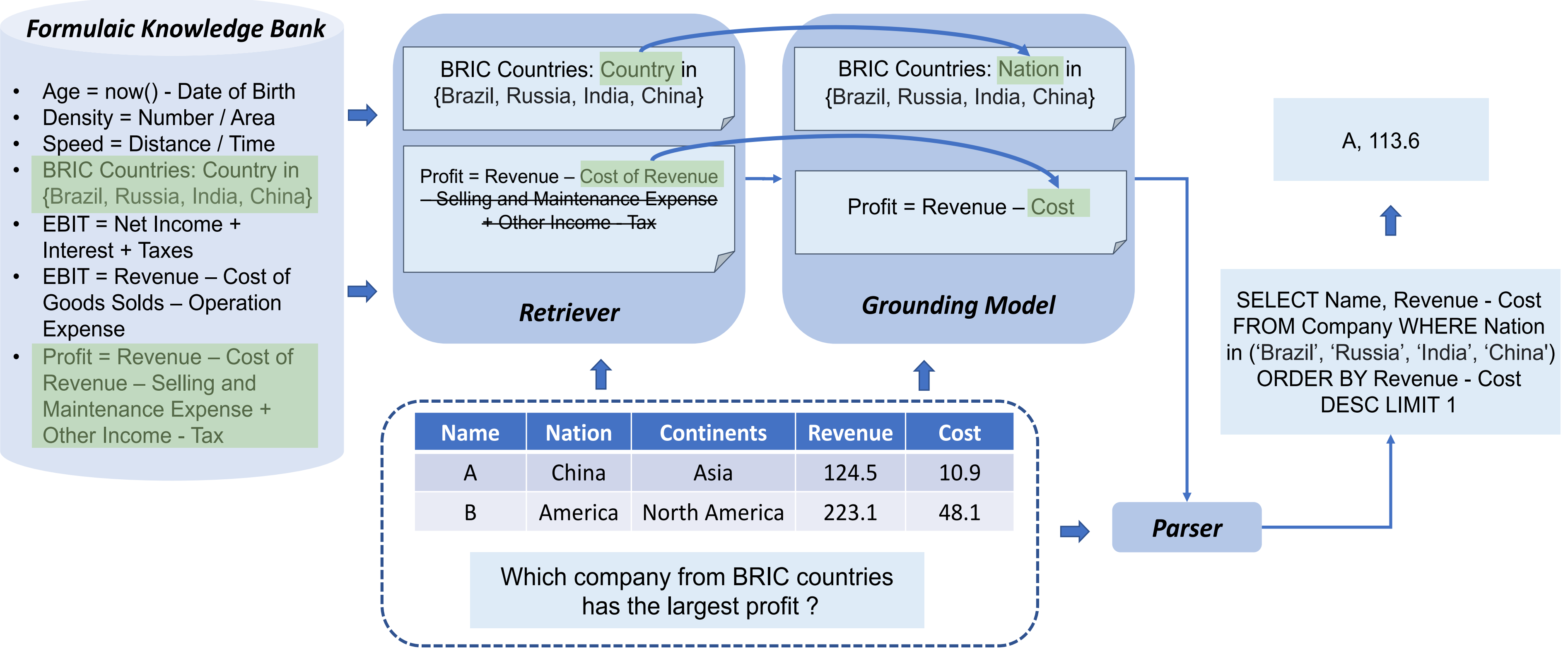
How to Calculate EBIT

1. EBIT = Net Income + Interest + Taxes.
2. EBIT = Revenue - COGS - Operating Expenses.
3. EBIT = Gross Profit - Operating Expenses.

Feb 25, 2022

Formulaic Knowledge

Motivation: When meeting unseen terminology, the human may first search the related mathematical knowledge or domain knowledge from textbooks or encyclopedias. **(Knowledge-Centric)**



Formulaic Knowledge Bank

- Age = now() - Date of Birth
- Density = Number / Area
- Speed = Distance / Time
- BRIC Countries: Country in {Brazil, Russia, India, China}
- EBIT = Net Income + Interest + Taxes
- EBIT = Revenue – Cost of Goods Solds – Operation Expense
- Profit = Revenue – Cost of Revenue – Selling and Maintenance Expense + Other Income - Tax

BRIC Countries: Country in {Brazil, Russia, India, China}

Profit = Revenue – Cost of Revenue ~~– Selling and Maintenance Expense + Other Income – Tax~~

Retriever

BRIC Countries: Nation in {Brazil, Russia, India, China}

Profit = Revenue – Cost

Grounding Model

Name	Nation	Continents	Revenue	Cost
A	China	Asia	124.5	10.9
B	America	North America	223.1	48.1

Which company from BRIC countries has the largest profit ?

Parser

SELECT Name, Revenue - Cost FROM Company WHERE Nation in ('Brazil', 'Russia', 'India', 'China') ORDER BY Revenue - Cost DESC LIMIT 1

A, 113.6

Model Architecture: ReGrouP

(1) **Retrieve** the formulaic knowledge item from the bank;

(2) **Ground** the concept of formulaic knowledge into schema elements;

(3) **Parse** the question with grounded formulaic knowledge into SQL.

Advantage: Knowledge-extensible without re-training the model.

❖ Experiment: Main Results & Case Studies

Model	Dev	Finance	Estate	Transportation	Average
Vanilla	69.3	8.7	5.7	6.9	22.7
REGROUP (w/o Grounding)	71.7	38.1	25.1	32.7	41.9
REGROUP	74.6	43.7	46.1	39.1	50.9
REGROUP (Oracle)	78.4	71.4	84.8	64.7	74.8

(1) ReGrouP exceeds the vanilla model by **28.2%**, which indicates the effectiveness of using formulaic knowledge;

(2) Grounding the formulaic knowledge improves the model by **9.0%**.

Future work

(1) Iterative filling in the blank of formulaic knowledge bank;

(2) Mitigating the gap between formulaic knowledge and specific schema via improving the grounding model;

(3) Driving the parser to fully make use of more complicated (e.g., commonsense) formulaic knowledge.

Vanilla Model Error	*Formulaic Knowledge*
Question: 东三省每省的一胎出生率是多少? (What is the first birth rate in each of the three northeastern provinces in China?) Schema : 省份 ｜ 婴儿出生率 | 二胎出生率 | 人口 (Province | Birth Rate | Second Birth Rate | Population) Vanilla: SELECT 婴儿出生率 FROM 各省人口出生及死亡率 WHERE 省份 = "辽宁" ReGrouP: SELECT 婴儿出生率 - 二胎出生率 FROM 各省人口出生及死亡率 WHERE 省份 IN ("辽宁" , "吉林" , "黑龙江")	**Grounded Formulaic Knowledge:** 东三省 : { 辽宁 , 吉林 , 黑龙江 } (Three Northeastern Provinces: { Liaoning , Jilin , Heilongjiang }) 一胎出生率 = 婴儿出生率 - 二胎出生率 (First birth rate = Birth rate - Second Birth Rate)
Retriever Error (43%)	***Retrieval Knowledge***
Question:息税前利润是多少? (Please return the Earnings Before Interest and Taxes) **Schema**: 收入 | 净收入 | 销售费用 |营业费用 ｜销售额 (Revenue| Net Income | Cost of Goods Sold Expenses | Operating Expenses | Sales) Gold SQL: SELECT 收入 - 销售费用 - 营业费用 FROM 报表 Pred SQL: SELECT 净收入 + 销售额 FROM 报表	Oracle Formulaic Knowledge: 息税前利润 = 收入 - 销售成本 - 营业费用 (Earnings Before Interest and Taxes = Revenue – Cost of Goods Sold – Operating Expenses) Retrieved Formulaic Knowledge: 息税前利润 = 净收入 + 利息 + 税 (Earnings Before Interest and Taxes = Net Income + Interest + Taxes)
Grounding Error (41%)	***Grounded Knowledge***
Question: A公司的流动资产是多少? (What is company A's current assets?) **Schema**:现金 | 应收款项 | 可销售证券|商品成本| 运营费用 (Cash | Trade Receivables | Marketable Securities | Cost of Goods | Operating Expenses) Gold SQL: SELECT 应收款项 + 可销售证券 +现金 FROM 报表 Pred SQL: SELECT 应收款项 + 现金 FROM 报表	**Undergrounded Formulaic Knowledge:** 流动资产 = 短期资本 + 应收帐款 + 股票 + 存款余额 (Current Assets = Short Term Capital + Debtors + Stock + Cash and bank) Correct Grounded Formulaic Knowledge: 流动资产 = 应收款项 + 可销售证券 + 现金 (Current Assets = Trade Receivables + Marketable Securities + Cash) Prediced Grounded Formulaic Knowledge: 流动资产 = 应收款项 + 现金 (Current Assets = Trade Receivables + Cash)
Parser Error (12%)	***Leveraging Knowledge***
Question: 哪个城市的房地产市场发展合理? (Which city's real estate market is developing reasonably?) **Schema**: 城市 | 吸纳率 | 空置率 (City | Commercial Housing Absorption Rate | Commercial Housing Vacancy Rate) Gold SQL: SELECT 城市 FROM 报表 where 空置率 > 15% and 空置率 < 30% Pred SQL: SELECT 城市 FROM 报表 where 空置率 > 15%	**Grounded Formulaic Knowledge:** 房地产市场良性发展 : 空置率 > 15% AND 空置率 < 30% (Good development of real estate market: Commercial Housing Vacancy Rate > 15% AND Commercial Housing Vacancy Rate < 30%)

Paper Code

Longxu Dou
lxdou@ir.hit.edu.cn
https://longxudou.github.io/