Putting Trade Flows Into Context

Simon Blöthner, Durgesh Nandini,
Mario Larch, Mirco Schönfeld

July 25, 2025

Abstract

Contextual, often qualitative data are arguably the most important influences on economic action. However, they are notoriously difficult to capture in empirical economic models. This is most prevalent when modeling international trade flows that are the result of unaccountably many individual decisions, each influenced by subjective experiences. To include these determinants, we rely on recent advances in large language models (LLMs) and knowledge graphs to capture the richness of these types of data, without the assumptions usually required to fit them into quantitative models. We present a methodology for extracting structured trade-related knowledge from unstructured natural language texts of regional trade agreements using LLMs. This allows us to gauge the relevance of provisions made in bilateral agreements and open up new dimension of possible inquiry into the nature of bilateral trade flows.

JEL Classification Codes: C45, C54, C55, F13, F17

Keywords: International Trade, Machine Learning, Knowledge Graphs, Trade Agreements

1 Introduction

The use of the market system of voluntary exchange comes with costs outside of those associated with the production of the goods and services advertised in them. Munger (2022) identifies these costs of doing business or transaction costs as:

- **triangulation**: finding and identifying of trading partners
- **transfer**: transferring of goods/services and the payments for them
- **trust**: credible enforcement mechanism of the agreed to contracts

Naturally, as markets grow in size and knowledge about potential trading partners decreases once individual firms and consumers have to interact with others outside of their country, culture, or local networks, these costs tend to increase. This means that nominally mutually beneficial exchanges do not occur, leaving market participants worse off. Problems like this are most prevalent in the context of international trade, where transaction costs tend to be at their highest, meaning that many of the benefits of international specialization remain unrealized. Bilateral trade agreements aim to decrease the transaction costs, by laying a foundational set of rules upon which trading partners can rely. But what exactly in these agreements is it that allow for less costly international trade? The precise content and its formulation can vary widely across agreements, meaning that merely treating it as a binary variable misses a large part of the important information contained in it. Furthermore, an agreements signed between two countries might have a different impact in the context of another pair. However, including this heterogeneity for inference in conventional empirical methods, as well as more flexible machine learning methods has so far been difficult, due to the qualitative and context dependent nature of the data. To address this, we rely on recent advances in large language models (LLMs) and knowledge graphs (KGs), both of which excel at the representation of data in unstructured formats, unlike conventional methods, which are limited to data structured in tabular formats. To explore the extent of this heterogeneity this we will rely on two novel methodologies. Firstly, we follow Nandini et al. (2024) in structuring our data as a KG, which represents entities and their relationships as nodes and edges (Ehrlinger and Wöß 2016; Hogan et al. 2021). In our case these entities could be countries and their relationships the bilateral trade flows between them. Secondly, we employ LLMs, built on deep learning architectures like the Transformer (Vaswani et al. 2017) that excel at extracting meaning from unstructured text, such as legal documents, by modeling relationships between words and phrases (Brown et al. 2020). These allows us to easily extract textual information from the actual trade agreements relevant to the respective bilateral relationships. This combination allows us to attribute relevance to particular provisions and clauses in the context of the countries and their trading partners, providing a rich insight into the nature and impact of the coming into force of an agreement.

We follow this introduction with an overview an overview of our data in Section 2, introduce our methodology in Section 3, which we follow with our results in Section 4 before concluding in Section 5.

2 Data

We use bilateral trade flow data from Borchert et al. (2021), which we limit to the years 1986 to 2016. Additional information comes from Gurevich and Herman (2018) in the form of gross domestic product and population, geographic distance between two countries from Mayer and Zignago (2011), as well as binary information about regional trade agreements from Egger and Larch (2008). This leaves us with over 2.5 million observations across the entire time frame. In addition to the standard measures, we require context-specific information about the contents of the trade agreements. For this we use the World Trade Organization's RTA data base and its machine readable form developed in Alschner, Seiermann, and Skougarevskiy (2017). This consists of 450 XML files containing meta data as well as the full text of the agreement between countries. [1] To get an overview of the of the data, we show a word cloud for the words contained in the agreements, which highlights the most frequently used words across all the agreements in the data set. This reveals important, but also widely varying concepts, that once implemented, could reduce the transaction costs related to international trade.

3 Methodology

Investigating bilateral trade flows through the lens of networks and graphs is perhaps the most natural way of looking at the process of exchange. Our nat-

[1] The full data set can be access here: https://github.com/mappingtreaties/tota.

Figure 1: Word cloud of the most frequent words across all contracts.

ural language speaks of trading hubs, clusters and nodes of economic activity. Descriptions like the decentralization of trade only make sense when looking at the connected nature of economic agents. Unlike the dyadic structure common to and required by conventional econometric models, which largely restricts the analysis to bilateral effects, graph-based methods allow for and perform best when looking at interactions between neighbors and how they propagate through the network (Newman 2018; Jackson 2008). Conventionally, this means that countries are represented as nodes, while their relations (usually trade flows) are the edges connecting these. Looking at international trade flows in the framework of graphs is not new to economics (De Benedictis and Tajoli 2011; Basile et al. 2018). However, these efforts were limited to the qualitative analysis of the structure of the underlying networks, not allowing for the inferential analysis of flows which conventional empirical regression-based methods excel at.

Advances in machine learning techniques, especially in the realm of (graph) embeddings, have made possible the the use of inferential methods in the framework of graphs, which we apply to networks of international trade flows. To make use of this we translate our bilateral, structured data into a KG of triples, which can look very distinct from the commonly used graph structure in network analysis. A KG is a way to add information and context to conventional graphs (Ehrlinger and Wöß 2016; Hogan et al. 2021). Following Nandini et al. (2024) we model the countries as nodes, which have edges to the different pieces of context information that we wish to add to it. The triples allow us to represent this context information in a subject, predicate, object structure. As an example, if we want to represent the relationship between the United States and Canada with a trade volume of y, we would create a triple for the dyad *USA_CAN has_volume y*, which links the United States and Canada in a directed manner. We repeat the process for all of the foundational economic data outlined in Section 2. To add the information contained in the texts of the trade agreements, we first have to extract the triple relationships from the agreements. For this task we make use of an LLM, in particular the Llama 3.1 model, which, when prompted and fed the texts, returns triple structures for an agreement. As an example, we could see an agricultural agreement linking Japan and Thailand, where both agree to adopt common commodity labeling standards be represented as *Japan adopts common_standard_for_commodities*, linking Japan (and conversely as Thailand) to the node of adoption of common labeling standards for commodities.

Once our KG is thus populated, we transform it into an embedding, which is a representation of the entities and their relations in a continuous vector space, to make it tractable for downstream learning algorithms (Cai, Zheng, and Chang 2018; Goyal and Ferrara 2018; Wang et al. 2017). These translate the embeddings into lower dimension spaces, where similarities and generalities can be learned. After training our model, we are then able to make inferential statements about edges not in the data.

4 Results

5 Conclusion

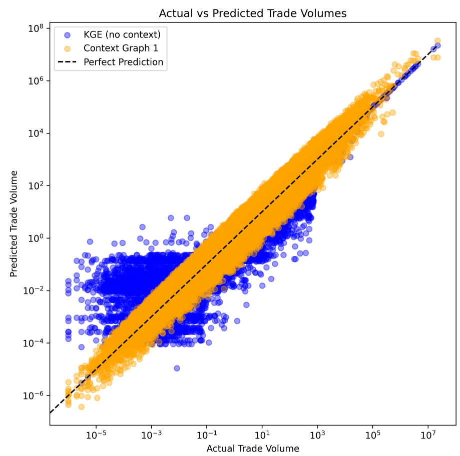

Figure 2: Comparison of the prediction performance of the KGE without RTA context information against the model with the text embeddings. The plot shows the actual trade flows against their prediction by our two models on a logarithmic scale. A perfect fit is indicated by the dotted line.

References

Alschner, Wolfgang, Julia Seiermann, and Dmitriy Skougarevskiy (2017). "Text-as-data analysis of preferential trade agreements: mapping the PTA landscape". In: *Centre for Trade and Economic Integration*.

Basile, Roberto, Pasquale Commendatore, Luca De Benedictis, and Ingrid Kubin (2018). "The impact of trade costs on the European Regional Trade Network: An empirical and theoretical analysis". In: *Review of International Economics* 26.3, pp. 578–609.

Borchert, Ingo, Mario Larch, Serge Shikher, and Yoto V Yotov (2021). "The international trade and production database for estimation (ITPD-E)". In: *International Economics* 166, pp. 140–166.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Cai, Hongyun, Vincent W Zheng, and Kevin Chen-Chuan Chang (2018). "A comprehensive survey of graph embedding: Problems, techniques, and applications". In: *IEEE Transactions on Knowledge and Data Engineering* 30.9, pp. 1616–1637.

De Benedictis, Luca and Lucia Tajoli (2011). “The world trade network”. In: *The World Economy* 34.8, pp. 1417–1454.

Egger, Peter H. and Mario Larch (2008). “Interdependent Preferential Trade Agreement Memberships: An Empirical Analysis”. In: *Journal of International Economics* 76.2, pp. 384–399.

Ehrlinger, Lisa and Wolfram Wöß (2016). “Towards a definition of knowledge graphs.” In: *SEMANTiCS (Posters, Demos, SuCCESS)* 48.1-4, p. 2.

Goyal, Palash and Emilio Ferrara (2018). “Graph embedding techniques, applications, and performance: A survey”. In: *Knowledge-Based Systems* 151, pp. 78–94.

Gurevich, Tamara and Peter Herman (2018). “The dynamic gravity dataset: 1948-2016”. In: *US International Trade Commission, Office of Economics Working Paper*.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. (2021). “Knowledge graphs”. In: *ACM Computing Surveys (Csur)* 54.4, pp. 1–37.

Jackson, Matthew O. (2008). *Social and Economic Networks*. Princeton University Press. ISBN: 9780691134406. URL: `http://www.jstor.org/stable/j.ctvcm4gh1` (visited on 07/25/2025).

Mayer, Thierry and Soledad Zignago (2011). “Notes on CEPII’s distances measures: The GeoDist database”. In: .

Munger, Michael C (2022). “Giants among us: do we need a new antitrust paradigm?” In: *Constitutional Political Economy* 33.4, pp. 445–460.

Nandini, Durgesh, Simon Blöthner, Mirco Schoenfeld, and Mario Larch (2024). “Multidimensional Knowledge Graph Embeddings for International Trade Flow Analysis”. In: *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KEOD*. INSTICC. SciTePress, pp. 63–73. ISBN: 978-989-758-716-0. DOI: `10.5220/0013028500003838`.

Newman, Mark (July 2018). *Networks*. Oxford University Press. ISBN: 9780198805090. DOI: `10.1093/oso/9780198805090.001.0001`. URL: `https://doi.org/10.1093/oso/9780198805090.001.0001`.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.

Wang, Quan, Zhendong Mao, Bin Wang, and Li Guo (2017). “Knowledge graph embedding: A survey of approaches and applications”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.12, pp. 2724–2743.