Paper Analysis

Mapping and Modeling the semantic space of Math concepts
Debray and Dahaene, 2024

Dhia Garbaya

October 2024

Abstract

This document presents a comprehensive analysis of the paper (Debray and Dehaene, 2024) by Samuel Debray and Stanislas Dehaene. The analysis covers the summary of the paper, research questions and proposed solutions, key observations, implementation results, insights, potential improvements, conclusions, and a personal extension using modern embedding models.

Contents

1 Summary

The paper (Debray and Dehaene, 2024) presents an extensive exploration of the semantic space of mathematical concepts. Recognizing the under-explored nature of mathematics within cognitive studies, that generally focus on elementary areas like numbers, fractions, and geometric shapes.
The authors constructed a 1000-word vocabulary of math terms from primary education to advanced university-level topics.
Using specifically the GloVe algorithm (Pennington et al., 2014), the study evaluates how well these models align with human judgments of familiarity among math concepts.
Further, they investigate the influence of educational background and level on these semantic representations, giving insights into how mathematical knowledge structures evolve with education.

2 Research Questions

The paper mainly addresses two questions:

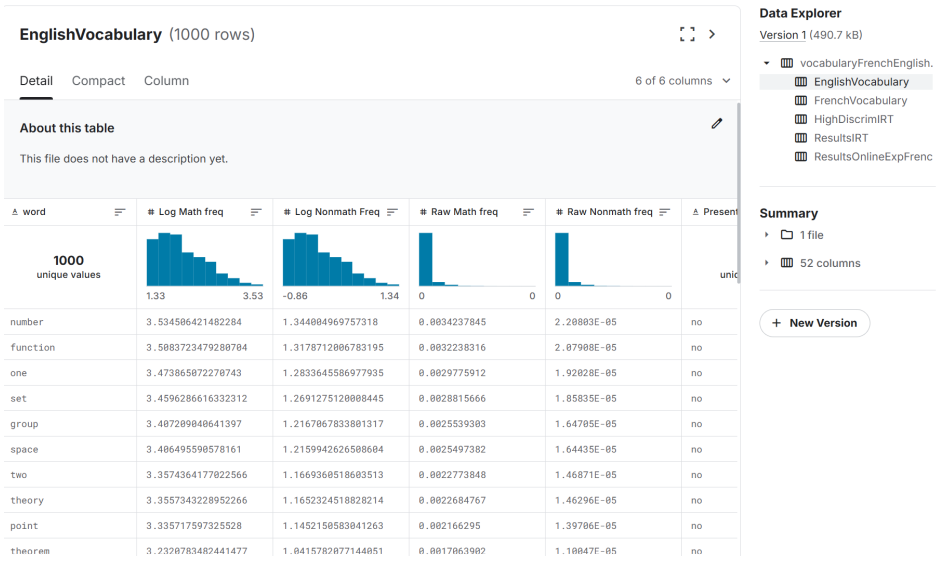
1. **Model fit to Human Judgments:** Can distributed semantic models, such as GloVe, provide a good representation of human judgments in the domain of mathematics?
2. **Academic level impact:** How does the fit between these semantic models and human judgments vary with the level of mathematical education of the participants?

3 Proposed Solution

To address the research questions, the authors propose the following methodology:

3.1 Vocabulary Construction

- **Corpus Selection:** The study begins by identifying mathematical content in French and English Wikipedia pages and extracting the 1000 most frequent mathematical terms in both languages.



- **Embedding Generation:** Using GloVe, they embed these terms *across three distinct corpora*:
  1. **Math Corpus:** only *math-related* pages.
  2. **Non-Math Corpus:** all *other* Wikipedia pages excluding.
  3. **Global Corpus:** a concatenation of the above two corpuses.

The idea of this separation is to avoid the impact non-math meanings of some concepts on the embedding. For ex. the word 'ball' will potentially loose its topoligical meaning and be similar to football (non-math meaning) when trained on the global corpus. Especially given the quite small embedding dimension used in the paper (50) likely unable to efficiently encode both meanings.

- **Additional Metadata:** for each word, they annotate frequency and grade of acquisition based on the French national curriculum, grammatical category, and polysemy.

3.2 Behavioral Data Collection

- An online experiment involving 1230 participants was conducted.
- **Familiarity Ratings:** Participants rated their familiarity with a subset of 429 words on a 9-level Likert scale.
- **Similarity Ratings:** Participants judged the semantic similarity between 3756 pairs of these words on a scale from 0 to 5.
- **Educational Assessment:** Were also reported the highest levels of mathematical education and the current mathematical skills of each participant.

3.3 Data Analysis

- **Familiarity Prediction:** Using linear regression, the study examines how word grade and word frequency predict familiarity ratings, modulated by participants' education levels.
- **Item Response Theory (IRT):** To refine the understanding of word difficulty and discrimination with participants' skills, IRT was used, particularly for a curated subset of 80 high-discriminability items (Lord, 1980).
- **Semantic Similarity Evaluation:** The correlation between human similarity ratings and GloVe similarities was estimated using Spearman's rank-order correlation.
- **Visualization:** Techniques like spectral clustering and t-SNE are utilized to visualize the high-dimensional semantic space, leading to clusters of different mathematical domains such as analysis, algebra, geometry.

3.4 Replication and Extension

The entire methodology is replicated in English to ensure the robustness of the findings across languages.

4 Observations and Ideas

4.1 Distributed Semantic Models and Human Judgments

- **Correlation results:** The study finds that GloVe embeddings account for approximately half of the variance in human similarity judgments (Spearman's $r_S \approx 0.66$ for French and ≈ 0.48 for English), indicating a significant but not complete alignment between semantic models and human cognitive representations of math concepts.
- **Dimensionality:** The optimal dimensionality of GloVe embeddings to predict human similarities plateaus at 50 dimensions, a compromise between complexity and performance. This was derived using explained variance.
- **Corpus impact:** Embeddings derived from the math-specific corpus outperform those from non-math and global corpora in predicting human judgments, especially for higher-grade and more advanced mathematical concepts. Which I find very **interesting and intuitive**.

4.2 Influence of Education on Semantic Representations

- **Familiarity Ratings:** Higher levels of mathematical education correlate with increased familiarity between words, especially those with higher grades of acquisition. Participants with advanced education show flatter familiarity curves, suggesting a good understanding of mathematical vocabulary.
- **Model Fit Across Education Levels:** The predictive performance of math embeddings improves with the participants' education level, confirming that academic level shapes the semantic representation of mathematical concepts.

4.3 Semantic Space Visualization

- **Cluster Identification:** The semantic space organizes math concepts into distinct clusters corresponding to different mathematical domains.
  For example, numbers form a certain cluster, and geometry terms occupy separate ones, aligning with neuroscientific findings of different cortical activations for different math domains (Amalric and Dehaene, 2016).
- **Numerical Magnitude Representation:** Embeddings of numbers exhibit a quasi-logarithmic structure, reflecting the mental number line observed in cognitive neuroscience (Dehaene, 1997). This confirms that GloVe captures aspects from numbers' cognition.
- **Cross-Concept Relationships:** The embeddings reveal meaningful relationships between concepts, such as the similarity between "three" and "triangle," indicating an underlying geometric association.
  I confirmed this in my code extension with modern embeddings.

5 Implementation and Results

Main implementation aspects include:

5.1 Vocabulary Curation

A manual review ensured the inclusion of relevant mathematical terms solving issues like polysemy.

5.2 GloVe Embeddings

By creating separate embeddings for math, non-math, and global contexts, the study accounts for polysemous words.

5.3 Online Experiment Design

The online experiment was well-structured, giving relevant word pairs according to the participant level.

5.4 Predictive Performance

- **Familiarity Ratings:** The regression models predict familiarity based on word grade and frequency. And indicates that education modulates the impact of these factors.
- **Similarity Ratings:** GloVe's cosine similarity is a better predictor of human similarity judgments than Euclidean distance, though both show good correlations, aligning with what I observed in my extension analysis. The fit improves with higher education levels.

5.5 Semantic Space Insights

- **Clustering Validation:** The t-SNE visualizations and spectral clustering effectively discern distinct mathematical domains, substantiating the cognitive segregation of math fields.
- **Number Concept Organization:** The logarithmic representation of numbers and their multiplicative relationships in the embeddingsshow that distributed semantics capture numerical cognition aspects.

6 Insights

6.1 Cognitive Representation of Math

The study shows that math concepts are semantically organized in a structured space that aligns partially with computational models.
The quasi-logarithmic representation of numbers and the clear clustering of math domains reflect cognitive structures.

6.2 Educational Impact on Semantic Structures

The modulation of familiarities and similarities by educational background shows the variable nature of mathematical semantics.
As individuals progress through education, their semantic network of mathematical concepts becomes more refined and expansive.

6.3 Modeling in Cognitive Neuroscience

This partial success of GloVe embeddings in predicting human judgments encourages to continue exploring representation models.
The remaining unexplained variance shows the necessity for better models or the need to add factors.

7 Potential Improvements

7.1 Enhancing Semantic Models

- **Advanced Embedding Techniques:** Exploring embeddings from newer models like transformer-based architectures (e.g., BERT (Devlin et al., 2018), Llama-3 (Dubey et al., 2024) could capture more nuanced semantic relationships and reduce the unexplained variance in human judgments.
- **Domain-Specific Training:** Training such embeddings on math corpora, such as textbooks or research papers present in open math datasets, might better capture the semantics of advanced math concepts.

7.2 Vocabulary Expansion and Refinement

- **Inclusion of Multi-Word Expressions:** Mathematical concepts often comprise multi-word terms (e.g., "vector space," "topological manifold"). Incorporating these into the vocabulary, possibly using phrase embeddings or composite vectors, could provide a more holistic semantic map.
- **Cross-Linguistic semantics:** Extending the study to include more languages would test the universality of the semantic structures.

8 Conclusion

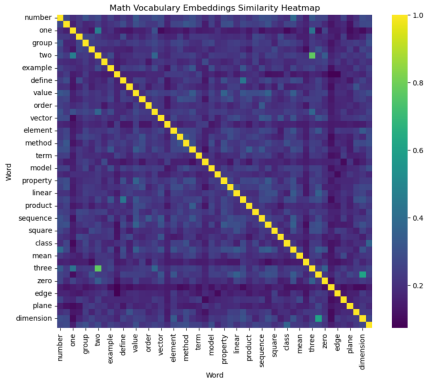
Debray and Dehaene's study provides a foundational framework for mapping the semantic space of mathematical concepts showing that these NLP techniques can serve neuroscience research. These insights can then turn out to be useful to improve the way we encode language.

9 Personal code extension

Given the availability of the authors' code and data, I thought it would be better and more effective to analyze modern embeddings, i.e those used in pretraining state-of-the-art decoder-only transformers, to evaluate their capacity to represent mathematical semantics compared to GloVe. Below is a summary of this extension and the preliminary observations:

9.1 Implementation Steps

1. **Embedding Model Setup:**
   - Loaded GPT-2 Large (770M parameters) and utilized the GPT-2 Byte Pair Encoding (BPE) fast tokenizer from HuggingFace.
   - Saved the embedding layer to streamline subsequent analyses.
2. **Pre-analysis:**
   - I had to write a utility function to ensure that mathematical terms are represented as single tokens by gpt2 tokenizer, to decide which concepts am I going to include in the analysis.
3. **Numbers:**
   - Analyzed embeddings of numbers from 1 to 100, noting that gpt-2 effectively represents each number as a single token.
   - **Patterns Observed:**
     - **Proximity Similarity:** Numerically close numbers exhibit higher similarity, aligning with intuitive mathematical relationships.
     - **Multiplicative Links:** Numbers that are multiplicatively related (e.g., 25 and 75) show significantly higher similarity, suggesting the model captures underlying arithmetic relationships.
     - **Frequency Effect:** Commonly used numbers (e.g., multiples of five and ten) have higher similarity, likely due to their frequent co-occurrence in training data.
     - **Contextual Influence:** Numbers embedded within diverse contexts (e.g., persons' ages, temperatures) have similarities influenced by non-math usage.
4. **Math Vocabulary Analysis:**
   - Extracted the English math vocabulary provided by the authors and selected the 50 most frequent single-gpt2-token words for analysis.
   - **Observations:**
     - **Semantic Closeness:** Words like "function" and "method," "define" and "definition" show high similarity, reflecting grammatical and semantic relationships.
     - **Numerical Semantics:** Primitive numbers like "one," "two," and "three" cluster together, indicating the model's grasp of basic numerical semantics.
     - **Conceptual Associations:** Terms like "three" and "dimension".



– An extension (not done yet) could be to use Latent Dirichlet Allocation (LDA) to extract topics.

9.2 Interpretation of Results

The analysis indicates that GPT-2 embeddings capture several meaningful relationships within math semantics, particularly for frequently used terms and numbers.
However, the model's pretraining on web-based data introduces challenges in distinguishing mathematical contexts from other usages, leading to mixed semantic effects (e.g., "37" relating to both numbers and temperatures).

9.3 Potential Next Steps

- **Clustering and Advanced Visualization:** Implement clustering algorithms.
- **Exploration more capable Models:** Utilize more recent and larger models like LLaMA 3.1 3B, which may offer richer vocabularies and improved math representation. Math full fine-tunes are also good candidates.
- **Tokenization:** Address the issue of multi-token representations for complex mathematical. Another point was made in the notebook remarks.

The extension using GPT-2 embeddings confirms that modern language models possess a foundational understanding of mathematical semantics, eventhough influenced by their context-rich training data.
Looking at context-specific representations is a nice idea to explore.

References

Amalric, M. and Dehaene, S. (2016). Origins of the brain networks for advanced mathematics in expert mathematicians. *Proceedings of the National Academy of Sciences*, 113:201603205.

Debray, S. and Dehaene, S. (2024). Mapping and modeling the semantic space of math concepts. *preprint*.

Dehaene, S. (1997). *The number sense*. Oxford University Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., and al (2024). The llama 3 herd of models.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.