Prov2Vec: Exploring Cultural and Linguistic Connections through Proverb Embeddings

Israel Avihail
avihaili@post.bgu.ac.il

Maidad Maissel
maidad@post.bgu.ac.il

Tomer Sagie
sagiet@post.bgu.ac.il

Iftach Shoaham
iftachs@post.bgu.ac.il

Abstract

Understanding the semantic relationships between proverbs across and within languages provides insights into cultural and linguistic connections. This paper introduces a novel approach, *Prov2Vec*, where we generate vector embeddings for proverbs using GPT-4 Mini explanations combined with few Token embeddings models. These embeddings are evaluated for semantic similarity using cosine similarity. Our work builds upon previous efforts in Hebrew-English translation of idioms but shifts focus to embedding-based analysis of proverbs, uncovering deeper connections within and across languages. The dataset comprises 5000 English proverbs, 1800 Hebrew proverbs, and 100 proverbs in Chinese, Arabic, and French. We highlight the methodology, showcase examples, and discuss the implications of our findings in computational linguistics and cultural studies.

1 Introduction

In the 20th century, linguistics radically shifted through the work of Saussure and Chomsky, moving from seeing language as a collection of fixed-meaning words to viewing it as a complex system. The key insight was that meaning emerges from relationships between signs and their structural organization, with Saussure introducing the arbitrary nature of the signifier-signified relationship and Chomsky developing the concept of surface and deep structures in sentences.

This new understanding transformed translation theory, as exemplified in Walter Benjamin's "The Translator's Task" (1923). Rather than seeking word-for-word equivalence, Benjamin argued that translation should capture the underlying patterns of meaning - the "pure language" beneath surface differences. This led to translation approaches that prioritized preserving structural and cultural patterns over literal word correspondence, in our research we tried to follow this kind of translation, in the field of idioms and proverbs.

Proverbs encapsulate the essence of cultural wisdom, reflecting shared experiences and values in concise expressions. Their semantic richness and cultural specificity make them ideal candidates for exploring linguistic and cultural connections. In this research, we present *Prov2Vec*, a novel framework for embedding proverbs to study semantic similarity both within a single language and across languages.

2 Motivation and Challenges

Translating proverbs between languages presents a unique set of challenges that extend beyond mere linguistic differences. Proverbs are deeply rooted in the cultural and historical contexts of their origin, making direct translations often inadequate or misleading. Several key difficulties arise in this process:

- Ambiguity and Multiple Interpretations: Many proverbs are inherently ambiguous, allowing for multiple interpretations depending on the context in which they are used. This flexibility can be lost in translation, as different languages may not have equivalent expressions that capture the full range of meanings.
- Cultural Context and Idiomatic Nuances: Proverbs often reflect specific cultural values, historical experiences, and social norms. A phrase that carries significant wisdom in one culture may be unfamiliar or even nonsensical in another. Successful translation requires not only linguistic expertise but also a deep understanding of the cultural background from which the proverb originates.
- Wordplay, Puns, and Rhymes: Many proverbs rely on linguistic devices such as wordplay, puns, or rhymes, which do not always have

direct equivalents in another language. For instance, a proverb that relies on a phonetic pun in English may not convey the same meaning when translated literally into another language. Adapting such proverbs while preserving their humor or poetic quality is a particularly demanding aspect of translation.

- Literal vs. Conceptual Translation: Understanding the meaning of a proverb in one language does not necessarily equate to finding a word-for-word equivalent in another. Literal translations can often strip a proverb of its intended wisdom, requiring translators to find functionally or culturally equivalent expressions that maintain the intended message while adapting to the target language.

3 Computational Challenges

From a computational perspective, standard Large Language Models (LLMs) struggle with deciphering the implicit and often metaphorical meanings embedded within proverbs. These models, which excel in syntactic and statistical patterns, frequently fail to grasp the deeper cultural and contextual layers that define a proverb's meaning. Furthermore, they are not inherently equipped to identify a proverb in one language and map it to a semantically equivalent proverb in another.

Approaches such as zero-shot and few-shot learning have also proven inadequate in this context. The presentation of a few examples in a given language does not necessarily aid in understanding the underlying semantic structure of proverbs in another language. Unlike standard phrase translations, proverbs require a deeper level of abstraction and reasoning that LLMs currently struggle to achieve.

Therefore, an elegant transformation framework is needed to effectively bridge this gap. Such a solution would involve mapping proverbs across languages in a way that preserves their semantic, cultural, and contextual significance rather than relying on direct translations or simplistic statistical associations. By leveraging advanced embedding techniques, conceptual mappings, and cross-lingual understanding mechanisms, it may be possible to create a more robust approach to proverb translation that overcomes the limitations of existing models.

4 Research Goals

The primary objective of this research is to tackle the challenge of translation and understanding proverbs across languages by leveraging semantic embeddings (Prov2Vec). Unlike traditional word-to-word translations, we aim to explore how culturally significant proverbs can be aligned across and within languages to reveal linguistic and cultural connections.

5 Research Question

How can proverbs in different languages be represented and compared to reveal linguistic and cultural connections?

Proverbs often encapsulate values, traditions, and societal nuances that transcend literal meaning. Our research investigates how embedding-based analysis can preserve the semantic richness of proverbs while uncovering connections within and across languages. This study contributes to understanding cross-linguistic relationships, the encoding of cultural knowledge, and applications in computational linguistics.

6 Illustrative Examples: Limitations of Naive LLM

To illustrate the limitations of a naive application of large language models in idiomatic translation, we present the following example:

Consider the Hebrew idiomatic expression "לתפוס שתי ציפורים במכה אחת", which literally translates to "Catch two birds with one blow". A naive LLM might provide this direct translation, as shown in the image above. However, this translation fails to convey the intended meaning of the idiom, which in English is "Kill two birds with one stone". Other examples can be seen in Figure 1.

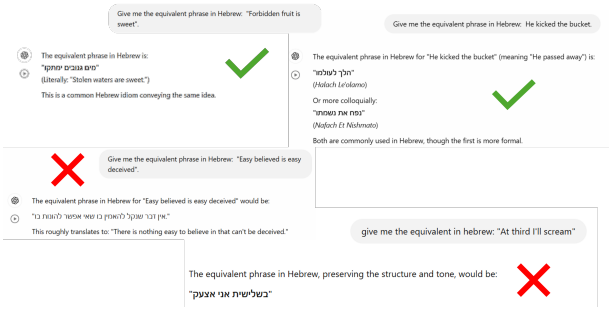

Figure 1: Examples of Hebrew idiomatic expression and its naive translation.

The issue here is that a naive LLM, without contextual understanding, tends to focus on word-for-word translation rather than capturing the idiomatic meaning. It overlooks the fact that the English equivalent involves a completely different phrasing that preserves the cultural and contextual essence of the expression. The translation fails to reflect the cultural significance and intended message, which is central to idiomatic expressions.

7 Dataset

Our dataset serves as a robust foundation for this exploratory study, comprising a diverse collection of proverbs across multiple languages. The dataset was compiled from multiple sources, primarily Kaggle and Wikipedia, where proverbs were extracted, curated, and refined to ensure consistency and reliability. The data collection process included filtering out duplicates and removing noise to enhance data quality. The final dataset includes:

- 4,437 English proverbs
- 1,867 Hebrew proverbs
- 300 Arabic proverbs
- 127 Chinese proverbs
- 90 French proverbs

8 Methodology

8.1 Computational Approach

To address the challenges of translating and analyzing proverbs across languages, we propose a computational approach that integrates the following key techniques:

- **GPT-4 Mini for Explanations:** Contextual explanations are generated for each proverb to enhance the embeddings with semantic richness.
- **Encoders Embeddings:** These embeddings translate explanations into high-dimensional representations to facilitate meaningful comparisons, we used several encoders and find that among those we checked, the *All MiniLM-L12-v2* was the best in the manner of capturing the meaning of the explanation paragraphs.
- **Cosine Similarity:** This metric quantifies the closeness between proverbs, allowing for cross-language and intra-language analysis.
- **Clustering Analyses:** Identifying clusters within embedding spaces provides insights into shared cultural themes and linguistic parallels.

This methodology enables a deeper exploration of cross-linguistic similarities and differences, advancing the field of computational linguistics by addressing challenges in cultural and contextual understanding. It also provide a solid way to translate idioms from one language to another, mostly English to Hebrew translation, since this was the focus in our research.

8.2 Prov2Vec Framework

The *Prov2Vec* framework consists of the following steps:

1. **Proverb Explanation:** Using GPT-4 Mini, we prompt the model to generate an explanation for each proverb, elucidating its meaning and context. This step ensures that embeddings capture semantic nuances.
2. **Embedding Generation:** The explanations are processed using *All MiniLM-L12-v2* embeddings, producing fixed-size vectors of dimension 384.
3. **Similarity Analysis:** We compute cosine similarity between embeddings to measure the semantic closeness of proverbs both within a language and across languages.
4. **Cross-Language and Intra-Language Analysis:** We analyze semantic clusters and relationships, seeking patterns and connections that highlight cultural and linguistic similarities.

This approach is particularly suited to the challenges of idiomatic translation because it combines the contextual understanding of pre-trained LLMs with task-specific adaptation. By focusing on the underlying meaning rather than literal translation, our method enables the accurate transfer of culturally rich expressions across languages.

Figure 2 illustrates the procedure we employ where a proverb is processed using GPT-4 Mini for explanation generation. The resulting explanations are embedded using the MiniLM-L12-v2 model, producing fixed-size 384-dimensional vectors. This procedure enables further comparative analysis between proverbs from different languages.

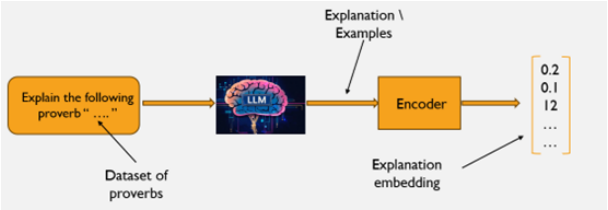

Figure 2: Illustration of the Prov2Vec Methodology.

The relevance of this approach to Natural Language Processing (NLP) lies in its focus on handling cultural and contextual complexities in language. Idiomatic expressions represent a unique challenge in NLP, as they require an understanding of meaning beyond words and syntax. By addressing this problem, our project explores the intersection of semantic understanding, cultural knowledge, and machine learning, which are critical for advancing NLP systems in real-world applications.

9 Translation Experiment

Our first experiment was to translate English idioms to Hebrew using our method, the same idioms sent to both GPT4o-mini and to Gemini with the following prompt:

"Give me a good translation of the following idiom, from English to Hebrew. Think what is the meaning of the idiom and then find as equivalent as possible idiom in hebrew. the idiom is: -The Idiom-"

The results show that in many cases our Prov2Vec method gave Hebrew idioms much more equivalence to the original, as shown in Table 1.

9.1 Exploratory Focus and Expected Findings

Given the exploratory nature of this study, our primary objective is to uncover meaningful patterns rather than conduct strict predictive evaluations. The research focuses on two key dimensions:

1. **Cross-Language Similarity:** where we expect proverbs with equivalent meanings across different languages to exhibit high cosine similarity in the embedding space. For example, proverbs such as "טובים השניים מן האחד" and its English counterpart "Two are better than one" should align closely in vector representations.

2. **Intra-Language Clusters:** where proverbs within the same language that share thematic or cultural significance—such as those related to wisdom or caution—are expected to form coherent clusters in the embedding space.

To assess these patterns, we employed a combination of quantitative and qualitative analysis. First, we examined the statistical distributions of cosine similarity scores to evaluate semantic alignment. Additionally, we performed outlier identification, analyzing proverbs with unexpected distances or alignments to gain insights into their unique semantic attributes.

9.2 Analysis of English Proverbs

For intra-language clustering in English, we utilized **Prov2Vec embeddings** (4,437 English proverbs) and applied **K-Means clustering** with **k = 20** to group semantically similar proverbs. Within each cluster, we calculated cosine similarity between all proverb pairs and filtered results with a score above **0.75**. Proverbs exceeding **0.85** similarity were often near-identical except for minor

Table 1: Translation Experiment Results: Examples of English idioms and the translation to Hebrew resulted from GPT4o-mini, Gemini, and Our method. The ilusinations in the LLMs translations are significant in many cases.

Proverb	Gpt 4o mini	Gemini	Our Approach
One good turn deserves another	"מעשה טוב דורש תגמול" "טובה תמורת טובה"	"מידה כנגד מידה" "טובה תחת טובה" "כגמולו ישולם לו"	"טובה תחת טובה" "מידה כנגד מידה"
Honey is sweet, but the bee stings	"הדבש מתוק, אבל הדבורים עוקצות"	"הדבש מתוק, אבל העוקץ צורב"	"לא מדובשך ולא מעוקצך"
Friend in need is friend indeed	"חבר בשעת צרה הוא חבר אמיתי"	"חבר בעת צרה הוא חבר אמיתי"	"יהי כבוד חברך חביב עליך כשלך" "לחמו נתן מימיו נאמנים"
Fair without, foul (false) within	"חיצוניות יפה, פנימיות מכוערת"	"פיו וליבו אינם שווים" "יפה מבחוץ, רקוב מבפנים"	"שקר החן והבל היופי"

Table 2: Cluster 1: human life, emotions, and experiences

Sim	Proverb Pairs	
0.76	little of what you fancy does you good	diseases are the interests of pleasures
0.75	it's better to have loved and lost	heart that once truly loves never forgets
0.76	death pays all debts	death is the grand leveler
0.77	man does not live by bread alone	hope is the poor man's bread
0.79	you can have too much of a good thing	diseases are the interests of pleasures

Table 3: Cluster 5: poverty, wealth, and money implications

Sim	Proverb Pairs	
0.75	poverty is not a shame	he is not poor that has little
0.76	money often unmakes the men	love of money is the root of all evil
0.77	money begets money	money is a good servant but a bad master
0.79	money often unmakes the men	money is the root of all evil

Table 4: Inner Language Chinese - Embedding Model Comparison

Model	NMI	V-M	Hom	FMI
bert-base	0.32	0.32	0.31	0.29
para-ML-L12	0.28	0.28	0.26	0.27
All-ML-L12	0.24	0.24	0.23	0.21
roberta-lg	0.24	0.24	0.24	0.20

word variations (e.g., *"wave a red rag to a bull"* and *"red rag to a bull"*), leading us to disregard them to maintain meaningful differentiation within the clusters. Inner results of different clusters are shown in figures 4,5,6:

9.3 Analysis of Chinese Proverbs

The dataset includes **127 Chinese proverbs**, which were categorized based on their underlying themes, such as **Wisdom, Friendship, Love, Family, Encouragement, Education, Literature, and Dragons**. This classification enables a structured analysis of how different cultural values are reflected in proverbial expressions. Each category encapsulates key aspects of Chinese philosophy, societal norms, and traditional beliefs.

For example, the proverb (Huà lóng diǎn jīng) literally means **"To dot the eyes of a painted dragon."** This expression originates from an ancient legend about an artist who painted dragons but left their eyes blank. Upon adding the pupils, the dragons miraculously came to life. The phrase symbolizes the importance of a **finishing touch** that brings completeness or vitality to a task. It falls under the **Dragons** category, as mythical creatures hold significant symbolism in Chinese culture.

Similarly, the proverb (Qīng guān nán duàn jiāwù shì) translates to **"Even a fair official finds it difficult to rule on family affairs."** This saying reflects the complexity of interpersonal relationships within families, emphasizing that even the most impartial judge may struggle to resolve domestic conflicts. This phrase was categorized under **Family**, as it highlights the deep-rooted cultural value of familial harmony and the challenges of mediating household disputes.

Assessment Methods:

- **Quantitative Analysis:** Statistical distributions of cosine similarity scores were analyzed to evaluate semantic alignment.
- **Outlier Identification:** Proverbs with unexpected distances or alignments were analyzed qualitatively to gain insights into their unique semantic attributes.

10 Initial Results and Discussion

Our initial experiments demonstrate the effectiveness of the *Prov2Vec* framework in identifying semantic alignments across languages. Key findings include:

Cross-Language Semantic Similarity:

- The Hebrew proverb "אל יתהלל חוגר כמפתח"

  ("Don't count your chickens before they hatch") and its English equivalent exhibited a cosine similarity score of 0.78, reflecting strong semantic alignment.

- Similarly, the Arabic proverb"القرش الأبيض بينفع في اليوم الأسود" ("A white coin is useful on a black day") aligned significantly with English and Hebrew equivalents, scoring above 0.81. Figure 3 illustrates these relationships, highlighting how cultural parallels manifest in shared themes.

Effectiveness of the Chain of Thought (CoT) Method:

Figure 3: Cosine Similarity Matrix for Arabic, English, and Hebrew Proverbs.

- Incorporating CoT explanations and examples produced nuanced embeddings. However, initial experiments showed that including both explanations and examples slightly decreased alignment scores (e.g., from 0.78 to 0.68). When examples were omitted during encoding, alignment scores improved significantly to 0.85, highlighting the potential of CoT.

11 Conclusion

This study demonstrates that the Prov2Vec framework is a groundbreaking approach for understanding semantic and cultural relationships between proverbs across and within languages. It effectively aligns proverbs in the embedding space, offering new insights into linguistic and cultural nuances.

Overall, Prov2Vec provides a promising foundation for tackling proverb translation and interpretation, advancing our understanding of linguistic and cultural expressions.

12 Future Work

Our initial results confirm the viability of *Prov2Vec* in uncovering linguistic and cultural connections via embedding-based analysis. For future work we suggest few options:

- Expand the dataset to include more proverbs in underrepresented languages such as French and Chinese.
- Experiment with state-of-the-art embedding models to enhance accuracy and semantic richness.
- The Chain of Thought (CoT) method showed significant potential in improving semantic alignment, though further testing is needed to refine its application. Future work will focus on clustering analyses to explore thematic and cultural groupings and systematically testing CoT with and without examples to better understand its impact.
- Use the Prov2Vec method in translation models, to have this an important feature of translating idiom to equivalent idiom.

These steps will build on our current findings, advancing the field of computational linguistics and cultural studies.

13 Code & Data Reference

The implementation code for this project is available on GitHub[1]. You can refer to the repository for detailed information on the model, algorithms, and how to run the code.

The dataset used in this project can be accessed through Google Drive[2]. Please ensure you have the necessary permissions to access and use the data for your research or experimentation or send a request for access if denied.

14 Reflection on the Project Process

14.1 What Did We Learn from the Process?

From the perspective of project execution and management, we learned that by integrating various computational approaches—such as prompt engineering, Chain of Thought (CoT), embeddings, and vector similarity comparisons in a vector space-combined with human thinking rather than relying solely on machine logic, we can achieve significantly improved results in language-related tasks.

14.2 Insights into Computational Approaches

We were exposed to different Large Language Models (LLMs) and their distinctions, particularly in how they handle semantic understanding of idioms and expressions. We discovered that in certain cross-linguistic meaning comprehension tasks, LLMs struggled, especially in performing direct semantic translation of idioms across different languages. This limitation necessitated human intervention in the form of CoT reasoning.

[1] https://github.com/iftach21/Prov2Vec

[2] https://drive.google.com/file/d/15vfFSbOqoiwMO8YeYtY5Fa0oSxNY5i88/view

Specifically, in our research, we introduced a tweak where we instructed the LLM to explain the intended meaning of the idiom rather than translating it literally. This served as a form of reasoning and CoT, effectively aligning idioms across languages and bridging the cultural gap caused by linguistic and cultural differences—where sentences that are identical in wording across languages might have entirely different meanings. This tweak allowed us to normalize the expressions, making them comparable by processing their explanations into embeddings. As a result, we achieved interesting and promising outcomes, demonstrating the effectiveness of our approach, which could be further developed in future research.

14.3 Conclusion

The rapid pace at which the field of Natural Language Processing (NLP) is evolving was evident throughout our work and the course itself. New breakthroughs and models emerged in real-time during the course, such as O1 and DeepSeek, which were released only recently.

15 Division of work

Tomer: Data collection, cleaning and arrangement, inclusion of a final report.

Israel: Creating the component that adds an explanation to the idiom.

Maidad: Creating the component that processes an explanation into an embedding.

Iftach: Planning experiments and building presentations.