Word Similarity and Relatedness Using Lexical Resources and Machine Learning

Aaryan Kashyap - 2023114006
Computational Linguistics II: Comp Semantics and Discourse parsing
International Institute of Information Technology

December 4, 2024

GitHub Repository link

Abstract

This project explores the implementation of various models measuring word similarity and relatedness using lexical resources such as WordNet, VerbNet, FrameNet, and hybrid approaches. The primary objective is to evaluate path-based and information content-based models, as well as syntactic and semantic models, in predicting word similarity. Over 300,000 word pairs were generated for training using lexical resources, and test evaluations were performed on datasets such as WordSim-353 and SimLex-999. Experimental results show that cosine similarity with a window size of 4 outperforms other models, achieving higher correlation with human judgments. While WordNet-based methods like Path Similarity, Wu-Palmer Similarity, and Resnik Similarity provide valuable insights, their performance is limited by WordNet's structure. The report provides detailed explanations of the theory, methodology, and results, offering potential avenues for future research in hybrid approaches.

Contents

1 Introduction

1.1 Research Problem

Word similarity and relatedness are critical tasks in Natural Language Processing (NLP), playing a foundational role in applications such as word sense disambiguation, machine translation, information retrieval, text classification, and semantic similarity measurement. These tasks involve quantifying the degree of similarity or relatedness between two words based on their meanings, a challenging endeavor given the complexity and variability of human language.

Despite significant advances in NLP, accurately measuring word similarity remains a persistent challenge. Existing models often struggle to capture nuanced relationships, such as those between words with subtle semantic differences or contextual dependencies. Additionally, the ambiguity inherent in natural language, where a single word may have multiple meanings, further complicates this task. The need for robust, context-sensitive methods for quantifying word similarity and relatedness is thus an ongoing research problem with wide-ranging implications for various NLP applications.

1.2 Literature Review

Over the years, researchers have developed several approaches to address word similarity and relatedness. These can be broadly categorized into lexical, statistical, and hybrid methods:

- **Lexical Resource-Based Methods:** These methods rely on structured linguistic resources such as WordNet and FrameNet. WordNet-based measures, such as Path Similarity and Wu-Palmer Similarity, leverage the hierarchical organization of synsets to quantify relationships between words. Similarly, FrameNet uses semantic frames to model word relationships in context. While these methods provide interpretable results and are effective for specific tasks, they are limited by the coverage and granularity of the underlying resources.

- **Statistical or Distributional Methods:** These methods exploit the distributional hypothesis, which posits that words with similar meanings tend to appear in similar contexts. By analyzing large text corpora, distributional methods generate vector representations of words (e.g., word embeddings) and calculate similarity using metrics like cosine similarity. Contextualized word embeddings, such as those generated by transformer models like BERT, have significantly improved performance by capturing semantic relationships in context.

- **Hybrid Methods:** Hybrid approaches aim to combine the interpretability of lexical methods with the robustness of statistical methods. For example, models may integrate WordNet's hierarchical relationships with corpus-based embeddings to improve performance. Despite their promise, hybrid methods often face challenges in balancing interpretability, computational efficiency, and accuracy.

Existing approaches, while effective in specific domains, often fail to generalize across tasks due to limitations such as sparse coverage in lexical resources or lack of contextual understanding in statistical methods. Moreover, the integration of lexical and distributional knowledge remains an open area of research.

1.3 Research Objectives and Hypotheses

This project aims to advance the field of word similarity measurement by exploring and evaluating multiple approaches, including lexical resource-based, distributional, and hybrid methods. The specific objectives are as follows:

- **Objective 1:** Implement and evaluate WordNet-based similarity measures, including Path Similarity, Wu-Palmer Similarity, and Resnik Information Content, to assess their effectiveness in quantifying word relationships.
- **Objective 2:** Incorporate FrameNet and VerbNet to introduce syntactic and semantic frame-based features into the similarity models, enabling the evaluation of frame-level semantics.
- **Objective 3:** Explore distributional methods, particularly cosine similarity, using word embeddings derived from the training corpus, with a focus on varying context window sizes to capture semantic nuances.
- **Objective 4:** Compare the performance of individual models with hybrid approaches, hypothesizing that combining lexical and distributional methods will yield higher correlation with human judgments of word similarity.

This research hypothesizes that hybrid approaches, which integrate the structural knowledge of lexical resources with the contextual richness of distributional methods, will outperform standalone models in capturing complex semantic relationships. The findings of this study are expected to contribute to the development of more accurate and interpretable word similarity measures, with potential applications in a wide range of NLP tasks.

2 Theoretical Framework: WordNet and FrameNet

2.1 WordNet: A Lexical Database for English

WordNet, developed at Princeton University by George A. Miller and his team, is a comprehensive lexical database designed to enhance the computational understanding of the English language. It groups words into sets of cognitive synonyms, referred to as *synsets*, and establishes intricate semantic relationships between these synsets. Unlike traditional dictionaries, WordNet emphasizes the semantic and lexical interrelations of words, making it a valuable resource for natural language processing (NLP) tasks.

2.1.1 Key Features of WordNet

- **Synsets:** A synset is a collection of words or phrases that share an identical or closely related meaning. For example, the words *car*, *automobile*, and *motorcar* belong to the same synset, representing the concept of a motorized vehicle. Synsets form the foundational unit of WordNet, enabling its semantic structure to be organized effectively.
- **Semantic Relations:** WordNet defines a network of relationships among synsets, offering a rich understanding of lexical semantics. Some of the most notable semantic relations include:
  - *Hypernymy (Is-A Relationship):* This relation denotes a generalization hierarchy. For instance, a *dog* is a hypernym of a *Labrador*, and similarly, *animal* is a hypernym of *dog*.
  - *Hyponymy (Subtype Relationship):* The inverse of hypernymy, this relation represents specialization. For example, a *sparrow* is a hyponym of a *bird*, and a *bird* is a hyponym of an *animal*.
  - *Meronymy and Holonymy:* These relations express part-whole hierarchies. For example, a *wheel* is a meronym (part) of a *car*, and a *car* is a holonym (whole) of a *wheel*.
  - *Antonymy:* WordNet captures opposites, such as *hot* and *cold*, providing valuable resources for understanding contrasts in meaning.
  - *Synonymy:* Synonymy refers to words with identical or nearly identical meanings, such as *happy* and *joyful*.
  - *Troponymy:* This relationship captures specific manners of performing actions. For example, *to whisper* is a troponym of *to speak*.
- **Hierarchical Organization:** WordNet organizes concepts into a hierarchical structure, where synsets are linked via hypernymy or hyponymy relations. At the top of the hierarchy are broad concepts such as *entity*, *event*, and *state*, while more specific concepts lie deeper in the taxonomy. This hierarchical organization allows for semantic generalization and specialization.
- **Lexical and Conceptual Coverage:** WordNet covers an extensive vocabulary of nouns, verbs, adjectives, and adverbs. Each word is associated with multiple senses, each of which belongs to a specific synset. For example, the word *bank* can refer to a financial institution or the side of a river, each with distinct synsets and relationships.

2.1.2 Applications of WordNet

WordNet's rich lexical database is widely used in various NLP applications, including but not limited to:

- **Word Sense Disambiguation:** Determining the correct sense of a word in a given context by leveraging the semantic relationships within WordNet.
- **Text Classification and Clustering:** Enhancing text analysis by identifying the semantic similarity between words.
- **Semantic Similarity Measurement:** Calculating the similarity between words or phrases using path-based or information content-based metrics.
- **Question Answering and Information Retrieval:** Using WordNet's semantic relationships to improve the retrieval of relevant information.

2.1.3 Limitations of WordNet

Despite its strengths, WordNet has some notable limitations:

- **Coverage Gaps:** WordNet's vocabulary may not adequately cover domain-specific or technical terms, limiting its applicability in specialized fields.
- **Static Nature:** As a manually curated resource, WordNet is less adaptable to rapidly evolving language usage compared to data-driven models.
- **Lack of Contextual Awareness:** WordNet does not account for contextual nuances, which are crucial in tasks involving polysemous words or phrases.
- **Relational Overlap:** The semantic relations defined in WordNet may overlap or lack precision in certain cases, leading to ambiguity in similarity calculations.

3 Methodology

3.1 Data Collection

- Training data: Over 300,000 word pairs were generated using lexical resources, including WordNet synsets and FrameNet relations.
- Test data: The test dataset consists of 2,070 word pairs with human-annotated similarity scores from WordSim-353 and SimLex-999.

3.2 WordNet-Based Models

3.2.1 Path Similarity

Path similarity calculates the inverse of the shortest path length between two synsets in WordNet:

$$\text{Path Similarity}(w_1, w_2) = \frac{1}{\text{Shortest Path Length}(w_1, w_2) + 1} \tag{1}$$

Words closer in the hierarchy have higher similarity scores. Limitations include handling unrelated words and dependency on WordNet's structure.

3.2.2 Wu-Palmer Similarity

Wu-Palmer similarity is defined as:

$$\text{Wu-Palmer Similarity}(w_1, w_2) = \frac{2 \cdot \text{Depth of LCS}}{\text{Depth}(w_1) + \text{Depth}(w_2)} \tag{2}$$

Where LCS is the lowest common subsumer of the two synsets.

3.2.3 Resnik Similarity

Resnik similarity measures shared information content:

$$\text{Resnik Similarity}(w_1, w_2) = -\log P(\text{LCS}(w_1, w_2)) \tag{3}$$

Here, $P(\text{LCS})$ is the probability of encountering the lowest common subsumer in a corpus.

3.3 Cosine Similarity

Cosine similarity was calculated using word embeddings generated with window sizes of 2 and 4. The formula is:

$$\text{Cosine Similarity}(w_1, w_2) = \frac{\vec{w_1} \cdot \vec{w_2}}{\|\vec{w_1}\| \cdot \|\vec{w_2}\|} \tag{4}$$

Experiments show that a window size of 4 outperforms size 2, indicating a better contextual understanding.

3.4 FrameNet Models

FrameNet models consider frame overlap and frame distance. Word pairs were assigned similarity scores based on shared frames and their distances.

4 Results

In this project, we evaluated the effectiveness of multiple lexical similarity measures using WordNet, FrameNet, and cosine similarity with varying window sizes. The training data consisted of over 50,000 sentences, providing a comprehensive lexical corpus for the model's learning phase. Additionally, the test dataset contained 2,070 word pairs annotated with semantic relationships such as synonyms, antonyms, and hyponyms, serving as ground truth for evaluation. Below, we discuss the results of each method in detail.

4.1 WordNet-Based Similarity Measures

WordNet-based similarity measures, including Path Similarity, Wu-Palmer Similarity, and Resnik Information Content, were implemented to evaluate semantic relationships between word pairs. The training data involved calculating semantic distances and probabilities for a large set of word pairs. However, the WordNet results were suboptimal for several reasons:

- **Path Similarity**: The similarity scores were heavily dependent on the hierarchy depth in the WordNet taxonomy. While some word pairs with direct hierarchical relationships (e.g., synonyms or hyponyms) achieved reasonable scores, others with more abstract connections performed poorly.
- **Wu-Palmer Similarity**: The Wu-Palmer measure, which calculates similarity based on the least common subsumer (LCS), showed moderate performance but struggled with pairs lacking a shared LCS in WordNet's hierarchy.
- **Resnik Similarity**: Resnik's measure, leveraging Information Content (IC), failed to capture nuanced similarities due to the sparsity of corpus-based probabilities.

Example Results:

Table 1: WordNet Similarity Example

Word 1	**Word 2**	**Path Sim.**	**Wu-Palmer Sim.**	**IC Resnik**
cost	destination	11	0.0	7.475
cost	include	inf	0.0	7.475
cost	all	inf	0.0	7.475

4.2 FrameNet-Based Predictions

FrameNet models were constructed using frame overlaps and distances as similarity metrics. The training data contained a set of word pairs annotated with frame-level features derived from the FrameNet semantic database.

Results:

- Predictions consistently produced low similarity scores (e.g., 0.1 across test pairs), indicating poor differentiation among relationships.
- Frame overlaps were rare in the training data, leading to sparse feature representations.

Challenges:

- The lack of extensive coverage in FrameNet frames for certain domain-specific or low-frequency words limited its applicability.
- The implementation may have suffered from incorrect frame mappings, affecting the overall quality of predictions.

4.3 Cosine Similarity-Based Approaches

Cosine similarity was applied to word embeddings derived from the training corpus. Two window sizes, 2 and 4, were evaluated to explore the impact of context size on similarity calculations.

Results for Window Size 2:

- Cosine similarities exhibited mixed results, with moderate performance for some synonym pairs (e.g., "recognized" and "acknowledge" with 0.6045) but low scores for others.
- Antonyms and unrelated word pairs (e.g., "close" and "distant") were often assigned scores close to zero, failing to capture semantic opposition.

Results for Window Size 4:

- Larger context windows significantly improved performance, particularly for synonym pairs. For instance:
  - "Accustomed" and "habituate" had a cosine similarity of 0.8968, compared to -0.0448 for window size 2.
  - "Right" and "redress" increased from 0.2737 (window size 2) to 0.3463.
- The additional context provided by a larger window appeared to capture semantic nuances more effectively.

Example Results:

Table 2: Cosine Similarity Example (Window Size 4)

Word 1	**Word 2**	**Cosine Sim.**
accustomed	habituate	0.8968
right	redress	0.3463
eruptive	igneous	-0.0403
raised	curse	nan

4.4 Limitations and Challenges

While cosine similarity (window size 4) showed the best performance, the project faced several limitations:

- **WordNet and FrameNet Coverage**: Both resources lacked adequate coverage for domain-specific words, leading to suboptimal results.

- **Implementation Challenges**: FrameNet predictions suffered from frame-mapping errors, and WordNet's hierarchical limitations resulted in undefined similarity scores for certain pairs.
- **Training Data Quality**: Although the training corpus of 50,000 sentences was extensive, the semantic diversity may not have been sufficient for robust model training.

4.5 Conclusion

The results demonstrate that context-based methods (cosine similarity with larger window sizes) outperform lexical database-based approaches like WordNet and FrameNet for semantic similarity tasks. Future work should focus on:

- Improving the implementation of FrameNet and WordNet models to address coverage and mapping issues.
- Incorporating pre-trained embeddings or transformer-based models for enhanced semantic representation.

5 Discussion

The results of this study highlight the strengths and limitations of different approaches to semantic similarity. Specifically, distributional methods, such as cosine similarity, consistently outperformed lexical database-based methods like WordNet and FrameNet. This outcome is not entirely unexpected, as distributional methods are designed to leverage contextual co-occurrence patterns in a corpus, which often capture semantic nuances more effectively than static lexical structures.

5.1 Strengths of Distributional Methods

The distributional approach, particularly cosine similarity with a larger context window (size 4), demonstrated superior performance in capturing semantic relationships. This was evident in the high similarity scores assigned to synonym pairs, such as "accustomed" and "habituate," which achieved a cosine similarity score of 0.8968. By incorporating a broader context, the model was able to detect subtle semantic associations that smaller windows or lexical approaches failed to identify.

The flexibility of distributional methods allows them to adapt to domain-specific vocabularies and dynamic language use, which static resources like WordNet and FrameNet often struggle to accommodate. Furthermore, the ability to compute similarities without requiring predefined hierarchical structures makes cosine similarity a versatile tool for a wide range of applications.

5.2 Limitations of Lexical Database-Based Models

WordNet-based methods, while theoretically robust, were hampered by several practical limitations:

- **Static Hierarchies**: WordNet's rigid taxonomy often failed to capture nuanced relationships between words, particularly for those not explicitly linked by direct or indirect hierarchical paths.
- **Sparse Information Content**: Resnik similarity relies on corpus-based information content, which was limited by the training corpus. As a result, many word pairs received low or undefined similarity scores.

FrameNet, on the other hand, struggled due to implementation challenges and coverage limitations:

- **Sparse Frame Overlaps**: Many word pairs in the test set did not share overlapping frames, leading to uniformly low similarity scores (e.g., 0.1 across most pairs).
- **Frame Representation Issues**: The training data revealed inconsistencies in frame assignment, likely contributing to the poor performance of FrameNet-based predictions.

These challenges underscore the need for more dynamic and scalable lexical resources that can better integrate with corpus-based methods.

5.3 Limitations of the Study

While the project yielded valuable insights, it was not without limitations:

- **Training Data Quality**: Although the dataset of 50,000 sentences was extensive, it lacked sufficient semantic diversity to fully capture complex word relationships.
- **FrameNet Coverage**: FrameNet's limited vocabulary and rigid frame structures restricted its applicability, particularly for low-frequency or domain-specific words.
- **Implementation Challenges**: Errors in frame mapping and semantic distance calculations likely affected the accuracy of the results, particularly for FrameNet-based models.

5.4 Implications for Future Research

The results of this study suggest several avenues for future research:

- **Integration of Pre-trained Models**: Incorporating pre-trained embeddings, such as those from BERT or Word2Vec, could enhance performance by providing richer semantic representations.
- **Hybrid Models**: Developing models that combine lexical resources like WordNet and FrameNet with distributional methods could mitigate the limitations of each approach.
- **Dynamic Lexical Resources**: Enhancing lexical databases with corpus-based statistics or crowdsourced data could improve their coverage and adaptability.
- **Evaluation Metrics**: Future studies should explore alternative evaluation metrics that better capture the nuances of semantic similarity.

6 Conclusion

This project successfully implemented and evaluated multiple approaches to word similarity and relatedness, including WordNet, FrameNet, and cosine similarity methods. Among these, cosine similarity with a larger context window (size 4) achieved the best results, demonstrating the effectiveness of distributional methods in capturing semantic relationships.

The study also highlighted the limitations of lexical resources like WordNet and FrameNet, which struggled with coverage and static structures. These findings underscore the need for hybrid approaches that can leverage the strengths of both lexical and distributional methods.

6.1 Key Contributions

- **Comparative Analysis**: This study provided a detailed comparison of lexical and distributional approaches to semantic similarity.
- **Implementation Insights**: The project revealed practical challenges in implementing FrameNet-based models, offering valuable lessons for future work.
- **Recommendations for Improvement**: By identifying the limitations of existing methods, this study outlined clear directions for enhancing word similarity models.

6.2 Future Directions

Building on these findings, future research should focus on integrating pre-trained models, enhancing lexical resources, and developing hybrid approaches. Such efforts have the potential to significantly advance the field of computational linguistics, enabling more accurate and scalable semantic similarity models.