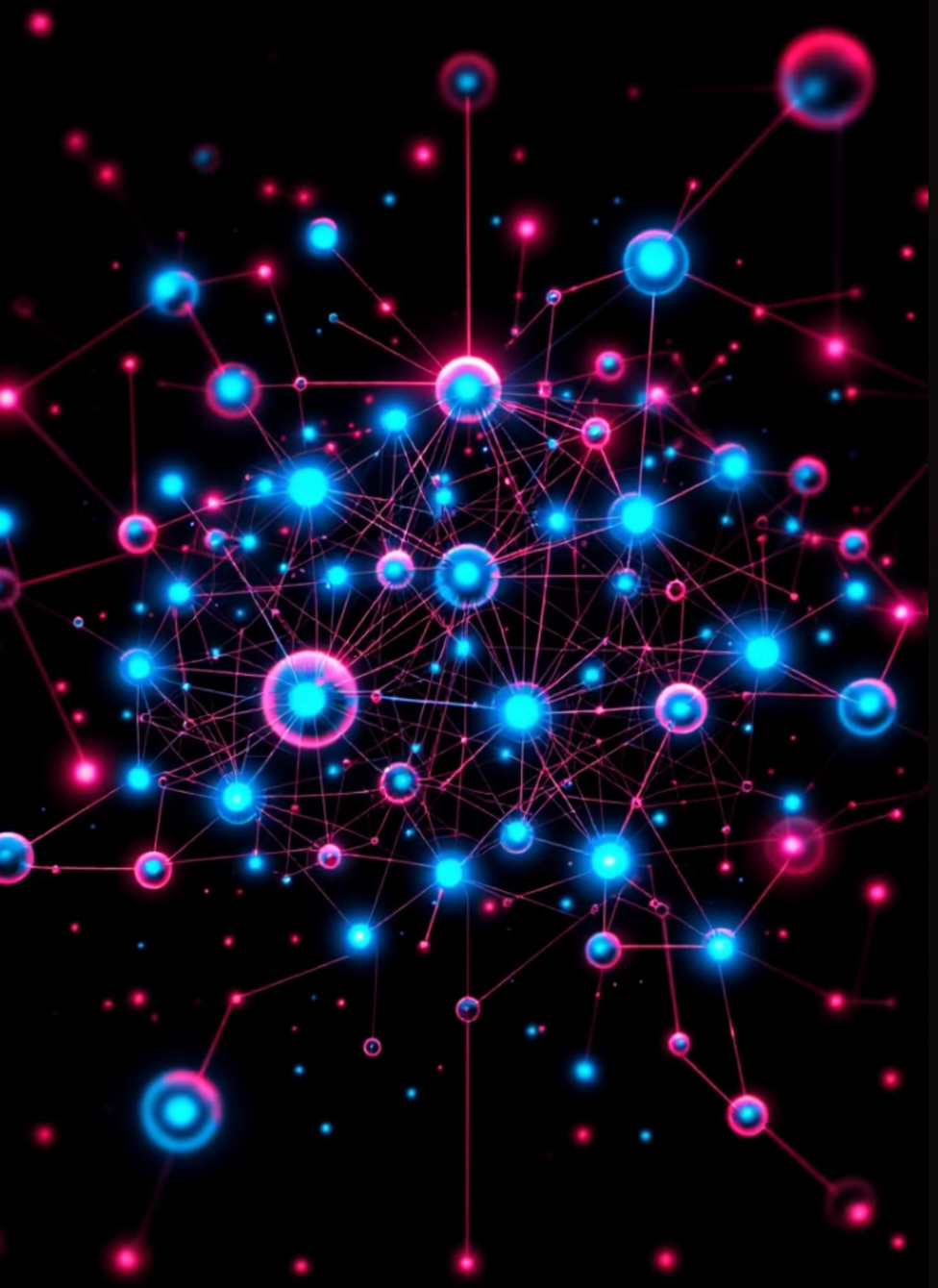


Optimizing A Minimal Core Vocabulary For Language Representation

This research investigates identifying a minimal core vocabulary that maintains expressive power in English. The study constructs a representative vocabulary using computational techniques, develops an algorithm to generate sentences with the core vocabulary, and measures its effectiveness. By integrating advanced language models, dimensionality reduction, and clustering methods, we optimize vocabulary size and evaluate efficacy. We analyze the relationship between similarity scores and vocabulary size using Cosine and Jaccard similarity metrics to assess semantic coherence.

by Roey Feingold,Yohai Haddad

Vocabulary Building Process

1

Corpus-Based Selection

Extract most frequent words from corpora like Brown, Reuters, and Gutenberg. Filter out stop words and non-alphabetic characters.

2

WordNet Enhancement

Expand vocabulary using WordNet to include hypernyms, covering broader semantic concepts.

3

Embedding Filtering

Apply GloVe and BERT embeddings to refine vocabulary, pruning words with high semantic overlap.

Sentence Generation Algorithm

1

Preprocessing

Tokenize input, remove stop words, identify out-of-vocabulary words

2

Word Substitution

Find closest words using cosine similarity of embeddings

3

Sentence Reconstruction

Rebuild sentence using substitutions while maintaining structure

Semantic Similarity Metrics

Cosine Similarity

Measures angle between word vectors in semantic space. Range: -1 to 1. Higher values indicate greater similarity. Useful for comparing word embeddings.

Jaccard Similarity

Measures overlap between word sets. Defined as intersection divided by union. Useful for evaluating lexical overlap between sentences.

Coverage and Compression Ratios

Coverage: proportion of words representable by core vocabulary. Compression: reduction in vocabulary size while retaining meaning.

Evaluation Methodologies

Embedding-Based Substitution

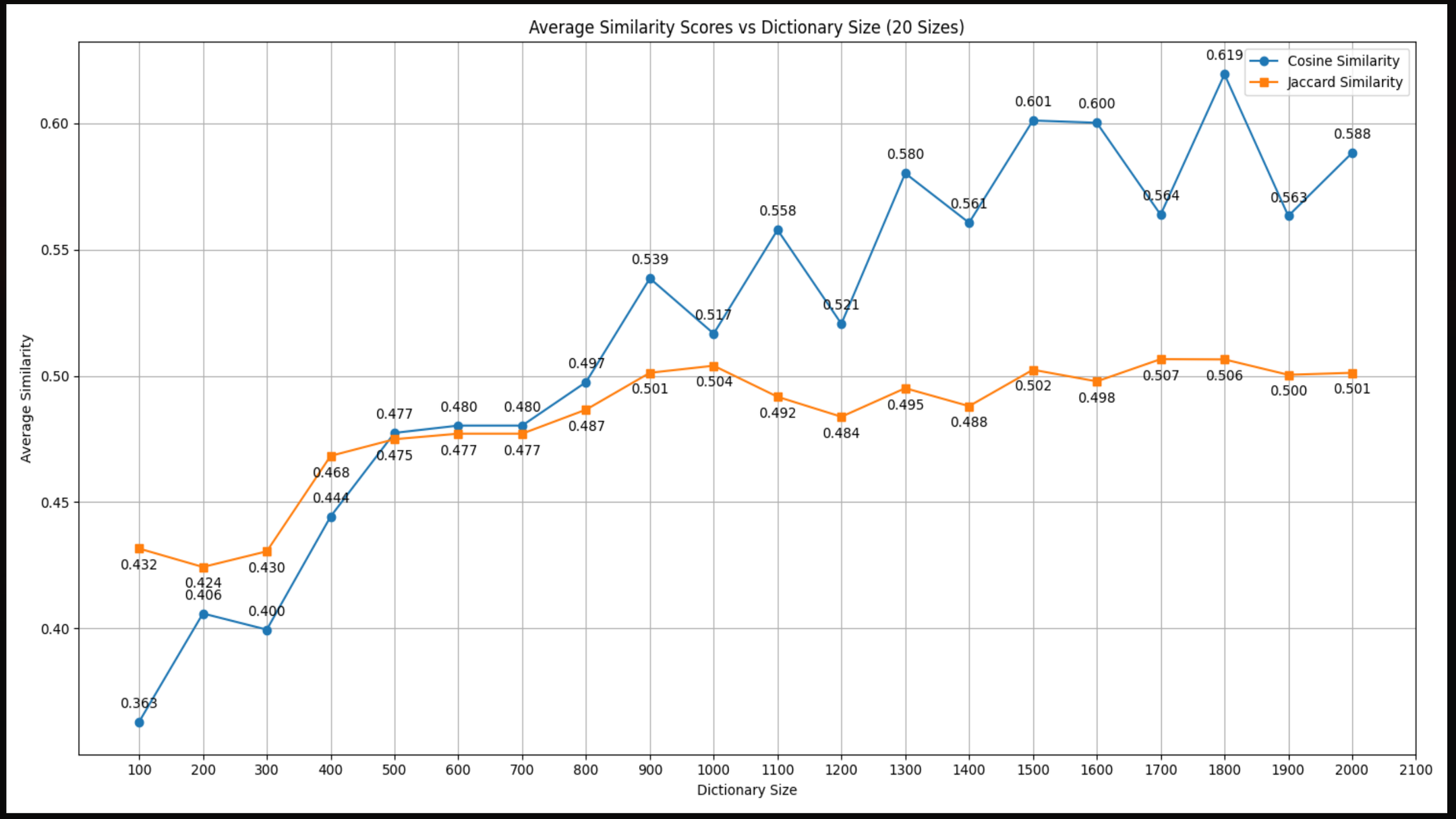
Use word embeddings to find semantically similar words in reduced vocabulary. Calculate cosine similarity between original and substituted sentences.

Back-Translation Analysis

Translate sentence to another language and back. Assess how well reduced vocabulary captures nuances across languages.

Human Evaluation

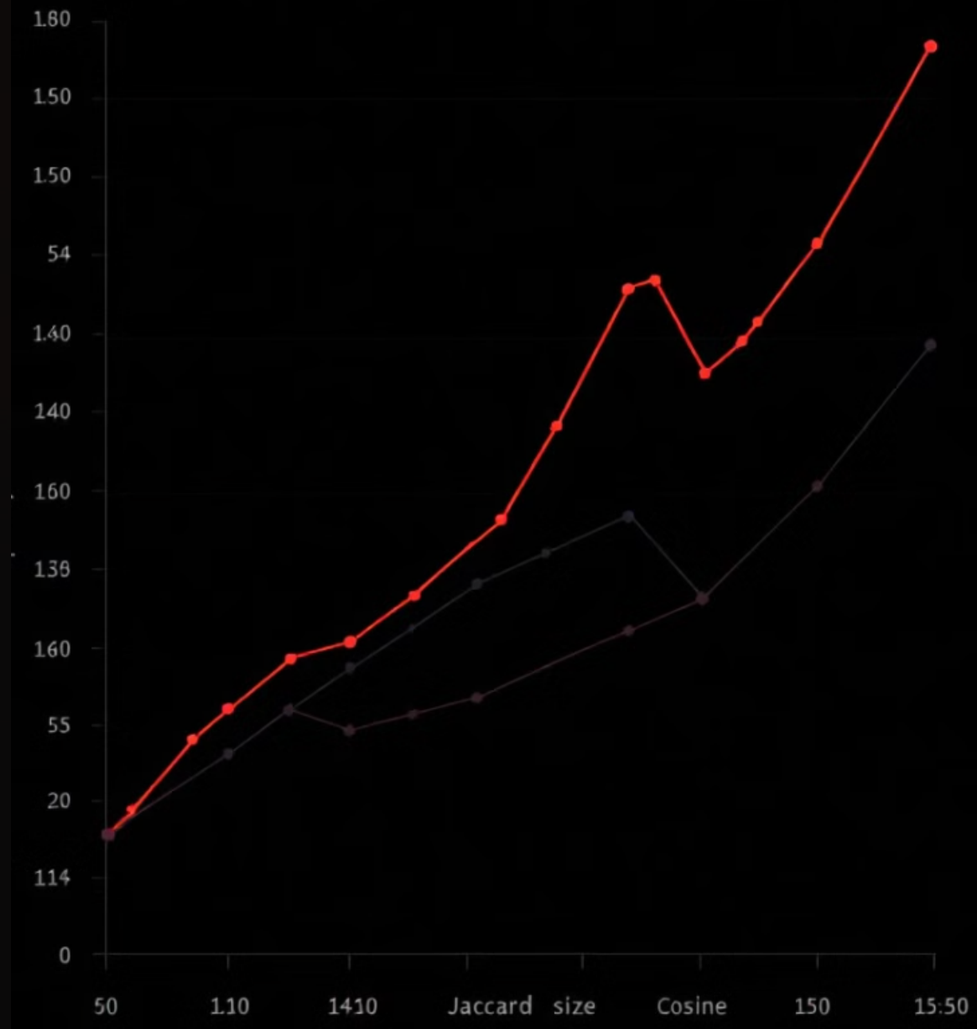
Have evaluators assess intelligibility and naturalness of sentences constructed using reduced vocabulary.

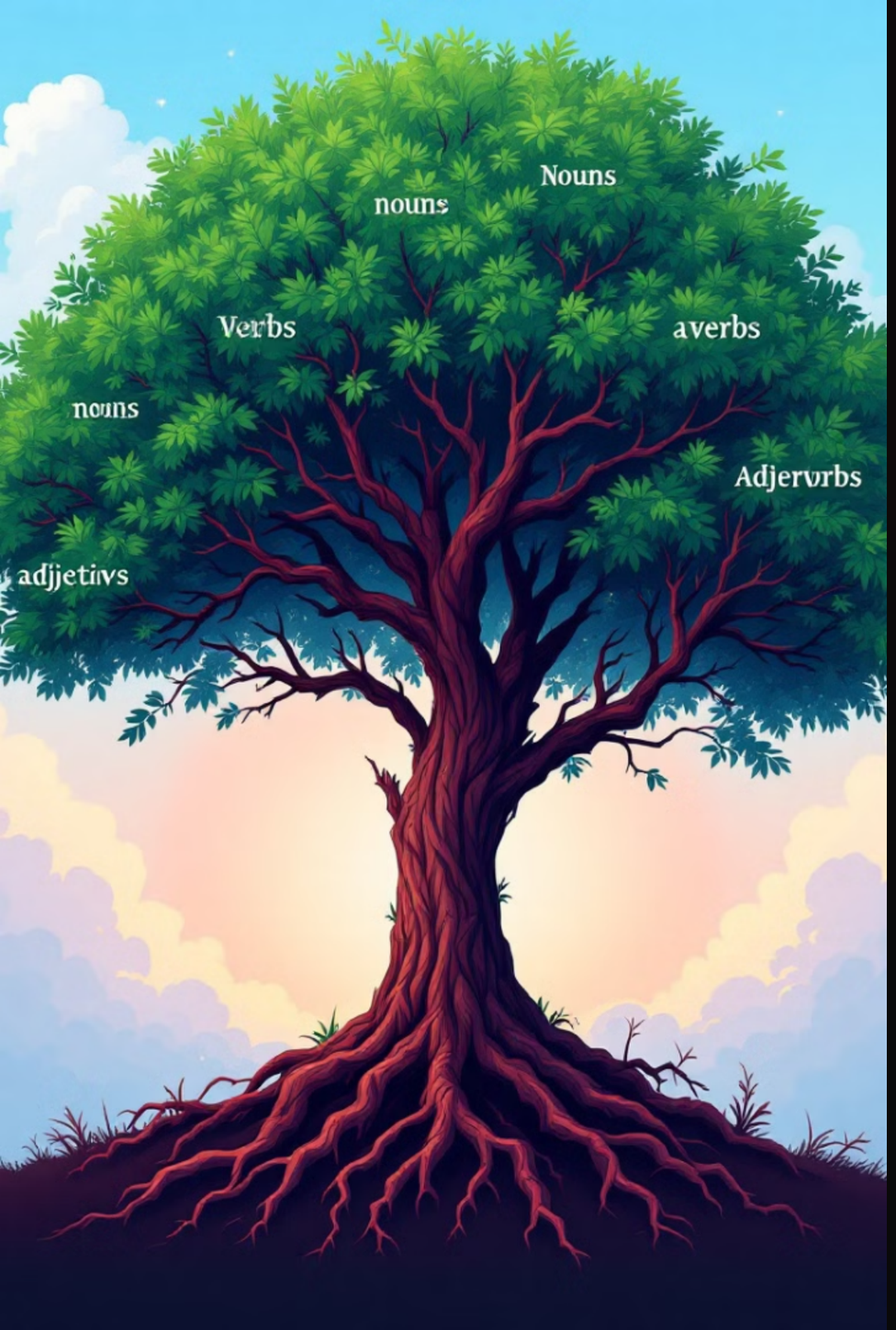

Average Similarity Scores vs Dictionary Size (20 Sizes)
Cosine Similarity
Jaccard Similarity
Average Similarity
Dictionary Size
0.60
0.55
0.50
0.45
0.40
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
0.363
0.406
0.400
0.444
0.477
0.480
0.480
0.497
0.539
0.517
0.558
0.521
0.580
0.561
0.601
0.600
0.564
0.619
0.563
0.588
0.432
0.424
0.430
0.468
0.475
0.477
0.477
0.487
0.501
0.504
0.492
0.484
0.495
0.488
0.502
0.498
0.507
0.506
0.500
0.501

SICK Dataset Results

Metric	SICK Creators	Our Approach
Cosine Similarity	0.6009	~0.619
Jaccard Similarity	0.4884	~0.501

Our approach using a limited vocabulary achieved comparable or slightly better results than the SICK dataset creators, who used unrestricted vocabulary. This demonstrates the effectiveness of our core vocabulary in maintaining semantic fidelity.





Optimal Vocabulary Size

1 1000 Words

Lower bound for retaining significant semantic information. Suitable for simplified language models and beginner language learning.

2 1500 Words

Balances common terms and some specialization. Ideal for intermediate learners and systems requiring richer language understanding.

3 2000 Words

Upper range showing consistent improvement in similarity scores. Accommodates more nuanced expression while maintaining efficiency.

Conclusion and Future Directions

Refinement

Further refine core vocabulary selection using advanced techniques like contextualized embeddings.

Cross-Linguistic

Apply approach to other languages to explore universality of findings.

NLP Applications

Develop more efficient language models, especially for low-resource languages.

Language Learning

Create simplified learning approaches focusing on mastering essential words first.

This research demonstrates the feasibility of reducing language to a core vocabulary while retaining communicative power. Future work should address limitations in specialized domains and subjectivity of similarity measures.