Similarity Measurements for Romanian Words

Rares Dan Tiago Goia

April 13, 2025

Abstract

This paper explores diverse approaches for measuring similarity between Romanian words, ranging from distributional semantics to orthographic and etymological methods. By synthesizing findings from three published papers related to the domain —(1) *Alternative Measures of Word Relatedness in Distributional Semantics*, (2) *An Etymological Approach to Cross-Language Orthographic Similarity*, and (3) *More Romanian Word Embeddings from the RETEROM Project*—I will provide a comparative framework for understanding how semantic, morphological, and historical dimensions influence word similarity in Romanian and related languages.

Contents

1 Introduction

For underutilized languages like Romanian, Natural Language Processing (NLP) requires accurate methods for determining word similarity. Accurate similarity metrics are essential for semantic analysis, information retrieval, translation, and a variety of other NLP tasks. This paper compares three distinct approaches to word similarity:

- **Semantic similarity in distributional spaces**, with alternative metrics beyond standard cosine similarity.
- **Etymological and orthographic similarity across languages**, emphasizing historical evolution and cognate detection.
- **Word embeddings and neural representation models** as developed in the RETEROM project, addressing challenges like rich morphology and inflectional variance.

2 Distributional Semantic Relatedness

2.1 Theoretical Framework

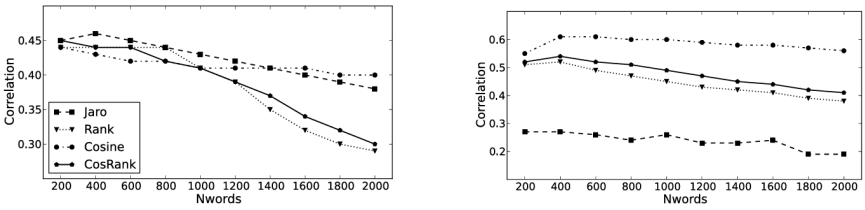
A. Ciobanu and A. Dinu propose novel alternatives to classical cosine similarity in measuring word similarity. While cosine similarity captures the angle between word vectors, it may not fully represent nuanced contextual relationships. In their work, words are represented as ranked lists of co-occurring terms (utilizing tf–idf weights), and a variety of metrics are explored:

- **Rank Distance**: Quantifies differences between word ranking lists.
- **CosRank Distance**: Merges ranking similarity with cosine-style vector comparison for improved nuance capture.
- **MeanRank Distance**: Averages differences over reversed rankings to reduce bias.
- **Jaro Distance**: Originally a string metric, adapted here to compare ranked contexts.
- **Additional Measures**: Jaccard and Dice coefficients are also considered to stress the degree of overlap between contexts, while PMI-based metrics help mitigate frequency biases.

2.2 Evaluation and Results

The method was tested on the WS-353 test set, comprising 353 word pairs rated by human judges. Distances were computed using raw frequency and tf–idf weights derived from the Wacky corpus. Key observations include:

- Rank and Jaro distances exhibited strong performance, particularly with smaller ranking sizes (up to 1000 terms).
- Cosine similarity's performance improves only at very high dimensional spaces.
- The CosRank distance, when used with tf–idf weights, achieved a high Spearman correlation (0.55) with human judgments.
- Context-dependent similarity measures, such as the Jaccard and Dice coefficients, provide complementary insights for capturing both substitutional and associative relationships.



(a) Results for experiments with raw frequencies

(b) Results for experiments with $tf - idf$ weights

Figure 1: Comparison of results for different similarity scores

2.3 Implications

The ranking-based and alternative metrics approach offers:

- Reduced computational overhead compared to high-dimensional vector models.
- Stability across different corpus domains.
- A flexible framework that can be extended to phrase-level or cross-linguistic semantic analysis.

3 Etymology-Based Orthographic Similarity

3.1 Motivation

Given Romanian's complex linguistic history with influences from Latin, Slavic, French, Italian, and even Turkish, traditional orthographic similarity measures (e.g., Levenshtein distance) often fail to capture deeper linguistic relationships. This study emphasizes the importance of **etymology** in determining orthographic similarity, particularly for cross-language comparisons.

3.2 Proposed Approach

The authors introduce an etymology-aware similarity metric that incorporates:

- Historical evolution of words.
- Orthographic transformations observed across languages.
- Automatic cognate identification using dictionaries, etymological databases, and translation APIs.

Additionally, the approach considers conventional string-based metrics such as:

- **LCSR (Longest Common Subsequence Ratio)**
- **Normalized Edit Distance**
- **Rank Distance for Strings**

3.3 Methodology and Findings

- **Data Collection and Preprocessing:** The process involves extracting word pairs from etymological databases, normalizing diacritics, and handling language-specific affixes.
- **Similarity Calculation:** Beyond simple Levenshtein distance, the method combines standard orthographic metrics with linguistic rules that account for historical changes.
- **Results:** Cognate identification was significantly enhanced, with similarity scores increasing up to 300% for languages such as Spanish. Romanian shows high orthographic similarity with French and Italian, and even Turkish words exhibit notable similarities due to shared loanwords.

3.4 Conclusion

The integration of etymological insights showcases a robust approach that outperforms purely orthographic methods in recognizing true linguistic relationships. The technique has promising applications in multilingual lexicography, machine translation, and AI-driven language learning tools.

Language	Parliament					Eminescu					Chronicles					RVR				
	%words	D		ND		%words	D		ND		%words	D		ND		%words	D		ND	
French	70.6	45.5	46.0	48.3	48.8	57.2	35.2	36.1	37.2	38.2	36.7	20.3	21.1	22.3	23.1	50.6	30.3	31.4	32.2	33.3
Latin	63.7	40.2		42.0		59.9	34.6		36.6		44.9	24.2		25.7		56.5	34.0		37.3	
Italian	48.5	28.1	33.4	29.1	34.5	44.7	26.9	30.2	27.9	31.2	31.7	19.6	20.3	20.7	21.4	41.4	23.4	26.2	25.2	28.0
Spanish	40.2	9.2	24.9	10.7	27.0	38.1	10.9	21.2	12.9	23.7	29.7	11.9	15.1	13.9	17.2	32.5	9.0	19.5	9.9	21.0
Portuguese	35.0	8.3	22.1	9.5	24.0	31.3	9.6	18.5	11.3	21.0	28.3	12.2	16.3	13.9	18.4	29.3	8.6	17.4	9.4	18.9
English	22.1	2.2	14.0	2.2	14.2	18.8	1.1	9.9	1.2	10.1	11.3	1.3	5.9	1.3	6.2	14.3	1.6	10.3	1.6	10.4
Provencal	17.7	9.6		9.8		20.7	11.3		11.6		21.8	13.0		13.4		16.8	9.7		10.5	
German	9.2	5.8		5.9		6.9	4.5		4.6		4.9	2.4		2.4		10.2	6.3		6.6	
Turkish	7.7	0.9	5.4	0.9	5.6	6.6	1.7	4.5	1.7	4.7	5.6	2.9	3.7	3.1	3.9	7.4	1.6	5.0	1.8	5.3
Russian	5.9	3.7		4.0		6.5	4.0		4.4		7.5	4.3		4.9		9.0	5.4		6.2	
Catalan	5.9	3.3		3.4		9.0	4.8		5.1		11.2	5.9		6.4		8.4	4.6		4.9	
Greek	4.8	2.9		3.0		6.0	3.6		3.7		4.5	2.6		2.7		4.6	2.5		2.6	
Albanian	4.8	2.6		3.0		6.7	3.7		4.0		9.1	4.9		5.3		8.4	4.2		4.8	
Bulgarian	4.0	2.6		3.0		7.4	4.7		5.5		10.6	6.8		7.8		11.8	7.2		8.4	
Slavic	4.9	2.3		2.5		6.6	3.4		3.8		12.1	6.5		7.7		9.8	5.0		5.7	
Old Slavic	3.8	2.2		2.7		6.1	3.3		4.3		11.9	6.8		8.7		9.5	5.2		6.0	
Hungarian	2.9	1.8		2.0		5.1	2.9		3.3		7.5	4.3		4.7		7.4	3.7		4.6	
Ruthenian	2.4	1.6		2.0		4.7	3.0		3.7		6.0	3.7		4.4		4.5	2.4		3.0	
Serbian	2.6	1.4		1.6		5.8	3.0		3.4		8.9	5.0		5.5		8.6	5.2		6.0	
Sardinian	1.7	1.0		1.0		3.3	1.7		1.8		4.0	2.0		2.1		2.6	1.4		1.5	

Table 3: Results for the Romanian datasets. In the *D* and *ND* columns we provide the average degrees of similarity for the datasets with and without diacritics. For languages for which we determine cognate pairs (besides etymons), we report both versions of the results, before and after cognate identification. In the *%words* column we provide the percentage of words having an etymon or a cognate pair in each language. The results are ordered according to the ranking of similarity for the corpus comprising the parliamentary debates after identifying cognates and with diacritics included.

Figure 2: Comparison of results for different datasets

4 Word Embeddings and Neural Representations

4.1 Methodology

The RETEROM project focuses on enhancing Romanian NLP through robust word embeddings. Utilizing the CoRoLa corpus—a large collection of contemporary Romanian texts—the project trains embeddings using:

- **CBOW (Continuous Bag-of-Words)**
- **Skip-gram models**

The study addresses the challenge of Romanian's rich morphology by comparing embeddings generated from full word forms with those derived from lemmatized data (both case-sensitive and lowercased).

4.2 Exploration and Evaluation

Embedding models are evaluated using cosine similarity to capture semantic relationships. The study also offers:

- **Text-based queries:** Similarity and analogy tasks (e.g., the classic analogy $king - man + woman = queen$).
- **Graph-based visualizations:** Tools such as t-SNE and interactive network graphs provide intuitive representations of the word vector space.

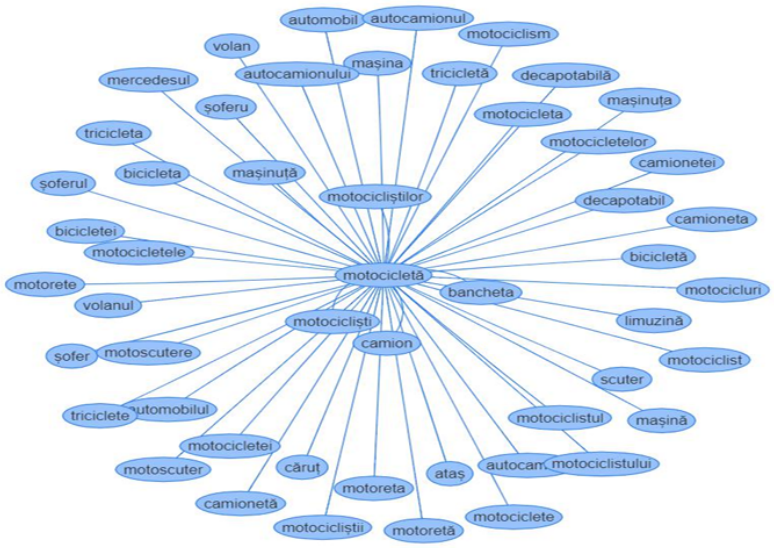


Figure 5. Graph based representation for the word "motocicletă" ("motorcycle") with 50 similar words

Figure 3: Word similarities for "motocicleta"

4.3 Results and Implications

- **Enhanced Semantic Representation:** Lemmatized embeddings mitigate inflectional variability, resulting in clearer semantic clusters.
- **Improved Interpretability:** The models demonstrate robust contextual clustering and facilitate the exploration of semantic relationships.
- **Applications:** These embeddings have been successfully applied in text analysis, information retrieval, and machine translation tasks, further solidifying their utility in Romanian NLP.

5 Comparative Analysis

Aspect	Dist. Semantics	Etymology	Embeddings
Precision	High (tf–idf, PMI)	Moderate	High
Morphology	Limited	Some handling	Strong (lemmatized)
Cross-Lingual	Weak	Strong	Weak
Interpretability	Medium	High	Low
Visualization	Basic	None	Advanced
Resources	Low	Medium	High
Strength	Nuanced context	Historical links	Deep semantics

Table 1: Comparison of Romanian word similarity methods

Explanation:

- **Semantic Precision:** Distributional methods combine multiple metrics to capture nuances, while embeddings excel at context-driven tasks. Etymological methods, though indirect, capture historical and cross-language relationships effectively.
- **Morphology:** Lemmatized embeddings provide an edge by reducing inflectional noise, whereas the etymological approach accounts for underlying orthographic transformations.
- **Cross-Language Applications:** The etymological approach is particularly strong in identifying cognate relations across languages.
- **Interpretability and Visualization:** Etymology-based measures are more transparent while neural embeddings, despite their abstract nature, benefit from modern visualization tools.
- **Resource Requirements:** Embedding models demand extensive corpora and computational resources compared to the lightweight distributional and etymology-aware methods.

6 Conclusion and Future Directions

Each methodology offers unique contributions:

- **Distributional Semantics:** Provides efficient and nuanced context-based similarity measures through alternative metrics.
- **Etymological Similarity:** Leverages historical and orthographic insights to enhance cross-language and linguistic relationship detection.
- **Word Embeddings:** Offer robust semantic representations ideal for modern NLP tasks, especially when incorporating lemmatized data to manage morphological complexity.

Future research could explore hybrid models that integrate distributional statistics, etymological insights, and neural embedding techniques to further enhance Romanian NLP applications such as machine translation, language learning, and digital lexicography.

References

- Ciobanu, A. M., & Dinu, L. P. (2013). *Alternative measures of word relatedness in distributional semantics.*
- Ciobanu, A. M., & Dinu, L. P. (2014). *An etymological approach to cross-language orthographic similarity.*
- Păiş, V., & Tufiş, D. (2020). *More Romanian Word Embeddings from the RETEROM Project.*