Lexical Semantic Change Detection

Nicole Chen, Tammy Yang, Tzu-Yu Ko

 # SemanticChange
 # ComputationalModels
 # NLPResearch
 # AI application

Research Objectives

1. **Detect and quantify lexical semantic changes in Spanish.**
2. **Measure the evolution of Spanish word meanings.**

Introduction

Traditionally, linguists relied on manual hand-annotated word approaches for vocabulary semantic change evaluation, but recent advancements in computer science and computational linguistics have introduced self-driving language models. Leveraging digitized historical documentation and large-scale corpora, this research focuses on building a model for detecting lexical semantic change, aiming to contribute to information access systems in fields such as digital journalism and online chatbots.

Data

We utilized two distinct language corpora: the old corpus (1810-1906) and the modern corpus (1994-2020), and they were processed using spaCy and the target words were selected.
Lexical Semantic Change Detection (LSCD) was applied to identify words experiencing shifts in meaning over time. This approach streamlined the selection of target words and the creation of annotated usage samples.

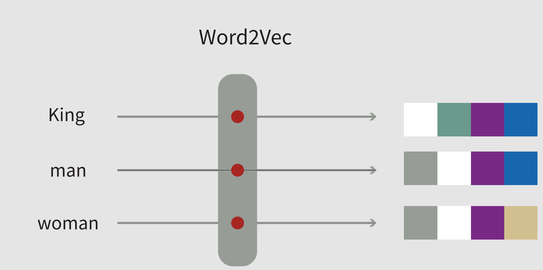
Corpus	Time Period	Tokens
Old Corpus	1810-1906	Around 13M
Modern Corpus	1994-2020	Around 22M

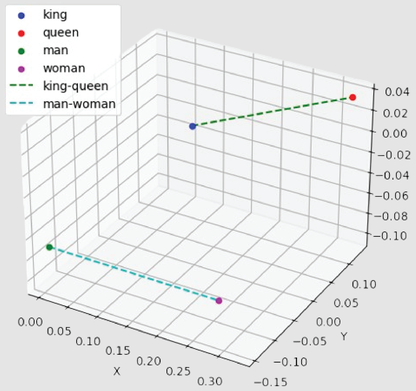
Discussion

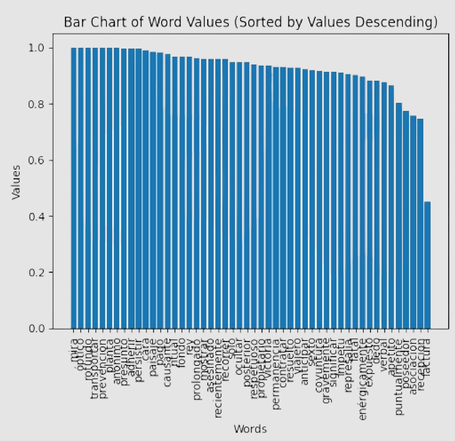
The study reveals semantic changes in words between old and modern eras, opening avenues for future research:
1. Cross-Linguistic Comparative Studies
2. Fine-Tuning for Specific Domains

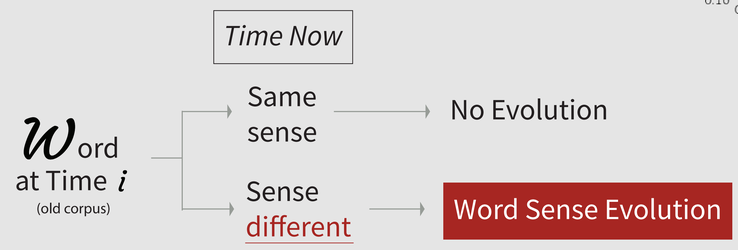
Additionally, the research suggests considering diverse applications beyond textual data, such as audio and visual modalities, for diachronic language data. This opens the door for:
3. Multimodal Semantic Change Detection
4. Implementation in Information Retrieval Systems

Models









Findings

Detect and quantify lexical semantic changes in Spanish

Using Skip-gram with Negative Sampling (SGNS), Orthogonal Procrustes (OP), and Cosine Distance (CD), our approach effectively quantified lexical semantic changes in Spanish. SGNS handled large vocabularies, and OP aligned vector representations, enabling the measurement of graded changes through cosine distance. The method identified shifts in word meanings by comparing vector representations in old and modern datasets. SGNS+OP+CD provided a robust framework for capturing subtle semantic nuances, offering a comprehensive understanding of how Spanish word meanings have evolved over time.

Measure the evolution of Spanish word meanings

To gauge the evolution of Spanish word meanings, we employed Spearman's Correlation Coefficient and COMPARE scores. Spearman's Correlation Coefficient assessed the strength and direction of the monotonic relationship between computed and golden scores, effectively capturing the model's consistency with expected outcomes. The COMPARE score, predicting negated diachronic usage relatedness, provided insights into semantic change by directly measuring relatedness between old and modern language usage. The mixed COMPARE group added depth to our analysis, considering pairs from both time periods.

Measurement	Value
Spearman's Correlation Coefficient	**0.543**
COMPARE	**0.561**