Semantic Text Similarity: A Feature Importance Analysis

Your Name

Universitat Politècnica de Catalunya

November 30, 2024

Outline

Introduction

- Semantic Text Similarity (STS) is crucial for many NLP tasks
- Challenge: Which features best capture semantic similarity?
- Our approach: Unbiased feature analysis using Random Forests

Methodology

- Generated 2000 potential features
- Used Random Forest's feature importance capabilities
- Let the data guide feature selection

Top Features

- Jaccard similarity dominates (7 of top 10)
- Common pipeline steps: lemmatization, stopwords, n-grams
- Top feature accounts for 20% importance

Conclusions

- Simple features can be highly effective
- Pipeline complexity isn't always better
- Character-level analysis with n-grams shows promise