Semantic Textual Similarity

Bruno Sánchez Gómez and María del Carmen Ramírez Trujillo

December 12, 2024

Table of Contents

1. Introduction
2. Methodology
3. Results
4. Conclusion

1. Introduction
2. Methodology
3. Results
4. Conclusion

Introduction

- intro

1. Introduction

2. Methodology

3. Results

4. Conclusion

- Methodology

1. Introduction
2. Methodology
3. Results
4. Conclusion

Model & Feature Overview

- Models: MLP, SVR, RFR.
- Features:
    - Lexical, Syntactic, Strings (individually).
    - Unrestricted (Lexical + Syntactic + Strings).
    - FeatureSelection, based on:
        - Pearson correlation for MLP/SVR
        - Feature importance for RFR
- Performance measured using Pearson correlation with the Gold Standard

Results Summary

Features	**MLP**	**SVR**	**RFR**
Lexical	0.607	0.681	0.728
Syntactic	0.666	0.658	0.661
Strings	0.674	0.676	0.685
Unrestricted	0.652	0.744	**0.757**
FeatureSelection	0.744	0.742	0.745

- Best performance: RFR with Unrestricted (0.757)
- Syntactic features less informative than Lexical/Strings
- Feature combination improves SVR/RFR
- MLP suffers from overfitting

Top Features: Pearson Correlation

Top 5 features based on Pearson correlation with the Gold Standard:

Feature	**Correlation**
lemmas_wn_aug_overlap	0.7233
normal_char_2gram	0.7216
lemmas_char_2gram	0.6902
sw_char_2gram	0.6876
sw_gst_5	0.6666

Three key feature types:

- WordNet-Augmented Overlap (Lexical)
- Character n-grams (String-based)
- Greedy String Tiling (String-based)

Top Features: Feature Importance

Top 5 features based on Feature Importance scores from RFR:

Feature	Importance
lemmas_wn_aug_overlap	0.4325
normal_char_2gram	0.1630
chunk_sim_s	0.0413
lemmas_weighted_overlap	0.0275
normal_char_5gram	0.0199

The top 2 features:

- Common with Pearson correlation table.
- Have significantly higher importance, indicating their dominance in sentence similarity prediction

Feature types: 2 Lexical, 1 Syntactic, 2 Strings-related.

1. Introduction
2. Methodology
3. Results
4. Conclusion

Conclusion

- Best performance: RFR with Unrestricted features (0.757 Pearson correlation).
- Key features: *WordNet-Augmented Overlap* and *Character n-grams*.
- Lexical and String-based features encode most of the relevant information for STS.
- Combining feature types (Lexical, Syntactic, Strings) significantly boosts performance.

References

- D. Bär, C. Biemann, I. Gurevych and T. Zesch. (2012). UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. 435-440.
- F. Sari, G. Glavaš, M. Karan, J. Snajder, B. Dalbelo Bašić. (2012). TakeLab: Systems for Measuring Semantic Text Similarity. Proceedings of SemEval-2012.