Semantic Textual Similarity between sentences

Daniel Sam Pete Thiyagu and Sanketh Kokkodu Balakrishna

Advisors: Prof. Brendan O'Connor, Abe Handler

Comp-Sci 585

Introduction

Problem:

- **Input:** two sentences in English
- **Output:** Score in the range 0-5 indicating similarity

Applications:

- Information Extraction
- Web search queries
- Summarization

Dataset:

- STS-Semeval data from the past 5 years, consisting of around 15000 pairs of sentences along with the Gold tags.

Goals:

- Get similarity scores which match with the Gold scores
- Get scores matching with human notion of similarity of sentences

Approaches

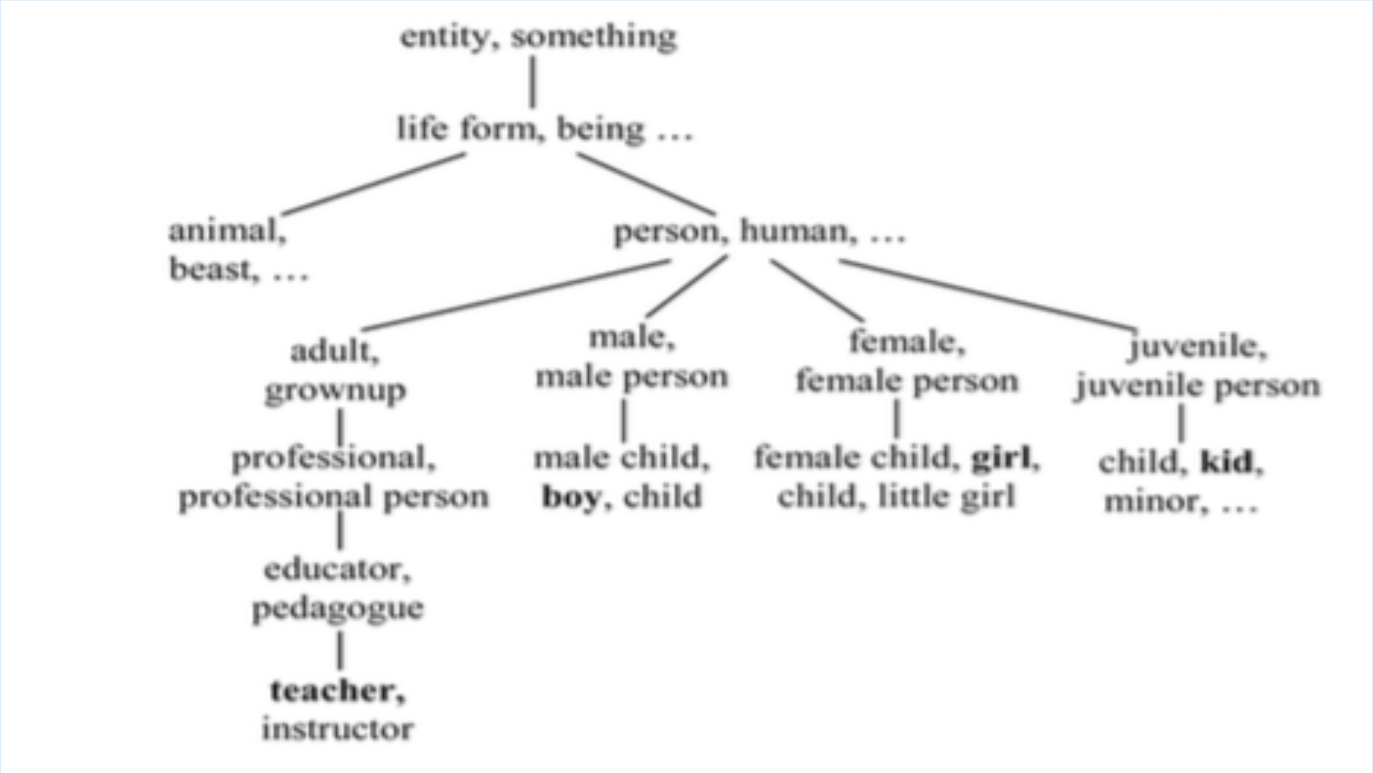
Two approaches we could have taken for this project:

- Stick to one method. Try variations and try to optimize to get the best possible score.
- Try and explore various techniques, to get a general idea of what works best.

We chose the second approach.

Method 1: Similarity using semantic and word order similarity

- Similarity is computed using two similarity measures, as indicated above.
- WordNet's path similarity is used to get a similarity score between two words. Uses the idea of synsets and path between nodes to give the score.



Synsets in WordNet

- Word order similarity depends on the ordering of the words. For example,
  T1: A quick brown dog jumps over the lazy fox.
  T2: A quick brown fox jumps over the lazy dog.
- Word order: r1 = {1 2 3 4 5 6 7 8 9}
  r2 = {1 2 3 9 5 6 7 8 4}

Method 2:Extract lexical and syntactic information from the sentences using:

- Lemma n-gram overlap
- POS n-gram overlap
- Character n-gram overlaps
- TF-IDF Weights

These are then used as features for computation using Jaccard co-efficient and Containment Co-efficient .

Method 3: Linear regression using the above features
Method 4: SVM using above features
Method 5: Multi Layer Perceptron(MLP)

Analysis

sentence pair: "The group is good" , "The group is good".
similarity score: 5/5 or 100%.

- Score as expected.

sentence pair: "The problem is simple" , "The problem is very easy".
similarity score: 2.72/5 or 54.47%.

- No similarity between "very" and any word in the first sentence.

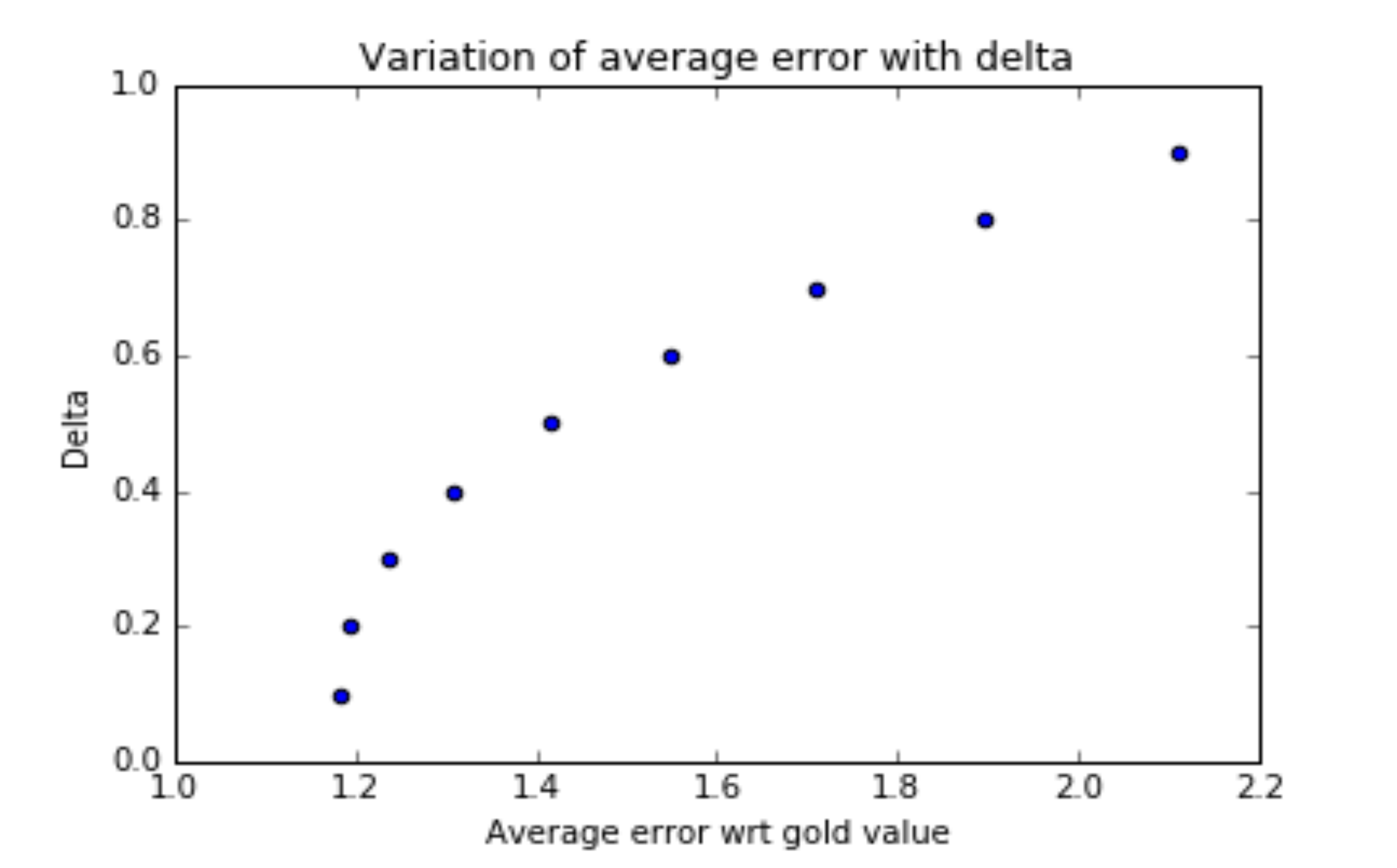
sentence pair: "Today is a Friday" , "That person is not related to me.".
similarity score: 0.69/5 or 13.81%.

- No similarity. Output score based on a common word "is".

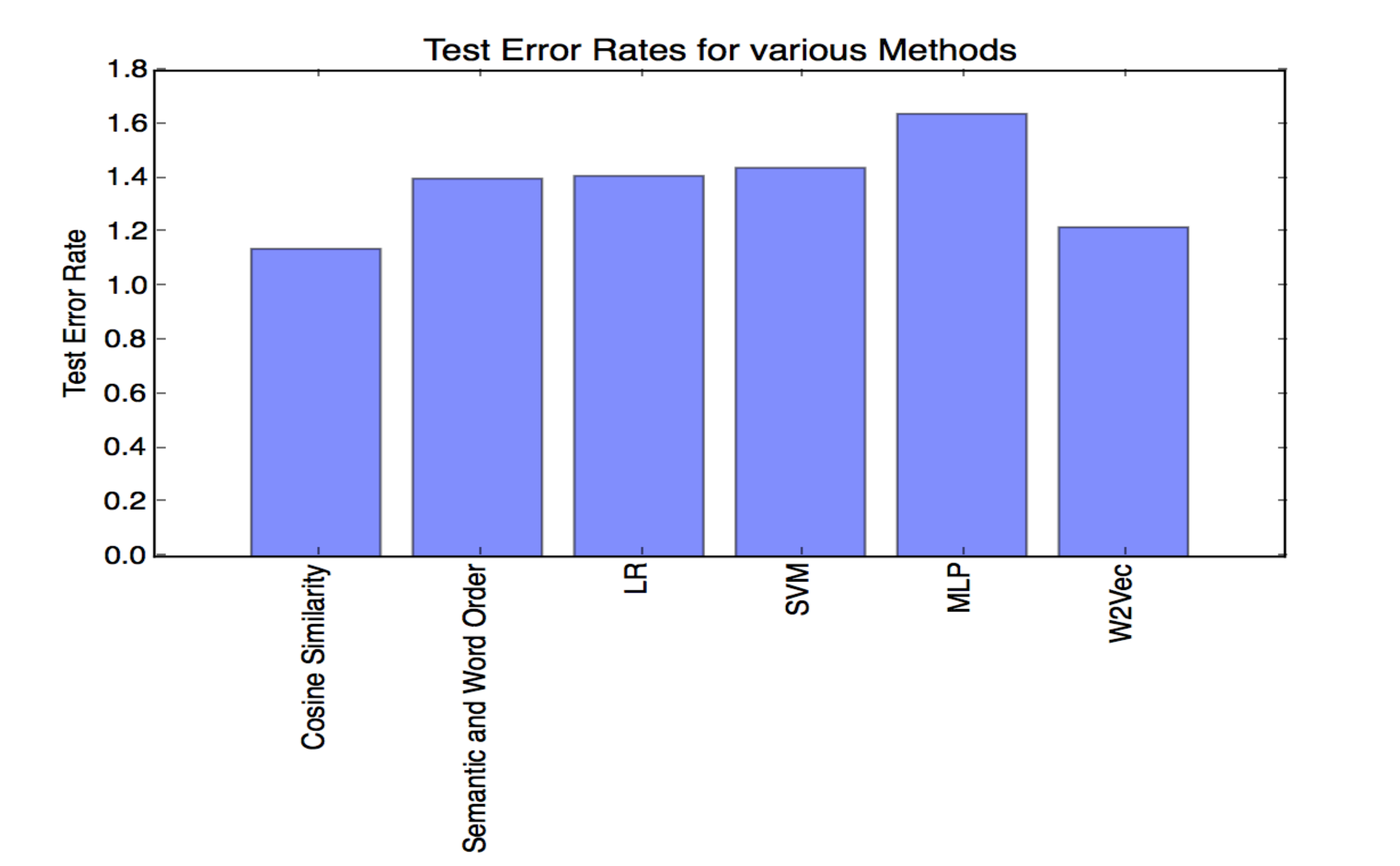
Shortcomings:

- **Phrase error**: sentence pair: "He is a Bachelor" , "He is an unmarried man".
  similarity score: 2.3/5 or 46%.
- **Word sense disambiguation**: sentence pair: "It's an Orange" , "Its Orange".
  similarity score: 3.91/5 or 78.23%.
- Phrase errors can be solved to some extent using Word2Vec similarity measures and a linear combination of vectors using cosine similarity. Sentence pair: "He is a bachelor" and "He is an unmarried man" .
  similarity score 3.5/5 or 70%.

Results



Error variance with increasing weight of semantics



Average Test Errors w.r.t gold scores using different methods

Additional Interests

- Similarity using word alignment.
- Comparing our models with human annotated sentences.

Conclusions

- Our methods did reasonably well, given the ambiguities and structure of English language.
- 65-80% achieved through different methods, when compared to gold standards.
- Gold standards cannot be assumed to be always right.

Bibliography

- Li, Yuhua et al. "Sentence Similarity Based on Semantic Nets and Corpus Statistics." *IEEE Trans. Knowl. Data Eng.* 18 (2006): 1138-1150.
- Brychcin, Tomas and Lukás Svoboda. "UWB at SemEval-2016 Task 1: Semantic Textual Similarity using Lexical, Syntactic, and Semantic Information." *SemEval@NAACL-HLT*(2016).