Semantic Textual Similarity

Mauro Vázquez Chas, Dániel Mácsai

December 12, 2024

Outline

Preprocessing

Preprocessing

- Lowercasing and removing punctuation
- Tokenization with and without stopwords
- Lemmatization with and without stopwords
- Synsets computed by Lesk's algorithm

Basic features

Basic features

- Character ratio, token ratio
- Levenshtein ratio (edit distance)
- Jaccard similarity
  - on tokens with and without stopwords
  - on lemmas with and without stopwords
  - on synsets from Lesk's algorithm

Utilizing synset similarity

Best possible pair

For each word in each sentence, we get the best similarity value considering all of its possible synsets and all the synsets from the words in the other sentence, all of this without modifying its post tags. After this, take the average, considering only the tokens with a valid wordnet pos tag. Methods we used for the synset similarities:

- Path similarity
- Wu-Palmer similarity
- Leacock-Chodorow similarity

Utilizing synset similarity

Using Lesk's algorithm

We use Lesk's algorithm to match synsets to tokens in the two sentences. After, for each synset in a sentence, we compute the highest similarity value using the synsets from the other sentence, and we take the average of these similarities. Methods we used for the synset similarities:

- Path similarity
- Wu-Palmer similarity
- Leacock-Chodorow similarity

Other features

N-grams

- On tokens, lemmas and characters
- Using Jaccard similarity for sets and cosine similarities on histograms

SentiWordNet

For both the positive and the negative scores we compute the following getting two features in total: For each sentence, we sum the scores of each Lesk synset. Afterwards, we substract the values of the sentences and normalize by the maximum number of synsets between the two sentences.

Models and additional ideas

Models

- SVM
- XGBoost
- Random Forest

Additional ideas

- Oversampling SMTeuroparl and MSRpar
- Feature selection with Random Forest based on importance
- Normalizing the features
- PCA on the features

Course of Actions

Initial approach

- Features
  - Basic features (without Levenshtein)
  - Synset best possible pair similarity (Path and Wu-Palmer)
  - Jaccard similarities for tokens and lemmas
  - N-grams jaccard similarities
- Using XGBoost (with a grid search for hyperparameters)

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling
- Added PCA
- Lesk Jaccard similarities

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling
- Added PCA
- Lesk Jaccard similarities
- SentiWordNet

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling
- Added PCA
- Lesk Jaccard similarities
- SentiWordNet
- Added N-Grams occurrence-based similarities

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling
- Added PCA
- Lesk Jaccard similarities
- SentiWordNet
- Added N-Grams occurrence-based similarities
- Ignore stopwords in Jaccard similarities

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling
- Added PCA
- Lesk Jaccard similarities
- SentiWordNet
- Added N-Grams occurrence-based similarities
- Ignore stopwords in Jaccard similarities
- Tried feature selection with Random Forest Importance

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling
- Added PCA
- Lesk Jaccard similarities
- SentiWordNet
- Added N-Grams occurrence-based similarities
- Ignore stopwords in Jaccard similarities
- Tried feature selection with Random Forest Importance
- Removed punctuation

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling
- Added PCA
- Lesk Jaccard similarities
- SentiWordNet
- Added N-Grams occurrence-based similarities
- Ignore stopwords in Jaccard similarities
- Tried feature selection with Random Forest Importance
- Removed punctuation
- Added lesk synset pair similarities

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling
- Added PCA
- Lesk Jaccard similarities
- SentiWordNet
- Added N-Grams occurrence-based similarities
- Ignore stopwords in Jaccard similarities
- Tried feature selection with Random Forest Importance
- Removed punctuation
- Added lesk synset pair similarities
- Tried Scaling Variables

Course of Actions

- Switched to Random Forest (with a grid search for hyperparameters)
- Manual feature selection
- Added Leackock-Chodorow similarity
- Added Resnik and Jiang-Conrath similarities
- Tried Support Vector Regression
- Tried Oversampling
- Added PCA
- Lesk Jaccard similarities
- SentiWordNet
- Added N-Grams occurrence-based similarities
- Ignore stopwords in Jaccard similarities
- Tried feature selection with Random Forest Importance
- Removed punctuation
- Added lesk synset pair similarities
- Tried Scaling Variables
- Removed PCA

Results

Dataset	**Pearson Correlation**
SMTeuroparl	0.5794
MSRvid	0.8375
MSRpar	0.6148
surprise.OnWN	0.7240
surprise.SMTnews	0.5576
All datasets	0.7382

Table: Testing Pearson Correlation on Different Datasets

Takeaways

- Usefulness of the different Features
- Gap with modern techniques
- Importance of cleaning and preprocessing
- Importance of effective implementation
- Importance of the choice of the model

Questions?