Investigating Lexical and Syntactic Differences in Written and Spoken English Corpora

Presented by

Aadya Ranjan
Faizanuddin
IIIT Hyderabad

June 18, 2024

Outline

Introduction

- Disparities between speaking and writing form an important narrative.

Introduction

- Disparities between speaking and writing form an important narrative.
- Challenges in deriving precise algorithms.

Introduction

- Disparities between speaking and writing form an important narrative.
- Challenges in deriving precise algorithms.
- Understanding these differences aids cognitive insights.

Introduction

- Disparities between speaking and writing form an important narrative.
- Challenges in deriving precise algorithms.
- Understanding these differences aids cognitive insights.
- Focus on Syntactic and lexical differences in speeches and writings :

  CoreNLP and BERT(Text analysis)

Introduction

- Research Questions:

Introduction

- Research Questions:
  1. Which syntactic features best differentiate written text from speech transcription?

Introduction

- Research Questions:

  1. Which syntactic features best differentiate written text from speech transcription?
  2. Which lexical features best differentiate written text from speech transcription?

Introduction

- Research Questions:

  1. Which syntactic features best differentiate written text from speech transcription?
  2. Which lexical features best differentiate written text from speech transcription?
  3. Do feature-based algorithms or BERT perform better at this task?

Related Work

- Woolbert (1922) and Olson (1996) explored differences between speaking and writing.

Related Work

- Woolbert (1922) and Olson (1996) explored differences between speaking and writing.
- Fairbanks (1944) and Mann (1944) used type-token ratios and part of speech analysis.

Related Work

- Woolbert (1922) and Olson (1996) explored differences between speaking and writing.
- Fairbanks (1944) and Mann (1944) used type-token ratios and part of speech analysis.
- Biber (1986a,b) analyzed linguistic features, revealing four textual dimensions.

Related Work

- Chafe and Tannen (1987) found variations in involvement and detachment based on context.

Related Work

- Chafe and Tannen (1987) found variations in involvement and detachment based on context.
- Freedman and Krieghbaum (2014-2017) used machine learning to analyze educational dialogues and writing styles.

Datasets

- Transcriptions of presidential speeches and books from George Washington.

Datasets

- Transcriptions of presidential speeches and books from George Washington.
- Texts are preprocessed to remove:
  1. Numbers
  2. Currency values
  3. Excess whitespace
  4. Chunked into 512 tokens using nltk to standardize length.

Features

- Morphological aspects:

Features

- Morphological aspects:
  1. Average syllables per word

Features

- Morphological aspects:
  1. Average syllables per word
  2. Average words per sentence

Features

- Morphological aspects:
  1. Average syllables per word
  2. Average words per sentence
  3. Average characters per word

Features

- Morphological aspects:
  1. Average syllables per word
  2. Average words per sentence
  3. Average characters per word
- Lexical aspects of text:

Features

- Morphological aspects:
  1. Average syllables per word
  2. Average words per sentence
  3. Average characters per word
- Lexical aspects of text:
  1. Lexical diversity

Features

- Morphological aspects:
  1. Average syllables per word
  2. Average words per sentence
  3. Average characters per word
- Lexical aspects of text:
  1. Lexical diversity
  2. Readability

Features

Lexical aspects of sentences:

Features

Lexical aspects of sentences:

1. Number of words in a sentence

Features

Lexical aspects of sentences:

1. Number of words in a sentence
2. Percentage of POS (verb, adjective, noun, adverb, coordinators)

Features

Lexical aspects of sentences:

1. Number of words in a sentence
2. Percentage of POS (verb, adjective, noun, adverb, coordinators)
3. Percentage of personal pronouns (first, second, and third)

Features

Syntactical aspects:

Features

Syntactical aspects:

1. Frequency and percentage of subordinate clauses

Features

Syntactical aspects:

1. Frequency and percentage of subordinate clauses
2. Depth of parse tree

Features

Syntactical aspects:

1. Frequency and percentage of subordinate clauses
2. Depth of parse tree
3. Frequency and percentage of noun phrases

Features

Syntactical aspects:

1. Frequency and percentage of subordinate clauses
2. Depth of parse tree
3. Frequency and percentage of noun phrases
4. Average length of noun phrases

Features

Syntactical aspects:

1. Frequency and percentage of subordinate clauses
2. Depth of parse tree
3. Frequency and percentage of noun phrases
4. Average length of noun phrases
5. Yes/no questions

Features

Syntactical aspects:

1. Frequency and percentage of subordinate clauses
2. Depth of parse tree
3. Frequency and percentage of noun phrases
4. Average length of noun phrases
5. Yes/no questions
6. Direct wh-questions

Features

Syntactical aspects:

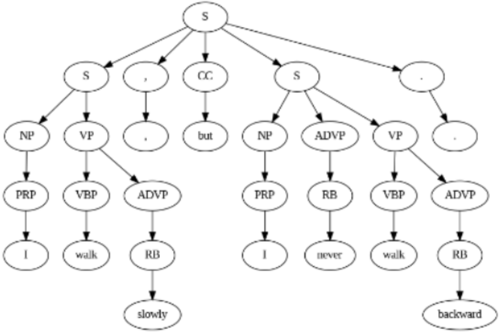
1. Frequency and percentage of subordinate clauses
2. Depth of parse tree
3. Frequency and percentage of noun phrases
4. Average length of noun phrases
5. Yes/no questions
6. Direct wh-questions

Text-level aspects:

1. Sentences

Features

- CoreNLP was used to parse sentences.
- Tree generated by CoreNLP for the sentence -I walk slowly, but I never walk backward.



Experiments

Three experiments were conducted to derive significant features from text data using various machine learning techniques.

- Experiment 1: Parse Trees and Feature Extraction:
  - Extracted features based on sentence parse trees.
  - Used SVM and Random Forest models for classification.
  - Removed highly correlated features (e.g., character count, verb count) to avoid redundancy.

Experiments

Three experiments were conducted to derive significant features from text data using various machine learning techniques.

- Experiment 2: Lexical Diversity and Readability:
  - Calculated metrics such as ARI, Flesch-Kincaid, TTR, and more.
  - Removed outliers (data points greater than 3 standard deviations from the mean).
  - Applied Random Forest to the refined metrics.
  - Retained impactful features like average sentence length, word length, and Maas, which measures lexical diversity unaffected by text length.

Experiments

Three experiments were conducted to derive significant features from text data using various machine learning techniques.

- Experiment 3: Used BERT to distinguish between writing and speaking styles:
  - Trained BERT model on sentences from original data.
  - Focused on transcribed speeches vs written books by US presidents.
  - Demonstrated BERTâs effectiveness, suggesting deep neural networks can enhance text classification.

Results

Experiment 1: Syntactic Features

```
Model Performance:

SVM Accuracy: 54%
Random Forest Accuracy: 61%
RF outperformed SVM and the other models.
```

Key features included length, noun percentage, verb percentage, and parse tree depth etc.

Results

Experiment 2: Lexical Diversity

```
Model Performance:

RF Accuracy with Complexity Metrics: 72.2%
Added Avg_Sentence & Word_Length: 79.2%
After Removing Correlated Features: 87.4%
Only Avg_Sentence & Word_Length: 92.9%
```

Key features such as word length, average sentence length, and Maas effectively distinguish speeches from written text more than complex methods.

Results

Experiment 3: Differentiating with BERT

```
Model Performance:

Accuracy: 90% using the ktrain library.
Batch size = 6, max features = 35,000.
BERT outperformed both SVM and RF
```

Data: Random under-sampling for balance. Split: 80/20. Validation: Random forest (max depth 15) for feature importance.

Tables

Table 1: Evaluation of Syntactic Models and BERT

	Labels	Precision	Recall	F1
SVM	Spoken	58.6%	24.3%	34.4%
	Written	52.2%	82.7%	64.0%
RF	Spoken	60.9%	61.0%	61.0%
	Written	61.0%	60.9%	60.9%
BERT	Spoken	89.9%	90.4%	**90.1%**
	Written	90.6%	90.1%	**90.3%**

Table 3: Hypothesis Testing for Lexical Features

Feature	p-value
maas	1.95e-9
Average Sentence Length	1.31e-5
Average Word Length	1.78e-4

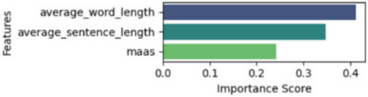
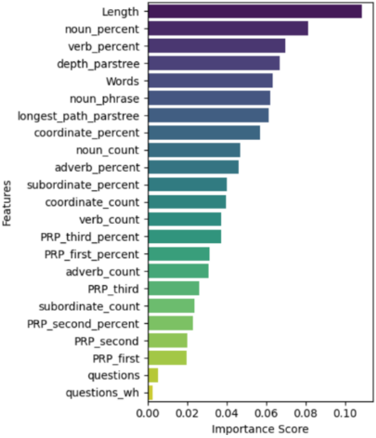


Figure 4: Feature Importance for Lexical Features



Conclusions

- 1. Syntactic Features Distinction:
    - Key features like sentence length and parse tree depth effectively differentiate spoken from written texts.
- 2. Effectiveness of Simple Lexical Metrics:
    - Average word and sentence length outperformed complex metrics, significantly enhancing the accuracy.
- 3. Superiority of BERT:
    - Achieved the highest accuracy, without extensive feature engineering.
    - Traditional models remain valuable for identifying features, aiding in the interpretability of text classification tasks.

Limitations and the Future Work

- Limited Access to Primary Sources and Complexity in Feature Coding
  - Restricted access to presidential books limited dataset diversity.
  - Complex feature coding due to ambiguous definitions.
- Expanding the Dataset and Adding More Linguistic Features:
  - Expand the dataset with more sentences.
  - Add features to improve model accuracy..
  - Develop better AI techniques for distinguishing speech from writing.


Thank you