Project Discussion (and classification example)

CS 485, Fall 2023
Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs485_f23/

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

[Slides by Laure Thompson]

Literary Pattern Recognition: Modernism between Close Reading and Machine Learning

Core Question:

What defines the English haiku in the modern period?

Long & So et al. 2016

Is this an English haiku?

Three spirits came to me
And dew me apart
To where the olive boughs
Lay stripped upon the ground;
Pale carnage beneath bright mist.

Long & So et al. 2016

Is this an English haiku?

Three spirits came to me
And dew me apart
To where the olive boughs
Lay stripped upon the ground;
Pale carnage beneath bright mist.

- It’s short
- It foregrounds a series of images rather than depict a narrative
- Images are drawn from nature

Long & So et al. 2016

The English haiku as statistical pattern

"This is not […] to reinforce the initial distinction we have made, but to **test its boundaries** and **determine what textual patterns are unique** to each group of texts."

Long & So et al. 2016

Dataset

Haiku – 400 poems

· A translation from a seminal text

· Self - identified as a haiku
i.e., “haiku” in title

· Identified explicitly as influence by Japanese short verse forms

· 2 categories: translation, adaptation

Non - Haiku – 1900+ poems

· Short poems from magazines during the later phases of the haiku’s reception
e.g., *Poetry Magazine* , *Harper’s Magazine* , *Lyric West*

· Short: <300 characters

Long & So et al. 2016

Features

Poem as Raw Text

So cold I cannot sleep; and as
I cannot sleep, I'm colder still.

Author Unknown; A 1902 translation by Basil Hall Chamberlain

Poem as a tokenized "bag-of-words"

['so', 'cold', 'i', 'can', 'not', 'sleep', 'and', 'as', 'i', 'can', 'not', 'sleep', 'i'm', 'colder', 'still']

Poem as "bag-of-words" without stopwords (i.e., function words)

['so', 'cold', 'sleep', 'colder', 'still']

Poem as labeled feature set (note that word-order is irrelevant)

[{'cold': True, 'colder': True, 'less_than_20_syl': True, 'sleep': True, 'still': True, 'so': True}, 'haiku']

FIGURE 4. Machine interpretable representations of a single haiku text. Note in the final representation that each feature is assigned a value of "True," indicating its presence in the original text. "Haiku" is the label assigned to the feature vector.

Long & So et al. 2016

Feature Analysis

```
sky = True          not-ha : haiku  =     5.7 : 1.0
shall = True        not-ha : haiku  =     5.0 : 1.0
sea = True          not-ha : haiku  =     5.0 : 1.0
man = True          not-ha : haiku  =     4.3 : 1.0
last = True         not-ha : haiku  =     3.7 : 1.0
snow = True         haiku : not-ha =      3.7 : 1.0
earth = True        not-ha : haiku  =     3.7 : 1.0
blue = True         not-ha : haiku  =     3.7 : 1.0
pass = True         not-ha : haiku  =     3.7 : 1.0
voice = True        haiku : not-ha =      3.7 : 1.0
white = True        not-ha : haiku  =     3.0 : 1.0
house = True        haiku : not-ha =      3.0 : 1.0
```

Long & So et al. 2016

Initial Results

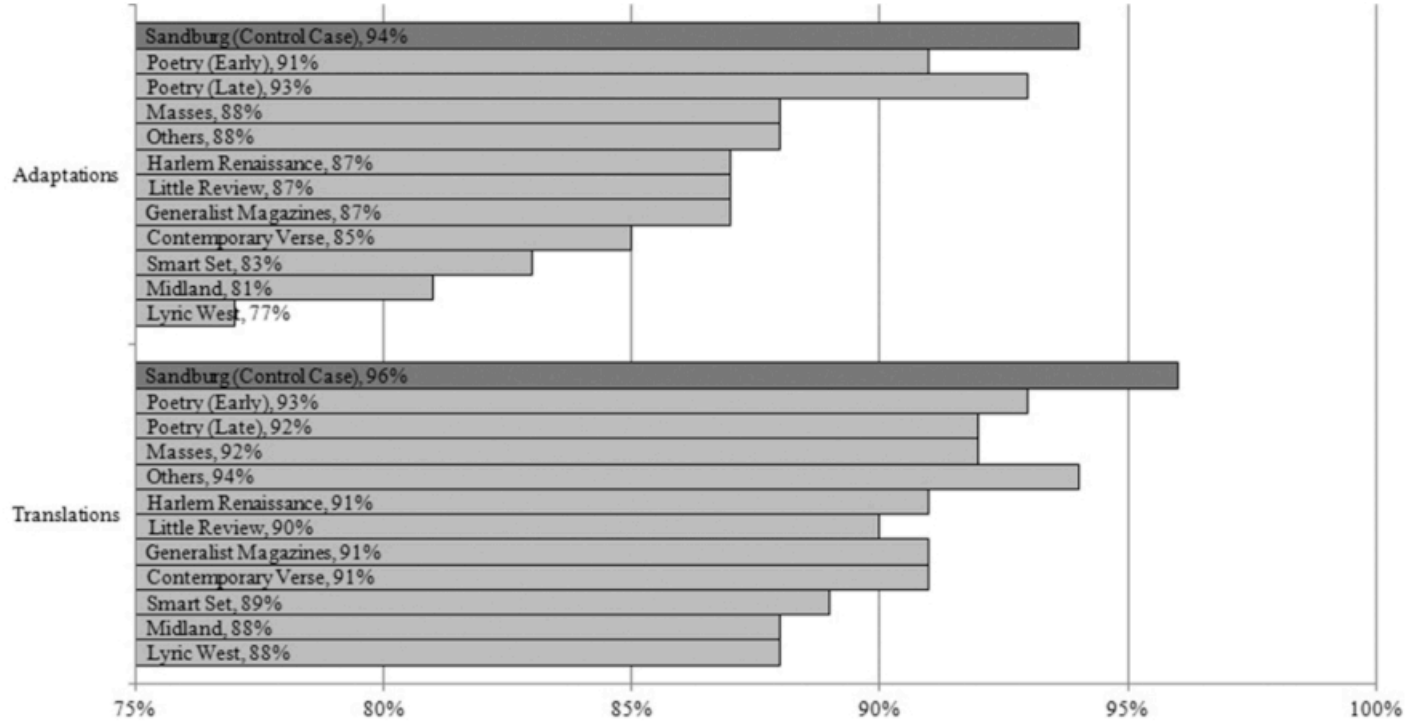


FIGURE 6. Average accuracy scores for one hundred classification tests. The top portion gives the scores for adapted haiku classified against the various short-poem corpora. The bottom portion gives the scores for the translated haiku.

Long & So et al. 2016

On Errors

“Rather than correct for the error, what if we consider how it **troubles the initial categorical distinction** built into the procedure? Or better yet, try to generate similar errors so as to **blur the distinction** ?”

“What the machine learning literature treats as misclassifications, then, we treat as **opportunities for interpretation** .”

Long & So et al. 2016

Misclassified Poems: Haiku in Waiting

Rain rings break on the pool
And white rain drips from the reeds
Which shake and murmur and bend;
The wind - tossed wistaria falls.

The read- beaked water fowl
Cower beneath the lily leaves;
And a grey bee, stunned by the storm,
Clings to my sleeve.

Long & So et al. 2016

Misclassified Poems: Machine Haiku

When she turns her head sidewise;
The line of her chin and throat
Running down her shoulder
Is as graceful as the undulating motion of the neck of a peacock
Is as smooth as the petals of a Marechel Niel rose.
And her voice
Sounds like a man
Cleaning the rust out of a boiler.

Long & So et al. 2016

Misclassified Poems: In Between

Out of the granite rock I've wrested life;
Fending the storm I've strengthened root and limb,
Crouching, I hold the plunging chasm's rim,
As I have braved a thousand years of strife.

Long & So et al. 2016

Final Projects

https://people.cs.umass.edu/~brenocon/cs485_f23/project.html

Project Overview

Investigate, analyze, and come to research findings about new methods, or insights on previously existing methods.

In groups of 2 - 4, you will either *build* a natural language processing system or *apply* them to some task.

Your project must: (1) use or develop a dataset, and

(2) report empirical results/analyses with this dataset

Project Components

Proposal: A 2 page document outlining the problem, your approach, possible dataset(s) and/or software systems to use.

Progress Report: A 4 - 8 page document that describes your preliminary work and results

Presentation: An opportunity to present your near- complete project to the class.

Final Report: An 8 - 12 page document that describes your project and final results.

Where to start

- What *core question(s)* are you trying to answer?
- How will you *operationalize* this question?
- What work are you building off of? What has been done before?
- What experiments will you run?
- How will you measure the success of these experiments?
  e.g., held - out accuracy, error analysis, manual evaluation, etc.

Where to look for related work?

NLP research papers:

- The ACL Anthology is a good place to start
- Some Resources:
  - On how to read research papers
  - On navigating the NLP research space

How to search for papers

- Search keywords in the ACL anthology, Google Scholar, Semantic Scholar
- Look at the papers that a paper references and those that cite it
- Examine other papers by a given author and their lab

Where to look for related work?

A standard web search can also be useful for finding…

- Research blog posts
- Datasets
- Related codebases
- Recorded Talks
- …and more!

Choice of emphasis

- Implementing and developing algorithms and features
- Defining a new linguistic / text analysis task, and tackling it with off - the- shelf NLP software
- Collect and explore a new textual dataset to address research hypotheses about it

A large variety of tasks

Detection Tasks

Classification Tasks

Prediction Tasks

- Predict external information from text (e.g. movie revenue, post popularity, stock volatility, etc.)

Structured Linguistic Prediction

- Relation, event extraction
- Narrative chain extraction
- Parsing

Text Generation Tasks

- Machine Translation
- Summarization & Normalization
- Poetry / Lyric generation

End - to - End Systems

- Question Answering
- Conversational dialogue systems

Visualization & Exploration

- Temporal analysis of events
- Topic modeling & clustering

For more dataset and task ideas

- See resources from Tues 10/3
- Shared task websites
  - SemEval: Series of semantic evaluation tasks.
    - SemEval 2023 tasks, 2022, 2021, etc.
      There may be access to data!
  - CoNLL shared tasks

Some projects from recent years

Text Classification

- Song genre classification using lyrics
- Comparing models for multi - labeled classification of book genres
- Distinguishing between 19 th and 20 th century literature
- Predicting political slant in news comments
- Classification of political views on Reddit
- Classifying BBC news articles into their section/category types
- Language classification

Some projects from recent years

Detection Tasks

· Paraphrase detection

· Toxicity level detection in social media posts

Prediction Tasks

· Estimating stock volatility from news articles

· r/ AmITheAsshole verdict prediction

· Predicting tweet popularity

Text Generation Tasks

· Text summarization for lectures

End - to - End Systems

· FAQ answering

· Medical diagnosis chatbot

Visualization & Exploration

· Sentiment analysis of songs throughout time

· Sentiment analysis of r/ wallstreetbets

JEOPARDY! Category Analysis

Evan Risas & Alisa Kotliarova

Task: Analyze each question-answer pair to determine which broad category it most closely fits, then predict category frequency for future Jeopardy games.

Dataset

200,000+ Jeopardy! Question & Answer pairs

Approach

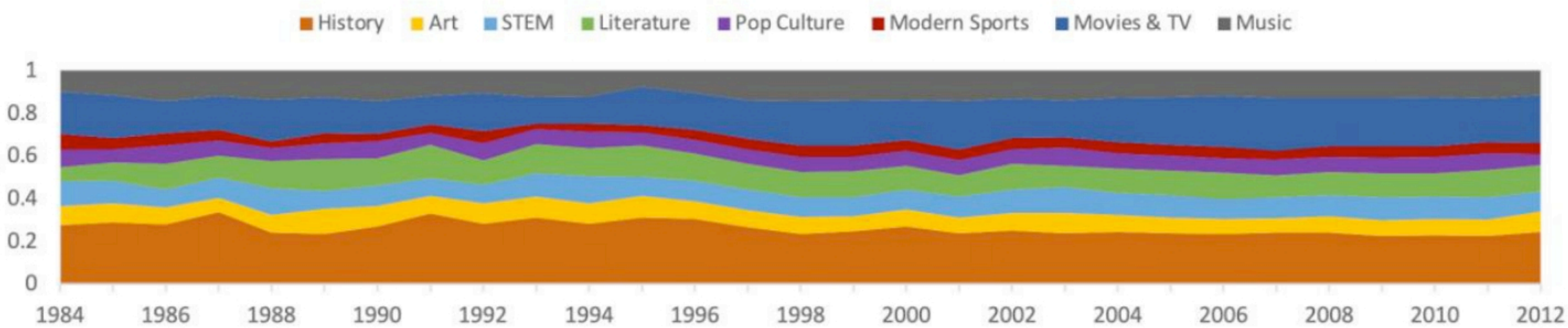
Classify into custom categories using NLP model built on Word2Vec

Observe

Examine category popularity over time

Predict

Predict future category frequency



Brainstorming Session

News article: classify harmful info / misinfo / propaganda
User interface???

Classify/extract values from a business's "About" page

Predict popularity of fan fiction

Sentiment over time - social media, real-world events

Translations to emojis / twitch emotes

AITA prediction

AI Text classification

text generation

What characteristics of movies make them successful?
By genre, etc.

[notes from lecture]