Novel2Vec: Characterising 19th Century Literature via Word Embeddings

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene

20 September 2016
A unique collaboration between the Insight Centre and Humanities Institute in UCD.

Our Objective

Study a collection of Irish and British novels from the 19th Century using both quantitative and qualitative methodologies to examine works in new ways.
This presentation will focus on our most recent work: The exploration of our textual corpus through the lens of word embeddings.

1. What are word embeddings?

2. How we’ve applied them to our corpus.

3. What insights have we gained as a result?
Computers aren’t adept at understanding natural language like humans.

Therefore, we convert words into vectors.

These vector representations of words are called **Word Embeddings**

\[
\text{vec(“king”) } - \text{ vec(“man”) } + \text{ vec(“woman”) } = \text{ vec(“queen”)}
\]

They can capture the underlying similarity between words and their semantic properties.
19th Century Word Embeddings

Novels Selected

- 6 novels by Jane Austen
- 3 by Charles Dickens - Bleak House, Oliver Twist, Great Expectations
- 3 by Arthur Conan Doyle - First 3 novels in the Sherlock Holmes series.
Methods: Data Preprocessing

- The text of each chapter of a novel is annotated by a literary scholar to identify all characters and their aliases.

"Don't keep coughing so, Kitty Bennet, for heaven's sake! Have a little compassion on my nerves. You tear them to pieces."

"Kitty Bennet has no discretion in her coughs," said her father (Mr. Bennet); "she times them ill."

"I do not cough for my own amusement," replied Kitty Bennet fretfully.

"When is your next ball to be, Elizabeth Bennet?"

- A character dictionary is then created, mapping all aliases for a character to their definitive name.
19th Century Word Embeddings

Methods: Data Preprocessing

- Part-of-speech tagging (POS tagging) was applied to each text using the Natural Language Toolkit (NLTK) PerceptronTagger Implementation.

```
1) DT the
    2) JJ little
    3) JJ yellow
    4) NN dog

5) VBD barked
6) IN at
7) DT the
8) NN cat

S
```
19th Century Word Embeddings

Methods: Word Embedding Generation

- **Word2vec** is a two-layer neural network that processes text into a set of feature vectors distributed within a dense dimensional space.
- For our purposes, we used the **Gensim** implementation.

- We visualised the generated word embeddings in 2 dimensional space using the dimensionality reduction technique known as **t-SNE** initialised with **PCA**.
19th Century Word Embeddings
Word2Vec: CBOW Versus Skip-Gram NS
19th Century Word Embeddings

Resulting Word Embeddings - Austen

CBOW 2

CBOW 5

SG 2

SG 5
19th Century Word Embeddings

Context Window Sensitivity - Dickens Corpus

Silhouette Score

-0.12
-0.08
-0.04
0
0.04
0.08
0.12
0.16

1f
3f
5f
10f
50f
100f

sg-5
sg-2
cbow-5
cbow-2

Δ sg
Δ cbow
19th Century Word Embeddings

Context Window Sensitivity

Silhouette Score

- Austen
- Dickens
- Doyle
- All

Silhouette Score Values:
- Δ sg: 2,5
- Δ cbow: 2,5

Austen: Δ sg: 2,5 = 0.14
Dickens: Δ cbow: 2,5 = 0.07
Doyle: Δ sg: 2,5 = 0.105
All: Δ cbow: 2,5 = 0.14

Insight Centre for Data Analytics
Siobhán Grayson | UCD
The 6 protagonists of Austen can be found grouped together.
Results

Charles Dickens

Unlike Austen, the protagonists within our Dickens dataset do not group together.
Results

Arthur Conan Doyle

Sherlock Holmes from the first book in the series does not map into the same embedding space as later versions of himself.
Results

Aggregated: Austen, Dickens, Doyle
Conclusions and Future Work

• We have generated, visualised, and explored word embedding representations for four different datasets consisting of 12 popular 19th century novels.

• Our results suggest that word embeddings can potentially act as a useful tool in supporting quantitative literary analysis.

• Providing new ways of representing and visualising well-known literary texts that complement traditional “close reading” techniques.

Future Work

• In future work, we hope to extend our analysis to diachronic word embeddings to discover how word usage within our corpus changes over time.
References


