Novel2Vec: Characterising 19th Century Fiction via Word Embeddings

Siobhán Grayson[1], Maria Mulvany[2], Karen Wade[2], Gerardine Meaney[2], and Derek Greene[1]

[1] School of Computer Science, University College Dublin, Ireland
{siobhan.grayson, derek.greene}@insight-centre.org
[2] Humanities Institute, University College Dublin, Ireland
{maria.mulvany, karen.wade, gerardine.meaney}@ucd.ie

Abstract. Recently, considerable attention has been paid to word embedding algorithms inspired by neural network models. Given a large textual corpus, these algorithms attempt to derive a set of vectors which represent the corpus vocabulary in a new embedded space. This representation can provide a useful means of measuring the underlying similarity between words. Here we investigate this property in the context of annotated texts of 19th-century fiction by the authors Jane Austen, Charles Dickens, and Arthur Conan Doyle. We demonstrate that building word embeddings on these texts can provide us with an insight into how characters group differently under different conditions, allowing us to make comparisons across different novels and authors. These results suggest that word embeddings can potentially provide a useful tool in supporting quantitative literary analysis.

1 Introduction

Within the last decade, substantial advances have been made within the field of computational linguistics, due in part to the evolution of neural networks. One particular natural language application of neural networks that has amassed considerable attention involves the use of *word embeddings*, where the original words from a corpus are mapped to corresponding vectors in a new high-dimensional space. We can subsequently analyse the associations between pairs or clusters of words within this space. The most popular approach in the literature has been *word2vec* [14], which uses a two-layer neural network model to learn from word contexts and transform the words in a corpus to a new set of vectors. This allows for the detection of contextually similar words without human intervention, since words that share common contexts will also have similar vectors in the new space which will be located close to one another. Using these concepts, word2vec has been incorporated into an extensive number of natural language processing applications (*e.g.* [15, 23]).

In parallel to the advances in NLP, an increasing number of humanities scholars are seeking to complement their literary research by incorporating computational techniques to provide alternative perspectives [9]. This particularly

benefits scholars who are interested in 'distant reading' [16], the practice of understanding literature from a macro-level viewpoint, as opposed to exclusively from a traditional micro-level 'close reading' standpoint. Distant reading offers new ways to challenge assumptions about genre, narrative and other aspects of literature, by facilitating the analysis of large-scale collections of literary works. Numerous approaches have been proposed and tested for this purpose, including those based on statistical topic models [10], character profiling [6], character frequency analysis [5, 22], and sentiment analysis [4].

In this paper, we explore the use of word embeddings to analyse four different datasets compiled from twelve popular 19th century novels written by the authors Jane Austen, Charles Dickens, and Arthur Conan Doyle. We compile these datasets from texts that have been manually annotated to include definitive character names[1]. In Section 3.3 we describe the construction of high-dimensional embedded spaces, which are created using the aggregation of texts on a per-author basis. We consider the effect of applying two variants of word2vec, a continuous bag-of-words strategy and a skip-gram strategy, on the extent to which these spaces exhibit a tendency to cluster words that are syntactically related rather than semantically related. In Section 4 we discuss the resulting word2vec models in relation to characterisation, by examining how the names of characters are positioned in the new embedded spaces. Finally, we consider constructing a single embedding which represents all twelve texts. This allows us to further examine the clustering of characters, especially in relation to novel protagonists, to learn whether characters are more likely to group if written by the same author or from the same book than by their role or function. Our results suggest that word embeddings, such as those generated by word2vec, can potentially provide a new way of studying well-known literary texts.

2 Related Work

2.1 Word Embeddings

In the general areas of natural language processing and text mining, the study of word co-occurrences has often been used to identify the linkages between words in unstructured texts. The motivation for this type of analysis comes from the distributional hypothesis in linguistics, which states that "a word is characterised by the company it keeps". The general goal of co-occurrence analysis is to quantify how words occur together, which can in turn help us to uncover the associative and semantic relations between words [3]. Such analysis can also help to support subsequent analysis tasks, such as topic modelling and data visualisation.

The relationship between pairs of words occurring within a fixed-size context window is a key component of popular word embedding methods such as word2vec [14]. Word2vec is essentially a shallow, two-layer neural network that

[1] The annotated texts were created as part of the "Nation, Gender, Genre" project. See `http://www.nggprojectucd.ie`

transforms textural data into a set of vector representations, each corresponding to a word distributed within the original high-dimensional feature space. These vectors typically have 50–300 dimensions, where the dimensionality is specified by the user. By training on a sufficiently large and coherent corpus of text, the idea is that the resulting model should provide a vector space where words with similar meanings are mapped to similar positions in that space. The models produced by word2vec have been used in a range of natural language processing applications, including machine translation [15], sentiment analysis [23], and topic modeling [17]. In more recent work, the word2vec approach was extended to learn from sentences as well as individual words, and has also been used to measure the similarity of entire documents [11]. Other similar embedding approaches have also been proposed such as *GloVe*, which also constructs a new representation for each word by aggregating pairwise word co-occurrence statistics from a corpus of text [18].

2.2 Analysis of Literary Texts

A range of computational methods have recently been applied to the quantitative study of literary texts. Notably, Moretti [16] analysed the plot structure of the works of Shakespeare by examining the interactions between characters on basis of shared dialogue. Several authors have focused on the problem of analysing literary texts at a macro level, without close reference to the texts themselves. Jockers and Mimno [10] applied topic modeling techniques to a corpus of 3,346 works of 19th-century fiction to identify broad themes common across the corpus in order to support distant reading. More recently, Reagan *et al.* [19] applied sentiment analysis to a collection of over 1,700 works of fiction from Project Gutenberg. By analysing the emotional arcs in these novels (*i.e.* trajectories of emotional valence), they identified six basic plot types with characteristic arcs.

Preliminary work has been done in applying word embedding methods to fictional texts to support literary analysis. This work has included a short analysis of word associations produced by a word2vec model built on 18th-century texts[2], and a visualisation of the nouns appearing in Jane Austen's *Pride and Prejudice*, generated using word2vec and the t-SNE visualisation method[3]. However, to the best of our knowledge, no previous studies have looked at using word embeddings to analyse and visualise 19th-century texts in any detail.

3 Methods

3.1 Data Preprocessing

In this paper we consider a collection of twelve novels from three 19th century novelists - six by Jane Austen, three by Charles Dickens, and three by Arthur Conan Doyle - sourced from Project Gutenberg. Initial data preparation involves

[2] `http://ryanheuser.org/word-vectors-1`
[3] `http://www.ghostweather.com/files/word2vecpride`

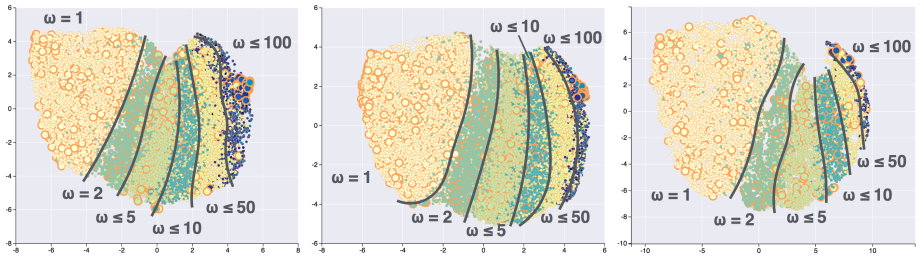
Table 1: Summary of character, word, sentence, and chapter numbers for each novel within this study.

Author	Novel	*#Chars.*	*#Words*	*#Sents.*	*#Chpts.*
Jane Austen	Northanger Abbey	94	75153	3523	31
	Pride and Prejudice	117	120262	5679	61
	Persuasion	136	81809	3606	24
	Sense and Sensibility	158	118149	4796	50
	Emma	193	156364	8438	55
	Mansfield Park	218	157800	6824	48
	Total	**916**	**709537**	**32866**	**269**
Charles Dickens	Oliver Twist	286	153990	8973	53
	Great Expectations	288	177043	9720	59
	Bleak House	516	341441	20292	67
	Total	**1090**	**672474**	**38985**	**179**
Arthur Conan Doyle	A Study in Scarlet	130	42497	2679	14
	The Sign of the Four	127	42410	2911	12
	The Hound of the Baskervilles	126	62448	3861	15
	Total	**383**	**147355**	**9451**	**41**
All	**Total**	**2389**	**1529366**	**81302**	**489**

the manual annotation of the novels, where literary scholars identify all character references in the text of each novel as described in [7]. For each annotated text, no two characters share the same definitive name. However, as each text is annotated in isolation, it is possible that characters from different novels may have the same name. For instance, the character Sherlock Holmes is present within three of our twelve novels where for each he is identified by the same definitive name. Thus, in order to ensure that each characterisation of Sherlock Holmes is represented separately, a unique identifier is assigned depending on which novel the character appears. This allows all characters from all novels to be distinguishable from each other. Part-of-speech tagging (POS tagging) was then applied to each text using the Natural Language Toolkit (NLTK) [2] PerceptronTagger implementation. This was to facilitate syntactical comparisons between the eventual *novel2vec* models. Finally, we divided our corpus into four different datasets, one for each collection of novels by author, and one consisting of all twelve novels compiled together. A summary of each dataset can be found in Table 1.

3.2 Word Embedding Generation

As outlined in Section 2.1, word2vec is a two-layer neural network that processes text into a set of feature vectors distributed within a high-dimensional space. Two different approaches exist within word2vec itself, the Continuous Bag-of-Words (CBOW) model and the Skip-Gram (SG) model. The essential difference between these models is in how they implement predictions. CBOW predicts words on the basis of the context in which they occur, i.e. the group of words which surrounds a given word. By contrast, SG predicts a target context for a given word. For the purposes of converting our textual datasets into



(a) Austen (b) Dickens (c) Doyle

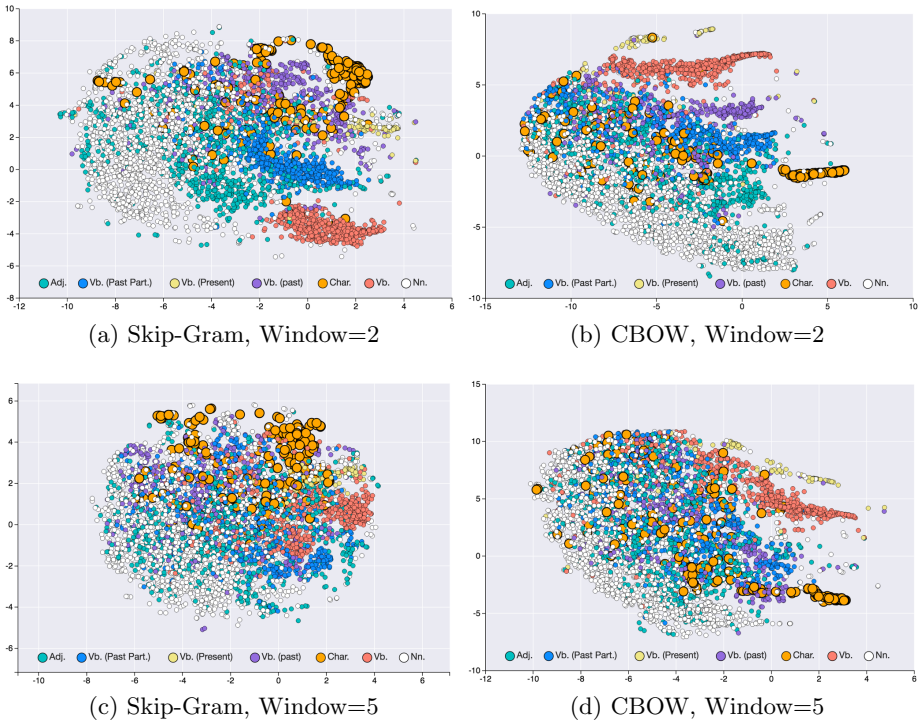
Fig. 1: (a) Austen (b) Dickens (c) Doyle ω =word frequency which increases as we move from left to right across each graph. All skip-gram with context window size 5.

vector word embeddings, we employ the Gensim [20] word2vec implementation. Word embeddings were generated using both the CBOW and SG models with 300 NN layers, and for context windows of size 2 and 5 in each case. This was repeated for different minimum word frequency counts (f_{min}) within the set $f_{min} = \{1, 3, 5, 10, 50, 100\}$. All other parameters were left at their default settings. We then visualised each of the generated word embeddings by reducing the learned vectors dimensionalities into 2D space using the dimensionality reduction technique known as t-Distributed Stochastic Neighbour Embedding (t-SNE) [13]. Although t-SNE is fundamentally probabilistic, initialising it with PCA stabilised how groups were spatially arranged across runs. An example of the resulting word embedding visualisation for skip-gram with context window of size 5 models, SG-5, is displayed in Fig. 1 for three of our datasets where $f_{min=1}$.

3.3 Evaluation

In Sections 3.1 and 3.2, we describe the preprocessing techniques applied to our corpus before compiling our model training datasets. One of the techniques, POS tagging, assigns each word in the text to its most likely grammatical class. In Fig. 2, SG and CBOW ($f_{min} = 5$) generated word embeddings for our Austen corpus have been visualised using t-SNE and are coloured according to their tagged grammatical class[4]. From visually inspecting each mapping within Fig. 2 we can see that a certain amount of clustering is occurring based on syntactical similarities, with the quality of clustering depreciating as the context window increases from 2 to 5. The difference between context window clustering is most noticeable within our skip-gram model embeddings, Fig. 2(a) and Fig. 2(c), while

[4] Clustering evaluated using word embeddings that are apart of the 6 highest frequently occurring POS tags within our corpus, along with independently tagged character embeddings. The POS tag that an embedding belongs to is indicated by the colour of the node within our tSNE visualisation.



(a) Skip-Gram, Window=2 (b) CBOW, Window=2

(c) Skip-Gram, Window=5 (d) CBOW, Window=5

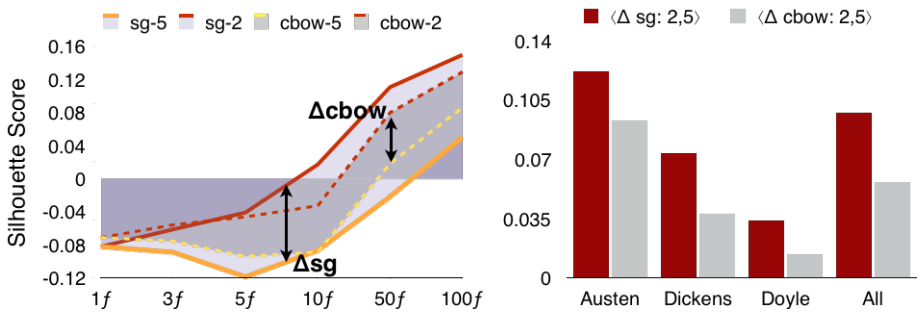
Fig. 2: Word embedding visualisation for the Austen dataset for $f_{min} = 5$ where words embeddings are coloured according to their grammatical class. Adjective: Green, Verb (Past Participle): Blue, Verb (Present): Yellow, Verb (Past): Purple, Character: Large Orange nodes, Verb: Red, Noun: White.

our CBOW mappings remain relatively unchanged in comparison, 2(b) and Fig. 2(d).

In order to quantify the extent to which syntactic clustering is occurring we have computed the mean silhouette coefficients for each of our models. The silhouette coefficient quantifies the quality of clustering by evaluating the cohesion and separation of clusters [21]. Cohesion is calculated using the mean intra-cluster distance, where the average pairwise distance, $\langle D_a \rangle$, is computed for a point p between all other points within the same cluster C_a. The separation between clusters is then found by finding the average pairwise distance, $\langle D_n \rangle$, between p and all points within the nearest neighbouring cluster C_n. The silhouette coefficient for point p is then defined as

$$S = \langle D_n \rangle - \langle D_a \rangle / max\{\langle D_a \rangle, \langle D_n \rangle\}$$

The silhouette score for each model is then computed by finding the mean silhouette coefficient for all samples. We can then use this to evaluate the context



(a) Dickens Dataset Scores (b) Overall Average Scores

Fig. 3: Context window sensitivity comparison between SG and CBOW models.

window sensitivity by finding the absolute difference between S computed for SG-2 and SG-5, and for CBOW-2 and CBOW-5, which we in turn use to find the average context window sensitivity of $|\Delta SG : 2, 5|$ and $|\Delta CBOW : 2, 5|$ across all f_{min} models, for each of our datasets, Fig. 3(b).

4 Results and Discussion

Understanding whether a model places syntactically or semantically similar words close together has been analysed extrinsically in previous work where models are evaluated using sets of 'question-word' analogy tasks [15, 1, 12]. In this paper, we take a different, slightly more intrinsic approach [24], by accessing how our models embed similar words together via cluster analysis. As described in Section 3.3, each model appears to vary in how much it syntactically clusters word embeddings. In order to quantify the extent to which syntactic clustering is occurring and compare the context window sensitivity of each model, we have computed the silhouette score for each case. In Fig. 3(a), the silhouette score for each model ($SG_{k=2}$, $SG_{k=5}$, $CBOW_{k=2}$, $CBOW_{k=5}$) is graphed across the range of different minimum word frequencies ($f_{min} = \{1, 2, 5, 10, 50, 100\}$) applied to our Dickens dataset. A silhouette score ranges between 1 and -1, where 1 is a perfect clustering score and -1 the lowest, suggesting that points have been assigned to incorrect clusters. Values close to 0 indicate that clusters are overlapping [21].

As we can see for $f_{min} = 1$, each model performs similarly, with $S \simeq -0.08$. As we filter out lower frequency words, cluster performance improves for SG-2 and CBOW-2, whilst performance of SG-5 and CBOW-5 decreases and only improves for each after $f_{min} = 10$. This in itself is an interesting result, demonstrating how $k = 5$ models still perform relatively poorly on clustering vectors grammatically in comparison to their $k = 2$ model counterparts, despite lower frequency, potentially noisy words being filtered out. Another observation is that each of our SG models are more sensitive to the application of different context

(a) SG, Window=2, $f_{min} = 5$ (b) SG, Window=5, $f_{min} = 5$

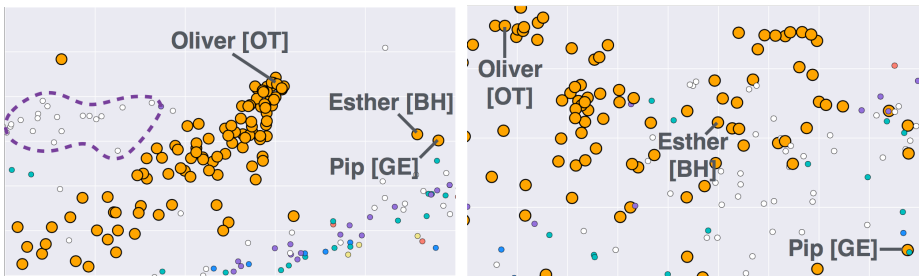
Fig. 4: Character embeddings from the Austen dataset.

window sizes, where the absolute difference between silhouette scores for SG-2 and SG-5 (ΔSG) are greater than the absolute difference between silhouette scores for CBOW-2 and CBOW-5 ($\Delta CBOW$), Fig. 3(a). This trend is present throughout each of our datasets, Fig. 3(b), suggesting that SG is capturing the nuances of our corpus better than CBOW models. This finding aligns with what has previously been observed by Mikolov et. al. [14], who note that SG models work well with small training sets and rare words, while CBOW performs slightly better for frequent words. Thus, taking these findings into account, we will now take a closer look at each of our SG character embeddings for $f_{min} = 5$[5] to see the difference between SG-2 syntactical similarities and learn whether SG-5 identifies semantic similarities within the context of our corpus.

4.1 Character Embedding Comparisons

Austen. By comparing the resulting embeddings produced by SG-2, Fig. 4(a), and SG-5, Fig. 4(b), for our Austen dataset, we find that character embeddings cluster closer to each other than to other non-character word embeddings in each case. Furthermore, on closer inspection of the resulting character clusters themselves, we find that character embeddings within the SG-2 model are more tightly grouped and isolated from non-character embeddings than those in the SG-5 model. We also observe in SG-2 the lead romantic interests from *Sense and Sensibility* occur in close proximity to each other, whereas the romantic interests of the protagonists from across different Austen novels group together in SG-5. We also note that family relations of protagonists from across Austen's novels occur close not only to each other but also nouns (white nodes) for family members such as 'Mother', 'Father', 'Sister', 'Brother', etc.

[5] Any larger and we might lose too much contextual information for each of our relatively small datasets.

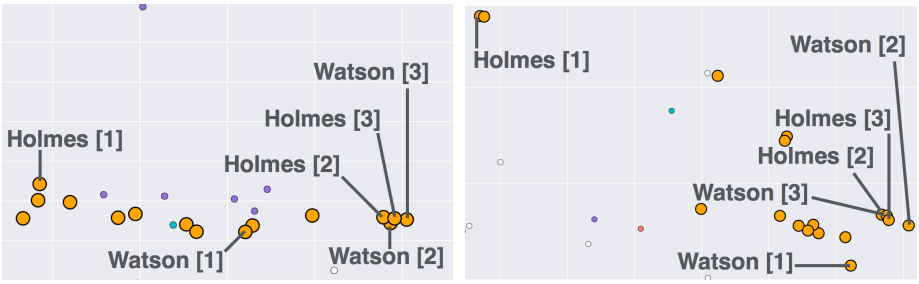


(a) SG, Window=2, $f_{min} = 5$ (b) SG, Window=5, $f_{min} = 5$

Fig. 5: Character embeddings from the Dickens dataset.

Dickens. Unlike Austen, the character embeddings of Dickens for SG-2 and SG-5 are quite different to one another. In SG-2 we observe that the protagonists Esther Summerson from *Bleak House* and Pip from *Great Expectations* appear close together but are distanced from the rest of the characters, including *Oliver*, the protagonist of *Oliver Twist*. Interestingly, what makes the protagonist of *Bleak House* and *Great Expectations* different is that each acts as narrator for part of their story, whereas the protagonist, Oliver, does not participate as a narrator within his text. Our SG-5 results again diverge from both SG-2 and Austen, in that protagonists do not occur in close proximity to each other. Instead, our protagonists occur close to other characters from within their corresponding novels. This may indicate that the protagonists in Dickens's novels are different from one another in some manner, while Austen's protagonists tend to resemble each other. It may be relevant to note that a cohesive cluster of nouns relating to people and family ('mother', 'farther', 'woman', etc) occurs close to the character embeddings of SG-2, but becomes dispersed amongst our character embeddings within SG-5, instead occurring near to characters that might be described by the nouns occurring close to them.

Doyle. At just 147,355 words, Doyle's novels comprise our smallest dataset; this is approximately the same amount of words found in just one novel by either Austen or Dickens. Because of this, we were curious to see how the recurring characters of Sherlock Holmes and John Watson would behave in the word embedding space. What we found for both models is that the characters of Sherlock Holmes and John Watson from the first novel, *A Study in Scarlet*, are embedded away from their counterparts that occur in subsequent novels of the detective series. In the case of SG-2, Holmes is located close to the characterisation of Mary Morstan from the second book in the series, *The Sign of the Four*, where she is first introduced. It is notable that the introduction and characterisation of Sherlock in book one of the series is similar to how Mary Morstan is introduced and characterised in book two. In other words, later manifestations of recurring characters appear to map onto each other as they are further devel-



(a) SG, Window=2, $f_{min} = 5$ (b) SG, Window=5, $f_{min} = 5$

Fig. 6: Character embeddings from the Doyle dataset.

(a) CBOW, Window=2, $f_{min} = 50$ (b) SG, Window=5, $f_{min} = 50$

Fig. 7: Character embeddings from the 'All' dataset. Words embeddings are coloured according to their grammatical class. Adjective: Green, Verb (Past Participle): Blue, Verb (Present): Yellow, Verb (Past): Purple, Character: Large Orange nodes, Verb: Red, Noun: White.

oped developed, perhaps because the author has gained a clearer idea of their characterisation and routine. However, the small size of our Doyle corpus means that we must proceed cautiously in attempting to interpret such phenomena. Adding the fourth text in the series (*The Valley of Fear*) to our corpus might, for example, reveal that the first manifestation of Sherlock maps onto a similar embedding space as Sherlock from the fourth novel in the series. Analysing further works by Doyle is a next proposed step for the project, in order to resolve such questions and also to assist in the task of determining the minimum corpus size required for generating meaningful results.

All. We now turn our attention to the opposite end of the scale to look at our largest textual dataset, consisting of all twelve novels. In this case, rather than comparing the difference between two SG model for $f_{min} = 5$ we explore the most extreme divergence in silhouette scores. This occurs between CBOW-2 ($f_{min} = 50$), which achieves the highest silhouette score out of all models for

this training set. We compare this to the lowest silhouette scoring for $f_{min} = 50$ which is not surprisingly SG-5. What is striking in this case is how the character embeddings go from a being a single, well-defined cluster of characters within the CBOW-2 model, Fig. 7(a), to breaking into two character embedding clusters. One of these consists solely of characters from Austen's novels, while the second is made up of characters from the works of Dickens and Doyle, Fig. 7(b). The small number of Sherlock Holmes texts in our study may affect this result; if our Doyle corpus was larger it might be the case that Doyle would cluster separately from Dickens. As to whether characters are grouping semantically, topically, or as a result of an author's writing particularities, it is hard to say at this stage and would require further textual data for comparison. However, we can conclude that they are not, in the case of SG-5 being grouped syntactically.

5 Conclusion

In this paper, we have generated, visualised, and explored word embedding representations for four different datasets consisting of 12 popular 19th Century novels by the authors Jane Austen, Charles Dickens, and Arthur Conan Doyle. In each case, we have analysed the effect of applying two variants of word2vec, a continuous-bag-of-words strategy and a skip-gram strategy. We first evaluated the differences from a cluster analysis perspective, finding that a context window size of 2 in each case resulted in a tendency for words that are syntactically related to group together. By contrast, context windows of size 5 tended to group characters and words that were more semantically or topically related close to each other.

We also devised a measure for assessing a model's context window sensitivity and found that skip-gram embeddings diverge the most for different context window sizes. We explored each of our novel embeddings in further detail, finding that syntactically, character vectors are very distinguishable from other grammatical categories of words within each *novel2vec* dataset. Our initial 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. In future work, we hope to extend our analysis to diachronic word embeddings [8] to discover how word usage within our corpus changes over time.

Acknowledgments. This research was partly supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, in collaboration with the Nation, Genre and Gender project funded by the Irish Research Council.

References

1. Bansal, M., Gimpel, K., Livescu, K.: Tailoring Continuous Word Representations for Dependency Parsing. Acl pp. 809–815 (2014)
2. Bird, S., Klein, E., Loper, E.: Natural language processing with Python. " O'Reilly Media, Inc." (2009)

3. Bordag, S.: A comparison of co-occurrence and similarity measures as simulations of context. In: Proc. International Conference on Intelligent Text Processing and Computational Linguistics. pp. 52–63. Springer (2008)
4. Elsner, M.: Abstract Representations of Plot Struture. LiLT (Linguistic Issues in Language Technology) 12(5) (2015)
5. Elsner, M.: Character-based kernels for novelistic plot structure. In: European Chapter of the Association for Computational Linguistics. pp. 634–644 (2012)
6. Flekova, L., Gurevych, I.: Personality Profiling of Fictional Characters using Sense-Level Links between Lexical Resources. In: Proc. Conference on Empirical Methods in Natural Language Processing. pp. 1805–1816 (2015)
7. Grayson, S., Wade, K., Meaney, G., Greene, D.: The sense and sensibility of different sliding windows in constructing co-occurrence networks from literature. In: Workshop Proc. 2nd Computational History and Data-Driven Humanities (2016)
8. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. In: Proc. of 54th ACL (2016)
9. Jockers, M.L.: Macroanalysis: Digital methods and literary history. University of Illinois Press (2013)
10. Jockers, M.L., Mimno, D.: Significant themes in 19th-century literature. Poetics 41(6), 750–769 (2013)
11. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proc. 31st ICML. p. 11881196 (2014)
12. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: ACL (2). pp. 302–308 (2014)
13. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9(Nov), 2579–2605 (2008)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Workshop Proc. ICLR (2013)
15. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: HLT-NAACL. vol. 13, pp. 746–751 (2013)
16. Moretti, F.: Network Theory, Plot Analysis. New Left Review 68, 80–102 (2011)
17. O'Callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications (ESWA) (2015)
18. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–43 (2014)
19. Reagan, A.J., Mitchell, L., Kiley, D., Danforth, C.M., Sheridan Dodds, P.: The emotional arcs of stories are dominated by six basic shapes. ArXiv e-prints (2016)
20. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA (May 2010)
21. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53–65 (1987)
22. Sack, G.: Simulating plot: Towards a generative model of narrative structure. In: 2011 AAAI Fall Symposium Series (2011)
23. Xue, B., Fu, C., Shaobin, Z.: A study on sentiment computing and classification of Sina Weibo with word2vec. In: 2014 IEEE International Congress on Big Data. pp. 358–363. IEEE (2014)
24. Yaghoobzadeh, Y., Schütze, H.: Intrinsic subspace evaluation of word embedding representations. In: Proc. 54th ACL (2016)