

NOVEL2GRAPH: Visual Summaries of Narrative Text Enhanced by Machine Learning

Vani K
vanik@idsia.ch

Alessandro Antonucci
alessandro@idsia.ch

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)
Lugano - Switzerland

Abstract

A machine learning approach to the creation of visual summaries for narrative text is presented. Standard natural language processing tools for named entities recognition are used together with a clustering algorithm to detect the characters of the novel and their aliases. The most relevant ones and their relations are evaluated on the basis of a simple statistical analysis. These characters are visually depicted as nodes of an undirected graph whose edges describe relations with other characters. Specialized sentiment analysis techniques based on sentence embedding decide the colours of characters/nodes and their relations/edges. Additional information about the characters (e.g., gender) and their relations (e.g., siblings or partnerships) are returned by binary classifiers and visually depicted in the graph. For those specialized tasks, small amounts of manually annotated data are sufficient to achieve good accuracy. Compared to analogous tools, the machine learning approach we present allows for a richer representation of texts of this kind. A case study to demonstrate this approach for a series of books is also reported.

1 Introduction

The analysis and interpretation of potentially long and complex literary texts with the goal of extracting readable and concise information is a challenging task even for trained human experts. Developing AI systems able to support (or even replace) humans in such activities is therefore an exciting challenge for research in natural language processing. In fact the analysis of literary texts involves various complex steps such as the identification of the main characters and relations, and the corresponding *typification* (e.g., gender, partnerships, goodness). Moreover the high variance in the style and the lexicon with frequent use of neologisms [MC⁺14] and figures of speech [Nyg06] further complicates the scenario. Manually created summaries based on graphs such as the one in Figure 1 can be therefore regarded as useful tools to better understand complex plots.

In the very recent years, *natural language processing* (NLP) is taking advantage of machine learning techniques based on neural networks with a deep structure and trained by huge amounts of text data [MSC⁺13]. Word and sentence *embedding* techniques are now the state-of-the-art representation models in NLP [LGD15, SJV19, WBGL15]. The main goal of this paper is to show how these modern machine learning tools as well as more traditional techniques (to be used if only small amounts of training data are available or needed) can significantly

Copyright © 2019 for the individual papers by the paper's authors. Copying permitted for private and academic purposes. This volume is published and copyrighted by its editors.

In: A. Jorge, R. Campos, A. Jatowt, S. Bhatia (eds.): Proceedings of the Text2StoryIR'19 Workshop, Cologne, Germany, 14-April-2019, published at <http://ceur-ws.org>

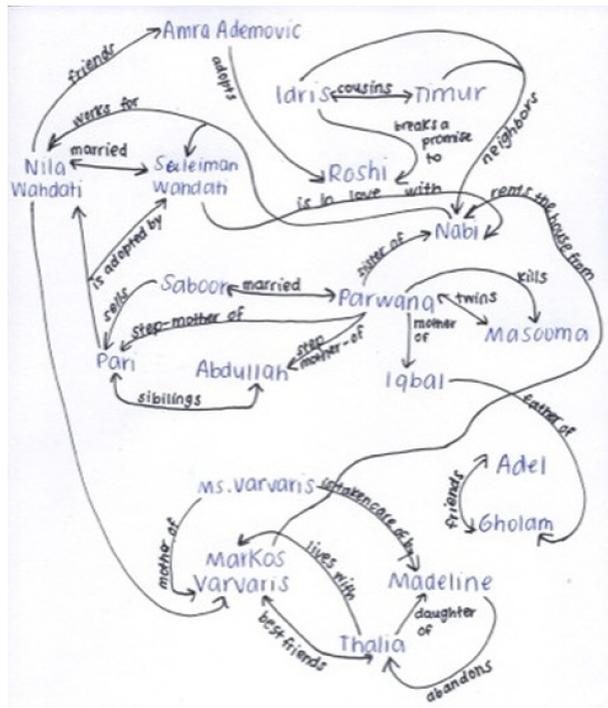


Figure 1: Manually created visual outline of *And the Mountains Echoed* by K. Hosseini

improve the quality of the information extracted from a narrative text, thus enhancing the readability of the corresponding summaries. In particular we focus on the creation of visual outlines based on graphs, analogous to what readers and specialists are often doing to understand the plot of complex novels as the one in Figure 1.

Our approach to literary text analysis encompasses various NLP techniques such as named entity recognition, syntactic parsing, semantic text analysis and NLP application tasks such as gender recognition, sentiment analysis and social network analysis. Most of these tasks have been deeply studied in the literature and standard tools are available for them. Yet, some tuning and adaptation to the specific case of narrative text is required, mainly because of the inherent nature of text content. For instance, sentiment analysis is generally intended to process short texts such as product reviews or social networks comments. Using those tools for narrative texts can produce inaccurate evaluations and (Section 4), as collecting dedicated training data might be problematic, tools for semantic learning can instead give good results [HM17, KIO⁺16, MPGC17, NS07, NS19]. Besides sentiment analysis, extensive textual pre-processing is needed for literary texts because of the difficulty in terms of interpretability, as experienced even by humans.

Overall, we present a combination of *natural language processing* (NLP) and *machine learning* (ML), with a particular focus on current trends of *deep learning* (DL), adapted to the analysis and visualization of textual narrative data. DL approaches are used to improve *natural language understanding* (NLU) of text together with specialized syntactic and semantic analyses to recognize the characters and their relations. Furthermore, for the extraction of the character and relationship sentiments, we use the classification and prediction approaches based on the *universal sentence embedding* (USE, [CYK⁺18]). USE helps in transfer learning using the transformer architectures [VSP⁺17] or deep averaging networks [IMBGDI15]. Moreover, family relations between the characters are reduced to binary classification tasks. The proposed method shows the strength of these techniques to achieve a good understanding of this particular kind of texts.

2 Existing Work

The automatic interpretation and visual analysis of literary texts has been explored in the past few years from various perspectives. Most of the studies in this area focus on different methodologies to detect the characters of a novel and identifying their relationships. One of the works focused on character detection proposes a new approach to named entity recognition based on co-reference resolutions and based on a set of heuristic

rules integrating the *Stanford CoreNLP*.¹[VJPR15]. A similar tool² for efficient character annotations and alias resolutions was proposed by [VDJ⁺16]. Sentiment analysis for the characters and their relations is another focus area. Besides the Stanford CoreNLP, the approach in [FPU15] uses also the *SentiWordNet* ontology³ to build character networks. In [PAHS⁺17], the literary characters and the network associations have been also studied, where the features that defines character qualities are identified, while in [NB13] sentiment relations between characters in Shakespeare’s plays have been processed using AFINN sentiment lexicon⁴. Some of the reported studies also presented a dynamical analysis of the novels, this taking into account that the character traits and relations varies over the pages. In [Els12], a kernel method to capture novelistic plots over time has been considered. In [KC12] a map of the emotions over time to indicate the information flow over time was used to predict the emotions. Gender based stereotypes in novels are studied in [Cri16] analysing the masculine and feminine adjectives and the gender-biased portrayal of characters. The present work should be intended as one of the first attempts to apply the recent advances of DL in NLP in the particular area of the creation of visual summaries for narrative text.s

3 Methodology

The algorithms we adopt to generate visual summaries of narrative text have been shaped as a single Python software tool called NOVEL2GRAPH. The associated workflow is depicted in Figure 2. The three sections here below describe the main NOVEL2GRAPH modules, while the very last step, i.e., the actual creation of the graph is based on the Graphviz free software.⁵

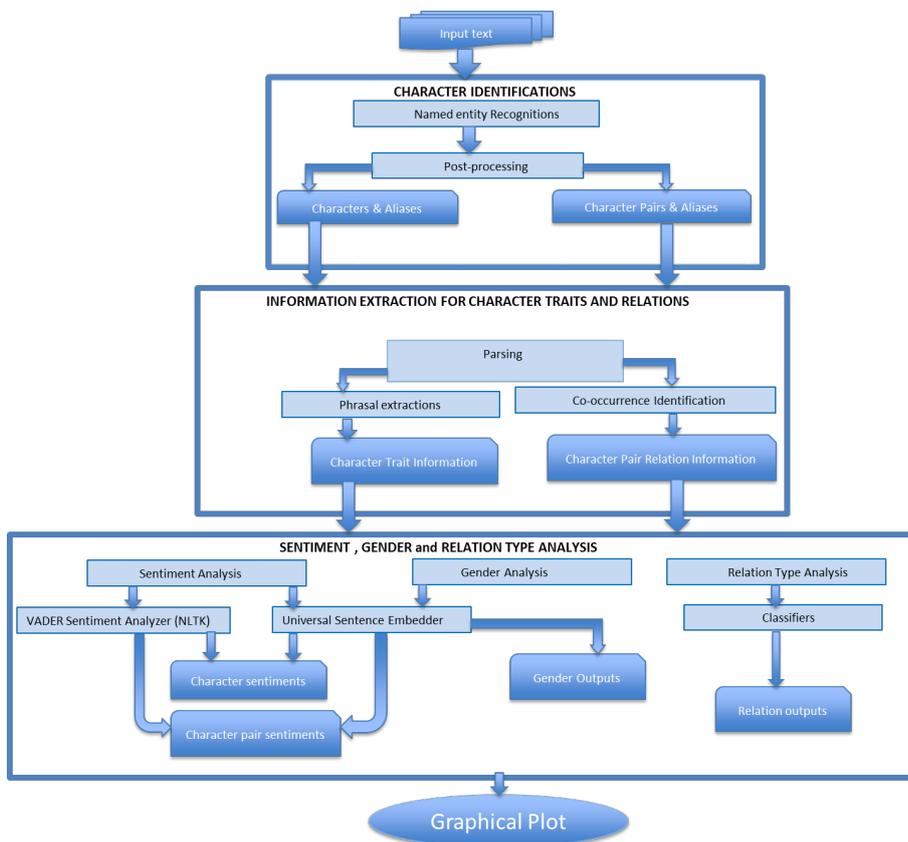


Figure 2: The NOVEL2GRAPH workflow

¹<https://stanfordnlp.github.io/CoreNLP>
²<https://github.com/hardik-vala/charles>
³<https://sentiwordnet.isti.cnr.it>
⁴<https://github.com/fnielsen/afinn>
⁵<https://www.graphviz.org>

3.1 Character Identification and Alias Clustering

The first module of the flow handles identification of the characters in novel. This also involves the recognition of *aliases*, as the author might refer to the same character in different ways. The Stanford *named entity recognizer* (NER)⁶ is initially used to detect all the *person* entities. Afterwards, part-of-speech tagging is applied to filter out the false positives returned by the NER. The basic pre-processing such as tokenisation and stopwords removal are also applied. The so-obtained list of persons is finally processed by the DBSCAN clustering algorithm [BK07] to group together the aliases of a same character. As a pairwise distance between the person names we use the Ratcliff-Obershelp algorithm [RM88]. A dictionary with the *representative* character names (corresponding to the longest alias of each group) as keys and the aliases as values is eventually created.

3.2 Character and Relation Traits Identification

This module focuses on the extraction of character *traits*, which are supposed to describe the role of the character or the quality of a relation between two characters. For such a typification, we extract the phrasal information from the text that contains the character (or an alias) name. A syntactic parse tree is therefore constructed by means of the *pattern parse tool*.⁷ The number of phrases related to a character decides the size of the corresponding node in the graphical visualization. As the phrases can be too short to include a co-occurrence, for character pair relations, we instead extract from the parse tree the sentences in which the character pair co-occurs. The number of sentences related to a character pair defines the size of the edge connecting the nodes of the two characters in the graph. This information is also used as the input for sentiment, gender and relation type analysis.

3.3 Sentiment, Gender and Relationship Type Analysis

This module focuses on the identification of (positive or negative) sentiments describing the character traits and the relationships with other characters. Phrasal and sentence information extracted from the previous module are processed by two different methods: (a) the NLTK-VADER sentiment analyser;⁸ (b) the Universal Sentence Encoder (USE).⁹ The goal is to compare the two approaches in terms of accuracy in the classification of the possibly intricate sentiments occurring in a narrative text. NLTK-VADER uses a rule-based lexicon method built up with words annotated in terms of positive or negative sentiments. The phrasal and sentence information is directly processed by the sentiment analyser and the corresponding *sentiscores*, giving positive, negative or neutral weights to the sentiments, are recorded. The tool eventually computes the *compound score* (i.e., the average of sentiscores) which defines the sentiment. The second approach, based on DL, requires a supervised training, and allows to formulate the problem as a classification task. USE helps to capture the sentence meanings and their relations representing them as embeddings at the sentence level. For the training, we use lists of adjectives collected online.¹⁰ Adjectives seem to be the most informative tags to reasonably describe a character or a relationship. Similar techniques based on USE are also used for gender identification, for which dedicated and sufficiently large training sets are already available. In both cases, we used the Keras model with 256 dense layers and a prediction layer with softmax activation, Adam optimisation, and categorical loss entropy. The model is trained with 10 epochs on 32 batches. Another focus is to extract the relation types, which can describe the family links or other relations such as siblings, friends, enemies, killers. For this task we created manually annotated data of small size, which are instead processed with classical ML algorithms and currently we focus only on the sibling relations. s

4 Empirical Analysis on a Case Study

For a preliminary validation of our methodology, we fed NOVEL2GRAPH with the six books of the *Harry Potter* series by J.K. Rowling. Due to the lack of space, in Figures 4 and 5 we depict the graphs of two books, with two thresholds: only the most important characters of each book are depicted; only relations expressed more than a given threshold are displayed. The size of nodes and the thickness of edges increase linearly with the number of occurrences in the text. Red denotes negative sentiments and blue the positive ones. Relatives recognized

⁶<https://nlp.stanford.edu/software/CRF-NER.shtml>

⁷<https://www.clips.uantwerpen.be/pages/pattern-en>

⁸https://www.nltk.org/_modules/nltk/sentiment/vader.html

⁹<https://tfhub.dev/google/universal-sentence-encoder/2>

¹⁰https://lesn.appstate.edu/fryeem/re4030/character_trait_descriptive_adje.htm

by relation type classifiers are connected by dashed edges. A small focus group of persons who carefully read the books was asked to decide the quality of the detected relations and sentiments and the comments have been unanimously positive.

For a quantitative evaluation, we focus on the sentiment analysis. The focus group was asked to annotate the sentiments of the characters and the relations as positive or negative. While NLTK-VADER was originally trained by a large data set, we train USE with a data set of only 4000 adjective conveying positive or negative sentiments. Sentiment analysis accuracy of NLTK-VADER and USE are compared in Figure 3. The specialized USE, outperforms the general purpose NLTK-VADER, which confirms the strength of transfer learning when used in specific domains such as narrative text. The combination of NLP and DL approaches seems therefore to ease the text understanding. USE gives similar accuracies (71% on average over all the six books) for gender analysis. As a first test for the relationships classification, a simple *naive Bayes* classifier with *bag-of-words* features and trained with a small data set of 100 manually annotated sentences was able to detect the notion of *sibling* with full accuracy.

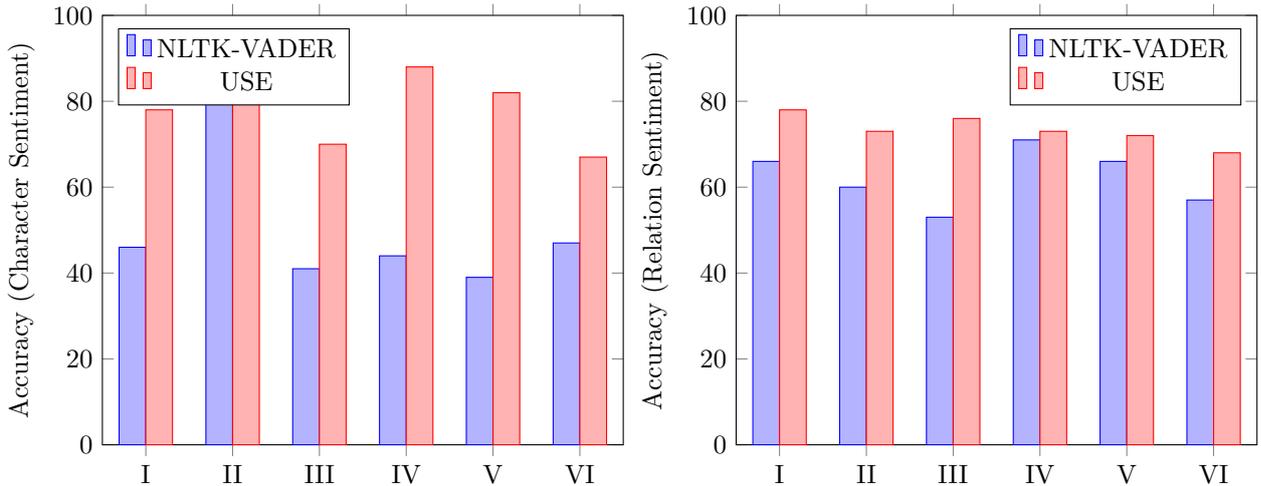


Figure 3: Comparison of NLTK-VADER and USE approaches to sentiment analysis

5 Conclusions and Future Work

The present work attempted to explore and analyse the synergy of natural language processing and deep learning for the analysis and summarization of literary texts. We see many directions to improve the techniques proposed here. First, we need to test the tool on an extensive benchmark of novels. The dynamic evolution of those graphs, either incrementally or by using a fixed-length window over the pages, could be expressed by an animation. To properly achieve that, the problem of a proper temporal positioning of the detected events, which are currently described only by their position in the text, should be tackled. Regarding the character identification module, deeper analysis based on co-reference resolutions and contextual analysis should be considered. Regarding the analysis of sentiments, gender and relationships, improvements can be achieved by recognizing the roles of the characters within the context. While semantic role labelling could be used for that by labelling triplets (subject, predicate, object) in a better way, in narrative texts the structure of the sentence may not follow a specific syntactic rule. In many cases, the literary texts contain many conversations enriched with semantic usages. To tackle the problem more effectively, it is important that the current cutting edge deep learning approaches are used in conjunction with the NLP techniques to achieve better text understanding.

Acknowledgements

We acknowledge Arpitha Prasad Bharathi, who helped us in evaluating our results with the Harry Potter series throughout the work, and Markus Zohner, who first suggested us the idea of developing visual outlines of novels by means of AI techniques.

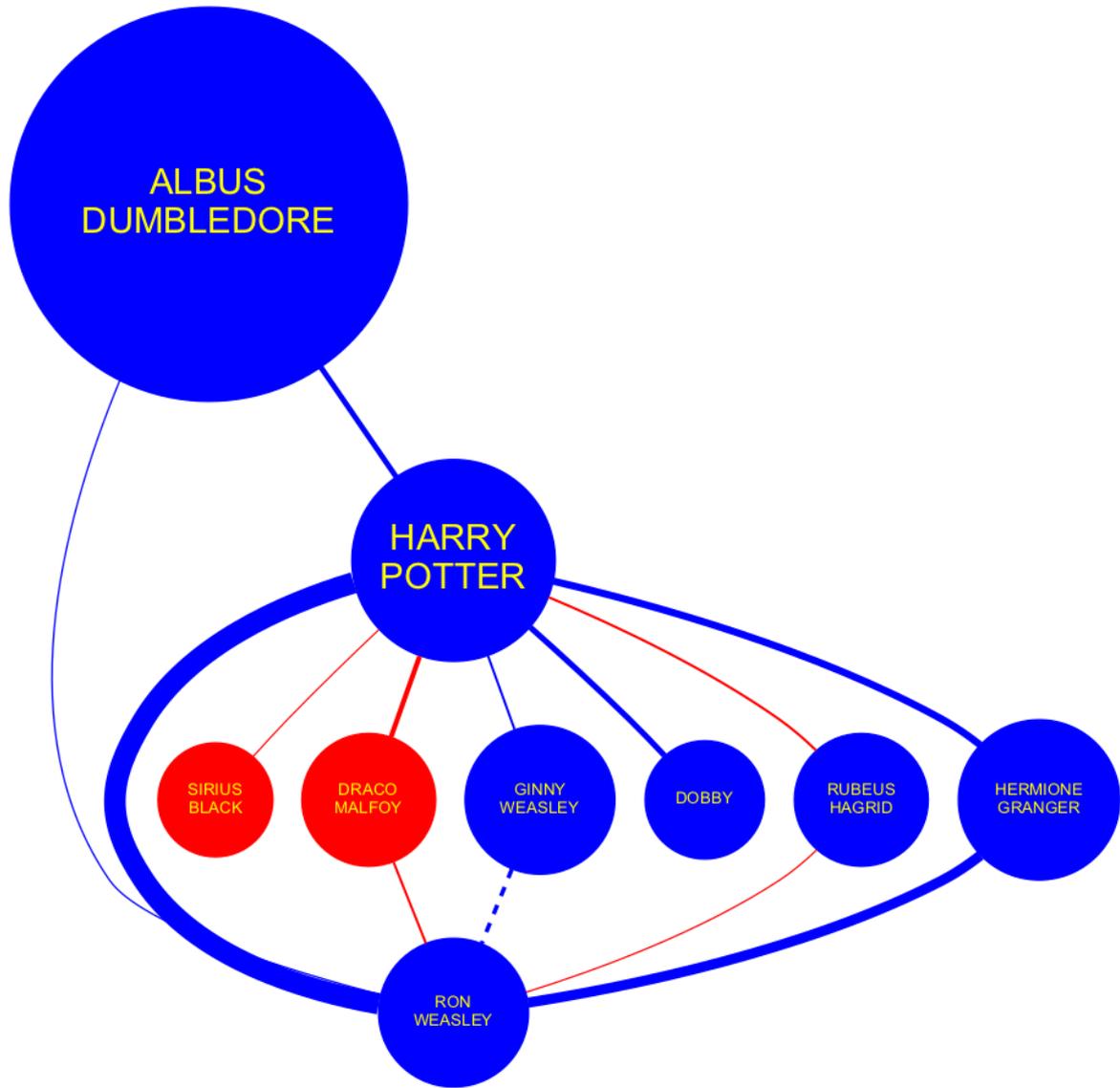


Figure 4: NOVEL2GRAPH output of the second book of the Harry Potter series

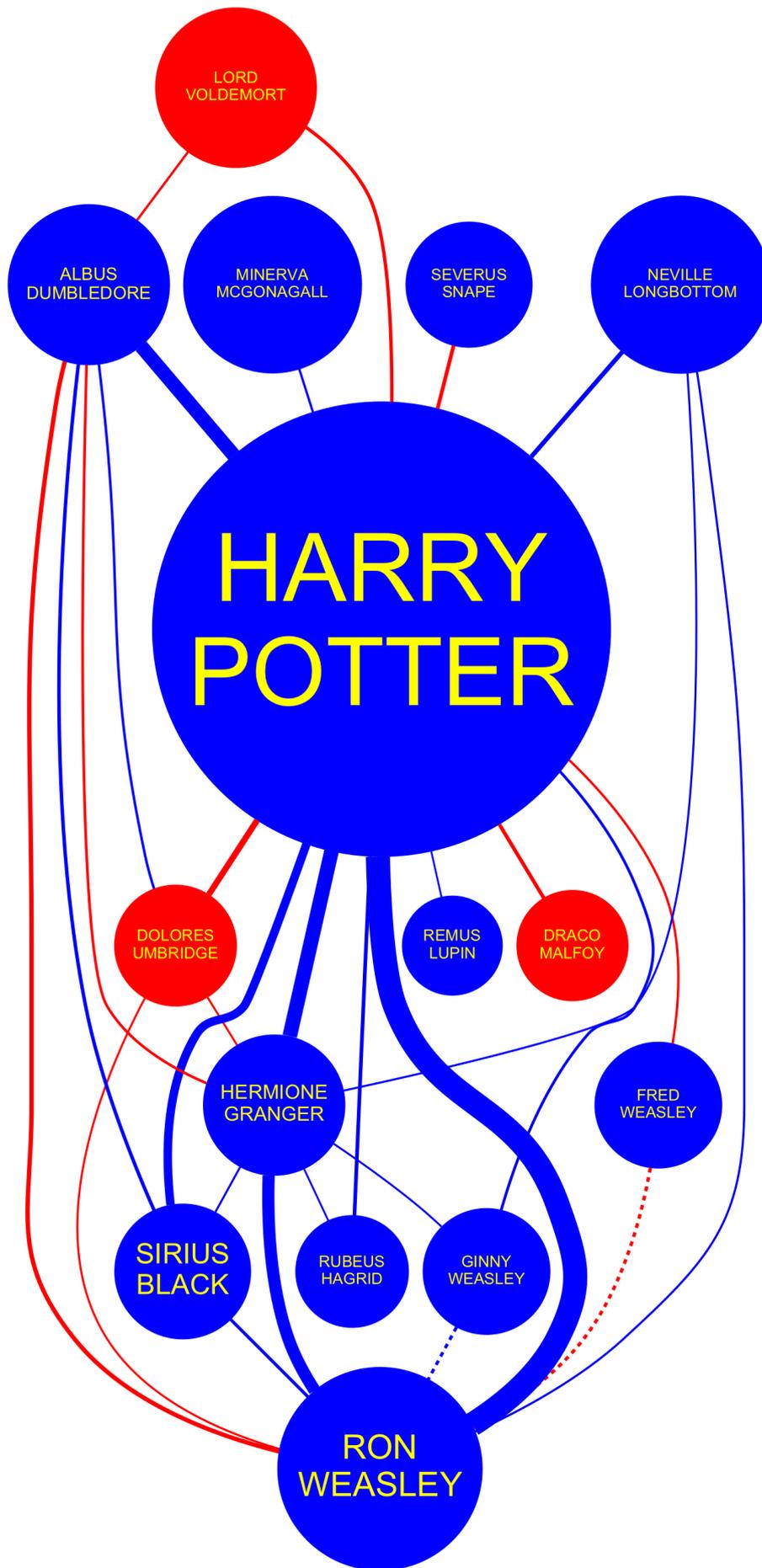


Figure 5: NOVEL2GRAPH output of the fifth book of the Harry Potter series

References

- [BK07] Derya Birant and Alp Kut. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.
- [Cri16] Rebecca Cripps. The persistence of gender-based stereotypes in the language of Harry Potter and the Philosopher’s Stone and Harry Potter and the Goblet of Fire, 2016.
- [CYK⁺18] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.
- [Els12] Micha Elsner. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644. Association for Computational Linguistics, 2012.
- [FPU15] Matt Fernandez, Michael Peterson, and Ben Ulmer. Extracting social network from literature to predict antagonist and protagonist. 2015.
- [HM17] Abdalraouf Hassan and Ausif Mahmood. Deep learning approach for sentiment analysis of short texts. In *Control, Automation and Robotics (ICCAR), 2017 3rd International Conference on*, pages 705–710. IEEE, 2017.
- [IMBGDI15] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1681–1691, 2015.
- [KC12] Tabitha Kirkland and William A Cunningham. Mapping emotions through time: How affective trajectories inform the language of emotion. *Emotion*, 12(2):268, 2012.
- [KIO⁺16] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387, 2016.
- [LGD15] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.
- [MC⁺14] Marina Martínez Carbajal et al. Neologisms in Harry Potter books. 2014.
- [MPGC17] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79, 2017.
- [MSC⁺13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. 2013.
- [NB13] Eric T Nalisnick and Henry S Baird. Character-to-character sentiment analysis in Shakespeare’s plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 479–483, 2013.
- [NS07] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.
- [NS19] Mohammad K. Nammous and Khalid Saeed. Natural language processing: Speaker, language, and gender identification with LSTM. In *Advanced Computing and Systems for Security*, pages 143–156. 2019.

- [Nyg06] Åsa Nygren. Essay on the linguistic features in J.K. Rowling’s Harry Potter and the Philosopher’s Stone, 2006.
- [PAHS⁺17] Andrew Piper, Mark Algee-Hewitt, Koustuv Sinha, Derek Ruths, and Hardik Vala. Studying literary characters and character networks. In *DH*, 2017.
- [RM88] John W. Ratcliff and David E. Metzener. Pattern matching: The gestalt approach. *Dr Dobbs Journal*, 13(7):46, 1988.
- [SJV19] Roger Alan Stein, Patricia A Jaques, and João Francisco Valiati. An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471:216–232, 2019.
- [VDJ⁺16] Hardik Vala, Stefan Dimitrov, David Jurgens, Andrew Piper, and Derek Ruths. Annotating characters in literary corpora: A scheme, the CHARLES tool, and an annotated novel. In *LREC*, 2016.
- [VJPR15] Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. Mr. Bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, 2015.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [WBGL15] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.