ARElight: Context Sampling of Large Texts for Deep Learning Relation Extraction

Nicolay Rusnachenko[1 [0000−0002−9750−5499]], Huizhi Liang[1 [0000−0003−4408−4528]], Maksim Kalameyets[1 [0000−0002−7873−2733]], Lei Shi[1 [0000−0001−7119−3207]]

1. School of Computing, Newcastle University, Newcastle upon Tyne,
{nicolay.rusnachenko,huizhi.liang,maksim.kalameyets,lei.shi}@newcastle.ac.uk

Abstract. The escalating volume of textual data necessitates adept and scalable Information Extraction (IE) systems in the field of Natural Language Processing (NLP) to analyse massive text collections in a detailed manner. While most deep learning systems are designed to handle textual information as it is, the gap in the existence of the interface between a document and the annotation of its parts is still poorly covered. Concurrently, one of the major limitations of most deep-learning models is a constrained input size caused by architectural and computational specifics. To address this, we introduce ARElight[1], a system designed to efficiently manage and extract information from sequences of large documents by dividing them into segments with mentioned object pairs. Through a pipeline comprising modules for text sampling, inference, optional graph operations, and visualisation, the proposed system transforms large volumes of text in a structured manner. Practical applications of ARElight are demonstrated across diverse use cases, including literature processing and social network analysis.

Keywords: Data Processing Pipeline · Information Retrieval · Visualisation

1 Introduction

Information Extraction (IE) in the domain of Natural Language Processing (NLP) involves a separate studies aimed on objects annotation (entities, events, etc.) in texts [3, 15] and establishing connections between objects [4, 8]. IE finds a significant application in text structurization, the knowledge base formation [2, 6]. One of the generalized concept for structuring raw texts is to form *pipeline* of sequential text transformations, with relation extraction[2] as a module [7, 26]. Another alternative to the vast number of solution follows the concept of *target-oriented systems*, aimed at applying specific machine learning architectures for the given raw input [13, 18, 25]. However, once texts become larger or their actual amount is massive, the direct application of these systems to the entire text becomes: (i) less informative for result analysis [20], and (ii) less effective.

[1] https://github.com/nicolay-r/ARElight
[2] http://deepdive.stanford.edu/relation_extraction

Partitioning large texts [11] that conveyed relations between mentioned objects represents a common solution for managing long-input problems in downstream systems [16, 26]. In this paper, we propose ARElight that follows bridges the gap in processing of large documents. Our system contributes by offering a scalable relation annotations, surpassing the similar slot-filling systems for processing large collections instead of single documents [11]. We demonstrate this system's ability to analyse sentiment relations in literature novel books, social networks.

2 The ARElight System Design

ARElight system represent a pipeline of further modules: (1) text sampler, (2) inference, (3) graph operations (optional), and (4) graph visualisation. Since the source of input information represent raw documents, *text sampler* module represent a core of the system. Figure 1 shows the pipeline architecture along with a detailed illustration of its core module.

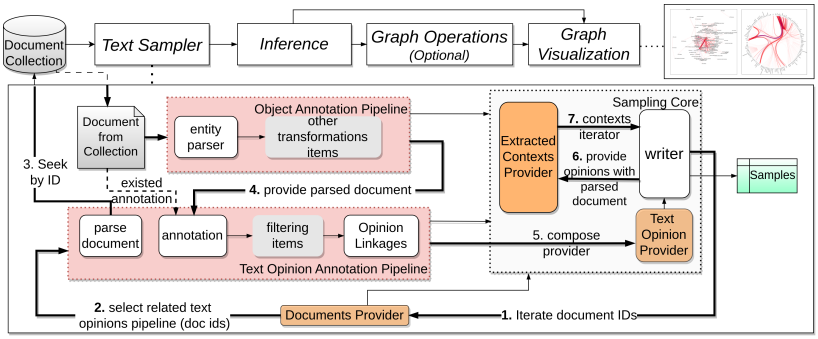

Fig. 1. ARElight-0.24.0 system design; *top*: the main application pipeline; *bottom*: architecture of the *text sampler* module with two separated pipelines for document content annotation (red blocks), data providers (yellow blocks); the process on *document collection sampling*, which is depicted in bold arrows, numbered from 1 to 7

Text Sampler. This module performs streaming extraction of context with mentioned object pairs in it from large amount of texts in the document collection. Unlike existed systems, the core module provide two separate *declarative pipelines*[3] that describe annotation for (i) objects[4] and (ii) pairs formation between objects[5]. To automatically extract mentioned objects in text, our annotation pipeline provides support for models from DeepPavlov [12]. The system supports pairs forming based on: (i) document level (object value), and (ii) context level (object indices) annotations. We consider *strategy pattern* [1] for the implementation of provider components (blocks in yellow color, Figure 1).

[3] `https://github.com/nicolay-r/AREkit/wiki/Task-Schemata`
[4] `https://github.com/nicolay-r/AREkit/wiki/Pipelines:-Text-Processing`
[5] `github.com/nicolay-r/AREkit/wiki/Pipelines:-Text-Opinion-Annotation`

Inference. This module performs samples classification, followed by automatic graph serialization necessary for visualization. For samples classification, we propose supporting language models [10]. We use OpenNRE [18] as target-oriented solution for contextual relation extraction. Our system supports BERT-based text classification models [17], pre-trained with OpenNRE and distributed as PyTorch checkpoints [9]. We present classified samples as undirected graphs $G = (V, E, W_e, W_v)$ where V and E represent vertices (found objects) and edges (found pairs), and W_e, W_v denote their respective weights (frequencies in text).

Graph operations. This module allows to serialize new graph from a pair of existing graphs. Our system supports three crucial operations:

Union ($G_1 \cup G_2$). The result graph contains all the vertices and edges that are in G_1 and G_2. The edge weight is given by $W_e = W_{e1} + W_{e2}$, and the vertex weight is its weighted degree centrality: $W_v = \sum_{e \in E_v} W_e(e)$.

Intersection ($G_1 \cap G_2$). The result graph contains only the vertices and edges common to G_1 and G_2. The edge weight is given by $W_e = \min(W_{e1}, W_{e2})$, and the vertex weight is its weighted degree centrality: $W_v = \sum_{e \in E_v} W_e(e)$.

Difference ($G_1 - G_2$). The result graph contains all the vertices from G_1 but only includes edges from E_1 that either don't appear in E_2 or have larger weights in G_1 compared to G_2. The edge weight is given by $W_e = W_{e1} - W_{e2}$ if $e \in E_1$, $e \in E_1 \cap E_2$ and $W_{e1}(e) > W_{e2}(e)$.

Visualisation. This module composes HTML page with visual user interface (UI) (Figure 2) and launches a web server to host it. The UI consists of (a) a dataset selector, (b) visualisation options, (c) visualisation model selector, and (d) two D3JS visualisation modes – force [28] and radial [27] graph templates.

3 Experiments and Demonstration

As a demo, we propose the analysis of texts' narratives across 3 distinct use cases: (Case 1) novel "War and Peace"$_{\text{Vol.1-3}}$ by Leo Tolstoy, (Case 2) pro-Russian/Ukrainian war comments on VK social network [24], and (Case 3) X/Twitter accounts. In particular, we extract sentiment relations (pos/neg) between objects in texts [14].

Collections preparation. We executed *text sampling+inference* scenario[6] for texts in Russian (Case 1-2), and in English (Case 3). For objects annotation[7], we consider BERT$_{\text{mult-OntoNotes5}}$ [12, 5]. For samples classification we adopt RuBERT [19] model, use fine-tuned on RuSentRel [14] and RuAttitudes [21] collections with NLI-prompt [22, 23] (Case 1-2). We automatically translate these collections in English to fine-tune BERT$_{\text{cased}}$ [17] (Case 3). These models are publicly available and automatically fetched upon scenario launch.

Graph operations. We present an example of graph analysis through various operations. For the Case 2, the aggregated pro-Russian/Ukrainian graph was obtained by employing the *Union* operation on narrative graphs of individual users, extracted from the dataset [24]. We utilized the *Intersection* operation to discern commonalities between the narratives of pro-Russian and pro-Ukrainian

[6] `https://github.com/nicolay-r/ARElight/wiki/Language-Specific-Application`

[7] We keep {`ORG,PERSON,LOC,GPE`} types and mask their values in text classification

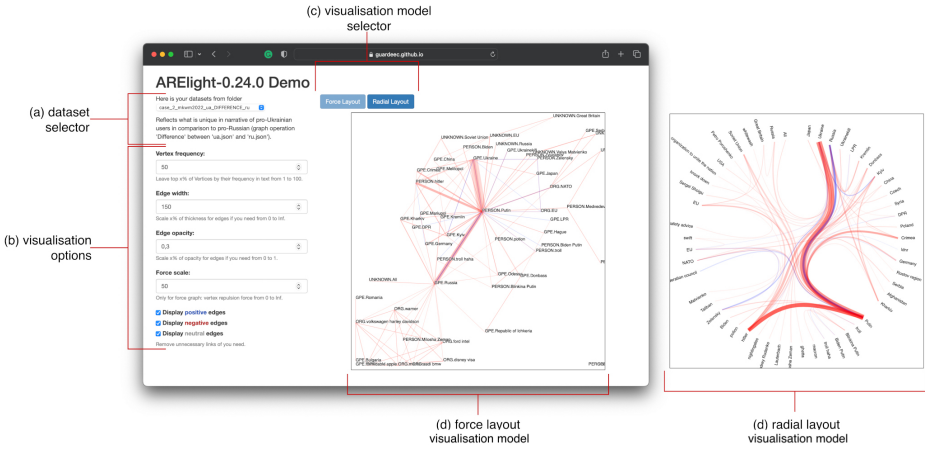

Fig. 2. The visual interface of the ARElight-0.24.0 web server

users (Case 2), as well as between Rishi Sunak and Boris Johnson on Platform X/Twitter (Case 3). The *Difference* operation facilitated understanding the unique aspects of one narrative compared to another. We applied it to extract differences between pro-Ukrainian and pro-Russian users' narratives and vice versa (Case 2), and to elucidate disparities between Rishi Sunak and Boris Johnson, and vice versa, on Platform X/Twitter (Case 3).

The visual interface is presented in Figure 2. For example, one can see that pro-Ukrainian users more often mention Putin with Hitler than pro-Russians (Case 2). As for the Case 1, "War and Peace"$_{\text{Vol.2}}$ is distinctly centers on life and relations between individuals, in contrast to more war-centric themes in other volumes. In Case 3, related to the differences between B. Jonson and R. Sunak, one can observe a higher number of positive UK-France pairs and significantly more UK-Ukraine pairs. You can explore all the three cases by following the demo project link: `https://guardeec.github.io/arelight_demo/template.html`.

4 Conclusion and Future Work

In this paper, we introduced the ARElight system, designed to facilitate the segmentation and analysis of large documents by converting them into smaller text parts associated with mentioned object pairs, and subsequently analyzing them as graphs. The system filters and samples text segments involving such pairs throughout a sequence of large documents, constructs graphs from them, and conducts graph operations and visualisations. The aim of the proposed system was to bridge the existing gap in the programming interface between a document and its subsequent annotation. The reusability of the system components has been demonstrated through use cases and showcases its applicability to various scenarios, ranging from the analysis of books to social media accounts.

Acknowledgments
This research is partially supported by UK Research and Innovation, United Kingdom through the Strategic Priority Fund as part of the Protecting Citizens Online programme. Grant: "AGENCY: Assuring Citizen Agency in a World with Complex Online Harms", EP/W032481/2 at Newcastle University.

References

[1] Erich Gamma et al. "Elements of Reusable Object-Oriented Software". In: *Design Patterns* (1995).

[2] Tetsuya Nasukawa and Jeonghee Yi. "Sentiment Analysis: Capturing Favorability Using Natural Language Processing". In: *Proceedings of the 2nd International Conference on Knowledge Capture*. K-CAP '03. Sanibel Island, FL, USA: Association for Computing Machinery, 2003, pp. 70–77. ISBN: 1581135831. DOI: 10.1145/945645.945658. URL: https://doi.org/10.1145/945645.945658.

[3] David Nadeau and Satoshi Sekine. "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[4] Iris Hendrickx et al. "SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 33–38. URL: https://aclanthology.org/S10-1006.

[5] Weischedel Ralph et al. *OntoNotes Release 5.0*. 2012. DOI: https://doi.org/10.35111/xmhb-2b84. URL: https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf.

[6] Marilyn Walker et al. "Stance classification using dialogic properties of persuasion". In: *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 2012, pp. 592–596.

[7] Christopher Manning et al. "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 55–60. DOI: 10.3115/v1/P14-5010. URL: https://aclanthology.org/P14-5010.

[8] Eunsol Choi et al. "Document-level Sentiment Inference with Social, Faction, and Discourse Context". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 333–343. DOI: 10.18653/v1/P16-1032. URL: https://aclanthology.org/P16-1032.

[9] Adam Paszke et al. "Automatic differentiation in PyTorch". In: 2017.

[10] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[11] Heike Adel et al. "DERE: A task and domain-independent slot filling framework for declarative relation extraction". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2018, pp. 42–47.

[12] Mikhail Burtsev et al. "DeepPavlov: Open-Source Library for Dialogue Systems". In: *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 122–127. DOI: 10.18653/v1/P18-4021. URL: https://aclanthology.org/P18-4021.

[13] Xu Han, Zhiyuan Liu, and Maosong Sun. "Neural Knowledge Acquisition via Mutual Attention between Knowledge Graph and Text". In: *Proceedings of AAAI*. 2018.

[14] Natalia Loukachevitch and Nicolay Rusnachenko. "Extracting sentiment attitudes from analytical texts". In: *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialogue-2018 (arXiv:1808.08932)* (2018), pp. 459–468.

[15] Vikas Yadav and Steven Bethard. "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 2145–2158. URL: https://aclanthology.org/C18-1182.

[16] Saizheng Zhang et al. "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2204–2213. DOI: 10.18653/v1/P18-1205. URL: https://aclanthology.org/P18-1205.

[17] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[18] Xu Han et al. "OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Ed. by Sebastian Padó and Ruihong Huang. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 169–174.

DOI: 10.18653/v1/D19-3029. URL: https://aclanthology.org/D19-3029.

[19] Yuri Kuratov and Mikhail Arkhipov. "Adaptation of deep bidirectional multilingual transformers for russian language". In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*. 2019.

[20] Nicolay Rusnachenko and Natalia Loukachevitch. "Neural Network Approach for Extracting Aggregated Opinions from Analytical Articles". In: *Data Analytics and Management in Data Intensive Domains*. Ed. by Yannis Manolopoulos and Sergey Stupnikov. Cham: Springer International Publishing, 2019, pp. 167–179. ISBN: 978-3-030-23584-0.

[21] Nicolay Rusnachenko, Natalia Loukachevitch, and Elena Tutubalina. "Distant Supervision for Sentiment Attitude Extraction". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (2019).

[22] Chi Sun, Luyao Huang, and Xipeng Qiu. "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 380–385. DOI: 10.18653/v1/N19-1035. URL: https://aclanthology.org/N19-1035.

[23] Nicolay Rusnachenko. "Language Models Application in Sentiment Attitude Extraction Task". Russian. In: *Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS), vol.33*. 3. 2021, pp. 199–222.

[24] Maxim Kolomeets. *Dataset with Russian-Ukrainian war related comments from top Russian media based on their VKontakte web pages*. 2022. URL: https://github.com/guardeec/datasets#mkwm2022.

[25] Gaku Morio et al. "Hitachi at SemEval-2022 Task 10: Comparing Graph- and Seq2Seq-based Models Highlights Difficulty in Structured Sentiment Analysis". In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1349–1359. DOI: 10.18653/v1/2022.semeval-1.188. URL: https://aclanthology.org/2022.semeval-1.188.

[26] Adam Roberts et al. "Scaling Up Models and Data with t5x and seqio". In: *arXiv preprint arXiv:2203.17189* (2022). URL: https://arxiv.org/abs/2203.17189.

[27] Mike Bostock. *D3js gallery: Hierarchical edge bundling*. 2023. URL: https://observablehq.com/@d3/hierarchical-edge-bundling.

[28] Observable. *D3js gallery: Force-directed graph*. 2023. URL: https://observablehq.com/@d3/force-directed-graph/2.