Multi-document Cohesion Network Analysis: Visualizing Intratextual and Intertextual Links

Maria-Dorinela Dascalu[1], Stefan Ruseti[1], Mihai Dascalu[1,2(✉)], Danielle S. McNamara[3], and Stefan Trausan-Matu[1,2]

[1] University Politehnica of Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania
{dorinela.dascalu, stefan.ruseti, mihai.dascalu, stefan.trausan}@upb.ro
[2] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044 Bucharest, Romania
[3] Department of Psychology, Arizona State University, PO Box 871104, Tempe, AZ 85287, USA
dsmcnama@asu.edu

Abstract. Reading comprehension requires readers to connect ideas within and across texts to produce a coherent mental representation. One important factor in that complex process regards the cohesion of the document(s). Here, we tackle the challenge of providing researchers and practitioners with a tool to visualize text cohesion both within (intra) and between (inter) texts. This tool, Multi-document Cohesion Network Analysis (MD-CNA), expands the structure of a CNA graph with lexical overlap links of multiple types, together with coreference links to highlight dependencies between text fragments of different granularities. We introduce two visualizations of the CNA graph that support the visual exploration of intratextual and intertextual links. First, a *hierarchical view* displays a tree-structure of discourse as a visual illustration of CNA links within a document. Second, a *grid view* available at paragraph or sentence levels displays links both within and between documents, thus ensuring ease of visualization for links spanning across multiple documents. Two use cases are provided to evaluate key functionalities and insights for each type of visualization.

Keywords: Cohesion Network Analysis · Semantic links · Lexical overlap links · Coreference links · Graph visualizations

1 Introduction

Comprehension is a difficult and challenging process, for which learners need to understand words and sentences, connect ideas and link them to prior knowledge, while creating a coherent mental representation of the read text. One important factor in the comprehension process regards the cohesion of text [1], which considers the degree to which there are semantic links between ideas within a text. Cohesion is higher when there are multiple ideas and words that overlap and when the connections between ideas are explicit. Low cohesion text is more challenging to understand, particularly for

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 80–85, 2020.
https://doi.org/10.1007/978-3-030-52240-7_15

low knowledge and less skilled readers [2]. The process of overcoming cohesion gaps is even more challenging when learners are faced with multiple documents that require establishing connections both within and between disparate text fragments. Making connections across multiple texts is considerably more difficult than doing so within a single text. Some text fragments may be semantically linked, while other may be isolated, distal, and thus more difficult to recognize or infer.

While text cohesion is recognized as an important factor in comprehension and learning from text, there is currently no technique or tool available to visualize cohesive links between documents. We address this gap here, introducing Multi-document Cohesion Network Analysis (MD-CNA). CNA [3] relies on advanced natural language processing techniques, together with Social Network Analysis [4] measurements applied on the cohesion graph, to model discourse structure in terms of semantic links. The MD-CNA graph is a multi-layered graph that establishes semantic links between text elements of different granularities (i.e., the entire text, paragraphs, or sentences), including hierarchical inclusion links and links among elements of the same level. MD-CNA can be used to model both local and global cohesion, as it reflects the underlying semantic content of discourse within a document or between multiple texts [5].

In this study, we extend the CNA graph with lexical overlap links of two types (i.e., topic and content), together with coreference links, to better highlight dependencies between text fragments at different levels. We also introduce visualizations that highlight filtered links from the extended CNA graph, both within and between documents.

2 Method

The CNA graph [3] is centered mainly on semantic links computed using various models (e.g., Latent Semantic Analysis [6], Latent Dirichlet Allocation [7], word2vec [8], FastText [9], or Glove [10]), that can be established either between text elements of the same level (e.g., among sentences), or between different layers of the hierarchy (e.g., sentences relating to the constituent paragraphs). The CNA graph was extended for this study with two new types of links. First, *lexical overlap* is computed as a Jaccard distance over a bag of word representation of the text elements. Two types of measurements are performed after preprocessing the text using in spaCy[1]. *Content overlap* considers the usage of content words which include useful information from the text (i.e., lemmatized forms of words having as part-of-speech one of the following: nouns, verbs, adjectives, or adverbs). *Topic overlap* considers a more constrained view which takes into account only lemmas responsible for text contextualization and inducing actions (i.e., only nouns and verbs).

Second, *coreference links* are identified using NeuralCoref[2], which includes a mentions-detection module based on rules built on top of spaCy, together with a feed-forward neural-network to identify relevant pairs of mentions. The resulting clusters of

[1] https://spacy.io.

[2] https://github.com/huggingface/neuralcoref.

co-referring mentions are used to enrich the CNA graph structure. All follow-up visualizations rely on this extended CNA graph that is rendered both using only one reference text, as well as sequences of documents.

3 Visualization Use Cases and Discussion

Two types of visualizations are introduced here, together with preliminary use cases to illustrate their extensive applicability. First, a *hierarchical view* groups nodes by granularity level (i.e., document, paragraph, and sentence), followed by the rendering of different types of links from the MD-CNA graph using a tree-structure of discourse (see Fig. 1). The corresponding use case explores differences between high-low cohesion documents from the study performed by McNamara, Louwerse, McCarthy and Graesser [11]. All types of links are filtered within the rendered visualizations by minimum similarity thresholds available for each type of link. These values can be easily adjusted within the user interface. For this use case, topic overlap was set at 0.4, content overlap was established at 0.3, and high level of semantic similarity (0.7) was imposed.

Sentences from the same paragraph share its color. The size of each node is proportional to its semantic degree – i.e., the sum of all in-bound and out-bound semantic links above a statically imposed threshold, which ensures a sufficiently high semantic relatedness based on the context and readers (for Fig. 1, we considered the average plus standard deviation of all links, at each analysis level). On mouse-over, the link is colored in red, and a tooltip is displayed containing relevant details, including: link type, inter-connected text elements, similarity value (for content and semantic links) or pairs within the coreference cluster identified between the two nodes. The text from Fig. 1.a has low cohesion – only 2 semantic links are above the imposed threshold (i.e., links between sentences 1.1–1.4, and 1.6–1.9 respectively). The text was modified to increase its cohesion and, as expected, there are considerably more links (2 versus 4 topic overlap links, 6 versus 6 content overlap links, 6 versus 14 semantic links, and 3 versus 8 coreference links; 17 versus 32 total links), covering more text elements which are distributed throughout the entire document. Moreover, the semantic degree of most nodes is higher, mainly in the first 2 paragraphs.

The hierarchical view depicts only within document links, as the input consists of one text. The views are useful for analyzing text structure and cohesion, both locally at sentence level, as well as globally, between paragraphs. We can also observe cohesive sections of text and potential cohesion gaps, further providing improvement recommendations in terms of structure. Importantly, MD-CNA affords a *visual illustration* of cohesive links within a document, affording greater ease for researchers and educators in recognizing text cohesion and potentially increasing it for students.

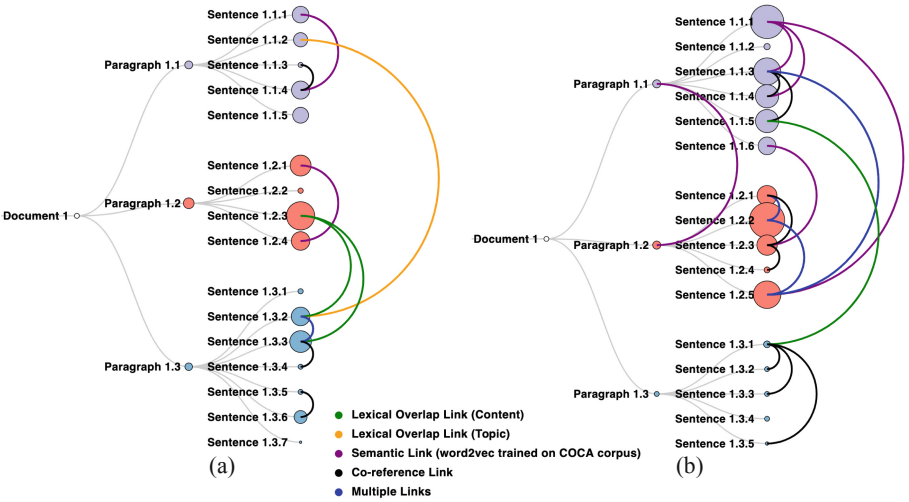

Fig. 1. CNA graph for a) low cohesion text; b) revised text having a high cohesion. (Color figure online)

Second, the *grid view* ensures ease of visualization for links spanning across multiple documents (see Fig. 2). The corresponding use case explores the task of multi-document comprehension on the collection of four documents used in the experiments performed by Nicula, Perret, Dascalu and McNamara [5]. Topic overlap was set at 0.1 due to a more diverse vocabulary, content overlap was kept at 0.2, while semantic similarity was increased to 0.75 to reduce the clutter generated by a dense semantic network.

This visualization shows connections both within (curved lines) and between documents (straight lines). The view can be rendered at two granularity levels (i.e., paragraph and sentence); for the second option, sentences have the same color as their corresponding paragraph. Documents are rendered as different columns in the grid, with constituent text elements displayed sequentially. As it can be observed, all documents are tightly related, with the 1st and 4th document containing many intratextual and intertextual semantic links. This second view enables researchers and educators to easily identify and trace semantically similar text segments between multiple documents, as well as to provide support to better target representative information (e.g., encourage bridging across multiple texts). This view can also be used to guide tutors to adequately order texts for presentation to learners, as well as to formulate comprehension questions that address a cohesive context spanning across multiple texts.

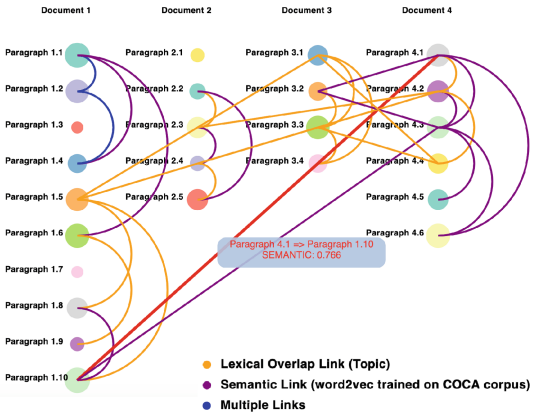

Fig. 2. CNA graph for multi-document analysis at paragraph level. (Color figure online)

In summary, the views provided in this study represent visual aids for researchers and educators to adequately evaluate and select texts to maximize cohesion flow and ease presentation of reading material. Our visualizations are designed to also scaffold readers to establish connections between texts and integrated concepts across documents, facilitating a more coherent understanding from separate sources of information.

Acknowledgments. This research was been funded by the "Semantic Media Analytics – SEMANTIC", subsidiary contract no. 20176/30.10.2019, from the NETIO project ID: P 40 270, MySMIS Code: 105976, the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125, the Institute of Education Sciences (R305A180144, R305A180261, and R305A190063), and the Office of Naval Research (N00014-17-1-2300 and N00014-19-1-2424). The opinions expressed are those of the authors and do not represent views of the IES or ONR.

References

1. McNamara, D.S.: SERT: self-explanation reading training. Discourse Process. **38**, 1–30 (2004)
2. O'Reilly, T., McNamara, D.S.: Reversing the reverse cohesion effect: good texts can be better for strategic, high-knowledge readers. Discourse Process. **43**(2), 121–152 (2007)
3. Dascalu, M., McNamara, D.S., Trausan-Matu, S., Allen, L.K.: Cohesion network analysis of CSCL participation. Behav. Res. Methods **50**(2), 604–619 (2018)
4. Scott, J.: Social Network Analysis. Sage, London (2017)
5. Nicula, B., Perret, C.A., Dascalu, M., McNamara, D.S.: Predicting multi-document comprehension: cohesion network analysis. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 358–369. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_30
6. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychol. Rev. **104**(2), 211–240 (1997)

7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(4–5), 993–1022 (2003)
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representation in vector space. In: Workshop at ICLR, Scottsdale, AZ (2013)
9. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
10. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: The 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014), vol. 14. ACL, Doha (2014)
11. McNamara, D.S., Louwerse, M.M., McCarthy, P.M., Graesser, A.C.: Coh-Metrix: capturing linguistic features of cohesion. Discourse Process. **47**(4), 292–330 (2010)