Exploring Dialogism Using Language Models

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

[1] University Politehnica of Bucharest, 313 Splaiul Independentei,
060042 Bucharest, Romania
{stefan.ruseti,dorinela.dascalu,dragos.corlatescu,
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, USA
dsmcnama@asu.edu

Abstract. Dialogism is a philosophical theory centered on the idea that life involves a dialogue among multiple voices in a continuous exchange and interaction. Considering human language, different ideas or points of view take the form of voices, which spread throughout any discourse and influence it. From a computational point of view, voices can be operationlized as semantic chains that contain related words. This study introduces and evaluates a novel method of identifying semantic chains using BERT, a state-of-the-art language model for computational linguistics. The resulting model generalizes to multiple relations including repetitions, semantically related concepts from WordNet (i.e., synonyms, hypernyms, hyponyms, and siblings), as well as pronominal resolutions. By combining the attention scores between words, word pairs are merged into connected components that denote emerging voices from the discourse. The introduced visualization argues for a more dense capturing of inner semantic links between words and even compound words in contrast to classical methods of building lexical chains.

Keywords: Dialogism · Semantic chains · Language models

1 Introduction

Dialogism is a philosophical theory introduced by Mikhail Bakhtin [1,2], centered on the idea that everything, even life, is dialogic, a continual exchange and interaction between voices: "Life by its very nature is dialogic... when dialogue ends, everything ends" [2]. Trausan-Matu et al. [3] extended the concept of voice for analyzing discourse, in general, and collaborative learning, in particular. They consider voices to be generalized representations of different points of view or ideas, which spread throughout the discourse, and influence it. Voices were subsequently operationalized by Dascalu et al. [4] as semantic chains that were

I. Roll et al. (Eds.): AIED 2021, LNAI 12749, pp. 296–301, 2021.
https://doi.org/10.1007/978-3-030-78270-2_53

obtained by combining lexical chains, i.e., sequences of repeated or related words, including synonyms or hypernyms [5]. Semantic chains propagate along sentences and help create narrative threads throughout the text.

Recent studies on building lexical chains consider word repetitions, synonyms, and semantic relationships between nouns [6]. Mukherjee et al. [6] used lexical chains to distinguish easy from difficult medical texts. Identifying lexical chains that signal a difficult sentence helps in the simplification process. Olena [7] proposed a method for identifying lexical chains based on graphs, in which the nodes represent the terms in the document, and the edges the semantic relations between them. More recently, Ruas et al. [8] combined lexical chains with word embeddings to classify documents.

We introduce and evaluate a novel operationalization of voices using BERT [9], a state-of-the-art language model. This model enhances even further the Cohesion Network Analysis graph from the ReaderBench framework [10,11] by integrating semantic links of related concepts, indicative of semantic flow [12].

2 Method

A specific dataset with examples of links was required to identify the attention heads from BERT capable of detecting semantic links between words that belong to the same chain. A set of simple heuristics were used to extract links from sample texts, for all pairs of words tagged as noun, verb, or pronoun that fulfil one of the following conditions: a) repetitions of words having the same lemma; b) synonyms, hypernyms, or siblings in the WordNet taxonomy [13]; and c) coreferences identified using spaCy[1]. The TASA corpus[2] was selected as reference due to its diversity and covered complexity levels. The "correct" pairs were extracted from the entire dataset using the previous rules, while the "incorrect" ones were randomly sampled with 10% probability from all pairs of words that were not selected (i.e., otherwise, the number of negative samples would have been one order of magnitude larger than "correct" semantic associations). In total, 49 million word pairs were extracted, out of which around 20 million were positive examples.

Transformer-based models, in particular BERT [9], build contextual representations of words by stacking multi-head attention layers. Besides state-of-the-art results obtained on a vast range of tasks in Natural Language Processing, these models also provide insights regarding the importance of words and the relations between them by looking at the attention values. Clark et al. [14] explored the interpretability of different attention heads from different layers of the BERT model. The authors show that attention heads can be used to identify specific syntactic functions or perform coreference resolution.

No single attention head is accurate enough to predict these kinds of semantic relationships between words. Therefore, a prediction model that learns to combine the attention values from all the attention heads between two words was

[1] https://www.spacy.io, Retrieved April 15th, 2021.

[2] http://lsa.colorado.edu/spaces.html, Retrieved April 15th, 2021.

trained on the dataset constructed based on TASA. By considering both directions of the attention heads, 288 scores were used in total, similar to the approach used by Clark et al. [14]. An issue to be tackled was the limited sequence length accepted by the pretrained BERT model (i.e., 512 tokens). Texts in the TASA dataset, but also in general, can be longer; thus, a sliding window was used to compute the attention weights for all pairs of words. The sliding window had a length of 256 for efficiency reasons, but also because semantic chains usually do not contain links that are too far apart. An overlap of 128 tokens was used so that words on different sides of the window could still be connected; if two different attention values are computed between the same two words (because of this overlap), the maximum value was used as the weight.

The previously described prediction model was used to score all pairs of words that are within a given distance in the text. The next step consisted of grouping these pairs of words into sets of semantically related words, i.e., semantic chains. In order to filter the links based on the predicted weight, a fixed threshold was experimentally set at 0.90. The semantic chains are selected in the form of connected components from the resulting graph.

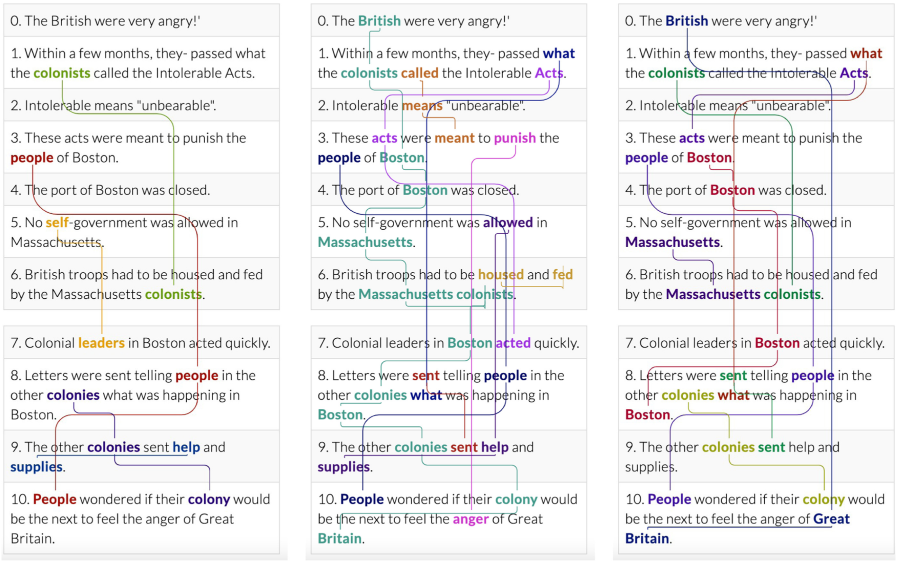
3 Results

Different architectures for identifying semantic links were trained and evaluated: a linear model that only computes one weight for each attention head, and Multi-Layer Perceptron (MLP) with one or two hidden layers. All models return one number passed through a Sigmoid activation (see Table 1).

Table 1. Link prediction results.

Model	Hidden layer size	Accuracy (%)
Linear	–	79.75
MLP	16	85.67
MLP	32	86.24
MLP	64	86.65
MLP	64, 64	87.43
MLP	128, 64	87.99

An interactive view developed using Angular 6 (https://angular.io) was introduced to display the semantic chains - see Fig. 1 for a text selected from the dataset described in McNamara et al. [15]. Each sentence is represented in a row, while rows are grouped in their corresponding paragraph. Words and links from a semantic chain share the same color. A higher density of the chains extracted with our method can be observed in contrast to classical lexical chains. Surprising relations not present in the constructed dataset can be seen in the generated chains. The linear model found connections between “colonists” and “Boston”, or between “help” and “supplies”, while the MLP model identified connections between “British” and “Great Britain” as a compound word. This example also shows that choosing the best model between linear and MLP is not straightforward, despite the substantial performance improvement of the latter on the word pairs dataset. Even though the linear model cannot perfectly learn the simple heuristics used to build the initial dataset, it can retrieve new insightful connections between words.



0. The British were very angry!'
1. Within a few months, they- passed what the **colonists** called the Intolerable Acts.
2. Intolerable means "unbearable".
3. These acts were meant to punish the **people** of Boston.
4. The port of Boston was closed.
5. No **self**-government was allowed in Massachusetts.
6. British troops had to be housed and fed by the Massachusetts **colonists**.
7. Colonial **leaders** in Boston acted quickly.
8. Letters were sent telling **people** in the other **colonies** what was happening in Boston.
9. The other **colonies** sent **help** and **supplies**.
10. **People** wondered if their **colony** would be the next to feel the anger of Great Britain.

(a)

0. The **British** were very angry!'
1. Within a few months, they- passed **what** the **colonists called** the Intolerable **Acts**.
2. Intolerable **means** "unbearable".
3. These **acts** were **meant** to **punish** the **people** of **Boston**.
4. The port of **Boston** was closed.
5. No self-government was **allowed** in **Massachusetts**.
6. British troops had to be **housed** and **fed** by the **Massachusetts colonists**.
7. Colonial leaders in **Boston acted** quickly.
8. Letters were **sent** telling **people** in the other **colonies what** was happening in **Boston**.
9. The other **colonies sent help** and **supplies**.
10. **People** wondered if their **colony** would be the next to feel the **anger** of Great **Britain**.

(b)

0. The **British** were very angry!'
1. Within a few months, they- passed **what** the **colonists** called the Intolerable **Acts**.
2. Intolerable means "unbearable".
3. These **acts** were meant to punish the **people** of **Boston**.
4. The port of **Boston** was closed.
5. No self-government was allowed in **Massachusetts**.
6. British troops had to be housed and fed by the **Massachusetts colonists**.
7. Colonial leaders in **Boston** acted quickly.
8. Letters were **sent** telling **people** in the other **colonies what** was happening in **Boston**.
9. The other **colonies sent** help and supplies.
10. **People** wondered if their **colony** would be the next to feel the anger of **Great Britain**.

(c)

Fig. 1. Visualizations of a) lexical chains [5], b) semantic chains using the linear model, and c) semantic chains using the MLP model.

4 Conclusions

A novel method for identifying semantic links is introduced using only the attention scores computed by BERT, a core task for operationalizing dialogism as a discourse model. Choosing which attention heads are relevant for this task and how to combine them was achieved by building a dataset with pairs of words with simple rules. The introduced visualization argues for a more dense capturing of inner semantic links between words and even compound words, which are quite sparse when considering manually defined synsets from WordNet. Our aim is to further extend this model with sentiment analysis features derived from local contexts captured by BERT, thus further enriching the analysis with the identification of convergent and divergent points of view.

Acknowledgments. This research was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number TE 70 PN-III-P1-1.1-TE-2019-2209, ATES – "Automated Text Evaluation and Simplification", the Institute of Education Sciences (R305A180144 and R305A180261), and the Office of Naval Research (N00014-17-1-2300; N00014-20-1-2623). The opinions expressed are those of the authors and do not represent views of the IES or ONR.

References

1. Bakhtin, M.M.: The Dialogic Imagination: Four Essays. The University of Texas Press, Austin and London (1981)
2. Bakhtin, M.M.: Problems of Dostoevsky's Poetics. University of Minnesota Press, Minneapolis (1984)
3. Trausan-Matu, S., Stahl, G., Sarmiento, J.: Supporting polyphonic collaborative learning. E-service J. Indiana Univ. Press **6**(1), 58–74 (2007)
4. Dascalu, M., Trausan-Matu, S., Dessus, P.: Voices' inter-animation detection with readerbench - modelling and assessing polyphony in CSCL chats as voice synergy. In: 2nd International Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 2nd International Conference on Systems and Computer Science (ICSCS), pp. 280–285. IEEE (2013)
5. Galley, M., McKeown, K.: Improving word sense disambiguation in lexical chaining. In: Gottlob, G., Walsh, T. (eds.) 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), pp. 1486–1488. Morgan Kaufmann Publishers, Inc. (2003)
6. Mukherjee, P., Leroy, G., Kauchak, D.: Using lexical chains to identify text difficulty: a corpus statistics and classification study. IEEE J. Biomed. Health Inform. **23**(5), 2164–2173 (2018)
7. Medelyan, O.: Computing lexical chains with graph clustering. In: Proceedings of the ACL 2007 Student Research Workshop, pp. 85–90 (2007)
8. Ruas, T., Ferreira, C.H.P., Grosky, W., de França, F.O., de Medeiros, D.M.R.: Enhanced word embeddings using multi-semantic representation through lexical chains. Inf. Sci. (2020)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Dascălu, M.: Analyzing Discourse and Text Complexity for Learning and Collaborating. SCI, vol. 534. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-03419-5
11. Dascalu, M., Trausan-Matu, S., McNamara, D., Dessus, P.: Readerbench - automated evaluation of collaboration based on cohesion and dialogism. Int. J. Comput.-Support. Collab. Learn. **10**(4), 395–423 (2015)
12. O'Rourke, S., Calvo, R.: Analysing semantic flow in academic writing. In: Dimitrova, V., Mizoguchi, R., du Boulay, B., Graesser, A. (eds.) Artificial Intelligence in Education. Building Learning Systems That Care: From Knowledge Representation to Affective Modelling (AIED 2009), pp. 173–180. IOS Press, Amsterdam, The Netherlands (2009)
13. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)

14. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does BERT look at? An analysis of BERT's attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 276–286 (2019)
15. 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)