TRoTR: A Framework for evaluating the recontextualization of text

Francesco Periti, University of Milan, francesco.periti@unimi.it
Pierluigi Cassotti, University of Gothenburg, pierluigi.cassotti@gu.se
Stefano Montanelli, University of Milan, stefano.montanelli@unimi.it
Nina Tahmasebi, University of Gothenburg, nina.tahmasebi@gu.se
Dominik Schlechtweg, University of Stuttgart, dominik.schlechtweg@ims.uni-stuttgart.de

EMNLP 2024

In a nutshell

- Introduced the **Topic Relatedness of Text Reuse** (TRoTR) framework to model recontextualization in text reuse.
- Defined two NLP tasks, Text Reuse in-Context (**TRiC**) and Topic variation Ranking across Corpus (**TRaC**).
- Developed a benchmark with human-annotated *topic relatedness* labels on biblical text reuses extracted from Twitter (now X).
- Proposed a new annotation process for modeling *topics* through relatedness in context pairs.
- Established a baseline evaluation of SBERT models, showing that the presence of common substrings can bias computational judgments.

Tasks

In the TRoTR tasks, instances of text reuse are presented within different contexts, each representing a new recontextualization of the original text

TRiC

Text Reuse in Context frames a text reuse *t* within two different contexts *c1* and *c2*. The goal is to assess the topic relatedness of *c1* and *c2*.

Subtask 1: *binary classification*
Subtask 2: *ranking*.

TRaC

Topic variation Ranking across Corpus frames a text reuse *t* within a corpus *C* that includes various contexts *c* where *t* occurs.

Annotation

Guidelines: Your task is to rate the degree of *topic relatedness* between two texts in which a text sequence is used. [...]

. . . check out our paper for full guidelines . . .

1 *context* **text reuse** *context*

2 *context* **text reuse** *context*

Scale	
4	Identical
3	Closely Related
2	Distantly related
1	Unrelated
–	Can't decide

We avoid explicit topic annotation by adopting the annotation paradigm from the Word-in-Context task (Pilehvar et al., 2019). Annotators are asked to rate topic relatedness instead of assigning labels.

- **TRiC labels (*subtask 2*):** we average the judgments of all annotators
- **TRiC labels (*subtask 1*):** we binarize the average judgment using a threshold of 2.5 (the midpoint of the scale)
- **TRaC labels:** we average the judgments of all annotators over all instances for a target

Experimental results

- **Bi-Encoder vs. Cross-Encoder:** we compared the performance of Bi-Encoder and Cross-Encoder architectures in our base TRiC task. Bi-Encoder models demonstrated superior results. Based on this, we decided to proceed with fine-tuning only on the Bi-Encoder model.
- **Pre-trained vs. Fine-tuned:** we compared the performance of pre-trained and fine-tuned models, with fine-tuning via *contrastive learning*. While fine-tuning improved performance over the baseline, the overall improvement remained moderate.
- **Standard vs. Masked:** to assess the impact of common substrings, we experimented by masking the shared text reuse. Consistently higher results were achieved when the common text reuse was masked. Our evaluation reveals that SBERT models exhibit a bias toward their pre-training focus on semantic similarity, influencing the computational judgment of topic relatedness between sentences.

Background

Text reuse: the reuse of existing written sources in the creation of a new text. (Clough et al., 2002)

Text reuse detection: text reuses are all assumed as "*topically related to the source*" (Hagen et al., 2011), the boundaries of reused text are unknown, and the goal is to detect text reuse across a diachronic corpus (Seo et al., 2008).

Recontextualization: the dynamic transfer-and transformation of a text from one discourse/text-incontext to another (Connolly, 2014).

Topic: our definition follows the popular notion of *what the text is about* (Bauwelinck et al., 2020).

Topic relatedness: TRoTR is grounded on a specific facet of semantic relatedness that considers *the extent to which two texts share a common topic*.

1
It's the wonderful pride month!! ♥ ♥ ♥ ♥ ♥ ♥ Honestly pride is everyday! Love is love don't forget I love you ♥. Remember this! John 15:12-13: "My command is this: Love each other as I have loved you. **Greater love has no one than this: to lay down one's life for one's friends**"

2
At a large Crimean event today Putin quoted the Bible to defend the special military operation in Ukraine which has killed thousands and displaced millions. His words "**There is no greater love than if someone gives soul for their friends**". And people were cheering him. Madness!!!

Consider three recontextualizations of the biblical passage ***John 15:13.***

- Text **1** has a different topic with respect to Text **2** and **3**.
- Text **2** and **3** are *topic related.*

3
"Freeing people from genocide is the reason, motive & goal of the military operation we started in the Donbas & Ukraine", Putin says, then quotes the Bible: "**There is no greater love than to lay down one's life for one's friends.**" It's like Billy Graham meets North Korea

Data

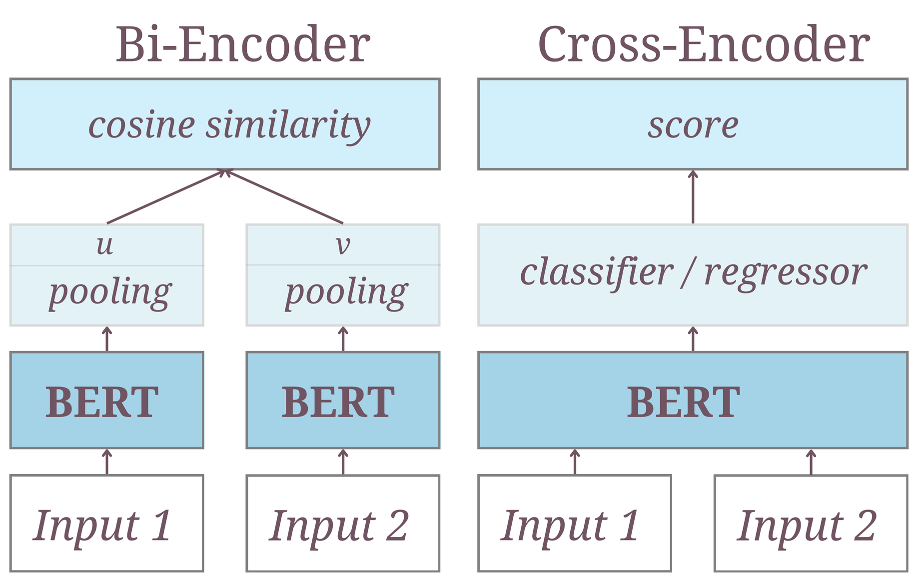
Biblical text reuse: Inspired by Moritz et al. (2016); we focus on text reuse in biblical passages due to their high context variety (Cheong, 2014). Moreover, they are often cited explicitly with references (e.g., *John 15:13*).

Tweets were collected through a manual search process, thus allowing us to avoid a Text Reuse Detection phase and its validation.

For a set of **42 target** passages, we collected **30 tweets** each.

10-fold validation: To strengthen the robustness of the evaluation, we generate ten randomized Train-Dev-Test splits and set the average performance across all the splits as reference for comparison.

Evaluation



TRiC *subtask 1*

We train a threshold classifier based on sentence similarities. The threshold is determined using the Dev set and then applied to the Test set. We evaluate the performance using the **F1-score**.

TRiC *subtask 2*

We use raw sentence similarities between sentence pairs. We evaluate the performance using **Spearman's correla-tion coefficient.**

TRaC

For each target reuse, we calculate the average similarity over all sentence pairs. We evaluate the performance using **Spearman's correlation coefficient.**

References

Nina Bauwelinck and Els Lefever. 2020. **Annotating Topics, Stance, Argumentativeness and Claims in Dutch Social Media Comments: A Pilot Study**. In Proc. of ArgMining, pages 8–18, Online. ACL.

Paul Clough, Robert Gaizauskas, Scott S.L. Piao, and Yorick Wilks. 2002. **Measuring Text Reuse**. In Proc. of ACL, pages 152–159, Philadelphia, Pennsylvania, USA. ACL.

Matthias Hagen and Benno Stein. 2011. **Candidate Document Retrieval for Web-Scale Text Reuse Detection**. In String Processing and Information Retrieval, pages 356–367, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jangwon Seo and W. Bruce Croft. 2008. **Local Text Reuse Detection**. In Proc. of SIGIR, page 571–578, New York, NY, USA. ACM.

John H. Connolly. 2014. **Recontextualisation, Resemiotisation and Their Analysis in Terms of an FDG-based Framework**. Pragmatics, 24(2):377–397.

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. **Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and its Application to Bible Reuse**. In Proc. of EMNLP, pages 1849–1859, Austin, Texas. ACL.

Pauline Hope Cheong. 2014. **Tweet the message? religious authority and social media innovation.** Journal of Religion, Media and Digital Culture, 3(3):1 – 19.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. **WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations**. In Proc. of NAACL, pages 1267–1273, Minneapolis, Minnesota. ACL.

. . . check out our paper for further references . . .

UNIVERSITÀ DEGLI STUDI DI MILANO
LA STATALE
UNIVERSITY OF GOTHENBURG
CHANGE IS KEY