



Deutsches  
Forschungszentrum  
für Künstliche  
Intelligenz GmbH



# Semantic Relations between Text Segments for Semantic Storytelling: Annotation Tool – Dataset – Evaluation

LREC 2022

Michael Raring, Malte Ostendorff, Georg Rehm

DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany  
michael.raring@gmail.com, {malte.ostendorff, georg.rehm}@dfki.de

---

# Semantic Storytelling

## Vision

Automatic and semi-automatic generation of stories based on extracted, processed, classified and annotated information from large content resources <sup>[1][2][3]</sup>

## Goal

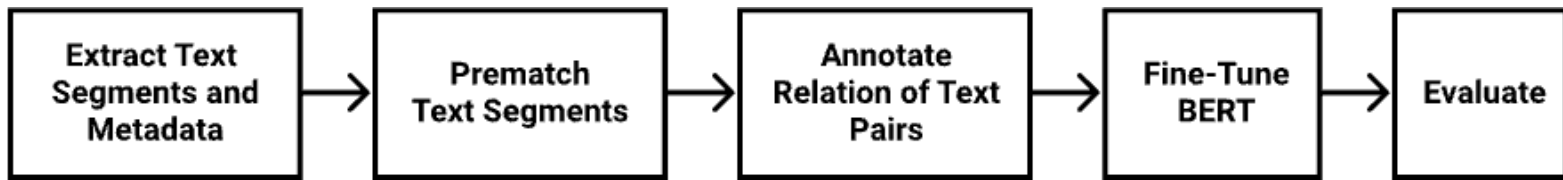
Support content curators in creating new storylines through the relevant information extracted and presented by a corresponding tool <sup>[3]</sup>

## Building Blocks / Core Process

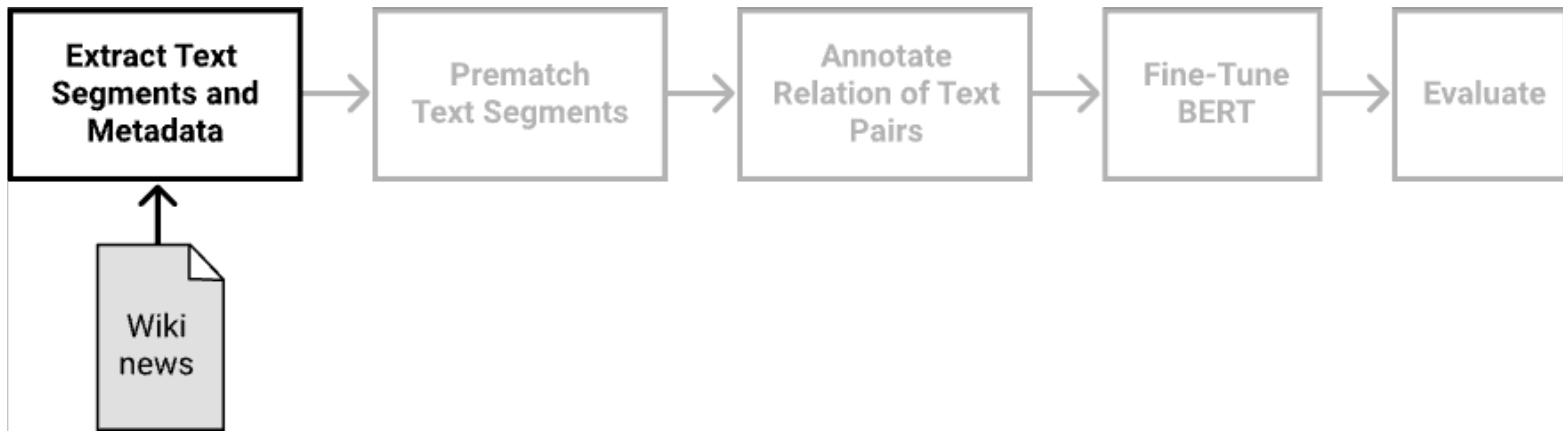
Determine...

1. the **relevance** of a segment
2. the **importance**
3. the **semantic relation** between two segments <sup>[2]</sup>

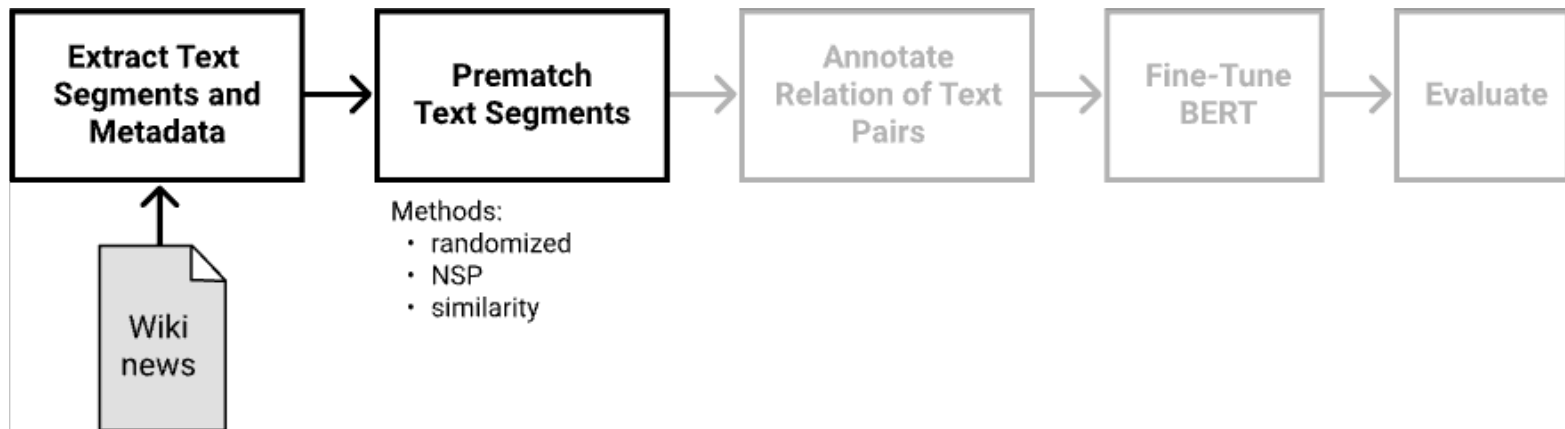
# Process of the Experiment



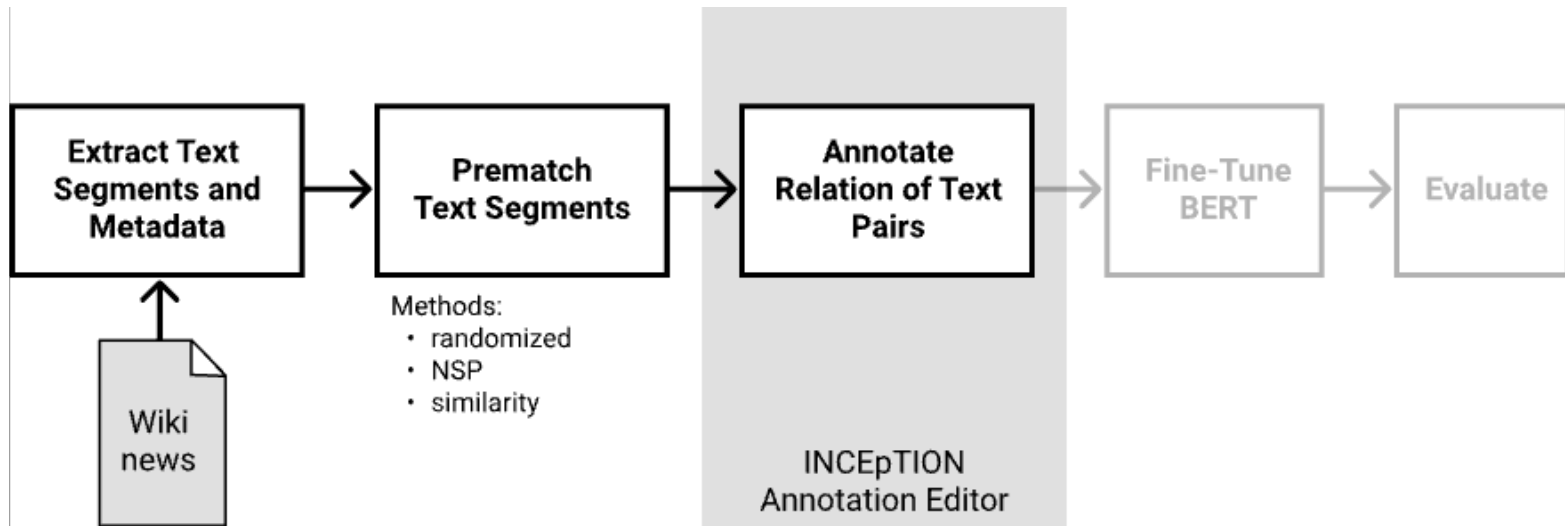
# Process of the Experiment



# Process of the Experiment



# Process of the Experiment



# Annotation Software

## INCEpTION

- Open-source web-based annotation platform <sup>[13]</sup>
- Annotation of various NLP-related features
- Project administration features
  - Annotation layers
  - Annotation classes
  - Annotators
  - Documents
  - Knowledge resources
  - Recommender systems
- ...
- Data structure: Apache UIMA CAS (Unstructured Information Management Architecture / Common Analysis System) <sup>[14]</sup>



# Annotation Editor

The screenshot displays the Annotation Editor interface. At the top, a navigation bar includes 'Pairwise Navigation' with left, right, and double-right arrows, a 'Progress' indicator at 16% (146/915), and buttons for 'Filter', 'Statistics', and 'Save'. A red box labeled '1' highlights this navigation bar.

Below the navigation bar are two text segments. The left segment contains the text: 'The unemployment rate in the United States has surpassed ten percent, according to figures released by the US Department of Labor on Friday.' Below this text is a '(1/1)' indicator and navigation arrows. A source document is listed as 'US unemployment rate surpasses 10%'. A blue box labeled '2' highlights this segment.

The right segment contains the text: 'Data from the United States Labor Department says that the unemployment rate in the U.S. has reached its highest level in over 25 years, namely 9.4%.' Below this text is a '(1/1)' indicator and navigation arrows. A source document is listed as 'US unemployment rate reaches 9.4 percent'. A blue box labeled '2' highlights this segment.

In the center, between the two segments, is a relationship editor. It features two dropdown menus. The top dropdown is set to 'none' and has a green arrow pointing right above it. The bottom dropdown is set to 'temporal' and has a green arrow pointing left above it. A green box labeled '3' highlights this relationship editor.



# Annotated Sentence Pairs

**12 (7)**

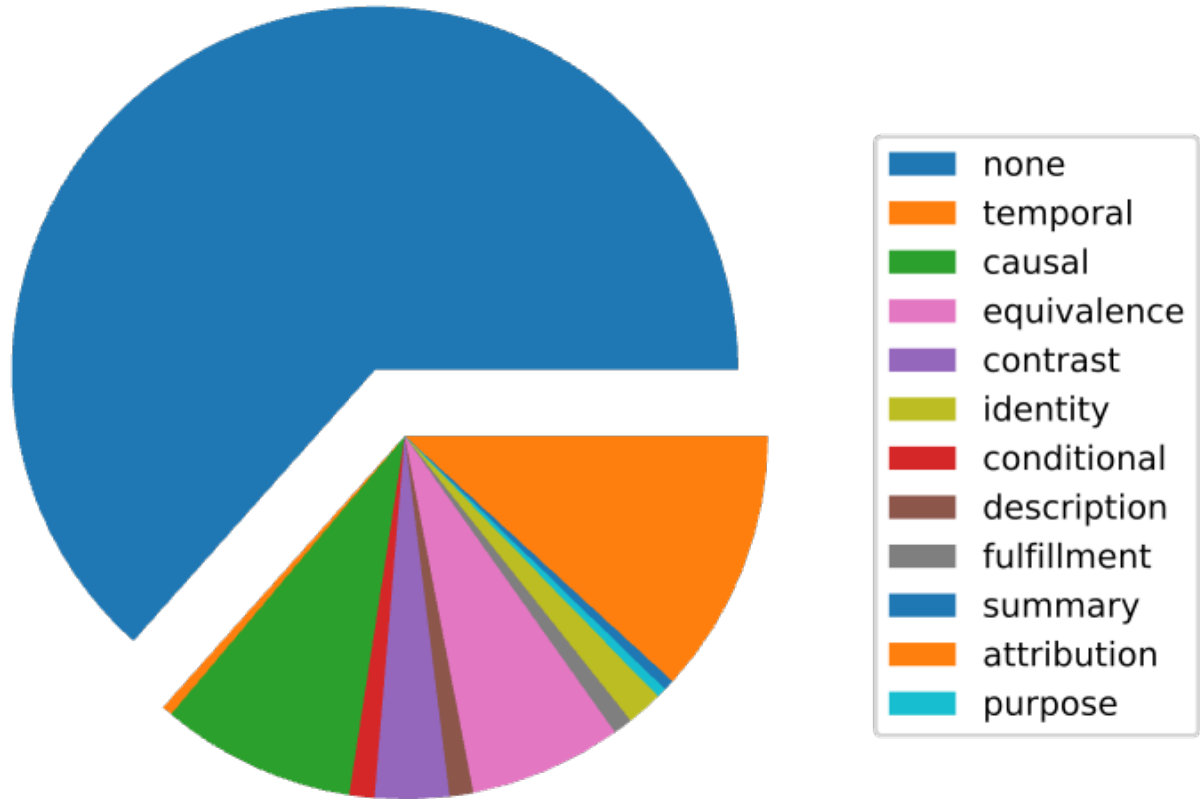
relation classes  
(subset)

**2,501**

sentence pairs

**2,638**

articles

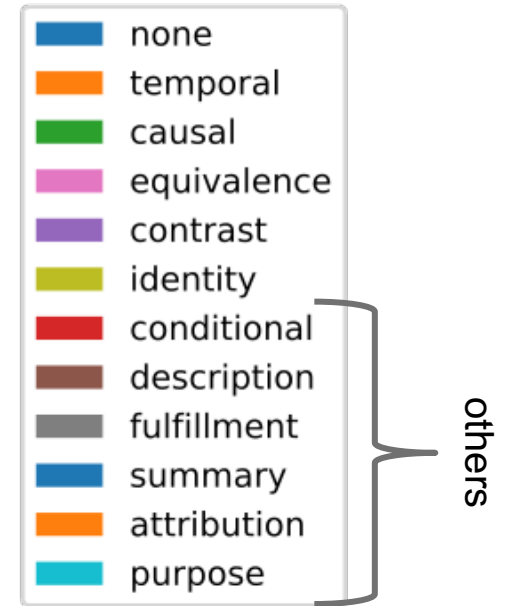
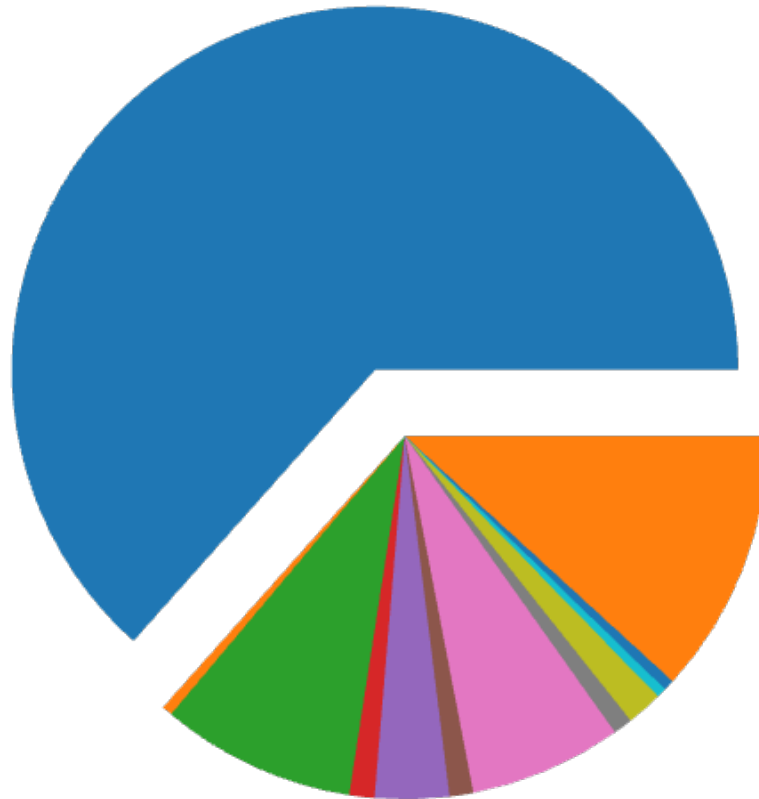


# Annotated Sentence Pairs

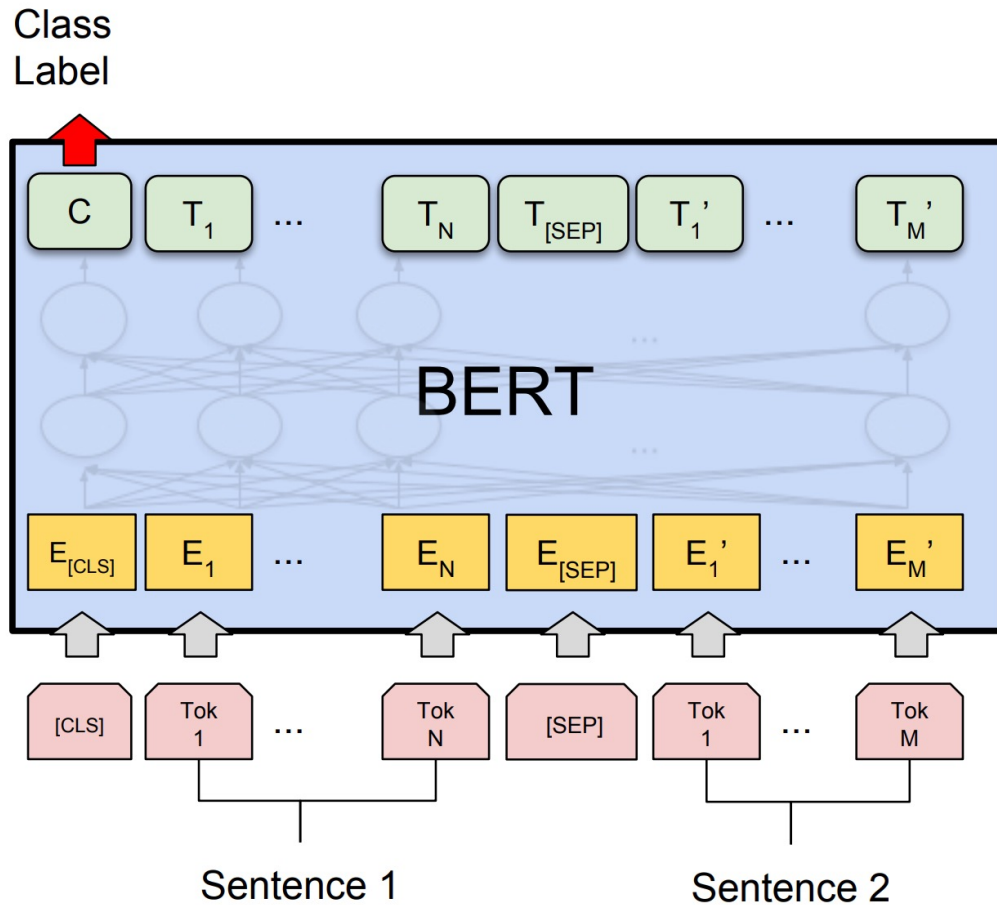
**12 (7)**  
relation classes  
(subset)

**2,501**  
sentence pairs

**2,638**  
articles



# Pretrained Language Model

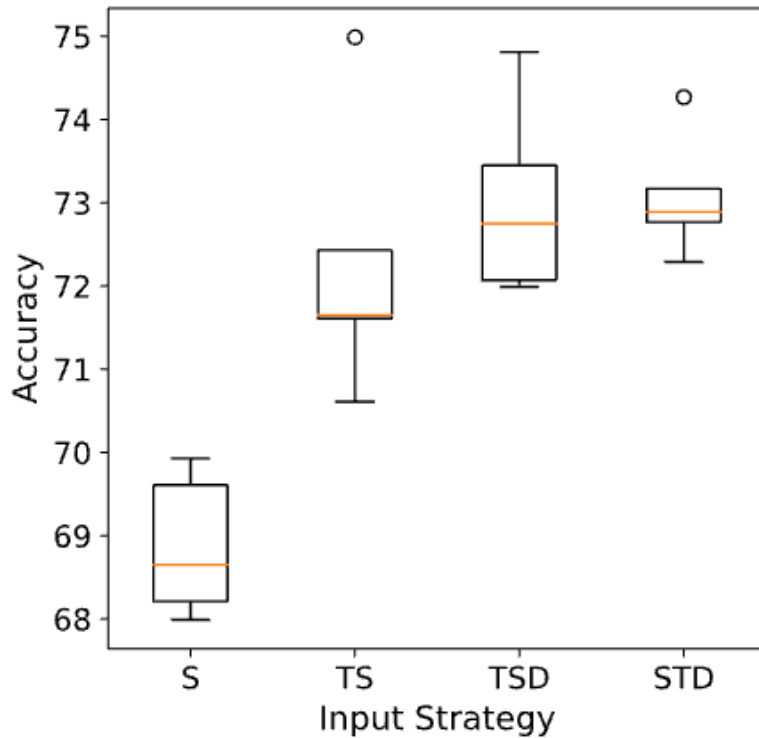


# Input Strategies

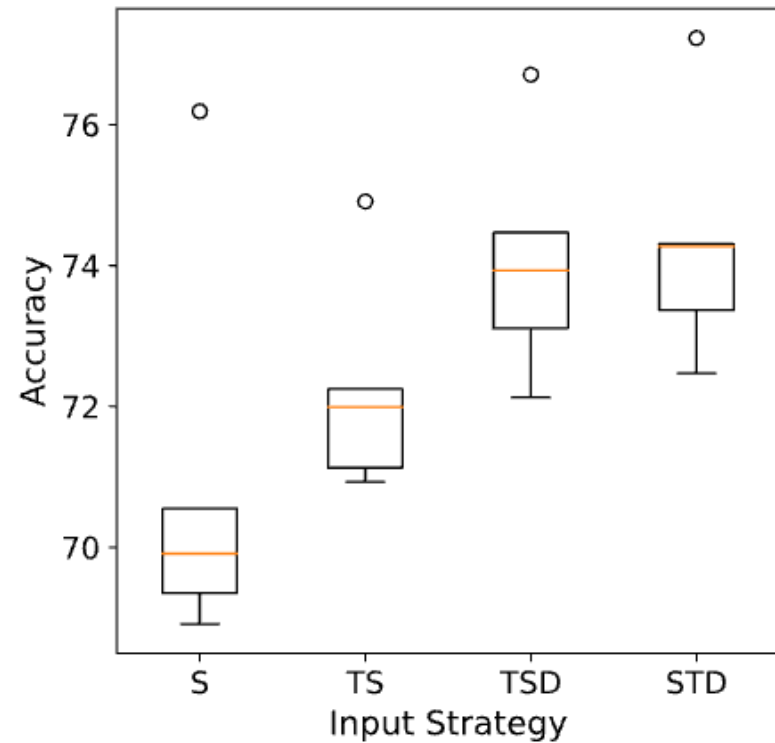
	Input Strategy	Example
S	Sentence.	However preliminary results based on 95% of the votes cast give Hamas' Change and Reform Party 76 seats, leaving Fatah with 43 seats.
TS	Title. Sentence.	<b>Hamas wins Palestinian election.</b> However preliminary results based on 95% of the votes cast give Hamas' Change and Reform Party 76 seats, leaving Fatah with 43 seats.
TSD	Title. Sentence. Date	<b>Hamas wins Palestinian election.</b> However preliminary results based on 95% of the votes cast give Hamas' Change and Reform Party 76 seats, leaving Fatah with 43 seats. <b>January 29, 2005</b>
STD	Sentence. Title. Date	However preliminary results based on 95% of the votes cast give Hamas' Change and Reform Party 76 seats, leaving Fatah with 43 seats. <b>Hamas wins Palestinian election.</b> <b>January 29, 2005</b>

# Results Depending on Input Strategy

a) Full classification

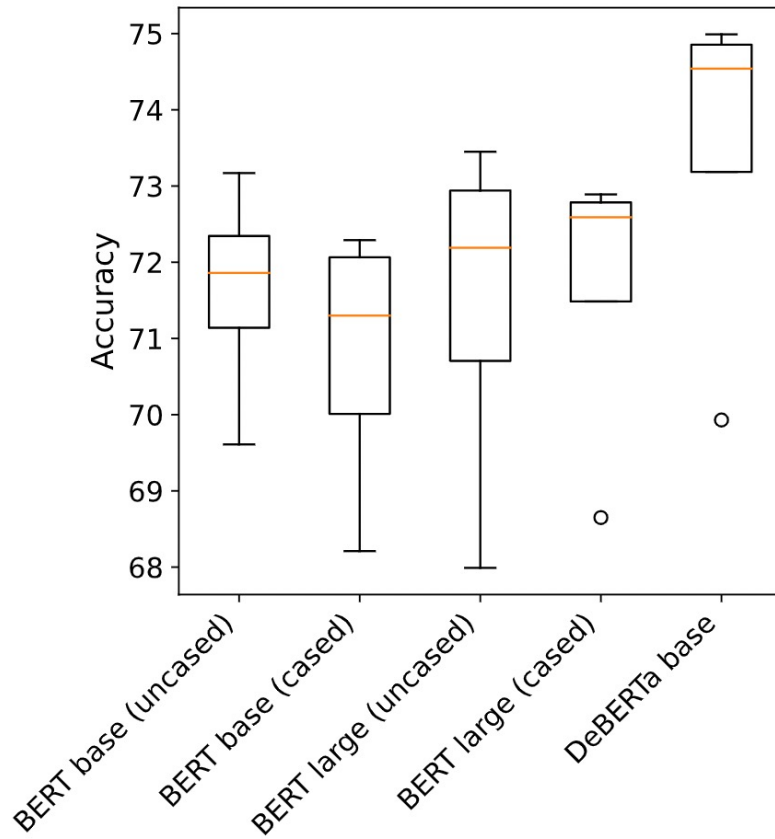


b) 7-classes classification

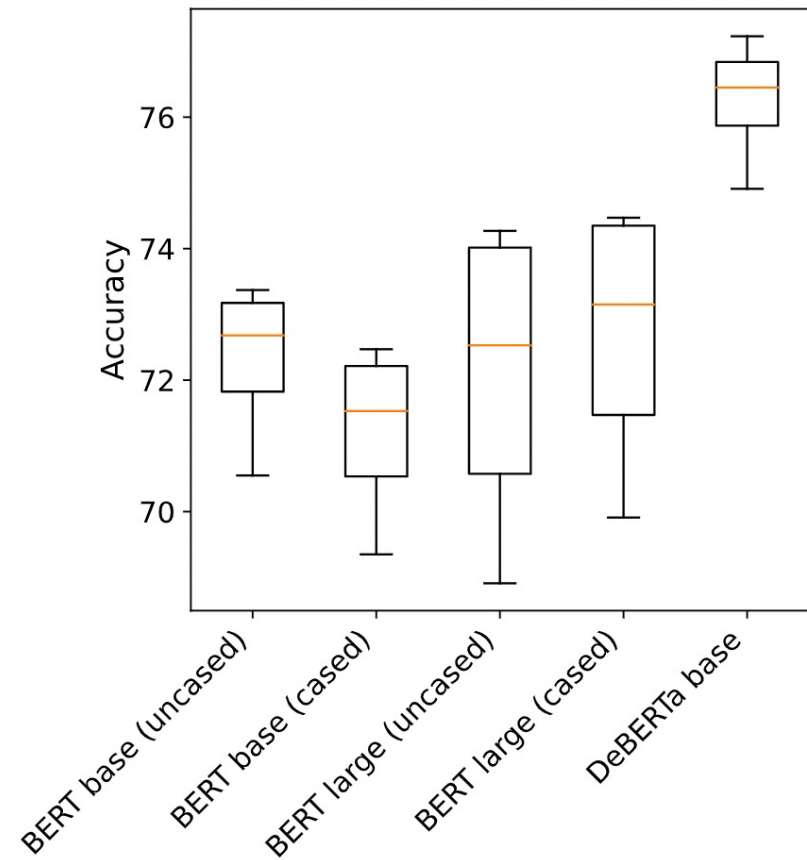


# Results Depending on Language Model

a) Full classification



b) 7-classes classification



# Class Results for Full Classification

Class	Precision	Recall	F1-score	Support
<i>None</i>	<b>87.0</b>	<b>86.0</b>	<b>86.5</b>	794
<i>Identity</i>	78.3	77.5	77.6	20
<i>Equivalence</i>	78.5	71.9	74.6	85
<i>Causal</i>	57.3	67.0	61.6	109
<i>Contrast</i>	55.0	66.1	59.9	41
<i>Temporal</i>	45.9	50.3	48.0	147
<i>Conditional</i>	24.1	21.4	21.5	14
<i>Description</i>	26.5	13.1	17.4	13
<i>Attribution</i>	20.8	8.3	11.8	6
<i>Fulfillment</i>	0.0	0.0	0.0	11
<i>Summary</i>	0.0	0.0	0.0	6
<i>Purpose</i>	0.0	0.0	0.0	6
<b>Micro avg.</b>	<b>74.6</b>	<b>75.0</b>	<b>74.6</b>	<b>1251</b>
<b>Macro avg.</b>	<b>39.5</b>	<b>38.5</b>	<b>38.2</b>	<b>1251</b>

# Class Results for 7-Classes Classification

Class	Precision		Recall		F1-score		Support
<i>None</i>	87.8	+0.8	<b>87.8</b>	+1.8	<b>87.8</b>	+1.3	794
<i>Identity</i>	<b>90.1</b>	+11.8	76.2	-1.3	82.2	+4.6	20
<i>Equivalence</i>	74.2	-4.3	68.4	-3.5	71.0	-3.6	85
<i>Causal</i>	59.9	+2.6	67.7	+0.7	63.4	+1.8	109
<i>Contrast</i>	66.0	+11.0	62.5	-3.6	64.1	+4.2	41
<i>Temporal</i>	50.5	+4.6	53.5	+3.2	51.9	+3.9	147
<i>Others</i>	41.6		33.5		36.9		55
<b>Micro avg.</b>	<b>77.3</b>	+2.7	<b>77.2</b>	+2.2	<b>77.2</b>	+2.6	<b>1251</b>
<b>Macro avg.</b>	<b>67.2</b>	+27.7	<b>64.2</b>	+25.7	<b>65.3</b>	+27.1	<b>1251</b>



# Conclusion

- We have developed an annotation editor for pairwise relations and manually annotated the semantic relation between **2500 sentence pairs** individually extracted from Wikinews articles
- We successfully trained a multi-class classifier to recognize semantic relations between text segments from different source documents by fine-tuning multiple versions of BERT's language model with our annotated dataset
- We ran all trainings with a **12-class & 7-class** setup. The 7-class classification performed better in micro & macro metrics
- DeBERTa (base) performed best with **accuracy and micro F1-score of 75% (12-class) and 77% (7-class)**
- Adding metadata such as the article name and publication date has on average led to slightly better and more stable results
- Future work:
  - **Dataset:** Improve imbalance, classes, pre-matching, pre-processing
  - **Classifier:** Use more recent models, test few-shot approaches

# Thanks! Any questions?



Michael Raring



[hub.cards/michael.raring](https://hub.cards/michael.raring)

[michael.raring@gmail.com](mailto:michael.raring@gmail.com)

In cooperation with:



Deutsches  
Forschungszentrum  
für Künstliche  
Intelligenz GmbH



Reproduce this research:



Editor Source Code + Scripts + Dataset + Best Model:

<https://github.com/DFKI-NLP/semantic-storytelling>

# References (1)

- [1] Rehm, G., Zaczynska, K., and Moreno, J. (2019). Semantic storytelling: Towards identifying storylines in large amounts of text content. In *Text2Story@ ECIR*, pages 63–70.
- [2] Rehm, G., Zaczynska, K., Moreno-Schneider, J., Ostendorff, M., Bourgonje, P., Berger, M., Rauenbusch, J., Schmidt, A., and Wild, M. (2020). Towards discourse parsing-inspired semantic storytelling.
- [3] Rehm, G., Zaczynska, K., Bourgonje, P., Ostendorff, M., Moreno-Schneider, J., Berger, M., Rauenbusch, J., Schmidt, A., Wild, M., Bo'ttger, J., Quantz, J., Thomsen, J., and Fricke, R. (2021). Semantic Storytelling: From Experiments and Prototypes to a Technical Solution. In Tommaso Caselli, et al., editors, *Computational Analysis of Storylines: Making Sense of Events*, Studies in Natural Language Processing, pages 240–259. Cambridge University Press, Cambridge, November.
- [4] Mann, W. C. and Thompson, S. A. (1987). *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.
- [5] Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.
- [6] Miltsakaki, E., Prasad, R., Joshi, A. K., and Webber, B. L. (2004). The penn discourse treebank. In *LREC*. Citeseer.
- [7] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The penn discourse treebank 2.0. In *LREC*. Citeseer.
- [8] Radev, D. (2000). A common theory of information fusion from multiple text sources step one: cross-document structure. In *1st SIGdial workshop on Discourse and dialogue*, pages 74–83.

## References (2)

- [9] Radev, D. R., Otterbacher, J., and Zhang, Z. (2004). CST Bank: A corpus for the study of cross-document structural relationships. In *Lrec*.
- [10] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.
- [11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.
- [12] He, P., Liu, X., Gao, J., and Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention.
- [13] Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, 06.
- [14] Ferrucci, D., Lally, A., Verspoor, K., and Nyberg, E., (2008). *Unstructured Information Management Architecture (UIMA) Version 1.0*. Oasis.

## Images

Slide 12 contains an image from [10]

Slide 19 has been designed using resources from Flaticon.com