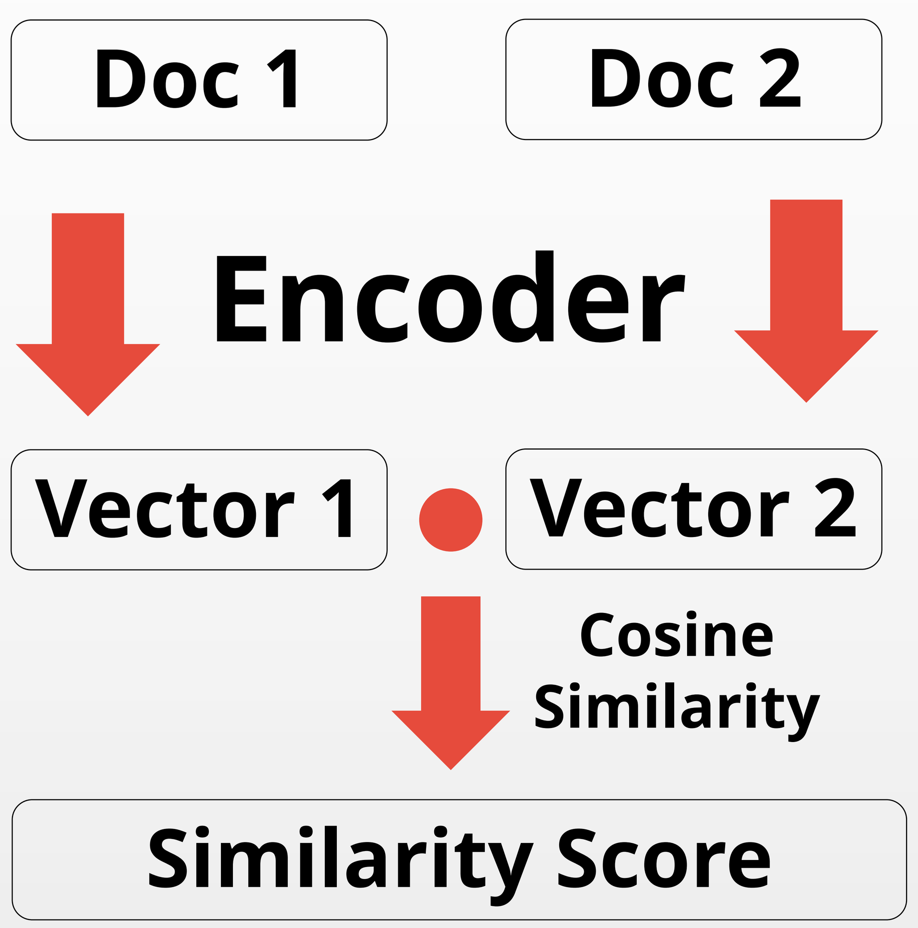
ExSiM: Explainable Methodology to Upgrade Sentence Similarity Metrics to Document-Level

Matthew “Hugh” C. Williams Jr., Shubhra “Santu” Karmaker

Common Paradigm for Metrics

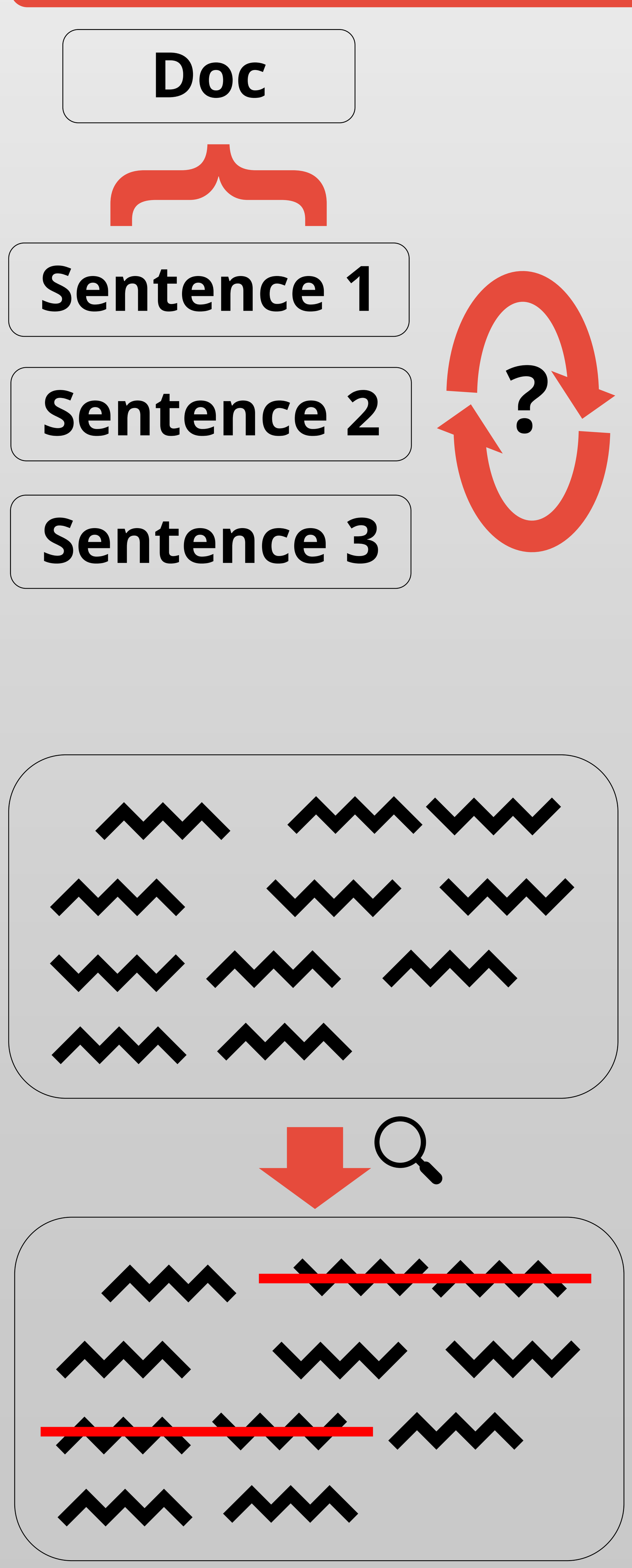


A singularly common way to determine document similarity is by using *encodings*.

Encoders: Deep-learning models that encode documents as vectors. For example, BERTScore.

Metric: Find the cosine similarity between two generated encodings.

Issues in Paradigm



Sentence Level: These encoders are trained for sentence level tasks. This means that there may be some properties of multi-sentence documents that are not properly considered.

- For example, consider how sentences can stand alone and have an ordering between them.

Inability to Account for Correctness: One can note that encoder paradigm is inherently commutative, as in it does not matter the order of the two documents. While this ought to be true in most cases, what about in situations where we know one document is correct and we are finding the similarity in the context of that?

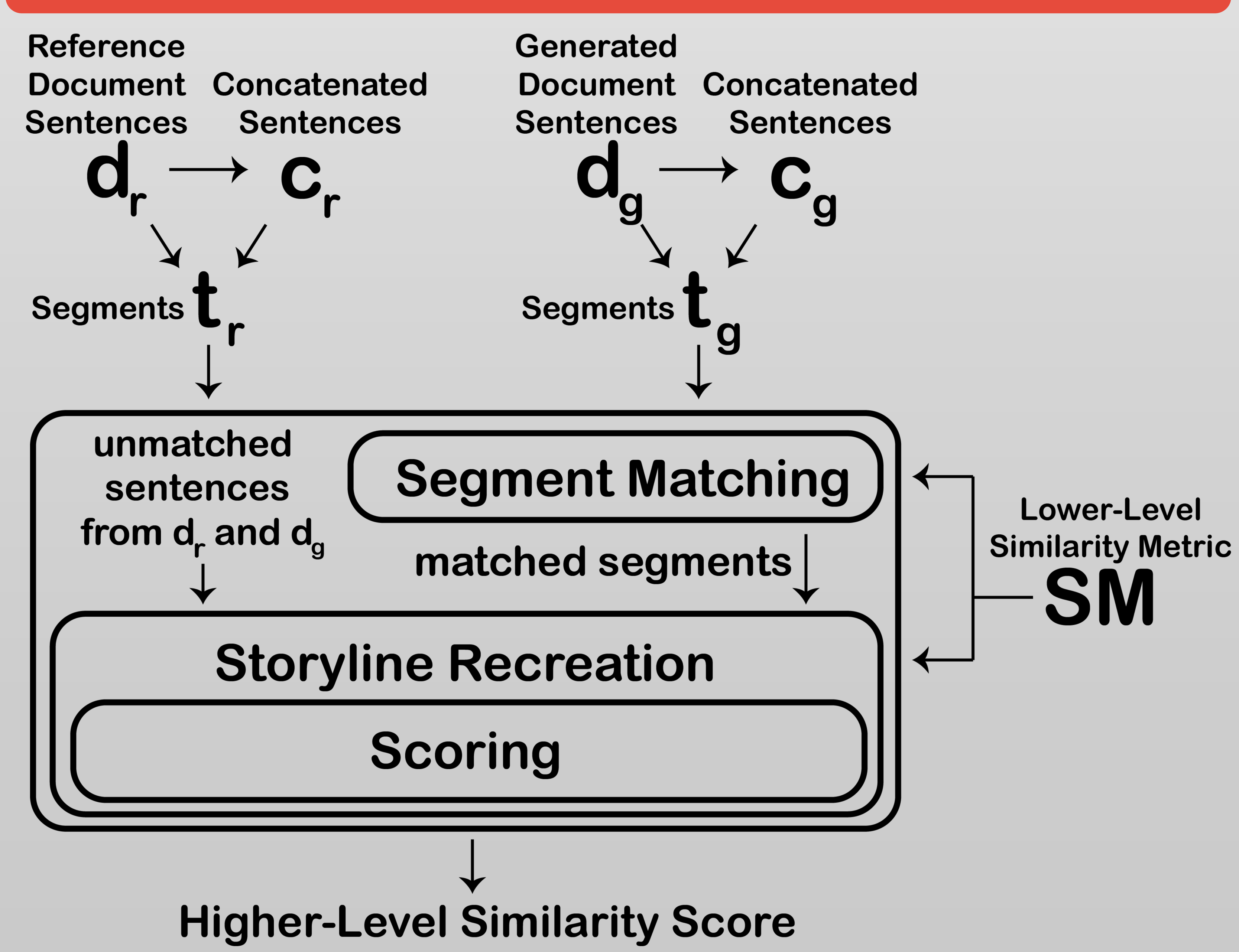
ExSiM's Solutions

To Sentence Level: ExSiM uses encoders, but only on sentence level where they are optimized. ExSiM itself, using an analytic methodology, converts these sentence level similarities into a document level one. It does by seeing how connections between adjacent sentences are preserved in the other document.

To Correctness: ExSiM can be non-commutative. This is chiefly important in cases where one document is denoted as correct. It can do this by mimicking how humans do comparisons: while reading one document, between sentences it goes back and looks for same idea in the other document. Note, this can be made commutative by averaging both non-commutative results.

Overall: ExSiM returns a *vector of metrics*, each element of which analyzes a different facet of the similarity of the two documents. One is sheer similarity while the rest are more novel.

ExSiM's Framework



Methodology: ExSiM uses simple algorithms to piece together results from a sentence similarity metric into a holistic and explainable document similarity metric.

Provisional Results

On Wikipedia Triplets Dataset

Given three articles *A*,*B*, and *C* such *metric(A,B)* > *metric(B,C)*, we can test a metric by seeing how accurately it preserves this inequality.

	Synthetic	Handpicked
MiniLM (Avg. SBERT)	77.1%	**94.0%**
BERTScore (Roberta)	76.0%	84.2%
ExSiM	**77.8%**	91.4%

On Human Annotated Dataset

Thanks to some Auburn students, we were able to rank a series of document pairings on how similar each pair was. Below is how well each metric correlates with human annotated similarities.

	Overall Similarity	Reordering Similarity
BERT	0.632	0.589
ExSiM	**0.768**	**0.8**
Commutative ExSiM	0.62	0.58

Qualitative Evaluation for Vector of Metrics

On the same dataset used above, we used ExSiM to compare a few sentence reordering models: BART, GPT-3.5, ReBART, and DistilBART.

- *Localized Storyline Similarity:* Each model showed slightly decreased performance towards end of generation.
- *Frequency of Splits and Fusions:* ReBART tended to fuse and not split, while rest were similar.
- *Coverage of Information, Information Preservation, Hallucination:* GPT-3.5 stood out well, its only error being that when it did produce extra sentences, albeit rarely, they were very hallucinatory.

References

1. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, & Yoav Artzi. (2020). BERTScore: Evaluating Text Generation with BERT.