Annotating and Modeling Fine-grained Factuality in Summarization

TAUR LAB @UT Austin

Tanya Goyal Greg Durrett

Synthetic training data for factuality fails to cover the range of errors made by summarization models.

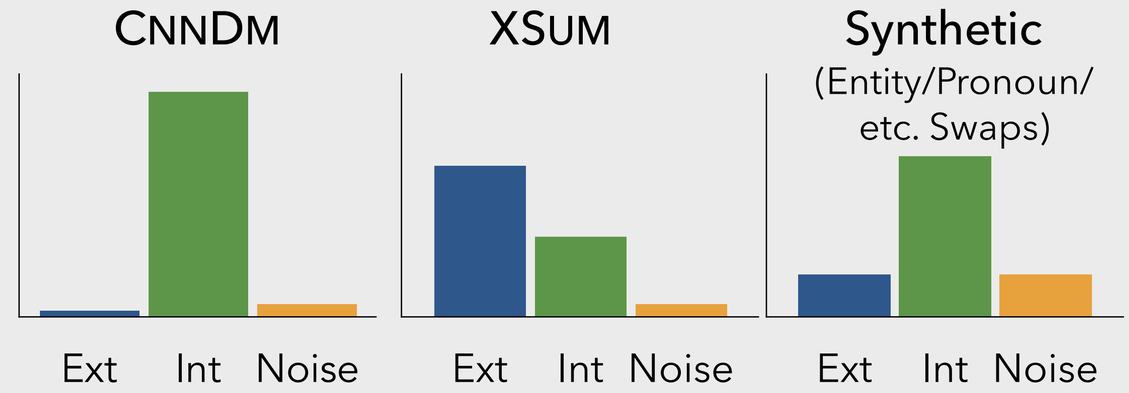
… he won two and lost three of his games …

Extrinsic Errors → … On Sunday, he won three of his games games…

Intrinsic Errors → … On Sunday, he won three of his games games…

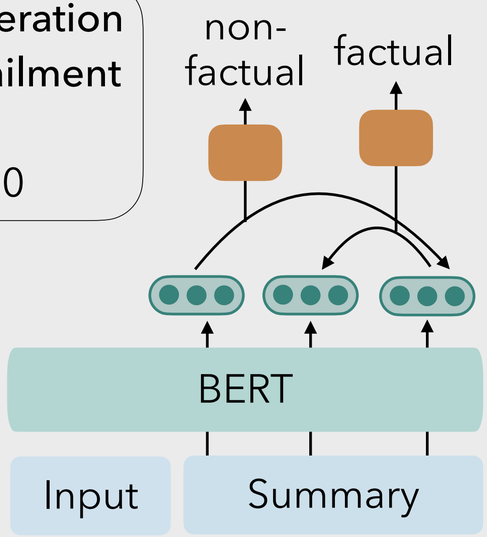
Noise → … On Sunday, he won three of his games games…

**more fine-grained error taxonomy in the paper



Fine-grained human annotation and modeling are needed to identify errors on tougher datasets!

Evaluating Factuality in Generation with Dependency-level Entailment
Goyal and Durrett,
Findings of EMNLP2020



Classification Accuracy Results

	Synthetic	Human
Majority	50.0	50.0
Sent-level	50.9	65.6
Dep-level	51.2	**78.7**

We use error localisation to train more factual summarization models on noisy data (e.g. XSUM).

	Avg. Factuality Score (Human)
Baseline	0.37
Ours	0.46