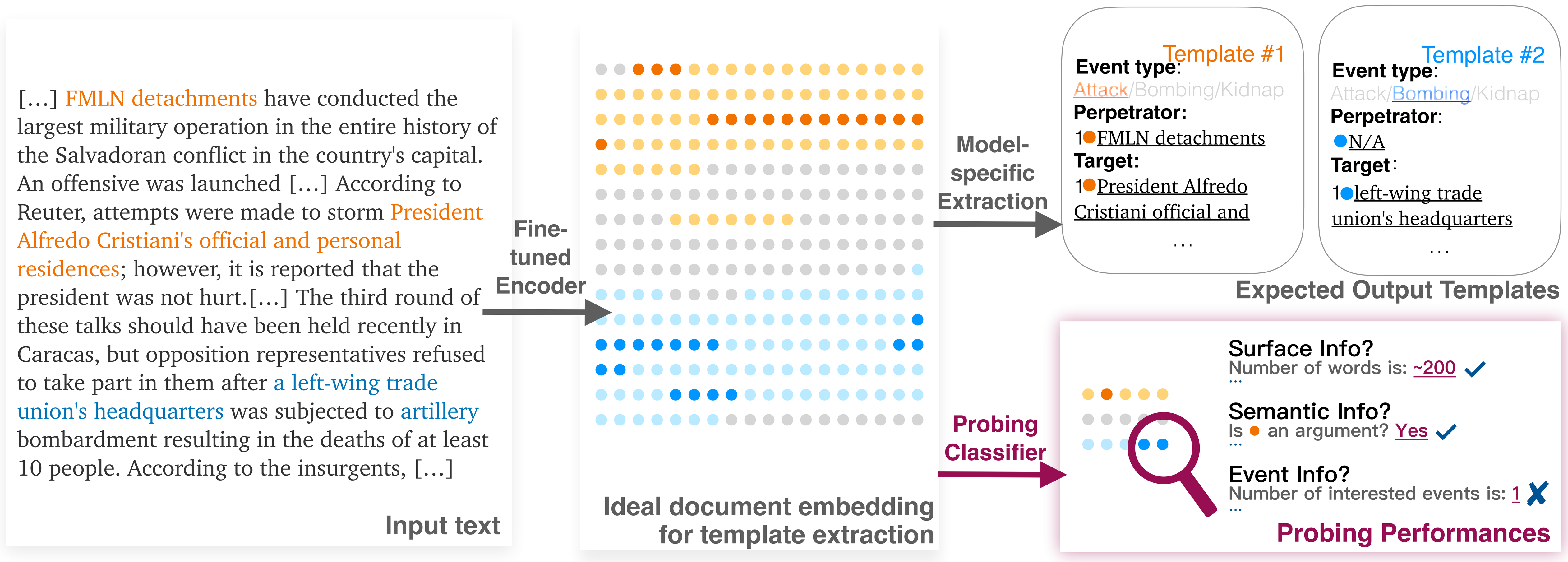
Probing Representations for Document-level Event Extraction

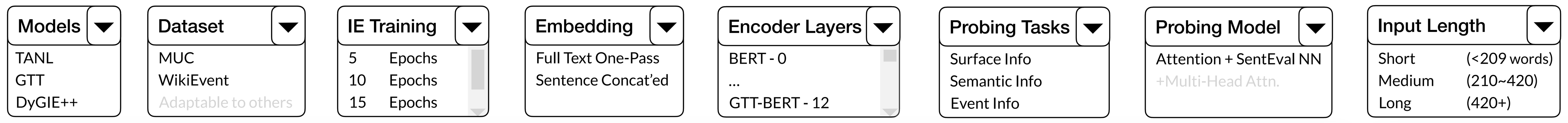
Cornell Bowers C·IS College of Computing and Information Science
Department of Computer Science, The University of Texas at Dallas

Barry Wang, Xinya Du, Claire Cardie

Document Embeddings and Document-Level Information Extraction



- **Document-level information extraction** is a task to extract structured "events" (or "templates") from unstructured input texts
- Modern systems for this task predominately encode inputs using a neural **encoder** (BERT, T5, etc), which is trained in IE task back prop
- We design **probing** tasks to understand **how encodings from the trained encoders make certain information more (or less) extractable**

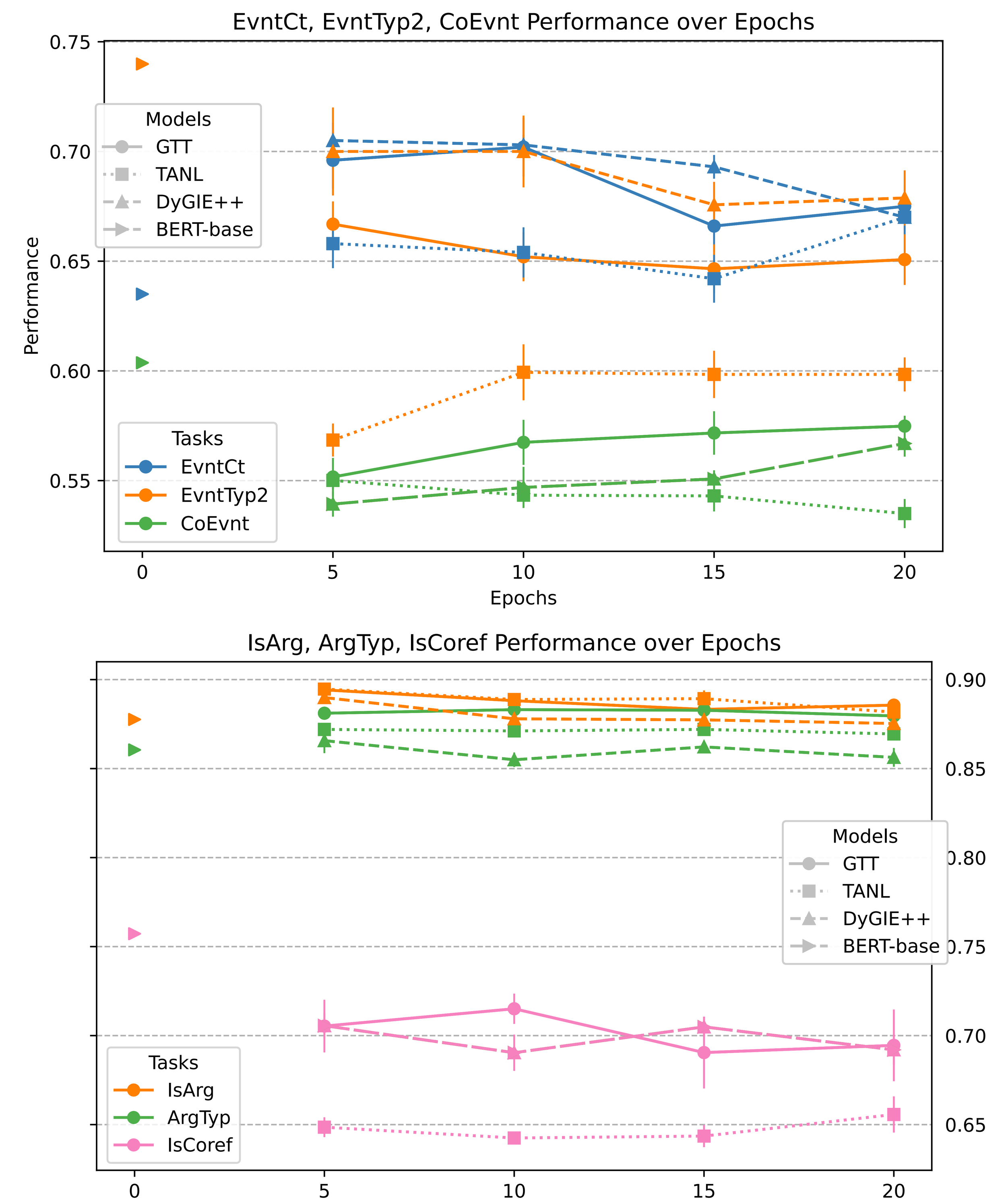


Probing Tasks

Category	Illustration	Task	Task Full Name
Surface	... -> #Words	WordCt	Word Count
	... -> #Sentences	SentCt	Sentence Count
Semantic	a.k.a. ?	Coref	Are Coreferent
	in Any ?	IsArg	Is an Argument
	-> Perpetrator? Victim? ...	ArgTyp	Argument Type
Event	-> Bombing/ Attack... ?	EvntTyp$_2$	Event Type
	both in Any ?	CoEvnt	Co-Event
	... -> #	EvntCt	Event Count

- 8 probing tasks across 3 levels on document embeddings
- Each non-surface probing task tests an encoding capability necessary for correct IE output. Surface results in appendix.
- We use probing model improved over previous works

Probing Performances over IE Framework Training Epochs



Probing accuracy on event (up) and semantic (down) info over document-level IE training epoch. 5 random seed results averaged (with std. error bars). Trained encoder gain and lose information in their generated embeddings as they are trained for the IE tasks.

Information Extraction Models

- **DyGIE++** (Wadden et al., 2019) is a discriminative multi-task framework. It achieves IE by enumerating and scoring sections (spans) of encoded text and using the relations between spans to detect triggers and events.
- **GTT** (Du et al., 2021) is a sequence-to-sequence event-extraction model that perform the task end-to-end, without using labeled triggers. It is trained to decode a serialized template, with tuned decoder constraints.
- **TANL** (Paolini et al., 2021) is a sequence-to-sequence multi-task model that "translates" input text to augmented languages. For IE, the in-text augmented parts identify triggers and roles. It uses a two-stage approach for event extraction by first extracting trigger then finding arguments for each trigger predicted.

Datasets

- **MUC 3/4:** Our document-level data source to create probing tasks, thanks to its rich coreference information. The dataset has 1300/200/200 training/validation/testing documents. Note that 44.6% of the inputs have no corresponding events. A keyword-based trigger was added to every template of the MUC dataset to make it compatible with TANL and DyGIE++.
- **WikiEvents** (Li et al., 2021) results are reported in appendix. It has 200/20/20 training/validation/testing inputs and has wider ranges of incident types.

Probing Performances with Different IE Frameworks and Embedding Method

Model (IE-F1)	Input	WordCt	SentCt	IsArg	ArgTyp	Coref	EvntTyp$_2$	CoEvnt	EvntCt	Avg
DyGIE++ (41.9)	FullText	58.6	47.0	87.1	83.8	64.7	**60.5**	73.6	67.2	67.8
	SentCat	57.4	58.9	87.5	85.6	69.2	56.7	67.9	67.0	68.8
GTT (49.0)	FullText	58.6	46.3	88.3	**88.5**	66.7	60.4	66.4	**68.3**	67.9
	SentCat	55.8	**58.9**	**88.6**	88.0	69.5	57.5	65.07	67.5	68.8
TANL (33.2)	FullText	54.2	43.3	88.2	86.8	66.6	57.8	60.0	65.8	65.3
	SentCat	34.3	40.8	88.2	87.0	65.6	53.5	59.8	67.0	62.0
BERT$_{base}$	FullText	**65.5**	45.0	87.8	86.1	**75.7**	60.4	**74.0**	63.5	69.7

Table 1: Probing Task Test Average Accuracy. IE frameworks trained for 20 epochs on MUC, and we run probing tasks on the input representations. We compare the 5-trial averaged test accuracy on full-text embeddings and concatenation of sentence embeddings from the same encoder to the untrained BERT baseline. IE-F1 refers to the model's F1 score on MUC test. Underlined data are the best in same embedding method, while bold, overall.

SentCat here refries to embedding sentences individually and then concatenating them. This can be more effective for IE tasks than using embeddings directly from a fine-tuned encoder designed for entire documents. This indicates encoders' poor capacity to capture of **discourse** information.

Trained encoders significantly enhance embeddings for **event detection** (higher accuracy in event count predictions (EvntCt↑)). Nevertheless, embeddings **lose information** for event typing (EvntTyp 2 ↓) and coreference (Coref↓).

Event Probing Performances over Different Input Lengths

Model	FullText Best	FullText Avg	Sent Best	Sent Avg
WordCount: ≤ 209				
DyGIE++	68.5	67.1	**69.7**	**68.8**
GTT	70.3	**68.7**	**72.1**	68.0
TANL	**71.8**	**70.2**	66.3	64.2
WordCount: 210-420				
DyGIE++	67.0	**65.7**	**67.6**	64.7
GTT	**67.6**	**67.0**	66.4	64.7
TANL	**64.8**	**62.0**	63.6	60.8
WordCount: ≥ 431				
DyGIE++	70.6	70.2	**74.2**	**72.1**
GTT	69.1	68.7	**71.5**	**70.2**
TANL	67.3	65.2	**69.7**	**68.3**

Table 2: EvntCt Probing Test Accuracy (%) 5 random seed averaged. When WordCount $\geq$431, both FullText and SentCat embeddings are truncated to the same length (e.g., BERT-base has a limit of 512) for comparison fairness. Concatenated sentence embeddings show an advantage on medium or long texts.

Probing Performances over BERT Layers

Probe acc. on event & semantic information over BERT layers as-is and from GTT trained over 18 epochs.

Takeaways

- Our work provides the first insights into document-level representations.
- Trained encoding improves on capabilities like event detection and argument labeling, but IE training compromises encoder's ability to encode coreference and event typing information.
- Current models marginally outperformed the baseline in capturing event information at best, uncovered by comparisons of IE frameworks.
- Encoder models struggle with document length and cross-sentence discourse, as concatenation of sentence embeddings yielded stronger probing performances.

EMNLP 2023
Findings, EMNLP 2023
See our Paper, Github Repo
Correspondence: Barry Wang