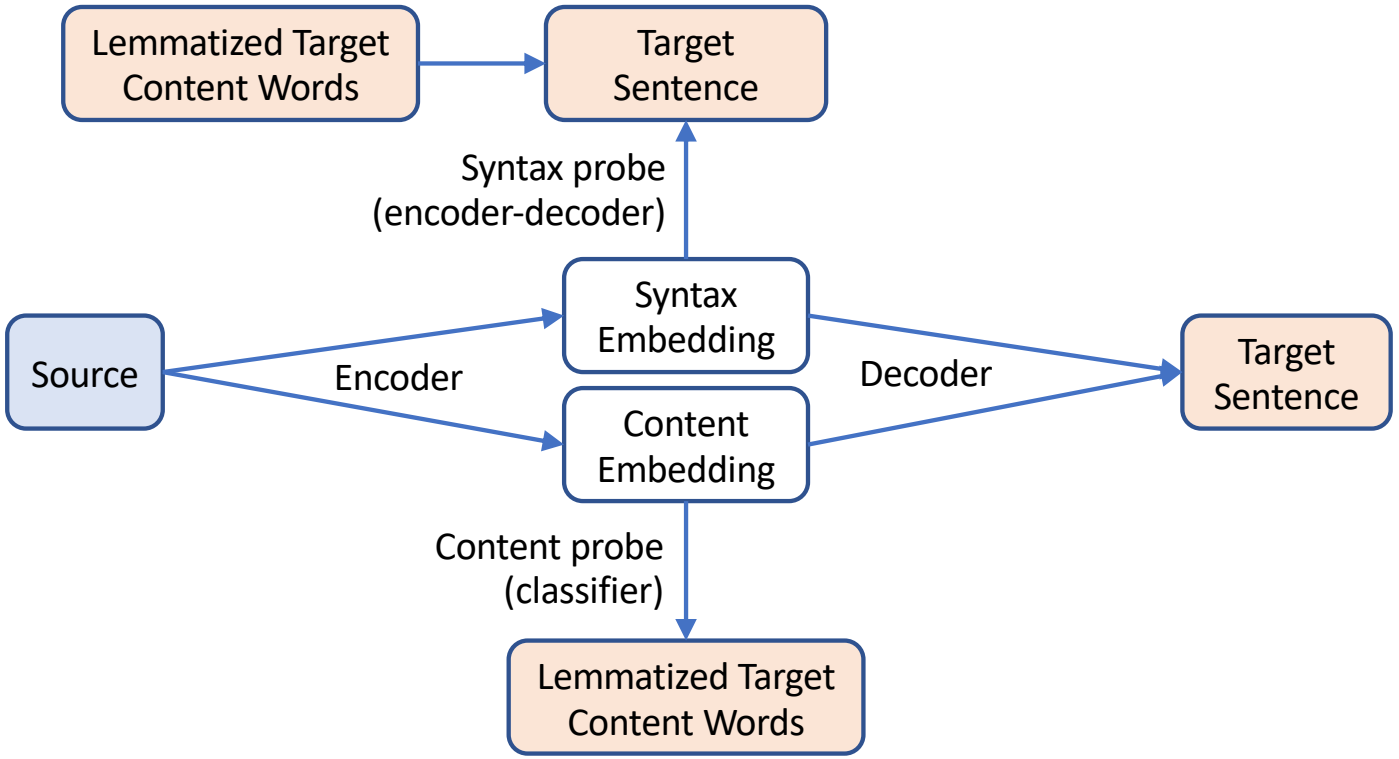
Scribendi – UW project

Plan

- Disentangled content-syntax embedding
  - Performance comparison
  - Error analysis
  - Potential papers
- Disentangled concepts-syntax embedding
- Disentangled knowledge graph-syntax embedding
- Timeline (Feb-April)

Disentangled Content-Syntax Embedding



Performance Comparison

- Scribendi dataset: 4M source-target pairs
- Public datasets: NUCLE, FCE, Lang-8, JFLEG, W&I, LOCNESS

Models	BLEU	GLUE	M^2	ERRANT
Scribendi baseline				
Content-Syntax baseline				
Content-Syntax w probes				
Content-Syntax w probes & attention				

- Model variants:
  - RNN vs transformer vs informer (non-autoregressive n log n transformer)

Component Error Analysis

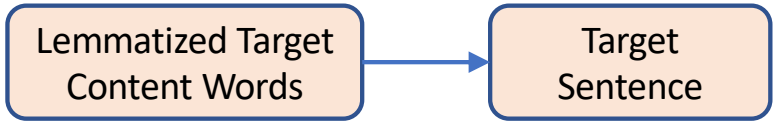
Models	Content probe accuracy	Syntax probe accuracy	Decoder accuracy	
	% correct lemmatized content words	BLEU	% correct lemmatized content words	BLEU
Content-Syntax w probes				
Content-Syntax w probes & attention				

Potential paper #1

- Disentangled content-syntax embedding
- EMNLP (Abstract deadline May 10)
- Contributions:
  - Improved GEC performance
  - New model that disentangles content and syntax in a self-supervised fashion
  - Automated component error analysis

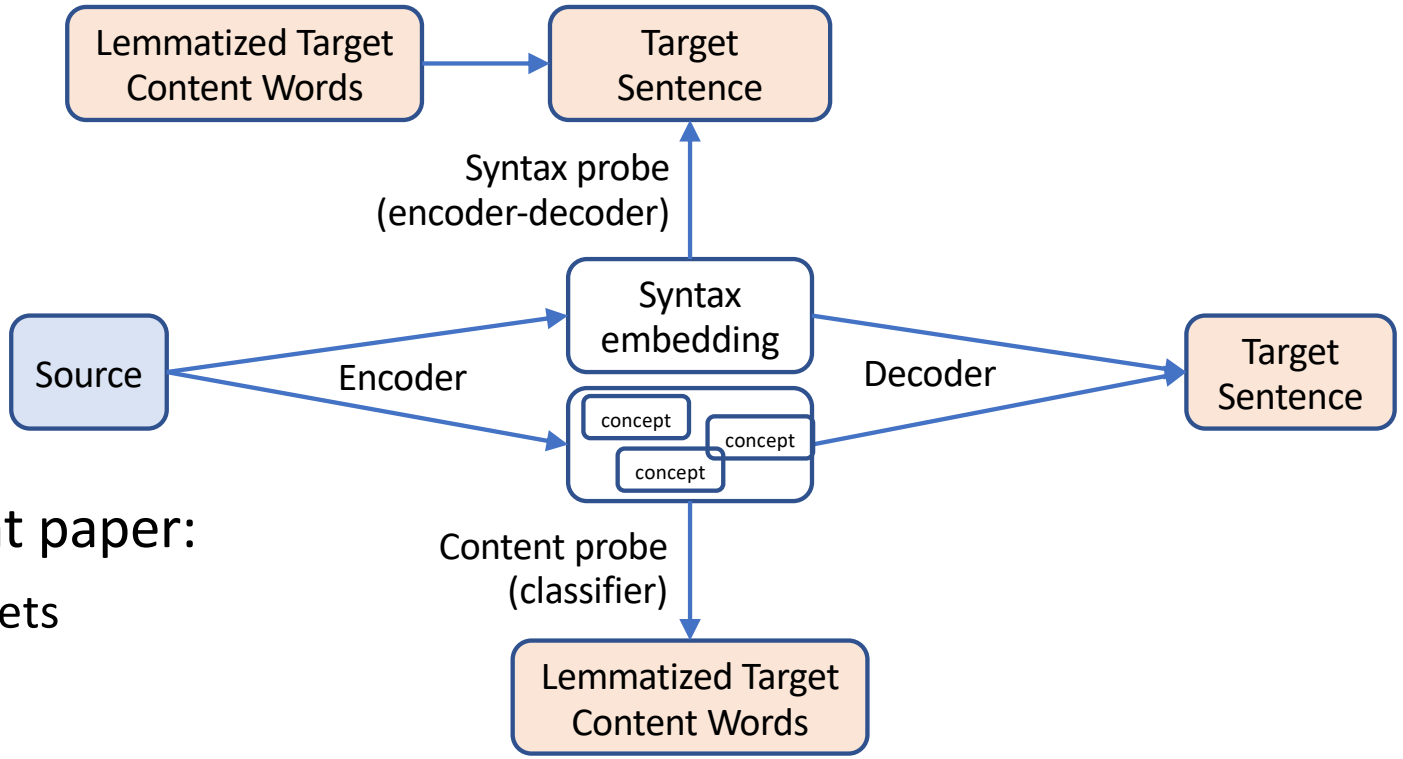
Potential paper #2

- Sentence generation from lemmatized keywords



- EMNLP (Abstract deadline May 10)
- Contributions:
  - Improved sentence generation
  - New model with pointer/copy network

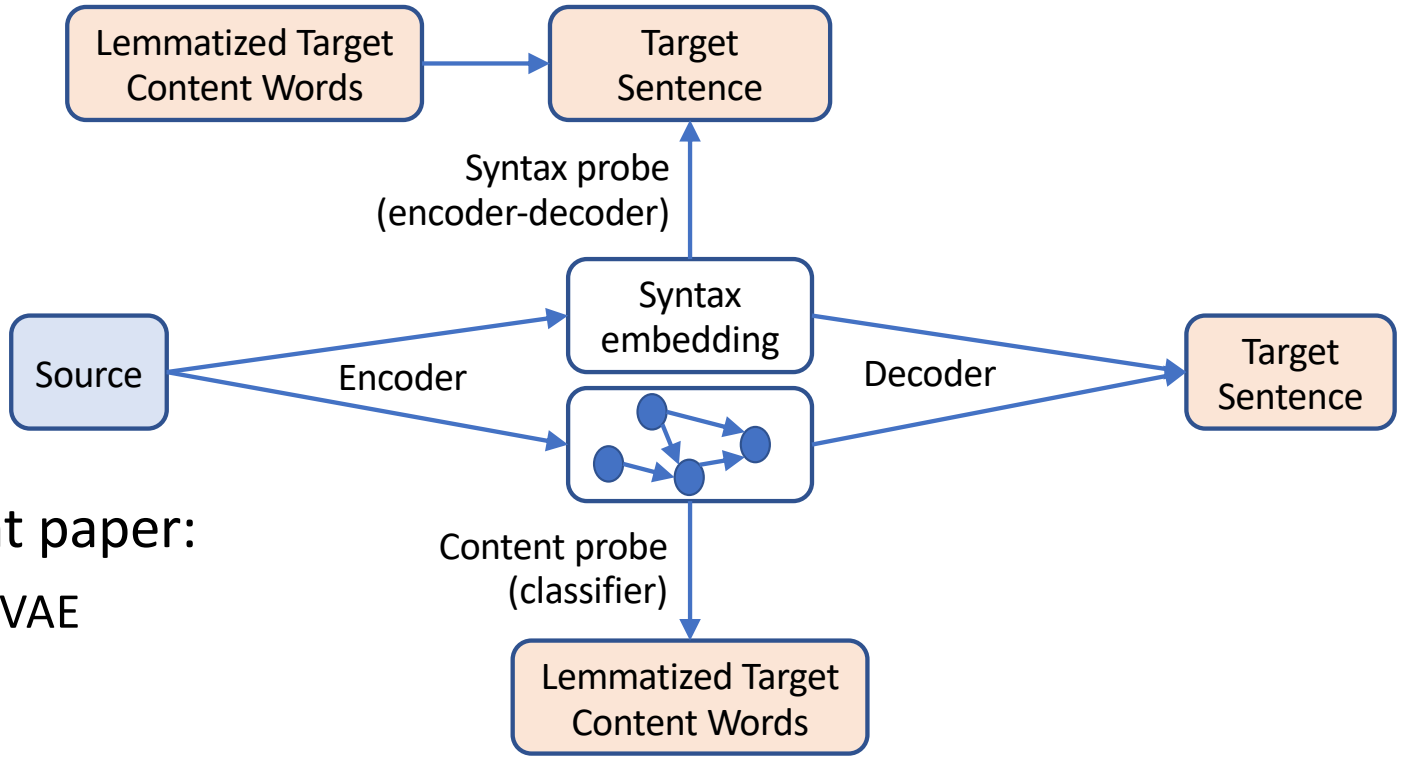
Disentangled Concepts-Syntax Embedding



Relevant paper:

- Deep Sets

Disentangled Knowledge Graph-Syntax Embedding



Relevant paper:

- Causal VAE

Timeline

- February: Disentangled content-syntax embedding
  - Performance comparison
  - Component error analysis
- March: Disentangled concepts-Syntax embedding
  - Model implementation
- April: Disentangled knowledge-graph syntax embedding
  - Model implementation
  - Paper submission EMNLP (Abstract deadline May 10)