Unsupervised Neural Text Simplification

Sai Surya[†] **Abhijit Mishra**[‡] **Anirban Laha**[‡] **Parag Jain**[‡] **Karthik Sankaranarayanan**[‡]
[†]IIT Kharagpur, India [‡]IBM Research
subramanyamdvss@gmail.com
{abhijimi,anirlaha,pajain34,kartsank}@in.ibm.com

2020/02/17 Paper reading
Slides made by Song

Overview

- Introduction
- Related work
- Motivation
- Proposed method
- Experiment setting
- Result
- Conclusion

Introduction

- Text Simplification (TS):

Original Text

Amnesty International accused the U.S. authorities of providing an "inhuman" treatment to Bradley Manning, a soldier accused of leaking "wires" of American diplomacy to the website Wikileaks.

Adapted Text (by trained editor)

United States treats a soldier in prison very badly.

The soldier is called Bradley Manning.

Bradley Manning is in prison for giving information about the Government of the United States to Wikileaks.

Wikileaks is a website which provides information on matters of public interest.

Simplify aspects:

1) Lexical: Complex words
2) Syntactic: Hierarchical structures

Related work

- 1. Rule based: Chandrasekar and Srinivas (1997)
- 2. Modular system: Canning and Tait (1999)
- 3. Data-driven:
    - Statistical: Phrase-based SMT (Specia, 2010; Stajner et al., 2015)
    - Neural: (Wang et al., 2016; Nisioi et al., 2017)

Motivation

Supervised VS Unsupervised

Method	Data	Limitation (Solution)
Supervised Neural TS (NTS) (Wang et al., 2016; Nisioi et al., 2017)	Parallel Original-Simplified data	1) No (less) parallel data for new language 2) Noise in current parallel dataset
Unsupervised Neural TS (UNTS)	Complex sentence dataset + Simple sentence dataset	1) Do not need original-simplified parallel data

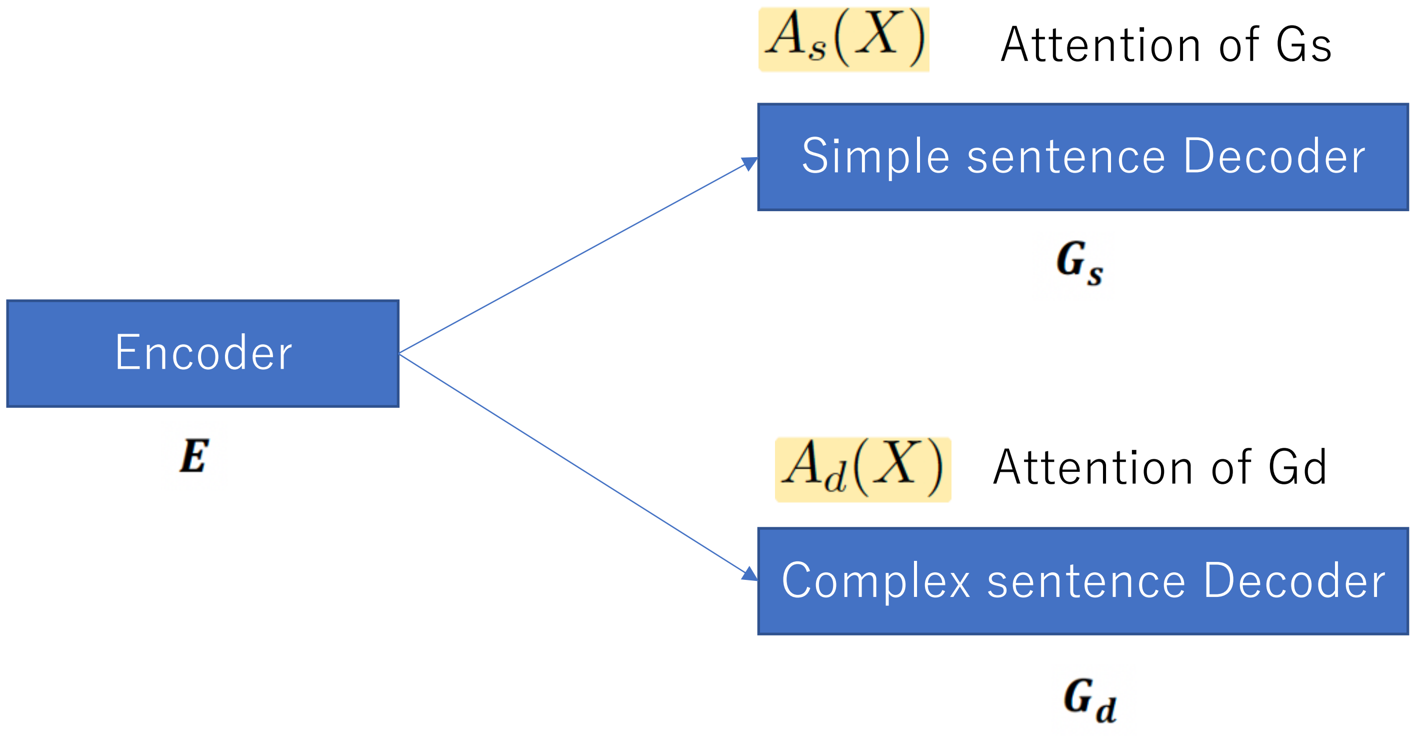
Motivation

Previous unsupervised method VS proposed

Method	Limitation (Solution)
Unsupervised lexical simplification (Paetzold and Specia, 2016)	Only complex → simple words No syntactic level simplification.
UNTS	Both lexical and syntactic level.

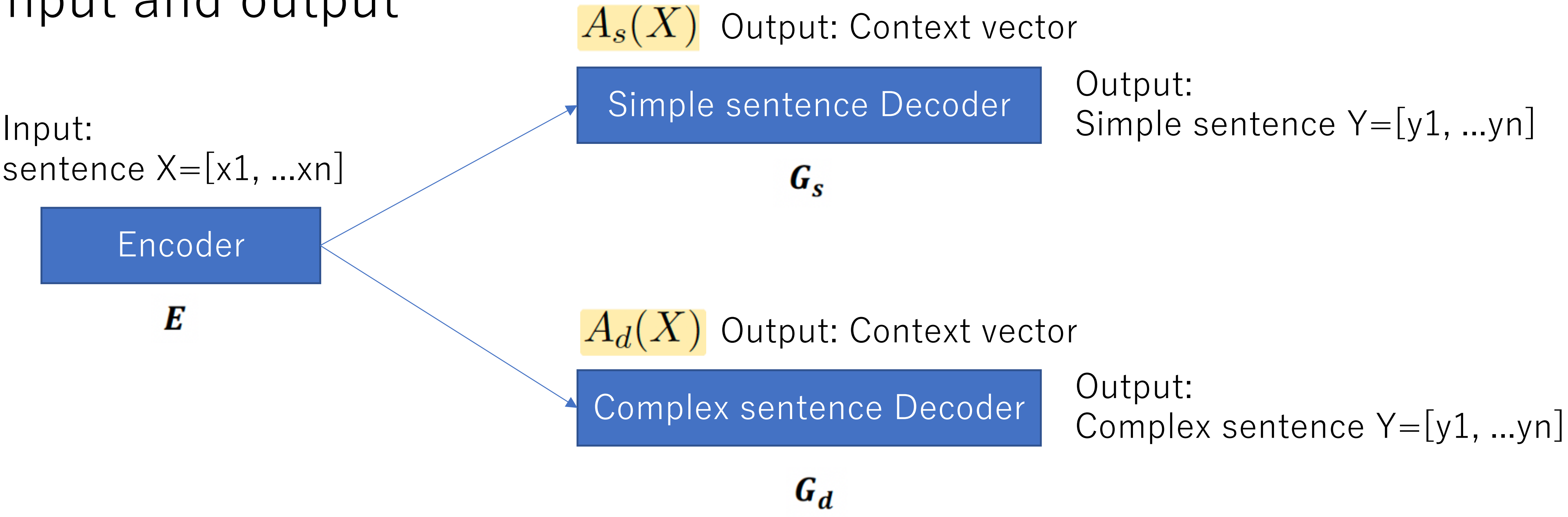
Proposed method

- Encode-attend-decode architecture



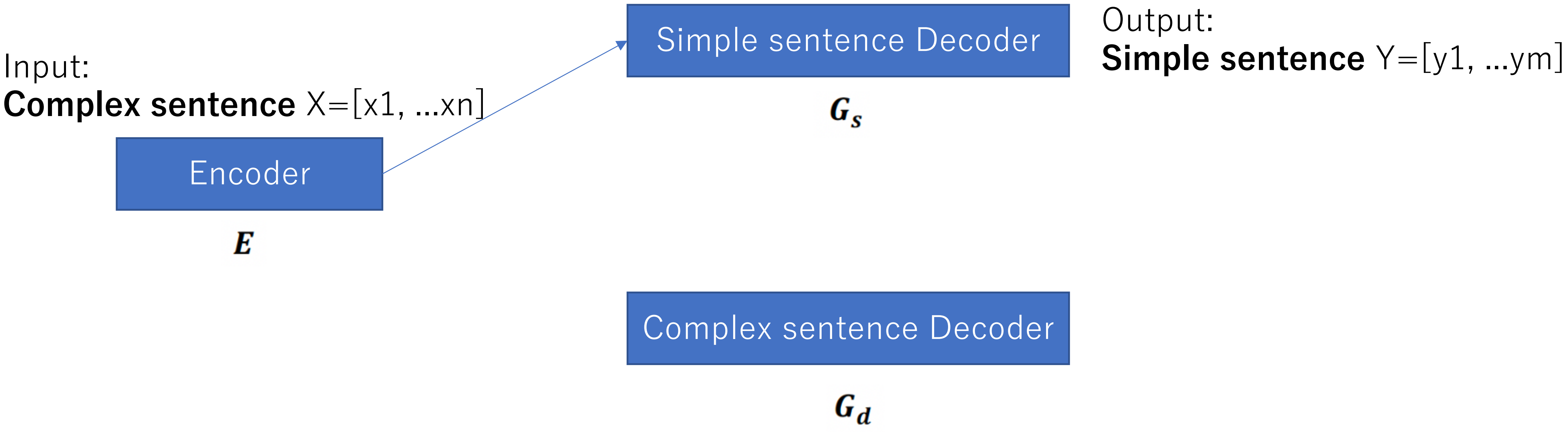
Proposed method

- Input and output



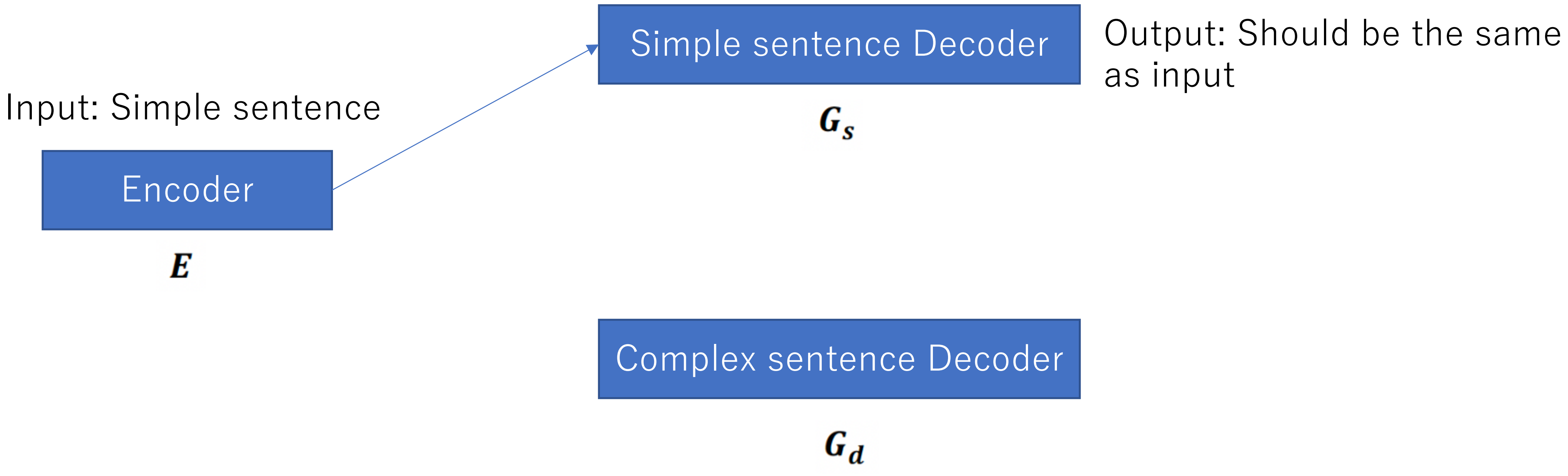
Proposed method

- Final goal:



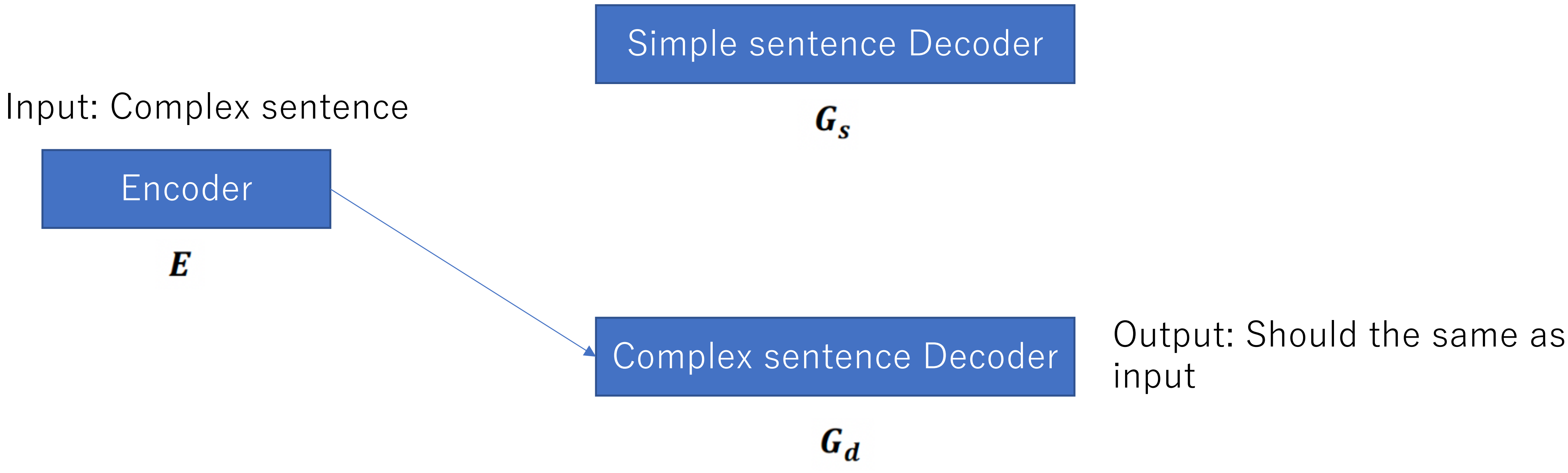
Proposed method

- Training goal (task):   (1 Reconstruction



Proposed method

- Training goal: (1 Reconstruction

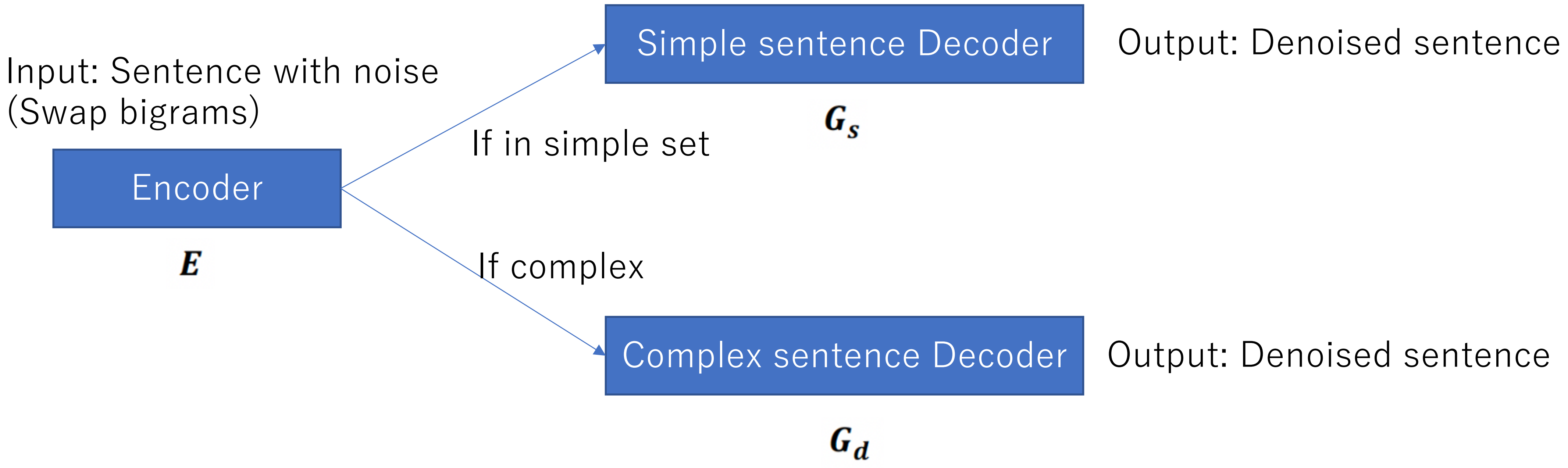


Loss:

$$\mathcal{L}_{rec}(\boldsymbol{\theta_E}, \boldsymbol{\theta_{G_s}}, \boldsymbol{\theta_{G_d}}) = -\mathbb{E}_{X_s \sim \boldsymbol{\mathcal{S}}}[\log P_{\boldsymbol{E}-\boldsymbol{G_s}}(X_s)] - \mathbb{E}_{X_d \sim \boldsymbol{\mathcal{D}}}[\log P_{\boldsymbol{E}-\boldsymbol{G_d}}(X_d)]$$

Proposed method

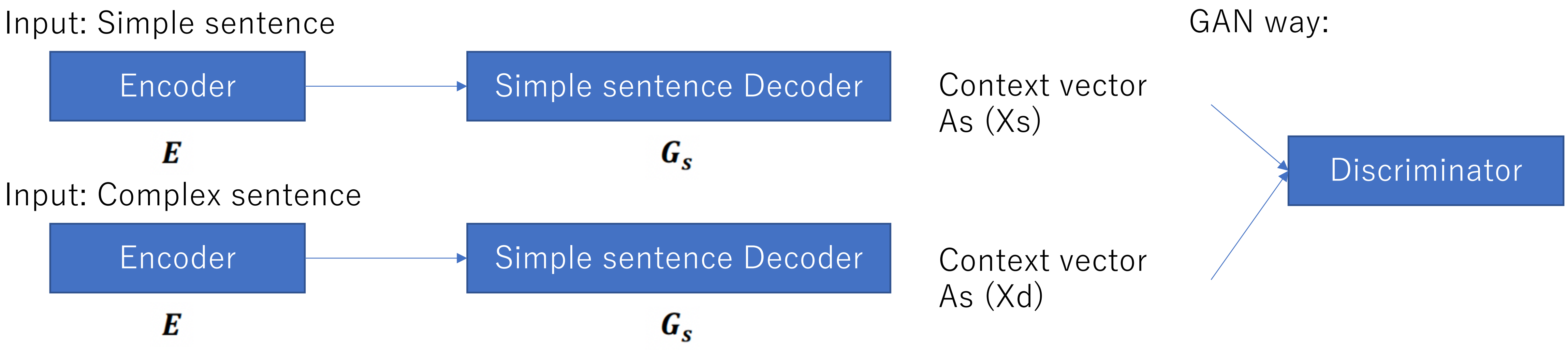
- Training goal: (2 Denoise



$$\mathcal{L}_{denoi} = -\mathbb{E}_{X_s \sim \boldsymbol{S}}[\log P_{\boldsymbol{E}-\boldsymbol{G_s}}(X_s|noise(X_s))] - \mathbb{E}_{X_d \sim \boldsymbol{D}}[\log P_{\boldsymbol{E}-\boldsymbol{G_d}}(X_d|noise(X_d))]$$

Proposed method

- Training goal:
  - (3 Simple decoder can process complex sentence



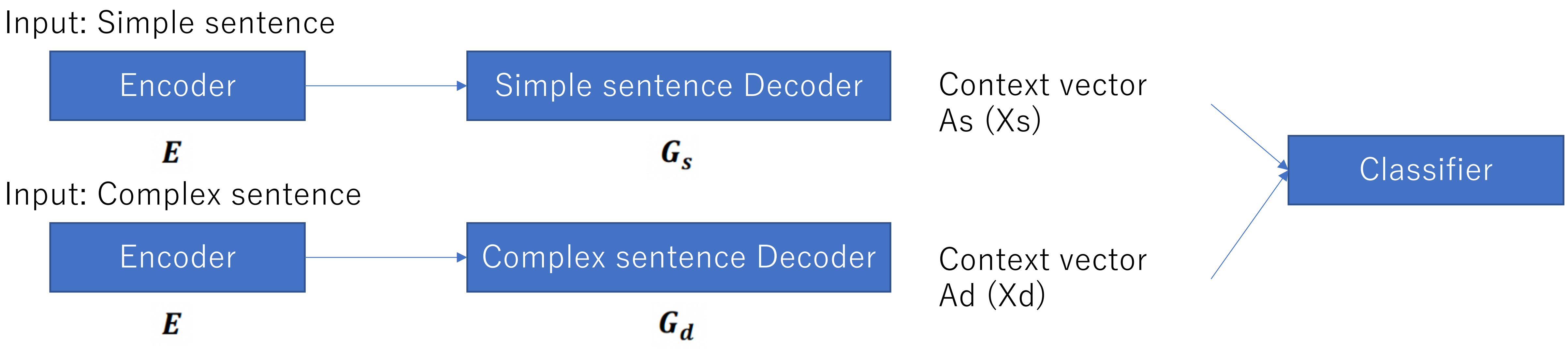
Make As(Xs) and As(Xd) with similar data distribution

$$\mathcal{L}_{adv,D}(\boldsymbol{\theta_D}) = -\mathbb{E}_{X_s\sim\boldsymbol{\mathcal{S}}}[\log(\boldsymbol{D}(A_s(X_s)))] - \mathbb{E}_{X_d\sim\boldsymbol{\mathcal{D}}}[\log(1-\boldsymbol{D}(A_s(X_d)))] \quad (3)$$

$$\mathcal{L}_{adv,G_s}(\boldsymbol{\theta_E},\boldsymbol{\theta_{G_s}}) = -\mathbb{E}_{X_d\sim\boldsymbol{\mathcal{D}}}[\log(\boldsymbol{D}(A_s(X_d)))] \quad (4)$$

Proposed method

- Training goal:
  - (4 Context vector of simple sentence should be different from that of complex sentence



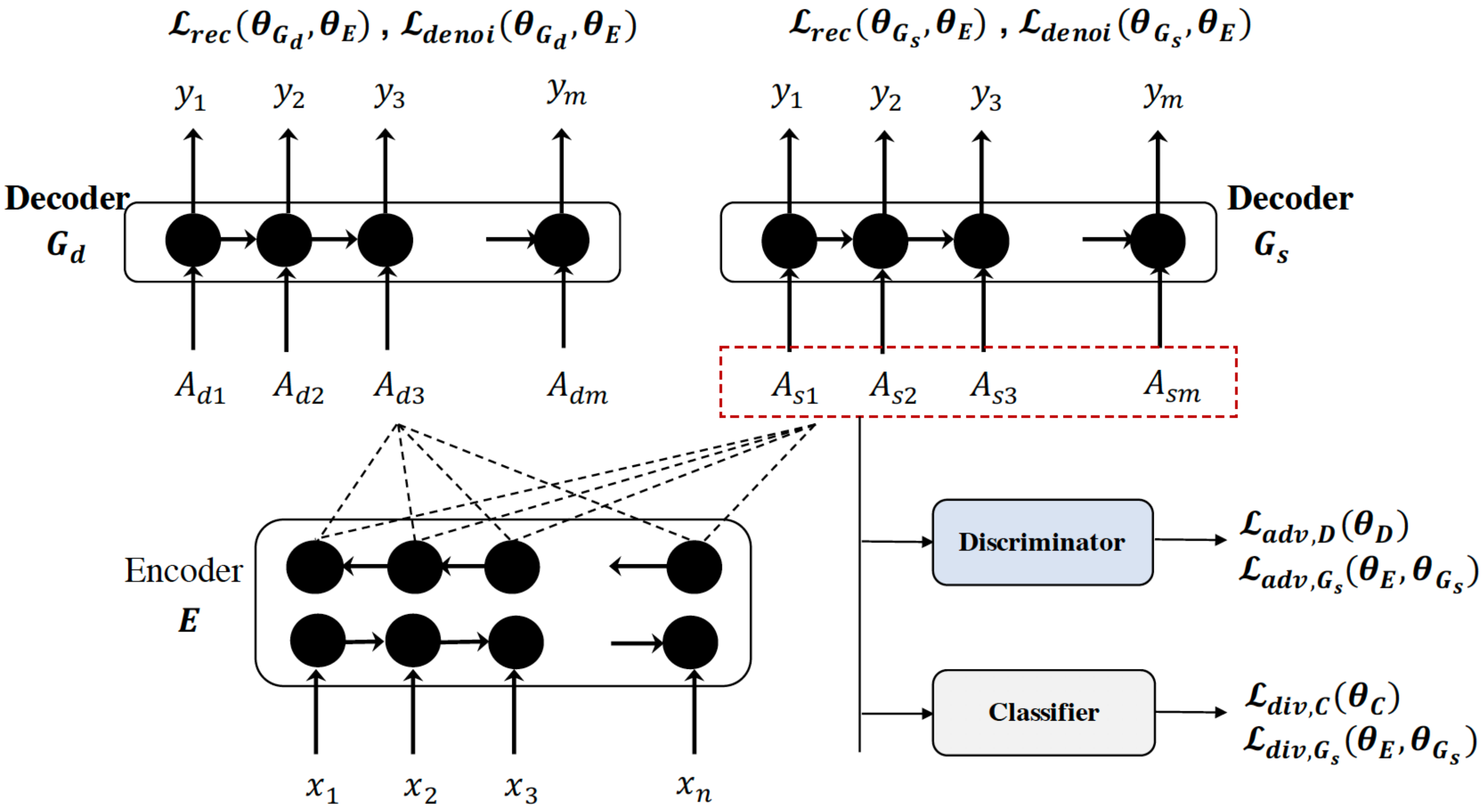
Make As(Xs) and Ad(Xd) with different data distribution

$$\mathcal{L}_{div,C}(\boldsymbol{\theta_C}) = -\mathbb{E}_{X_s\sim\boldsymbol{\mathcal{S}}}[\log(\boldsymbol{C}(A_s(X_s)))] - \mathbb{E}_{X_d\sim\boldsymbol{\mathcal{D}}}[\log(1-\boldsymbol{C}(A_d(X_d)))] \quad (5)$$

$$\mathcal{L}_{div,G_s}(\boldsymbol{\theta_E},\boldsymbol{\theta_{G_s}}) = -\mathbb{E}_{X_d\sim\boldsymbol{\mathcal{D}}}[\log(\boldsymbol{C}(A_d(X_d)))] \quad (6)$$

Proposed method

- Graph from the paper



Training process

Algorithm 1 Unsupervised simplification algorithm using denoising, reconstruction, adversarial and diversification losses.

Input: simple dataset $\boldsymbol{\mathcal{S}}$, complex dataset $\boldsymbol{\mathcal{D}}$.

Initialization phase:

repeat

Update $\boldsymbol{\theta_E}, \boldsymbol{\theta_{G_s}}, \boldsymbol{\theta_{G_d}}$ using $\mathcal{L}_{denoi}$

Update $\boldsymbol{\theta_E}, \boldsymbol{\theta_{G_s}}, \boldsymbol{\theta_{G_d}}$ using $\mathcal{L}_{rec}$

Update $\boldsymbol{\theta_D}, \boldsymbol{\theta_C}$ using $\mathcal{L}_{adv,D}$ $\mathcal{L}_{div,C}$

until specified number of steps are completed

Adversarial phase:

repeat

Update $\boldsymbol{\theta_E}, \boldsymbol{\theta_{G_s}}, \boldsymbol{\theta_{G_d}}$ using $\mathcal{L}_{denoi}$

Update $\boldsymbol{\theta_E}, \boldsymbol{\theta_{G_s}}, \boldsymbol{\theta_{G_d}}$ using $\mathcal{L}_{adv,G_s}$, $\mathcal{L}_{div,G_s}$, $\mathcal{L}_{rec}$

Update $\boldsymbol{\theta_D}, \boldsymbol{\theta_C}$ using $\mathcal{L}_{adv,D}$, $\mathcal{L}_{div,C}$

until specified number of steps are completed

Experiment setting

Dataset

	Dataset	Description
Train	en-wikipedia	Readability score: categorize complex simple sentences
Dev (Xu et al., 2016)	2000 sentences	each with 8 references
Test (Xu et al., 2016)	359 sentences	each with 8 references

Experiment setting

Evaluation

	Metric
Model selection	Automatic metric such as: BLEU and Word Difference
Test set evaluation	Human evaluation: 1) Simpleness: [0-1] 2) Grammaticality: fluency [1-5] 3) Relatedness: meaning [1-5] (higher->better)

Results

X% sentences are simplified

The higher, the better

	System	Simpleness	Fluency	Relatedness
Proposed semi-supervised	UNTS+10K	**57%**	4.13	3.93
Proposed Unsupervised	UNTS	47%	3.86	3.73
Unsupervised NMT	UNMT	40%	3.8	**4.06**
Supervised Neural TS	NTS	49%	4.13	3.26
Syntax based MT	SBMT	53%	**4.26**	**4.06**
Phrase-based SMT	PBSMT	53%	3.8	3.93
Unsupervised lexical simplification system	LIGHTLS	6%	4.2	3.33

Table 3: Average human evaluation scores for simpleness and grammatical correctness (fluency) and semantic relatedness between the output and input.

Conclusion:

- **Novel attempt** towards unsupervised text simplification.
- (Including) Novel model and training process.
- Better than or competitive to these baselines