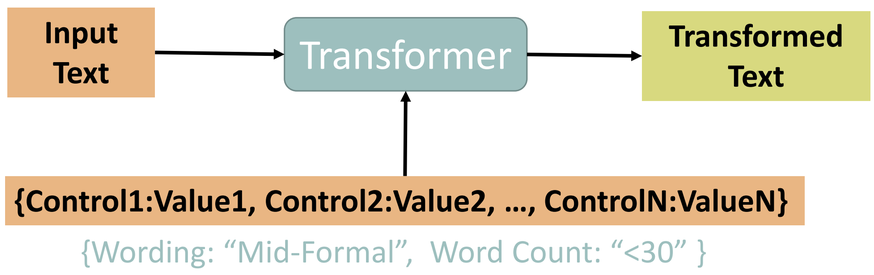
Unsupervised Controllable Text Formalization

Parag Jain, Abhijit Mishra, Amar Prakash Azad, Karthik Sankaranarayanan
{pajain34, abhijimi, amarazad, kartsank}@in.ibm.com, IBM Research, India

Task: Natural Language Transformation



- We experiment on transforming a given text to a more formal style

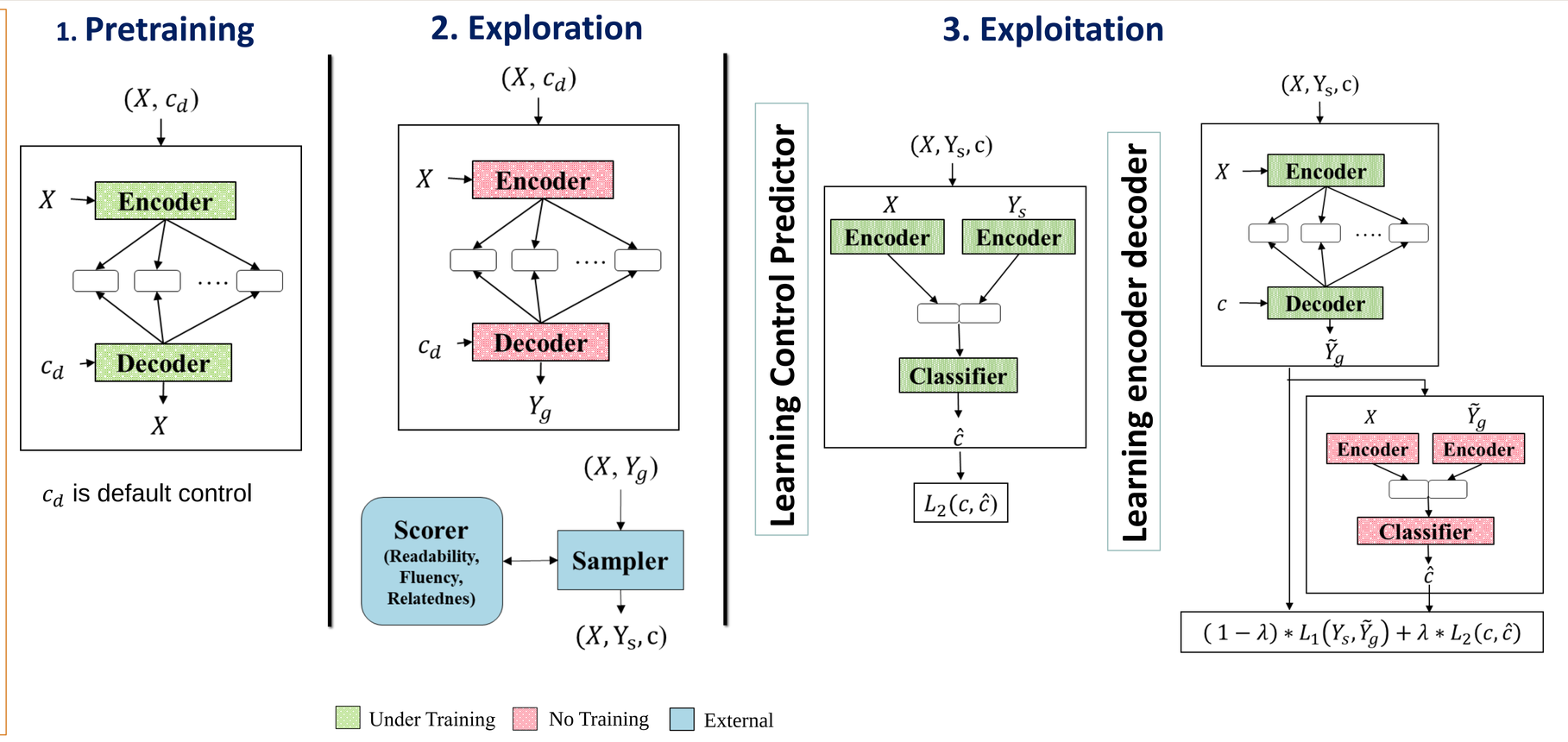
Overview

- Framework for controllable natural language transformation.
- Features:
  - unsupervised training scheme, handles infeasibility to annotate data for each <*input*, *output*, *controls*> triples
  - preservation of language semantics
  - use Off-the-self NLP modules for verification and scoring
  - control the degree of the intended attribute desired at the output.
  - learning to incorporate multiple control inputs (which can be dependent)

System Description

Training scheme

1. **Pretraining**: as an autoencoder for better initialization
2. **Exploration**:
   a) Synthesize instances of input text, output text and appropriate control <X, Y, c>
   b) Sample k outputs which maximizes the cumulative score
3. **Exploitation**:
   a) Train control predictor based on sampled $< X, Y, c >$ instances
   b) Train Enc-Dec using sampled $< X, Y, c >$ and control predictor



Sampling and Score calculation

Sampling:

- Sample k sentences which maximizes the cumulative score

$$Y_s = argmax_Y \left\{ G(X,Y)\ Y \in \left\{ Y_g, Sample_k\ (Y_g) \right\} \right\}$$

Cumulative score calculation:

- Cumulative score for the sampled sentence is generated as,

$$G(X,Y) = \beta_s r_s(X,Y) + \beta_f r_f(X,Y) + \beta_d r_d(X,Y)$$

$r_s(X,Y)$ is document similarity
$r_f(Y)\ is\ fluency$
$r_d\ (Y)\ is\ Flesch - Kincaid\ readability$

Control Determination:

- Control value (c) for the newly sampled example is determined as,

$$c = \begin{cases} 1, \text{if } c_r < \rho_1 \\ 2, \text{if } \rho_1 < c_r < \rho_2 \\ 3, \text{if } c_r > \rho_2 \end{cases} \qquad c_r = {}^{r_d(Y_s)}/_{r_d(X)}$$

Example input and transformed sentences

Mode – Ctrl	Input sentence	Input sentence
	18 year old who abandoned her child in a hospital later got custody	the first sync after upgrading will be slow
WithPred - Mid	18 year old who unpopulated her kid in a infirmary resultant got custody	the first synchronize afterward upgrading will be idle
WithPred - High	18 year old who deserted her tyke in a infirmary resultant got detention	the first synchronize afterward upgrading will be laggard
NoPred - Mid	18 class old who deserted her child in a infirmary accompanying got detention	the introductory synchronize afterward upgrading will be goosy
NoPred - High	18 class old who deserted her child in a infirmary accompanying got detention	the introductory synchronize afterward upgrading will be goosy
OneShot - Mid	18 yr old who untenanted her tyke in a hospital subsequently got detention	the eightieth sync later upgrading bequeath be tedious
OneShot - High	18 class old who deserted her tyke in a hospital subsequently got detention	the eightieth sync later upgrading bequeath be tedious

Dataset and Results

Curated Dataset [Code & data: https://github.com/parajain/uctf]

- 14432 sentences which are simple and informal in nature

Comparison with Existing Unsupervised Method

Mode	CTRL WITHPREDICTOR		CTRL NOPREDICTOR		CTRL ONESHOT		Mueller et al., 2017
Formalness Control	Mid	High	Mid	High	Mid	High	None
Readability	0.568	0.583	0.538	0.538	0.554	0.554	0.33
Relatedness	0.72	0.74	0.77	0.77	0.78	0.78	0.05
LM Score	0.34	0.34	0.32	0.32	0.30	0.30	0.16

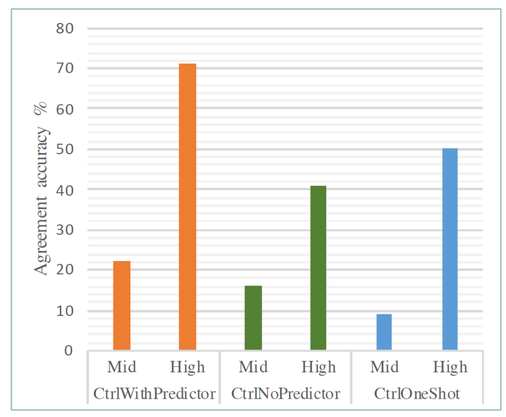
Average test-set scores (normalized between [0 - 1])

- Average readability grade of the input (0.54)

Human Evaluation

- Rank readability of different control outputs
- 80.2% agreement between human rated rank labels and ranking based on output control value

Output control agreement accuracy



Indirect Comparison with Supervised Systems

- Reversed-simplification Task
- Average readability 0.50

System	BLEU	Relatedness	Readability
Mueller et al.	5.09	0.41	0.38
Seq2Seq (skyline)	38.37	0.17	0.71
Formalness-Mid (Ours)	21.81	0.58	0.52
Formalness-High (Ours)	21.14	0.57	0.74

References

- Mueller, J.; Gifford, D.; and Jaakkola, T. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In ICML, 2017
- Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In ICML, 2017.
- Nisioi, S.; Stajner, S.; Ponzetto, S. P.; and Dinu, L. P. 2017. Exploring neural text simplification models. In ACL, 2017.