Improvement of Attention Based Formality Style Transfer Model with In-domain Data Augmentation

COSE461 Natural Language Processing
Final Presentation

통계학과 2019150419 기다연
컴퓨터학과 2019320062 이지수

Table of Contents

INTRODUCTION

1.1 Formality Style Transfer

Formality Style Transfer (FST)?

: conversion of a formal piece of text into an informal piece of text, vice versa

Related Work

❶ Zhang et al., (2020)

: Novel data augmentation methods for FST based on Seq2Seq model architecture

1. Back translation (French)
2. Formality Discrimination
3. Multi-task transfer

❷ Etinger and Black (2019)

: Highlight importance of in-domain training for FST models

INTRODUCTION

1.2 Problems of Related Methods

Problems

1. Inadequacy of training data
2. Lack of parallel corpora
3. Importance of In-domain training

In-domain Data Augmentation

1. Synonym Replacement (SR)
2. Round trip translation (RTT)

DATA

2.1 GYFAC Data

GYFAC? Grammarly's Yahoo Answers Formality Corpus

		Informal to Formal		*Formal to Informal*	
	Train	Tune	Test	Tune	Test
E&M	52,595	2,877	1,416	2,356	1,082
F&R	51,967	2,788	1,332	2,247	1,019

GYFAC dataset Statistics

Informal:	*I'd say it is punk though.*
Formal:	*However, I do believe it to be punk.*
Informal:	*Gotta see both sides of the story.*
Formal:	*You have to consider both sides of the story.*

Examples of GYFAC dataset

❶ Largest parallel corpus of human-labeled formality style transfer

❷ Sentences from 2 different domains

- Entertainment & Musics (E&M)
- Family & Relationships (F&R)

❸ Dataset size

- train dataset: 50k
- validation dataset: 3k
- test dataset: 1.5k

DATA

2.2 Augmented Data

Original: *I have never seen the show but I am worried that she will win because she is a Scientologist* Augmented: *I have never take in the show but I am distressed that she will pull ahead because she is a Scientologist*
Original: *i can't sign in at the chatroom* Augmented: *I am unable to sign in at the chat room.*

Examples of augmented dataset
(Synonym Replacement)

Original: *Perhaps it is because men are afraid to hurt girl's feelings.* Augmented: *Maybe because men are afraid of hurting girls' feelings.*
Original: *I presume it to be so, but avatars make anyone look great.* Augmented: *I guess that's the case, but the avatars make everyone look good.*

Examples of augmented dataset
(Round trip translation)

❶ Synonym Replacement

: replacement of certain words in formal input with their synonyms through **WordNet** (*155k*) according to their cosine similarity and likeness

❷ Round trip translation

: translation from English (source language) ↔ French (target language)

- train dataset size: 210k for each

PRE-PROCESSING

Methods

1. Indication of **start/end** of text sequence with '<' and '>' symbols (initiation/termination)
2. **Tokenization** by allocating unique identifiers for each vocab
3. Exclusion of common **punctuations** (;?:)
4. **Zero post-padding**
5. Filter out sentences greater than 150 after **data length** distribution examination

encoder input	**decoder input**	**decoder output**
<I do not intend to be mean.>	<I don't want to be mean.	I don't want to be mean.>
<I mean that you have to really be her friend.>	<and i mean Really be her friend.	and i mean Really be her friend.>

MODEL

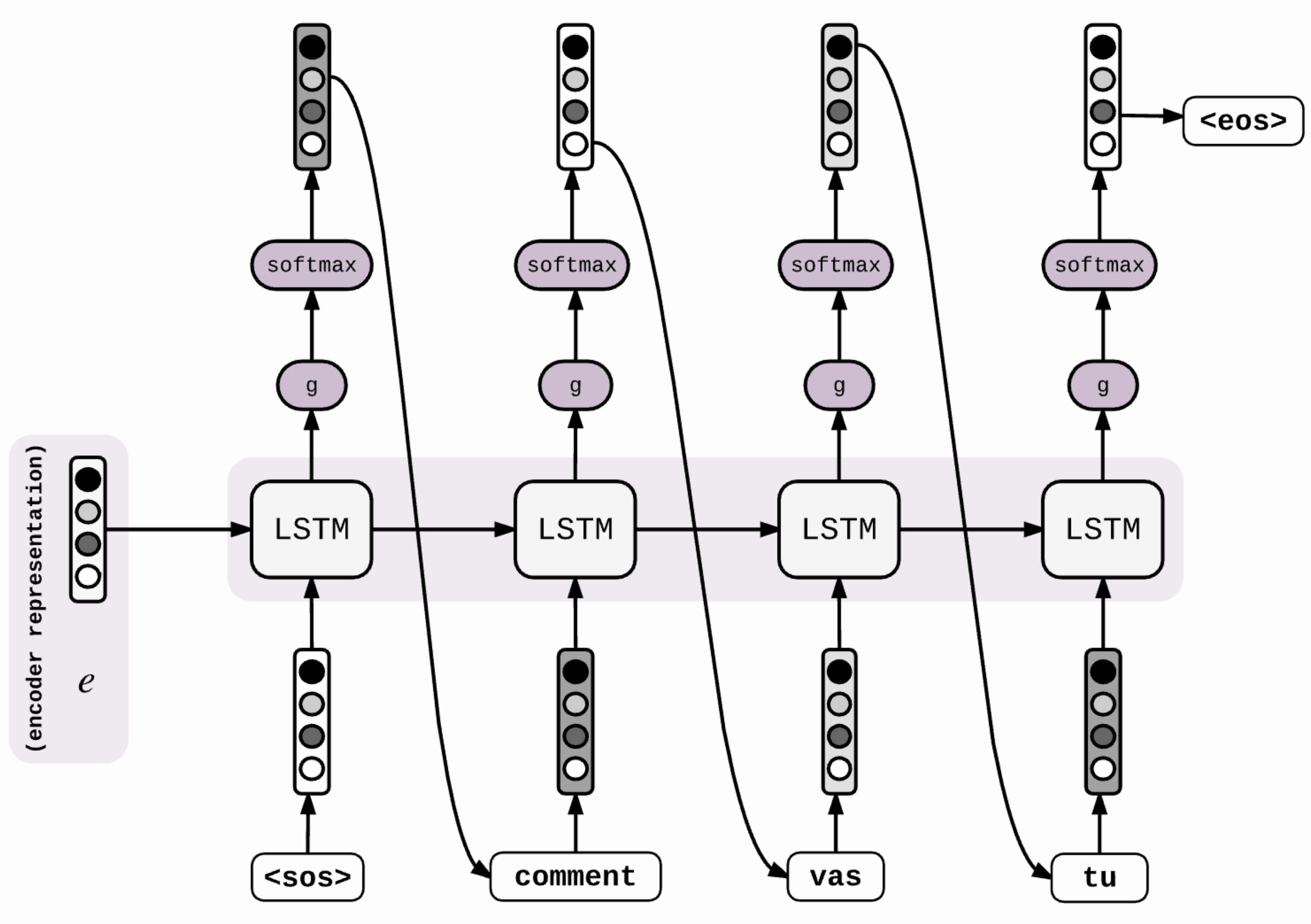
4.1 Sequence to sequence (seq2seq)

Encoder

- 3 Unidirectional LSTM layers
- input: input sequence + initial state of encoder
- ouput: recent step's hidden state + current state

Decoder

- time-step decoder: 2 Unidirectional LSTM layers + Dense layer
- combine preceding decoder with attention weights



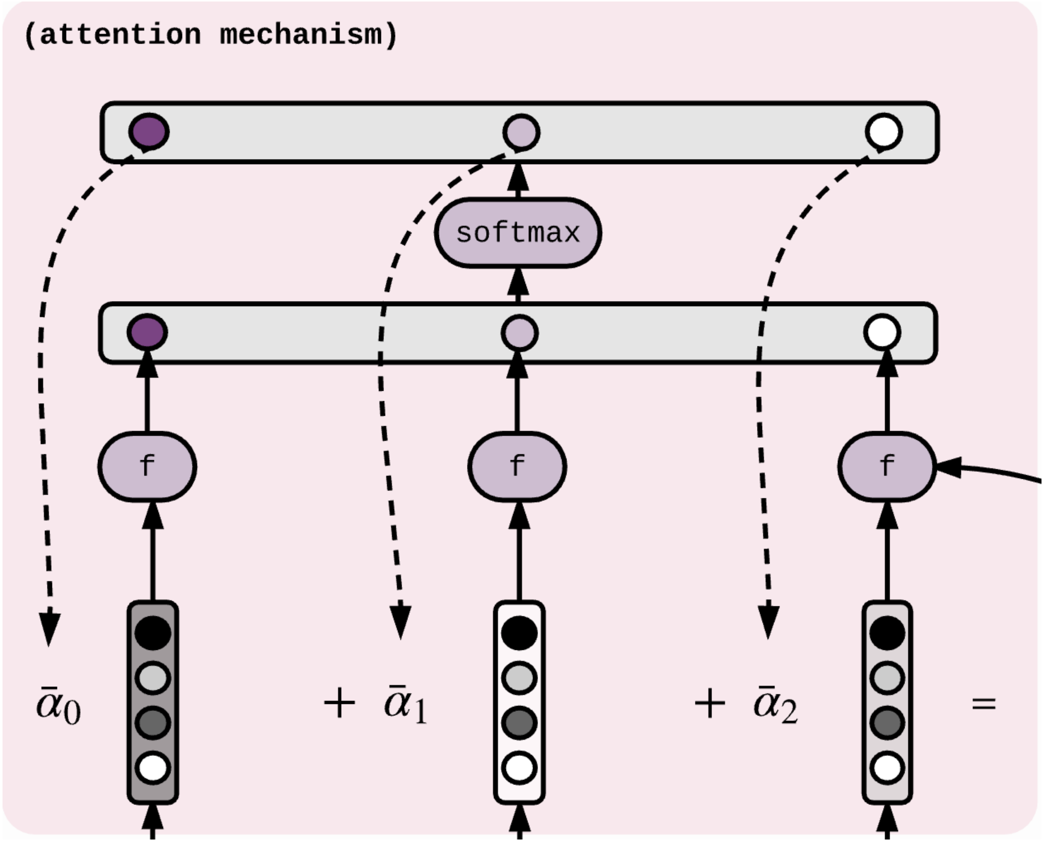
MODEL

4.2 Attention Mechanism

Why Attention?

: **Bottleneck** problem of seq2seq models

(difficultly to capture input vector information as the length of sequence increases)



Attention?

: gives greater weights to different parts of the input at distinct steps

Scoring function

difference in weights to the output encoder tokens

1. Dot
2. General
3. Concatenate

$$f(h_{t-1}, e_{t'}) = \begin{cases} h_{t-1}^T e_{t'} & \text{dot} \\ h_{t-1}^T W e_{t'} & \text{general} \\ v^T \tanh(W[h_{t-1}, e_{t'}]) & \text{concat} \end{cases}$$

EXPERIMENTS

5.1 Evaluation metric

BLEU score

Bilingual Evaluation Understudy

: Method to quantify the quality of generated sentences from one natural language to another

❶ Calculate correspondence between model's predicted output ↔ gold answer

❷ Averaged score over whole corpus of generated sentence

❸ Estimation of the model's overall quality

$$BLEU = exp(\sum_{n=1}^{N} w_n \log p_n)$$

EXPERIMENTS

5.2 Experimental setting

1. Attention-based transformer as seq2seq model (shared vocab of 20K BPE tokens)
2. Filteration of sentences longer than 150
3. Adam optimizer
4. Initial learning rate: 0.01 (reduce when metric stops improving)
5. Total number of epochs: 12
6. Early stopping

RESULTS

Result Table

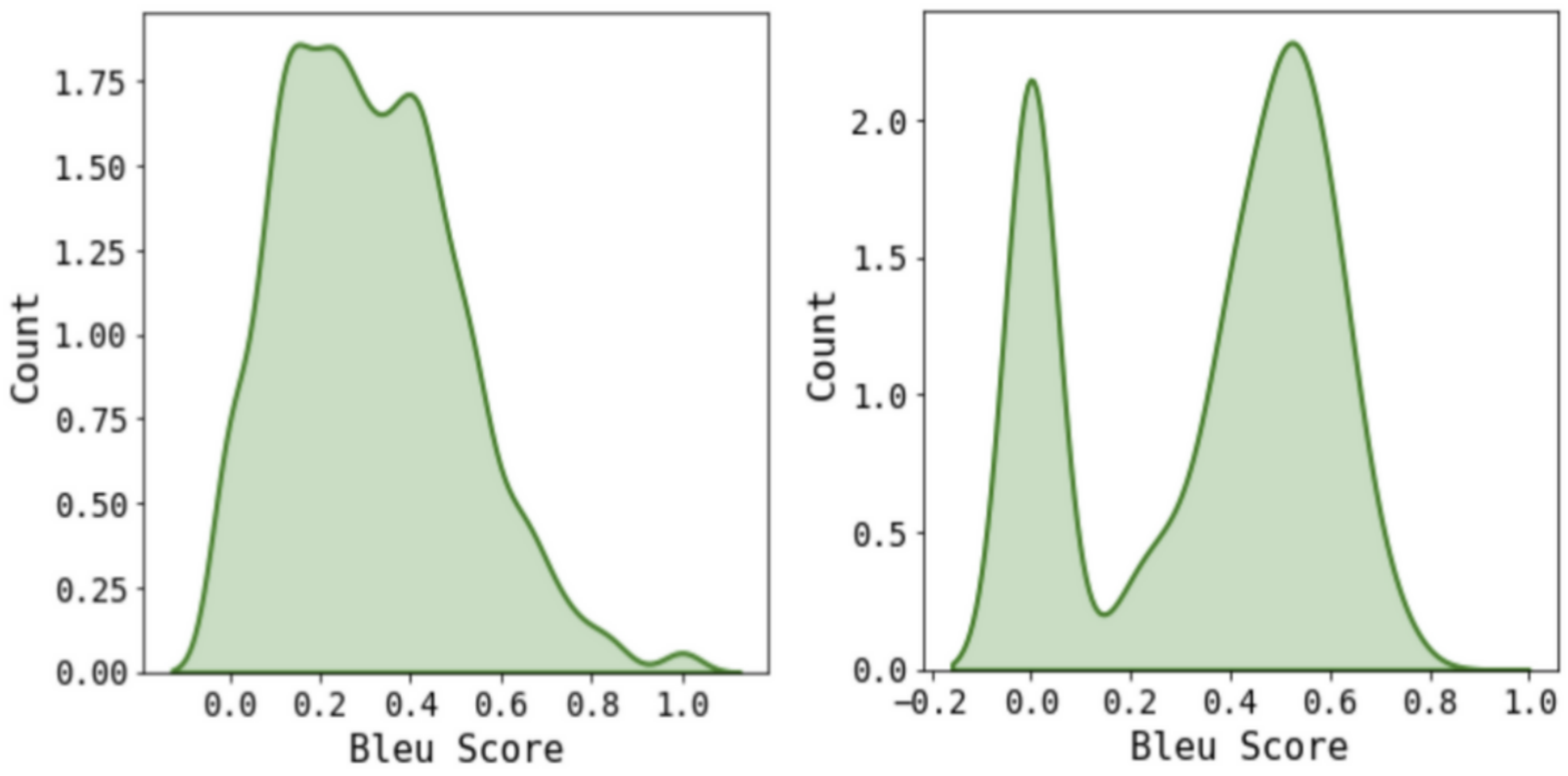
Model	Scoring Function	BLEU Score
Baseline	Dot General Concat	0.308 0.269 0.202
Synonym Replacement (SR)	Dot General Concat	0.313 0.350 0.313
Round-trip Translation (RTT)	Dot General Concat	**0.358** 0.332 0.300

Table 6: Final results (12 epochs)

ANALYSIS

7.1 Quantitative Analysis

❶ Skewed BLEU score distribution



❷ Loss values

Model	Scoring Function	BLEU Score	Train loss	Validation loss
Baseline	Dot General Concat	0.308 0.269 0.202	0.302 0.519 0.498	0.400 0.490 0.514
Synonym Replacement (SR)	Dot General Concat	0.313 0.350 0.313	0.347 0.419 0.347	0.369 0.452 0.369
Round-trip Translation (RTT)	Dot General Concat	0.358 0.332 0.300	0.348 0.459 0.565	0.351 0.449 0.546

Table 7: Training and Validation loss values (12 epochs)

ANALYSIS

7.2 Quantitative Analysis

❶ TOP 5 BEST & WORST for:

(1) Baseline
(2) Synonym Replacement
(3) Round Trip Translation

❷ General prediction

Best Prediction: *i traveled there and he was present.* Expected Output: *i went and there he was.*
Worst Prediction: *cause it's buy one take one.* Expected Output: *its because is because it is belle it is belle it is belle it is belle*

PATTERN

(1) Constant Repetition of certain word/phrase
(2) Loss of context (Example 'Girl' to 'Boy')

ABLATION STUDIES

8.1 Number of Epochs Analysis

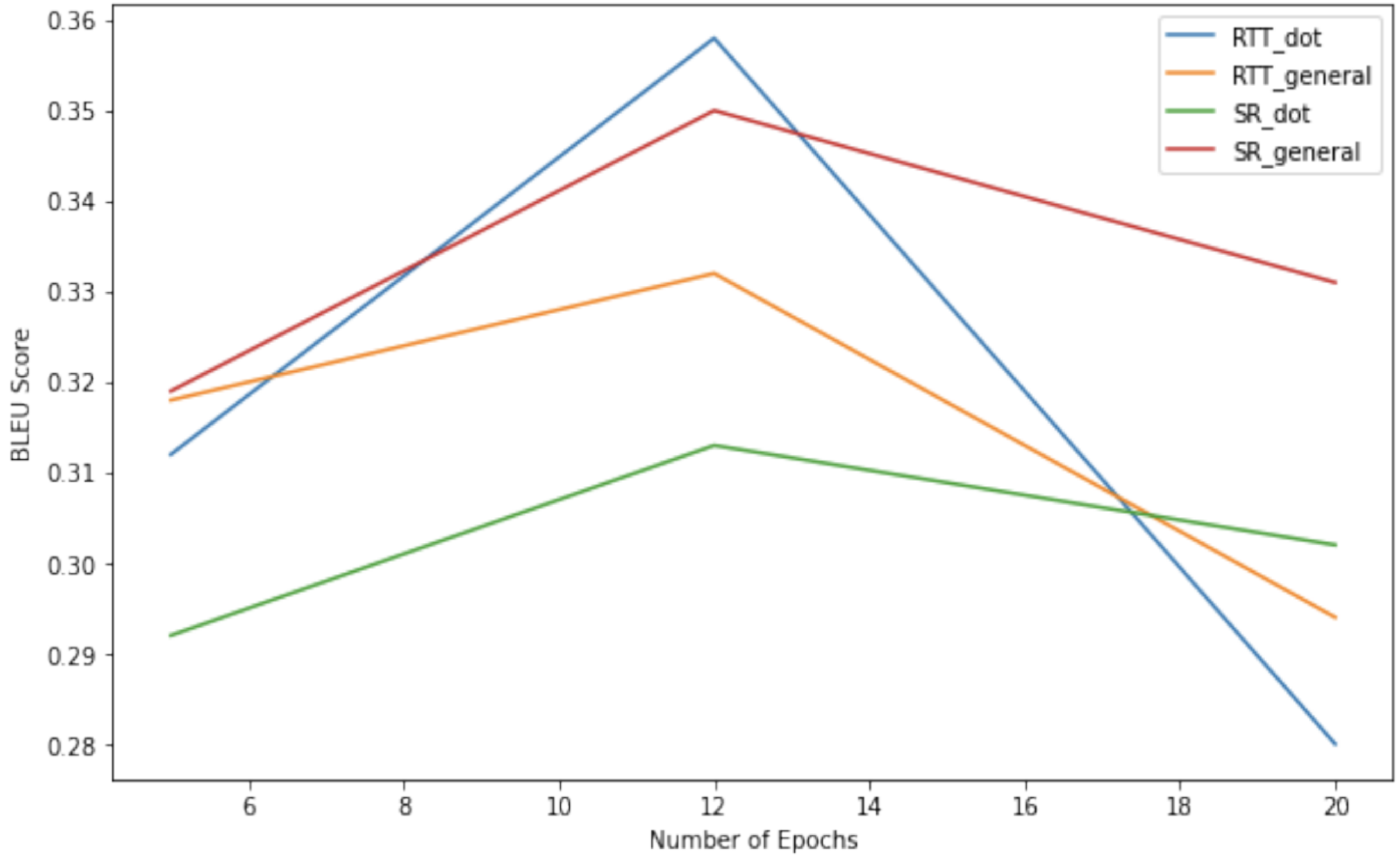
BLEU score itself is INSUFFICIENT!

- SKEWED distribution
- AVERAGED bleu score

CHANGE THE NUMBER OF EPOCHS:

❶ Beyond Epoch 12 : DECREASE

- OVERFITTING



(1) SR (Dot, General)

(2) RTT (Dot, General)

ABLATION STUDIES

8.2 Pivot Language Analysis

PIVOT LANGUAGES

1. FRENCH
2. CHINESE
3. SPANISH

Model	Scoring Function	French (Fr)	Chinese(Zh)	Spanish(Es)
BLEU Score	Dot	0.358	0.316	0.2719
	General	0.332	0.300	0.301
	Concat	0.300	0.304	0.302

ENGLISH vs FRENCH :

Identical sentence structure

CONCLUSION

9.1 Conclusion

❶ **In-domain Data augmentation methods for FST**

- Synonym Replacement
- Round Trip Translation

❷ **Sequence-to-Sequence Model with Attention**

- Scoring variants:
  Dot, General, Concatenate

DATA AUGMENTATION IMROVES PERFORMANCE

CONCLUSION

9.2 Further Investigations

1. Bidirection Transformation
   - Formal to Informal, Informal to Formal

2. Bidirectional LSTM
   - Contextual advatnage

3. Decoder Improvement
   - Beam Search
   - Extract context or phrase similarities better

4. Pre-trained language model
   - BERT or GPT
   - Substantial improvement

COSE461 NLP

Final Presentation

Improvement of Attention Based
Formality Style Transfer Model
with In-domain Data Augmentation

THANK YOU:)

통계학과 2019150419
기다연
컴퓨터학과 2019320062
이지수