Faculty of Computers and Information Science
Cairo University

Style-Transfer Text Paraphrasing

M.Sc Thesis Proposal Presentation

Ahmed Hani Ibrahim
-
Supervised By
Prof. Aly Fahmy

June 22, 2018

Content

Introduction
Task Definition

- **Paraphrasing** is core problem in Natural Language Processing that refers to texts that convey the same meaning but **with different expressions**
- We can consider it as a **transformation** for a given text with keeping the semantic it
- **Style-Transfer Paraphrasing** preserves the writer's style of writing while generating the paraphrase
- In other words, Style-Transfer Paraphrasing is the regular Text Paraphrasing **conditioned** on the writing style

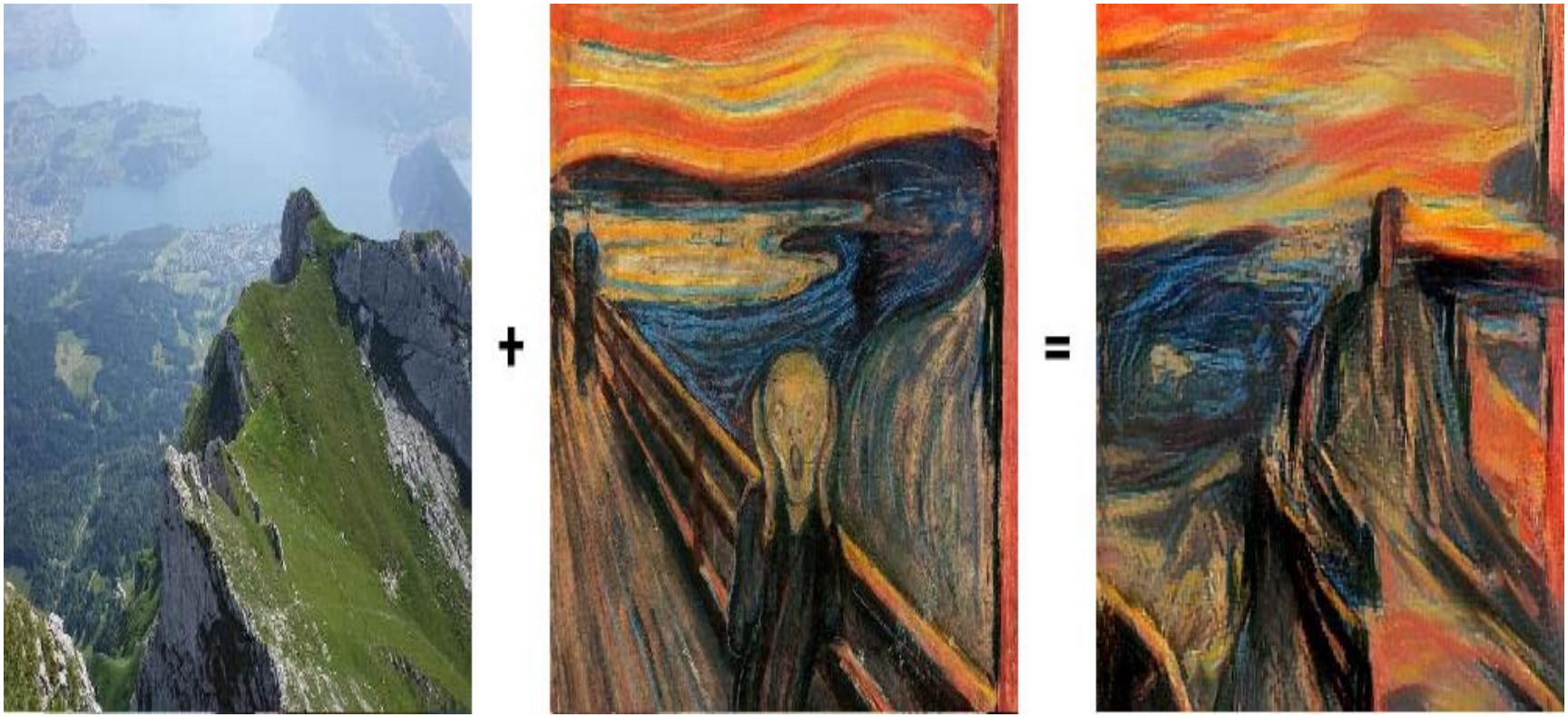
Paraphrase Examples

- How far is Earth from Sun
- What is the distance between Sun and Earth
- How many miles is it from Earth to Sun
- Distance between Earth and Sun

Style-Transfer Examples (Shakespeare Poems)

- JULIA: What shall by these things were a secret fool, That still shall see me with the best and force?
- DUKE SOLINUS: Merchant of Syracuse, plead no more, I am not partial to infringe our laws,
- SCENE III: An ante-chamber. The COUNT's palace.
- SCENE I: Venice. A street.

A Neural Algorithm of Artistic Style (Leon et al. 2015)



Introduction
Problem Formulation

Given a document ***D1*** in a style ***S1***, and a separate style ***S2***, can we produce a new document ***D2*** in style ***S2*** which preserves the meaning of ***D1***

Source Text	Trump	Twain	Shakespeare
It is obvious today that America has defaulted on this promissory note	**That's** obvious today that America has defaulted with **respect to** this promissory note.	It is **clearly evident the present day** that America has defaulted with **respect to** this promissory note	It is **very** obvious the **present day** that America has defaulted with **respect to** this promissory note

Motivation and Challenges

Motivation

- Paraphrases has numerous applications such as **Information Extraction**, **Question Answering**, **Semantic Search** and **Dialogue-based Systems**
- It can be used as a part of **Plagiarism Detection** for author copyrights ownership and **Text Similarity**
- It can be used for **Text Grammar Correction**
- Fix long sentences to short sentences while keeping their semantics
- **Data Augmentation** for several Natural Language Processing tasks such as **Sentiment Analysis** and **Author Identification and Recognition**
- Exploring current Machine Learning techniques and their capabilities in text generation and understanding

Motivation and Challenges

Challenges

Due to the complexity of natural language, automatically generating accurate and diverse paraphrases for a given sentence is still very challenging

- Keeping the structure and the language's grammar while generating texts is kind of hard problem
- Imitating the writing style while keeping the semantic representation of the paraphrased text is unexplored area as far as our knowledge and it still needs a lot of researches to reach a satisfactory results
- Unsupervised Learning problems are always considered to be challenging problems, specifically in text
- Semantic level representation of sentences and words are still under exploration and researching

Available Datasets

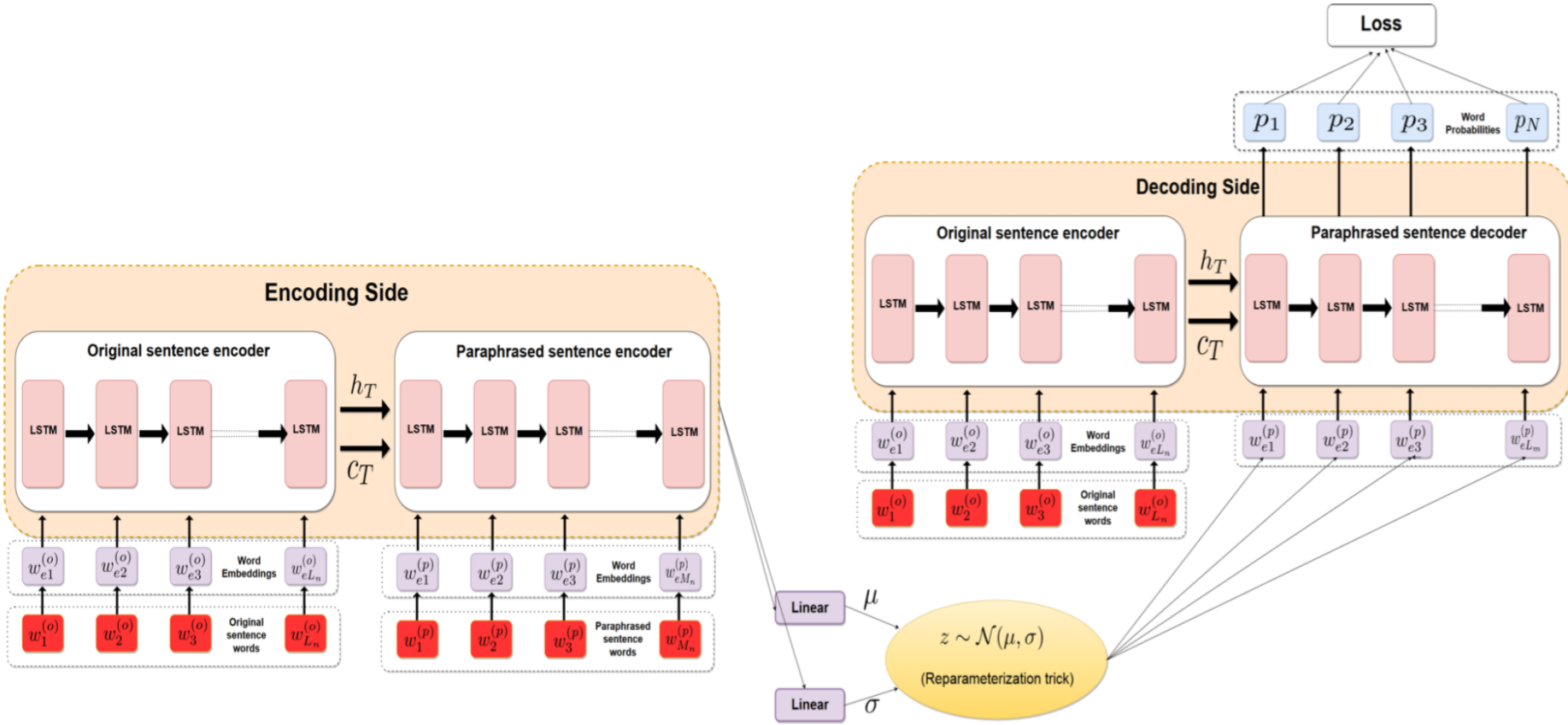
Paraphrases Datasets			
Dataset Name	**Number of Sentences**	**Number of Unique Words**	**Source**
PPDB	6.8 M	1.5 M	Pennsylvania University
Quora Questions Pair	400 K	111 K	Quora & Kaggle
MSCOCO	600 K	233 K	Microsoft

Style-Transfer Datasets			
Dataset Name	**Number of Sentences**	**Number of Unique Words**	**Source**
Donald Trump's	-	423 K	Stanford University
Mark Twain's	-	939 K	Stanford University
Charles Dickens	-	3 M	Stanford University
Shakespeare	-	52 K	Oxford University

Recent Existing Paraphrasing Approaches

Sequence-to-sequence Variational Autoencoder Model

A Deep Generative Framework for Paraphrase Generation.
(Gupta et al. 2017)



Recent Existing Paraphrasing Approaches

Sequence-to-sequence Variational Autoencoder Model Cont.

Encoder Side

- Two of LSTMs encoders
- the first one converts the original sentence into vector representation (Skip-thought)
- The second one encodes the paraphrased sentence as well
- The two vector representations are fed into Feedforward Network to estimate the VAE's mean and variance

Sampling

- Use the estimated parameters to produce a sample from a distribution that is parameterized by the estimated mean and variance

Recent Existing Paraphrasing Approaches

Sequence-to-sequence Variational Autoencoder Model Cont.

Decoder Side

- The VAE's output side uses an LSTM decoder which takes as input the latent representation and the vector representation of the original sentence
- Both latent representation and the original sentence representation are used to reconstruct the paraphrased sentence
- In the testing phase, we only concern on the decoder side and ignore the encoding side. We take the decoding side and feed it with the input sentence that we want to get its paraphrased version

Recent Existing Paraphrasing Approaches

Results and Evaluation

Model	MSCOCO			Quora Dataset		
	BLEU	METEOR	TER	BLEU	METEOR	TER
Seq-to-Seq (Sutskever, Vinyals, and Le 2014)	16.5	15.4	67.1	-	-	-
With Attention (Bahdanau, Cho, and Bengio 2014)	18.6	16.8	63.0	-	-	-
Seq-toSeq (Sutskever, Vinyals, and Le 2014)	28.9	23.2	56.3	-	-	-
Bi-directional (Graves, Jaitly, and Mohamed 2013)	32.8	24.9	53.7	-	-	-
With Attention (Bahdanau, Cho, and Bengio 2014)	33.4	25.2	53.8	-	-	-
Residual LSTM (Prakash et al. 2016)	37.0	27.0	51.6	-	-	-
Seq-to-seq VAE	41.7	31.0	40.8	17.4	22.2	54.9

Recent Existing Paraphrasing Approaches

Results and Evaluation Cont.

Source	What is my old Gmail account ?
Reference	How can you find all of your Gmail accounts ?
Generated	Is there any way to recover my Gmail account ? How can I find my old Gmail account number? How can I get the old Gmail account password ?

Source	What are my options to making money online ?
Reference	How can we earn money through online ?
Generated	How can I make money online ? What are ways of earning money online ? How can I profitable earn money online ?

Recent Existing Paraphrasing Approaches

Other Models

- Neural Paraphrase Generation with Stacked Residual LSTM Networks (Prakash et al. **2016**)
- Paraphrase Generation with Deep Reinforcement Learning (Zichao et al. **2018**)
- Adversarial Example Generation with Syntactically Controlled Paraphrase Networks (Mohit et al. **2018**)

Evaluation Metrics
BLEU Score

BLEU stands for **Bilingual Evaluation Understudy**

-

Generated Text	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

MaxCount(the, Reference 1, Reference 2) = 2

$$Score = \frac{2}{7}, (BadGeneration!)$$

- For each unique word in the generated text, we get the maximum count of it in the references, then sum them all normalized by the total number of candidate words
- This could be modified to work on **n-grams** instead of unigrams

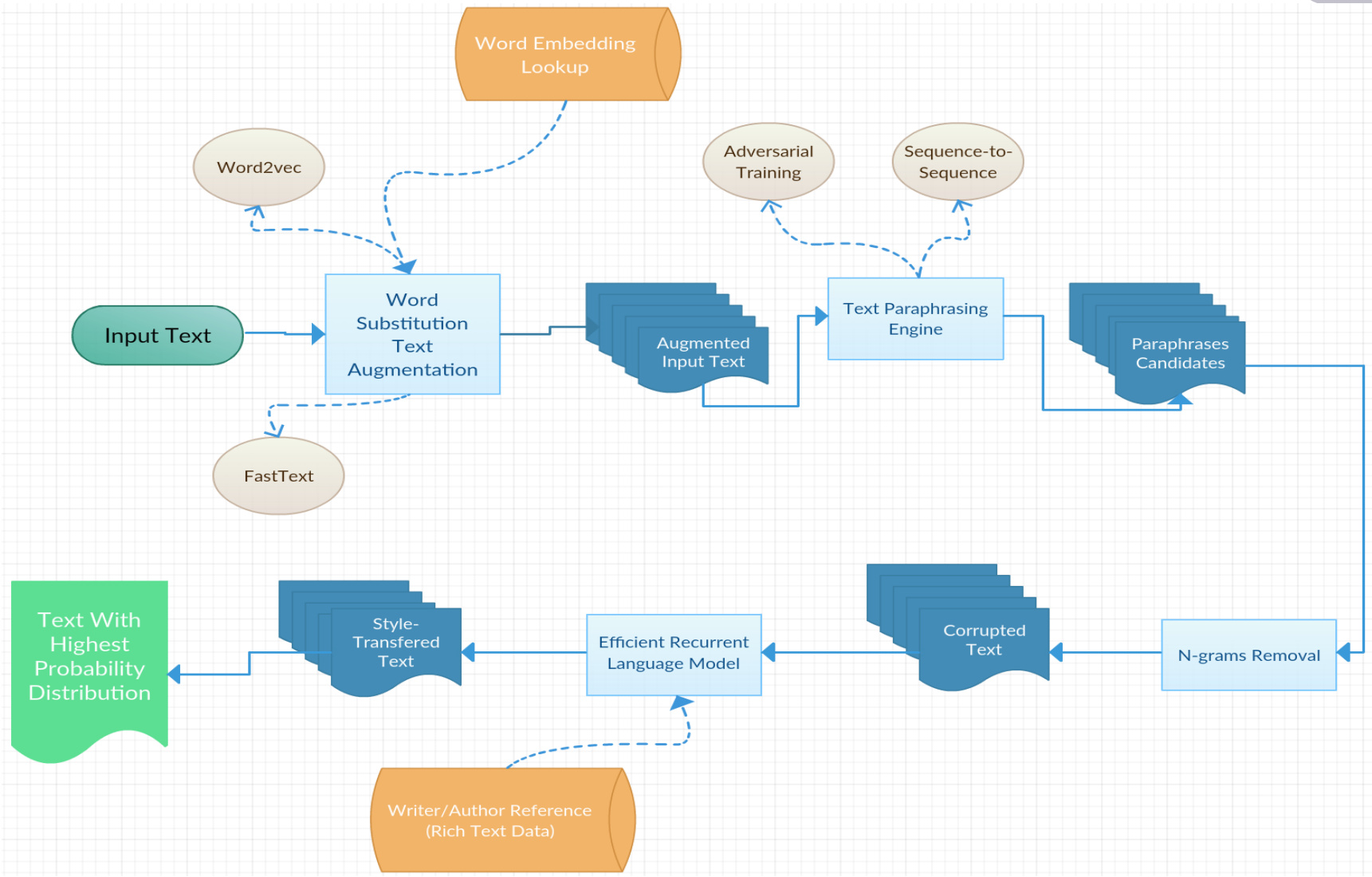
Evaluation Metrics

BLEU, METEOR, TER

- **METEOR** stands for **Metric for Evaluation of Translation with Explicit Ordering**
- **TER** stands for **Translation Error Rate**
- Regularly, **BLEU**, **METEOR** and **TER** are the evaluation metrics that are used in any Text Generation task
- **METEOR** word matches between input and output semantic equivalent
- **METEOR** relies more on words ordering instead of **BLEU's** n-grams approach
- **METEOR** has better correlation with human judgments, specially in short sentence level
- **BLEU** works better in large sentences and paragraphs
- **TER** works better in character-level matching cases

Proposed Approaches

Non-Parallel Corpora based Approach



Proposed Approaches

Non-Parallel Corpora based Approach

Pros

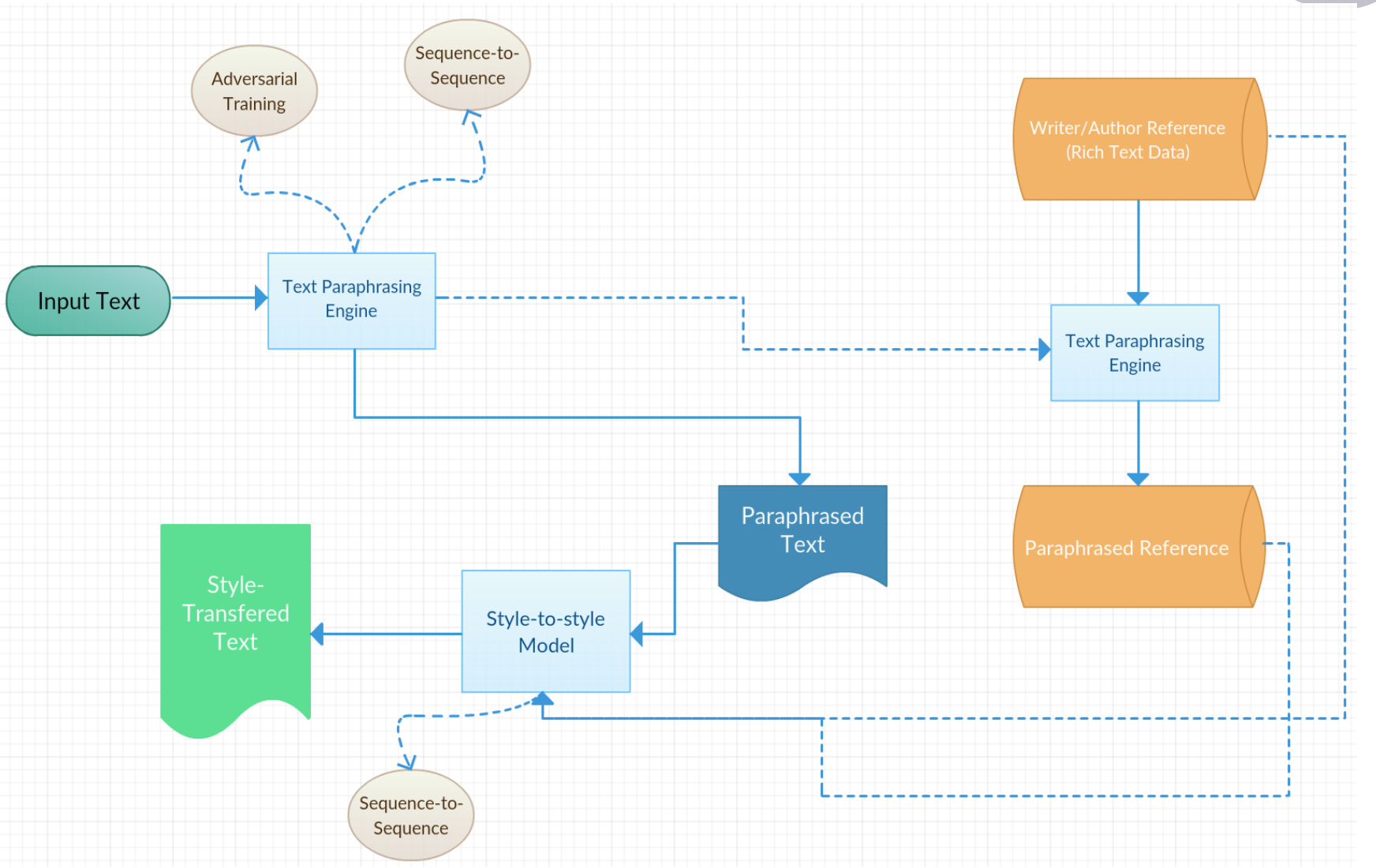
- **Doesn't** depend on parallel corpus that contains 2 corresponding text with different styles
- Depends on **Language Modeling** which isn't hard to be implemented nowadays
- Less complex compared to the Parallel Corpora approach
- The **Word Substitution trick** is commonly used nowadays for different type of tasks like **Sentiment Analysis**

Cons

- Needs an efficient Test Paraphrasing engine
- Needs and efficient **Word Embedding** model
- Taking the text with the highest probability could be misleading with some cases in text with **a lot of short consecutive sentences**

Proposed Approaches

Parallel Corpora based Approach



Proposed Approaches
Parallel Corpora based Approach

Pros

- A trained model for every author's style
- Produces one text instead of multiple ones
- Depends on **Transfer Learning** technique which is commonly used in Images and Texts nowadays

Cons

- Training a style transfer model for each writer consumes a lot of time compared to the Language Model
- Increases the **accumulative error** compared to the first approach (Text Paraphrasing Error + Paraphrase-style Transfer)
- Needs a very efficient **Text Paraphrasing model** as it will be used for 3 times **(Text to Paraphrase Text, Style-text to Paraphrase Text and Paraphrase Text to Style-text)**
- Needs a lot of author's samples

References

- A Deep Generative Framework for Paraphrase Generation - https://arxiv.org/abs/1709.05074
- Neural Paraphrase Generation with Stacked Residual LSTM Networks - https://arxiv.org/abs/1610.03098
- Paraphrase Generation with Deep Reinforcement Learning - https://arxiv.org/abs/1711.00279
- Sequence to Sequence Learning with Neural Networks - https://arxiv.org/abs/1409.3215
- Generating Sentences from a Continuous Space - https://arxiv.org/abs/1511.06349
- Variational Recurrent Auto-Encoders - https://arxiv.org/abs/1412.6581
- METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments - https://www.cs.cmu.edu/ alavie/papers/BanerjeeLavie2005-final.pdf
- BLEU: A Method for Automatic Evaluation of Machine Translation http://aclweb.org/anthology/P/P02/P02-1040.pdf
- Style Transfer with Non-Parallel Corpora http://prosello.com/papers/style-transfer-s16.pdf
- Facebook Research Lab Forum Discussions http://forums.fast.ai/t/exploring-style-transfer-for-text/2055/6

Thank you!