Dataset Generation, and Evaluation for Natural Question Generation

By: Amirhossein Razavi

Capstone Project 4088

Supervisor: Professor Aijun An

Introduction

- Natural Language processing (NLP) is a branch of Artificial Intelligence that
  - Analyzes and understands language as humans do
- Question generation from text is a task in NLP, which concerns
  - How to generate questions from text documents
- Applications of question generation:
  - Education
  - Reading comprehension
  - Development of knowledge base for conversational systems

Motivation

- The most effective methods for question generation:
  - Neural sequence-to-sequence methods
- Such methods require a large set of labelled training data:
  - A set of Questions and Answers (QA pairs)
- There are some existing datasets, e.g., SQuAD, NewsQA
  - Most of QAs in these datasets are *not natural,* meaning:
    - Questions are obtained by looking at the answers through, e.g., crowdsourcing.
  - Most datasets are *paragraph-based* with each question having:
    - A paragraph as long answer
    - A word or phrase as short answer

Motivation (Cont'd)

- Sentence-based natural QA datasets are lacking
- Some neural question generation systems can only take a single sentence as answer to generate questions.

Objectives

- Create a QA dataset that
  - Contains natural questions
    - Practical to use in the question generation tasks in the real world
  - Contains a single sentence as an answer

- Evaluate two state-of-the-art neural question generation methods:
  - BART
  - T5

Steps Taken in This Project

1. Dataset Generation
   - Original dataset
   - Data parsing
   - Coreference Resolution
2. Natural Question Generation
3. Evaluation

Original dataset

- The Google's Natural Questions dataset was used to help the development of this natural question generation task
- The dataset contains 307,373 (question, Wikipedia page, long answer, short answer) quadruples
- The dataset was originally designed to help spur development in open-domain question answering

Question:

who played will on as the world turns?

Short Answer:

Jesse Soffer

Long Answer:

William " Will " Harold Ryan Munson is a fictional character on the CBS soap opera As the World Turns . He was portrayed by Jesse Soffer on recurring basis from September 2004 to March 2005 , after which he got a contract as a regular . Soffer left the show on April 4 , 2008 and made a brief return in July 2010.

Original dataset (Cont'd)

- The reasons that this dataset was chosen:
  1. The dataset is large enough for the task to be done
     - 307,373 quadruples
  2. It has a broad range of topics
  3. As the questions (or queries) were asked by the real users, these questions represent real natural questions

Data Parsing

Only the data that:

1. Had both short answer and long answer
2. Had the short answer existed in the long answer
3. Had the sentence of short answer being less than 50 words
4. Did not contain any kind of HTML tag (list, table, etc.)
5. The corresponding question starts with a ‘wh’ question word
6. Had the “sentence” of short answer contain only one sentence

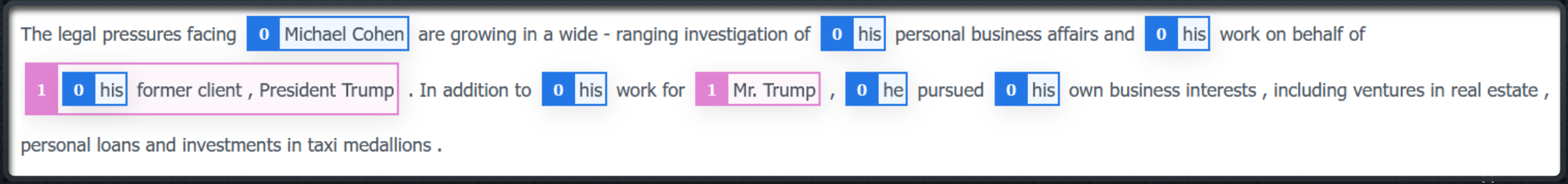
Data Parsing (Cont'd)

- An example of data parsing results:
  - Question: WHAT YEAR DID THE MOVIE DEUCES COME OUT
  - Long answer: DEUCES IS AN AMERICAN CRIME DRAMA WRITTEN AND DIRECTED BY JAMAL HILL . THE FILM STARS LARENZ TATE , MEAGAN GOOD , LANCE GROSS AND SIYA . THE FILM IS EXECUTIVE PRODUCED BY QUEEN LATIFAH FOR HER PRODUCTION COMPANY FLAVOR UNIT ENTERTAINMENT . DEUCES PREMIERED ON NETFLIX ON APRIL 1 , 2017 .
  - Original short answer from the dataset: 2017
  - Short answer gotten from data parsing: DEUCES PREMIERED ON NETFLIX ON APRIL 1 , 2017
- Let me name the dataset generated to be "Sentence-based Natural Question Dataset" or "SNQD"

Dataset	Size of the dataset
SNQD	59756

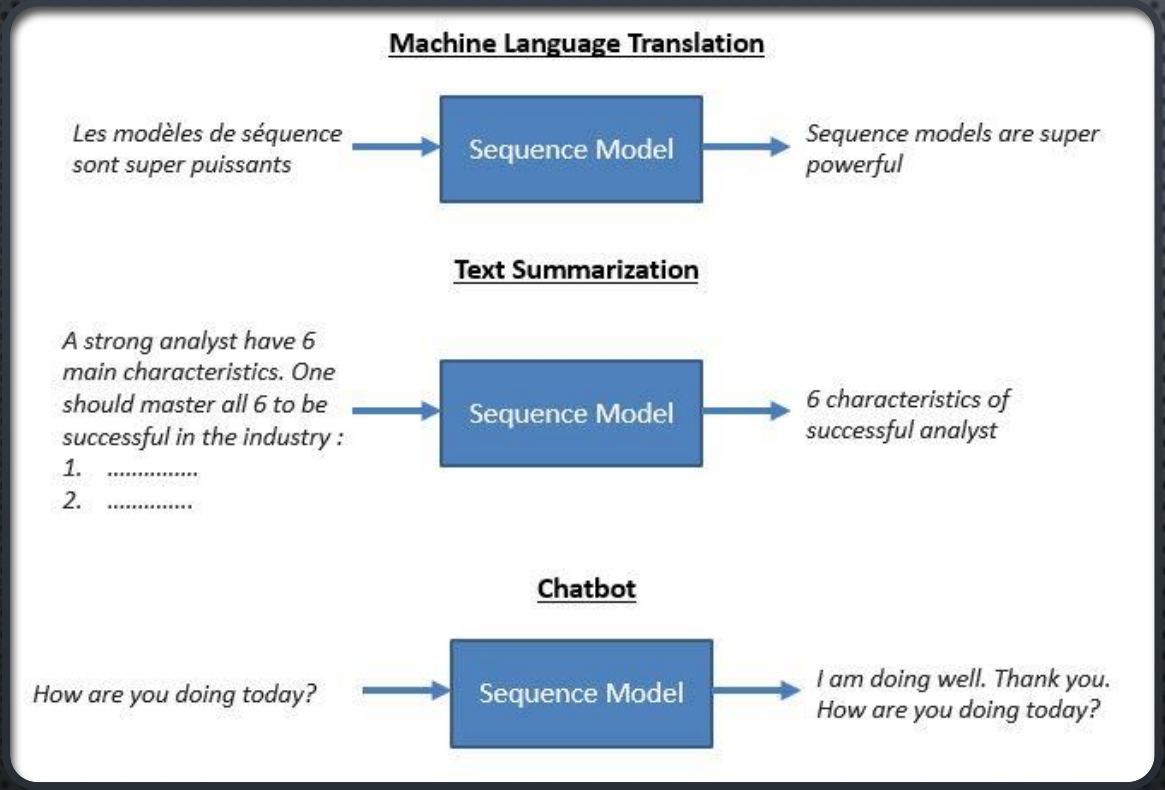
Coreference Resolution

- Coreference Resolution is the task of finding all expressions that refer to the same entity in a text
- It is useful for the NLP tasks that contain natural language understanding such as question answering
- The AllenNLP library was used for this part of the project which returns clusters of the representation of each span in the document



Steps Taken in the Project

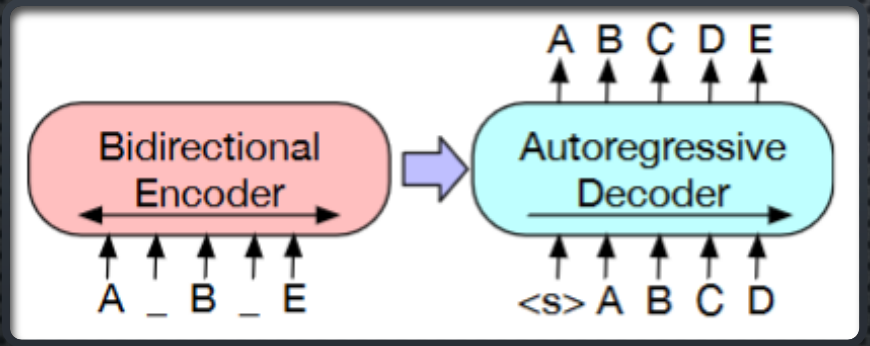
1. Dataset Generation
2. Natural Question Generation
   - Sequence-to-Sequence models
   - BART Algorithm
   - T5 Algorithm
3. Evaluation



Sequence-to-Sequence models

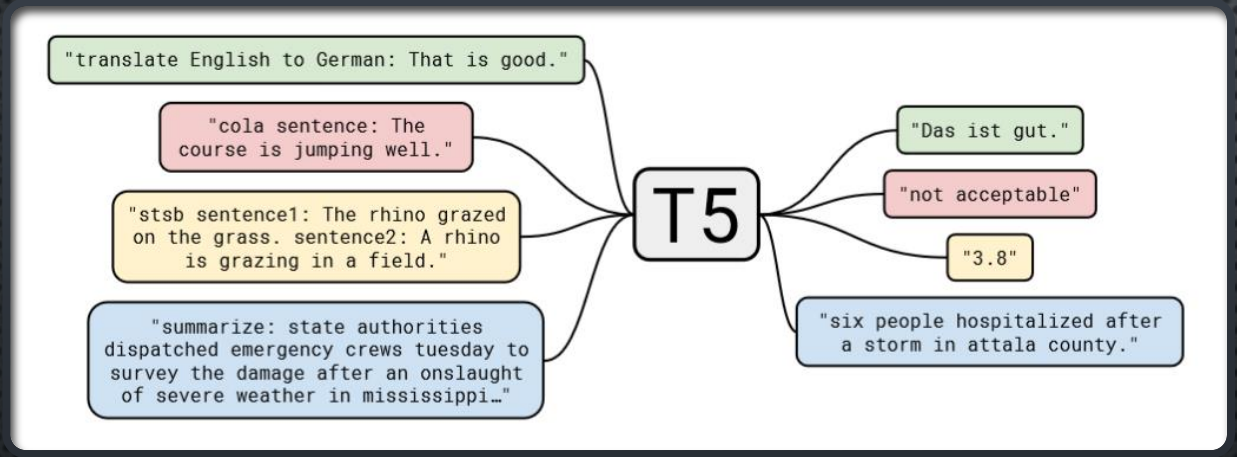
- These models take a sequence as input and output another sequence, as in, e.g., machine translation
- These sequences can be paragraphs, sentences, words, letters, time series, etc.
- Question Generation is a sequence-to-sequence task

BART Algorithm



- BART is a denoising autoencoder for pretraining sequence-to-sequence models
- BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text
- It uses a standard Transformer-based neural machine translation architecture which, despite its simplicity can be seen as generalizing BERT, GPT, and other recent pretraining schemes

T5 Algorithm



- T5 stands for Text-To-Text Transfer Transformer
- T5 is an encoder-decoder model and converts all NLP problems into a text-to-text format
- T5 is trained using teacher forcing
- It works well on a variety of tasks without any major modification

Steps Taken in the Project

1. Dataset Generation
2. Natural Question Generation
3. Evaluation
   - Without implementation of coreference resolution
   - With implementation of coreference resolution
   - Duplicate short answer problem
   - Human evaluation

Evaluation

- Types of evaluation:
  1. Automatic evaluation
     - BLEU Score (Bilingual Evaluation Understudy):
       - first introduced for the machine translation task.
       - measures how similar a candidate text sequence is to a reference text sequence. It is based on n-gram matching (n=1, 2, 3, or 4).
     - METEOR Score (Metric for Evaluation of Translation with Explicit Ordering):
       - based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.
       - Use stemming and synonymy matching by aligning them to ground truth
     - ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation)
       - Proposed for evaluating text summarization and machine translation.
       - $ROUGE_L$: based on the Longest Common Subsequence (LCS) matching.
  2. Human evaluation

Evaluation (Cont'd)

- Giving our evaluation process a test-run to check its effectiveness
- The dataset used in this test is called 'Car Manual' dataset
- This dataset was chosen due to its complete sentences and small test set (train: 4360, dev: 360, test: 312, Question-Answer pairs)
- Automatic evaluation:

Dataset	Model	BLEU 1	BLEU 2	METEOR	$ROUGE_L$
Car Manual	BART	0.560175	0.498418	0.391080	0.576190
	T5	0.548173	0.478553	0.370160	0.544810

- Human evaluation:
  - In most cases the generated questions were correct even if it differed from the targeted question (small change in grammatical structure, or etc.)

Without implementation of coreference resolution

- The sentences could contain words like ‘he’, ‘she’, ‘they’, ‘their’, etc. which would cause the sentence to be incomplete
- For example: THE NAME IS DERIVED FROM PATRONYMIC FORM OF THE NAME HENDRY , WHICH IS A SCOTTISH FORM OF HENRY
- A true natural question cannot be generated by having incomplete sentence

Dataset	Model	BLEU 1	BLEU 2	METEOR	$ROUGE_L$
SNQD (0-50)	BART	0.323440	0.225199	0.166647	0.300929
SNQD (0-30)		0.366399	0.263268	0.188732	0.354553
SNQD (10-30)		0.346400	0.242911	0.179168	0.322215
SNQD (30-50)		0.376816	0.276793	0.198042	0.375410

With implementation of coreference resolution

- The sentences are complete, natural questions can be generated from these sentences
- For example: HENDERSON IS DERIVED FROM PATRONYMIC FORM OF THE NAME HENDRY , WHICH IS A SCOTTISH FORM OF HENRY
- The word count in most sentences has changed
- As the sentences are complete, no need to test the dataset based on different count of words in sentences

Dataset	Model	BLEU 1	BLEU 2	METEOR	$ROUGE_L$
Coref Resolved SNQD	BART	0.511272	0.391012	0.261638	0.498220
	T5	0.464178	0.342072	0.227008	0.452510

Comparing with/without implementing coreference resolution

- Targeted question: WHEN DOES THE NEXT EPISODE OF DYNASTY COME OUT
- Without implementing coreference resolution:
  - Sentence: ON APRIL 2 , 2018 , THE CW RENEWED THE SERIES FOR A SECOND SEASON , WHICH IS SET TO PREMIERE ON OCTOBER 12 , 2018
  - BART: WHEN DOES THE NEW EPISODE OF RIVERDALE COME OUT
  - T5: WHEN DOES THE NEW SEASON OF THE TV SHOW COME OUT

1. With implementing coreference resolution:
   - Sentence: ON APRIL 2 , 2018 , THE CW RENEWED DYNASTY FOR A SECOND SEASON , WHICH IS SET TO PREMIERE ON OCTOBER 12 , 2018
   - BART: WHEN DOES DYNASTY COME BACK FOR SEASON 2
   - T5: WHEN DOES THE NEW SEASON OF DYNASTY START

Duplicate short answer problem

- Multiple different question corresponding to the same text exists in the dataset
- Benefit: It gives the system chance of understanding the texts in multiple ways
- Drawback: The possibility of having the same sentence in the training and testing dataset
- The approaches taken to evaluate this issue:
  1. Duplicates allowed (Dupes)
  2. Only unique sentences be present in the dataset (Unique)
  3. Keeping up to three questions corresponding to the same sentence (Moderate)

Duplicate short answer problem (Cont'd)

Dataset	Number of data
Coref Resolved SNQD (Dupes)	59756
Unique CRSNQD	45605
Moderate CRSNQD	55863

Dataset	Model	BLEU 1	BLEU 2	METEOR	$ROUGE_L$
Coref Resolved SNQD	BART	0.511272	0.391012	0.261638	0.498220
	T5	0.464178	0.342072	0.227008	0.452510
Unique CRSNQD	BART	0.456743	0.337215	0.224961	0.449741
	T5	0.450142	0.327546	0.217379	0.438352
Moderate CRSNQD	BART	0.499597	0.378458	0.252593	0.487689
	T5	0.465117	0.343105	0.226187	0.453794

Human evaluation

- Although BART's performance scores were higher, T5 in my opinion did better at generating questions
- How can that be the case:
  - The question generated is correct but different from the targeted question
  - The text in questions contain words that are not in the sentence, and the model might have learned it from a possible duplicate in training dataset
  - The question and short answer might have little to no overlap

Human evaluation (Cont'd)

For example:

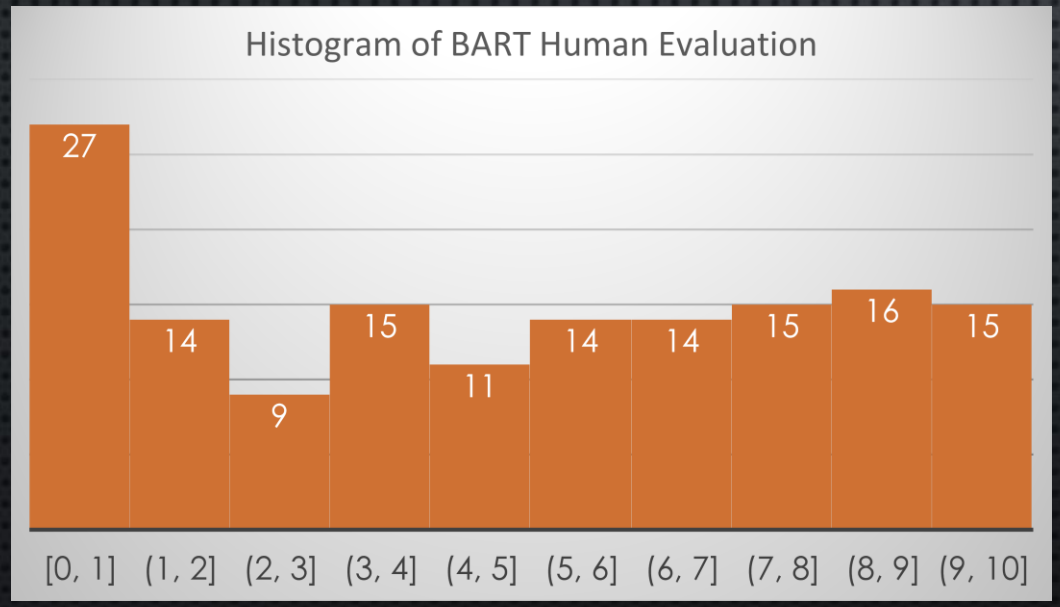
- Sentence: The 2017 World Series began October 24 and Game 7 was played on November 1 , in which the Houston Astros defeated the Los Angeles Dodgers , 5 - 1 , to capture the Houston Astros's first World Series championship in franchise history
- Question generated using BART: when does the world series for baseball start
- Question generated using T5: when does the world series start in 2017
- Targeted question: how many games in the major league baseball playoff series

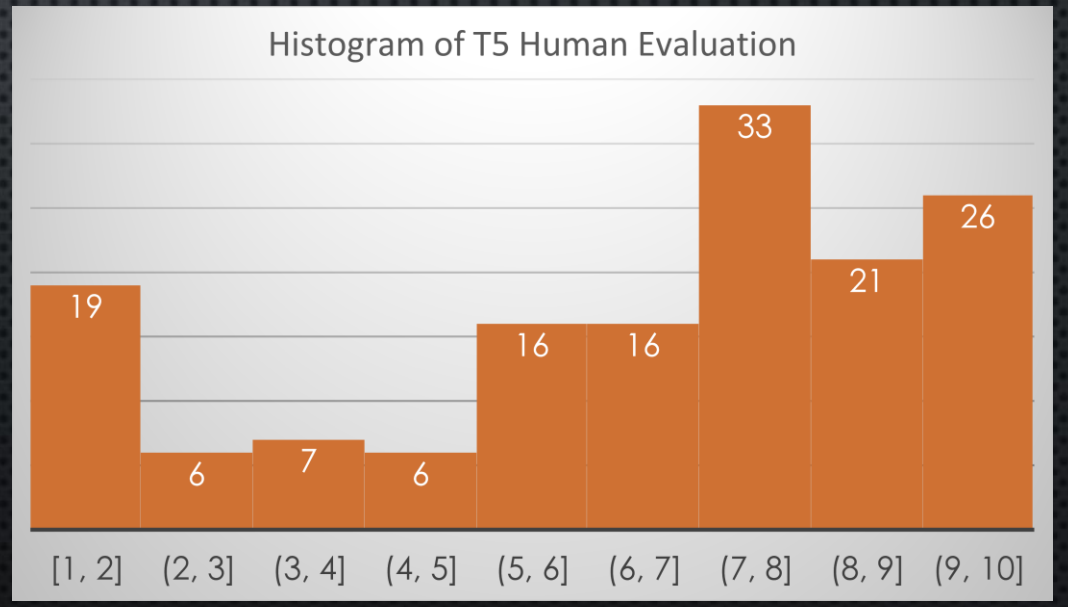
Human evaluation (Cont'd)

- The performance scores by themselves, might not represent the actual performance of the models
- Basically NLP is about teaching the computer to 'think' like a human so humans can be good judges for the purposes of this project
- This human evaluation was done by 5 different people each having 30 sentences, BART generated questions, and T5 generated questions (same for all 5 people) to score from 1 to 10
- The 30 sentences, and their corresponding generated questions were selected from the Coref Resolved SNQD dataset randomly

Human evaluation (Cont'd)

- The results of the human evaluation done:





Summary of Contributions

- Developed a new QA dataset (Coref Resolved SNQD) for natural question generation (QG)
  - Can be shared with the NLP community for training neural QG models
- Evaluated two state-of-the-art models using both
  - Automatic evaluation with BLEU, $ROUGE_L$ and METEOR
  - Human evaluation
- Findings:
  - BART performs better in automatic evaluation
  - T5 performs better in human evaluation
- The findings indicate more research needs to be done to improve automatic evaluation

What I Learned in this Project

- First time working on an NLP or machine learning project
- First time to use Google Cloud's servers
- I learned:
  - Text processing techniques, such as parsing, coreference resolution.
  - The state-of-the-art text generation models (BART and T5)
  - Using BART and T5 to train models for question generation
  - How to evaluate text generation using the standard measures (BLEU, $ROUGE_L$, METERO)

Possible Future Works:

- Using the generated dataset on other question generation algorithms, especially the sequence-level QG systems
- Checking for the possible structural grammar issues in the generated questions
- Checking whether the keywords in the sentences are present in the question generated or not

References:

- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, Slav Petrov; Natural Questions: A Benchmark for Question Answering Research. Transactions of the Association for Computational Linguistics 2019; 7 453–466. doi: https://doi.org/10.1162/tacl_a_00276
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, & Luke Zettlemoyer. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. http://aclanthology.coli.uni-saarland.de/pdf/P/P02/P02-1040.pdf.

- Automatic Neural Question Generation using Community-based Question Answering Systems, written by Tina Baghaee. https://opus.uleth.ca/handle/10133/5004
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Baltimore, Maryland, pages 376–380. http://www.aclweb.org/anthology/W14-3348
- Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. Association for Computational Linguistics, Barcelona, Spain, pages 74–81. http://aclweb.org/anthology/W/W04/W04-1013.pdf.
- Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P.J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv, abs/1910.10683.

- https://github.com/Maluuba/nlg-eval
- https://ai.google.com/research/NaturalQuestions/dataset
- https://github.com/patil-suraj/question_generation
- https://huggingface.co/docs/transformers/model_doc/t5
- https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html
- *Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.*
- *https://www.analyticsvidhya.com/blog/2018/04/sequence-modelling-an-introduction-with-practical-use-cases/*
- *https://www.analyticsvidhya.com/blog/2020/08/a-simple-introduction-to-sequence-to-sequence-models/*