Text Generation, Editing and Summarization

BING Lidong (邴立东)
https://lidongbing.github.io/

R&D Center Singapore
Machine Intelligence Technology
Alibaba DAMO Academy

Define Smarter Tomorrow.

National University of Singapore, Jan. 2019

DAMO: Discovery, Adventure, Momentum and Outlook

Rooted in Science, Innovate for Applications

"Must outlive Alibaba", "Serve at least 2 billion people worldwide", "Future-oriented and use technology to solve the challenges of the future".

5 Research Areas | 14 Laboratories

- **Machine Intelligence Technology**
  - Speech Lab
  - Language Technology Lab
  - Vision Lab
  - Decision Intelligence Lab
  - City Brain Lab
- **Data Computing**
  - Computing Technology Lab
  - Data Analytics & Intelligence Lab
  - Database & Storage Lab
- **Robotics**
  - Intelligence Transportation Lab
- **Financial Technology**
  - Financial Intelligence Lab
  - Blockchain Lab
  - Biometrics Lab
- **X Laboratory**
  - Quantum Lab
  - AI Labs

Machine Intelligence Technology at DAMO

Hundreds of Researchers and Engineers
in Hangzhou, Beijing, Seattle, Silicon Valley and Singapore

Speech Processing

- Speech Recognition
- Speech Synthesis
- Voice Biometrics
- Human-Machine Interaction

Natural Language Processing

- Semantic Analysis
- Sentiment Analysis
- Text Classification
- Question and Answering, Chatbot
- Machine Translation

Image/Video Analytics

- Product Identity & Search
- Face Recognition
- Object Recognition
- Scene Recognition
- Video Search

Optimization & Decision Making

- Predictive Inventory Optimization
- Delivery Assignment Optimization
- Manufacturing Scheduling
- Predictive Maintenance

NLP R&D at Alibaba

NLP research has made great progress from using complex sets of human rules, statistical natural language processing techniques to deep learning nowadays

Missions of Alibaba's NLP R&D:

1. Support all the demands of NLP techniques and applications in Alibaba's eco-system (new-retail, finance, logistics, entertainment etc.)
2. Enable Alibaba's business partners with NLP solutions
3. Advance the State-of-the-Art NLP research with colleagues from both academia and industries

Alibaba-DAMO-NLP: 100 employees (e.g., former tenured Professors and senior researchers) in 6 locations all over the world.

R&D Center Singapore

An international R&D team with the focus on developing cutting edge speech and language processing technologies, including **ASR, TTS, NLP, and MT**.

Paying special attention to the areas of **multilingual speech and language processing**, including:

- Speech recognition and synthesis of multiple languages
- **NLP technology for SouthEast Asian languages (SEAL)**
- **Machine translation systems for SEAL**

AliNLP

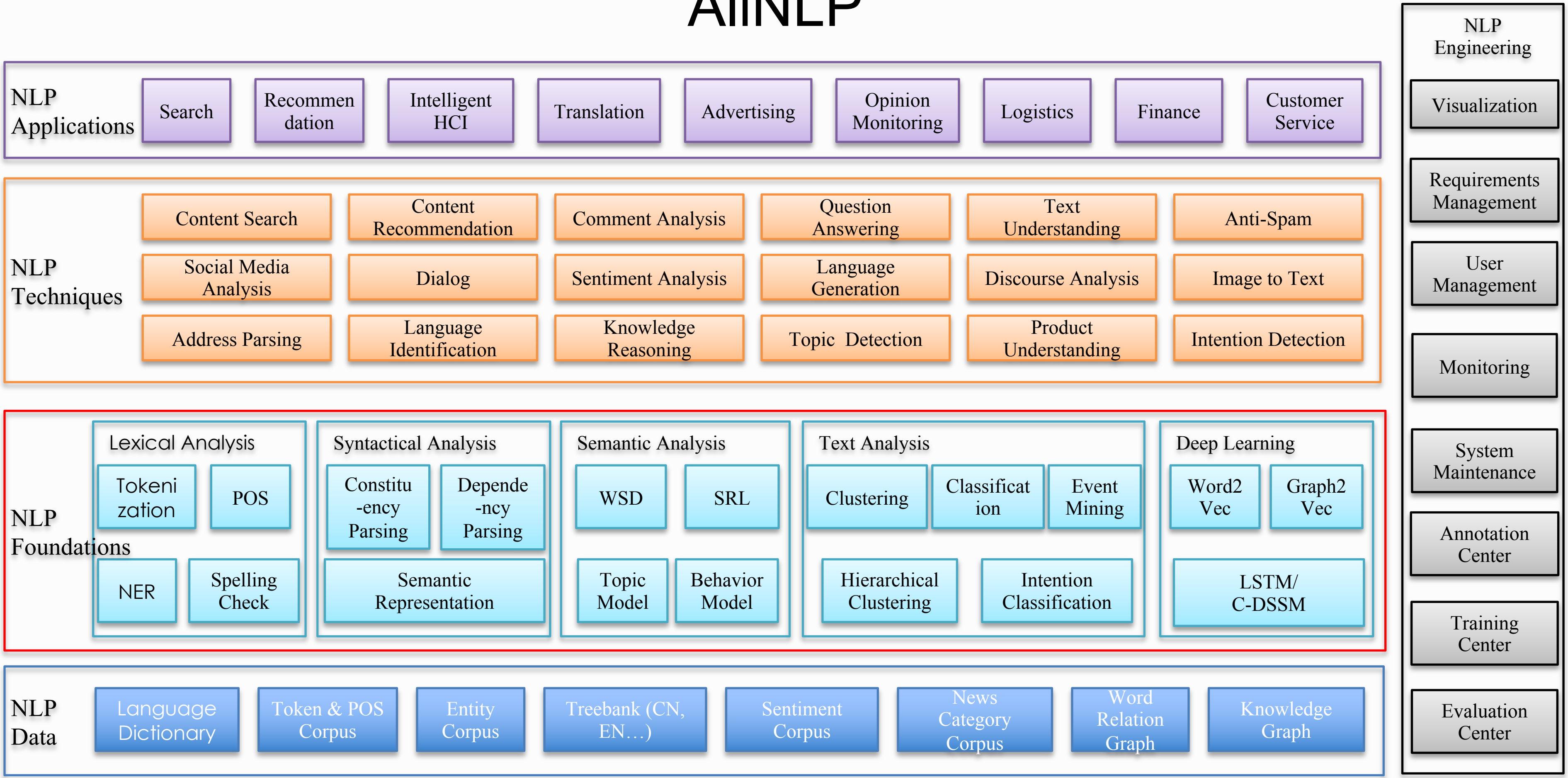
AliNLP is a large-scale NLP platform for the entire Alibaba Eco-system. The platform covers major aspects of NLP such as data collecting/processing techniques and multilingual algorithms for lexical, syntactic, semantic, document analysis, and distributed representation of text

Used in 350+ business scenarios (Jan, 2019) with more than 2000Billion+ API calls per day.

Some key characteristics:

- Utilizing behavior data instead of demanding human annotations for NLP algorithms
- Utilizing multiple correlated tasks for improving effectiveness of individual tasks of the complex Alibaba eco-system

AliNLP

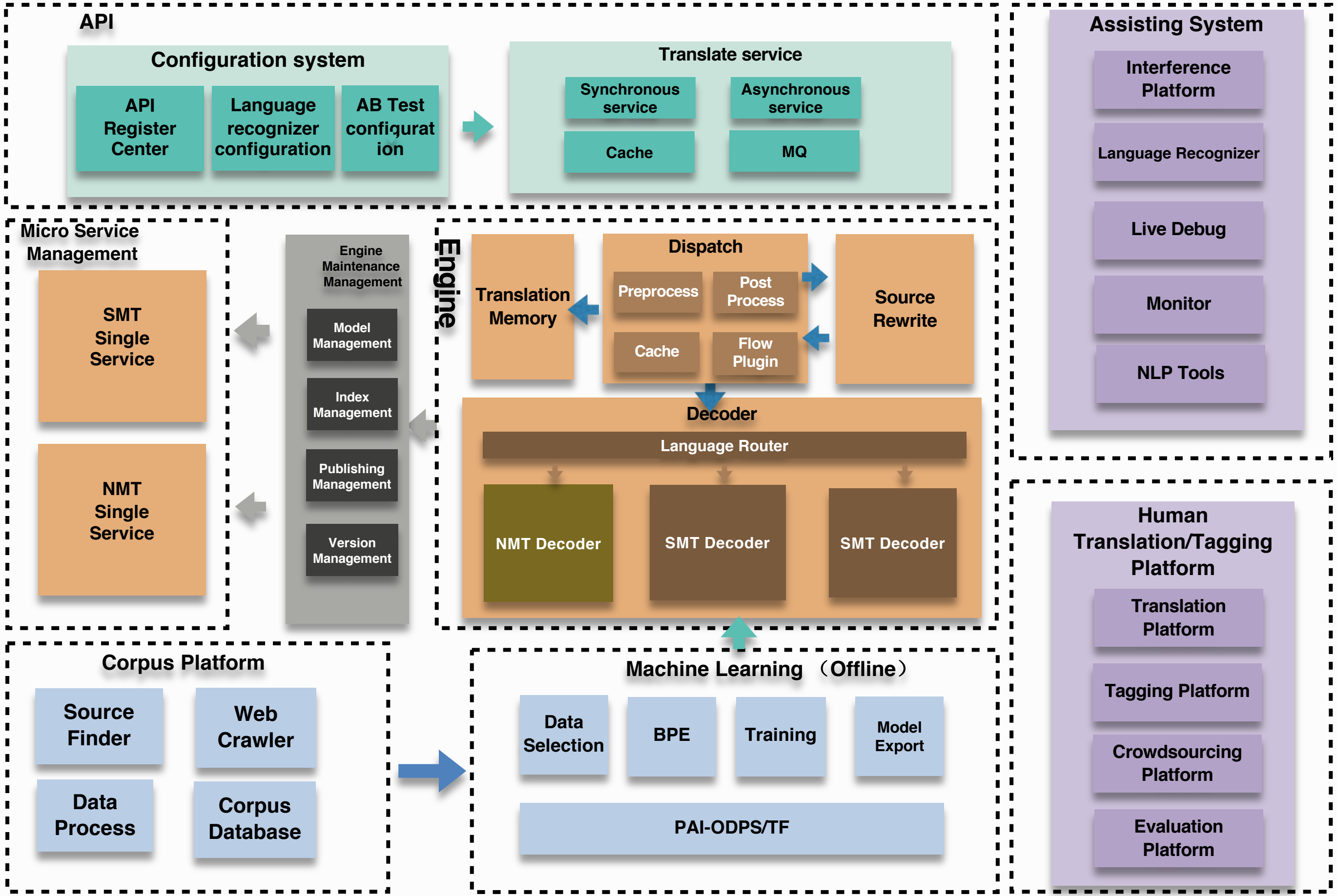


Define Smarter Tomorrow.

Machine Translation at Alibaba

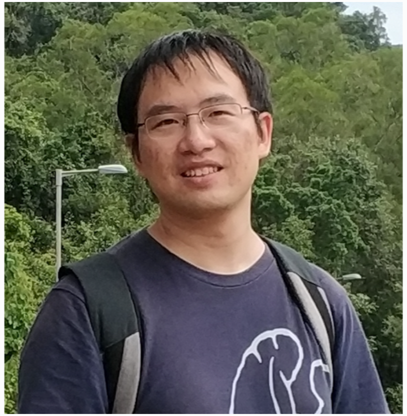
2017- 2018 :

- Support AliExpress, Alibaba.com and Lazada. Processing 250 billion requests in the whole year (60% increase)
- Translating 20 trillion words in the whole year ($2 billion if using Google)
- In WMT'18 got No. 1 in 5 MT tasks for automatic evaluation



Define Smarter Tomorrow.

Text Generation, Editing and Summarization



Fu Zihao (付子豪)

Gao Yifan (高一帆)

Li Juntao (李俊涛)

Li Piji
(李丕绩)

Liao Yi
(廖亿)

Text Generation: A Longstanding Topic

- Congenitally, focus on addressing two questions separately. (Reiter and Dale, 1997; Jurafsky and Martin, 2014).
  - What to say
  - How to say
- Leading to systems with explicit
  - content selection
  - macro- and micro-planning
  - surface realization components
- Encoder-decoder generation systems
  - Reduce the manual effort
  - Allow more flexible surface form of the text
  - Increase the fluency of NLG outputs
  - Take more types of input: Text, Audio/music, Image, Video

BUT: it is risky to deploy application, since it is not fully controllable.

Text Generation: An Even Broader Topic Now

- More to less: abstractive summarization, headline generation
- QA and Chatbot: question, answer, and response generation
- Text editing: sentiment transfer, formality editing, modern to Shakespeare, spoken text smoothing
- Less to more: KB triples to text, story generation
- ABC to αβγ: machine translation. Others ……

Text Summarization

- Single Document Summarization
  - Generate summary for a single input document
- Multi-Document Summarization
  - Generate summary for multiple docs about the same topic

- General Methods
  - Extractive
    - Select some original sentences
  - Compressive
    - Delete some words and phrases from the extracted sentences
  - Abstractive
    - Sentence fusion or generation

Abstractive Summarization

- The objective
  - Generate abstractive summaries as human does
  - Condense text more aggressively than extraction and compression
- Human-written summaries are more abstractive (Cheung and Penn, 2013; Jing and McKeown, 2000).
- Bottleneck
  - Natural language understanding and generation
  - Not many works on abstractive summarization before DL
    - Sentence clustering and fusion within each cluster (Barzilay and McKeown 2005; Filippova and Strube, 2008; Filippova, 2010)

Phrase Selection and Merging for Abstractive MDS

- More fine-grained syntactic units than sentences
  - noun/verb phrases (NPs/VPs)

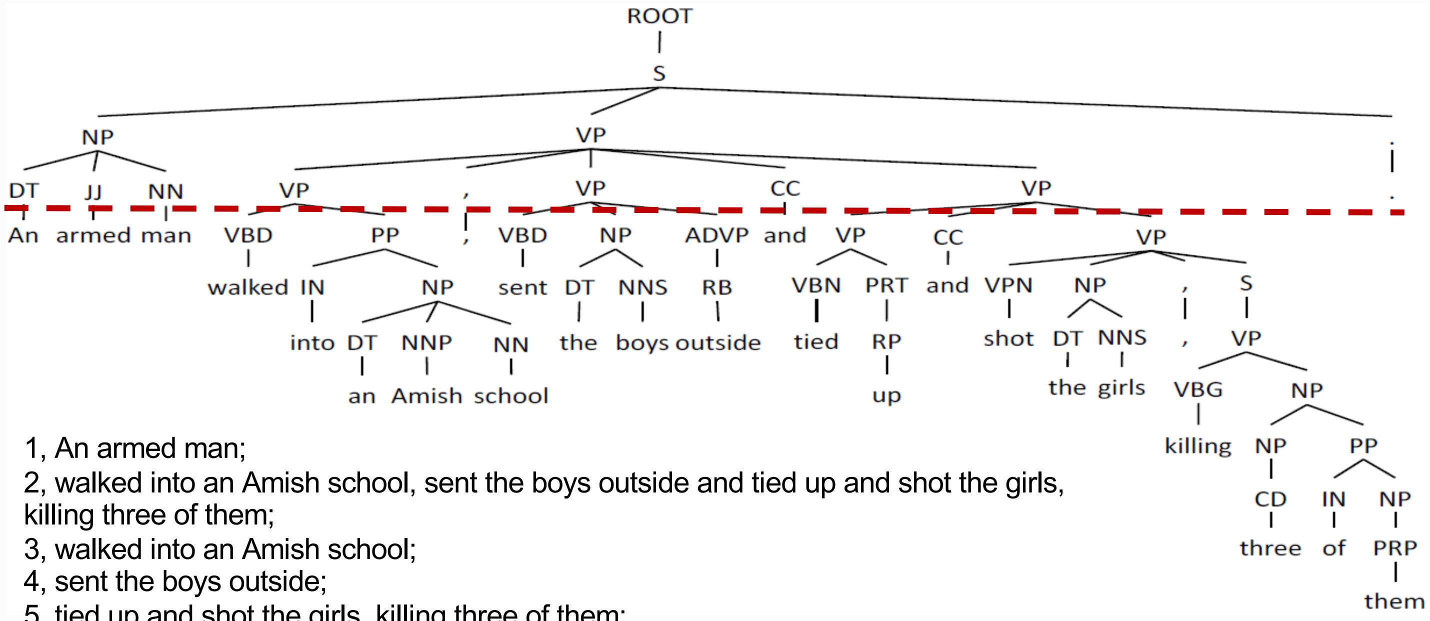
- This idea is based on two observations
  - The major constituent phrases loosely correspond to the concepts and facts.
  - The summary writer re-organizes the key concepts and facts to form new sentences for the summary

Concept	+	Fact
an armed man		walked into an Amish school

- Two components
  - Phrase extraction and salience calculation
  - Summary generation

Methods – Phrase Extraction

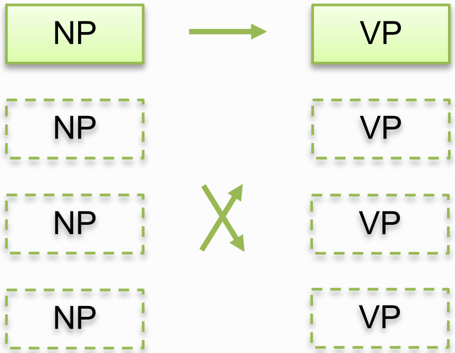
- Employ Stanford parser to obtain a constituency tree for each sentence
- Extract noun phrases (NPs) and verb-object phrases (VPs)
  - Recursively extract NPs and VPs until level 2



1, An armed man;
2, walked into an Amish school, sent the boys outside and tied up and shot the girls, killing three of them;
3, walked into an Amish school;
4, sent the boys outside;
5, tied up and shot the girls, killing three of them;

Methods – Summary Generation

- Compatibility Relation
  - To indicate whether an NP and a VP can be merged and form a new sentence.
  - Find alternatives for NP and VP



$$\tilde{\Gamma}[p,q] = \begin{cases} 1 & \text{if } N_p \in \tilde{\mathbf{N}}_i \wedge \Gamma[i,q] = 1 \\ 1 & \text{if } V_q \in \tilde{\mathbf{V}}_j \wedge \Gamma[p,j] = 1 \\ 1 & \text{if } \Gamma[p,q] = 1 \\ 0 & \text{otherwise} \end{cases}$$

  - Coreference resolution for NPs
  - Jaccard Index as the similarity measure for VPs, with a threshold value

Methods – Summary Generation

- Phrase-based content optimization
  - Jointly considering salience and diversity with **Integer Linear Programming (ILP)** framework

$$\max\{\sum_i \alpha_i S_i^N - \sum_{i<j} \alpha_{ij}(S_i^N + S_j^N)R_{ij}^N + \sum_i \beta_i S_i^V - \sum_{i<j} \beta_{ij}(S_i^V + S_j^V)R_{ij}^V\}$$

Salience

Diversity

  - α_i and β_i are selection indicators for NP and VP
  - S_i^N and S_i^V are the salience scores
  - R_{ij}^N and R_{ij}^V are the similarity of NP pairs and VP pairs

- *The phrase selection is governed by a set of constraints, to generate valid sentences*

Methods – Summary Generation

- Sentence generation constraints
  - NP validity, $\tilde{\gamma}_{ij}$ indicates whether both N_i and V_j are selected to construct a new sentence

$$\forall i, j, \alpha_i \geq \tilde{\gamma}_{ij}; \;\; \forall i, \sum_j \tilde{\gamma}_{ij} \geq \alpha_i.$$

  - VP legality

$$\forall j, \sum_i \tilde{\gamma}_{ij} = \beta_j$$

  - Not i-within-i

$$\begin{aligned} &\text{if } \exists V_k \rightsquigarrow V_j, \text{then } \beta_k + \beta_j \leq 1, \\ &\text{if } \exists N_k \rightsquigarrow N_j, \text{then } \alpha_k + \alpha_j \leq 1 \end{aligned}$$

Methods – Summary Generation

- Sentence generation constraints
  - Sentence number: avoid of too many short sentences

$$\sum_i \alpha_i \leq K$$

  - Length: summary length

$$\sum_i \{l(N_i) * \alpha_i\} + \sum_j \{l(V_j) * \beta_j\} \leq L$$

  - Phrase co-occurrence
  - Short sentence avoidance
  - Pronoun avoidance

Experiments – Case Study

- Sentence 1,2,4 are compressive, 3 is a new abstractive one.

[1:C] {An armed man (25)} {walked into an Amish school (25)} {tied up and shot the girls, killing three of them. (25)} [2:C] {A man who laid siege to a one-room Amish schoolhouse (64)} {told his wife shortly before opening fire that he had molested two young girls who were his relatives decades ago (64)} {was tormented by dreams of molesting again. (64)} [3:N] {Charles Carl Roberts IV (84)} {killed himself as police stormed the building (85)} {left what they described as rambling notes for his family. (150)} [4:C] {The gunman (145)} {was not Amish (145)} {had not attended the school. (145)} [5:O] {The shootings (148)} {occurred about 10:45 a.m. (148)} [6:O] {Police (149)} {could offer no explanation for the killings. (149)}

(25): (NP An armed man) (VP (VP walked into an Amish school), ~~(VP sent the boys outside)~~ and (VP tied up and shot the girls, killing three of them)), ~~(NP authorities) (VP said)~~.
(64): (NP (NP A man) who laid siege to a one-room Amish schoolhouse), ~~(VP killing five girls)~~, (VP (VP told his wife shortly before opening fire that he had molested two young girls who were his relatives decades ago) and (VP was tormented by "dreams of molesting again")), (NP ~~authorities) (VP said Tue)~~.
(~~145): According to media reports~~, (NP the gunman) (VP (VP was not Amish) and (VP had not attended the school)).

(84): On Monday morning, (NP Charles Carl Roberts IV) (VP (VP entered the West Nickel Mines Amish School in Lancaster County) and (VP shot 10 girls), (VP killing five)).
(85): (NP Roberts) (VP killed himself as police stormed the building).
(150): (NP Roberts) (VP left what they described as rambling notes for his family).

Sequence to Sequence Neural Network Model

Sequence to Sequence Learning with Neural Networks. NIPS 2014

- Abstractive summarization became the main stream after seq2seq NN models

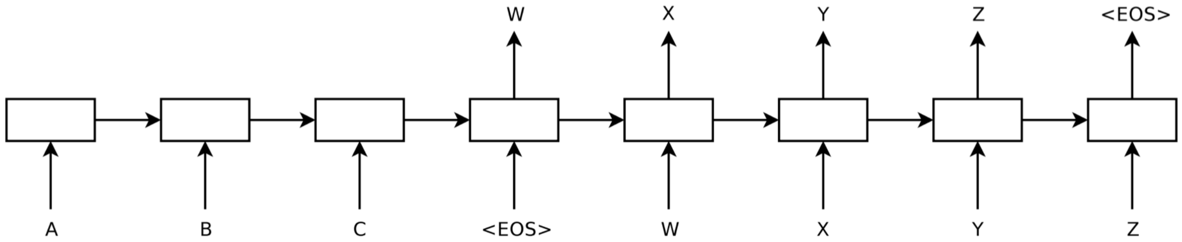
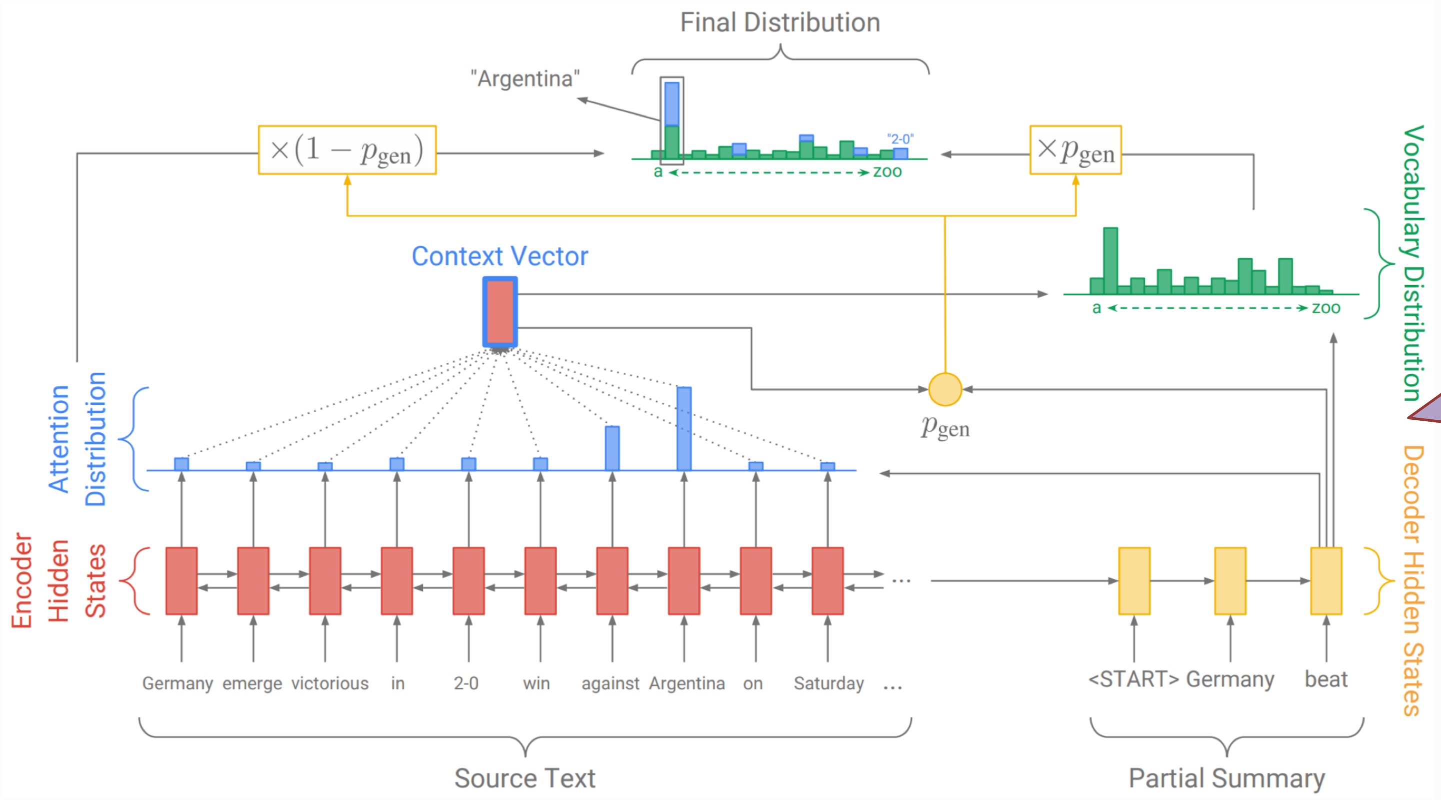


Figure 1: Our model reads an input sentence “ABC” and produces “WXYZ” as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

Pointer-Generator Networks for Abstractive Summarization

Get To The Point: Summarization with Pointer-Generator Networks. ACL 2017



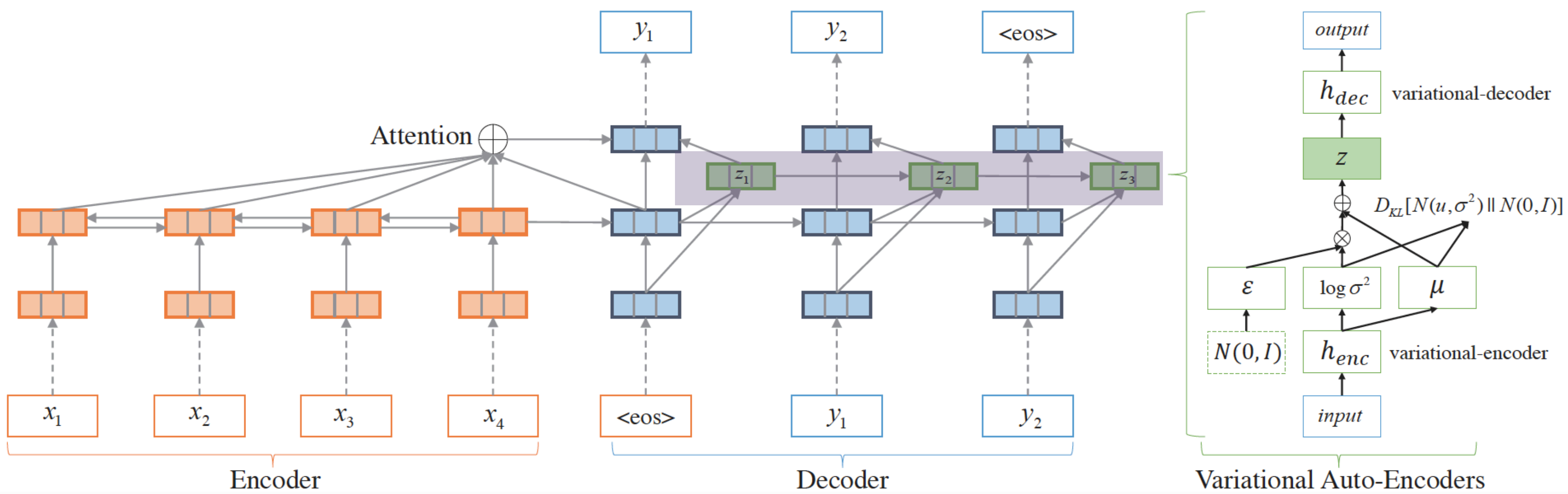
Need large training data.

Basically from a single input document to a summary.

Modelling Latent Structures for Headline Generation

- Headline generation: give the first few sentences of a news, to generate its headline
- Latent structures in human-written summaries/headlines
  - "Apple sues Qualcomm for nearly $1 billion": Who (Apple) Action (sues) What (Qualcomm for …)
  - "St. Louis' public library computers hacked": What-Happened
- In typical Seq2seq discriminative models (Lopyrev, 2015; Rush et al., 2015; Nallapati et al., 2016)
  - The calculation of decoding states is deterministic.
  - Such deterministic transformation limits the representation ability of the latent structure.

Deep Recurrent Generative Decoder (DRGD)



- DRGD captures the latent structure within the encoder-decoder architecture.
- DRGD adds historical dependencies on the latent variables of VAEs.

Text Summarization by Considering Reader Interests

Table 1: Examples of the text summarization. The text in red denotes the focused aspect by the good summary, while the text in blue is described by the bad summary. The text with underline is the focused aspect by reader comments.

document	On August 28, according to a person familiar with the matter, Toyota Motor Corporation will invest 500 million U.S. dollars into the Uber, a taxi service company, with a valuation of up to 72 billion U.S. dollars. The investment will focus on driverless car technology. However, its development path is not smooth. In March of this year, a Uber driverless car hit a woman and caused her death. In last year, Softbank also invested into Uber with a valuation of $48 billion.
comments	Toyota's investment in Uber is a wise choice.
	$500 million investment is really a lot of money!
good summary	Toyota invests $500 million into Uber with a valuation of $72 billion
bad summary	An Uber driverless car hits a passerby to death

- Reader-aware multi-document summarization: An enhanced model and the first dataset. Workshop on New Frontiers in Summarization 2017.
- Abstractive Text Summarization by Incorporating Reader Comments. AAAI 2019

Types of QA Tasks

- Similar question retrieval: more on question side
  - Community QA
  - Frequent AQ
- Reading comprehension: more on answer side, answer selection or extraction
  - Extraction based, such as SQuAD
  - Multiple choice, such as RACE
- Knowledge-based QA
  - KB QA
  - Product QA, using product key attributes, description, review, previous questions
- And others, such as visual QA

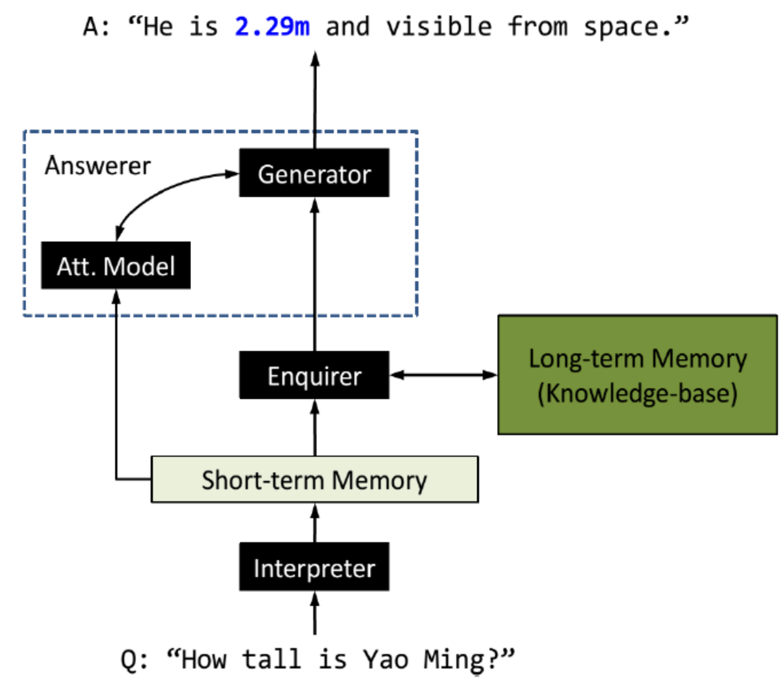
Answer Generation: GENQA

Neural Generative Question Answering. IJCAI 16

- Generate answers to simple factoid questions by accessing a knowledge base
  - Enquire a knowledge-base to retrieve the triples related to the question
  - The decoder generates a common word (e.g., is) or a term (e.g., "John Malkovich") retrieved from KB

Question & Answer	Triple (*subject*, *predicate*, *object*)
Q: *How tall is Yao Ming?* A: *He is 2.29m and is visible from space.*	`(Yao Ming, height, 2.29m)`
Q: *Which country was Beethoven from?* A: *He was born in what is now Germany.*	`(Ludwig van Beethoven, place of birth, Germany)`
Q: *Which club does Messi play for?* A: *Lionel Messi currently plays for FC Barcelona in the Spanish Primera Liga.*	`(Lionel Messi, team, FC Barcelon)`

Answer Generation: GENQA



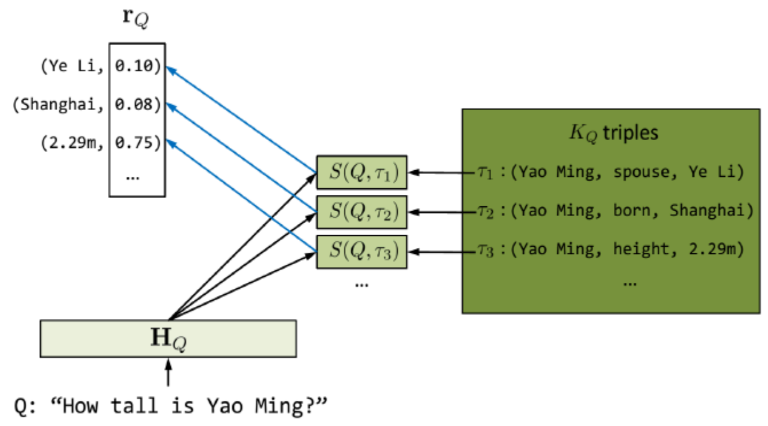


Figure 2: Enquirer of GENQA.

The enquirer **may not be able to get correct knowledge**, and the generator may still **make mistakes when selecting terms from long term memory**

Produce SPARQL from Natural Question

Template-based Question Answering over RDF Data. WWW 2012.

- Convert the natural question into a SPARQL query

Who produced the most films?

Generate? to enquire more precise knowledge

```
SELECT ?x WHERE {
   ?x ?p ?y .
   ?y rdf:type ?c .
}
ORDER BY DESC(COUNT(?y)) LIMIT 1 OFFSET 0
```

Slots:

- ⟨?c, class, films⟩
- ⟨?p, property, produced⟩

```
SELECT ?y WHERE {
   ?x rdf:type onto:Film .
   ?x onto:producer ?y .
}
ORDER BY DESC(COUNT(?x)) OFFSET 0 LIMIT 1
```

Note: Inconsistent symbols

Reading Comprehension of Multiple Choice Questions

RACE: Large-scale Reading Comprehension Dataset From Examinations. EMNLP 2017.

Article:

. . .

The Yanomami live along the rivers of the rainforest in the north of Brazil. They have lived in the rainforest for about 10,000 years and they use more than 2,000 different plants for food and for medicine. But in 1988, someone found gold in their forest, and suddenly 45,000 people came to the forest and began looking for gold. They cut down the forest to make roads. They made more than a hundred airports. The Yanomami people lost land and food. Many died because new diseases came to the forest with the strangers.

. . .

In 1987, they closed fifteen roads for eight months. No one cut down any trees during that time. In Panama, the Kuna people saved their forest. They made a forest park which tourists pay to visit. The Gavioes people of Brazil use the forest, but they protect it as well. They find and sell the Brazil nuts which grow on the forest trees.

Question:

Those people built roads and airports in order to _ .

A. carry away the gold conveniently (**Answer**)

B. make people there live a better life (**Distractor**)

C. stop spreading the new diseases (**Distractor**)

D. develop the tourism there (**Distractor**)

- SOTA system's performance: 59% (by Dec 2018)
- Some questions could be really difficult, e.g.:
  - "what's the best title for this article?"
  - "what do you learn from this article?"

Distractor Generation

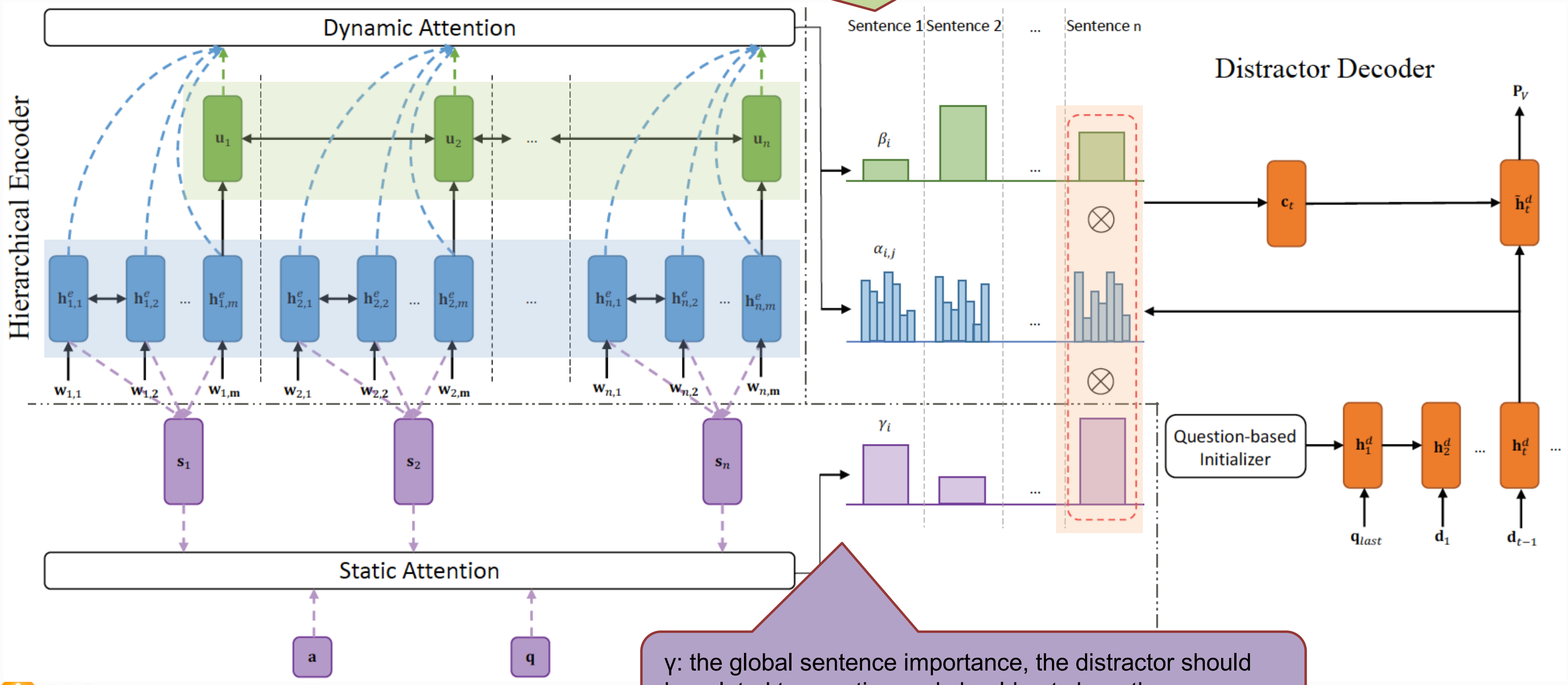
- We proposed ***a new task: distractor generation for MCQ***
  - Given an article, a pair of question and its correct option
  - The goal is to generate context and question related, grammatically consistent wrong options, i.e. distractor

- Motivations
  - aid the preparation of MCQ reading comprehension datasets
  - helpful to alleviate instructors' workload in designing MCQs for students.

- Should it be useful for boosting the performance of MCQ RC systems?
  - We did observe improvement by applying generated question-answer pairs to train models to solve SQuAD questions. (Yang et al. 2017)

Hierarchical Encoder-Decoder Framework

β: emphasize the sentence that is important for generating the current distractor
α: word importance in each sentence



γ: the global sentence importance, the distractor should be related to question and should not share the same semantic meaning with the correct answer.

Dataset

- RACE (Lai et al. 2017)
  - 27,933 articles from English examinations of Chinese students from grade 7 to12.
- Exclude trivial distractors, having no semantic relevance with the article

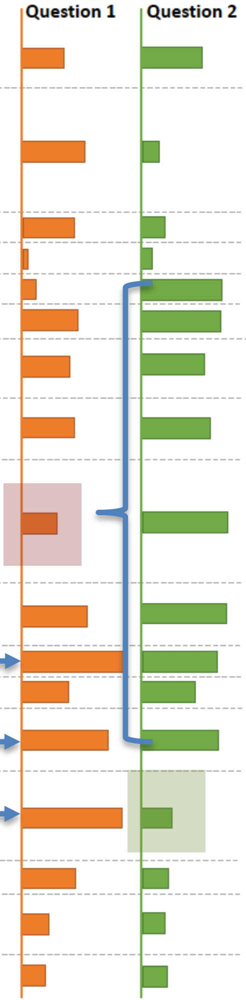
# Train Samples	96501
# Dev Samples	12089
# Test Samples	12284
Avg. article length (tokens)	347.0
Avg. distractor length	8.5
Avg. question length	9.9
Avg. answer length	8.7
Avg. # distractors per question	2.1

Dataset: https://github.com/Evan-Gao/Distractor-Generation-RACE

Define Smarter Tomorrow.

Article:

1. Dear friends, The recent success of children's books has made the general public aware that there's a huge market out there.
2. And there's a growing need for new writers trained to create the $3 billion worth of children's books bought each year... plus stories and articles needed by over 650 publishers of magazines for children and teenagers.
3. Who are these needed writers?
4. They're ordinary people like you and me.
5. But am I good enough?
6. I was once where you might be now.
7. My thoughts of writing had been pushed down by self-doubt, and I didn't know where to turn for help.
8. Then, I accepted a free offer from the Institute to test my writing ability, and it turned out to be the inspiration I needed.
9. The promise that paid off The Institute made the same promise to me that they will make to you, if you show basic writing ability: you will complete at least one manuscript suitable to hand in to a publisher by the time you finish our course.
10. I really didn't expect any publication before I finished the course, but that happened.
11. I sold three stories.
12. And I soon discovered that was not unusual at the Institute.
13. Since graduation, I have written 34 nationally published children's books and over 300 stories and articles.
14. Free test and brochure We offer a free ability test and will send you a copy of our brochure describing our recognized home-study courses on the basis of one-on-one training.
15. Realize your writing dream today.
16. There's nothing sadder than a dream delayed until it disappears forever.
17. Sincerely, Kristi Hill Institute of Children's Literature



Static Attention Distribution

Question 1: You are promised to publish one manuscript when you _ .

Options:

A. show basic ability B. finish the course

C. have sold three stories D. have passed the test

Seq2Seq:

1. have made a mistake
2. have written a lot of books
3. have been writing a newspaper

HRED:

1. have finished the course
2. have a free test
3. have been opened

Our Model:

1. have sold three stories
2. write a book
3. have passed the test

Question 2: Why does Kristi Hill mention her own experience of attending the courses?

Options:

A. To introduce the home-study courses.

B. To show she has realized her dream.

C. To prove she is a qualified writer.

D. To promote the writing program.

Seq2Seq:

1. To show she is a successful publisher.
2. To show how inspiring her books are.
3. To show her interest in writing books.

HRED:

1. To encourage readers to buy more books.
2. To show she wanted to improve her reading skills.
3. To prove she is a well-known courses publisher.

Our Model:

1. To prove she is a qualified writer.
2. To show her great achievements in literature.
3. To encourage readers to be interested in writing.

Question Generation

- Generate natural language questions from different forms of content, e.g. knowledge triples, tables, sentences, or images

- The generated question should be answerable with the given content as clue

- Applications:
  - convert customized contents of specific domain into Q-A pairs, used in domain QA or service bot
  - generate rich Q-A pairs to help human construction of QA dataset.
  - applying generated QA pairs to enhance models to solve SQuAD questions.

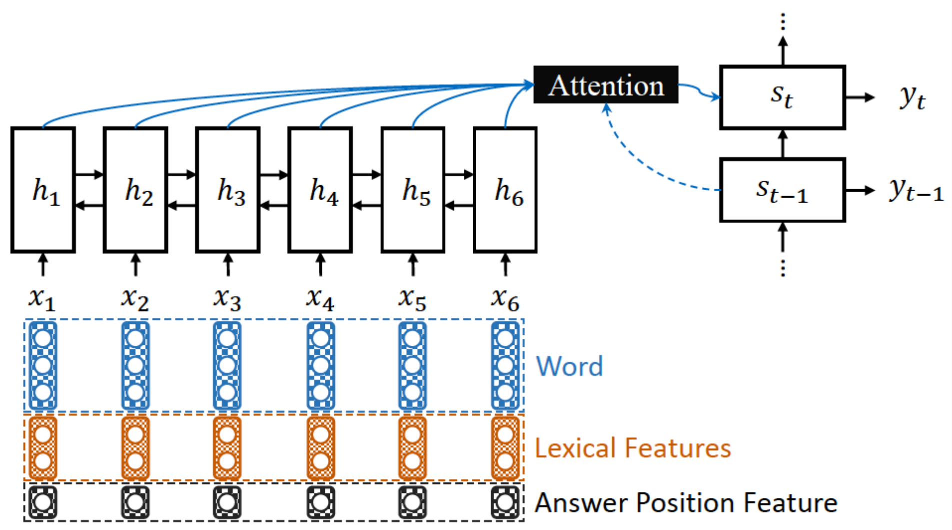
Question Generation from Free Text

[1] Learning to Ask: Neural Question Generation for Reading Comprehension. ACL 2017
[2] Neural Question Generation from Text: A Preliminary Study. arXiv:1704.01792

- Generate SQuAD-like question

- Formulated as a seq2seq problem [1]
  - Given an input sentence to generate a natural question related to information in the sentence

- Consider answer information [2]
  - BIO tags to denote answer
  - Lexical feature: POS and NER



Difficulty Controllable Question Generation

- SQuAD questions have different difficulty levels
  - Q1 and Q2 are easy
  - Q3 is hard
- Can we control the difficulty of generated questions?

S1: Oxygen is a chemical element with symbol O and atomic number 8.
Q1: (*Easy*) What is the atomic number of the element oxygen?
A1: 8

S2: It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements.
Q2: (*Easy*) Of what group in the periodic table is oxygen a member?
A2: chalcogen

S3: The electric guitar is often emphasised, used with distortion and other effects, both as a rhythm instrument using repetitive riffs with a varying degree of complexity, and as a solo lead instrument.
Q3: (*Hard*) What instrument is usually at the center of a hard rock sound?
A3: The electric guitar

Difficulty Controllable Question Generation for Reading Comprehension. arXiv:1807.03586

Difficulty Controllable Question Generation

- The task
  - Given a sentence, a text fragment (answer) in the sentence, and **a difficulty level**
  - To generate a question that is asked about the fragment and satisfy the difficulty levels

- Challenges
  - no existing QA dataset has difficulty labels for questions.
  - for a single sentence and answer pair, we want to generate questions with diverse difficulty levels, but SQuAD only has one given question for each sentence and answer pair
  - no metric to evaluate the difficulty of questions.

Data Preparation

- Two difficulty levels: **Hard** and **Easy**

- Automatic labelling protocol for data preparation
  - employ two RC systems, namely R-Net (Wang et al., 2017) and BiDAF (Seo et al., 2017)
  - A question would be:
    - labelled with Easy if both R-Net and BiDAF answer it correctly under the exact match metric
    - labelled with Hard if both systems fail to answer it.
  - The remaining questions are eliminated for suppressing the ambiguity.

	Train	Dev	Test
# easy questions	34,813	4,973	4,937
# hard questions	24,317	3,573	3,442
Easy ratio	58.88%	58.19%	58.92%

Define Smarter Tomorrow.

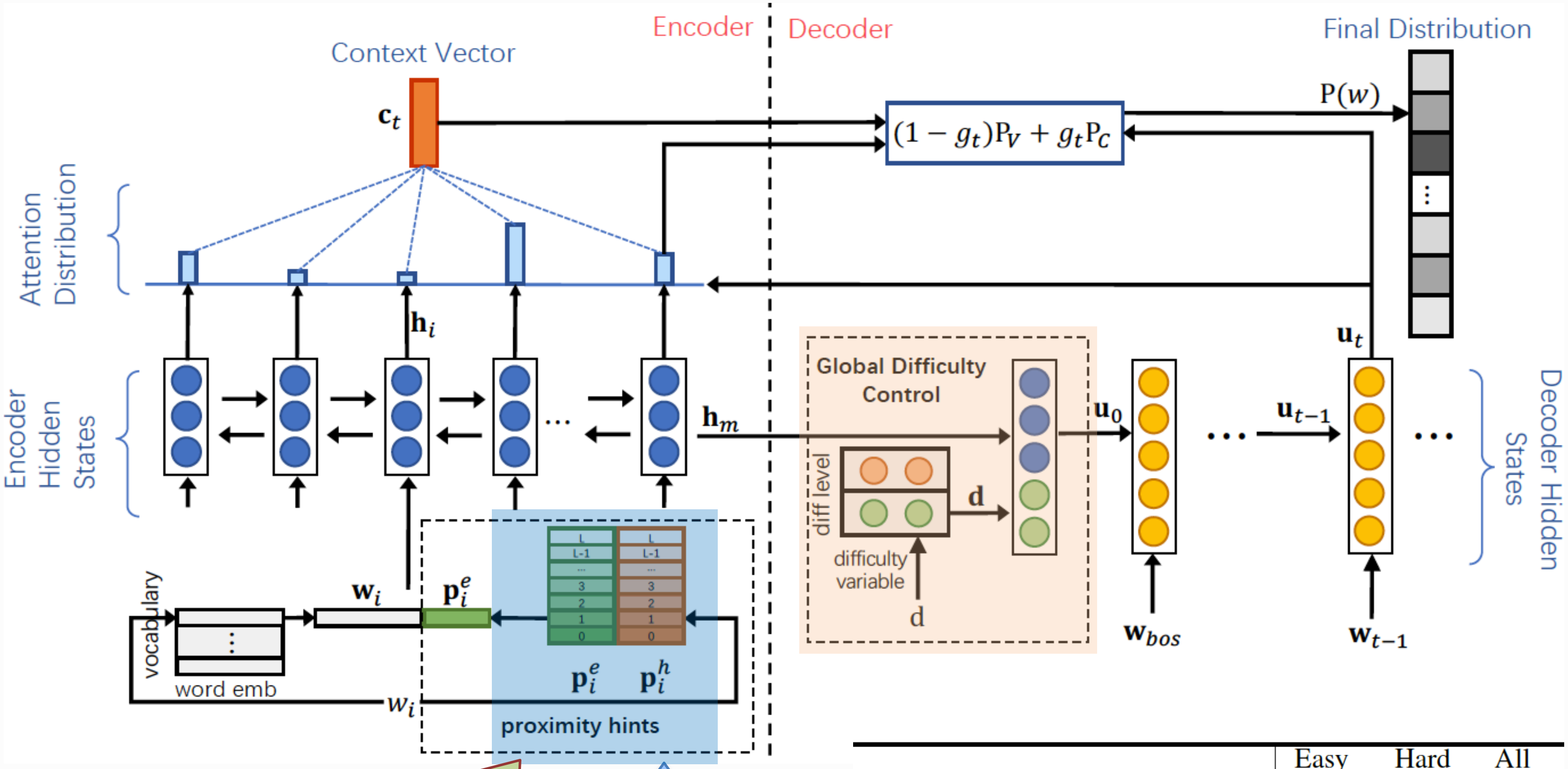
Reasonability of Labelling Protocol

- Sugawara et al. (2017) manually labelled 100 questions by analyzing the skills needed to answer a question correctly.

	# examples		# AVG skill	
	Easy	Hard	Easy	Hard
BiDAF	64	36	1.20	1.44
R-Net	70	30	1.11	1.7
Our protocol	55	21	1.16	1.81

Note that for BiDAF and R-Net, the total number of questions is 100, but for our protocol, 76 questions are left after filtering out the ambiguous ones.

Our Framework for DQG



	Easy	Hard	All
Avg. distance of question words	7.67	9.71	8.43
Avg. distance of all sentence words	11.23	11.16	11.20

QWPH: Only capture the position info
DLPH: Also consider the difficulty levels

Evaluation

- **Difficulty of the generated questions**. For easy questions, higher score indicates better difficulty-control, while for hard questions, lower indicates better.

	Easy Questions Set				**Hard** Questions Set			
	R-Net EM	R-Net F1	BiDAF EM	BiDAF F1	R-Net EM	R-Net F1	BiDAF EM	BiDAF F1
Ans	82.16	87.22	75.43	83.17	34.15	60.07	29.36	55.89
QWPH	82.66	87.37	76.10	83.90	33.35	59.50	28.40	55.21
QWPH-GDC	84.35	88.86	77.23	84.78	31.60	57.88	26.68	54.31
DLPH	85.49	89.50	78.35	85.34	28.05	54.21	24.89	51.25
DLPH-GDC	**85.82**	**89.69**	**79.09**	**85.72**	**26.71**	**53.40**	**24.47**	**51.20**

- **The results of controlling difficulty**. The scores are performance gap between questions generated with original difficulty label and questions generated with reverse difficulty label.

	Easy Questions Set				**Hard** Questions Set			
	R-Net EM	R-Net F1	BiDAF EM	BiDAF F1	R-Net EM	R-Net F1	BiDAF EM	BiDAF F1
QWPH-GDC	7.41	5.72	7.13	5.88	6.45	5.47	6.13	5.10
DLPH	12.41	9.51	11.28	8.49	12.01	10.45	10.51	9.37
DLPH-GDC	**12.91**	**9.95**	**12.40**	**9.23**	**12.68**	**10.76**	**11.22**	**9.97**

Case Study

- Our model
  - give more hints (shorter distance) when asking easier questions
  - give less hints (longer distance) when asking harder questions.

Input 1: prajñā is the wisdom that is able to extinguish afflictions and bring about bodhi . (*Easy Question)*
Human: (4.5) prajna is the wisom that is able to extinguish afflictions and bring about what ?
Ans: (13.0) what is prajñā ?
DLPH-GDC: (6.2) prajñā is able to extinguish afflictions and bring about what ?
DLPH-GDC (reverse): (7.3) what is prajñā able to bring ?

Input 2: the electric guitar is often emphasised , used with distortion and other effects , both as a rhythm instrument using repetitive riffs with a varying degree of complexity , and as a solo lead instrument . (*Hard Question)*
Human: (16.0) what instrument is usually at the center of a hard rock sound ?
Ans: (5.5) what is often emphasised with distortion and other effects ?
DLPH-GDC: (25.7) what is a solo lead instrument ?
DLPH-GDC (reverse): (2.5) what is often emphasised ?

"(xx.x)": Avg. distance of question words to answers, i.e. proximity hints

Text Editing: Style Transfer

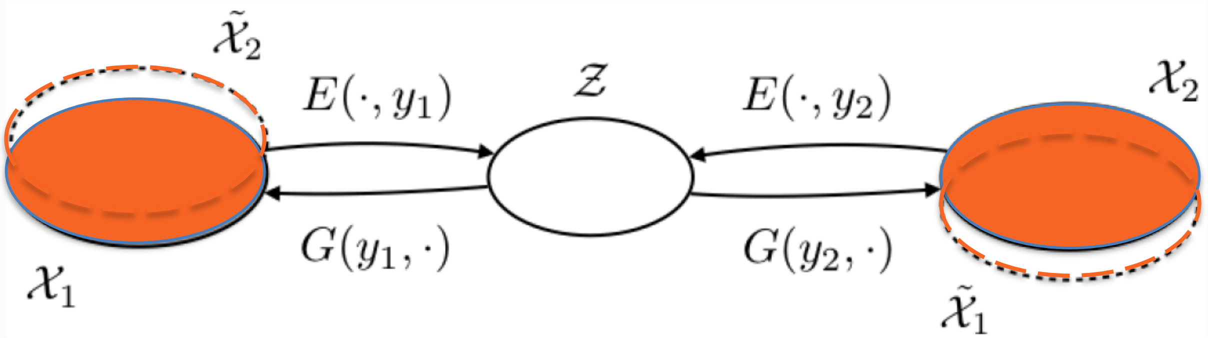
- Text editing:
  - ASR smoothing
  - MT postediting
  - grammar correction
- Text Style
  - Modern vs Shakespeare
  - Positive vs negative
  - Formal vs informal (e.g. News vs tweets)
  - Chinese in Taiwan vs Mainland China
  - Chinglish vs Singlish vs English

Style Transfer from Non-Parallel Text

Style Transfer from Non-Parallel Text by Cross-Alignment. NIPS 2017

- Setting
  - access to two corpora of sentences with the same distribution of content, but rendered in different styles (sentiment polarities).

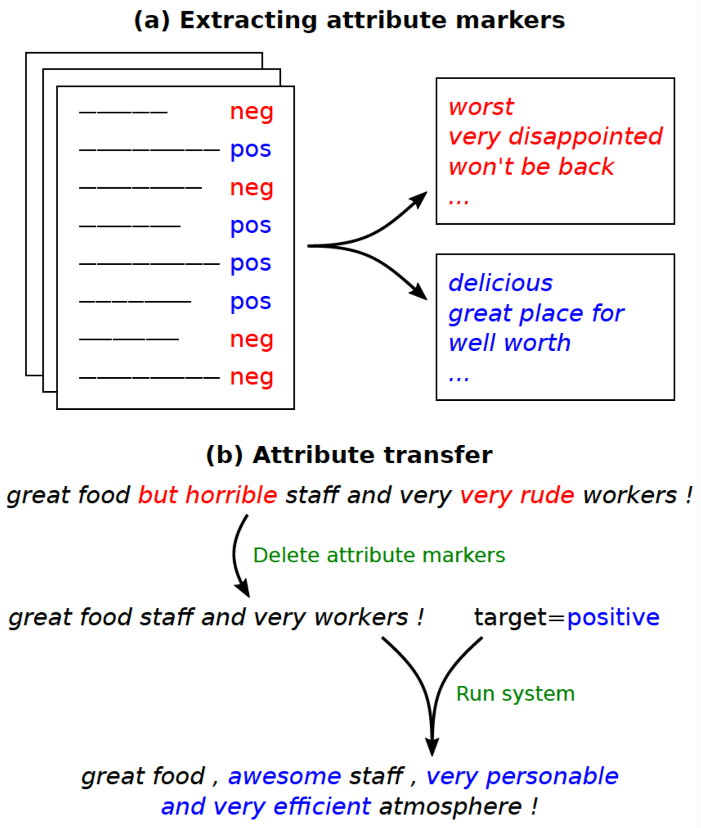
- Approach
  - map a sentence in one style to a style-independent content vector
  - then decode it to a sentence with the same content but a different style.



Delete, Retrieve, Generate

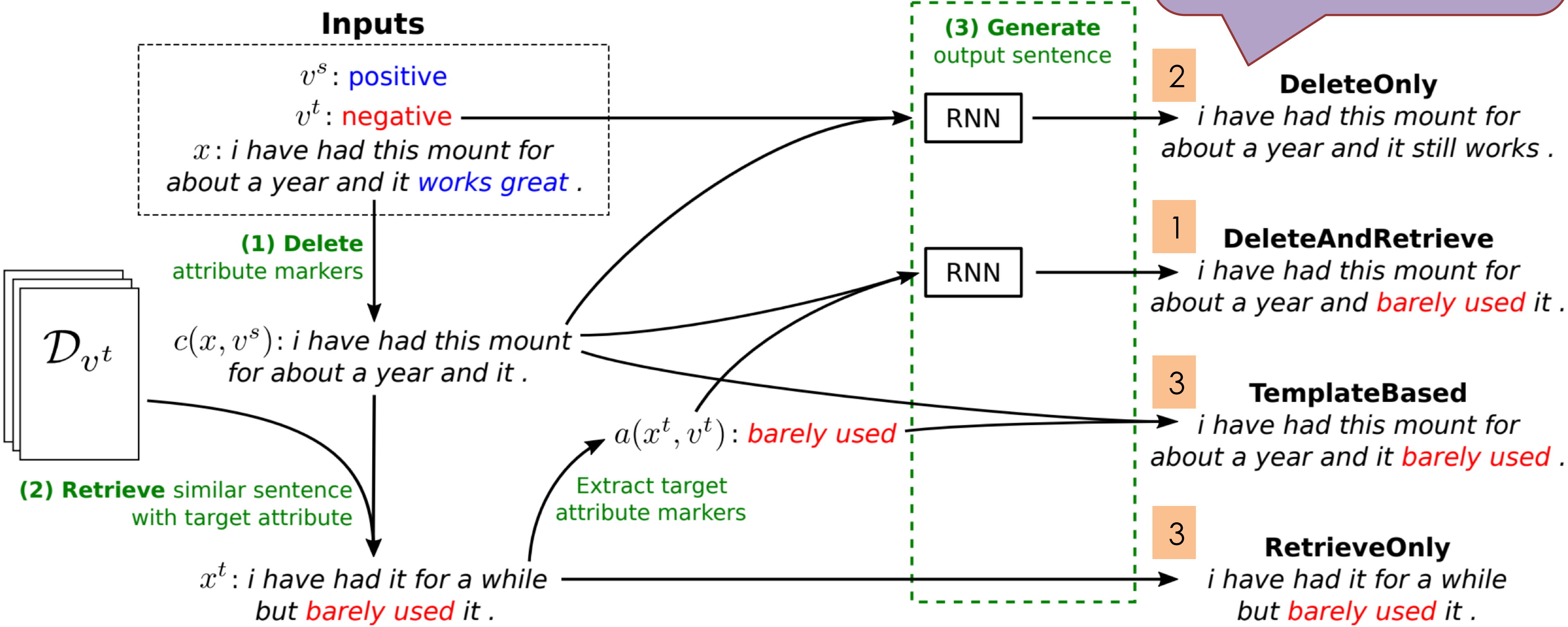
Delete, Retrieve, Generate- A Simple Approach to Sentiment and Style Transfer. NAACL-HLT 2018

- First, identify attribute markers from an unaligned corpus.
- Remove markers of the original attribute.
- Generate a new sentence conditioned on the remaining words and the target attribute.
- Four "Run systems", next page.



Delete, Retrieve, Generate

DeleteAndRetrieve is the best among the four methods on all three datasets, evaluated from grammar, content preservation, and target attribute.



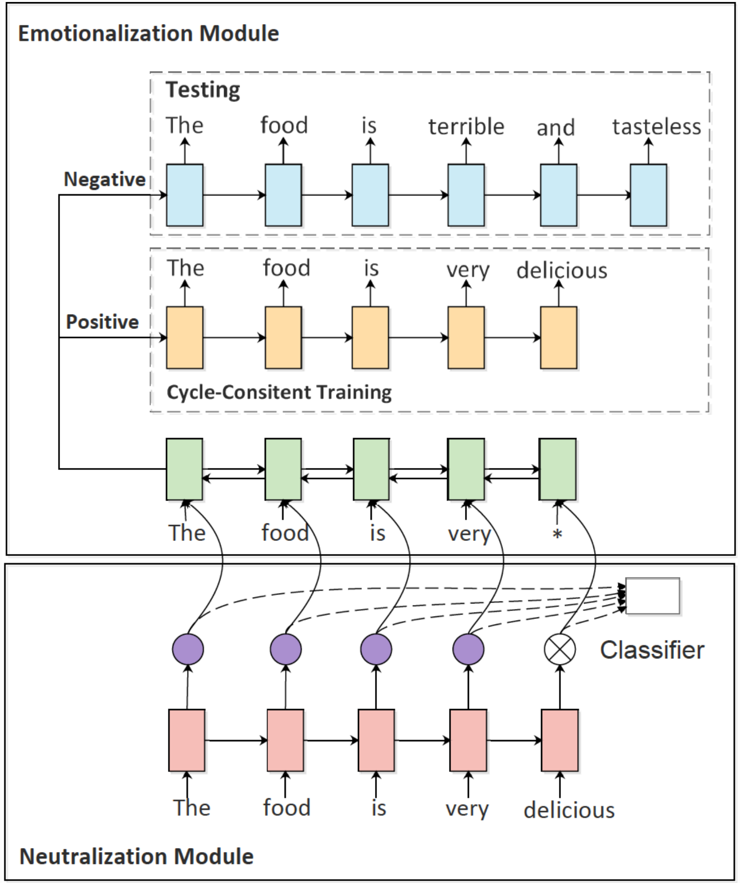
Unpaired Sentiment-to-Sentiment Translation

Unpaired Sentiment-to-Sentiment Translation: A Cycled Reinforcement. ACL 2018

- Motivation: keep content better

- The **neutralization** filters out emotional words.
- The **emotionalization** adds sentiment to the neutralized semantic content for sentiment-to-sentiment translation.

- **RL to reward** the neutralization module based on the feedback from the emotionalization module. Rewards:
  - whether the generated text matches the target sentiment;
  - evaluating the content preservation performance.



Quantifiable Sequence Editing (QuaSE)

- In this setting, a sentence is associated with a numerical outcome value measuring a certain property of the sentence.
  - outcome of a review is its rating
  - outcome of an advertisement is its degree of attractiveness

Example	
"The food is terrible"	–outcome(rating): 1
"The food is OK"	–outcome(rating): 3
"The food is extremely delicious"	–outcome(rating): 5

Define Smarter Tomorrow.

Problem Definition of QuaSE

- Input:
  - an original sentence X_0 with outcome R_0
  - a target outcome R^*

- Output:
  - A sentence X^* with outcome R^*, by revising X_0

Example

Input:

- X_0 = "The food is terrible" with R_0=1.

Output:

- when R^* is set to 5, X^* = "The food is extremely delicious"
- when R^* is set to 3, X^* = "The food is OK"

Characteristics and Challenges

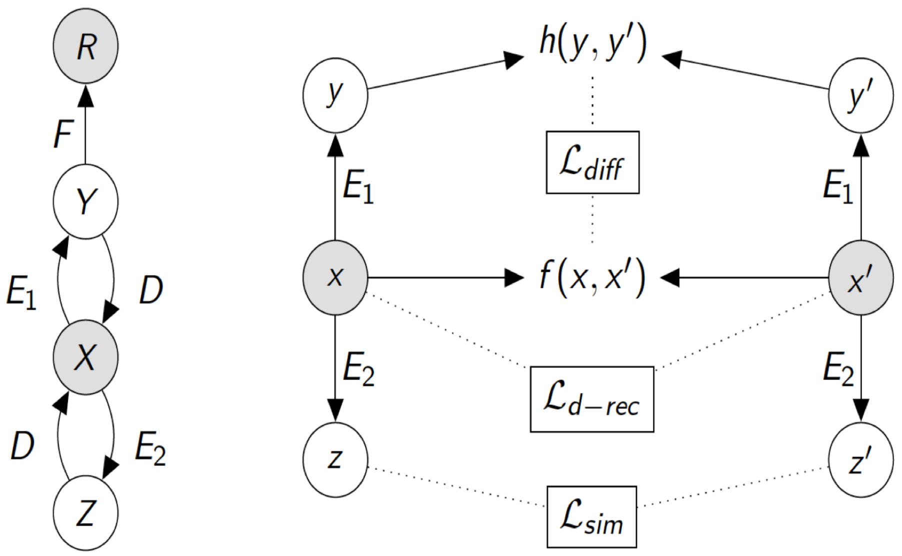
- The outcome here is **numerical** (continuous), other than categorical (discrete).
  - it is impossible to construct two corpora as counterpart of each other
- The editing operation is **under a quantifiable guidance** (target outcome).
  - When different target outcomes are specified, the corresponding generated sentences are different

- Challenges
  - A model should be able to perceive the association between an outcome and its relevant wordings.
  - When performing editing, the model should keep the content, and only edits the outcome-related wordings.
  - We do not have parallel data as in NMT. i.e. [original sentence, target sentence]

QuaSE Model

- **Modelling single sentence** to disentangle a sentence into two latent factors
  - Y : Outcome variable: captures the outcome properties of the sentence
  - Z: Content variable: captures the content of the sentence

- **Exploiting pseudo-parallel sentences** to further enhance the disentanglement, and the dual reconstruction structure further enhances the capability of generating expected output



Modelling Single Sentence

- Encode each sentence *X* into outcome variable *Y* and a content variable *Z* via VAE
- Losses
  - L_{rec}: reconstruction loss, mimic a seq2seq model where both ends are the sentence itself.

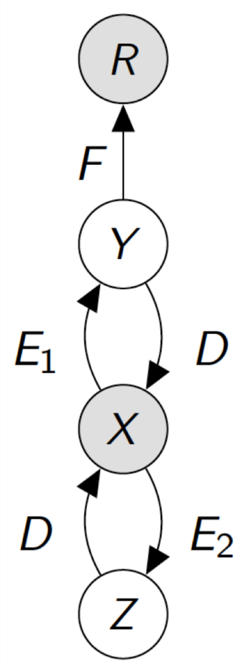
$$\mathcal{L}_{rec} = -\mathbb{E}_{Y,Z\sim q(Y|X),q(Z|X)}[\log p(X|Y,Z)]$$
$$= H(X, D(E_1(X), E_2(X)))$$

  - L_{kl}: KL-divergence between the posterior distribution and prior distribution, viewed as a regularization term avoiding overfitting

$$\mathcal{L}_{kl} = KL[q(Y|X)|p(Y)] + KL[q(Z|X)|p(Z)]$$

  - L_{mse}: we use a full-connected neural network F to project Y to R.

$$\mathcal{L}_{mse} = (R - F(Y))^2 = (R - F(E_1(X)))^2$$



Modelling Pseudo-Parallel Sentences

- Pseudo-Parallel Sentences: A pair of sentences
  - with high overlapping of wordings (e.g. Jaccard Index > 0.7)
  - with substantially different outcomes

As a result, two sentences in a pair have the same or very similar content, but different outcome values.

x	I will never come back to the restaurant.
x'	I will definitely come back to the restaurant, recommend!

Modelling Pseudo-Parallel Sentences

- Wording difference vs outcome different
  - Wording difference. E.g. inc(x; x'): embedding of "definitely" or "recommend", dec(x; x'): embedding of "never"

    $$f(x, x') = inc(x, x') \oplus dec(x, x')$$

  - Outcome difference

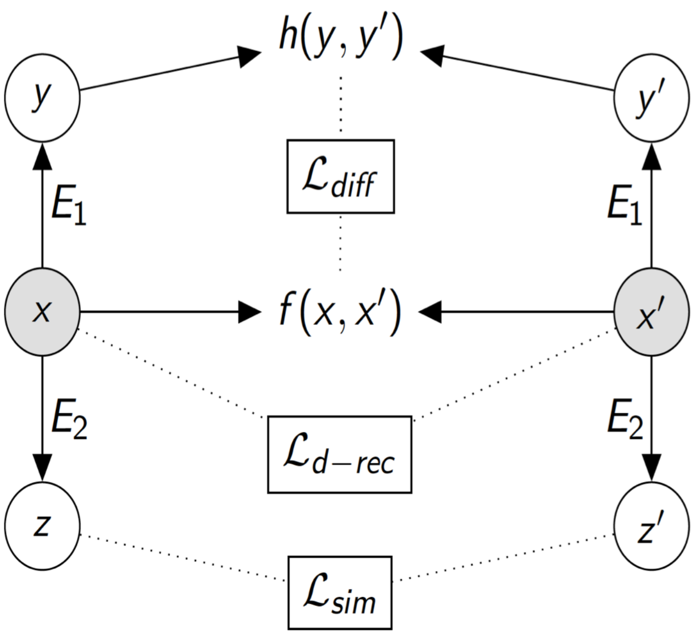
    $$h(y, y') = y - y' = E_1(x) - E_1(x')$$

  - Loss (U is a regression network)

    $$\mathcal{L}_{diff} = ||h(y, y') - U[f(x, x')]||^2$$

- The content of pseudo-parallel sentences are similar.

  $$\mathcal{L}_{sim} = ||z - z'||^2 = ||E_2(x) - E_2(x')||^2$$



Modelling Pseudo-Parallel Sentences

- Since x shares similar content with x', its content factor z, when combined with the outcome factor y' of x', should nearly reconstruct x'.

$$\begin{aligned}\mathcal{L}^{d-rec}_{x';x} &= H(x', D(E_1(x'), E_2(x))) \\ &= H(x', D(y', z))\end{aligned}$$

$$\mathcal{L}_{d-rec} = \mathcal{L}^{d-rec}_{x';x} + \mathcal{L}^{d-rec}_{x;x'}$$

- Joint training

$$\begin{aligned}\mathcal{L}_{joint} =& \lambda_{rec}\mathcal{L}_{rec} + \lambda_{kl}\mathcal{L}_{kl} + \lambda_{mse}\mathcal{L}_{mse} + \\ & \lambda_{diff}\mathcal{L}_{diff} + \lambda_{sim}\mathcal{L}_{sim} + \lambda_{d-rec}\mathcal{L}_{d-rec}\end{aligned}$$

Editing under Quantifiable Guidance

General Strategy

Given a input sentence X_0 and target outcome R^*:

- Use encoders E_1 and E_2 to obtain Y_0 and Z_0.
- Modifies the outcome variable Y_0 to Y^*, such that the outcome associated with Y^* is R^*
- Fix the content variable Z_0
- Reconstruct the revised sentence by feeding the decoder D with Y^* and Z_0

How to determine Y^*

- Y follows the Gaussian distribution $Y \sim \mathcal{G}(Y_0 = E_1(X_0), \sigma)$
- Choose $\mathcal{C} = \{Y : \mathcal{G}(Y|E_1(X_0), \sigma) > \tau\}$ as the feasible range for Y^*, where τ is a threshold.
- Finally, Y^* is determined as follows: $Y^* = \arg\min_{Y \in \mathcal{C}}(F(Y) - R^*)^2$

Results of QuaSE on the Yelp dataset

	MAE					Edit Distance				
	T=1	T=2	T=3	T=4	T=5	T=1	T=2	T=3	T=4	T=5
Original	2.218	1.237	0.825	0.927	1.781	N/A	N/A	N/A	N/A	N/A
S2BS	1.683	0.944	0.756	0.757	1.302	6.643	5.342	4.939	5.005	6.229
Our Model	1.416	0.629	0.740	0.537	0.940	7.919	4.7	3.450	4.13	8.009

- T =1,2,3,4,5 refer to the target ratings
- Our model achieves smaller MAE values than S2BS
- Both models are able to handle QuaSE, by comparing with Original
- Relations between MAE and Edit Distance

Case Study & Human Evaluation

	Sentence
E.g. 1	this tire center is amazing .
T=1	this tire center is horrible .
T=3	this tire center is really good .
T=5	this tire center is amazing .
E.g. 2	horrible food !
T=1	horrendous
T=3	their food amazing !
T=5	amazing delicious food ! recommend !
E.g. 3	decent food and wine selection , but nothing i will rush back for .
T=1	decent food and wine selection , but nothing i will rush for no .
T=3	decent food and wine selection , but i will never look back for .
T=5	decent food and wine selection , but excellent service, will return !
E.g. 4	our first time and we had a great meal , wonderful service .
T=1	our first time and we had a terrible meal , stale service .
T=3	our first time and we had a great meal , we have service .
T=5	our first time and we had a great meal , wonderful service .

	Content Preservation (Range: [0, 2])	Fluency (Range: [1, 4])
TST	1.02	2.56
S2BS	0.70	2.53
Our Model	1.38	2.48

More Works on Text Style Transfer

Paper list: https://github.com/fuzhenxin/Style-Transfer-in-Text

- Sequence to Better Sequence: Continuous Revision of Combinatorial Structures, ICML-2017, [paper], [code]
- Toward Controlled Generation of Text, ICML-2017, [paper], [official code], [unofficial code]
- Style Transfer from Non-Parallel Text by Cross-Alignment, NIPS-2017, [paper], [code]
- Zero-Shot Style Transfer in Text Using Recurrent Neural Networks, Arxiv-2017, [paper], [code]
- Style Transfer in Text: Exploration and Evaluation, AAAI-2018, [paper], [code]
- Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer, NAACL-2018, [paper], [code]
- SHAPED: Shared-Private Encoder-Decoder for Text Style Adaptation, NAACL-2018, [paper]
- Sentiment Transfer using Seq2Seq Adversarial Autoencoders, project for CSYE7245 Northeastern University, [paper]
- Style Transfer Through Back-Translation, ACL-2018, [paper], [code]
- Unpaired Sentiment-to-Sentiment Translation: A Cycled Reinforcement Learning Approach, ACL-2018, [paper], [code]
- Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer, ACL-2018, [paper]
- Unsupervised Text Style Transfer using Language Models as Discriminators, NIPS-2018, [paper]
- Disentangled Representation Learning for Text Style Transfer, Arxiv, [paper], [code]
- Language Style Transfer from Sentences with Arbitrary Unknown Styles, Arxiv, [paper]
- What is wrong with style transfer for texts? Arxiv, [paper]
- Style Transfer as Unsupervised Machine Translation, Arxiv, [paper]
- Learning Sentiment Memories for Sentiment Modification without Parallel Data, EMNLP-2018, [paper], [code]
- Style Transfer Through Multilingual and Feedback-Based Back-Translation, Arxiv, 2018, [paper]
- Structured Content Preservation for Unsupervised Text Style Transfer, OpenReview, 2018, [paper]
- Adversarially Regularized Autoencoders, ICML-2018, [paper], [code]
- Unsupervised Controllable Text Formalization, Arxiv, 2018, [paper]
- Large-scale Hierarchical Alignment for Author Style Transfer, Arxiv, 2018, [paper]
- Learning Criteria and Evaluation Metrics for Textual Transfer between Non-Parallel Corpora, Arxiv, 2018, [paper]
- Multiple-Attribute Text Style Transfer, Arxiv, 2018, [paper]
- Content preserving text generation with attribute controls, NIPS, 2018, [paper]
- Multiple-Attribute Text Rewriting, ICLR, 2019, [paper]
- Iterative Matching and Translation for Non-Parallel Style Transfer, UNK, 2018, [paper]
- QuaSE: Sequence Editing under Quantifiable Guidance, EMNLP, 2018, [paper]

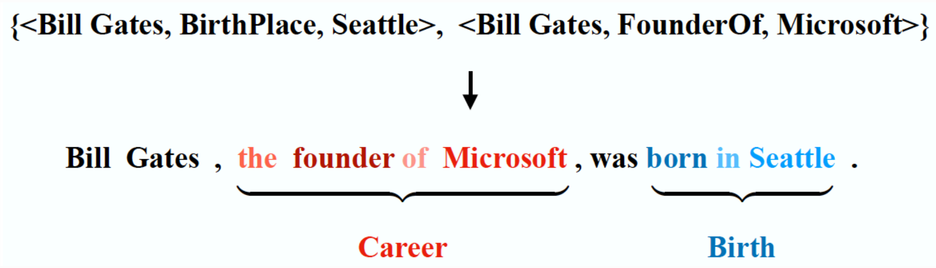
From Knowledge to Text

- WikiBIO to text
  - Paper: *Neural text generation from structured data with application to the biography domain*. EMNLP 2016
  - Given infobox to generate the first sentence of the person's bio
  - The task is a bit too simple
- WikiTable to text
  - Paper: *Table-to-Text: Describing Table Region with Natural Language*. AAAI 2018
  - Given a more complex table, to generate description for specified table region/columns

WebNLG Task

- The WebNLG Challenge: Generating Text from RDF Data
- E.g.
  - Data: (Johm E Blaha BIRTHDATE 1942-08-26) (Johm E Blaha BIRTHPLACE San Antonio) (Johm E Blaha OCCUPATION Fighter)
  - Text: John E Blaha, born in San Antonio on 1942-08-26, worked as a fighter pilot

- Our Dynamic Topic Attention Model



Open Questions

- The black-box nature of generic encoder-decoder models
  - Reference errors in the generated texts
  - Inter-sentence incoherence
  - A lack of fidelity to the source material
- Bring back the advantage of conventional methods
  - Learning sentence templates with neural models
  - Learning macro- and micro-planning

Thanks!

Define Smarter
Tomorrow.