Recent Advances in Natural Language Generation

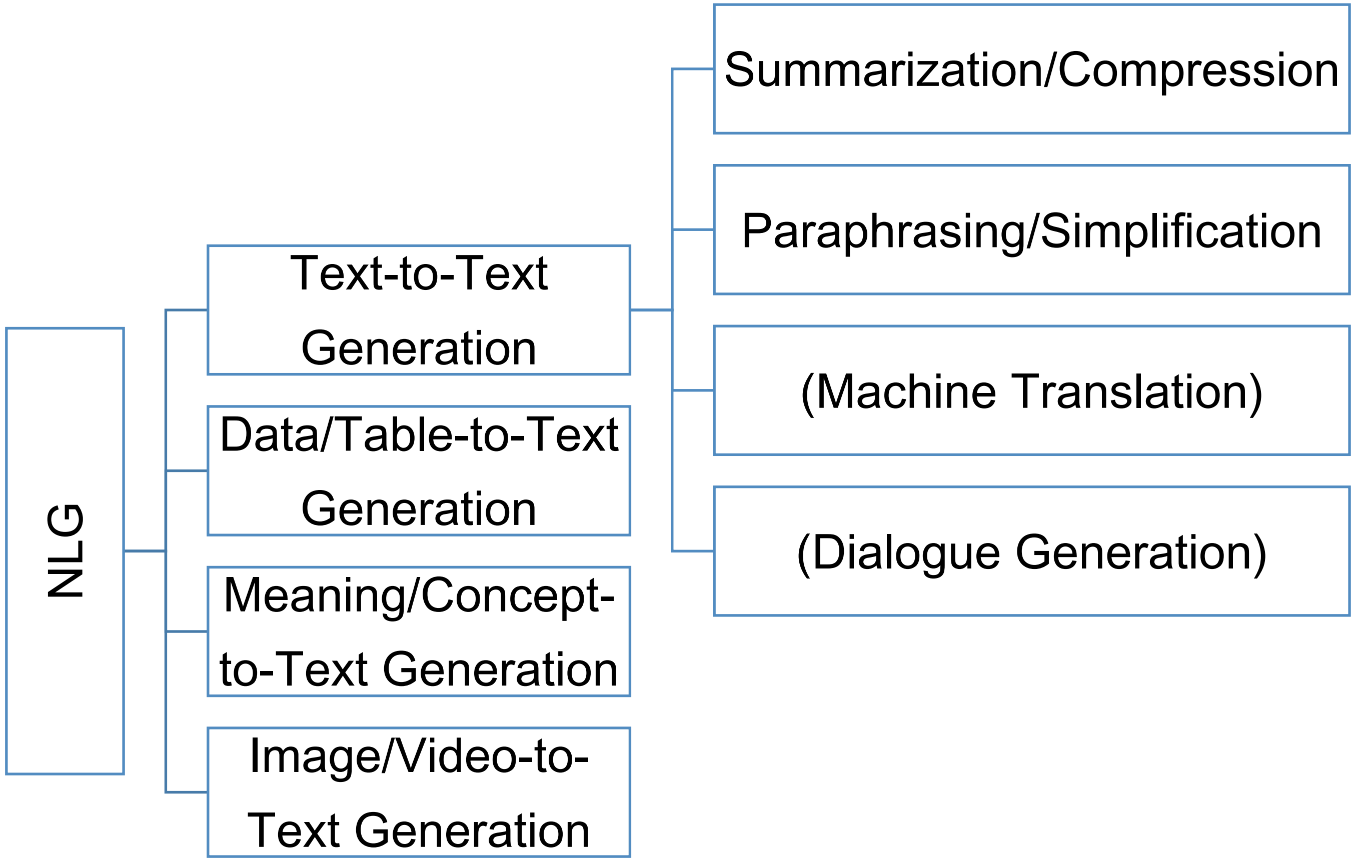
Xiaojun Wan

Peking University
https://wanxiaojun.github.io/

August 24, 2019

Taxonomy of NLG tasks

Creation vs. Re-Creation



WASHINGTON — Directly contradicting much of the Trump administration's position on climate change, 13 federal agencies unveiled an exhaustive scientific report on Friday that says humans are the dominant cause of the global temperature rise that has created the warmest period in the history of civilization.

Over the past 115 years global average temperatures have increased 1.8 degrees Fahrenheit, leading to record-breaking weather events and temperature extremes, the report says. The global, long-term warming trend is "unambiguous," it says, and there is "no convincing alternative explanation" that anything other than humans — the cars we drive, the power plants we operate, the forests we destroy — are to blame.

The report was approved for release by the White House, but the findings come as the Trump administration is defending its climate change policies. The United Nations convenes its annual climate change conference next week in Bonn, Germany, and the American delegation is expected to face harsh criticism over President Trump's decision to walk away from the 195-nation Paris climate accord and top administration officials' stated doubts about the causes and impacts of a warming planet.

	2006	2005	2004	2003	2002
Sales	\$128.3	\$97.2	\$74.6	\$61.9	\$68.3
Gross Margin	\$71.0 55%	\$53.2 55%	\$40.1 54%	\$25.7 42%	\$15.9 23%
Selling/Admin	\$40.2	\$31.8	\$25.9	\$27.0	\$33.8
Net R&D	\$15.4	\$12.2	\$12.4	\$27.0	\$12.0
Earnings (Loss) from Operations	\$8.0	\$(0.02)	\$(10.6)	\$(39.6)	\$(74.8)
Net Earnings (Loss)	\$8.1	\$(91.6)	\$(8.4)	\$(55.0)	\$(308.5)
Net Earnings per Share	\$0.12	\$(0.02)	\$(0.13)	\$(0.87)	\$(5.09)

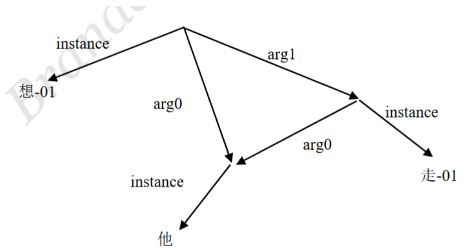
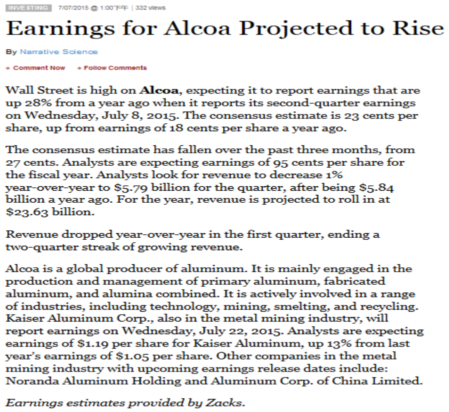


图 1：“他想走”的 CAMR 表示

Applications of NLG

News Summarization



Earnings for Alcoa Projected to Rise

By Narrative Science

Wall Street is high on **Alcoa**, expecting it to report earnings that are up 28% from a year ago when it reports its second-quarter earnings on Wednesday, July 8, 2015. The consensus estimate is 23 cents per share, up from earnings of 18 cents per share a year ago.

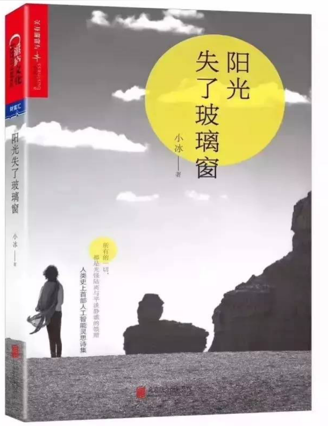
The consensus estimate has fallen over the past three months, from 27 cents. Analysts are expecting earnings of 95 cents per share for the fiscal year. Analysts look for revenue to decrease 1% year-over-year to $5.79 billion for the quarter, after being $5.84 billion a year ago. For the year, revenue is projected to roll in at $23.63 billion.

Revenue dropped year-over-year in the first quarter, ending a two-quarter streak of growing revenue.

Alcoa is a global producer of aluminum. It is mainly engaged in the production and management of primary aluminum, fabricated aluminum, and alumina combined. It is actively involved in a range of industries, including technology, mining, smelting, and recycling. Kaiser Aluminum Corp., also in the metal mining industry, will report earnings on Wednesday, July 22, 2015. Analysts are expecting earnings of $1.19 per share for Kaiser Aluminum, up 13% from last year's earnings of $1.05 per share. Other companies in the metal mining industry with upcoming earnings release dates include: Noranda Aluminum Holding and Aluminum Corp. of China Limited.

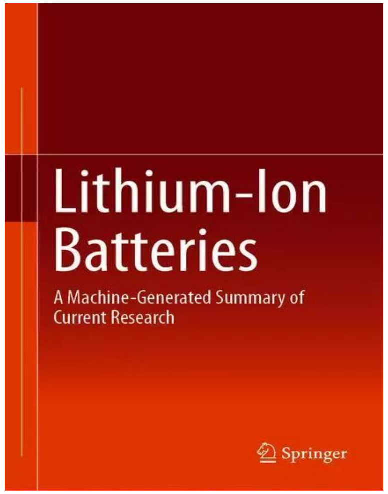
Earnings estimates provided by Zacks.

News Writing



Poetry Generation

Product Title /Specification Generation



Book Writing

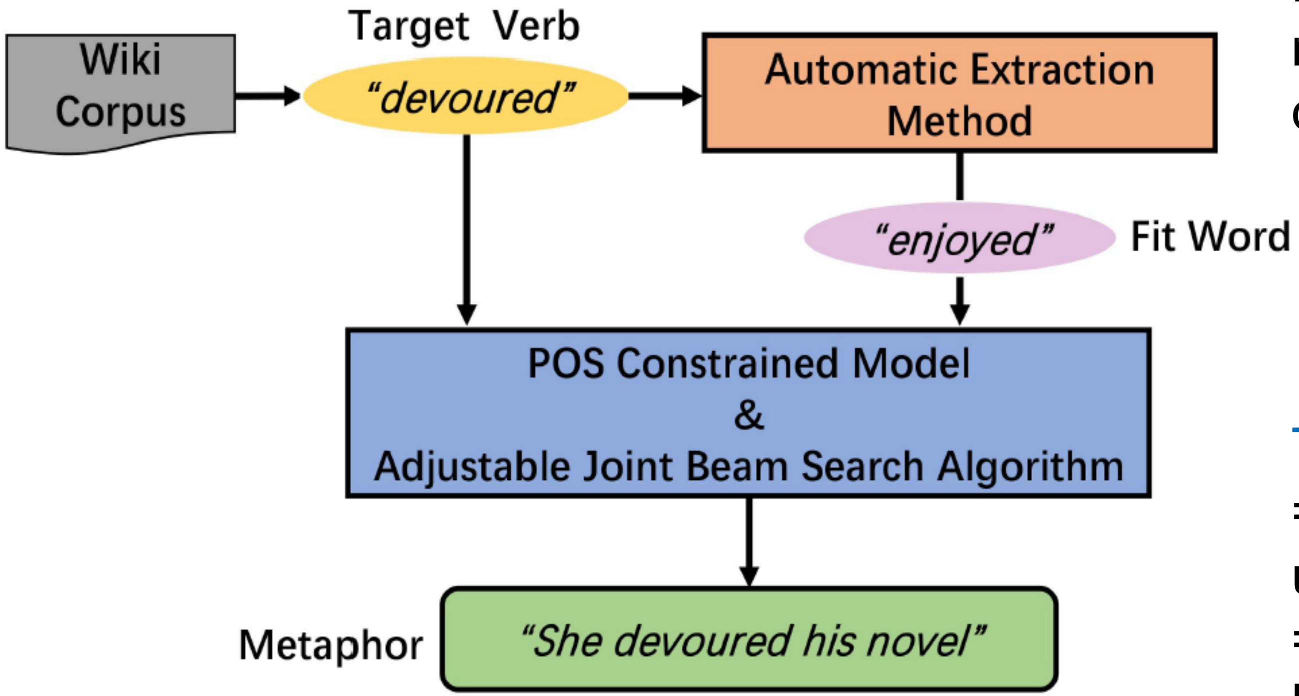
Applications of NLG

Recent Trends in NLG

- Creative Text Generation (Pun, Metaphor, Story, Poetry, …)
- Controllable Text Generation (w/ length, lexical, syntactic constraints)
- Generating Texts with Special Attributes (Sentiment and style transfer)
- Cross-Modal Text Generation (Image/Video caption & comment generation)
- Question Generation / Learning to Ask

Recent Trends in NLG

◆Creative Text Generation (Pun, Metaphor, Story, Poetry, …)



Metaphors are a form of figurative language, which refers to words or expressions that mean something different from their literal definition

Target Verb: absorbed   Fit Word: learn
=>Our Result: he absorbed his studies at the university of birmingham .
=>Gold Metaphor: he absorbed the knowledge or beliefs of his tribe .

[Yu and Wan, NAACL 2019]

Recent Trends in NLG

- Controllable Text Generation (w/ length, lexical, syntactic constraints)

X: his teammates' eyes got an ugly, hostile expression.
Y: the smell of flowers was thick and sweet.
Z: the eyes of his teammates had turned ugly and hostile.

X: we need to further strengthen the agency's capacities.
Y: the damage in this area seems to be quite minimal.
Z: the capacity of this office needs to be reinforced even further.

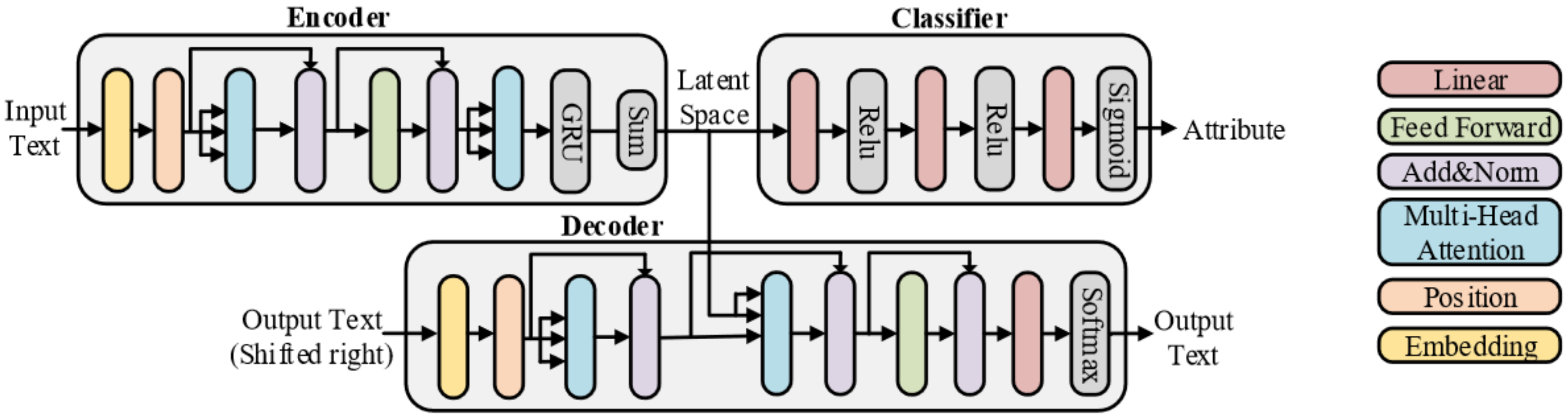
Controllable Paraphrase Generation with a Syntactic Exemplar [Chen et al., ACL 2019]

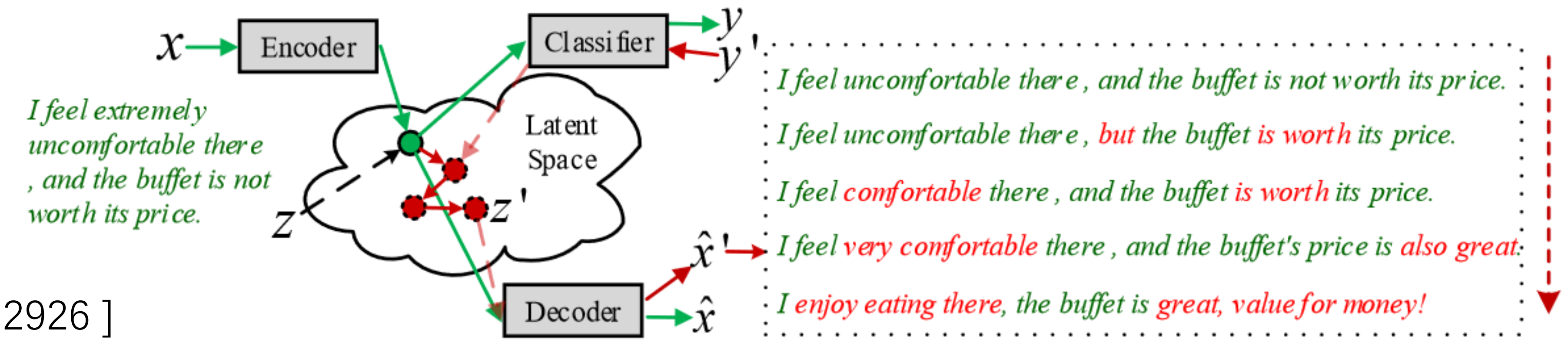
Input sentence & labels	Strategy	Generated Paraphrase
what is the value of a 1961 us cent?	All	what is the 1961 nickel 's value?
what is the value of a 1961 us cent?	Phrase	*what is the price of a 1961 nickel?*
what is the value of a 1961 us cent?	Sentence	*what is the 1961 us cent 's value?*
what is the value of a 1961 us cent?	Phrase	*what is the value of a 1961 nickel?*
what is the value of a 1961 us cent?	Sentence	*how much is a 1961 us cent worth?*

Decomposable Neural Paraphrase Generation [Li et al., ACL 2019]

Recent Trends in NLG

◆Generating Texts with Special Attributes (Sentiment and Style transfer)

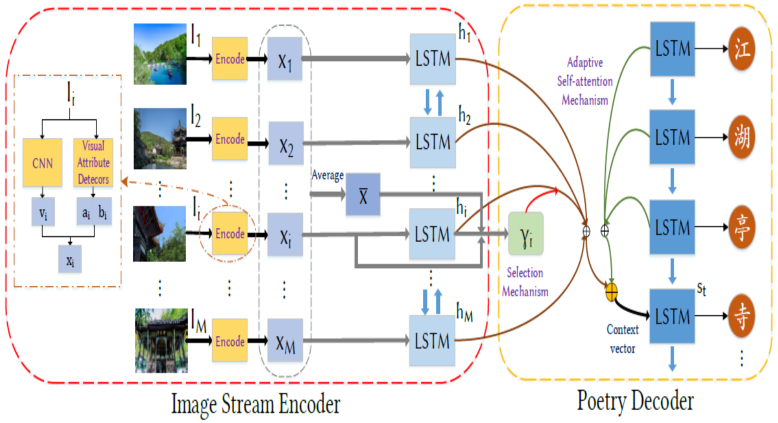




[Wang et al., arXiv:1905.12926]

Recent Trends in NLG

◆Cross-Modal Text Generation (Image/Video caption & comment generation)



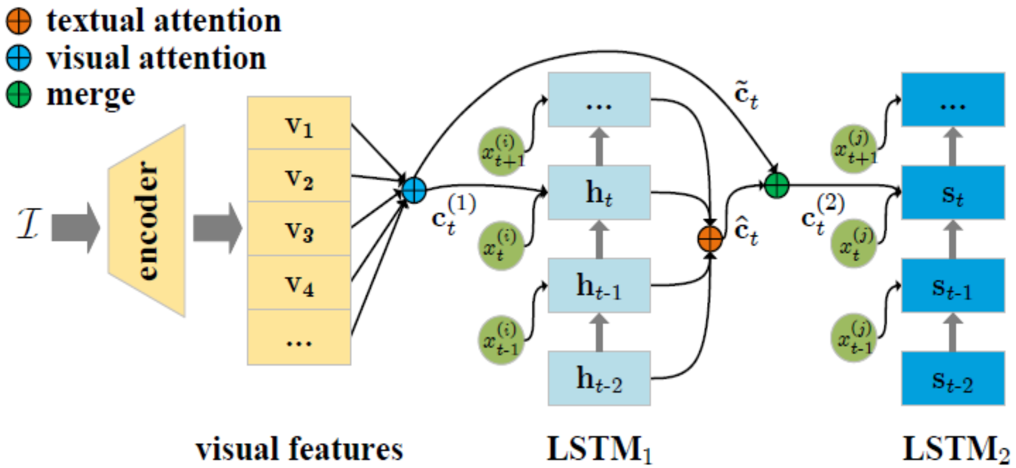
高阁登临万里遥，山川水色自萧条。
江湖有约同游赏，洛浦无情共寂寥。
古寺荒芜秋雨湿，孤城寒树暮潮销。
故园不见今年晚，一夕西风吹鬓凋。

高阁亭下水面平，
寺壁秋深古殿清。
廊庙不知天地窄，
桥边谁识此时情。

[Liu et al., ACMMM 2018]

Recent Trends in NLG

◆Cross-Modal Text Generation (Image/Video caption & comment generation)



Attention: a person riding a motorcycle on a road
Ours (IR): a person on a motorcycle in a race
Ours (Tdiv): a man riding a motorcycle down a road next to a pile of hay
Human: a person on a motorcycle riding beside hay bales

Attention: a black and white photo of an airplane
Ours (IR): an airplane sitting on a runway with people on the ground
Ours (Tdiv): a black and white photo of an airplane parked at an airport with people standing around
Human: an airplane with people under the wings at a field

[Liu et al., ICCV 2019]

Recent Trends in NLG

◆Question Generation / Learning to Ask

Fixed-answered questions:

Who invented the car? (Standard answer: Karl Benz)

Open-answered questions:

What do you think of the self-driving car? (No standard answer)

OpenQG**: Generating open-answered questions** based on given news that are suitable to arouse **open discussions**.

[Chai et al., ACL 2019]

Question Analysis

- Motivation
  - The more answers a question receives in online forums, the better it is for open discussions.
  - Many factors can influence the number of answers.
- Dominated variables
  - Prior works: how language use affects the reaction that a piece of text generates.
  - Four dominated variables: topic, author, posted time & **language use**
  - How language use makes a question get more answers.

    Control the influence of topic, author and posted time.

OQRanD

- OQGanD (Open Question Ranking Dataset).
  - Based on 11.5M open-domain questions from an in housed Zhihu database.
  - Contains 22K question-pairs. In each pair, the different in topic, time and author is controlled, and the main difference is language use.

Questions	# answers
你的家乡有什么初次尝试不太容易接受的美食么? (Is there any food that is hard to accept by foreigners for the first time in your hometown?)	1
有哪些在自己家乡很正常但在外地人眼里是黑暗料理的美食? (Which foods are normal in your hometown but are dark cuisine in the eyes of foreigners?)	89

The Effect of Language Use

- Method
  - Perform significant tests (one-sided paired t-test) on different linguistic features.
  - Find 33 features pass the significant tests.

- Some interesting features & conclusions
  - Length of questions: Ask concise questions.
  - # nouns: use less nouns to ask one topic a time.
  - # verbs: use more verbs to make the question vivid.
  - # honorifics: interact with readers naturally, do not use too many honorifics.
  - # positive words: questions with positive emotions are often more popular.
  - LM results: use familiar expressions. Distinctive expressions may attract attention, but “common language” can make a question better understood.

Question Evaluation Model

- Question ranking task.
  - Train a score model F_s which inputs a question and outputs a score.
    The larger score, the more answers are expected.
  - By comparing the two F_s values for each question-pair in OQGenD, we can predict which question gets more answers.

- Results
  - Statistical machine learning models:
    LR, RF & SVM
    N-gram word and POS features.
  - Deep learning methods:
    RNN & CNN
    word and POS embeddings.

Model	Accuracy	
	traditional	traditional+ours
LR	78.61%	82.33%
RF	81.70%	87.74%
SVM	79.02%	**87.96%**
RNN	74.68%	-
CNN	**83.18%**	-

Question Evaluation Model

- Question ranking task.
  - Train a score model F_s which inputs a question and outputs a score.
    The larger score, the more answers are expected.
  - By comparing the two F_s value each question-pair in OQGenD, we can predict which question gets more answers.

- Results
  - Statistical machine learning models:
    The 33 added features can help a lot.
  - Deep learning methods.
    Only use automatic-learned features.

Model	Accuracy	
	traditional	traditional+ours
LR	78.61%	82.33%
RF	81.70%	87.74%
SVM	79.02%	**87.96%**
RNN	74.68%	-
CNN	**83.18%**	-

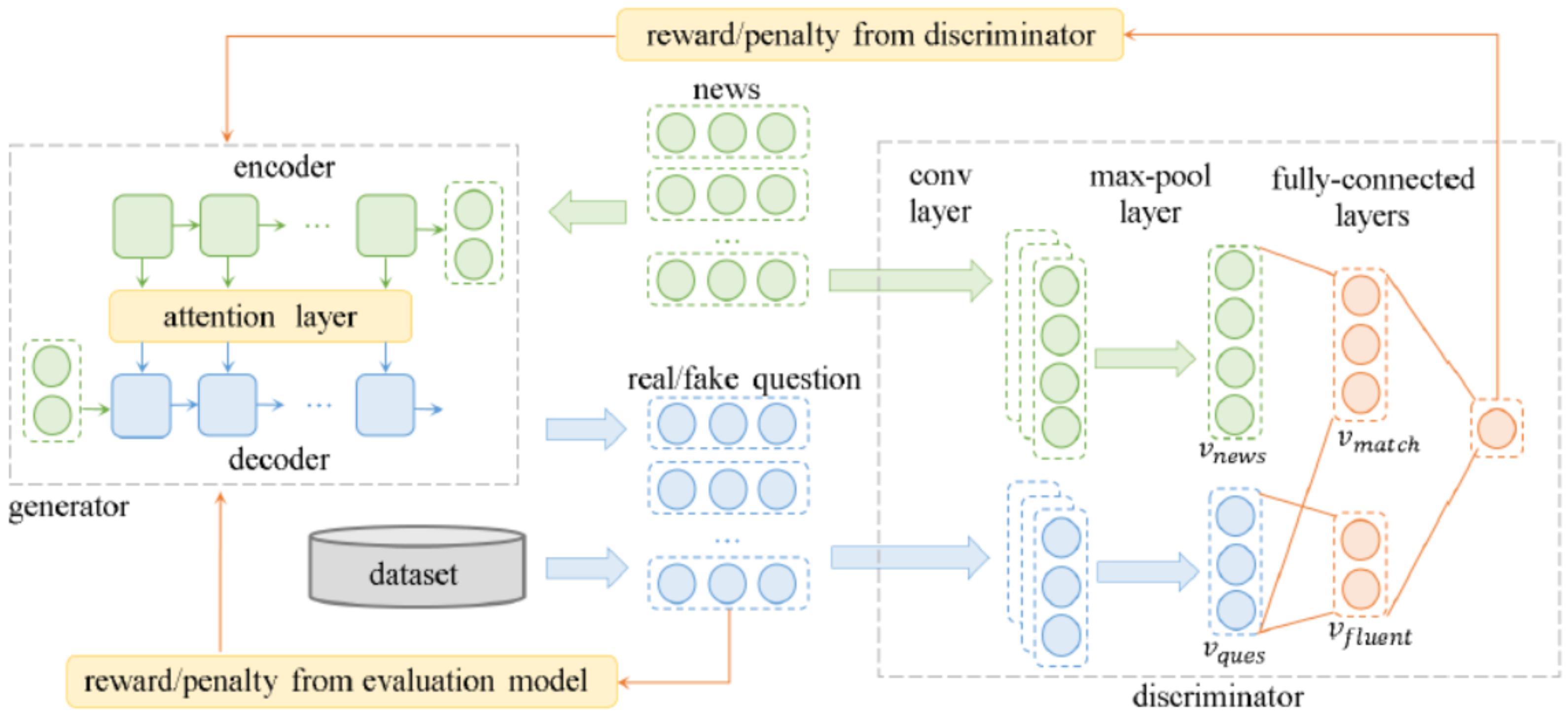
OQGenD

- OQGenD (Open Question Generation Dataset)

news	最后一次世界杯，C罗和梅西谁会赢。C罗和梅西谁更强？这个问题自两人出道就争论至今。2018年俄罗斯世界杯，…… (Who will win the last World Cup between Ronaldo and Messi? Who is stronger, Ronaldo or Messi? This issue has been debated since the beginning of their career. The 2018 World Cup in Russia ...)
gold questions	最后一次世界杯，C罗和梅西谁会赢？ (Who will win the last World Cup between Ronaldo and Messi?)
	最后一次世界杯，C罗会战胜梅西吗？ (Will Ronaldo defeat Messi in the last World Cup?)
	最后一次世界杯，C罗会输给梅西吗？ (Will Ronaldo lose to Messi in the last World Cup?)
	最后一次世界杯，梅西会输给C罗吗？ (Will Messi lose to Ronaldo in the last World Cup?)
	最后一次世界杯，梅西会战胜C罗吗？ (Will Messi defeat Ronaldo in the last World Cup?)

- Each news corresponds with more than one open-answered questions.
- In total: 9K news and 20K (news, question) pairs.

Model (Architecture)



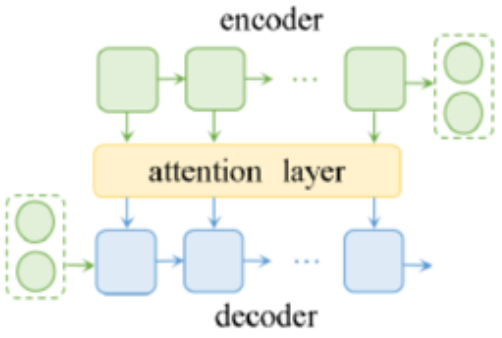
Based on the conditional generative adversarial network

- Generator: inputs news, generates (fake) questions.
- Discriminator: input (news, questions), predicts how likely it comes from real-world dataset.

Both of the generator & discriminator are conditioned on news.

Model (Details)

- Generator G_θ:
  - Sequence to sequence model with attention mechanism.



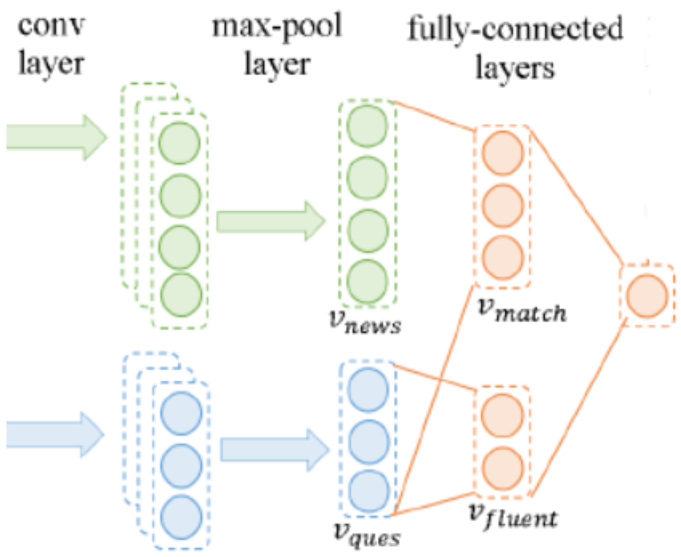
- Discriminator D_φ:
  - Embed input (news, question) into (v_{news} , v_{ques}).
  - High level representations:

$$\boldsymbol{v}_{match} = \boldsymbol{W}_m \left[\boldsymbol{v}_{news}; \boldsymbol{v}_{ques}\right] + \boldsymbol{b}_m$$

$$\boldsymbol{v}_{fluent} = \boldsymbol{W}_f \, \boldsymbol{v}_{ques} + \boldsymbol{b}_f$$

  - Get the final prediction:

$$D_\phi(X, Y_D) = \sigma(\boldsymbol{W}_{proj} \left[\boldsymbol{v}_{match}; \boldsymbol{v}_{fluent}\right] + \boldsymbol{b}_{proj})$$



Input of D_φ: (news, real-questions), (news, fake-questions)
(news, shuffled real-questions)

Object Functions

- Discriminator D_φ: $J_D(\phi) = -\mathbb{E}_{(X,Y)\sim P_{\text{real data}}} \log D_\phi(X,Y)$
$$-\mathbb{E}_{(X,Y)\sim P_{\text{fake data}}} \log(1 - D_\phi(X,Y))$$

- Generator G_θ:
  - Text generation is a discreate process, cannot directly use $D_\phi(X, Y_D)$ to train G_θ.
  - Use the idea of reinforcement learning.

    Policy π: the generator.

    State s_t: the generated text.

    Action a_t: generating the next word.

    Reward r_t: perform Monte-Carlo search,

Vanilla version: only use the results from our discriminator.

$$\boxed{r_t = \frac{1}{k}\sum_{i=1}^{k} D_\phi(\hat{Y}_{MC}^{(i)}, X)}$$

  - Get object function for generator by using policy gradient:

$$J_G(\boldsymbol{\theta}) = -\mathbb{E}[\sum_t r_t \cdot \log \pi(a_t|s_t)]$$

Object Functions

- Discriminator D_φ: $J_D(\phi) = -\mathbb{E}_{(X,Y)\sim P_{\text{real data}}} \log D_\phi(X,Y)$
  $-\mathbb{E}_{(X,Y)\sim P_{\text{fake data}}} \log(1 - D_\phi(X,Y))$

- Generator G_θ:
  - Text generation is a discreate process, cannot directly use $D_\phi(X, Y_D)$ to train G_θ.
  - Use the idea of reinforcement learning.

    Policy π: the generator.

    State s_t: the generated text.

    Action a_t: generating the next word.

    Reward r_t: perform Monte-Carlo search,

Full version: add the predictions from our question evaluation model

$$\boxed{r_t = \frac{1}{k}\sum_{i=1}^{k}(\gamma D_\phi(\hat{Y}_{MC}^{(i)}, X) + (1-\gamma)F_s(\hat{Y}_{MC}^{(i)}))}$$

  - Get object function for generator by using policy gradient:

$$J_G(\boldsymbol{\theta}) = -\mathbb{E}[\sum_t r_t \cdot \log \pi(a_t|s_t)]$$

Experiments

Models	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE$_L$	METEOR	F_s (SVM)
Seq2seq	$36.35^{*}_{\diamond}$	$20.25^{*}_{\diamond}$	$14.90^{*}_{\diamond}$	$13.22^{*}_{\diamond}$	$36.72^{*}_{\diamond}$	$21.57^{*}_{\diamond}$	$-2.28_{\diamond}$
CopyNet	$37.89^{*}_{\diamond}$	$21.09^{*}_{\diamond}$	$15.77^{*}_{\diamond}$	$14.07^{*}_{\diamond}$	$38.05^{*}_{\diamond}$	$22.63^{*}_{\diamond}$	$-1.80^{*}_{\diamond}$
SeqGAN	$38.51_{\diamond}$	$22.29_{\diamond}$	$16.97^{*}_{\diamond}$	$14.92^{*}_{\diamond}$	$38.40_{\diamond}$	$23.13^{*}_{\diamond}$	$-1.67^{*}_{\diamond}$
SentiGAN	$37.25^{*}_{\diamond}$	$21.52^{*}_{\diamond}$	17.24^{*}	15.60	$36.85^{*}_{\diamond}$	23.57	$-2.42^{*}_{\diamond}$
Ours (vanilla)	**39.67**	**23.62**	$18.01_{\diamond}$	$16.00_{\diamond}$	$\mathbf{39.87}_{\diamond}$	$24.52_{\diamond}$	$-1.89_{\diamond}$
Ours (full)	39.35	23.25	**18.62**	**16.44**	39.10	**24.96**	**-1.54**

Table 5: Results for openQG. $*$ ($\diamond$) denotes that our vanilla (full) model differs from the baseline significantly based on one-side paired t-test with $p < 0.05$.

- Traditional automatic question evaluation metrics.
- Predictions from our question evaluation model, F_s.

  The higher value, the better for open discussions.

  We use the SVM model as our F_s.

My Concerns about NLG

Evaluation

Usability

Thanks !