Multi-Fact Correction in Abstractive Text Summarization

Yue Dong[1] Shuohang Wang[2] Zhe Gan[2] Yu Cheng[2]
Jackie Chi Kit Cheung[1] Jingjing Liu[2]

[1]Mila / McGill University {yue.dong2@mail, jcheung@cs}.mcgill.ca
[2]Microsoft Dynamics 365 AI Research {shuowa, zhe.gan, yu.cheng, jingjl}@microsoft.com

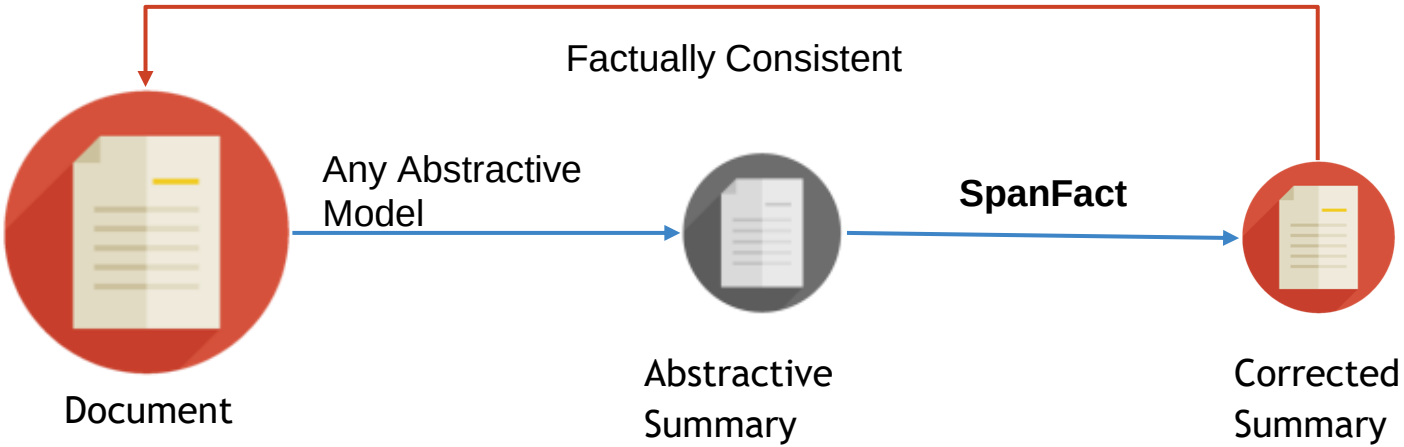
Factual Inconsistency in Abstractive Summaries

- On average, 8 - 30% of abstractive summaries by different systems are factually inconsistent w.r.t the source document (Kryscinski et. al., 2019)
- Need models to improve the factual consistency

CNNDM Source	Bottom-up Summary
(CNN) About a quarter of a million Australian homes and businesses have no power after a ``once in a decade'' storm battered Sydney and nearby areas. About 4,500 people have been isolated by flood waters as ``the roads are cut off and we won't be able to reach them for a few days,"...	a quarter of a million australian homes and businesses have no power after a decade.

What is SpanFact?

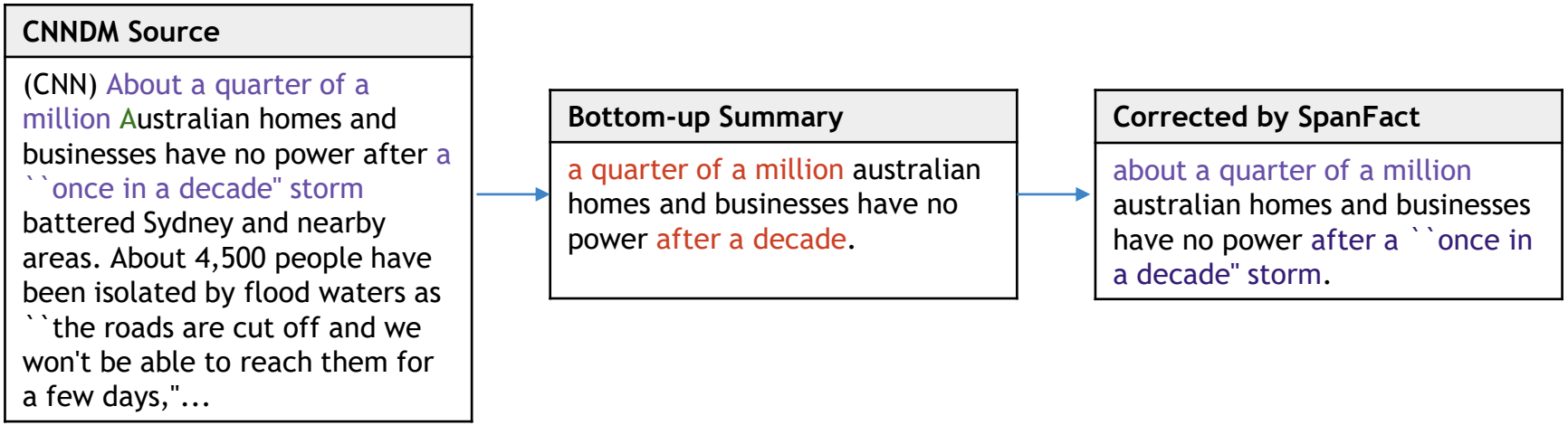
- A suite of two neural-based factual correctors
  - Post-editing models follows the principle that a summary should be
    - Informative (e.g. high ROUGE w.r.t. reference summary)
    - Correct (high factual scores w.r.t. the source document)



Mila

What is SpanFact?

- A suite of two QA-inspired factual correctors
  - Focus on entity-type error correction, a major source of hallucinated errors in abstractive summaries (Kryscinski et. al. 2019; Maynez et al., 2020)
  - Performing span selection to replace wrong entities



Why SpanFact as a Post-editing Model?

- Designing a fact-aware summarization model is expensive
- Hard and slow to incorporate latest advances in NLP (e.g. pre-trained models)
- Hard to achieve SOTA performance
  - Small boosts on factual scores (Kryscinski et. al. 2019; Wang et al., 2020)
  - Huge ROUGE drop of 12-35% (Cao et al. 2018; Zhu et al., 2020)

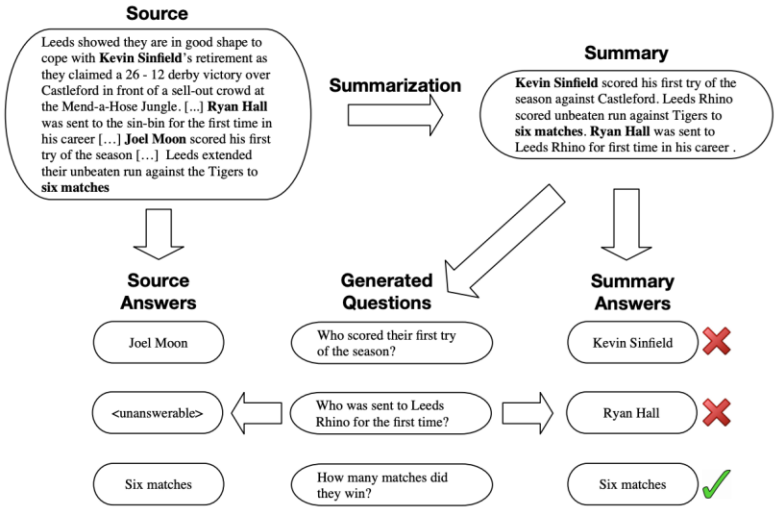
We propose:

Light-weight post-editing correctors that work on *any* abstractive systems

SpanFact: Inspired by QA for Reasoning

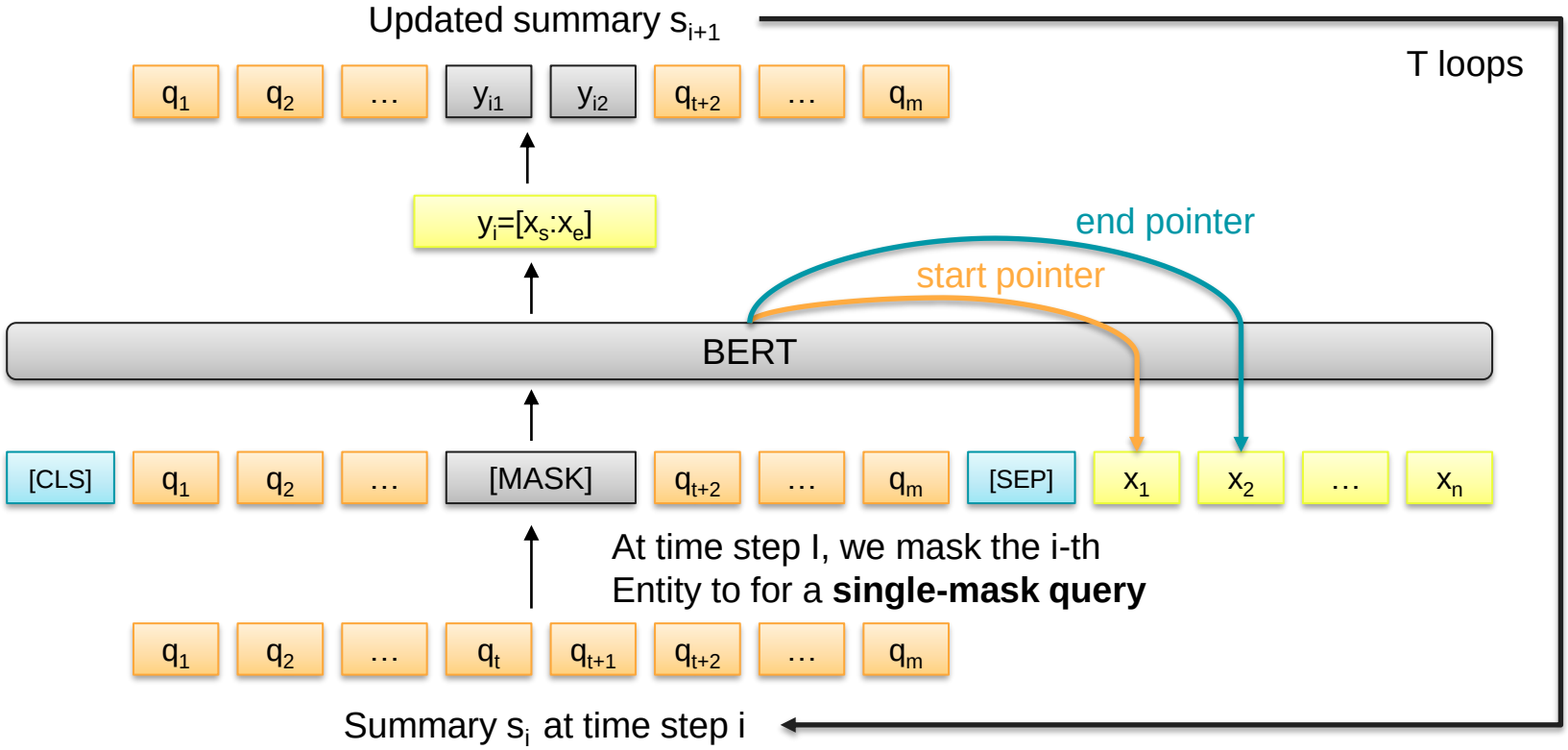
- QA systems have been already successfully used for factual consistency evaluation (Wang et al., 2020; Durmus et al.,2020)

- Pre-trained QA models achieve high performance on extractive QA (90%+ on SQuAD)

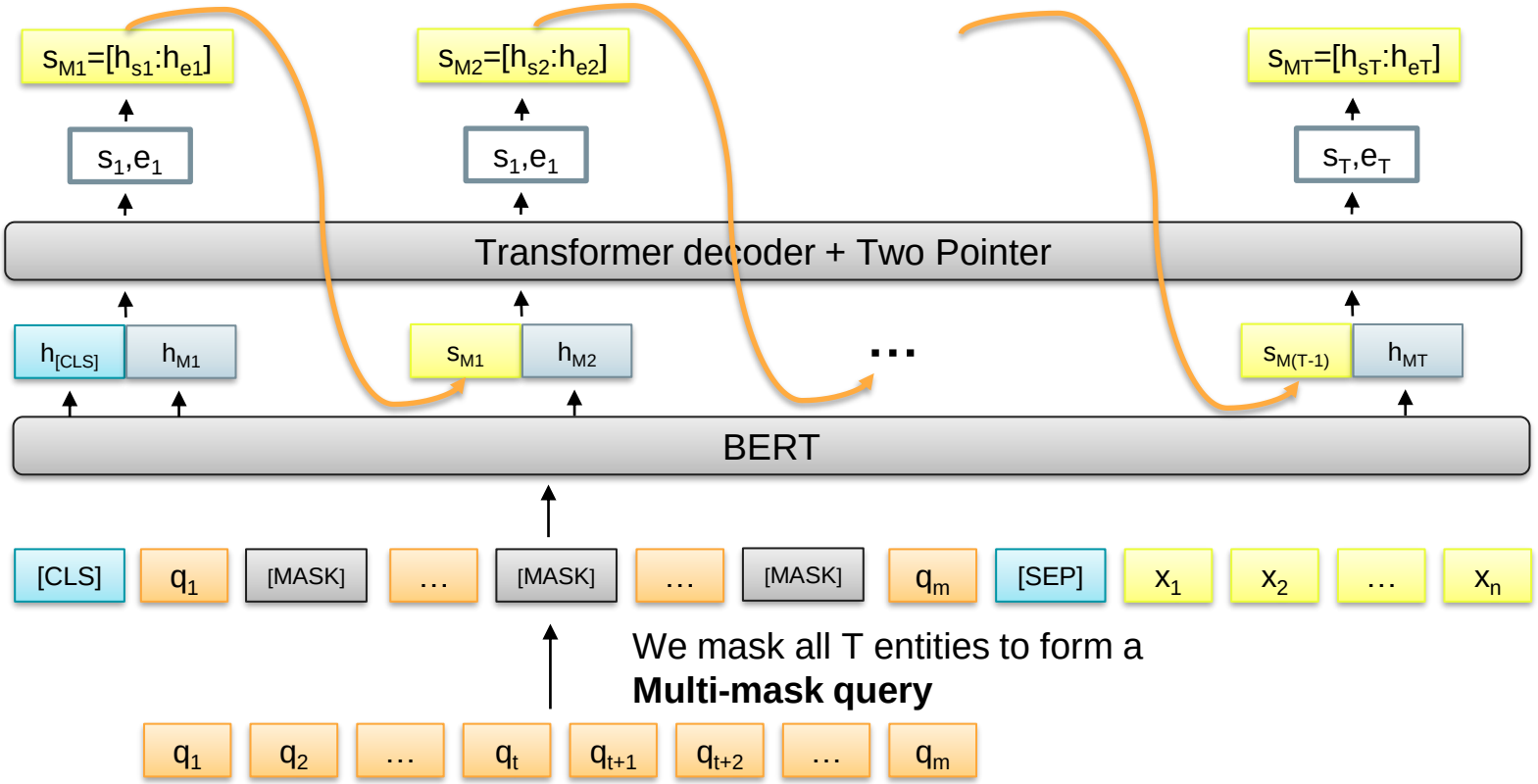


QAGS Wang et al., 2020

QA-Span Iterative Model



Auto-regressive Model



Training Data Creation

Learn to predict multi-token span for factual consistency correction

Source: -- the partnership started as a single shop on oxford street in london, opened in 1864 by john lewis. today the partnership is an organization with bases throughout the uk, with supermarkets and department stores, employing approximately 67,100 people. all 67,100 permanent staff are partners who own 26 john lewis department stores, 183 waitrose supermarkets, an online and catalogue business, john lewis direct a direct services company - greenbee, three production units and a farm. every partner receives the same scale of bonus, based on a fixed percentage of their annual wage. the bonus for 2006 was 18\% equivalent to 9 weeks pay, which was rolled out for every employee. chairman sir stuart hampson retired at the end of march 2007, his successor is charlie mayfield. hampson's salary for january 26, 2006 to january 26, 2007 was \$1.66 million which included the partnership bonus of \$250,000. john lewis' consolidated revenue for the last financial year was \$11.4 billion. **Target summary**: john lewis partnership began as a shop on london's oxford street in 1864. all 67,100 employees are partners in the organization and own shares.	***Iterative masking and span selection:*** **Query:** john lewis partnership began as a shop on [MASK]'s oxford street in 1864. all 67,100 employees are partners in the organization and own shares. **Answer Start:** 65 **Answer End:** 71 **Answer Text:** london ***Sequential masking and span selections:*** **Query:** [MASK] began as a shop on [MASK]'s [MASK] street in [MASK]. all [MASK] employees are partners in the organization and own shares. **Answer Start:** [-1,65,48,83,239] **Answer End:** [-1,71,54,87,245] **Answer Text:** ['john lewis partnership', 'london', 'oxford', '1864', '67,100']

Datasets and Abstractive Summarization Models

Summaries generated by abstractive summarization models are used as queries for error correction

- CNN/DailyMail (Hermann et al., 2015)
  - BertSumAbs, BertSumExtAbs, TransformerAbs (Liu and Lapata, 2019)
  - Bottom-up (Gehrmann et al., 2018)
- XSum (Narayan et al., 2018)
  - BertSumAbs, BertSumExtAbs and TransformerAbs (Liu and Lapata, 2019)
- Gigaword (Graff et al., 2003; Rush et al., 2015)
  - pointer-generator (See et al., 2017)
  - base and full GenParse models (Song et al., 2020)

Factual Correction Baseline

Two-encoder **Pointer Generator** (Split Encoder) (Shah et al., 2020)

- Masking all the entities in the system summary
- Uses dual encoders to copy and generate from both the source and the masked query for fact update
- Regenerate the mask query based on the source (generate one token at a time)

Our model's advantage:

- Local edits only on the masks
- Multi-token span selections

Evaluation Metrics

- Informativeness
  - ROUGE -1,-2,-L
- Factual Consistency
  - FactCC (Kryscinski et. al. 2019)
    - Classifier trained on weakly-supervised data
    - Decide whether a claim sentence is factually consistent with the source
  - QAGS (Wang et al., 2020)
    - F1 scores based on matched answers from the source and the generated summary
    - Answers generated from a summary should be similar to those generated from the source
    - We use our reimplementation QAQG, as the code and model were not available

Results on CNNDM

- Our models QA-Span and Auto-regressive
  - Boost factual consistency measures (QGQA and FactCC) by large margins
  - with only small drops on ROUGE (important content selection)
- QA-Span (iterative model) is better than Auto-regressive model
- Similar trends on Gigaword

Datasets	QGQA	FactCC sent	ROUGE 1	ROUGE 2	ROUGE L
Bottom-up	70.58	73.66	**41.24**	**18.70**	**38.15**
Split Encoders	70.22	73.15	39.78	17.87	37.01
QA-Span	**74.15**	**76.60**	41.13	18.58	38.04
Auto-regressive	72.78	74.42	41.04	18.48	37.95
BertSumAbs	72.68	76.76	**41.67**	**19.46**	**38.79**
Split Encoders	72.13	76.43	40.21	18.38	37.87
QA-Span	**74.93**	**78.69**	41.53	19.28	38.65
Auto-regressive	74.34	77.58	41.45	19.18	38.57
BertSumExtAbs	74.15	79.22	**41.87**	**19.41**	**38.94**
Split Encoders	73.67	79.12	40.55	18.41	38.45
QA-Span	**75.94**	**80.97**	41.75	19.27	38.81
Auto-regressive	75.19	79.89	41.68	19.16	38.74
TransformerAbs	73.79	80.51	**39.96**	**17.63**	**36.90**
Split Encoders	73.11	79.54	38.83	16.51	35.71
QA-Span	**75.70**	**82.82**	39.87	17.50	36.80
Auto-regressive	75.21	81.64	39.81	17.40	36.75

Results on XSum

- Our models boost factual consistency measures by large margins with a slight drop in ROUGE
- XSum is for extreme summarization, many entities in the reference are not in the source
- Auto-regressive model performs better

Datasets	QGQA	FactCC sent	ROUGE 1	ROUGE 2	ROUGE L
BertSumAbs	12.78	23.60	**37.78**	**15.84**	**30.50**
Split Encoders	**24.65**	24.19	34.22	13.76	27.86
QA-Span	23.85	23.90	36.44	14.56	29.38
Auto-regressive	24.14	**25.08**	36.24	14.37	29.22
BertSumExtAbs	13.62	23.12	**38.25**	**16.16**	**30.87**
Split Encoders	**25.17**	24.67	35.66	13.98	27.93
QA-Span	24.52	23.96	36.86	14.82	29.70
Auto-regressive	24.96	**25.10**	36.67	14.64	29.53
TransformerAbs	7.00	24.15	**29.86**	**10.05**	**23.78**
Split Encoders	11.77	24.78	28.14	8.65	22.70
QA-Span	12.88	24.44	29.51	9.67	23.45
Auto-regressive	**13.89**	**25.75**	29.45	9.59	23.40

Human Evaluation

Pairwise comparison of CNNDM summaries enhanced by different correction strategies

- Three annotators for each pair
- summaries from our two models are chosen more frequently as the factually correct one compared to the original.
- The preferences are comparable between iterative and auto-regressive correction models

BertAbs	Better	Worse	Same
QA-Span vs. original	**28.6%**	18.7%	52.7%
Auto-regressive vs. original	**31.3%**	16.7%	52%
QA-Span vs. Auto-regressive	26%	27.3%	46.7%
TransformerAbs	Better	Worse	Same
QA-Span vs. original	**38%**	11.3%	40.7%
Auto-regressive vs. original	**36%**	19.3%	44.7%
QA-Span vs. Auto-regressive	32.7%	28%	39.3%
Bottom-up	Better	Worse	Same
QA-Span vs. original	**34%**	12%	54%
Auto-regressive vs. original	**31.4%**	13.3%	55.3%
QA-Span vs. Auto-regressive	41.3%	32%	26.7%

Table 6: Human evaluation results on pairwise comparison of factual correctness on 450 (9×50) randomly sampled articles.

In Summary

Gigaword Source	all the 12 victims including 8 killed and 4 injured have been identified as senior high school students of the second senior high school of ruzhou city, central china's henan province, local police said friday.
Pointer-Generator Summary	12 killed, 4 injured in central china school shooting.
Corrected by SpanFact	8 killed, 4 injured in central china school shooting.
XSum Source	st clare's catholic primary school in birmingham has met with equality leaders at the city council to discuss a complaint from the pupil's family. the council is supporting the school to ensure its policies are appropriate...
BertAbs Summary	a muslim school has been accused of breaching the equality act by refusing to wear headscarves.
Corrected by SpanFact	a catholic school has been accused of breaching the equality act by refusing to wear headscarves.

Post-editing correctors

- Based on QA systems (pre-trained QA model)
- Performing span selection to replace wrong entities

Work on any abstractive systems

- Boosts on factual scores
- Without sacrificing ROUGE

Future Work:

- Errors beyond entities
- Generalization beyond domains that QA models are trained on

Thank You!

Please attend our Q&A session