Supplementary Material:
Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation

Han Guo[*]   **Ramakanth Pasunuru**[*]   **Mohit Bansal**

UNC Chapel Hill

{hanguo, ram, mbansal}@cs.unc.edu

1 Dataset Details

CNN/DailyMail Dataset   CNN/DailyMail dataset (Hermann et al., 2015; Nallapati et al., 2016) is a large collection of online news articles and their multi-sentence summaries. We use the original, non-anonymized version of the dataset provided by See et al. (2017). Overall, the dataset has $287,226$ training pairs, $13,368$ validation pairs and, $11,490$ test pairs. On an average, a source document has 781 tokens and a target summary has 56 tokens.

Gigaword Corpus   Gigaword is based on a large collection of news articles, where the article's first sentence is considered as the input document and the headline of the article as output summary. We use the annotated corpus provided by Rush et al. (2015). It has around 3.8 million training samples. For validation, we use $2,000$ samples and for test evaluation we use the standard test set provided by Rush et al. (2015). Following previous work, we keep our vocabulary size to $50,000$ frequent words.

DUC Corpus   We use the DUC-2002[1] document summarization dataset for checking our model's generalizability capabilities. DUC-2002 corpus consists of 567 documents with one or two human annotated reference summaries. We also tried beam retuning using DUC-2003[2] as a validation set, which consists of 624 documents with single human annotated reference summaries.

SNLI corpus   We use the Stanford Natural Language Inference (SNLI) corpus (Bowman et al.,

---

[*] Equal contribution.

[1] https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

[2] https://www-nlpir.nist.gov/projects/duc/guidelines/2003.html

2015) for our entailment generation task. Following Pasunuru and Bansal (2017), we use the same re-splits provided by them to ensure a zero train-test overlap and multi-reference setup. This dataset has a total of $145,822$ unique premise pairs out of $190,113$ pairs, which are used for training, and the rest of them are divided equally into validation and test sets.

SQuAD Dataset   We use Stanford Question Answering Dataset (SQuAD) for our question generation task (Rajpurkar et al., 2016). In SQuAD dataset, given the comprehension and question, the task is to predict the answer span in the comprehension. However, in our question generation task, we extract the sentence from the comprehension containing the answer span and create a sentence-question pair similar to Du et al. (2017). The dataset has around 100K sentence-question pairs from 536 articles.

2 Training Details

The following training details are common across all models and datasets. We use LSTM-RNN in our sequence models with hidden state size of 256 dimension. We use 128 dimension word embedding representations. We do not use dropout or any other regularization techniques, but we clip the gradient to allow a maximum gradient norm value of 2.0. We use Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001. Also, we share the word embeddings representation of both encoder and decoder in our models. All our tuning decisions (including soft/hard and layer-specific sharing decisions) were made on the appropriate validation/development set.

CNN/DailyMail: For all the models involving CNN/DailyMail dataset, we use a maximum encoder RNN step size of 400 and a maximum decoder RNN step size of 100. We use a mini-

Input Document: *john hughes has revealed how he came within a heartbeat of stepping down from his job at inverness as the josh meekings controversy went into overdrive this week . the caley thistle boss says he felt so repulsed by the gut-wrenching predicament being endured by his young defender - before he was dramatically cleared - that he was ready to walk away from his post and the games he loves , just weeks before an historic scottish cup final date . keen cyclist hughes set off on a lonely bike ride after hearing meekings had been cited for the handball missed by officials in the semi-final against celtic , and admits his head was in a spin over an affair that has dominated the back-page headlines since last sunday . inverness defender josh meekings will be allowed to appear in scottish cup final after his ban was dismissed . only messages of support awaiting him on his return from footballing friends brought him back from the brink of quitting . hughes , who lives in the black isle just north of inverness , said : ' i came in here this morning after a day off . i turned my phone off and was away myself , away out on the bike with plenty of thinking time : a great freedom of mind . ' i was that sick of what has been going on in scottish football i was seriously contemplating my own future . i 'm serious when i say that . ' i had just had it up to here and was ready to just give it up . if it was n't for what happened when i turned my phone back on , with the phone calls and texts i received from people i really value in football , that my spirits picked up again . ' the calls and texts came in from all over the place , from some of the highest levels across the game . i 've had phone calls that have really got me back on my feet . ' i would n't like to name them all , but there were a lot of good people and a good few close friends in the football fraternity . meekings was not sent off and no penalty was given as inverness went on to beat celtic 3-2 after extra-time . ' they were saying : " you need to lead from the front , you need to fight it . " that restored and galvanised that focus and drive in me . and , if that was how i was feeling , how was the boy josh meekings feeling ? it should never have come to this . ' meekings was cleared to play in the final by the judicial panel yesterday , but hughes insists this ' unprecedented ' sfa wrangle must be the catalyst for change in scottish football 's governance . although those who sit on the panel are drawn from many walks of life , ranging from former players and coaches to ex-refs and members of the legal profession , hughes said he wants ' real football people ' drafted in instead of the ' suits ' he claims lack understanding of the nuances and spirit of the professional game . and he seemed to point a thinly-veiled finger of accusation at sfa chief executive stewart regan by alleging that compliance officer tony mcglennan was a mere ' patsy ' in the process . (...)*

Ground-truth: *Inverness defender josh meekings has won appeal against one-match ban . the 22-year-old was offered one-game suspension following incident . however , an independent judicial panel tribunal overturned decision . inverness reached the scottish cup final with 3-2 win over celtic .*

See et al. (2017): *Josh meekings has been cleared to play in the scottish cup final .The englishman admitted he was fortunate not to have conceded a penalty and been sent off by* `referee steven mclean` *for stopping leigh griffiths net-bound effort on his goal-line . Meekings was not sent off and no penalty was given as inverness went on to beat celtic 3-2 .*

Our Baseline: *Josh meekings cleared to play in the scottish cup final on* `may 30` *. Inverness defender josh meekings will be allowed to appear in scottish cup final . Meekings was not sent off and no penalty was given as inverness went on to beat celtic 3-2 .*

Multi-task: *Josh meekings has been cleared to play in the scottish cup final . Inverness defender josh meekings will be allowed to appear in scottish cup final after his ban was dismissed . Inverness went on to beat celtic 3-2 after extra-time .*

Figure 1: Example showing summaries generated by See et al. (2017), our baseline, and 2-way multi-task model with summarization and entailment generation. Boxed-red highlighted words/phrases are not present in the input source document in any paraphrasing form. As shown, both See et al. (2017) and the baseline generate extraneous information that is not entailed by the source documents ("referee steven mclean" and "may 30"), but our multi-task model avoids such unrelated information to generate summaries that logically follow from the source document.

Input Document: *bending and rising in spectacular fashion , these stunning pictures capture the paddy fields of south east asia and the arduous life of the farmers who cultivate them . in a photo album that spans over china , thailand , vietnam , laos and cambodia , extraordinary images portray the crop 's full cycle from the primitive sowing of seeds to the distribution of millions of tonnes for consumption . the pictures were taken by professional photographer scott gable , 39 , who spent four months travelling across the region documenting the labour and threadbare equipment used to harvest the carbohydrate-rich food . scroll down for video . majestic : a farmer wades through the mud with a stick as late morning rain falls on top of dragonsbone terraces in longsheng county , china . rice is a staple food for more than one-half the world 's population , but for many consumers , its origin remains somewhat of a mystery . the crop accounts for one fifth of all calories consumed by humans and 87 per cent of it is produced in asia . it is also the thirstiest crop there is - according to the un , farmers need at least 2,000 litres of water to make one kilogram of rice . mr gable said he was determined to capture every stage of production with his rice project - from the planting to the harvesting all the way down to the shipping of the food . after acquiring some contacts from experts at cornell university in new york and conducting his own research , he left for china last may and spent the next four months traveling . he said : ' the images were taken over a four month period from april to july last year across asia . i visited china , thailand , vietnam , laos and cambodia as part of my rice project . video courtesy of www.scottgable.com . breathtaking : a paddy field worker toils on the beautiful landscape of dragonsbone terraces in longsheng county , china . farmers ' procession : a rice planting festival parade takes place near the village of pingan in guangxi province , china . ' the project is one part of a larger three part project on global food staples - rice , corn and wheat . i am currently in the process of shooting the corn segment . ' the industrialisation of our food and mono-culture food staples have interested me for some time so that 's probably what inspired me to do this project . ' i shot the whole project using a canon slr and gopros . the actual shooting took four months and then post production took another four more months . ' the reaction to my work has been incredibly positive - i was able to secure a solo gallery show and create quite a bit of interest online which has been great . ' family crop : a hani woman in traditional clothing sits on top of her family 's rice store in yunnan province , china . arduous labour : employees of taiwan 's state-run rice experimental station are pictured beating rice husks by hand as the sun shines on them . mr gable spent months learning mandarin chinese in preparation for his trip , but the language barrier was still his greatest challenge . (...)*

Ground-truth: *the spectacular photos were taken at paddy fields in* china , thailand , vietnam , laos *and* cambodia *. photographer* scott gable *spent four months travelling region to document the process of harvesting the crop .* rice *accounts for one fifth of all calories consumed by humans but crop is often still cultivated in primitive way .*

See et al. (2017): *the pictures were taken by professional photographer* scott gable *, 39 , who spent four months travelling across the region documenting the labour and the arduous life of the farmers who cultivate them . the images were taken over a four month period from april to july last year across asia . mr gable said he was determined to capture every stage of production with his* rice *project .*

Our Baseline: rice *is a staple food for more than one-half the world 's population . crop accounts for one fifth of all calories consumed by humans and 87 per cent of it is produced in asia .*

Multi-task: *in a photo album that spans over* china , thailand , vietnam , laos *and* cambodia *, extraordinary images portray the crop 's full cycle from the primitive sowing of seeds to the distribution of millions of tonnes for consumption . the crop accounts for one fifth of all calories consumed by humans and 87 per cent of it is produced in asia .*

Figure 2: Example showing summaries generated by See et al. (2017), our baseline, and 2-way multi-task model with summarization and question generation. All the unboxed-green highlighted words/phrases correspond to the salient information (based on the cloze-blanks of the original CNN/DailyMail Q&A task/dataset (Hermann et al., 2015)). As shown, our multi-task model is able to generate most of this saliency information, while the outputs from See et al. (2017) and baseline missed most of them, especially the country names.

Input Document: *celtic have written to the scottish football association in order to gain an ' under-standing óf the refereeing decisions during their scottish cup semi-final defeat by inverness on sunday . the hoops were left outraged by referee steven mclean ś failure to award a penalty or red card for a clear handball in the box by josh meekings to deny leigh griffith ś goal-bound shot during the first-half . caley thistle went on to win the game 3-2 after extra-time and denied rory delia ś men the chance to secure a domestic treble this season . celtic striker leigh griffiths has a goal-bound shot blocked by the outstretched arm of josh meekings . celtic ś adam matthews -lrb- right -rrb- slides in with a strong challenge on nick ross in the scottish cup semi-final . ' given the level of reaction from our sup-porters and across football , we are duty bound to seek an understanding of what actually happened , ćeltic said in a statement . they added , ' we have not been given any other specific explanation so far and this is simply to understand the circumstances of what went on and why such an obvious error was made . however , the parkhead outfit made a point of congratulating their opponents , who have reached the first-ever scottish cup final in their history , describing caley as a ' fantastic club ánd saying ' reaching the final is a great achievement . ćeltic had taken the lead in the semi-final through defender virgil van dijk ś curling free-kick on 18 minutes , but were unable to double that lead thanks to the meekings controversy . it allowed inverness a route back into the game and celtic had goalkeeper craig gordon sent off after the restart for scything down marley watkins in the area . greg tansey duly converted the resulting penalty . edward ofere then put caley thistle ahead , only for john guidetti to draw level for the bhoys . with the game seemingly heading for penalties , david raven scored the winner on 117 minutes , breaking thousands of celtic hearts . celtic captain scott brown -lrb- left -rrb- protests to referee steven mclean but the handball goes unpunished . griffiths shows off his acrobatic skills during celtic ś eventual surprise defeat by inverness . celtic pair aleksandar tonev -lrb- left -rrb- and john guidetti look dejected as their hopes of a domestic treble end .*

Ground-truth: *`celtic` were defeated 3-2 after extra-time in the `scottish cup` semi-final . `leigh griffiths` had a goal-bound shot blocked by a clear handball. however, no action was taken against offender `josh meekings` . the `hoops` have written the `sfa` for an 'understanding' of the decision .*

See et al. (2017): *`john hartson` was once on the end of a major `hampden injustice` while playing for celtic . but he can not see any point in his old club writing to the scottish football association over the latest controversy at the national stadium . `hartson` had a goal wrongly disallowed for offside while `celtic` were leading 1-0 at the time but went on to lose 3-2 .*

Our Baseline: *`john hartson` scored the late winner in 3-2 win against `celtic` . celtic were leading 1-0 at the time but went on to lose 3-2 . some fans have questioned how referee steven mclean and `additional assistant alan muir` could have missed the infringement .*

Multi-task: *`celtic` have written to the scottish football association in order to gain an ' understand-ing ' of the refereeing decisions . the `hoops` were left outraged by referee steven mclean 's failure to award a penalty or red card for a clear handball in the box by `josh meekings` . celtic striker `leigh griffiths` has a goal-bound shot blocked by the outstretched arm of josh meekings .*

Figure 3: Example of summaries generated by See et al. (2017), our baseline, and 3-way multi-task model with summarization and both entailment generation and question generation. The boxed-red highlighted words/phrases are not present in the input source document in any paraphrasing form. All the unboxed-green highlighted words/phrases correspond to the salient information. See detailed discussion in Fig. 1 and Fig. 2 above. As shown, the outputs from See et al. (2017) and the baseline both include non-entailed words/phrases (e.g. "john hartson"), as well as they missed salient information ("hoops", "josh meekings", "leigh griffiths") in their output summaries. Our multi-task model, however, manages to accomplish both, i.e., cover more salient information and also avoid unrelated information.

batch size of 16. We initialize the LSTM-RNNs with uniform random initialization in the range $[-0.02, 0.02]$. We set λ to 1.0 in the joint cross-entropy and coverage loss. Also, we only add coverage to the converged model with attention and pointer mechanism, and make the learning rate from 0.001 to 0.0001. During multi-task learning, we use coverage mechanism for primary (CNN/DailyMail summarization) task but not for auxiliary tasks (because they do not have traditional redundancy issues). The penalty coefficient γ for soft-sharing is set to 5×10^{-5} and 1×10^{-5} for 2-way and 3-way multi-task models respectively (the range of the penalty value is intuitively chosen such that we balance the cross-entropy and regularization losses). In inference time, we use a beam search size of 4, following previous work (See et al., 2017).

Gigaword: For all the models involving Gigaword dataset, we use a maximum encoder RNN step size of 50 and a maximum decoder RNN step size of 20. We use a mini-batch size of 256. We initialize the LSTM-RNNs with uniform random initialization in the range $[-0.01, 0.01]$. We do not use coverage mechanism to our Gigaword models. Also, we set our beam search size to 5, following previous work (Nallapati et al., 2016).

DUC: For the CNN/DM to DUC domain-transfer experiments where we allow the beam sizes of all models to be individually re-tuned on DUC-2003, the chosen tuned beam values are $10, 4, 3$ for the multi-task model, baseline, and See et al. (2017), respectively.

2.1 Multi-Task Learning Details

Multi-Task Learning with Question Generation Two important hyperparameters tuned are the mixing ratio between summarization and entailment generation, as well as the soft-sharing coefficient. Here, we choose the mixing ratios 3:2 between CNN/DailyMail and SQuAD, 100:1 between Gigaword and SQuAD. Intuitively, these mixing ratios are close to the ratio of their dataset sizes. We set the soft-sharing coefficient γ to 5×10^{-5} and 1×10^{-5} for CNN/DailyMail and Gigaword, resp.

Multi-Task Learning with Entailment Generation Here, we choose the mixing ratios 3:2 between CNN/DailyMail and SNLI, 20:1 between Gigaword and SNLI. We again set the soft-sharing coefficient γ to 5×10^{-5} and 1×10^{-5} for CNN/DailyMail and Gigaword, resp.

| **Premise:** *People walk down a paved street that has red lanterns hung from the buildings.* |
| **Entailment:** *People walk down the street.* |
| **Premise:** *A young woman on a boat in a light colored bikini kicks a man wearing a straw cowboy hat.* |
| **Entailment:** *A young woman strikes a man with her feet.* |

Figure 4: Output examples from our entailment generation model.

| **Input:** *The college of science was established at the university in 1865 by president father patrick dillon.* |
| **Question:** *In what year was the college of science established ?* |
| **Input:** *Notable athletes include swimmer sharron davies , diver tom daley , dancer wayne sleep , and footballer trevor francis .* |
| **Question:** *What is the occupation of trevor francis ?* |

Figure 5: Output examples from our question generation model.

Multi-Task Learning with Question and Entailment Generation Here, we choose the mixing ratios and soft-sharing coefficients to be 4:3:3 and 5×10^{-5} for CNN/DailyMail, and 100:1:5 and 1.5×10^{-6} for Gigaword respectively.

3 Analysis

3.1 Entailment Generation Examples

See Fig. 4 for interesting output examples by our entailment generation model.

3.2 Question Generation Examples

See Fig. 5 for interesting output examples by our question generation model.

References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *ACL*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multi-task video captioning with video and entailment generation. In *ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.