An Entity-Driven Framework for Abstractive Summarization (Supplementary Material)

Eva Sharma[1]* **Luyang Huang**[2]* **Zhe Hu**[1]* and **Lu Wang**[1]
[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115
[2]Department of Electrical and Computer Engineering, Boston University, Boston, MA 02215
[1]evasharma@ccs.neu.edu, hu.zhe@husky.neu.edu, luwang@ccs.neu.edu
[2]lyhuang@bu.edu

A Entity-based Coherence Model

In Section 2.3, we explain that, since the coherence model outputs pairwise coherence scores, for a summary containing more than two sentences, we use the average of all adjacent sentence pairs' scores as the final summary coherence score. Here, we explain the correlation study we conducted to show that the average aggregation performed in this case, works reasonably well.

We first sample 300 such NYT summaries from our test set. Considering the original summary containing two sentences as a positive sample, for each summary, we construct a negative sample by reversing its original sentence order. If a summary contains more than two sentences, we further construct a "middle" sample by shuffling its sentence order such that it is different from positive and negative samples. We give each positive sample a coherence score of 3 (perfect score), a middle sample 2, and a negative sample 1 (worst score). We call this ideal coherence score. For each of these samples, we also get the coherence score given by our model using average aggregation to score each sample summary. Finally, we calculate the Spearman rank-order correlation coefficient between the ideal coherence score and our model score using SciPy[1]. The final correlation coefficient is 0.87, with p-value 1.43e-237, which shows that the two coherence scores are highly correlated. This implies that aggregated coherence score for a summary can be used as a reliable proxy for its overall coherence score.

A.1 Entity-base Coherence Model Evaluation

To investigate the robustness of our coherence model, we constructed several test sets to evaluate its performance on different dimensions of coherence, which we describe in detail here. We also considered three baselines as a control for some of these dimensions. Each test instance in our test sets had two sentences, except for the SHUF.. We describe our test data construction and baselines as:

PAIR: We construct this test set in the same way as the training data for our model (as in Section 4.1 in the main paper).

CONN.: To test whether model learns to discriminate between proper and improper usage of discourse connectors, we replace the discourse connectors connecting two sentences with a connector of different type in the incoherent pairs (We borrow the list of discourse connectors and rules from Geva et al. (2019)). We implement TOP20CON, as a baseline for this test set, which labels a pair of sentences using top 20 most frequently used discourse connectors as more coherent.

REF. AND ENT.REP.: Here, we want to detect whether model learns to discriminate between proper and improper usage of pronouns and entity repetition. For pronouns (Ref.), we swap the pronominal entity mention with the respective nominal entity mention to construct incoherent pairs. For entity repetition (Ent.Rep.), we replace the pronominal entity mention in the second sentence with the respective nominal entity mentioned in first sentence to construct incoherent pairs. For both of these respective test sets, we use ECHAIN as a baseline, that labels a pair of sentences as more coherent if they have one or more entity mentions co-referred.

OVER. To confirm that model learns that coherent pair of sentences share topical content, we force incoherent sentence pairs to have no content overlap. For this, we consider COSSIM, that labels a pair of sentences with higher cosine similarity as more coherent.

* These authors contributed equally. Work done while LH was at Northeastern University.

[1]https://docs.scipy.org/doc/scipy/reference/index.html

SHUF. To get the final coherence score for a summary, we take an average over the coherence scores of all the adjacent sentence pairs in a summary. Therefore, to reliably test model's performance on test instances containing more than two sentence, considering a human written summary as coherent, we shuffle sentences of each summary to construct respective incoherent pairs.

Model	PAIR.	CONN.	REF.	ENT.REP.	OVER.	SHUF.
RANDOM	50.0	50.0	50.0	50.0	50.0	50.0
TOP20CON	48.6	49.5	-	-	-	51.6
ECHAIN	54.5	-	61.9	43.1	-	48.5
COSSIM	50.7	-	-	-	79.2	55.7
Wu & Hu	41.5	50.6	42.1	34.6	52.5	19.0
OURS	**84.1**	**73.0**	**94.2**	**93.4**	**85.7**	**95.4**

Table 1: Accuracy of Entity-based Coherence model compared to different baselines and Wu and Hu (2018). The listed test sets are constructed from NYT test set and contain 1000 samples each.

Table 1 lists the performance of our coherence model and Wu and Hu (2018) on aforementioned test sets, along with four baseline systems. Our model achieves 95.4% accuracy on SHUF., which implies that it accurately captures overall coherence of a summary. Moreover, we find that our model performs significantly better on all other test sets too, compared to often worse than random performance of Wu and Hu (2018) model on the same. We attribute this to the fact that Wu and Hu (2018) data construction can not capture different dimensions of coherence. We also find our model to perform better than the three baselines on respective datasets. For instance, on CONN., our model performs significantly better than RANDOM and TOP20CON indicating that it is not overfitting to the data, instead learns the pattern of discourse connector usage.

B Human Evaluation Guideline

In our human evaluation, each human annotator is presented with 30 news articles. The annotators are asked to evaluate five summaries for each article on the following aspects on scale of 1 - 5 (1 being very poor and 5 being very good). More detailed instructions are in Table 2.

- **Informativeness**: whether the summary provides enough and necessary content coverage from the input article.
- **Grammaticality**: whether the summary has no obvious grammatically incorrect sentences (e.g., fragments, missing components) that make the text difficult to read.
- **Coherence**: weather the summary is coherent and well-organized.

C Sample Output

We include 4 complete sample outputs for different models in Figures 1 to 4.

References

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. DiscoFuse: A Large-Scale Dataset for Discourse-Based Sentence Fusion. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ArXiv preprint arXiv:1902.10526.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

	Informativeness:
1	Not relevant to the article e.g., *"less than three months after state voters defeat proposal that would have outlawed most abortions . legislators , introduced scaled-back version"*
3	Relevant, but misses the main point of the article e.g., *"vermont supreme court rules isabella miller-jenkins has two mothers . The Vermont ruling , he added , illustrates that same-sex marriage or civil unions will inevitably clash with other states ."*
5	Successfully captures the main point of the article e.g., *"vermont supreme court rejected host of arguments from isabella's biological mother lisa miller that her former lesbian partner janet jenkins should be denied parental rights . decision conflicts with one from court in virginia , where miller and her daughter , who is 0 , now live"*
	Grammaticality:
1	Summary contains fragments, missing components, extra punctuations e.g., *"court rejects host of arguments from isabella 's biological mother , lisa miller , that her former lesbian partner janet jenkins should be denied parental rights . miller and to , to"*
3	Relatively minor grammatical errors e.g., *"isabella miller-jenkins , vermont supreme court , rules rejects host of arguments from isabella 's biological mother , lisa miller ."*
5	Correct Grammar. e.g., *"vermont supreme court rejected host of arguments from isabella's biological mother lisa miller that her former lesbian partner janet jenkins should be denied parental rights ."*
	Coherence:
1	Completely incoherent with no across sentence organization e.g., *"isabella miller-jenkins , vermont supreme court , isabella miller-jenkins , vermont supreme court."*
3	Understandable, but sentences can be slightly re-ordered for it to read better e.g., *"vermont supreme court rejects arguments from lisa miller , biological mother of 0-year-old isabella miller-jenkins , that her former lesbian partner , janet jenkins , should be denied parental rights . isabella was born in virginia in 0 , after miller was impregnated with sperm from anonymous donor whom jenkins helped select. decision conflicts with one from court in virginia , where miller and her daughter now live ."*
5	Completely coherent. e.g., *"vermont supreme court rejected host of arguments from isabella 's biological mother lisa miller that her former lesbian partner janet jenkins should be denied parental rights . decision conflicts with one from court in virginia , where miller and her daughter , who is 0 , now live"*

Table 2: Sample summaries with explanations on human evaluation aspect scales.

ARTICLE:
A Backyard Mayan Temple May Be Doomed. WHEN Barbara Winston had a stone replica of a Mayan temple built in her backyard, it was meant to give her a sense of peace. Instead, it started a war. In August 0, Mrs. Winston, who lives on a 0-acre estate with her husband, Bruce, had a stonecutter erect a $ 0,0 replica of Guatemala's Temple of the Great Jaguar at the rear of their property. While the temple's scale is one-seventeenth of the original, it is by no means small, with granite blocks forming three staircases rising nine feet to a rectangular platform. It is also 0 feet from the horse barn and riding ring of the Winstons' neighbor, Diane Lewis , who complained to the Town of Bedford that the temple detracted from the enjoyment of her property. In September, the Zoning Board of Appeals rejected the Winstons' request for a variance and said that the temple was a structure requiring permits and a 0-foot setback from adjacent property. It questioned why the Winstons, with 0 acres to choose from , put the temple so close to Ms. Lewis's property. The ruling left the Winstons with three choices: tear down the temple, move it , or fight. (...) In their lawsuit, the Winstons argue that the temple was not a structure as defined in the ordinance. Even if it were, they said, the town should grant the variance because of the temple's spiritual significance. Joel Sachs, the town attorney, said the temple clearly met the definition of a structure . (...)

HUMAN:
Bruce and Barbara Winston file suit against Zoning Board in Bedford, NY, after board rejects their request for variance for stone replica of Mayan temple that couple had built on their 0-acre estate. board, after complaint from neighbor Diane Lewis, says temple is structure that requires permits and 0-foot setback from adjacent property

POINTGEN + COV:
arbara winston, who lives on 0-acre estate with her husband, bruce, has stonecutter erect $ 0,0 replica of guatemala's temple of great jaguar at rear of their property. it is by no means small, with granite blocks forming three staircases rising nine feet to rectangular platform

DEEPREINFORCE:
article on barbara winston, $ 0,0 replica of guatemala's temple of great jaguar at rear of their property. of zoning board of appeals, to request for variance and of that temple, to structure requiring permits and 0-foot from adjacent property.

SENECA:
winston, who lives on 0-acre estate with her husband, bruce. erect $ 0,0 replica of guatemala's temple of great jaguar at rear of their property. zoning board of appeals rejects winstons' request for variance and says that temple, is structure requiring permits and 0-foot setback from adjacent property

SENECA + R_{Coh}:
winston, who lives on 0-acre estate with her husband, bruce, has stonecutter erect $ 0,0 replica of guatemala's temple of great jaguar at rear of their property. zoning board of appeals rejects winstons' request for variance and says that temple is structure requiring permits and 0-foot setback from adjacent property

Figure 1: Sample summaries for an NYT article. Our models capture salient information which is missed by comparisons. Numbers are replaced with "0".

ARTICLE:
TV Ad for Pfizer's Painkiller Criticized by Consumer Group. A new television commercial for Pfizer's painkiller Celebrex that has attracted attention for both its length and innovative marketing approach is now also being criticized for its message. Public Citizen, a consumer group, asked the Food and Drug Administration yesterday to ban the commercial, charging that it gives consumers a false impression that the prescription drug has no more safety risk than some other painkillers. Celebrex is in the same class of drugs as the Merck pill Vioxx, which was withdrawn in 0 because of its link to cardiovascular problems. At that time, the F.D.A. also asked Pfizer to suspend its television advertising for Celebrex, and the company complied. The new Celebrex ad, which can also be seen on a Pfizer Web site celebrex.com, represents a return to television for the product after a hiatus of more than two years. It was broadcast for the first time last Monday on World News With Charles Gibson on ABC. It was two and a half minutes rather than the usual 0 seconds , and was the only ad during last Monday's program. Mr. Gibson announced that the new format with less advertising would be repeated on several Mondays this month. Pfizer has also bought last night's and next Monday's World News broadcast, but had said all along that it would not disclose the content of those ads until they were broadcast. Last night's advertisement involved smoking cessation and directed viewers to a Web site, www.mytimetoquit.com that links to information about Chantix, Pfizer's prescription stop-smoking drug. (...)

HUMAN:
New television commercial for Pfizer's painkiller Celebrex that has attracted attention for its length and innovative marketing approach is being criticized for its message; Public Citizen asks Food and Drug Administration to ban commerical, charging that it gives consumers false impression that drug has no more safety risk than some other painkillers; Pfizer defends ad

POINTGEN + COV:
public television commercial for pfizer's painkiller celebrex that has attracted attention for both its length and innovative marketing approach is now being criticized for its message. new celebrex ad, which can also be seen on pfizer web site celebrex.com, represents return to television for product after hiatus of more than two years

DEEPREINFORCE:
new television commercial for pfizer's painkiller celebrex that has attracted attention for both its length and innovative marketing approach is criticized for its message. public citizen, food and drug administration to ban commercial, charging that it gives consumers false impression that prescription drug has no more safety than some other painkillers.

SENECA:
new television commercial for pfizer's painkiller celebrex that has attracted attention for both its length and innovative marketing approach is now being criticized for its message. public citizen, consumer group asks food and drug administration to ban commercial, charging that it gives consumers false impression that prescription drug has no more safety risk than some other painkillers. fda asks pfizer to suspend its television advertising for celebrex

SENECA + R_{Coh}:
new television commercial for pfizer's painkiller celebrex that has attracted attention for both its length and innovative marketing approach is now also criticized for its message. public citizen, asked food and drug administration to ban commercial, charging that it gives consumers false impression that prescription drug has no more safety risk than some other painkillers. fda asks pfizer to suspend its television advertising for celebrex, and company complied

Figure 2: Sample summaries for an NYT article. Comparison model contains grammatical errors; our model is more coherent and with less redundant information. Numbers are replaced with "0".

ARTICLE:
many of us may not have the billionaire lifestyle or scientific prowess of marvel's tony stark, but we can now at least get his trademark hi-tech gloves at home. laser expert patrick priebe created a working iron man-style arm and hand that fires beams from the back of the wrist or from the wearer's palm. and in a video, the contraption is shown popping balloons and lighting matches from feet away. scroll down for video. laser expert patrick priebe built a working iron man-style arm and hand that fires beams from the back of the wrist or from the wearer's palm (pictured lighting a match). the gadget was created by wuppertal-based mr priebe, who designs and builds metal laser gadgets to order online. he posts videos to youtube as anselmofanzero. his laser gloves and laser gauntlet/arm are available to order in varying wavelengths. earlier this year, mr priebe built a laser weapon, inspired by the gatling gun. the gun is made from aluminium and shoots rounds of 1.4-watt class 4 blue lasers. the inventor said, although it is only strong enough to pop balloons, it is still too dangerous to be used outside. a video shows the rapid-fire weapon popping 26 balloons mounted on a wall. eight aa batteries power the laser motor and the gun's turret is powered by four ball bearings controlled by a mechanism under the gun. as the barrel spins, the lasers - powered by lithium-ion batteries - are shot out. these range from the blue 405nm-type laser model to one fitted with 445nm, or 650nm. he also offers a ' low power green' version on his website. the gadgets are powered by lithium-ion cells and output ranges between 0.2 w and 1.2 w. prices are not known. (...)

HUMAN:
iron man-style arm and hand was built by laser expert patrick priebe. it fires beams from the back of the wrist or from the wearer's palm. in a video, the contraption is shown popping balloons and lighting matches . gadget is powered by lithium-ion cells and can be ordered from mr priebe.

POINTGEN + COV:
laser expert patrick priebe created a working iron man-style arm and hand that fires beams from the back of the wrist or from the wearer 's palm. the gadget was created by wuppertal-based mr priebe, who designs and builds metal laser gadgets to order online. he posts videos to youtube as anselmofanzero.

DEEPREINFORCE:
laser expert patrick priebe built a working iron man-style arm. the gadget was created by wuppertal-based mr priebe, who designs and builds metal laser gadgets. patrick priebe created a working iron man-style arm and hand that fires beams from the wearer's palm. laser gloves and laser gauntlet/arm are available to order in varying wavelengths. earlier this year , he built a laser weapon, inspired by the gatling gun. the contraption is shown popping balloons and lighting matches from feet away.

SENECA:
laser expert patrick priebe created a working iron man-style arm and hand that fires beams from the back of the wrist or from the wearer's palm. the gadget was created by wuppertal-based mr priebe , who designs and builds metal laser gadgets to order online.

SENECA + R_{Coh} + R_{Ref} + R_{App}:
laser expert patrick priebe created a working iron man-style arm and hand that fires beams from the back of the wrist or from the wearer's palm. the gadget was created by wuppertal-based mr priebe , who designs and builds metal laser gadgets to order online. in a video, the contraption is shown popping balloons and lighting matches from feet away.

Figure 3: Sample summaries for an CNN/DM article. Our model overlap most with human summaries with information missed by comparisons. Comparison contains sentence which is less coherent and readable .

ARTICLE: if you're in your early 40s, own a flash car and have started listening to taylor swift and one direction, you are likely to be having a midlife crisis. streaming music service spotify believes it has identified the average age of midlife crises at 42. staff analysed data and found users aged around 42 drop their usual playlists – which usually contain hits from their youth – in favour of today's chart toppers from the likes of rihanna and sam smith. streaming music service spotify believes it has identified the average age of midlife crises at 42 (file picture). spotify and its rivals in the streaming music world are working hard to understand the tastes of their listeners, so they can make better recommendations for them (file picture). 'during the teenage years, we embrace music at the top of the charts more than at any other time in our lives. as we grow older, our taste in music diverges sharply from the mainstream up to age 25, and a bit less sharply after that,' explained the company on its insights blog. 'we're starting to listen to "our" music, not "the" music. music taste reaches maturity at age 35. 'around age 42, music taste briefly curves back to the popular charts – a musical midlife crisis and attempt to harken back to our youth , perhaps?' the findings come from a study conducted by ajay kalia, who oversees spotify's 'taste profiles' product, which tries to understand people's tastes based on their listening habits. spotify and its rivals in the streaming music world are working hard to understand the tastes of their listeners, so they can make better recommendations for them. (...)

HUMAN:
spotify believes it has identified the average age of midlife crises at 42. staff analysed data and found users aged 42 drop their usual playlists . start listening to today 's chart toppers , such as rihanna and sam smith

POINTGEN + COV:
streaming music service spotify believes it has identified the average age of midlife crises at 42. spotify and its rivals in the streaming music world are working hard to understand the tastes of their listeners.

DEEPREINFORCE:
spotify believes the average age of midlife crises at 42. staff analysed data and found users aged 42 drop their usual playlists. spotify believes it has identified the average age of midlife. findings come from a study conducted by ajay kalia.

SENECA: staff analysed data and found users aged around 42 drop their usual playlists in favour of today's chart toppers from rihanna and sam smith . streaming music service spotify believes it has identified the average age of midlife crises at 42.

SENECA + R_{App}:
streaming music service spotify believes it has identified the average age of midlife crises at 42. staff analysed data and found users aged around 42 drop their usual playlists – in favour of today's chart toppers from rihanna and sam smith .

Figure 4: Sample summaries for an CNN/DM article. Our models are able to capture important information along with correct entities . Comparisons suffer from either losing important information or redundancy.