Adaptive Pseudo-Parallel Data Generation for Summarization

Johannes Eschbach-Dymanus
University of Heidelberg / Student Nr. 3363713
eschbach@cl.uni-heidelberg.de

1 Introduction

This project investigates the effectiveness of adaptive pseudo-parallel data generation for summarization. The goal is to alleviate the problem of data-scarcity from which the discipline of automated text summarization often suffers. With only few large-scale data sets available, domain, task or language specific models often lack sufficient data to be effectively trained.

The problem of data-scarcity, however, is not exclusive to automatic summarization. In machine translation, language pairs often lack sufficient parallel data to effectively train a model with. Pham et al. (2021) propose to generate synthetic parallel data through Meta Back Translation. In **Back Translation**, a backward model is pre-trained on the available data in target-to-source direction and is then used to generate the source side for unpaired target-language data. The desired forward model's (source-to-target direction) training data is then complemented with those synthetic data pairs. In **Meta Back Translation**, however, the backward model is not only pre-trained, but also continuously updated during the forward model's training to have it generate parallel data which best facilitates the forward model's learning. This can be achieved by training the backward model with rewards through reinforcement learning.

The original motivation behind Back Translation is to increase the language quality of the generated sequences. Sennrich et al. (2015) propose to inflate the data-set with target-language sequences so that the decoder's language modeling can be improved with the additional data. Naturally, the target-language sequences need source-language sequences to be paired with to prevent detrimental effects. To acquire those, the additional target sequences are simply back translated.

In this project, I apply Meta-Back Translation to summarization data and investigate its effectiveness and apparent difficulties. The project's code is based on the descriptions provided by Pham et al. (2021). I want to disclaim that the authors' code is unfortunately not publicly available[1]. Hence, observed shortcomings of the method might entirely be fault of my own implementation.

2 Prior Work

Tardy et al. (2020) have used a Back Summarization approach to generate meeting transcripts for unaligned meeting summaries. Through this procedure they improved Rouge scores of their summarization model by a large margin. To the best of my knowledge, this is the only paper applying such procedure to summarization data. Furthermore, I could not find any paper applying the equivalent of Meta Back Translation to summarization data.

3 Adaptive Pseudo-Parallel Data Generation

Pham et al. (2021) derive their idea from a probabilistic framework. When transferred to the task of text summarization, the motivation behind the procedure can be described as follows: Let $P(S|D)$ be the conditional distribution of summary sequences S and document sequences D. We aim to approximate this distribution with a forward model θ by maximizing the objective:

$$J(\theta) = \mathbb{E}_{s,d\sim P(D,S)} \log P(s|d;\theta) \qquad (1)$$

With $P(D,S) = P(S)P(D|S)$ the equation can be refactored as

$$J(\theta) = \mathbb{E}_{s\sim P(S)}\mathbb{E}_{d\sim P(D|s)} \log P(s|d;\theta) \qquad (2)$$

[1]The link to the code in Pham et al. (2021) is not working, so I have contacted the authors. The authors replied that the release of the code is delayed due to internal approval issues by google research.

If we approximate $P(S)$ by uniformly sampling from a corpus C_s of summaries and furthermore derive an approximate distribution $\hat{P}(D|S)$, the objective becomes

$$J(\theta) = \mathbb{E}_{s\sim C_s}\mathbb{E}_{d\sim\hat{P}(D|S)} \log P(s|d;\theta) \quad (3)$$

The notion of relying on a corpus of summaries C_s to alleviate the problem of data scarcity probably seems paradoxical. However, using summary-like texts such as the lead sentences in documents might be just good enough. Theoretically speaking, any short text could serve as a summary as long as it has high information density. The approximated distribution $\hat{P}(D|S)$ stems from a backward model ψ, which has been pre-trained on the available summarization data-set $\mathcal{D} = \{(s, d)\}$. With $\hat{P}(D|S)$ modeled by $P(D|S;\psi)$ and a data-set $\mathcal{D}$, we can formulate a bi-level optimization problem with an inner loop where we wish to minimize the objective (dependent on ψ)

$$\theta^*(\psi) = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{s\sim C_s}\mathbb{E}_{\hat{d}\sim P(D|S;\psi)}\ell(\hat{d}, s;\theta) \quad (4)$$

with $\ell(\hat{d}, s;\theta) = -\log P(s|\hat{d};\theta)$ as loss function, and an outer loop where we wish to optimize ψ to minimize the loss of $\theta^*(\psi)$ on $\mathcal{D}$

$$\psi^* = \underset{\psi}{\operatorname{argmin}}\, \ell(\mathcal{D}, \theta^*(\psi)) \quad (5)$$

With $\hat{d} \sim P(d|s;\psi)$ sampled and learning rate η, the inner loop's update rule goes as usual

$$\theta_t = \theta_{t-1} - \eta_\theta \nabla_\theta \ell(\hat{d}, s;\theta_{t-1}) \quad (6)$$

Equation 6 directly depends on the sample $\hat{d}$ from the distribution parameterized by ψ. The outer loop must therefore optimize the parameters ψ to minimize the forward model's loss on the training data $\mathcal{D}$:

$$\psi_t = \psi_{t-1} - \eta_\psi \nabla_\psi \ell(\mathcal{D};\theta_t(\psi)) \quad (7)$$

The gradient in 7 can be computed as

$$\nabla_\psi \ell(\mathcal{D};\theta_t(\psi)) \approx - [\nabla_\theta \ell(\mathcal{D};\theta_t)^T \cdot \nabla_\theta \ell(\hat{d}, s;\theta_{t-1})] \cdot \nabla_\psi \log P(\hat{d}|s;\psi) \quad (8)$$

Pham et al. (2021) provide a lengthly derivation for this gradient in their appendix. However, one could also simply interpret it from a reinforcement learning perspective. If we treat Eq. 8 as a REINFORCE update, the reward for generating sample $\hat{d}$ would be

$$R(\hat{d}) = \nabla_\theta \ell(\mathcal{D};\theta_t)^T \cdot \nabla_\theta \ell(\hat{d}, s;\theta_{t-1}) \quad (9)$$

which is simply the dot product between the forward model's gradient on the generated data at time step $t-1$ and the gradient on the actual data from $\mathcal{D}$ at time step t. In other words, we want ψ to generate data pairs which nudge θ into a similar direction as the data from $\mathcal{D}$ does. Naturally, loss and reward are in an inverse relation, hence the minus in equation 8.

4 Implementation

The training loop consists of four steps. To keep the description simple, the text below ignores batch size. The real implementation is compatible with batch updates.

4.1 Sample $\hat{d} \sim P(d|s;\psi)$

In the first step, the backward model ψ receives a summary s from the corpus C_s. The model's forward pass returns the logits over the vocabulary for each token in the output sequence $\hat{d}$. Pham et al. (2021) do not specify how exactly they sample the generated text but simply state that "instead of designing the distribution $\hat{P}(x|y)$ by applying actions such as sampling or beam-search to $P(x|y;\psi)$, we let $\hat{P}(x|y) = P(x|y;\psi)$", which makes it seem like they simply use the argmax over the vocabulary for each token in the output sequence. Later on, however, they keep speaking of the "sampling of pseudo-source $\hat{x}$" and even emphasise the variance caused by the sampling process, which in my understanding contradicts the original statement. In my implementation, I apply the Gumbel-Softmax Trick to each token and then select the argmax word from the vocabulary. I decided to use this approach rather than multinomial sampling, as it allows more flexibility with the REINFORCE update (see 4.4).

Furthermore, Pham et al. (2021) do not clarify whether the text is generated token by token autoregressively, as it would usually be done in a generation process, or whether the word for each token is selected independently. Given that a differentiable token by token generation would be very expensive to have in a training loop (and difficult to implement), I decided to simply sample the words for each token independently.

4.2 Update $\theta_t = \theta_{t-1} - \eta_\theta \nabla_\theta \ell(\hat{d}, s; \theta_{t-1})$

Next, I create the artificial data pair with s as label and $\hat{d}$ as input. The attention mask is created manually by setting it to 0 where $\hat{d}$ is padded and 1 otherwise. This artificial pair is now fed into the forward model θ and the gradient $\nabla_\theta \ell(\hat{d}, s; \theta)$ is cloned in its current state. The model is then updated according to equation 6. While the theoretical analysis in section 3 uses $-\log P(s|\hat{d}; \theta)$ as loss for demonstration purposes, my implementation uses a regular Cross-Entropy loss between the prediction and label.

4.3 Update $\theta_{t+1} = \theta_t - \eta_\theta \nabla_\theta \ell(\mathcal{D}; \theta_t)$

In the third step, a data pair from the summarization data-set $\mathcal{D}$ is fed into the forward model θ and the gradient of the Cross-Entropy loss is stored again. Although not explicitly mentioned in the theoretical part, the forward model is updated with this gradient as well. After all, we aim to only complement our data-set with additional pseudo-parallel data and not purely rely on such data alone.

4.4 REINFORCE ψ

According to Pham et al. (2021), we should now calculate the dot product of the forward model's gradients at the two updates to obtain our reward. If not normalized, however, the dot product can have a considerable range of values - Bart-Base has approximately 140 million parameters - and finding an appropriate learning rate would prove difficult. As the motivation behind the update is to have the pseudo-parallel data push the model weights in the same direction as the data set $\mathcal{D}$, I decided to use the cosine similarity of the two gradients instead. To reduce variance, I subtract the moving average of rewards from the reward to obtain an "advantage-reward".

It is not quite clear again, what exactly (Pham et al., 2021) consider to be the probabilities $P(\hat{d}|s; \psi)$ they want to reinforce. As each word $\hat{d}_1$ to $\hat{d}_K$ in the sequence $\hat{d}$ is independently sampled from the respective token's probability distribution[2], we can simply reinforce them independently.

For each token, it seems intuitive to me to only reinforce the probability of the sampled word $\hat{d}_k$ itself and not the entire probability distribution over the vocabulary the word is sampled from. By applying Gumbel Softmax with minimal temperature to the decoder output scores, I receive a one-hot encoding of the sampled word which can be backpropagated. To instead reinforce the entire distribution for each token, I simply apply a regular Softmax with $\tau = 1$ to the output scores. Pham et al. (2021) mention nothing concrete in their paper, so I decided to simply test both approaches. I will refer to the former configuration with the one-hot encoding as METAHARD and to the latter approach as METADIST. The loss in total is calculated as follows:

$$\ell \approx -1 \cdot \text{cos-similarity}(\nabla_{\theta_{t-1}}, \nabla_{\theta_t}) \cdot \frac{1}{K} \sum_{k=1}^{K} H\left(P_k(V|\psi),\, \mathcal{U}(V)\right) \quad (10)$$

where V is the vocabulary, $P_k(V|\psi)$ is the probability distribution over the vocabulary at token k under model ψ, $\mathcal{U}(V)$ is the uniform distribution over the vocabulary, and $H\left(P_k(V|\psi),\, \mathcal{U}(V)\right)$ is the cross entropy between the two distributions. For METAHARD, $P_k(V|\psi)$ is a one-hot encoding of the sampled word. For METADIST, it is the softmax distribution of the model output scores. As entropy is maximal when the random variable is uniformely distributed, we can also instead rephrase the loss as

$$\ell \approx -1 \cdot \text{cos-similarity}(\nabla_{\theta_{t-1}}, \nabla_{\theta_t}) \cdot \left(\log(|V|) - \frac{1}{K} \sum_{k=1}^{K} H(P_k(V|\psi))\right) \quad (11)$$

If the cosine similarity is negative, the gradient of the loss will push P_k towards uniformity. If it is positive, it will reinforce the current distribution by pushing it away from uniformity.

5 Data

I randomly sampled 3% of data pairs of the CNN/DM summarization data-set (Hermann et al., 2015) to compile $\mathcal{D}$. With roughly 8600 data pairs remaining, we can reasonably speak of "scarce data". The corpus of summaries C_s is created by taking the lead sentences of each document in the CNN/DM set[3]. This approach is reasonable, as

[2] The distributions themselves are of course auto-regressively determined by the decoder, but this taken into account by the loss back-propagation.

[3] I should mention a minor implementation mishap here: For every document I chose as many full leading sentences as possible so that the resulting pseudo-summary contains

Configuration	Rouge1	Rouge2	RougeL	RougeLSum
Baseline	40.981	18.368	28.075	38.017
NoMeta	**41.433****	**18.903****	**28.660****	**38.531****
MetaHard	40.771	18.344	28.007	37.844
MetaHard-Pos	40.978	18.497	28.138	37.994
MetaDist	**41.304***	**18.760***	**28.574****	**38.380***
MetaDist-Pos	40.847	18.353	28.012	37.906

Table 1: Rouge evaluation metrics on the CNN/DM test-set. Boldfaced scores outperform the baseline with statistical significance (*$p < 0.1$ and **$p < 0.05$)

when confronted with data scarcity in a certain domain, one would usually lack summaries and not the documents themselves.

6 Models & Parameters

I use the pre-trained Bart-Base model by Facebook (Lewis et al., 2019) as basis for all fine-tuning. First, the backward model ψ is fine-tuned on the selected CNN/DM subset. For fine-tuning, I use batch size 4 per device, learning rate $5\mathrm{e}{-5}$ with 400 warm-up steps and then polynomial decay.

The same setting is used for all training configurations of forward model θ, except that the batch size per device is reduced to 2 due to memory issues. For the continuous updating of the backward model ψ during training, I reduced the learning rate to $5\mathrm{e}{-6}$, as the reinforcement learning acted unstable at higher learning rates.

I compare three different training configurations. The first one, referred to as NoMeta, utilizes the data pairs generated by the pre-trained backward model ψ, but does not continue to update ψ. MetaHard and MetaDist further update the backward model during training, but differ in terms of implementation of the reinforcement update. The configuration MetaHard uses the Gumbel Softmax Trick with $\tau = 0$ to effectively reinforce only the probability of the sampled word for each token[4]. The configuration MetaDist reinforces each generated token's probability distribution over the vocabulary.

less than 148 tokens, the maximal target length set for the models. Unfortunately, I did not use the BART tokenizer for this, but an NLTK one. This was rather clumsy, as the BART tokenizer byte-pair encodes the inputs. Ultimately, this resulted in many of the pseudo-summaries having more than 148 "BART tokens" which in turn caused them to be truncated when fed as input into the backward model

[4]In PyTorch, we can achieve this effect by setting *hard* = *True* in torch.nn.functional.gumbel_softmax

7 Results

Table 1 shows the Rouge metrics achieved by the models on the full CNN/DM test set. The forward model trained with the NoMeta and MetaDist configuration outperformed the baseline, a Bart-Base model finetuned regularly on the scarce data $\mathcal{D}$. Both models outperform this baseline with statistical significance with error chance $p < 0.05$ in all metrics. The backward models ψ of configurations MetaHard-Pos and MetaDist-Pos are trained with only positive reward by clamping the loss of the update to a maximum of 0. In both cases, the performance deteriorated.

Overall, Meta Back Summarization does not bring any advantage over the more straight forward Back Summarization approach. If Meta Back Summarization is conducted however, it is more effective to reinforce the modeled probability distribution the word is sampled from rather than just the probability of the word.

8 Discussion & Learned Lessons

The reinforcement learning of backward model ψ does not achieve the desired result. Right away from the start, the reward is more often than not lower than its current moving average. In other words, the loss consistently increases rather than decreases throughout the training. Despite the design of the update described in equation 11, however, the generated documents do not turn into random text, but instead get shorter and shorter until all that remains are copies of the summaries. It appears that the decoder ends up unlearning its pre-training and defaults to the encoding as exemplified in table 2. Both MetaHard and MetaDist display this behaviour. I attempted to counteract it by clamping the loss to a maximum of 0, thereby updating ψ with positive rewards only. As shown in table 3, this causes the decoder to lose its language modeling capacity and makes it produce pure noise.

It is hard to explain why the Rouge scores of METADIST are superior to the baseline and almost match the ones of NOMETA. A quick look at the generated documents does not grant any insights to why it outperforms METAHARD, as both configurations generate nearly identical documents (see table 4). I have not conducted repeated measurements nor optimized hyper-parameters, so it might simply be that the models are unreliable in that regard and the current configuration of hyper-parameters just happened to work well with METADIST but not with METAHARD.

Ultimately, I must declare that I was unable to successfully transfer "Meta Back Translation" to the task of automatic summarization. I am unsure whether this is due to a possibly badly formulated reinforcement reward or due to the inherit difference between the two tasks.

The reinforcement update stated in Equation 11 was the result of weeks of trial an error. What I later realized, however, was that there seems to be a general consensus that reinforcement learning with transformers is difficult (Parisotto et al., 2019) and specialized architectures have recently been developed (Chen et al., 2021). Initially, I tried to stay as close as possible to Pham et al. (2021), but the output by the backward model ψ degenerated almost immediately into noise. Replacing the dot product between the gradients with the cosine-similarity and introducing advantage based reinforcement learning delayed the degeneration during training, but did not quite solve it. Initially, I also tried to reinforce the output distribution over the vocabulary the following way

$$\ell \approx -1 \cdot \text{cos-similarity}(\nabla_{\theta_{t-1}}, \nabla_{\theta_t}) \cdot \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|V|}\sum_{i=1}^{|V|} P_k(v_i|\psi) \quad (12)$$

which makes very little sense, as the sum of probabilities from a distribution always add up to exactly one. Comparing the distribution over the vocabulary $P_k(V|\psi)$ with the uniform distribution through a Cross Entropy measurement (as formulated in Equation 10 and simplified in Equation 11) significantly improved the generation.

There are probably a number of control mechanisms I could have tried to implement. For instance, I could have kept updating the backward model with the data pairs from the ground truth data $\mathcal{D}$ to counteract the degeneration. Alternatively, I could have included some sort of length controllability to punish ψ for creating shorter and shorter documents. I decided not to do any of those, as it simply would feel like damage control for a failed experiment at this point.

The alternative explanation would be that the entire procedure simply does not work on summarization data. There could be a number of reasons to why not: In a translation setting, the source and target language contain the equivalent information. In summarization, the backward model is tasked to hallucinate an entire document, which is a much harder task. The original purpose of "BackTranslation" was furthermore to increase the language quality of the target side by inflating the data set with target side sequences. In summarization data, it is the target side which is inherently scarce, so I had to rely on pseudo summaries generated from the lead sentences of the documents. There might be a too large difference between such pseudo summaries and true summaries, which would render the procedure useless.

Both of these suspicions, however, are invalidated by the good performance of the NOMETA configuration. Despite the lack of proper summaries, the pre-trained backward model generated document-summary pairs which boosted the performance of the forward model. The generated documents most often contain the information from the summary redundantly over and over, phrased in various ways (see table 4 and 2). My assumption is that such trivial data pairs allow the forward model θ to improve its capacity to abstractively rephrase information. The scarce ground truth data $\mathcal{D}$ alone might not provide enough training data for this subtask.

If anything in the training procedure is incompatible with summarization data at all, it must therefore be the "Meta" aspect, the continuous updating of the backward model through reinforcement learning. One possible explanation would be that in a translation setting, source and target side have a different vocabulary and therefore the copy-paste behaviour the backward model ψ degenerated to during training cannot even occur. In fact, whatever caused it, might even be desirable for translation.

As a control, I should apply my training setup to a translation task and confirm whether Pham et al. (2021) results can be reproduced this way. Unfortunately I was not able to do this within the time-frame set for this project.

9 Conclusion

When facing scarce data in a generic summarization setting, we can boost the performance of our summarization model by supplementing the data with pseudo-parallel data pairs generated from summary-like texts. However, the backward model which generates the documents from the summary-like texts can not be successfully further fine-tuned through continuous updates during the summarization model's training. This is likely due to difficult nature of reinforcement updates. Further research could be done on whether Meta Back Summarization can be employed in more simple tasks such as generating queries for document-summary pairs in a query-focused summarization setting.

References

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *CoRR*, abs/2106.01345.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Emilio Parisotto, H. Francis Song, Jack W. Rae, Razvan Pascanu, Çaglar Gülçehre, Siddhant M. Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, Matthew M. Botvinick, Nicolas Heess, and Raia Hadsell. 2019. Stabilizing transformers for reinforcement learning. *CoRR*, abs/1910.06764.

Hieu Pham, Xinyi Wang, Yiming Yang, and Graham Neubig. 2021. Meta back-translation.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data.

Paul Tardy, Louis de Seynes, François Hernandez, Vincent Nguyen, David Janiszek, and Yannick Estève. 2020. Leverage unlabeled data for abstractive speech summarization with self-supervised learning and back-summarization. *CoRR*, abs/2007.15296.

Table 2: Effect of Reinforcment updates on ψ (METAHARD)

Generated Document at Training-Iteration 100

Summary: ”And he wonders why it is always him. Just when you thought you had seen it all with Mario Balotelli, he takes the breath away on Anfield’s biggest European night in five years. With Liverpool submerging at half-time, the evening having become an exercise in damage limitation rather than one to tell the grandchildren about as they trailed 3-0, Balotelli decided to swap shirts with Real Madrid defender Pepe as he sauntered off towards the dressing room. Not clever. Not clever at all. Having endured a dismal, uncomfortable start to his time on Merseyside, Balotelli had a glorious chance to show why Brendan Rodgers took a £16million gamble on him in August and made him a

Generated Document: (CNN) – Mario Balotelli is the man who takes the breath away on Anfield’s biggest European night in five years. He is the player who has taken the breath of life on Merseyside for the past five years, and he is the one who takes it all away on the biggest European evening in five-years. And he has done it again. With Liverpool submerging at half-time, the evening having become an exercise in damage limitation rather than one to tell the grandchildren about as they trailed 3-0, Balotell decided to swap shirts with Real Madrid defender Pepe as he sauntered off towards the dressing room. Not clever. Not stupid at all. Having endured a dismal, uncomfortable start to his time on Merseyide, Balottelli had a glorious chance to show why Brendan Rodgers took a £16million gamble on him in August and made him a star. He was the man that took the breath from Liverpool’s greatest European night since August. He has taken it away for the last five years and he has been the one that takes it away from the club. VIDEO Scroll down to watch MarioBalotelli take on Real Madrid in the Champions League. He has been a star for Liverpool since August, but he has not been the man to show the grandchildren that he can be a star. He is a man who has been an inspiration for Liverpool fans since August. He was one of Brendan Rodgers’ first signings in the summer, but now he has become the man they want to see most. He will be missed by the rest of the squad, and the club will be without him for a long time. But he is not the only one who has made the most of his time at Anfield. He can also be the man the club want to be without, and that is exactly what he did on Saturday night. He had been part of the Liverpool squad for the first time since the summer transfer window opened in August, and it was a glorious opportunity to show that Brendan Rodgers has taken a huge gamble on the player in August. And it is a chance that he has taken away from Liverpool. He took a huge chance on the evening with Liverpool’s second-half comeback against Real Madrid at Anfield on Saturday evening. He took the moment to show how Brendan Rodgers (left) took a big gamble on Mario Balottell (right) in August. It was his first European game since joining Liverpool from Real Madrid. He had a dismal start to the season, but the club have now made him one of the most important players in the world. He made a glorious impression on the Liverpool fans as they took to the pitch to celebrate with him on Saturday. It is a moment that has taken Liverpool fans by storm. It is the moment when Brendan Rodgers takes a big chance on him and makes him a superstar. It has taken Brendan Rodgers a great deal of time to take a big risk on him, but it has taken him a great amount of time. It was a moment when he took a massive chance on Liverpool’s biggest European moment in five year. It took him a lot of time and effort to get the ball out of his hands, but when he did, he took it away. He showed that he is a player who can be an inspiration to Liverpool fans, and not just because of the way he plays, but also because he has shown that he wants to be a hero for the club, and has shown why he is so much more than just a player. He shows that when he plays for Real Madrid, he is also a player that wants to play for the team, not just for the players, but for the fans, too. And that is why he has made such a huge impact on the club and has been such a part of so many of the team. It has been one of his most memorable moments in his career. The man who made the greatest impact on Liverpool has been so important to the club that he will always be remembered as one of Liverpool’s most influential players. It will be remembered by the club as the man of the year, and by the fans as the person who helped bring the club to the top of the English Premier League. [...]

Generated Document at Training-Iteration 3000

Summary: ”By. Meghan Keneally. It appears Kris Humphries has struck a deal to auction off the engagement ring that he gave to Kim Kardashian during their ill-fated romance, MailOnline has learned. The NBA star is said to be selling the Lorraine Schwartz ring that he gave to Kardashian at an upcoming Christie’s auction in New York. The ring is expected to sell for between $300,000 and $500,000- which is less than a quarter of the price that the ring was said to have cost originally. Bling: Kim’s former 20-carat sparkler was used by Humphries to propose to Kim back in May 2011. Return: Kim returned her former engagement ring to Kris following their divorce

Generated Document: By. Meghan Keneally. It appears Kris Humphries has struck a deal to auction off the engagement ring that he gave to Kim Kardashian during their ill-fated romance, MailOnline has learned. The NBA star is said to be selling the Lorraine Schwartz ring that she gave to Kardashian at an upcoming Christie’s auction in New York. The ring is expected to sell for between $300,000 and $500,000- which is less than a quarter of the price that the ring was said to have cost originally. Bling: Kim’s former 20-carat sparkler was used by Humphries to propose to Kim back in May 2011. Return: Kim returned her former engagement ring to Kris following their divorce

Table 3: Effect of Only Positive Reinforcment Updates on ψ (METAHARD-POS)

Generated Document at Training-Iteration 8500

Summary: "Two former presidents reflected on their greatest regrets in office Monday, each looking back to issues that continue to plague the nation years later. Former presidents and political rivals Bill Clinton and George H.W. Bush now share philanthropic efforts. Former Presidents George H.W. Bush and Bill Clinton appeared together at a question-and-answer forum before the National Automobile Dealers Association in New Orleans, Louisiana. Asked his biggest regret after leaving office, Bush said he now wonders whether he should have tried to get Saddam Hussein to leave office at the end of the first Gulf War in 1991.

Generated Document: By. Daily Daily Mail Daily Daily Daily News Daily Daily Page Daily Page Page Page Daily Daily New Daily Daily NY Daily NY NY Daily Daily Post Daily New NY Daily News NY Daily New New NY NY NY New Daily NY New New New Daily Post NY Daily Page NY Daily Post New NY New NY Post Daily Daily & Daily Daily N NY Daily & NY Daily Times NY Daily Mail NY NY Post NY NY & NY NY Times NY NY News Daily NY & Daily NY Post New Daily & New NY Times New NY & New Daily News New Daily New & Daily New Post NY New & NY New N NY NY/ Daily NY/N NY Daily/ Daily Daily/ New NY/ NY Daily / Daily NY / NY NY / New NY / Daily Daily Times New Daily/ NY NYN NY/New NY NY and NY NYE NY/F NY/G NY/E NY NY-N NY NYF NY NYG NY NYH NY NYS NY NYA NY NYR NY NYP NY NYO NY NYD NY NYNY NY NYM NY NYW NY NYB NY NYT NY NYL NY NYJ NY NYZ NY NY N NYE New NYN New NYF New NYE N NYN N NYF N NYG N NYA New NYP New NYNYN NYE Daily NYN. NY NYFW NY NYY NY NYU NY NYC NY NY NEW NYE NEW NYN NEW NY NY EST NY NYAL NY NY LA NY NYK NYNN NYN USA NY NYLA NY NYX NYN EST NYN U NY NY USA NYN LA NYNNY NYN & NYE EST NYF NEW NYG NEW NYS NEW NYNY NEW NYA NEW NYH NEW NY NEW NEW NEW NYO NEW NYF LA NYE LA NYS N NYH N NYS EST NYE USA NYG LA NYG New NYG USA NYE NYC NY NY NYC NYN NYC NYE & NYN COM NY NYV NY NY. NY NYWW NY NY. NYN Daily NYNYE NYNH NYNY NYNG NYNF NYNZ NYNA NYNE NYS New NYA N NYNYS NYNU NYNO NYNL NYNJ NYNW NYNP NYNAL NYNOL NY NYI NYNS NYE. NYENY NYFNY NYEN NYGNY NYGN NYNYF NYE COM NYN DE NY NYOL NYNR NYNKY NYN WN NYO New NYS USA NYF USA NYP NEW NYL NEW NYZ NEW NYR NEW NYM EST NYNYNY NYS LA NYP N NYZ N NYO N NYYN NYS. NYSNY NYPNY NYNYG NYG EST NYG COM NYE U NYN. NYNT NYNEX NYNUS NYNOT NYNIN NY NYOT NY NYEX NY NY COM NYS COM NYNYA NYE WNNN NEW NEW N NYW NEW NYYG NYE DE NEW NYAL NEW NYLA NEW NYU NEW NYW N NYM NEW NYT NEW NY N NEW NY COM NEW NYOLY NYE PLN NYA USA NYO EST NYS U NYE G NYN TRANS NY NYRO NYN CALL NY NYAZ NYN P NYN PLNN EST NEW NYB NEW NYP EST NYA EST NYW EST NYO USA NYS WNNY NEW NEW U NYG U NYH EST NYR EST NYM N NYR N NYP USA NYA. NYGG NYF EST NYHN NYZN NYFN NYHNY NYHG NYPN NYWN NYPY NYG WNYE NYE F NYNAR NYNAG NYN CENT NY NYUSA NYNX NYE P NYEE NYG G NY NYEA NYN PO NYNCH NYNET NYNRO NY NYSC NYNM NYNAC NYN/ NYNQ NYNWW NYN-NYN NEW N NEW NEW COM NYG PLNY NYO COM NYF COM NYP COM NYZ EST NYL EST NYYZ NYE IN NEW NYFW NEW NYD NEW NY TRANS NYN EX NY NYKYN NYYO NYE TRANS NYNYO NYS TRANS NYE EX NYNUG NYN G NYE AWN NYR U NYF U NYS P NYG P NYNYH NYE R NYN F NYE E NYN O NYE H NYN H NYE O NYN IN NY NYET NY NYOR NYNOR NYE BU NYN BAR NYN CON NYE CON NY NYST NYN WWN NYU U NYYF NYG H NYS BAR NYE BAR NY NYON NYE WWNN NN NYM U NYNYP NYE ZNNNYN N N NY

Table 4: Comparison of Backward Models after Training Completion

Sample Data from CNN/Daily Mail

Summary: ”There has been an unusually high number of sea lions stranded since January,” NOAA representative says.speculation is mothers are having difficulty finding food, leaving pups alone too long or malnourished.

Document: (CNN)Wildlife services in California are being pushed to their limits this year. Since January 2015, every month has set a record in sea lion ”strandings,” mostly sea lion pups, according to the National Oceanic and Atmospheric Administration. ”There has been an unusually high number of sea lions stranded since January,” said Justin Greenman, assistant stranding coordinator for NOAA on the West Coast. ”Stranding does happen, but just to give you perspective, 1,800 [sea lion] pups have been responded to this year alone. We responded to 1,600 strandings total during the entire year in 2013,” he said. Stranding is the official term to describe marine life that ”swim or float into shore and become beached or stuck,” according to NOAA. Strandings are taking a toll on the resources available in coastal counties from San Diego to Santa Barbara. Local care facilities have taken in more stranded sea lions this year than 2004-12 combined, and it is only mid-March. Greenman said he expects the problem to continue beyond April, when weaning normally occurs, when the pups are 10 or 11 months old. Dave Koontz, director of communications for SeaWorld San Diego, said SeaWorld has rescued nearly 500 sea lions this year. ”This is a new record for Sea World,” Koontz said. ”In 1983 we rescued 474.” Some of the sea lions responded to have had to be euthanized. ”They [sea lion pups] have to be able to eat and fish on their own before they can be released back into the wild, and a lot of these pups haven’t even been weaned,” Greenman said. Greenman said California has had warmer weather than usual this year, and, while NOAA is still conducting studies on the Channel Islands to get a more proven explanation, warmer water drives the food source farther out or deeper into the ocean, where the colder water is. When food is farther away, the mothers are away from the pup too long in search of food, and return with little food or too few nutrients for a growing sea lion. ”We have been seeing emaciated or dehydrated sea lions show up on beaches,” Greenman said. However, he said, the species has made a comeback since the Marine Mammal Protection Act of 1972. California’s sea lion population has grown to 300,000 from an estimated population of 10,000 in the 1950s, according to the Washington Department of Fish & Wildlife. People who observe stranded sea lions are advised not to touch them or attempt to rescue them, because it can be dangerous and it is illegal. Instead, call any of the rescue agencies listed on NOAA West Coast Region’s website. If the animal has died, the local dead animal pickup service should be alerted.

Generated Documents

METADIST: ”There has been an unusually high number of sea lions stranded since January,” NOAA representative says. The speculation is mothers are having difficulty finding food, leaving pups alone too long or malnourished.

METAHARD: (CNN) – There has been an unusually high number of sea lions stranded since January,” NOAA representative says. The speculation is mothers are having difficulty finding food, leaving pups alone too long or malnourished.

Pre-Trained/NOMETA: (CNN) – There has been an unusually high number of sea lions stranded since January, according to the National Oceanographic and Atmospheric Administration (NOAA). The number is likely to rise as mothers are having difficulty finding food, leaving pups alone too long or malnourished, the agency said in a statement. The agency said the number of stranded sea lions has been increasing since January. ”This is a result of the fact that mothers have been unable to find food for their pups,” the agency statement said. ”It is a reflection of the increasing number of animals stranded in the wild.” The agency has been monitoring the situation for several months, but the agency has not released a cause for concern. ”We do not have a specific cause,” the statement says. ”However, we do have a number of factors that may be contributing to the increase in the numbers of stranded animals.” The NOAA statement says the agency is working with the U.S. Fish and Wildlife Service to determine the cause of the increase. ”There is no indication that this is related to the lack of food for the sea lions, but it is possible that this may be related to a lack of availability of food,” the NOAA statement said in an email. ”In addition, there may be an increase in food availability for the animals, as well as a decrease in the availability of water for the pups.”