Semantic Extractor-Paraphraser based Abstractive Summarization

Anubhav Jangra*
IIT Patna, India
anubhav0603@gmail.com

Raghav Jain*
DTU, India
raghavjain106@gmail.com

Vaibhav Mavi*
IIT Delhi, India
vaibhavg152@gmail.com

Sriparna Saha
IIT Patna, India
sriparna.saha@gmail.com

Pushpak Bhattacharyya
IIT Bombay, India
pushpakbh@gmail.com

Abstract

The anthology of spoken languages today is inundated with textual information, necessitating the development of automatic summarization models. In this manuscript, we propose an extractor-paraphraser based abstractive summarization system that exploits semantic overlap as opposed to its predecessors that focus more on syntactic information overlap. Our model outperforms the state-of-the-art baselines in terms of ROUGE, METEOR and word mover similarity (WMS), establishing the superiority of the proposed system via extensive ablation experiments. We have also challenged the summarization capabilities of the state of the art Pointer Generator Network (PGN), and through thorough experimentation, shown that PGN is more of a paraphraser, contrary to the prevailing notion of a summarizer; illustrating it's incapability to accumulate information across multiple sentences.

1 Introduction

Over the past few years, the Internet has become the most convenient and preferred form of information sharing worldwide. The evolution of technology has made it possible for anyone to convey their knowledge, opinions and ideals to the world, resulting in an increasing surge of information hindering users from accessing desired content. This increasing need to obtain key information makes the task of summarization paramount. Text is the most widely adopted form of communication, be it for personal messaging[1] or for broadcasting, owing to its ability to convey almost any concept, its general flexibility to suit everyone's needs, and its less storage requirement (opposed to other modes of communication like audio and video). Text summarization is a problem at the very core of natural language processing, and has various applications in the spoken languages, including summarization of conversations, and public speeches.

Some works have been done in the field of reinforcement learning based text summarization (Dong et al., 2018; Liu et al., 2018), the most prominent architecture being extractor-abstractor (EXT-ABS) model (Chen and Bansal, 2018). Inspired from this architecture, in this manuscript, we have proposed an extractor-paraphraser system that uses semantic information overlap as the underlying guidance strategy. The model is further enhanced to surpass its limits using reinforcement learning, for which we have proposed a novel semantic overlap based reward function. Word Mover Similarity (WMS) (Clark et al., 2019) is utilized to evaluate semantic similarity across generated sentences and the true ground truth summary sentences.

We assume that paraphrasing is a relatively simpler task than abstractive summarization, with the underlying intuition that paraphrasing is a subproblem within abstractive summarization. To bolster our hypothesis, experiments are conducted on the extractor-abstractor (EXT-ABS) model (Chen and Bansal, 2018) and the Pointer Generator Network (PGN) (See et al., 2017), which is used as the basic abstraction unit in the former architecture. The results are rather staggering and reveal that the PGN model also paraphrases input document sentences, albeit implicitly. The major contributions of the paper are as follows:

- A novel semantic overlap based reward function is proposed for reinforcement of extractor-paraphraser model.
- To the best of our knowledge, we are the first ever to discover the fact that PGN networks are indeed doing an implicit extraction-paraphrasing operation, revealing the true nature of existing abstractive summarization models.

* means equal contribution.

[1] https://news.gallup.com/poll/179288/new-era-communication-americans.aspx

Proceedings of the 17th International Conference on Natural Language Processing, pages 191–199
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)

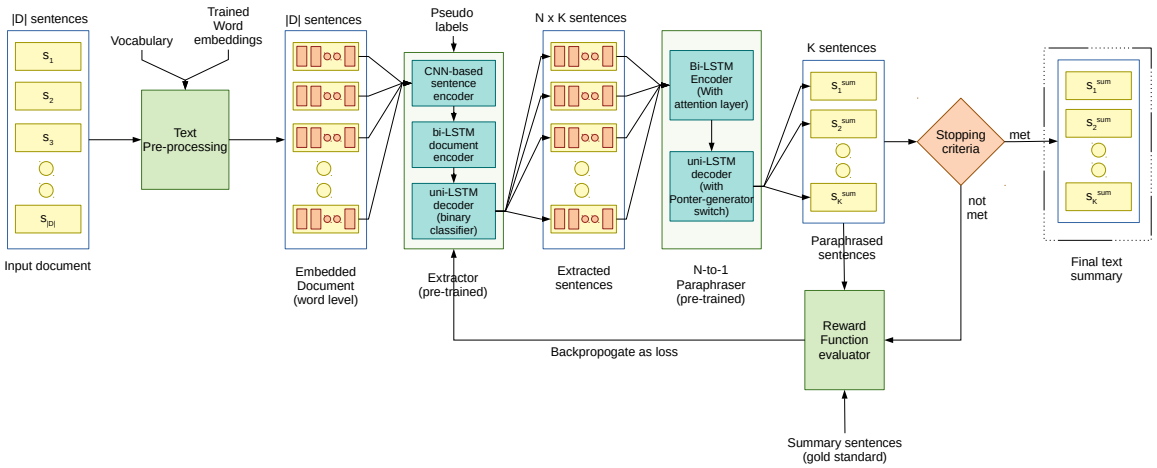

Figure 1: Proposed model architecture.

The rest of this paper is structured as follows: In Section 2 we have discussed related works of automatic text summarization. In Section 3 we have described the proposed model, and in Section 4 we have stated the experimental setup and the datasets used. A thorough discussion and state the results are provided in Section 5, followed by the conclusion and future work in Section 6.

2 Related Work

Automatic text summarization has been extensively researched over more than three decades, and has shown a lot of progress and promise over the course of time. Various approaches have been explored to tackle both extractive and abstractive summarization. Initial research (Paice, 1990; Kupiec et al., 1995) focused on extractive summarization due to its easier setup. Various techniques ranging from integer linear programming (Galanis et al., 2012), graph based approaches (Mihalcea and Tarau, 2004; Mihalcea, 2004), genetic algorithms (Saini et al., 2019a,b), and neural networks (Nallapati et al., 2017; Zhang et al., 2016) have been adopted to solve the extractive summarization task. The majority of the research in abstractive summarization revolves around deep learning (See et al., 2017; Chopra et al., 2016; Nallapati et al., 2016). Liu et al. (2018) proposed a generative adversarial network based model to generate document abstracts. A handful of works however also use ILP (Banerjee et al., 2015) and graph-based (Ganesan et al., 2010) techniques to attempt to solve the problem. A lot of domain specific summarization techniques have also been explored, like radiology findings summarization (Zhang et al., 2018), across-time summarization (Duan and Jatowt, 2019), movie review summarization (Zhuang et al., 2006), book summarization (Mihalcea and Ceylan, 2007), and customer review based opinion summarization (Pecar, 2018). Lately, multi-modal summarization (Jangra et al., 2020a,b; Zhu et al., 2020; Saini et al., 2020) has also gained popularity .

Recently, people have also explored reinforcement learning to tackle the problem of automatic text summarization in both extractive (Dong et al., 2018; Gao et al., 2019) and abstractive domains (Xiao et al., 2020; Chen and Bansal, 2018). Chen and Bansal (2018) have proposed an extractor-abstractor architecture, separating the relevant data searching part and the paraphrasing part to individual modules. In this work, we have proposed a system inspired from Chen and Bansal (2018), stressing on the significance of semantic information over the traditional syntactic overlap. The literature on text summarization is rich, and has an abundance of survey papers (Yao et al., 2017; Gambhir and Gupta, 2017) to get an in depth overview of the domain.

3 Proposed Method

Problem Definition: Given the training data {X, Y} where $X = \{d_1, d_2, ..., d_N\}$ is the set of input documents and $Y = \{y_1, y_2, ..., y_N\}$ is the set of corresponding output summaries, the task of automatic summarization is defined as the problem of discovering a function $f : X \mapsto Y$, such that $f(d_i) = y_i; \forall i \in \{1, 2, ..., N\}$.

We have proposed an extractor-paraphraser framework, which is inspired from the extractor-abstractor (EXT-ABS) framework introduced by

Chen and Bansal (2018). The summarization function $f(.)$ is approximated as the composition, $f(d_i) = h(g(d_i))$, where the functions $g(.)$ and $h(.)$ are modeled as the *extractor* and the *paraphraser* components of the model, respectively. Given an input document $d_i = \{s_1^{d_i}, s_2^{d_i}, ..., s_{|d_i|}^{d_i}\}$, the extractor $g(.)$ extracts relevant sentences, acting as the primary noise filter. These extracted set of sentences are fed to the paraphraser $h(.)$, which accumulates the information into a concise gist of the extracted sentences, simulating the natural language generation module in the proposed system (Fig. 1). The paraphraser in our system is capable of summarizing multiple extracted sentences to generate one sentence, in contrast with its predecessor 'abstractor' from EXT-ABS model (Chen and Bansal, 2018) which rephrases one extracted sentence at a time. The proposed framework consists of an 'n-to-one paraphraser', that compiles n sentences into one sentence, generating a richer summary. Formally, given a document $d_i = \{s_1^{d_i}, s_2^{d_i}, ..., s_{|d_i|}^{d_i}\}$, the extractor is defined as:

$$\begin{aligned} &g : X \mapsto Z^{|y_i| \times n}, \\ &\text{s.t. } g(d_i) = \{k_{j,l} \,|\, 1 \leq k_{j,l} \leq |d_i|\}_{l=1,j=1}^{n,|y_i|} \end{aligned} \tag{1}$$

where $k_{p,q}$ represents the index of the q^{th} extracted sentence corresponding to the p^{th} sentence in the final gold summary. For modelling the same, we have used an encoder-decoder model, where the encoder consists of a temporal convolutional model (Kim, 2014) cascaded with a bidirectional LSTM network (Schuster and Paliwal, 1997) and the decoder is a uni-directional LSTM model (Hochreiter and Schmidhuber, 1997) (Fig. 1).

The paraphraser function $h(.)$ is defined as follows:

$$h(\{k_{j,l}\}_{l=1,j=1}^{n,|y_i|}, d_i) = \{p(\oplus_{l=1}^{n} s_{k_{j,l}}^{d_i})\}_{j=1}^{|y_i|} \tag{2}$$

where $\oplus$ represents the concatenation of sentences, $k_{j,l}$ represents the extractor output and p is written as:

$$p(s_j^{ext}) = s_j^{y_i} \tag{3}$$

where $s_j^{ext} = \oplus_{l=1}^{n} s_{k_{j,l}}^{d_i}$ and $s_j^{y_i}$ is the j^{th} gold summary sentence. As shown in Fig. 1, a pointer-generator framework (See et al., 2017) is used to model the paraphraser.

3.1 Training the Submodules

In order to train the paraphraser beforehand, we need to mimic the output of an extractor[2]. To fulfil this requirement, *exemplary-extracted sentences* are generated using gold summaries. Word-mover distance (WMD) (Kusner et al., 2015) is adopted to create these exemplary-extracted sentences, with the motivation that it captures the semantic overlap between sentences better than its predecessors. For a training pair $\{d, y\}$, the labels are generated as:

$$\begin{aligned} &\forall l \in \{1, 2, ..., n\} : \\ &\quad \forall s_j \in y : \\ &\quad\quad k_{j,l} = argmin_i\{WMD(s_i^d, s_j^y)\}; \\ &\quad\quad \forall s_i^d \in d - \{s_{k_{j',0}}^d\}_{j'=1}^{j-1} \end{aligned} \tag{4}$$

where $k_{j,l}$ represents the l^{th} exemplary- extracted sentence corresponding to the j^{th} gold summary sentence, s_j^y.

The n exemplary-extracted sentences corresponding to each summary sentence are concatenated and fed into the paraphraser as the input and the summary sentence is fed as the output. We argue that this should allow the paraphraser to generate information rich summaries. To facilitate the expected behaviour during the testing phase, we require the extractor to feed the paraphraser with input similar to the exemplary-extracted sentences. Hence, we first pre-train the extractor on these exemplary-extracted sentences as opposed to random initialization. Cross-entropy loss is used for training the paraphraser and pre-training the extractor.

3.2 Extractor agent

Enhancement using reinforcement learning: Here, the extractor is trained as part of an actor-critic model (Mnih et al., 2016) which takes an action based on the current state and current value of parameters to maximize a given reward at each time step, where the action is to extract n sentences, $\{k_{t,l}\}_{l=1}^{n}$, and the state refers to the set of document sentences, d_i, and already extracted sentences, $\{d_{k_{1,l}}, d_{k_{2,l}}, ..., d_{k_{t-1,l}}\}$; $l \in \{1, 2, ..., n\}$. The predicted sentences are concatenated and passed to

[2]Note that the extractor's output can also be used to train the paraphraser at this step, however, since the extractor's weights get fine-tuned using reinforcement learning at a later stage, training paraphraser on extractor's non-ideal outputs might lead to unsatisfactory performance of the paraphraser.

the paraphraser to get an output sentence. A novel semantic-based reward function using word mover distance (WMD) is used as the reward function. Since the reward is to be maximised, WMD needs to be converted to a similarity function, for which, a generalised version of word mover similarity (WMS), proposed in (Clark et al., 2019), is used. Formally, at a time step t, given an action $j_{t,l}$, and a summary sentence, s_t, the short term reward, r is calculated as[3]:

$$r = \frac{a+1}{a + e^{b \times WMD(s_t^y, p(s_t^{ext}))}} \quad (5)$$

where $s_t^{ext} = \oplus_{l=1}^{n} s_{k_{t,l}}^{d}$, and $\oplus$ represents the concatenation of sentences, $\{s^d, s^y\}$ are the sentences belonging to a training pair and a and b are the hyper-parameters introduced[4].

To avoid redundant phrases and words, tri-gram avoidance through beam search (Paulus et al., 2017) is applied at a sentence level. For a fair comparison, details regarding the beam search reranking and other nuances of implementation are kept the same as (Chen and Bansal, 2018).

4 Experiments

4.1 Dataset

For all the following experiments we have used the CNN / DailyMail dataset (Nallapati et al., 2016), which contains online news articles, with the bullet highlights treated as the gold standard summaries. The experiments in this work are conducted on the non-anonymized version of this dataset. The dataset consists of 277,226 training, 13,368 validation and 11,490 test article-summary pairs. An article contains $\sim$780 tokens per document, whereas the summary consists of $\sim$ 56 tokens with the average number of sentences per summary being $\sim$3.75. An article sentence, on average contains $\sim$30 tokens.

4.2 Comparative methods

To highlight the superiority of the proposed semantic-overlap based methodology over existing syntactic measures, we set the value of $n = 1$ for initial experiments. To compare the complexity of the summarization task with paraphrasing objectively, the experiments are further extended for $n = 2$ as well [5]. We evaluate our model with sufficient baselines [6], including:

Pointer-Generator network [PGN]: An encoder-decoder attention based framework for abstractive summarization (See et al., 2017).

Pre-trained extractor - Extractor-Abstractor framework [EXT − ABS (Ext only)]: The pre-trained extractor of the extractor-abstractor framework proposed by Chen and Bansal (2018) on ROUGE-L based exemplary-extracted sentences.

Extractor-Abstractor framework without reinforcement [EXT − ABS (w/o RL)]: The extractor-abstractor framework proposed by Chen and Bansal (2018) at a stage before reinforcing the extractor.

Extractor-Abstractor framework [EXT− ABS + $\mathbf{RL_X}$]: The complete extractor-abstractor framework proposed Chen and Bansal (2018) including X as the reward function for reinforcement learning (X is either ROUGE-L or the reward function defined in Eq. 5) along with beam search.

Pre-trained extractor - One-to-one extractor-paraphraser model [O2O (Ext only)]: The pre-trained extractor of the proposed extractor-paraphraser framework (with $n = 1$) on WMD based (Eq. 4) exemplary-extracted sentences.

One-to-one extractor-paraphraser model without reinforcement [O2O (w/o RL)]: A particular setting of the proposed methodology where $n = 1$, without any reinforcement or beam search.

One-to-one extractor-paraphraser model [O2O+ $\mathbf{RL_X}$]: A particular setting of the proposed methodology where $n = 1$, including extractor, paraphraser, and X as the reward function for reinforcement learning (X is either ROUGE-L or the reward function defined in Eq. 5) and beam search.

Pre-trained extractor - Two-to-one extractor-paraphraser model [$\mathbf{M2O_{n=2}}$ (Ext only)]: The pre-trained extractor[7] of the proposed extractor-paraphraser framework (with $n = 2$) on WMS based (Eq. 4) exemplary-extracted sentences.

[3]Note that we obtain the word mover similarity proposed in (Clark et al., 2019) for the case $a = 0$, and $b = 1$.

[4]The hyperparameters are set to $a = 1$ and $b = 0.5$ after extensive experimentations, and are used throughout this paper.

[5]We have limited our work to $n = 2$ since the Two-to-one baselines did not perform efficaciously. Experiments for $n > 2$ would be done in future works.

[6]Statistical analysis on all the variations of the proposed model has been done.

[7]For fair comparison of extraction capabilities across all models, we limit the model to output 4 sentences during evaluation on test dataset.

Table 1: Evaluation scores for the generated text summary using ROUGE, METEOR and Word Mover Similarity (WMS). 'Id-ext' refers to the ideal extractor experiments. The '-' denotes unavailability of a score.

Models	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	WMS
Extractive baselines					
$EXT-ABS\,(Ext\,only)$	40.17	18.11	36.41	**22.81**	-
$O2O\,(Ext\,only)$	**40.97**	**18.42**	**37.35**	21.86	**14.28**
$M2O_{n=2}\,(Ext\,only)$	40.19	17.98	36.60	21.62	14.24
Abstractive baselines					
PGN	39.53	**17.28**	36.38	18.72	13.36
$EXT-ABS\,(w/o\,RL)$	38.38	16.12	36.04	**19.39**	-
$O2O\,(w/o\,RL)$	**39.82**	17.05	**37.21**	19.24	**13.81**
$M2O_{n=2}\,(w/o\,RL)$	32.82	11.29	31.08	16.13	12.92
Reinforced models					
$EXT-ABS+RL_{ROUGE}$	40.88	17.8	38.53	20.38	13.7
$O2O+RL_{ROUGE}$	**41.32**	**18.16**	**38.89**	20.52	14.56
$EXT-ABS+RL_{WMS}$	40.82	17.82	38.45	**21.47**	14.42
$O2O+RL_{WMS}$	41.2	18.12	38.81	21.34	**14.6**
$M2O_{n=2}+RL_{ROUGE}$	39.71	16.7	37.32	18.25	13.52
Ablation experiments (Ideal Extractor)					
$Id-ext\,EXT-ABS$	49.73	**26.53**	47.2	24.36	19.34
$Id-ext\,O2O$	**50.04**	26.31	**47.34**	**24.61**	**20.14**
$Id-ext\,M2O_{n=2}$	47.00	23.37	44.23	22.22	17.99

Two-to-one extractor-paraphraser model without reinforcement [$\mathbf{M2O_{n=2}\,(w/o\,RL)}$]: A specific case of the proposed many-to-one paraphrasing where $n = 2$. This specific model comprises of only the extractor and paraphraser modules.

Two-to-one extractor-paraphraser model [$\mathbf{M2O_{n=2}+\,RL_X}$]: A specific case of the proposed many-to-one paraphrasing where $n = 2$; including X as the reward function for reinforced extractor (X is either ROUGE-L or the reward function defined in Eq. 5) and beam search.

Ideal extractor-abstractor model [$\mathbf{Id-ext\ EXT-ABS}$]: To determine the performance of the abstractor component, the $EXT-ABS$ baseline is evaluated with the assumption that the extractor is ideal, subsequently feeding the exemplary-extracted sentences generated for the test data using ROUGE-L score to the paraphraser directly.

Ideal extractor-paraphraser model [$\mathbf{Id-ext\ O2O}$]: Similar to the $Id-ext\,EXT-ABS$ setup, the 'n-to-1' paraphraser is evaluated with $n = 1$, given that the extractor performs ideally (generating exemplary-extracted sentences as proposed in Eq. 4).

Ideal extractor-paraphraser model [$\mathbf{Id-ext\ M2O_{n=2}}$]: The 'n-to-1' paraphraser with $n = 2$ and the exemplary-extracted sentences assumed as the extractor (generating exemplary-extracted sentences as proposed in Eq. 4).

5 Results

Results of different baselines and the proposed approach are discussed in this section. Table 1 illustrates that our proposed techniques perform better than the rest of the systems. We have used ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and word mover similarity (WMS) (Clark et al., 2019) as evaluation metrics. We believe that ROUGE as an evaluation metric is incapable of judging the quality of an abstractive summary due to its emphasis on syntactic overlap over semantic overlap (Liu et al., 2016; Clark et al., 2019; Novikova et al., 2017). To overcome this, we have also used WMS as an evaluation metric.

5.1 Semantic information overlap

It is noticed that the proposed paraphraser in model $O2O\ (w/o\,RL)$ outperforms the abstractor from $EXT-ABS\ (w/o\,RL)$ in terms of ROUGE scores, while scoring marginally less in terms of METEOR. The extractor counterparts $EXT-ABS\,(Ext\,only)$ and $O2O\,(Ext\,only)$

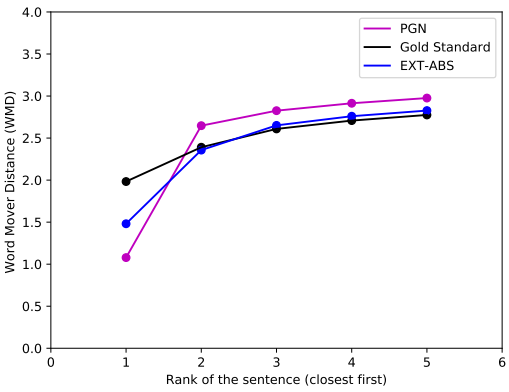

Figure 2: Average sentence distance (word mover distance) scores for most similar sentences.

also portray a similar tendency, with a wider gap in METEOR scores. We also observe that in the reinforced extractor models, the $O2O + RL_X$ setting surpasses the $EXT - ABS + RL_X$ setting in almost all the metrics[8]. A similar trend is also observed in the ideal extractor experiments, where the $Id - ext\, O2O$ model beats the $Id - ext\, EXT - ABS$ model in WMS while keeping other evaluation scores comparable. The above mentioned observations illustrate the true capabilities of using semantic overlap based exemplary-extracted sentences.

Keeping the main model same as the $EXT - ABS$ framework, and changing the reward function from ROUGE-L to WMS (Eq. 5), it is observed that the latter ($EXT - ABS\, RL_{WMS}$) attains significantly better METEOR and WMS scores, while maintaining comparable ROUGE scores. However, the true capabilities of the WMS reward function (Eq. 5) come into play when we attach it with our extractor-paraphraser framework; $O2O\, RL_{WMS}$ bests every other models in terms of WMS, while keeping ROUGE and METEOR comparable with the best attained scores.

One critical observation is that the reward function introduces a bias in the evaluation process. It can be clearly observed from the fact that the model $EXT - ABS + RL_{ROUGE}$ obtains better ROUGE scores while it pales in comparison to $EXT - ABS + RL_{WMS}$ in terms of WMS. Since we stress that semantic information overlap is more significant than the syntactic overlap, we believe that WMS is better suited for the evaluation task as well as a better choice for the reward function.

[8]Here $X \in \{ROUGE, WMS\}$, which remains same when comparing the two models

It is established by the fact that the models using WMS as the reward function attain comparable ROUGE scores as well (while the reverse is not true), indicating that incorporating semantic information can assist in capturing syntactic information as well. Examples of generated summaries for the $EXT - ABS\, RL_{ROUGE}$ and $O2O\, RL_{WMS}$ models are illustrated in Fig. 3. An important observation in the generated summaries is that the former model produces the name "shao li" which is not present in the input document or the gold standard summary, whereas this mistake is avoided by the $O2O\, RL_{WMS}$ model.

5.2 Summarization vs paraphrasing

Theoretically the $M2O_{n=2}$ setting should surpass the $O2O$ setting, since the former has extra input information at the paraphraser stage that the latter lacks. However, it is observed that this does not happen; in actuality, the $M2O_{n=2}$ model is outperformed by the $O2O$ model in all aspects (Table 1). After manual scrutiny of *exemplary-extracted sentence* pairs fed to the paraphraser and the generated sentences, it is observed that the expected accumulation of information does not take place. To quantify this observation, WMD based overlap of information is computed between the $exemplary - extracted\, sentence$ and the generated sentence. It is discovered that on average, the more similar sentence has a WMD of **1.775** while the other one obtains a value of **2.894**, illustrating the inability of the paraphraser to combine information across the two sentences into one.

Hence, we hypothesize that the PGN model intrinsically paraphrases one input sentence to generate the corresponding sentence in the generated summary, innately mimicking the extractor paraphraser behaviour. An experiment is formulated to evince the truth of this proposed hypothesis. For this experiment, three different collections of documents are used: 1) ground truth summary, 2) summaries generated by PGN (See et al., 2017), and 3) summaries generated by the $EXT - ABS$ model (Chen and Bansal, 2018), all corresponding to the data in the test set. For every summary sentence, its semantic overlap (using WMD) is computed with every document sentence and distances of the closest α document sentences are reported in an increasing order. The results of this experiment can be seen in Fig. 2, and it is noticed that the average gap between similarities with respect to first

and second most similar sentences is the highest in the case of the PGN model (approx. 1.57), while the value is around 0.41 for the gold standard summaries. The curve for gold standard has a steady slope, where as the slope for PGN rises quickly from the first point to the second point, and then follows a constant growth afterwards, illustrating a very high similarity with the most similar sentence, thus endorsing the hypothesis that the PGN intrinsically paraphrases one input sentence to generate one sentence in final summary.

Gold Standard:
tong shao, 20, was an international student from china attending iowa state university. her body was found in the trunk of her car in iowa city on september 26. police believe it had been for three weeks. she died of blunt force trauma and asphyxiation. her boyfriend, xiangnan li, 23, was the last to see her, but flew to china on september 8, before shao was officially missing. according to tong 's father, an arrest warrant has now been issued. however li has disappeared.
EXT-ABS + RL_{ROUGE}:
police are issuing a warrant for the girl's boyfriend. police found her body stuffed in the trunk of her toyota camry. shao li, 23, was listed as a person of interest in the case. tong shao was a chemical engineering student at iowa state university. xiangnan li, 23, found murdered in september after going missing. the girl 's body was discovered in iowa last year. tong shao, 20, was found dead in her car for three weeks.
O2O + RL_{WMS}:
tong shao went missing in september 2014. police found her body stuffed in the trunk of her toyota camry. shao's boyfriend, xiangnan li, 23, was listed as a person of interest. police are issuing a warrant for the girl's boyfriend. a 20-year-old daughter was found murdered in iowa last year.

Figure 3: Example of generated summaries for $EXT-ABS+RL_{ROUGE}$ model, $O2O+RL_{WMS}$. The text in red denotes the novel information that is not present in the gold summary, the text in green denotes the information overlap that is present exclusively in the generated summary and the text in blue denotes the information covered in all three scenarios.

5.3 Error Analysis

As the results illustrate in Table 1, overall the one-to-one setting of the proposed n-to-one paraphraser performs better than the two-to-one setting, revealing a major blockade of the existing sequential frameworks - the incapability of combining multiple sentences into one, deviating from the ideal notion of 'abstraction'. We also observe that the extractors outperform their corresponding $(w/o\,RL)$ counterparts, and are competing with the complete models as well to some extent, portraying the difficulty of abstractive summarization over sentence extraction, along with the inability of current evaluation metrics like ROUGE and METEOR to focus on semantic information over syntactic subtleties. As elucidated by the ideal extractor experiments, the paraphraser has great potential for even achieving state-of-the-art results in the summarization task, provided the extractor works ideally. Our future work would focus on mollifying the loss introduced at the extraction step, and obtaining a better harmony in the extractor-paraphraser network as a whole.

6 Conclusion

In this paper, we propose the extractor-paraphraser model, that comprises of the n-to-one paraphraser and a novel word mover similarity based reward function. We show that the semantic overlap based techniques surpass the strong baselines that rely on syntactic information overlap. We also develop experiments to unveil that the state-of-the-art pointer-generator network (PGN) indeed paraphrases input sentences intrinsically, and is unable to merge two sentences into one agglomerate sentence when explicitly conditioned to do so, changing our perception of sequence-to-sequence abstractive summarization models. We show that the existing works are still nowhere close to mimicking a true human summary, portraying the difficulty of true abstraction over its simplified formulation of extracting and then paraphrasing.

Acknowledgement: Dr. Sriparna Saha would like to acknowledge the support of Early Career Research Award of Science and Engineering Research Board (SERB) of Department of Science and Technology, India to carry out this research.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Multi-document abstractive summarization using ilp based multi-sentence compression. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. *arXiv preprint arXiv:1809.09672*.

Yijun Duan and Adam Jatowt. 2019. Across-time comparative summarization of news articles. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 735–743. ACM.

Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of COLING 2012*, pages 911–926.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Yang Gao, Christian M Meyer, Mohsen Mesgar, and Iryna Gurevych. 2019. Reward learning for efficient reinforcement learning in extractive document summarisation. *arXiv preprint arXiv:1907.12894*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Anubhav Jangra, Adam Jatowt, Mohammad Hasanuzzaman, and Sriparna Saha. 2020a. Text-image-video summary generation using joint integer linear programming. In *European Conference on Information Retrieval*, pages 190–198. Springer.

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammad Hasanuzzaman. 2020b. Multi-modal summary generation using multi-objective optimization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1745–1748.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2018. Generative adversarial network for abstractive text summarization. In *Thirty-second AAAI conference on artificial intelligence*.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 170–173.

Rada Mihalcea and Hakan Ceylan. 2007. Explorations in automatic book summarization. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 380–389.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Chris D Paice. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Samuel Pecar. 2018. Towards opinion summarization of customer reviews. In *Proceedings of ACL 2018, Student Research Workshop*, pages 1–8.

Naveen Saini, Sriparna Saha, Pushpak Bhattacharyya, and Himanshu Tuteja. 2020. Textual entailment–based figure summarization for biomedical articles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1s):1–24.

Naveen Saini, Sriparna Saha, Dhiraj Chakraborty, and Pushpak Bhattacharyya. 2019a. Extractive single document summarization using binary differential evolution: Optimization of different sentence quality measures. *PloS one*, 14(11):e0223477.

Naveen Saini, Sriparna Saha, Anubhav Jangra, and Pushpak Bhattacharyya. 2019b. Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm. *Knowledge-Based Systems*, 164:45–67.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Liqiang Xiao, Lu Wang, Hao He, and Yaohui Jin. 2020. Copy or rewrite: Hybrid summarization with hierarchical reinforcement learning. In *AAAI*, pages 9306–9313.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336.

Yong Zhang, Meng Joo Er, Rui Zhao, and Mahardhika Pratama. 2016. Multiview convolutional neural networks for multidocument extractive summarization. *IEEE transactions on cybernetics*, 47(10):3230–3242.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9749–9756.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50.