Evaluation of Cross Domain Text Summarization

Liam Scanlon*, Shiwei Zhang, Xiuzhen Zhang, Mark Sanderson
Computer Science, School of Science, RMIT University, Australia
liam.scanlon@unimelb.edu.au,{shiwei.zhang,xiuzhen.zhang,mark.sanderson}@rmit.edu.au

ABSTRACT

Extractive-abstractive hybrid summarization can generate readable, concise summaries for long documents. Extraction-then-abstraction and extraction-with-abstraction are two representative approaches to hybrid summarization. But their general performance is yet to be evaluated by large scale experiments. We examined two state-of-the-art hybrid summarization algorithms from three novel perspectives: we applied them to a form of headline generation not previously tried, we evaluated the generalization of the algorithms by testing them both within and across news domains; and we compared the automatic assessment of the algorithms to human comparative judgments. It is found that an extraction-then-abstraction hybrid approach outperforms an extraction-with-abstraction approach, particularly for cross-domain headline generation.

KEYWORDS

Text summarization, Headline generation, Evaluation

ACM Reference format:
Liam Scanlon*, Shiwei Zhang, Xiuzhen Zhang, Mark Sanderson. 2020. Evaluation of Cross Domain Text Summarization. In *Proceedings of 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, China, July 25–30, 2020 (SIGIR '20),* 4 pages.
https://doi.org/10.1145/3397271.3401285

1 INTRODUCTION

Hybrid summarization approaches generate readable, concise summaries and outperform traditional extractive methods. Two such summarization models that were recently proposed: one uses Reinforcement Learning (RL) [1] and the other, inconsistency loss (IL) [4]. The **RL model** adopts an extraction-then-abstraction hybrid approach, whereas the **IL model** combines by extraction-with-abstraction. Both models were reported to outperform traditional extractive summarizers for multi-sentence summarization tasks.

In this paper we evaluate the general performance of hybrid summarization approaches. We test the RL and IL models on the task of headline generation [13], which is to generate a short, one sentence article summary. We further evaluate the generalisation of the models across different news domains.

Summarization evaluation metrics – such as ROUGE [7], BLEU [10], and METEOR [3, 8] – reward summaries containing a high word (*n*-gram) overlap with a reference (human) summary [8]. Human evaluation is sometimes employed as a complement to evaluate summary readability or informativeness. However, human comparison of summary system output is not widely adopted. We gather such comparisons to rank the relative performance of the headlines generated by the two models. We find that:

- The extraction-then-abstraction RL approach outperforms the extraction-and-abstraction IL based model, particularly for cross domain applications.
- Human evaluation reveals that the METEOR and ROUGE-2 metrics correlate best with human judgements.

2 RELATED WORK

Hybrid summarization [1, 4] uses *extractive* salient sentence selection as an input to an *abstractive* summarizer module, that subsequently rewrites the chosen salient sentences. To bridge the abstractor and extractor modules, different end to end training modules are often implemented. In such systems, the number of novel *n*-grams increase, compared to past work [15]. Such increases denote a model's abstractiveness, resulting in much better performance [1].

The de facto evaluation standard is the *n*-gram overlap metric ROUGE and its variants: ROUGE-1/ROUGE-2 (1 & 2 word matches) and ROUGE-L (Longest Common Subsequence). Experiments correlating ROUGE with human judgements found that ROUGE-SU, ROUGE-L had the highest correlation on the DUC dataset, followed closely by ROUGE-1, 2 & 3 [6, 14].

Some researchers opt to use METEOR as an additional metric [5]. Originally used in machine translation, it differs from ROUGE by allowing more match flexibility with word stem variation and WordNet synonymy [3]. While other metrics have been created [3, 10, 12], there remains a concern with the reliability of automated metrics [8], consequently, human (crowd-sourced) judgements complement automatic metrics in recent work [1, 4].

3 THE TWO MODELS

The **RL model** [1] is claimed to deliver state-of-the-art scores and better generalisation performance. The model has a pre-training stage for both the extractor and the abstractor, and a reinforcement training stage for bridging the extractor and the abstractor. To pre-train the extractor and the abstractor, sentences of a document are labelled. In **RL model**, the most similar document sentence in terms of ROUGE-L_{recall} to a ground-truth summary sentence is extracted as the positive class, the rest are treated as negative, allowing the pre-training of the extractor to be formulated as a classification task. Additionally, the positive document sentence is paired with the summary sentence for pre-training the abstractor.

At the reinforcement training stage, a policy gradient method, Advantage Actor-Critic (A2C), is applied to optimise the model,

*Work done at RMIT University.

SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401285

Model	R-1	R-2	R-L
Pointer Generator [15]	39.53	17.28	36.38
IL [4]	40.68	17.97	37.13
RL [1]	40.88	17.80	38.54

Table 1: ROUGE scores for summarizers on CNN/DailyMail

where the extractor is treated as a RL agent. The approach encourages the extractor to choose sentences that the abstractor will compress into sentences having high ROUGE matches with the ground truth, discouraging the extractor from choosing sentences that will result in low ROUGE scores. In addition, language fluency [11] is improved by limiting the abstractor's word level training weights from being influenced by the bridging module's policy gradients.

The **IL model** [4] also claims state-of-the-art scores and the best readability and informativeness on CNN/DailyMail. The extractor of the **IL model** outputs sentence-level attentions which are then combined with word-level attentions in the abstractor. In order to ensure that word-level attentions are consistent with sentence-level attentions, inconsistency loss is proposed, which results in sentences with high attention also having high word-level attentions. In terms of training, this approach is similar to the **RL model**, which also has a three-stage training: pre-training modules separately, fine-tuning the abstractor, and end-to-end training of the whole model. Similar to the **RL model**, the **IL model** also uses ROUGE-L$_{recall}$ to calculate scores for labelling sentences of a document. When pre-training the extractor, it has been formulated as a classification task just like the one in **RL model**, but with a sigmoid cross entropy loss. The abstractor module is derived from a pointer-generator network [15], which can generate the summary by copying words (including OOV words) from a given document. For the IL function, sentence-level and word-level attention are combined to make attention to any single word dependent on the whole sentence, with the aim of reducing word/sentence attention inconsistency (L_{inc}). Apart from including a coverage mechanism in the abstractor from [15] to reduce repetition, the model also introduces a coverage loss measurement to directly penalise word level repetition (L_{cov}). In final end to end training, the four loss functions: the extractor, abstractor, coverage, and inconsistency functions, are all minimised. This unique IL function during end-to-end training can be written as:

$$L_{e2e} = \lambda_1 L_{ext} + \lambda_2 L_{abs} + \lambda_3 L_{cov} + \lambda_4 L_{inc}$$

The authors choose to predefine fixed hyper-parameters to each function, weighting the extractor loss function heavily (λ_1), as well as decreasing their learning rate by 15x to alleviate subpar performance of the extractor.

As both models were researched simultaneously, neither model was compared to the other. Both claim their method of end to end training is superior to prior models due to their unique improvements in the field, however while their raw ROUGE scores are similar (see Table 1), it is unknown which of the two generate truly better summaries in a direct comparison. In addition, both models were developed, trained and tested on the CNN/DailyMail dataset for multi-sentence summarization, but [1] claims to also perform well specifically when generalising to other datasets.

4 IN-DOMAIN AND CROSS-DOMAIN EVALUATION AND HUMAN JUDGEMENTS

We use two corpora, Gigaword and CNN/DailyMail, for training and testing of the summarizers, as well as human judgements to examine automated metrics.

The Gigaword corpus [1] has annotated one-sentence summaries, or headlines, for news articles and is used for evaluation of headline generation models. Gigaword has types of news articles different from CNN/DailyMail. Compared with CNN/DailyMail, on average, Gigaword had longer words (4.6 vs 4.4 char), shorter sentences (25 vs 28 words) and shorter articles (19 vs 27 sent). To be consistent with the size of the CNN/DailyMail dataset, a subset (312,085) of the Gigaword corpus is used for evaluation with a split of train (287,227), validation (13,368), and test (11,490). In-domain models were trained and tested on Gigaword, while cross-domain models were trained on CNN/DailyMail and evaluated on Gigaword. The performance of four pairs of models is compared:

(1) Ideal performance: **RL** vs **IL**; trained and tested on Gigaword (in-domain).
(2) Generalisation performance of **RL** vs **IL**; tested on Gigaword (cross-domain)
(3) **RL**: in-domain (ID) vs cross-domain (XD) performance.
(4) **IL**: in-domain (ID) vs cross-domain (XD) performance.

Gigaword annotations are used for tokenisation and sentence separation. Similar to past preprocessing [2, 13], we create an input-summary pair from the headline and first sentence of each article. Articles were tokenised with only 7-bit ASCII characters ($\leq U007F$) retained. Short sentences and headlines were found to cause problems with the summarizers, therefore, only articles with the following conditions were used: a sentence has > 5 words; a headline > 35 characters; the article must contain > 450 characters and > 5 sentences.

For training RL, a *word2vec* embedding was generated. After the extraction, labels were created using the CNN/DailyMail data script make-extraction-labels.py [1], both the abstractor and extractor were trained simultaneously, using the *word2vec* embedding. Each module performed the best on the validation set after 108k and 15k iterations respectively. The final bridging module (RL) was trained over the two models, using hyper-parameters as reported in the paper [1]. The model was used as the in-domain model for this approach (ID RL). For the cross-domain model, the best pretrained model – from the original paper – was used (XD RL).

For training of IL, as above, the three modules were trained separately, using their IL function contributions in the last module. All parameters reported in their paper were adhered to, including concurrent evaluation of the validation set during training, adding 1k iterations of coverage training on top of the abstractor, and using the best trained extractor module but the last trained abstractor module. Hyperparameters were also set at the values used in the paper. This model was used as the in-domain model for this approach (ID IL). The cross-domain model was the best pretrained model provided, same as above (XD IL).

We design user studies to obtain human judgements to directly compare summaries to examine the reliability of automatic scoring

[1] https://catalog.ldc.upenn.edu/LDC2012T21

		R1	R2	RL	RSU	M	SL	SD
ID	RL	19.6	5.7	17.8	.071	9.6	70.4	40.4
ID	IL	22.1	6.0	20.0	.080	10.2	52.5	10.8
XD	RL	24.6	8.6	22.1	.094	13.3	84.9	24.3
XD	IL	10.1	1.4	8.4	.016	7.4	151.1	83.0

Table 2: Automatic metric scores across all 11,490 instances. R1: unigram-R(OUGE), R2: bigram-R, RL: longest-common-sequence-R, RSU: skip-bigram-plus-unigram-R, M: METEOR, SL: Sentence Length and SD: SL standard deviation.

metrics. Four groups of 100 pairs of headlines were judged. The articles that headlines were drawn from were randomly picked from the same test set. As automatic summarisation metrics are sensitive to sentence length, it was ensured that selected pairs had the same or similar (within 3 characters) sentence lengths. Table 2 details average sentence lengths of the whole test set, Table 3 sentence lengths of the sentences picked for the survey.

Judges read a reference summary and two headlines, without indication which summarizer generated which headline. The judges selected the best matching headline or they indicated the pairs were equally good/bad. The four authors judged the pairs, taking on average 21 seconds to assess each.

We examined the agreement between the four judges (labelled ‘A’, ‘B’, ‘C’, ‘D’) using Cohen’s Kappa. Kappa values between each pair of judges were 0.45 (A/B), 0.47 (A/C), 0.47 (A/B), 0.36 (B/D), 0.39 (B/C), 0.54 (C/D). Judge B’s average kappa values showed the most disagreement with others (0.41). The total average pairwise Cohen’s kappa agreement is 0.44, meaning the judges were in moderate agreement, well within desired ranges [9]. With these results, we can be confident in our survey results.

5 EXPERIMENTS AND RESULTS

The evaluation packages `pyrouge` and METEOR [3] were used. Versions were those specified in past papers [1, 4]. As longer summaries can get higher scores, summary lengths (SL) are shown. Training/decoding was run on a Intel Xeon Gold 6132, 256GB RAM, Red-Hat RHEL-v7.4, NVIDIA Tesla V100 Tensor Core. Code link.[2]

5.1 Overview of results

Table 2 contains the results for all test instances. we see XD RL achives higher scores than XD IL despite the SL for RL being much shorter. This result is in-line with past work on the generalisation performance of RL. the creators of IL ([4]) on the other hand, mention nothing about generalisation. For ID IL outputs shorter summaries but scores better than ID RL.

In comparing ID and XD scores, ID IL scores better than its cross-domain version. In contrast, XD RL scores better than ID RL. Although this could be due to the RL’s strong emphasis on generalisation or XD RL producing slightly longer summaries.

Comparing ID RL and ID IL (first row of Table 3) shows no difference between the models. The second row compares the models cross-domain. Here we see that even with equal length sentences, IL scores poorly compared to RL over all metrics, confirming [1]’s claims that their RL model generalises well. To illustrate, the following example summary shows a reference headline in bold, followed by the RL and IL summaries.

		R1	R2	R3	RL	RSU	M	SL	C
ID	RL	27.4	11.4	6.4	25.5	.133	12.4	47.0	87
ID	IL	27.5	10.4	6.1	25.8	.129	12.4	47.0	83
XD	RL	25.4	10.5	5.1	22.7	.104	13.7	81.4	247
XD	IL	7.7	1.0	0.1	6.7	.015	5.5	81.3	25
ID	RL	23.0	9.2	6.7	20.6	.108	11.9	65.5	73
XD	RL	23.8	9.3	4.2	22.0	.099	13.0	65.6	173
ID	IL	25.1	8.4	3.9	23.5	.101	11.1	50.0	153
XD	IL	9.3	1.5	0.4	8.5	.024	5.4	50.4	45

Table 3: Automatic metric scores for 400 human judged instances. C: number of times chosen.

death toll in china landslide rises to 34.
death toll from landslide in southwest china’s guizhou province rose to 34.
ten people missing after the disaster, which occurred early friday, it said.

Both summaries are coherent and concise, however the RL summary is the most informative and direct. Here the extractor module for IL selects the wrong sentence from the article to generate its headline from, possibly due to words it overjudged as important, such as ‘missing’ and ‘disaster’.

The measures for ID RL and XD RL are similar. Considering that Gigaword has types of articles that are unseen in CNN/DailyMail, this gives the in-domain model a better chance to generate the correct headline for such articles. When the models correctly identify ‘news summary’ articles they get a full headline match, which increases ROUGE-3 (3-gram matches). We speculate such behaviour explains why ROUGE-3 for ID RL is higher than XD RL.

Examining the two domains for IL, as expected it performs better in-domain. An example is shown below. The headline is for a ‘news summary’ article. When IL is trained in-domain on Gigaword, the article style is recognised and an appropriate headline is generated, albeit with the wrong country. Cross-domain training appears to cause the model to fail.

major news items in leading new zealand newspapers.
major news items in leading chinese newspapers.
preparing to meet her sister for the first time.

5.2 Human judgements on the 400 instances

For in-domain comparison, survey results agree with the metrics. The judges preferred ID IL and ID RL about equally though with the most non-preferences (230). Note, no preference does not distinguish between equally good or equally bad headlines, the outputs are similarly effective.

For cross-domain, survey results also agreed with the metrics, with the fewest non-preferences (128). Judges were ten times more likely to choose RL over IL, indicating the scale of RL’s superiority in this comparison. Analysis of headline pairs was unable to pinpoint a clear reason why IL was much worse, as there were cases where either the extractor or the abstractor was at fault.

[2] https://github.com/scanlia/Evaluation-of-Cross-Domain-Text-Summarization

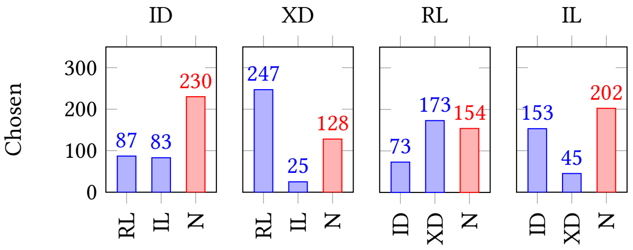

Figure 1: Aggregated human preference judgements across the same 400 instances measured in Table 3. The blue bars show preferences, the red bars show no preference.

The in/cross-domain comparison for RL showed the survey results not completely in line with the metrics. Judges preferred XD RL over ID RL (173 vs 73), despite the in-domain model's better R-3,-SU scores. METEOR and R-2 were the only metrics that markedly increased enough for the cross-domain model in relation to the magnitude of the survey scores. One such representative example is shown below (in-domain followed by cross-domain).

china takes asian women's volleyball crown.

- south korea beat china in the women's volleyball world championships here on sunday.
- china trounced japan 3-0 to claim sixth title in a row with all six match victories.

All judges preferred the XD generated headline (third line) to the ID model. However, the factually wrong ID headline (second line) scores better in all metrics, an example of automatic scoring failing to correlate with human judgements.

Finally, examining the in/cross-domain comparison for IL, the metrics correlate well with the judgements, both clearly agreeing that the in-domain headlines are better than the cross-domain headlines. We speculate that the reason for poor cross-domain performance is the large amount of noise in the Gigaword dataset. This included stock prices or sports results where there are more numbers than actual text, most likely throwing off the in-domain RL training. Furthermore, the good generalisation performance of RL could well be sensitive to and highly tuned for training on the specific CNN/DailyMail dataset.

5.3 Discussions

Across our results we find that RL – as an end-to-end training method for this type of hybrid approach – works better than IL. Although RL is only slightly better in an in-domain setup, it is strikingly better at cross-domain. This can be attributed to a simpler and more elegant policy-reward based Markov decision process, as well as the use of block word-level abstractor changes during end to end training of RL [1]. While IL, showing marked improvements in reported AB tests [4], the final loss function has no ability to train its fixed end to end training hyper-parameters.

Both models trained in-domain seem to perform well on the Gigaword headline dataset. The models are not like a fully *abstractive* network, which applies the whole input text to the abstractor and outputs a single headline. Instead, the models rely on an *extractive* salient sentence selection process that categorises a sentence as fully relevant to the core meaning of the text, and applies an abstractor to compress this sentence. This results in excellent abstractor focus and performance, but forfeits the ability to understand and interpret pieces of information across multiple sentences in the article, and sometimes output headlines lose underlying context that a fully *abstractive* model would gather in its headline.

6 CONCLUSION

Two state of the art text summarisation approaches were examined in a novel headline generation task on the Gigaword newswire corpus. Our study showed that both models work well when trained and tested in the same domain, but the RL model worked notably better than the IL, when tested across domains. The improvement was consistent across both automatic metrics and comparative human judgements in both domains.

The success of RL suggests that limiting the word-level attention in the abstractor from being modified during final training works well. Cross-domain generalisation of RL was so good it outperformed its own in-domain model, suggesting noise in the Gigaword dataset negatively affected the in-domain training.

Human judgements were found to mostly agree with the *n*-gram based metrics apart from in-domain and cross-domain comparisons for RL. The strength of inter-rater agreements indicated that the scores from this evaluation to be reliable.

7 ACKNOWLEDGEMENTS

This research is supported in part by Australian Research Council Discovery Projects DP200101441 and DP180102687.

REFERENCES

[1] Chen, Y.C., Bansal, M.: Fast abstractive summarization with reinforce-selected sentence rewriting. In: ACL (2018)

[2] Chopra, S., Auli, M., Rush, A.M.: Abstractive sentence summarization with attentive recurrent neural networks. In: Proc NAACL. pp. 93–98 (2016)

[3] Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation. pp. 376–380 (2014)

[4] Hsu, W.T., Lin, C.K., Lee, M.Y., Min, K., Tang, J., Sun, M.: A unified model for extractive and abstractive summarization using inconsistency loss. In: ACL (2018)

[5] Kim, S.N., Baldwin, T., Kan, M.Y.: Evaluating n-gram based evaluation metrics for automatic keyphrase extraction. In: Proceedings of the 23rd international conference on computational linguistics. pp. 572–580 (2010)

[6] Lin, C.Y.: Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough? In: NTCIR (2004)

[7] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)

[8] Liu, C.W., Lowe, R., Serban, I.V., Noseworthy, M., Charlin, L., Pineau, J.: How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023 (2016)

[9] Munoz, S.R., Bangdiwala, S.I.: Interpretation of kappa and b statistics measures of agreement. Journal of Applied Statistics **24**(1), 105–112 (1997)

[10] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318 (2002)

[11] Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304 (2017)

[12] Rankel, P.A., Conroy, J.M., Schlesinger, J.D.: Better metrics to automatically predict the quality of a text summary. Algorithms **5**(4), 398–420 (2012)

[13] Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685 (2015)

[14] Saggion, H., Poibeau, T.: Automatic text summarization: Past, present and future. In: Multi-source, multilingual information extraction and summarization, pp. 3–21. Springer (2013)

[15] See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368 (2017)