Evaluation of Personalized Summarization

Rahul Vansh
DA-IICT
Gandhinagar, Gujarat
202111035@daiict.ac.in

Prof. Sourish Dasgupta
DA-IICT
Gandhinagar, Gujarat
sourish_dasgupta@daiict.ac.in

***Abstract*—This study aims to demonstrate that evaluating a summarizer model's personalization is not the same as evaluating its accuracy. Since accuracy-based measures like ROUGE do not take subjectivity into account while evaluating the personalized summarizer model, we developed a metric e-DINS$_{sub}$ to evaluate the degree of personalization by taking into account both the user profile and the summary. Analysis has demonstrated the consistency and reliability of e-DINS$_{sub}$.**

***Index Terms*—text summarization, adaptive summaries, personalized summaries, evaluation metric**

I. Introduction

There are several metrics that evaluate the model in terms of quality (focusing on the readability) and accuracy of the summary [1]. None of them evaluate how well a summarization model can capture user preference w.r.t. subjectivity (user's individual perception of saliency). Measuring the degree of personalization is essential to know how well a summarization model can adapt user preference while generating a personalized summary.

In paper [2], which was Exdos [3] based personalized summarization model were measured in terms of iterative convergence. Here convergence happens when the user stop giving feedback on the same document. But in this case, no quantitive measure to differentiate two different users (having different interests).

Microsoft proposed PENS framework[4], which can generate personalized headlines based on user profiles. Headlines can be considered as personalized summaries. To evaluate their model ROUGE was calculated between user-written personalized summary and model-generated personalized summary. We argue that ROUGE is an accuracy measure, it can not measure the degree of personalization. Fig. 2 (a) shows that the model has high accuracy since the user profile and summary have a small difference, indicating that the model captures users' interests quite well. However, personalization is low since their summary pair has less difference compared to the user profile pair. Fig. 2 (a) shows that the model has high accuracy while personalization is also high as the summary and the user profile pair have almost equal differences. So a model may have high accuracy but still have a low degree of personalization. The same is supported by our findings; more information on this will be provided in the results and discussion section. Thus, accuracy measures are not adequate for measuring the degree of personalization.

The accuracy-centric measure focuses on the difference between user-written and model-generated summaries (shown by the red circle in fig 2). In contrast, personalization-centric measure focuses on the relationship between model-generated summaries and user profiles.

Our contributions are as follows:

- Demonstrated that evaluating a summarizer model's personalization is not the same as evaluating its accuracy
- Proposed evaluation metric to measure the degree of personalization w.r.t subjectivity.
- Analysis performed to show consistency and reliability of the proposed measure.

II. Proposed evaluation measure

In this section, we'll go over how our proposed evaluation measure was designed.

A. Defining Insensitivity

Given document $\mathbf{D}$, and user profile details $\mathbf{u}$, let summarization model $M_{\boldsymbol{\theta},\mathbf{u}}$ generate the best estimated personalized summary $\widehat{S}$

$M_{\boldsymbol{\theta},\mathbf{u}} : \mathbf{D}, \mathbf{u} \mapsto \widehat{S}_u$

Given two different user profiles, $\mathbf{u}_i$ and $\mathbf{u}_j$, summarization model $M_{\boldsymbol{\theta},\mathbf{u}}$ is (weakly) *Insensitive-to-Subjectivity* iff $\forall(\mathbf{u}_i, \mathbf{u}_j), \ni f^U_{dist}(\mathbf{u}_i, \mathbf{u}_j)^* > \tau^U_{max}$:

$$f^S_{sim}(M_{\boldsymbol{\theta},\mathbf{u}}(\mathbf{D}, \mathbf{u}_i), M_{\boldsymbol{\theta},\mathbf{u}}(\mathbf{D}, \mathbf{u}_j)) < \tau^S_{min}$$

where

- f^U_{dist} : User profile distance function
- f^S_{sim} : Summary similarity function
- τ^U_{max} : Max. limit for two different user profiles to be mutually indistinguishable
- τ^S_{min} : Min. limit for two generated summary w.r.t two different users to be mutually distinguishable

* : $f^U_{dist}(\mathbf{u}_i, \mathbf{u}_i) = 0$ and $f^U_{dist}(\mathbf{u}_i, \mathbf{u}_j) \in [0, 1]$

Example of how we can compare Degree-of-Insensitivity w.r.t subjectivity DINS$_{sub}$ of different summarization models $M_{\boldsymbol{\theta}_x,\mathbf{u}}$, $M_{\boldsymbol{\theta}_y,\mathbf{u}}$ and $M_{\boldsymbol{\theta}_z,\mathbf{u}}$. If we have a metric that gives a score based on how insensitive the model is, i.e., how poorly a model generates a personalized summary, then as per table I, high score indicates poor personalized summary score given by metric, while low score indicates model generated better personalized summary for that news article. Expected DINS$_{sub}$

		Manual Evaluation	Automated Evaluation	
			Reference-based	Reference-free
Accuracy		1. Factoid (2003) 2. Pyramid (2004) 3. SEE (2003)	1. Cosine similarity (1989) 2. BLUE (2002) 3. ROUGE (2004) 4. Unit overlap (2002) 5. Latent-based (2009) 6. Semi-automated pyramid (2018) 7. Automated pyramid (2017)	1. KL divergence (1959) 2. Jensen–Shannon divergence (1991) 3. FRESA (2013) 4. SummTriver (2018) 5. Summary likelihood (2013) 6. SUPERT (2020)
Quality		1. DUC 2005 readability 2. TAC 2008 readability	Sum-QE(2019)	
Degree of personalization	w.r.t. subjectivity			e-DINS$_{Sub}$

Fig. 1. Different evaluation metrics[1]

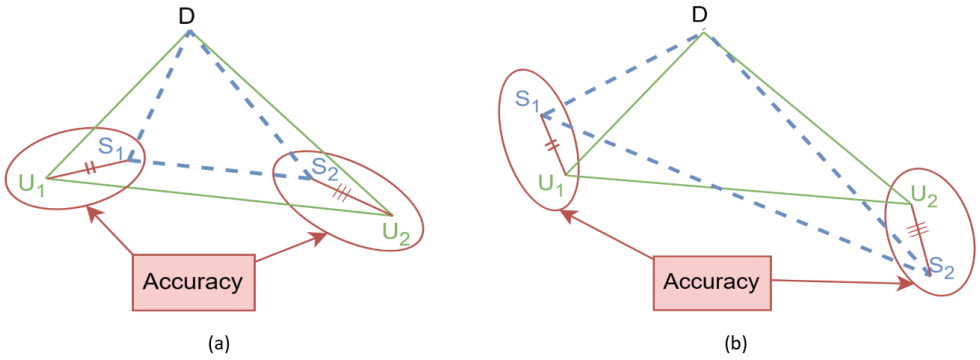

Fig. 2. (a) shows accuracy is high but personalization is low Accuracy. (b) shows both accuracy and personalization are high

for a model $M_{\boldsymbol{\theta}_x,\mathbf{u}}$ is sample-average of its DINS$_{sub}$ score of all news over all user pairs.

B. Degree of insensitivity w.r.t subjectivity (DINS$_{sub}$):

Deviation of summary $\boldsymbol{S}_i$ denoted by $Dev(S_i)$ calculates how other summary S_j deviates from S_i, weighted by the deviation between S_i and S_j w.r.t. deviation between S_i and source document D. Jensen–Shannon divergence used to find deviation between two distributions.

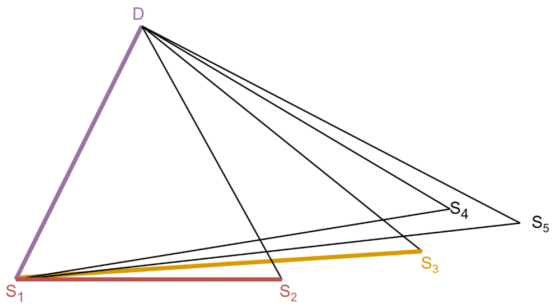

Fig. 3. Illustration of how other summaries deviates from document compare to summary S_1

$$Dev(S_i) = \sum_{j=1}^{n} \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}} * D_{JSD}(S_i||S_j)$$

$$w_{ij} = \frac{D_{JSD}(S_i||S_j)}{D_{JSD}(S_i||D)}$$

where

- n: number of personalized summaries for a single document
- w_{ij}: weight of summary w.r.t. source document
- $D_{JSD}(X||Y)$: Jensen Shannon divergence between distribution X and Y
- D: source document

Degree-of-Insensitivity w.r.t. subjectivity D-INS$_{sub}$ tells how insensitive model is in generating different summaries of same document based on the user's interests (subjectivity). High DINS$_{sub}$ indicates that model is insensitive to incorporate the user's interest while generating summary for that user.

We define the Degree-of-Insensitivity w.r.t. Subjectivity (DINS$_{Sub}$) as :

$$D\text{-}INS_{Sub} = \frac{1}{m*n^2} \sum_{Di}^{m} \sum_{S_j}^{n} Dev(S_j)$$

where

- m: number of documents
- n: number of personalized summaries for single document
- S_j: summary generated for user j

$\mathbf{u}_i$	$\mathbf{u}_j$	**Document**	DINS$_{sub}$ of $M_{\boldsymbol{\theta}_x,\mathbf{u}}$	DINS$_{sub}$ of $M_{\boldsymbol{\theta}_y,\mathbf{u}}$	DINS$_{sub}$ of $M_{\boldsymbol{\theta}_z,\mathbf{u}}$
Bob	**Alice**	$\mathbf{News_1}$	0.43	0.27	**0.61**
		$\mathbf{News_2}$	0.32	**0.86**	0.52
		$\mathbf{News_3}$	**0.58**	0.51	0.39

TABLE I
EXPECTED DINS$_{sub}$ FOR A MODEL $M_{\boldsymbol{\theta}_x,\mathbf{u}}$ IS SAMPLE-AVERAGE OF ITS DINS$_{sub}$ SCORE OF ALL NEWS OVER ALL USER PAIRS

- $Dev(S_j)$: Deviation of summary S_j from all other corresponding summaries of the same document wrt the original document

DINS$_{sub} \epsilon [0, 1]$

C. Effective degree-of-insensitivity w.r.t. subjectivity (e-DINS$_s$ub)

Penalty factor (PF) represents the differences in user profiles and their model-generated summaries. It penalizes the model score if the deviation between the user profile pair is low, but the deviation between the summary pair is high and vice versa.

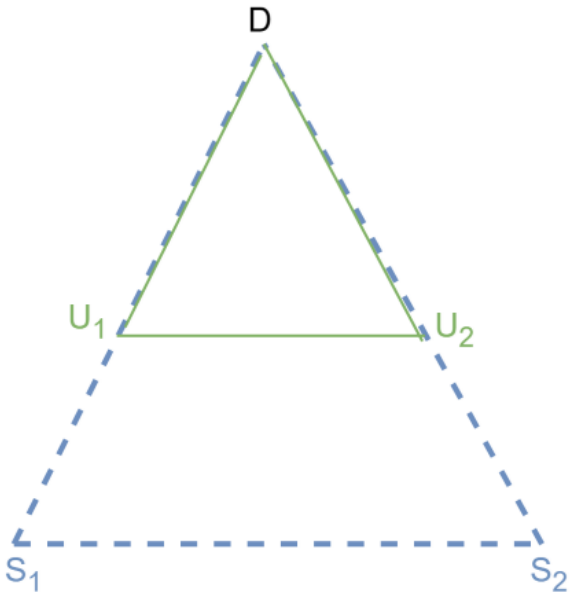

Fig. 4. deviation between summary S_j and S_k is high, but deviation between user profile U_j and U_k is low

If the summary S_j and S_k deviations are not as high as the deviation between the user profile U_j and U_k, as illustrated in the figure, the penalty factor penalizes model score. Since the summary is generated based on the respective user profile only, so if the deviation between two user profiles is low, then generated summary should also have a low deviation.

$$PF = 1 - \frac{1}{m*n^2} \sum_{Di}^{m} \sum_{S_j,U_j}^{n} \sum_{S_k,U_k}^{n} \frac{min(X,Y)}{max(X,Y)}$$

where

- $X = \frac{w^s_{jk}}{\sum_{p=1}^{n} w^s_{jp}} * D_{JSD}(S_j||S_k)$
- $Y = \frac{w^u_{jk}}{\sum_{p=1}^{n} w^u_{jp}} * D_{JSD}(U_j||U_k)$
- n: number of summaries
- S_j: model generated summary
- U_j: gold reference summary
- D_i: document
- w^s_{ij}: weight of model generated summary w.r.t. source document
- w^u_{ij}: weight of gold reference summary w.r.t. source document
- $D_{JSD}(x||y)$: Jensen Shannon divergence between distribution x and y

$PF \epsilon [0, 1]$

Effective degree-of-insensitivity w.r.t. subjectivity e-DINS$_{sub}$ considers penalty factor PF along with D-INS$_{sub}$.γ_{sub} is subjectivity constant which controls how much importance we want to give to PF.

$$e\text{-}DINS_{sub} = DINS_{sub} + \gamma_{Sub} * PF$$

where

- γ_{sub}: subjectivity constant; $\gamma_{sub} \epsilon [0, 1]$
- $D_{INS} \epsilon [0, 1]$
- $e\text{-}D_{INS} \epsilon [0, 2]$

III. EXPERIMENTAL SETUP

A. Dataset

We used test set of PENS(Personalized News headlineS) dataset [4]. Format of dataset is as per given in table II. Headlines can be considered as TLDR summary. 103 english native speakers collected click behaviors and more than 20,000 manually written personalized headlines of news articles, regarded as the gold standard of user-preferred titles. Test set created in 2 stages: In the first stage, Each user reads 1000 news articles and selects at least 50 articles in which he is interested. Headlines are randomly selected and arranged by their first exposure time. In the second stage, each person writes their preferred headlines for 200 articles without knowing the original news title. These news articles are excluded from the first stage.

We need model generated summary to calculate e-DINS$_{sub}$. We used the PENS framework to generate model summaries. Training of the model is done in two phases. In the first phase, the Pointer-Generator Network [5] trained on the actual headline of the article. While in the second phase, the policy model is optimized on the reward of generated personalized headlines where reward includes the degree of personalization, fluency, and factualness[4]. Different kinds of user profile (in form of user embedding) is used. Each of these is injected into the PENS framework in 3 ways, as shown in 5 : 1) To initialize the decoder's hidden state of the headline generator.

Column	Example	Description
userid	NT1	The unique ID of 103 users
rewrite_titles	'Legal battle looms over Trump EPAś rule...	The manuallywritten news headlines for the exhibited news articles and can be split by '#TAB#'
posnewID	N24110,N62769, ...	The exhibited news for each user at the second stage
clicknewsID	N108480,N38238, ...	The user's historical clicked news collected at the first stage

TABLE II
TEST SET FORMAT

2) To personalize attentive values of words in news body 3) To affect the choice between generation and copying.

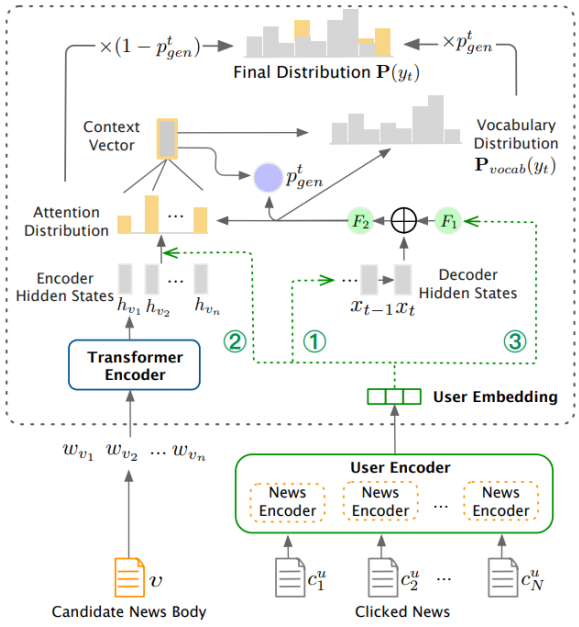

Fig. 5. Three kinds of user profile injections[4]

Consider the example shown in fig. 6, in which underlined words and colored words represent the correlated words in the manually-written headlines, clicked news, and the generated headlines, respectively.

Case 1. Original Headline:	Venezuelans rush to Peru before new requirements take effect
Pointer-Gen:	Venezuelans rush to Peru
user A written headline:	New **requirements** set to take effect causes **Venezuelans** to rush to **Peru**
NAML+HG for user A:	**Peru** has stricter entry **requirements** for escaping **Venezuelans** on that influx.
Clicked News of user A:	1. Peru and Venezuela fans react after match ends in a draw 2. Uruguay v. Peru, Copa America and Gold Cup, Game threads and how to watch
user B written headline:	**Venezuelan migrants** to Peru face danger and discrimination
NAML+HG for user B:	Stricter entry requirements on **Venezuelan migrants** and **refugees**.
Clicked News of user B:	1. Countries Accepting The Most Refugees (And Where They're Coming From) 2. Venezuelan mothers, children in tow, rush to migrate

Fig. 6. Example of personalized headline

B. Computing e-DINS$_{sub}$ using human gold reference summaries from PENS

We need distribution for model-generated personalized summary and user profile to find divergence in DINS$_{sub}$ and e-DINS$_{sub}$. Here, we used the user-written personalized summary as the user profile because the summary contains keyword in which the user is most interested in the article.

Words that occur in summary but not the original text are known as out-of-Vocabulary (OOV) words. The following are the steps to manage OOVs. - Find the similarity of OOV word with all words in the document using BART[6] embedding and cosine similarity

- Add bias, where $bias = 1 - \sqrt{max\_sim\_score}$. If a highly similar word is found in the document, then the bias will be small. If bias has the highest score among all the words in the document, it indicates no highly similar word is found in the document.
- Apply softmax across all similarity scores, which will act as the probability of having OOV word in the document.
- A score of 0 is assigned to an OOV if the bias has the highest score, meaning that no words in the text are highly similar to that OOV. If not, calculate the score by (ratio of word in model summary) / (average of score of all words in doc). Fig. 9 also denotes the same.

- Document : red cat on red tall table ⇒ *Preprocessing* ⇒ [red, cat, red, tall, table]
- User Summary : cat on table ⇒ *Preprocessing* ⇒ [cat, table]
- Model Summary: red cat on *desk* (OOV) ⇒ *Preprocessing* ⇒ [red, cat, desk]

Vocab	Doc distribution	User summary distribution	Model summary distribution
	ratio = word count in doc / total number of words in doc	ratio of word in user summary / ratio of word in doc	ratio of word in model summary / ratio of word in doc
red	2/5 = 0.4	0 *(absent)*	(1/3) / (1/5) = 1.66
cat	1/5 = 0.2	(1/2) / (1/5) = 2.5	(1/3) / (1/5) = 1.66
tall	1/5 = 0.2	0 *(absent)*	0 *(absent)*
table	1/5 = 0.2	(1/2) / (1/5) = 2.5	0 *(absent)*
desk	UNK	UNK	(1/3) / ? = ? **(OOV)**

Fig. 7. Generate distributions from summaries

Words in doc	Similarity of each word in doc with desk	Softmax
red	0.537	0.2
cat	0.405	0.175
tall	0.613	0.216
table	0.892	0.285
bias	1 - (0.89)$^{1/2}$ = *0.056*	0.124

Probability of having desk in doc → 0.876

Vocab	Model summary distribution
	ratio of word in model summary / ratio of word in doc
red	(1/3) / (1/5) = 1.66
cat	(1/3) / (1/5) = 1.66
tall	0 *(absent)*
table	0 *(absent)*
desk (OOV)	(1/3) / 0.876 = 0.381

Fig. 8. OOV handling in summary

IV. RESULTS AND DISCUSSION

Table IV shows the comparison of scores between Personalization vs. Accuracy w.r.t user profiles models.

$$f(multiplier) = \begin{cases} 1/((ratio\ of\ word\ in\ model\ summary)/(sum\ of\ all\ words\ in\ doc)) & ;\ element\ with\ maximum\ similarity\ score \neq bias \\ 0 & ;\ if\ bias\ have\ higher\ score\ than\ most\ similar\ word\ in\ doc \end{cases}$$

$where,$
$bias = 1 - \sqrt{max\_sim\_score}$

Fig. 9. Approach to handle OOV

User embedding used in PENS	**e-DINS$_{sub}$**	**ROUGE 1**	**ROUGE 2**	**ROUGE L**
PENS + EBNR [7]	**0.765984**	25.13	9.03	20.73
PENS + NAML [8]	0.767602	**27.49**	0.14	**21.62**
PENS + NRMS [9]	1.861643	26.15	**9.37**	21.03

Personalization vs. Accuracy w.r.t user profiles models

A. Consistency of e-DINS$_{sub}$ as a degree of personalization measure

Fig 10 shows that in the case of NRMS+PENS, when the subjectivity constant is set to 1, and the experiment is performed multiple times, then we consistently get $\approx$ 0.94 DINS$_{sub}$ score and $\approx$ 1.8 e-DINS$_{sub}$ score.

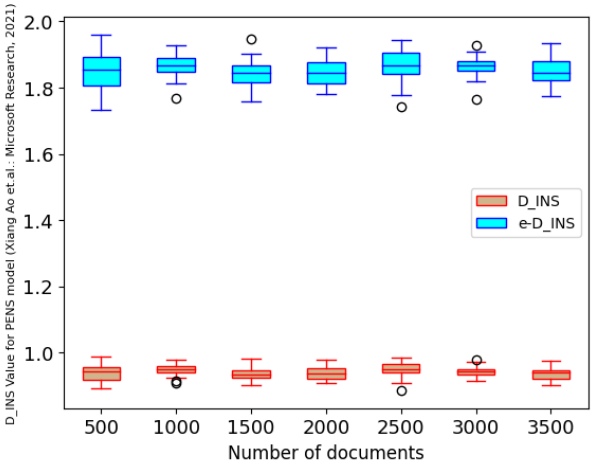

Fig. 10. Score need to have human agreement

Fig 11 shows that the PENS framework is not truly personalized w.r.t. subjective judgment of users' gold reference summaries.

B. Reliability of e-DINS$_{sub}$ as a degree of personalization measure

As shown in fig. 12, correlation with humans is required to prove that the proposed metric is reliable. The reliability of e-DINS$_{sub}$ with the human agreement can be checked in two ways. 1. Human agreement via ROUGE (Indirect way) 2. Human agreement via a survey (Direct way).

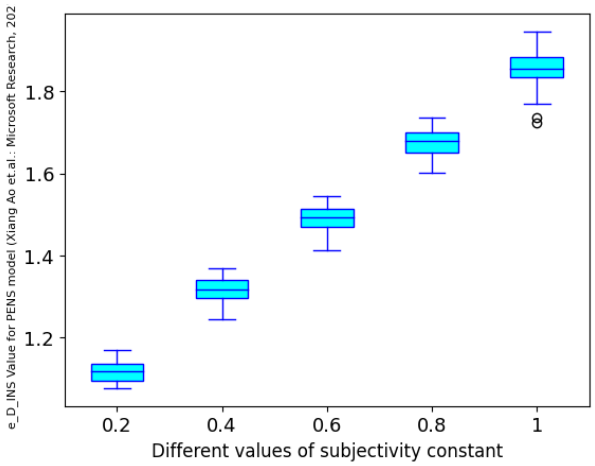

Fig. 11. e-DINS$_{sub}$ increases as subjectivity constant γ_{sub} increases

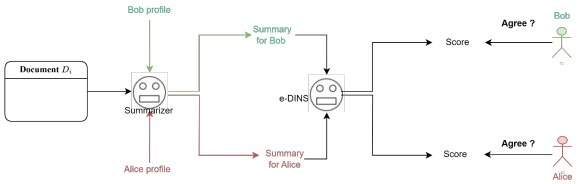

Fig. 12. Score need to have human agreement

C. Human agreement via ROUGE

ROUGE score is a widely used metric with a high correlation with human judgment compared to other measures of accuracy [10]. So let say if Bob agrees with ROUGE, then Bob has to agree with e-DINS$_{sub}$. This is how ROUGE allows us to obtain indirect human agreement.

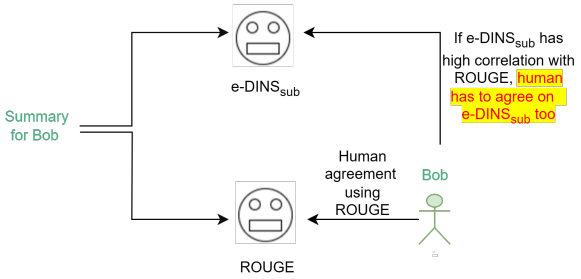

Fig. 13. OOV handling in summary

Fig. 14, 15 and 16 shows the correlation between e-DINS$_{sub}$ with ROUGE using Pearson, Kendall and Spearman. This

experiment was performed on NRMS+PENS. It's observed that when we holdout is high i.e., 80% in that case, we get a high correlation between e-DINS$_{sub}$ and also have low variation while calculating results multiple times.

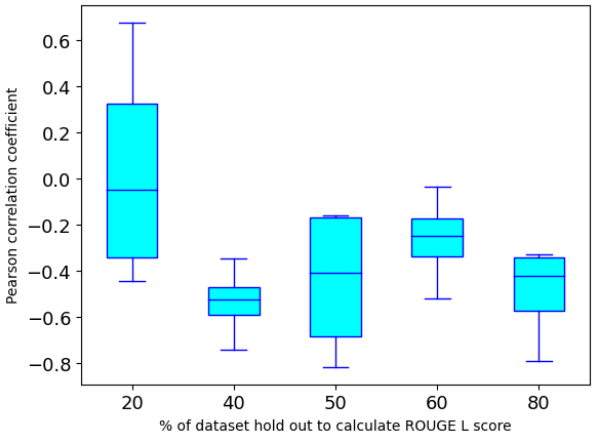

Fig. 14. Pearson correlation between e-DINS$_{sub}$ with ROUGE L

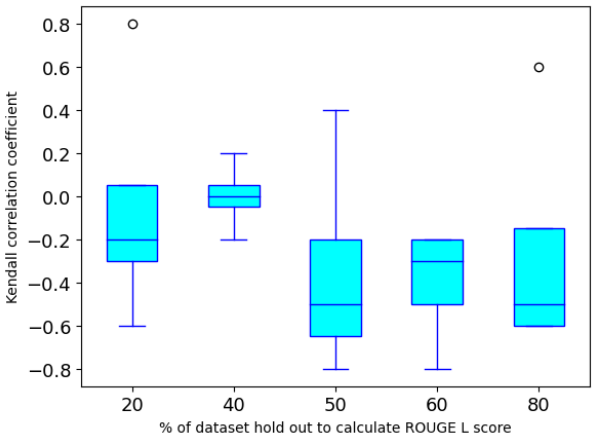

Fig. 15. Kendall correlation between e-DINS$_{sub}$ with ROUGE L

V. Future Work

There can be many variations of the same formula, or a distinct approach can be developed while still pursuing the same goal. For example, we considered differences in terms of divergence, where one can also calculate differences in terms of vector space.

A. Human agreement via survey

Rather than obtaining indirect human agreement via ROUGE, one can obtain direct human agreement by taking human responses via survey. A survey may include assigning similarity scores to user-written or model-generated summary pairs. This similarity score can be used to find DINS$_{sub}$ and e-DINS$_{sub}$. More specifically, the score we get from the distribution comparison will be replaced by the similarity score that we get via the survey, which also tells the difference

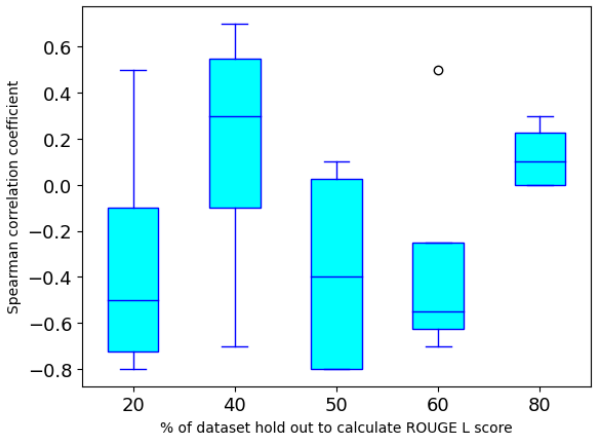

Fig. 16. Spearman correlation between e-DINS$_{sub}$ with ROUGE L

between the two summaries. Hence, the value of JSD used in Div and PF will be replaced with a similarity score.

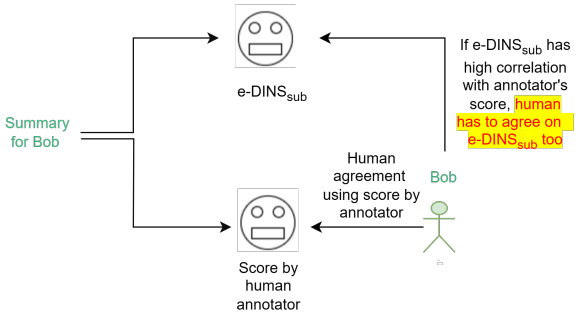

Fig. 17. OOV handling in summary

B. Metric that can measure accuracy and personalization

If the goal is to have a metric that gives a score based on accuracy and personalization, then in our case it is possible to formulate such a way that the metric is a function of e-DINS$_{sub}$ and ROUGE.

C. Analysis of model that includes prompt or instruction

Besides the PENS framework, we can use e-DINS$_{sub}$ to evaluate the degree of personalization in prompt-based models like OpenAI's Davinci and Stanford's Alpaca. We can give a few samples in the prompt that includes document and personalized summary written by a user, and then ask the model to learn about the user's interests from these examples and create a personalized summary for that user for the unseen document.

Models in which there is provision to instruct the model using conversation. Instruction-based models like chatGPT can also be evaluated using e-DINS$_{sub}$, in which a document was given to the bot and instructed to generate a summary. Once the summary is generated, user-written personalized summary shows the bot saying that the user would have generated a

summary this way. This is how the model learns the user's interest from the conversation.

VI. Conclusion

Our motive in the paper was to show that measuring accuracy is not the same as measuring personalization. In fact, a model may have high accuracy but still have a low degree of personalization. ROUGE does not consider subjectivity, so there must be a metric that takes into account the user profile in addition to the summary to determine the degree of personalization.

References

[1] D. O. Cajueiro, A. G. Nery, I. Tavares, *et al.*, *A comprehensive review of automatic text summarization techniques: Method, data, evaluation and coding*, 2023. arXiv: 2301.03403 [cs.CL].

[2] S. Ghodratnama, M. Zakershahrak, and F. Sobhanmanesh, "Adaptive summaries: A personalized concept-based summarization approach by learning from users' feedback," *CoRR*, vol. abs/2012.13387, 2020. arXiv: 2012.13387. [Online]. Available: https://arxiv.org/abs/2012.13387.

[3] S. Ghodratnama, A. Beheshti, M. Zakershahrak, and F. Sobhanmanesh, "Extractive document summarization based on dynamic feature space mapping," *IEEE Access*, vol. 8, pp. 139 084–139 095, 2020. DOI: 10.1109/ACCESS.2020.3012539.

[4] X. Ao, X. Wang, L. Luo, Y. Qiao, Q. He, and X. Xie, "PENS: A dataset and generic framework for personalized news headline generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 82–92. DOI: 10.18653/v1/2021.acl-long.7. [Online]. Available: https://aclanthology.org/2021.acl-long.7.

[5] A. See, P. J. Liu, and C. D. Manning, *Get to the point: Summarization with pointer-generator networks*, 2017. arXiv: 1704.04368 [cs.CL].

[6] M. Lewis, Y. Liu, N. Goyal, *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. [Online]. Available: https://aclanthology.org/2020.acl-main.703.

[7] S. Okura, Y. Tagami, S. Ono, and A. Tajima, "Embedding-based news recommendation for millions of users," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, Halifax, NS, Canada: Association for Computing Machinery, 2017, pp. 1933–1942, ISBN: 9781450348874. DOI: 10.1145/3097983.3098108. [Online]. Available: https://doi.org/10.1145/3097983.3098108.

[8] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, and X. Xie, *Neural news recommendation with attentive multi-view learning*, 2019. arXiv: 1907.05576 [cs.CL].

[9] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, and X. Xie, "Neural news recommendation with multi-head self-attention," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6389–6394. DOI: 10.18653/v1/D19-1671. [Online]. Available: https://aclanthology.org/D19-1671.

[10] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, *Benchmarking large language models for news summarization*, 2023. arXiv: 2301.13848 [cs.CL].