PSED: A Dataset for Selecting Emphasis in Presentation Slides

Amirreza Shirani[†], **Giai Tran**[†], **Hieu Trinh**[†], **Franck Dernoncourt**[‡], **Nedim Lipka**[‡], **Jose Echevarria**[‡], **Thamar Solorio**[†], **and Paul Asente**[‡]
[†]University of Houston [‡]Adobe Research
[†]{ashirani,gltran,httrinh,tsolorio}@uh.edu
[‡]{franck.dernoncourt,lipka,echevarr,asente}@adobe.com

Abstract

Emphasizing words in presentation slides allows viewers to direct their gaze to focal points without reading the entire slide, retaining their attention on the speaker. Despite many studies on automatic slide generation, few have addressed helping authors choose which words to emphasize. Motivated by this, we study the problem of choosing candidates for emphasis by introducing a new dataset containing presentation slides with a wide variety of topics. We evaluated a range of state-of-the-art models on this novel dataset by organizing a shared task and inviting multiple researchers to model emphasis in slides.

1 Introduction

Presentation slides have become so commonplace that researchers have developed resources for designing effective slides (Alley and Robertshaw, 2004; Alley and Neeley, 2005; Jennings, 2009). These guidelines cover advice on the overall style, such as choosing colors and font size to ensure readability from a distance, as well as ways to help the content stand out more distinctly. However, recommendations to enhance the slides' communication power could improve authoring even more.

Our goal is predicting emphasis words in presentation slides. Emphasis uses special formatting like **boldface** or *italics* to make words stand out. Well-designed emphasis can significantly increase the viewers' retention by guiding their focus to a few words (Alley and Robertshaw, 2004). Instead of reading the entire slide, they can read only the emphasized parts, keeping their attention on the speaker and their speech, as Figure 1 illustrates.[1]

The Emphasis Selection (ES) task was initially introduced by Shirani et al. (2019) with a focus on short written text in social media, and later became a SemEval 2020 task (Shirani et al., 2020b). In this paper, we focus on presentation slides, introducing a new corpus as well as automated emphasis prediction approaches. We are among the first to use the content of the slides to provide automated design assistance.

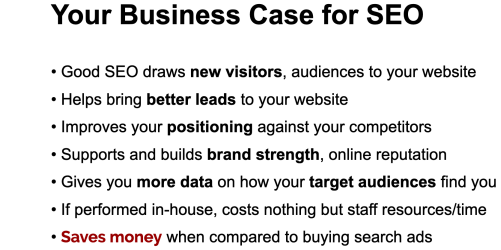

Figure 1: The slide uses special formatting to emphasize salient content.

Task Characteristics Emphasis selection poses new challenges specific to presentation slides. They can have different structures, and authors may follow traditional styles, or modern styles with more visual content. Slides cover a wide range of topics, from technical, marketing, and legal presentations to children's illustrations. The requirement to generalize to different domains and cover a variety of topics poses new challenges and encourages developing robust language understanding models. We rely only on input text without additional context from the user or the rest of the design. The task is highly subjective, but the goal is straightforward: use natural language understanding techniques to discover the most most common interpretation of a slide page and to generate emphasis that makes the page easier to understand quickly.

Benchmarking The Task Instead of providing baselines for the proposed dataset, we organized a shared task and invited researchers to work on the new corpus. Section 6 describes the top-performing methods. By examining the challenges of the

[1]Source: Web Marketing for Fundraisers: Get Found, Get Traffic, Get Ahead (http://www.fundraising123.org/files/web-marketing-for-fundraisers-get-found-get-traffic-get-ahead652.pdf)

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4314–4320

dataset, we provide different analysis components.

2 Related Work

Prior work explored automatically generating presentation slides from documents such as scientific articles (Beamer and Girju, 2009; Wang et al., 2017; Hu and Wan, 2013; Shibata and Kurohashi, 2005; Sravanthi et al., 2009). These projects assume that a slide page is a summarization of some part of the paper, and many summarization methods have been proposed to improve the effectiveness.

Other studies provide guidelines or alternatives to traditional designs to communicate a presentation's content more effectively (Alley and Robertshaw, 2004; Jennings, 2009; Alley et al., 2006; Atkinson, 2005; Doumont, 2005). These create slides with sentence headlines and visual elements to reinforce ideas and increase the audience's retention of the information during presentation.

Many applications provide design assistance for images and text, but most use only basic heuristics. Recent work uses AI-based models to recommend design attributes based on the content (Zhao et al., 2018b,a; Shirani et al., 2020a).

Shirani et al. (2019) introduced Emphasis Selection for written text in visual media. The proposed model with an end-to-end sequence tagging architecture utilizes label distribution learning (LDL) (Geng, 2016) to handle the task's subjectivity, and predicts emphasis scores for short written texts. They trained and evaluated the model against a collection of social media short texts from Adobe Spark[2]. Later on in SemEval 2020 (Shirani et al., 2020b), 31 teams proposed novel approaches to model emphasis more effectively. The organizers augmented the social media dataset with a large dataset of short quotations. Top-performing teams (Huang et al., 2020; Morio et al., 2020; Singhal et al., 2020) used rich contextualized pre-trained language models such as ERNIE 2.0 (Sun et al., 2020), XLMRoBERTa (Conneau et al., 2019), XLNet (Yang et al., 2019), and T5 (Raffel et al., 2019).

This study focuses on a new domain, presentation slides, where emphasis serves a different purpose than in social media. For social media the main purpose is to draw the audience's attention, while for presentations, the main purpose is to help the audience better understand the content. Identifying emphasis in presentations brings unique challenges due to differences in topic, length, and document structure.

3 Task Definition

Given a sequence of tokens in a slide page, $C = \{x_1, ..., x_n\}$, the task is to compute a real value $y_i \in [0, 1]$ for each x_i in C, indicating the degree to which the token needs to be emphasized.

4 Data Collection

The Presentation Slides Emphasis Dataset (PSED)[3] is a collection of presentation slides covering a wide range of topics, from technical slides on various topics to non-technical ones such as children's material. Each instance in PSED represents one slide page along with eight annotations. We only focused on English slides. To cover a wide range of topics and areas, we collected data from different sources such as websites with .ORG and .GOV domains and slides from the ACL anthology.[4] We pre-processed all slide pages to make sure they included clean pieces of text. We removed slides that only had equations, mathematical formulas, tables, or figures and used the PDFMiner Python library[5] to extract the text. Quality control steps ensured the text and the slide matched.

4.1 Annotation Process

In an MTurk experiment, we asked nine annotators to label each page. We showed the image of the slide as well as the corresponding text and asked workers to select words to emphasize as if they were preparing the slides for their own presentation. Ten percent of the hits included quality questions to make sure the annotators read the slides.

We observed a low Fleiss' Kappa score (Shrout and Fleiss, 1979) of 0.1414 on the dataset. A closer examination revealed that the dataset included some technical and domain-specific slides that were not entirely understandable to a general audience. Therefore, we removed slides with a score below -0.05 and the overall score increased to 0.1797. We also noticed that many cases included at least one annotator with a very different selection. To provide a more consistently-annotated data

[2] https://spark.adobe.com

[3] The dataset along with the annotations can be found here: https://github.com/RiTUAL-UH/Predicting-Emphasis-in-Presentation-Slides-Shared-Task.

[4] https://www.aclweb.org/anthology/

[5] https://github.com/pdfminer/pdfminer.six

set for training, we removed the annotator for each slide with the lowest agreement to the other annotators. The final dataset contains annotations from eight annotators and has a Fleiss' Kappa score of 0.2092. Such a score is similar to the score reported in (Shirani et al., 2020b) and indicates the existence of multiple points of view about emphasis in the dataset. Table 1 shows an example of a bullet point annotated with BIO annotation data. It shows that there is more agreement selecting words such as "risk" and "management" compared to the others.

Table 1: An example bullet point along with emphasis probabilities. "B" indicates the beginning of the emphasis, "I" the inside, and "O" non-emphasis words. "Freq." shows the frequencies of "B", "I" and "O". "Emphasis Probs.", shows the emphasis probability ("B+I") over eight annotations.

Words	Freq. [B,I,O]	Emphasis Probs. [B+I]
•	[0,0,8]	0.0
Demonstrate	[1,0,7]	0.125
how	[0,0,8]	0.0
operational	[1,0,7]	0.125
agencies	[1,0,7]	0.125
are	[0,0,8]	0.0
using	[0,0,8]	0.0
NASA	[2,0,6]	0.25
data	[0,1,7]	0.125
for	[0,0,8]	0.0
risk	[3,0,5]	0.375
management	[3,3,2]	0.75

5 Data Analysis

Table 2 provides more information on the number of slides, sentences, and words in the PSED dataset. The dataset contains 1,776 high-quality slides, randomly divided among training, development and test sets of 1,241, 180, and 355 instances respectively.

Table 2: Dataset Statistics

Section	#Slides	#Sentences	#Words
Train	1241	9645	96934
Dev	180	1251	12822
Test	355	2754	28108
Total	1776	13650	137864

Table 3 describes the length of instances in the PSED dataset, giving the minimum, mean, and maximum number of words in slides for each split.

As previous research has suggested, word types have a significant role in the selection of appropriate emphasis. Therefore, in this section, we examine the role of part-of-speech tags (POS) in this task. Specifically, we choose the top 20 POS tags, which frequently occur in the training and development sets, to analyze the feature's effectiveness. We used spaCy library[6] to obtain POS tags for all tokens. To examine how the emphasis probabilities are distributed, we divided them evenly into four intervals. Figure 2 shows the occurrence of the top 20 POS tags for all token labels in our training and development sets. POS tags such as "IN", ",", ".", and ":" are more favored to have low emphasis probabilities (0–0.25). Interestingly, some POS tags like "DT", "CD," and "VBZ" have zero words in the highest emphasis probability interval (0.75–1.0). Overall, most POS tags fall into the lowest emphasis probability, and the difference lies in the (0.25–0.5) interval, where POS tags like "NN", "NNS," and "VBG" mostly appear. Similar to POS tags, other hand-crafted features such as punctuation and upper-case tokens helped improve the results of some models. This motivated us to examine the degree of emphasis probability for different lexical features. Figure 3 shows the average emphasis scores for each category in the training and development sets. Comparing all lexical features, "Uppercase_start" has the highest average emphasis score, and "Contain_numbers" and "Punctuation" have the lowest. This indicates some general trends for emphasis with respect to word categories.

Table 3: Statistics on the length of the samples computed in words

Section	Min	Mean	Max
Train	13	78	180
Dev	15	71	164
Test	17	79	181

We also performed an error analysis to examine how the length of slides can affect the prediction. The results show that longer slides are more challenging due to having more options to select.

5.1 Evaluation Metric

For better comparison with previous work in ES, we followed an evaluation method similar to Shirani et al. (2020b). This metric is specifically designed to meet the subjectivity of the task.

[6] https://spacy.io/usage/linguistic-features

Table 4: Top-performing Models with Their Ranks and Score

Teams	Best Method	*RANK*	Score 1	Score 5	Score 10
UBRI-604	XLMRoBERTa Large + Lexical Features	**0.525**	**0.335** (1)	**0.686** (1)	0.554 (2)
DeepBlueAI	Ensemble of BERT, SciBERT, ERNIE 2.0	0.519	0.330 (2)	0.667 (3)	**0.559** (1)
Cisco	Ensemble of XLNet, RoBERTa + POS Tags	0.518	0.330 (2)	0.675 (2)	0.551 (3)
Baseline	BiLSTM+ELMo	0.475	0.301 (3)	0.634 (5)	0.489 (5)
Zouwuhe	N/A	0.474	0.285 (4)	0.638 (4)	0.500 (4)

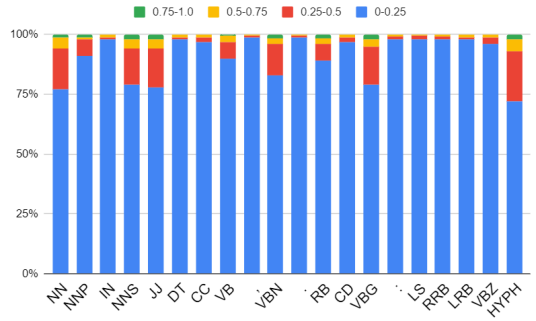

Figure 2: Frequencies of the top 20 POS tags in the 0–0.25, 0.25–0.5, 0.5–0.75, 0.75–1.00 probability intervals. Vertical values correspond to the percentage of tag counts over the total number of words in the training and development sets.

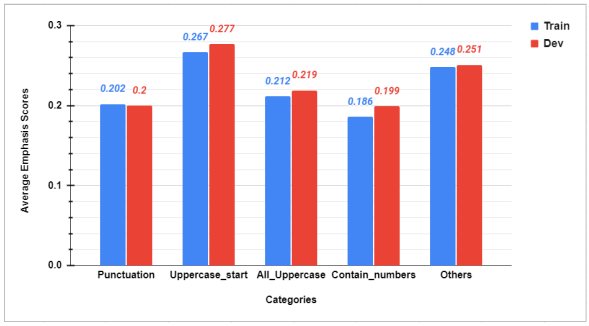

Figure 3: The figure shows average emphasis scores on the training and development sets for four different lexical features.

Match$_m$ For each slide page x in the test set D_{test}, we select a set $S_m^{(x)}$ of $m \in \{1, 5, 10\}$ words with the top m probabilities according to the ground truth. Similarly, we select a prediction set $\hat{S}_m^{(x)}$ for each $m \in \{1, 5, 10\}$, based on the prediction probabilities. Match$_m$ is defined as:

$$\text{Match}_m := \frac{\sum_{x \in D_{test}} |S_m^{(x)} \cap \hat{S}_m^{(x)}|/m}{|D_{test}|}$$

To rank models, we compute the average value of Match$_m$ for all m values and call this averaged value (*RANK*). We treat words in the ground truth with the same probability equally, so if the model predicts either of the tokens, we consider it as a correct answer.

6 Performance Benchmarks

To better examine the challenges of the dataset and benchmark the task, we organized a shared task and invited the community to participate in modeling emphasis in this new domain.[7]

Different novel and interesting solutions for this particular task were proposed. Table 4 shows the scores and the best methods for the top three teams. The most popular approach was ensemble Transformer-based models. Many hand-crafted features such as Part-of-speech (POS) tags, keywords, and lexical features (such as words with capital letters and punctuation) were explored to improve the models' performance. We describe and compare top-performing approaches next.

The top-performing team, *UBRI-604* (Hu et al., 2021), by proposing end-to-end Transformer-based approach, ranked in the first place with *RANK* score of (0.525). Different rich Transformer-based pre-trained language models were explored during the experiment, such as ALBERT (Lample and Conneau, 2019), GPT-2 (Radford and Wu, 2019), RoBERTa (Liu et al., 2019), ERNIE 2.0 (Sun et al., 2020), XLNet(Yang et al., 2019), XLMRoBERTa and BERT (Devlin et al., 2019). Comparing the results of all seven models, XLMRoBERTa performed the best. Besides pre-trained language models, *UBRI-604* leveraged lexical features such as capitalized words and punctuation, for further improvement.

DeepBlueAI team stood in second place (0.519), a *RANK* score that was 0.006 lower than the first team's. *DeepBlueAI* introduced an ensemble Transformer-based model with two fully-connected layers combined with POS tags embedding and hand-crafted features. The ensemble model takes advantage of BERT, SciBERT (Beltagy et al., 2019) and ERNIE 2.0 pre-trained language models by taking the average of the scores predicted by these models.

[7]CAD21 shared task: https://competitions.codalab.org/competitions/27419

Lastly, *Cisco* (Ghosh et al.), with a score 0.001 lower than the second team, ranked third. *Cisco* explored two approaches based on BiLSTM+ELMo (Shirani et al., 2019) architecture and Transformer-based pre-trained models with the base model of RoBERTa and XLNet. They enriched the ELMo contextual embedding in BiLSTM+ELMo model by incorporating a character-level BiLSTM Network. Their results show an increase of 0.026 when POS tags and keyphrases are added to the model, showing the effectiveness of these two features for this task. *Cisco*'s best score on the evaluation phase used an ensemble of XLNet and RoBERTa, giving them third place. They boosted the model further in the Post Evaluation phase by ensembling XLNet and BiLSTM+ELMo models and incorporating hand-crafted features like POS and Keyphrase.

We used the same baseline model (DL-BiLSTM+ELMo) introduced in Shirani et al. (2019) to better show the challenges of PSED dataset. This model achieved *RANK* score of 0.475 (Table 4) which is 0.275 lower than the reported score by Shirani et al. (0.75).[8] With a sequence-labeling architecture, this model utilizes ELMo contextualized embeddings (Peters et al., 2018) and two BiLSTM layers to label emphasis. The Kullback-Leibler Divergence (KL-DIV) (Kullback and Leibler, 1951) is used as the loss function during the training phase.

7 Discussion

The PSED dataset contains slides with different lengths. To better examine how the length of slides can affect the prediction, we performed an error analysis to examine this relationship. We divided the test set into three groups based on the instances' lengths, namely <60, 60–90, and >90 tokens. Then we computed the average Match_m scores over all shared task submissions, four in total, for every example in each group. As shown in Table 5, short slides always achieve better scores compared to medium and long slides. This indicates that predicting emphasis in longer instances is more challenging. This is due there being more options (words) to select for emphasis.

Many slides in the PSED dataset contain scientific words. Besides using pre-trained models, trained on a general domain, some teams decided to handle scientific words differently. For example, *DeepBlueAI* explored using the SciBERT (Beltagy et al., 2019) model, which is pre-trained on scientific articles. On the other hand, *Cisco* explored training a scientific keyword predictor and used the output as a feature to the model. Extending the proposed approaches to more efficiently address the diverse vocabulary of the dataset is an important future direction.

Table 5: Length vs. Performance on the test set. The average scores over all submissions are used for computing the performance. Short: (<60 tokens, 112 slides), Medium: (60–90 tokens, 126 slides), Long: (>90 tokens, 116 slides)

Length/Scores	*RANK*	Score 1	Score 5	Score 10
Short	**0.601**	**0.42** (1)	**0.634** (1)	**0.75** (1)
Medium	0.55	0.349	0.589	0.713
Long	0.485	0.293	0.526	0.635

8 Conclusion

We presented a new dataset for emphasis selection on presentation slides, posing new challenges for modeling emphasis. We created a shared task and invited researchers to model emphasis for presentation slides. We provided different data analyses on the dataset and summarized the insights gained from the shared task. A future extension could explore more robust techniques to address the challenges in the PSED dataset because of its diversity in topic, structure, and length.

9 Ethics

The proposed data in this work is collected from public domain sources and do not intrude on user privacy. For the manual work in annotation process, crowd workers were fairly compensated ($0.55 reward per response, which is over the US minimum wage).

Acknowledgments

We thank the reviewers for their thoughtful comments and efforts towards improving our work. We also thank Andrew Greene for his help in creating the corpus.

References

Michael Alley and Harry Robertshaw. 2004. Rethinking the design of presentation slides: Creating slides

[8] Match_m for $m \in \{1, 2, 3, 4\}$ is used in Shirani et al. (2020b).

that are readily comprehended. In *ASME International Mechanical Engineering Congress and Exposition*, volume 47233, pages 445–450.

Michael Alley, Madeline Schreiber, Katrina Ramsdell, and John Muffo. 2006. How the design of headlines in presentation slides affects audience retention. *Technical communication*, 53(2):225–234.

Michael P. Alley and Kathryn A. Neeley. 2005. Discovering the power of powerpoint: Rethinking the design of presentation slides from a skillful user's perspective. In *2005 ASEE Annual Conference and Exposition, Conference Proceedings*, pages 12325–12340.

Cliff Atkinson. 2005. *Beyond Bullet Points: Using Microsoft PowerPoint to Create Presentations That Inform, Motivate, and Inspire (Bpg-Other)*. Microsoft Press.

Brandon Beamer and Roxana Girju. 2009. Investigating automatic alignment methods for slide generation from academic papers.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jean-Luc Doumont. 2005. Slides are not all evil. *Technical communication*, 52:64–70.

X. Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.

Sreyan Ghosh, Sonal Kumar, Harsh Jalan, Hemant Yadav, and Rajiv Shah. Cisco at AAAI-CAD21 shared task: Predicting emphasis in presentation slides using contextualized embeddings. *arXiv preprint arXiv:2101.11422*. https://archive.org/details/CAD21.

GangQiang Hu, Chao Feng, HaoWen Lin, and JianGeng Chang. 2021. UBRI-604 at AAAI-CAD21 shared task: Predicting emphasis in presentation slides. https://archive.org/details/CAD21.

Yue Hu and Xiaojun Wan. 2013. Ppsgen: learning to generate presentation slides for academic papers. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer.

Zhengjie Huang, Shikun Feng, Weiyue Su, Xuyi Chen, Shuohuan Wang, Jiaxiang Liu, Xuan Ouyang, and Yu Sun. 2020. ERNIE at SemEval-2020 task 10: Learning word emphasis selection by pre-trained language model. *arXiv preprint arXiv:2009.03706*.

Ann Jennings. 2009. Creating marketing slides for engineering presentations. *Technical Communication*, 56.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 task 10: Emphasis distribution fusion on fine-tuned language models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1658–1664.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford and Jeffrey Wu. 2019. Rewon child, david luan, dario amodei, and ilya sutskever. 2019. *Language models are unsupervised multitask learners*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Tomohide Shibata and Sadao Kurohashi. 2005. Automatic slide generation based on discourse structure analysis. In *International Conference on Natural Language Processing*, pages 754–766. Springer.

Amirreza Shirani, Franck Dernoncourt, Paul Asente, Nedim Lipka, Seokhwan Kim, Jose Echevarria, and Thamar Solorio. 2019. Learning emphasis selection for written text in visual media from crowd-sourced label distributions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1167–1172, Florence, Italy. Association for Computational Linguistics.

Amirreza Shirani, Franck Dernoncourt, Jose Echevarria, Paul Asente, Nedim Lipka, and Thamar Solorio. 2020a. Let me choose: From verbal context to font selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 8607–8613, Online. Association for Computational Linguistics.

Amirreza Shirani, Franck Dernoncourt, Nedim Lipka, Paul Asente, Jose Echevarria, and Thamar Solorio. 2020b. SemEval-2020 task 10: Emphasis selection for written text in visual media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1360–1370, Barcelona (online). International Committee for Computational Linguistics.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420.

Vipul Singhal, Sahil Dhull, Rishabh Agarwal, and Ashutosh Modi. 2020. IITK at SemEval-2020 task 10: Transformers for emphasis selection.

M. Sravanthi, Ravindranath Chowdary, and P. Kumar. 2009. SlidesGen: Automatic generation of presentation slides for a technical paper using summarization.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *AAAI*, pages 8968–8975.

Sida Wang, Xiaojun Wan, and Shikang Du. 2017. Phrase-based presentation slides generation for academic papers. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 196–202. AAAI Press.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.

Nanxuan Zhao, Ying Cao, and Rynson W.H. Lau. 2018a. Modeling fonts in context: Font prediction on web designs. *Computer Graphics Forum (Proc. Pacific Graphics 2018)*, 37.

Nanxuan Zhao, Ying Cao, and Rynson W.H. Lau. 2018b. What characterizes personalities of graphic designs? *ACM Transactions on Graphics (Proc. of SIGGRAPH 2018)*, 37.