ユーザーの操作性を考慮した論文要約スライドの半自動生成

久保谷 善記 [1]　森島 繁生 [2]
[1] 早稲田大学　[2] 早稲田大学理工学術院総合研究所
yoshikikubotani@akane.waseda.jp　shigeo@waseda.jp

概要

スライドには，視覚的要素を用いた説明により認知負荷を下げ，聴衆の理解を促す効果がある．そのため，日頃から難解な内容を説明する必要のある研究者にとって，論文をスライドに要約し発表する機会は少なくない．しかし，研究者は研究作業や論文執筆などに時間が取られるため，スライド作成に割く十分な時間を確保できない場合がある．本研究では，このような問題を解決するため，論文からスライドを半自動で生成するシステムを提案する．作成者の裁量で自動化の度合いが調節できるよう，任意の作成段階からのスライド生成を実現する．

1　はじめに

コミュニケーションにおいては，視覚的な情報を用いることで，相手に対して自身の考えや意見を正確かつ効果的に伝達することができる．実際に，意思伝達を行う際には文章，音声，図表など様々な媒体が使用されるが，伝える内容が複雑な場合には，記述情報と視覚的情報を組み合わせることで，より円滑なコミュニケーションが期待できる．プレゼンテーションスライド（スライド）はこれら二つの情報を統合して扱うことのできる代表的な媒体として知られているが，コンパクトなテキスト（記述的情報）とその配置（視覚的情報）を用いることで聴衆の認知負荷を下げることができるため，教育現場である学校から職場まで幅広く使われている．

同様の理由で，学術的な場における意思伝達にもスライドは頻繁に使用されており，難解な内容を多く扱う研究者にとって必要不可欠な媒体である．特に，学会や研究室，輪読会・勉強会といった場においては，研究者は自身の研究や技術的文書から得た情報など，高度に専門的な内容の発表・共有を行う必要があるため，記述のみでは理解困難な内容を他の研究者に伝える手段として，スライドは重宝されている．

このように，研究者にとってスライドは情報発信に欠かせない媒体であるが，その作成には多くの時間と労力を要する．まず，研究者は日々の研究活動や論文の調査・執筆などで忙しく，スライドの作成に十分な時間をかけることは難しい．特に，近年は広く学問領域全体で論文数が増加しており，最新の手法に追随するために必要な論文調査の量は膨大である [1]．また，専門的な内容をわかりやすくスライドとしてまとめるには経験と技術を要する．具体的には，作成者がどのような目的で情報発信を行うのか，聴衆がどの程度の前提知識を持っているか，発表時間や全体のページ数に制約はないかなどの条件を考慮する必要があり，全ての条件を満たすようにページごとの説明の粒度や構図の配置を決定するには，経験者であっても試行錯誤を要する．

近年は，スライド作成にかかる労力を低減させるべく，論文要旨をとらえたスライドの自動生成に関する手法が提案されてきた．しかしその多くは，スライドテキストの生成に主眼を置き要素の配置を適切に扱っていない [2–5]，または逆にレイアウトの生成のみを扱っている [6] などの制約があり，テキストとレイアウト両者の情報を同時に扱うことには課題が残る．また，両者を同時に扱っている研究も，論文からスライドを生成する過程に人間が関与することは考慮しておらず，作成者の意図や好みを反映した自由度の高い生成は実現できていない [7]．

本研究では，多忙な研究者のスライド作成を補助する目的で，論文からスライドを半自動で生成するシステムを提案する．スライド作成者の熟練度やデザイン嗜好によって自動化の度合いが調節できるよう，タイトルのみが与えられた状態からだけでなく，追加で任意のスライド要素が与えられた状態からのスライド生成を実現する．

2　関連研究

スライドの自動生成に関する議論は，主として自然言語処理分野において論文の要旨要約という視点

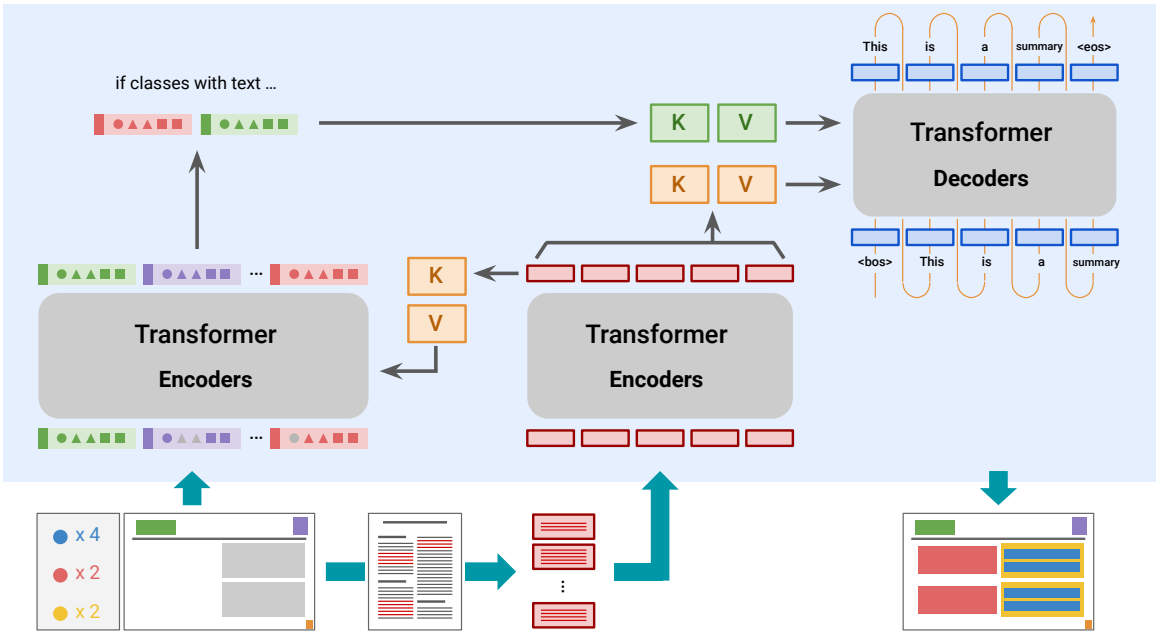

図 1: 提案手法の概要図

から行われてきた．2000 年代初頭の初期の研究においては，形態要素分析や決定的なルールによるスコアリングなどを用いて，スライドに載せるに相応しい長さのテキストを作成する研究 [3,8] が主流であった．その後は，サポートベクター回帰 (SVR) などの回帰モデルを訓練することで論文中の文章の重要度を学習し，整数線形計画法（ILP）によってスライドに割り当てるという手法が広く用いられるようになった [4,5,9]．また近年では，Encoder-Decoder 型のニューラルネットワークを用いた抽象型要約による研究が提案されている [2,7]．しかし，上記の先行研究には，取得したテキストをルールベースで配置していたり，テキストと図以外の要素をレイアウト生成の対象には含めていないといった問題がある．

3 提案手法

以降の議論のため，ここで本稿内で使用する記号について説明する．論文 P_i とそれと対を成すスライド D_i の集合を $\{P_i, D_i\}_{i=1}^{I}$ と置く．この時，スライド D_i が J 枚のスライドページからなるとすると，$D_i = \{P_j\}_{j=1}^{J}$ である．また，スライドページ P_j 内に配置される N 個の要素は，クラス C_n，サイズ (W_n, H_n)，位置 (X_n, Y_n) の 3 属性によって表現されるとする．ここでクラスとは，タイトルテキストや箇条書き，図や表というように，配置される要素がスライド内でどのような役割を果たすかを簡潔に表すものである．スライドの要素がテキストを含むクラスに属する場合，そのテキストは T_n で表記する．論文の文章については，4 文を 1 つのスニペット S として定義し，文章単位でなくスニペット単位で論文のテキストデータを扱った．従って，論文 P_i が K 組のスニペットからなるとすると，$P_i = \{S_k\}_{k=1}^{K}$ となる．なお，過度に複雑な表記を避けるため明示的に記載はしていないが，スニペット数 K やスライドページ数 J は論文やスライドごとに異なるため i に依存し，スライドページ内の要素の数 N はページごとに異なるため i と j に依存している．

ある論文からその要旨を捉えたスライドを生成する際には，以下の 3 つの点に配慮して作成する必要がある．

1. 現在のスライドページに論文のどの内容が載るべきか
2. 載るべき内容に沿って，スライドページをどのようなレイアウトとするか
3. レイアウトをもとに，実際にどの程度の粒度のテキストを載せるか

これらの要件をもとに，我々は図 1 に示すような Attention 機構を用いたモデルを提案する．まず要件 1 を満たすための前処理として，スライドページに関連するスニペットを論文から取得する．具体的には，式 (1), (2) に示すように，事前学習済みの言語モデルを使って論文のスニペット S_k と各スライドページのスライドタイトル T_n　s.t. C_n = title のそれ

ぞれから埋め込み $\boldsymbol{E}_s^{(k)}, \boldsymbol{E}_t^{(n)}$ を取得する．その後，式 (3) のように埋め込み同士で内積を取り，各スライドページ P_j ごとに内積スコア R_{kj} の上位 20 個のスニペットを用意する．スライドページの内容を最も端的に表すのはスライドタイトルであるため，この操作によって論文からスライドページに関わりのある情報だけを抽出できる．なお，埋め込み表現獲得のための言語モデルには Sentence-BERT [10] を使用した．

$$\boldsymbol{E}_s^{(k)} = \text{Sentence-BERT}(S_k) \tag{1}$$

$$\boldsymbol{E}_t^{(n)} = \text{Sentence-BERT}(T_n) \quad \text{s.t. } C_n = \text{title} \tag{2}$$

$$R_{kj} = \boldsymbol{E}_s^{(k)} \cdot \boldsymbol{E}_t^{(n)} \tag{3}$$

提案モデルは大まかに 2 つのモジュールに分かれており，それぞれレイアウト生成と論文要旨要約を担当する．レイアウト生成モジュールでは，Gupta らの研究 [11] に倣い，部分的にマスクされた系列 $[C_1, W_1, H_1, X_1, Y_1, \cdots C_N, W_N, H_N, X_N, Y_N]$ の回帰問題としてレイアウト生成を定義した．ネットワークには Xiang ら [12] の提案する双方向型の Transformer Encoder をベースとして用いており，系列内での関係を Self-Attention 機構によって捉えている．また，要件 2 で述べたように，各スライドページのレイアウトはそのページに載る内容に依存するべきである．従って，論文要旨要約モジュールの Encoder の出力とレイアウト生成モジュールの隠れ状態との間で Cross-Attention を取ることで，前処理で取得した内積スコア上位 20 組のスニペットとレイアウトとの関係を捉えている．一方，論文要旨要約モジュールには，訓練済みの BART [13] をベースとしたネットワークを用いている．要件 3 を満たすために，レイアウト生成モジュールで推定されたレイアウトのクラス $\hat{C}_n$，サイズ $(\hat{W}_n, \hat{H}_n)$，位置 $(\hat{X}_n, \hat{Y}_n)$ の埋め込みと BART の Decoder の隠れ状態との間で Cross-Attention を取っている．このようにすることで，Decoder は推定されたレイアウトの情報を参照しつつ文章要約を行うことが可能になる．

上記のアプローチでは，レイアウト系列のマスクの割合により，スライドの初期状態の調節が可能である．従って，目的や好みに合わせてスライド内の要素の数やクラス，サイズや位置を自由に編集することで，スライド作成者は生成過程に直接関与することができる．

表 1: 生成レイアウトの要素重複度比較

初期状態	IoU の平均値 (↓)
(a) タイトルのみ	**0.072**
(b) タイトル+他要素のクラス	0.078
(c) タイトル+他要素のクラス・サイズ	0.091
正解レイアウトの要素重複度	0.066

表 2: 生成-正解レイアウトの類似度比較

初期状態	類似度 (↑)
(a) タイトルのみ	0.079
(b) タイトル+他要素のクラス	0.123
(c) タイトル+他要素のクラス・サイズ	**0.134**

4 実験設定

提案手法の有効性を検証するためには，論文のテキストと，それに対応するスライドのテキストとレイアウトのすべてを含むデータセットが必須である．しかし，我々が調べた限りでは，そのようなデータセットは公開されていなかった．従って，ACL Anthology [1)] で公開されている発表資料から，小規模なデータセットを作成して実験を行った．データセットは 24 報の OCR された論文と，それに対応する同数のスライドからなる．総計 2885 個のスライドの要素はタイトル，箇条書き，見出し，その他テキスト，図，表，アイコンの 8 クラスのいずれかでアノテーションされており，これを訓練とテストでデータを 4:1 に分割して実験に用いた。

実験は，レイアウト生成モジュールと論文要旨要約モジュールの機能を確認するために，生成されたレイアウトとテキストについて行った．レイアウトについては，異なる作成段階から生成した結果を，2 つの指標で評価した．具体的には，(a) タイトルのみ指定，(b) タイトルとその他の要素のクラスを指定，(c) タイトルとその他の要素のクラス・サイズを指定，の 3 つの作成段階を初期状態としてレイアウトを生成した．生成されたレイアウトにおける要素同士の重複具合を確認するための指標として，式 (4) に示されるスライドページごとの IoU を算出し，その平均値をもって比較した．また，レイアウト類似度 [14] を用いることで，生成レイアウトと正解レイアウトの間の類似度を評価した．これは，クラスやサイズ，位置をもとに 2 つのレイアウトの要素間で関連度を算出し，関連度が最大となるマッチングの結果から，レイアウト全体としての類似度を 0 か

1） https://aclanthology.org/

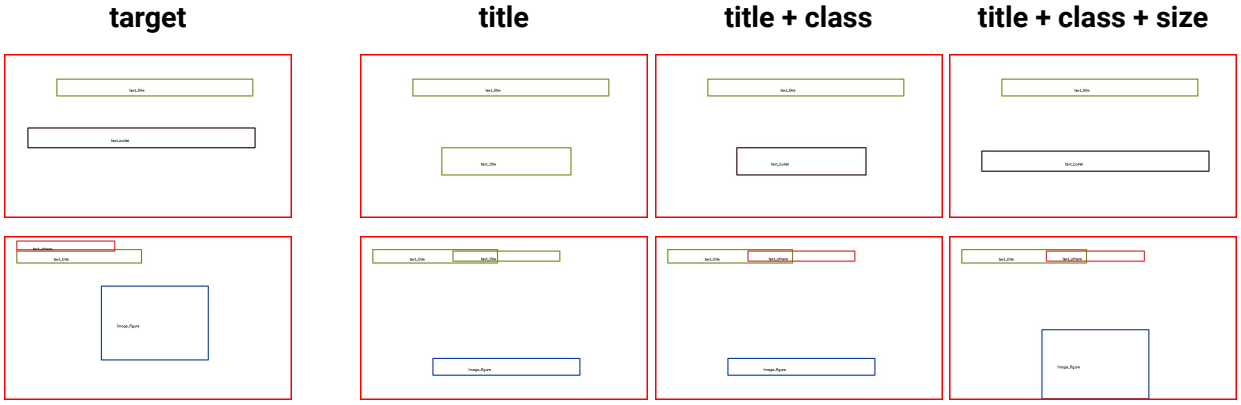

図 2: レイアウト生成結果の描画

表 3: 生成-正解テキストの類似度比較

	P	R	F
ROUGE-1	4.67	15.27	6.66
ROUGE-2	0.21	1.00	0.33
ROUGE-L	4.26	14.05	6.07

ら 1 のスコアで表す指標である．

$$\mathrm{IoU}_j = \frac{P_j\text{内の要素が重複した領域}}{P_j\text{内の要素が存在する領域}} \tag{4}$$

テキストについては，生成されたテキストと正解スライドテキストの類似度を ROUGE スコアで比較した．なお，生成テキストと正解テキストとの対応が取れなければ指標の計算ができないため，論文要旨要約モジュールに渡すレイアウトの情報には，レイアウト生成モジュールの生成結果ではなく正解レイアウトを使用した．

5 結果

スライド要素のページ内重複度とレイアウトの類似度のそれぞれについて，3 つの作成段階を初期状態として比較した結果を表 1，表 2 に示す．表 2 の結果から，完成に近い状態から生成を始めることで，正解レイアウトに近いレイアウトを生成できていることがわかる．一方で，要素の重複度は条件を与えて生成することで増加した．これは，ページ内の領域を多く占める可能性のあるクラスの要素が強制的に配置され，要素同士がより重なりやすくなったことが原因と考えられる．

また，生成されたスライドテキストが正解テキストとどれくらい類似していたかの結果を表 3 に示す．Precision に比べて Recall が高いが，これは生成されたスライドテキストが正解テキストに比べて長いためである．実際に，評価に使用したデータについて，生成テキストと正解テキストの平均単語数を計算したところ，前者は 50.63 words，後者は 7.52 words であった．また 2-gram についての ROUGE スコアが低いことから，生成テキストは正解テキストに含まれる 2 語以上のフレーズを再現することができていないことがわかる．

最後に，各条件のもと生成されたレイアウトと正解レイアウトを描画して比較した結果を図 2 に示した．条件が増えるごとに正解に近いレイアウトが生成できていることが見て取れる．ここではスペースの都合上載せていないが，要素の数が多い場合や複雑なレイアウトを持つ場合には，規則性のない配置になってしまっていた．これは，データ数が 1500 程度の小規模なデータを使用していたため，スライドのデータ分布が持つ特徴を完全には学習できなかったことを示している．

6 おわりに

本研究では，研究者にとって不可欠なスライドの作成コストを下げるため，論文からスライドを半自動で生成可能なシステムを提案した．レイアウトとテキストの両者を互いに考慮した生成を可能にするため，Attention 機構を用いたモデルを採用し，評価のため小規模なデータセットを作成して実験を行った．実験では，小規模なデータセットを用いたことから，全体として性能は高くはなかったものの，任意の編集段階からスライドを生成可能なモデルの可能性を示した．今後の課題として，より大規模なデータセットを用いて実験を行うことや，より多様な種類のモーダルを考慮した生成の仕組みを模索することが挙げられる．

参考文献

[1] Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. **Journal of the Association for Information Science and Technology**, Vol. 66, No. 11, pp. 2215–2222, 2015.

[2] Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy X.R. Wang. D2S: Document-to-Slide Generation Via Query-Based Text Summarization. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2021.

[3] Tomohide Shibata and Sadao Kurohashi. Automatic Slide Generation Based on Discourse Structure Analysis. In **International Conference on Natural Language Processing**, Vol. 3651, pp. 754–766, 2005.

[4] Yue Hu and Xiaojun Wan. PPSGen: Learning-Based Presentation Slides Generation for Academic Papers. **Knowledge and Data Engineering, IEEE Transactions on**, Vol. 27, pp. 1085–1097, 2015.

[5] Athar Sefid, Jian Wu, Prasenjit Mitra, and C. Lee Giles. Automatic Slide Generation for Scientific Papers. In **SciKnow@K-CAP**, 2019.

[6] Chuhao Jin, Hongteng Xu, Ruihua Song, and Zhiwu Lu. Text2Poster: Laying Out Stylized Texts on Retrieved Images. In **IEEE International Conference on Acoustics, Speech and Signal Processing**, pp. 4823–4827, 2022.

[7] Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. Doc2ppt: Automatic presentation slides generation from scientific documents. Vol. 36, No. 1, pp. 634–642, 2022.

[8] Mostafa Shaikh, Mitsuru Ishizuka, and Md Tawhidul Islam. 'Auto-Presentation': a multi-agent system for building automatic multi-modal presentation of a topic from World Wide Web information. In **IEEE/WIC/ACM International Conference on Intelligent Agent Technology**, pp. 246–249, 2005.

[9] A. Ashray Bhandare, Chetan J. Awati, and Sonam Kharade. Automatic era: Presentation slides from Academic paper. **International Conference on Automatic Control and Dynamic Optimization Techniques**, pp. 809–814, 2016.

[10] Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In **Empirical Methods in Natural Language Processing**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.

[11] K. Gupta, J. Lazarow, A. Achille, L. Davis, V. Mahadevan, and A. Shrivastava. LayoutTransformer: Layout Generation and Completion with Self-attention. In **Proceedings of the International Conference on Computer Vision**, pp. 984–994, 2021.

[12] Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. BLT: Bidirectional Layout Transformer for Controllable Layout Generation. In **Proceedings of the 17th European Conference on Computer Vision**, pp. 474–490, 2022.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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**, pp. 7871–7880, 2020.

[14] Akshay Gadi Patil, Omri Ben-Eliezer, Or Perel, and Hadar Averbuch-Elor. READ: Recursive autoencoders for document layout generation. In **CVPR 2020 Workshop on Text and Documents in the Deep Learning Era**, 2020.