

## 機械学習を用いた要修正なスライドの自動発見

川瀬 卓也<sup>†</sup> 大平 茂輝<sup>‡</sup> 長尾 確<sup>†</sup><sup>†</sup>名古屋大学 大学院情報学研究所<sup>‡</sup>名古屋大学 情報基盤センター

## 1. はじめに

研究発表では、パワーポイント等で作成したスライドを使用することが多い。発表内容の正確な理解と良い議論のためにスライドが果たす役割は大きい。発表に不慣れな場合や十分な推敲時間が確保できない場合、内容が不十分で分かりにくいスライドになってしまう。

そこで本稿では、スライドから抽出した特徴を用いて、修正が必要なスライドを自動的に発見する手法を提案する。

## 2. 学習対象となるスライドおよびデータ

本研究の学習対象となるスライドは、学生が学会発表本番までに繰り返し行う発表練習において使用したスライドとする。著者らの研究室では、独自の発表練習支援システムを構築・運用している[1]。発表者が発表練習を行う際に、他の参加者は所持しているタブレットから指摘を行う。指摘は大きく4つのタイプに分けられ、タイプ毎に用意された具体的な内容の指摘を選択することで簡便な入力が可能となっている。そのため、参加者は発表技術や内容、スライド中の文や図などの要素に対してさまざまな指摘を行うことができる。指摘タイプとタイプ毎の具体的な内容の例を図1に示す。システムは、指摘タイプとその内容をスライドと共に指摘データとして記録している。

本研究では、「スライドの内容に対する指摘」のデータが多く存在するスライドを修正が必要とみなし、そのようなスライドを分類することで要修正なスライドの自動発見を行った。

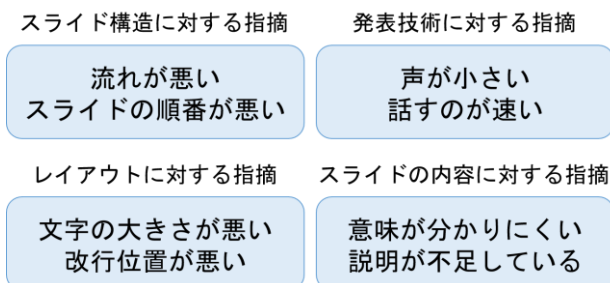


図 1. 指摘タイプと具体的な内容例

Automatic Detection of Presentation Slides that Need Correction Based on Machine Learning

<sup>†</sup>KAWASE, Takuya (kawase@nagao.nagoya-u.ac.jp)<sup>‡</sup>OHIRA, Shigeki (ohira@nagoya-u.jp)<sup>†</sup>NAGAO, Katashi (nagao@nuie.nagoya-u.ac.jp)

## 3. スライドの特徴抽出

スライドの自動評価を行う研究例としては、栗原らの研究などが挙げられる[2]。本研究では、これらの関連研究や一般的に用いられるスライド作成のノウハウなどを参考にした特徴を用いる。また、これらの特徴の他に、発表練習支援システムで使用され、指摘に基づいて修正されたスライドを比較することで、修正が必要なスライドが持つ新たな特徴を発見する。以下では、スライド比較による特徴抽出について述べる。

## 3.1 比較するスライドの収集

一般的に、最初に作成されたスライドに比べて、何度も修正を行ったスライドは、分かりにくい点や内容が不十分な点がかかり減少していることが考えられる。また、発表練習支援システムで記録された指摘データを確認した結果、発表練習を繰り返すごとにスライドの内容に対する指摘の数が少なくなる傾向が見られた。この結果から、古いスライドと新しいスライドを比較することで、修正が必要なスライドが持つ特徴を発見できると考えられる。そこで、同一タイトルのスライドで、更新日時が最古のものと最新のものを収集した。

収集したスライドは、過去2年間の発表練習に使用された、同一タイトルの最古のスライドと最新のスライド16ペアである。スライドの比較はページ単位で行うため、ページの見出しが一致し、その中でページ内容が異なる275ペアのページを取得した。スライド比較による修正箇所の確認は人手で行う必要があり、比較するスライドの内容が偏るのを防ぐため、頻出する見出しとその他の見出しでそれぞれ修正が著しいページを人手で選択した。選択したページ数(ペア数)の内訳を表1に示す。

表 1. 選択したページ数(ペア数)の内訳

見出し	ペア数
背景	5 ペア
目的・アプローチ 考察 まとめ(おわりに)	各 3 ペア
その他	10 ペア

表 2. 比較したスライドの修正内容

箇条書きの同一レベルの項目数の削減	箇条書きの文末表現の統一
箇条書きの深さの削減	連続出現した単語の削除(二重表現)
長文を箇条書きに分割	長文の短縮・削除
説明の追加	強調部分の制限
図の追加・削除	

スライドの比較は、学生と教員各1名で行った。上記のスライド内で見られた修正内容を表2に示す。

### 3.2 既存研究を参考にした特徴

著者らの研究室で作成されたスライドが持つ特徴のみでなく、より一般的なスライドにも適用できる特徴を発見するため、視覚的要素に基づく特徴[2]を参考にした。また、横林ら[3]は、文節の係り受けの関係が文の読みやすさに関連すると述べており、スライド中の文の係り受けの関係から得られる特徴を発見した。

横林らは独自の係り受けの複雑さの指標を使用していたが、スライド中の文は一般的な文章に比べ短いものが多く、文献の指標をそのまま適用することが難しいと判断した。以上の結果、指標をより単純化した以下の3つの特徴の抽出を行う。

- 正規化した係り受け距離の総和の最大値
- 係り受け距離の最大値
- 同一文節に係る文節数の最大値

### 3.3 分類実験のための特徴調整

3.1 項で確認された修正内容の中から機械的に取得できる特徴を選択し、3.2 項で抽出した特徴と合わせてスライドの分類実験に用いるために特徴を調整した。ここでは、4 節で述べる実験の際の特徴選択において、分類に有効であると考えられる特徴を下の表3に示す。

表 3. 本研究で用いたスライドの特徴

同レベル箇条書きの項目数の最大値	箇条書きの深さの最大値
1 ページ中の文の最大長	図・テキストの領域が占める割合
正規化した係り受け距離の総和の最大値	同一文節に係る文節数の最大値
二重表現の有無	

## 4. 要修正なスライドの自動発見

### 4.1 データセット

本実験では、要修正なスライドの分類をページ単位で行うため、「スライドの内容に対する指摘」が多いページを要修正なスライドとした。

要修正と判断する指摘数の閾値を 2, 3, 4 と変化させて実験を行った。要修正ではないスライドは「スライドの内容に対する指摘」が 0 であるページとした。各閾値とページ数の内訳を表 4 に示す。

表 4. スライドのページ数の内訳

要修正ではないスライド		449
要修正なスライド	閾値	2
		466
		3
		265
		4
		152

### 4.2 実験条件

本実験では、スライドの自動発見を行う分類器にランダムフォレストを用いた。これは、分類に有効な特徴を選別する特徴選択を容易に行うためである。分類結果の検証には 10 分割交差検証を用いた。また、要修正ではないスライドが多い場合、その中からランダムにサンプリングを行い、データ数を均一に揃えた。本実験は、要修正なスライドと、要修正ではないスライドの 2 値分類を行った。

### 4.3 実験結果

分類実験の結果を表 5 に示す。本実験で用いたデータセットでは、指摘が 3 個以上あるスライドページを要修正とした場合が最良の結果であり、学習モデルはこの閾値を採用することとする。

表 5. 実験結果

閾値	適合率	再現率	F 値
2	0.62	0.56	0.59
3	0.64	0.61	0.62
4	0.59	0.57	0.59

## 5. おわりに

本稿では、スライドの特徴から要修正なスライドの自動発見を行う手法を提案した。今後は、スライド分類の性能向上を目指しつつ、スライド分析による発表者へのプレゼンテーション支援手法の考案と実装などを行う予定である。

### 参考文献

- [1] 小林, 大平, 長尾: 聴き手から効果的に指摘を収集しフィードバックを容易にする発表練習システム, 情報処理学会第 77 回全国大会講演論文集, pp. 621-622, 2015.
- [2] 栗原, 加藤, 大浦: SlideChecker: プレゼンテーション資料の基礎的な定量的自動評価手法, WISS 論文集, pp. 89-94, 2009.
- [3] 横林, 菅沼, 谷口: 係り受けの複雑さの指標に基づく文の書き換え候補の生成と推敲支援への応用, vol. 45, No. 5, 情報処理学会論文誌, pp. 1451-1459, 2004, 5.