

スライド作成技術向上を目的とした文章構造解析手法の研究 Text Structure Analysis Methods to Improve Presentation Material Creation Techniques

磯邊 健太[†] 山北 謙信[†] 河崎 隆文[‡] 大倉 裕貴[‡] 岩本 健嗣[‡]
Kenta Isobe Kenshin Yamakita Takafumi Kawasaki Yuki Okura Takeshi Iwamoto

1. はじめに

研究発表や進捗報告等の場面において、内容を端的かつ正確に理解できるようにするためには、発表スライドに含まれる文章構造をレイアウトに反映させることが求められる。作成者へのフィードバックによる支援を実施することで、発表者のスライド作成技術向上が見込まれるが、発表スライドに含まれる文章構造の理解が行われていないため、現状では支援が困難である。そこで、本研究では作成支援、および作成技術向上を目的に、発表スライドに含まれる文章構造解析と、発表スライド分類手法の考案と検証を実施する。

2. 発表スライドの分類形式

文章構造の自動推定を目指す黒橋ら [1] の研究では、対象とする科学技術文の構造を説明するため、並列や対比をはじめとする 11 種類の結束関係を定義している。本研究では、研究発表等の発表スライドに存在するレイアウト構造を明らかにするため、黒橋らの結束関係や過去に作成された発表スライドを参考に、表 1 に示す 5 種類の分類形式を定義した。図 1 には分類形式毎の作成例を示す。

表 1 発表スライドの分類形式

分類形式	概要
並列	要素間に関連がありつつ、独立して説明ができるもの
順序	時系列や順序に沿って、要素を順番に説明する必要があるもの
比較	要素を並べた際に、対応する事柄同士を比較できるもの
複合	並列・順序・比較のうち、2 つ以上を組み合わせ作成されたもの
独立	タイトルスライドや目次など、発表スライドに特有なもの

3. 提案手法

発表スライドの分類にあたり、本研究にて定義する判断条件に基づき、各分類形式に属する確率を算出する手法を提案する。これは作成技術向上を意図した手法であり、明記された判断条件を、作成者が知識として習得することが狙いである。はじめに、梅澤ら [2] の研究で定義された、2 文間の接続関係を判別する文間接続関係判定ルールを参考

[†] 富山県立大学大学院 工学研究科 電子・情報工学専攻。
Graduate School of Engineering, Electrical and Computer Engineering, Toyama Prefectural University.

[‡] 富山県立大学 情報工学部 情報システム工学科。
Faculty of Information Engineering, Department of Information Systems Engineering, Toyama Prefectural University.

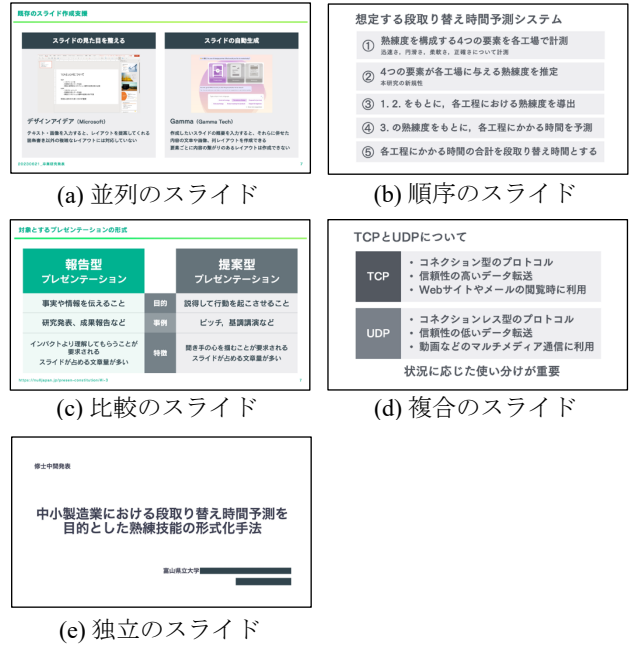


図 1 発表スライドの分類形式と作成例

に、文末表現や接続詞に注目した 26 種類の判断条件を定義する。次に、1 枚の発表スライドに含まれる文章を対象に形態素解析を実施し、解析結果を用いて、判断条件に合致する項目を算出する。ここでは、26 種類の判断条件に対し、合致する場合を 1、合致しない場合を 0 とする論理値の集合を取得する。そして、この論理値の集合を説明変数とする多項ロジスティック回帰分析を実施する。ここでは、発表スライドが並列・順序・比較・独立の分類形式に属する確率をそれぞれ取得する。以上を踏まえ、最も高い確率となる分類形式を、この発表スライドにおける分類形式として採用する。また、複合の導出には、閾値を用いた判断を実施する。具体的には、最も高い確率と次に高い確率の差が一定の閾値の範囲に収まった場合に、2 つの分類形式が含まれる複合の発表スライドとして判断する。

4. 評価実験

実験 1 では実装した分類モデルを評価する。また、実験 2 では分類に用いる判断条件を絞り込むことによる、分類の精度向上の有無を検証する。

4.1 実験 1

はじめに、分類モデルの評価に向け、正解ラベル付きのデータセットを作成した。修士課程の学生が作成した研究発表スライド 82 枚を対象に、学生へのアンケート調査によって正解ラベルを作成した。アンケートでは発表スライドを 1 枚ずつ提示し、該当する分類形式と選択の理由を収集した。アンケートの結果、複合以外の発表スライド 60

枚と、複合の発表スライド 22 枚に対する正解ラベルを取得した。複合の発表スライドは、2 つの分類形式が双方とも高い確率で推定される必要がある。そのため、複合の発表スライドについては、2 つの正解ラベルを作成することで、合計で 104 つの正解ラベルが得られた。このうち 45 つが並列、33 つが順序、19 つが比較、7 つが独立に分類された。そして、層化 5 分割交差検証を用いて分類モデルの学習と予測を実施した。正解率をはじめとする指標を用いて、分類モデルを評価した。表 2 に評価指標の一覧を、表 3 に予測した際の混同行列を示す。

表 2 分類モデル評価指標 (実験 1)

	1回目	2回目	3回目	4回目	5回目	平均
正解率	0.4286	0.5238	0.4286	0.4286	0.35	0.4319
適合率	0.25	0.2636	0.2094	0.2136	0.1771	0.2227
再現率	0.2738	0.3452	0.2659	0.3056	0.25	0.2881

表 3 混同行列

		予測ラベル			
		並列	順序	比較	独立
正解ラベル	並列	27	17	0	1
	順序	15	18	0	0
	比較	11	8	0	0
	独立	7	0	0	0

ここで、分類モデルの正解率は平均で 43.19%となった。表 3 の混同行列を確認すると、別の関係を持っているにも関わらず、多くの場合で並列の関係を予測する傾向がみられた。また、複合の発表スライドについて、予測結果を加味し 1%の閾値を用いた判断を試みたが、最も高い確率と次に高い確率の差が発表スライドごとに大きく異なり、複合関係の導出には至らなかった。しかし、複合の発表スライド 22 枚中 10 枚について、最も高い確率と次に高い確率で、2 つの分類形式が導出されていた。そのため、複合の検出方法を見直すことで、正しく複合関係が導出されると考えられる。

さらに、判断条件が分類の導出に与える影響を評価するため、オッズ比を用いた評価を実施した。並列の分類形式に注目すると、26 種類中 11 種類の判断条件について、オッズ比が 1 より大きい値を示した。そのほかの分類形式に対しても、順序では 8 種類、比較では 5 種類、独立では 4 種類の判断条件について、オッズ比が 1 より大きい値を示した。これらの判断条件は、各分類形式の導出に影響を与えるものであると解釈できる。しかし、各分類形式の導出に共通して用いられる判断条件が確認されており、特定の分類形式の導出に判断条件を利用するためには、判断条件の見直しが求められる。

4.2 実験 2

実験 1 では、26 種類の判断条件を用いて分類の導出を行った。しかし、オッズ比の低い判断条件も存在したことから、実験 2 では判断条件の絞り込みを行うことで、分類形式の導出精度向上を試みる。判断条件の絞り込みに向けて、Concept Explorer [3] を用いた形式概念分析によって、結論部に分類形式が出現するアソシエーションルールを出力し

た。ここで出力されるアソシエーションルールは「 $\langle 5 \rangle 3 5 12 14 23 = [80\%] \Rightarrow \langle 4 \rangle$ 並列;」のような形式で表現され、No. 3, 5, 12, 14, 23 の判断条件に該当する 5 枚の発表スライドのうち、4 枚 (80%) が並列の発表スライドに該当することを示している。そこで、アソシエーションルールが成立する枚数を 2 枚以上に限定し、残ったアソシエーションルールに出現する判断条件を分類モデルの説明変数として利用した。この結果、26 種類の判断条件を 15 種類まで削減した。

導出した 15 種類の判断条件を利用して、実験 1 と同様に分類モデルの学習と予測を実施した。表 4 には評価指標の一覧を示す。

表 4 分類モデル評価指標 (実験 2)

	1回目	2回目	3回目	4回目	5回目	平均
正解率	0.4286	0.5238	0.4286	0.4286	0.35	0.4319
適合率	0.25	0.2636	0.2094	0.2136	0.1771	0.2227
再現率	0.2738	0.3452	0.2659	0.3056	0.25	0.2881

ここで、実験 1 で実施した評価と比較して、正解率を始めとする評価指標の改善は確認できなかった。原因として、今回の実験で除外した 11 種類の判断条件が、実験 1 の学習の時点で既に分類の導出に影響を与えておらず、アソシエーションルールによる絞り込みとして十分に機能しなかったことが考えられる。また、今回導出した 15 種類の判断条件のうち 14 種類の判断条件が、実験 1 でオッズ比を用いて導出した判断条件と共通しており、重要な判断条件を導出するための手法として、オッズ比とアソシエーションルールが有効であると考えられる。

5. おわりに

本研究では、発表スライドに含まれる文章の内容に応じたレイアウトの作成支援を目指し、発表スライドの文章構造解析を実施した。発表スライドに 5 種類の分類形式を定義した上で、26 種類の判断条件に基づく多項ロジスティック回帰分析を実施し、43.19%の正解率で発表スライドの予測を行うことを確認した。また、オッズ比やアソシエーションルールの指標を用いて、分類の導出に影響を与える判断条件について確認した。今回、実験を通して正解率が向上しなかった理由として、正解データに対し、考慮すべき判断条件数が多く、十分に学習が行われていなかった可能性がある。また、オッズ比による評価では、複数の分類形式に影響を与える判断条件があることが確認されており、判断条件として十分に機能していなかったと考えられる。今後は、精度向上に向けた取り組みとして、正解データの追加や判断条件の見直しを実施する。また、導出した関係を、発表スライドのレイアウトに適用させる手法や、発表者に対する効果的な作成支援の手法について、検討を進める。

参考文献

- [1] 黒橋禎夫, 長尾眞, “表層表現中の情報に基づく文章構造の自動抽出,” 自然言語処理, Vol.13, No.3 (1994).
- [2] 梅澤俊之, 原田実, “センタリング理論と対象知識に基づく談話構造解析システム DIA,” 自然言語処理, Vol.18, No.1 (2011).
- [3] Yevtushenko, Serhiy A. “System of data analysis ‘Concept Explorer’,” Proceedings of the 7th national conference on Artificial Intelligence KII-2000, (2000).