論文 LaTeX 原稿からのプレゼンテーション資料自動生成

宮本 雅人　　酒井 浩之　　増山 繁
豊橋技術科学大学　　知識情報工学系
E-mail: sakai@smlab.tutkie.tut.ac.jp, masuyama@smlab.tutkie.tut.ac.jp

1　はじめに

研究のプレゼンテーションでは，限られた時間の中で，研究成果をよく理解してもらうために，プレゼンテーション資料（以下スライドとする）の準備が必要不可欠である．しかし，スライドの作成には多くの時間と手間を要する．そのため，多くの研究者がスライド作成の効率化を望んでいる．本研究では，論文 LaTeX 原稿からスライドを自動生成する手法を提案し，研究者の負担を軽減することを目的とする．具体的には，まず LaTeX 原稿から不要な情報を削除し，重要度の計算に基づくスライドへの内容の割り当てを行ない，最後に接続詞を利用した箇条書き生成を行なう．

2　関連研究

関連研究として，羽山らの研究 [3] では，隠れマルコフモデルを用いた論文とスライドの対応付け手法を提案している．しかし，この研究はスライド自動生成のための準備段階であり，スライド生成は行なっていない．安村らの研究 [2] では，スライド作成支援システムを提案している．このシステムでは，単語の重要度を計算し，それに基づいて計算したセクションの重要度に比例するように，ユーザに指定されたスライド枚数を割り当てる．本手法とは，単語の重要度の計算法やスライドへの割り当て方法が異なる．また，本手法では，文献 [2] では扱われていない並列関係の文を自動的に箇条書きにする手法も提案している．

3　提案手法

本手法では大きく分けて，3 つの Step で構成される．

Step 1: LaTeX ファイルの解析．

Step 2: 内容のスライドへの割り当て．

Step 3: 接続詞を利用した箇条書き生成．

以下にそれぞれの Step について順に説明する．

3.1　LaTeX ファイルの解析

本研究では LaTeX 形式の論文を対象とする．LaTeX ファイルは，書式を指定するためのコマンドを用いる．使用するコマンドは著者による差はあるものの，章を表す section や図を表す figure などの基本的なコマンドは，多くの著者が共通に使用する．そのため，LaTeX ファイルはある程度定型的な構造をしている．そこで，LaTeX ファイルの構造を利用すれば，スライド生成に必要な情報を特定することが可能である．この過程では，具体的には，スライド生成に必要な情報のみ残し，不要な情報を削除する処理を行なう．

表 1 に必要な情報と不要な情報の例を示す．

表 1: スライド生成に必要な情報と不要な情報の例

必要な情報	論文タイトル (title) セクションのタイトル (section) 本文 図 (figure)，表 (table)，式 (eqnarray など)
不要な情報	参考文献 (thebibliography) 謝辞 (ack) 概要 (abstract) 必要ない LaTeX コマンド (footnote など)

なお，現時点では，評価のために集めた，工学系学生の 5 つの論文に出現する LaTeX コマンドのみに対応している．

3.2　スライドへの割り当て

3.2.1　名詞の重要度の計算

解析を終えたファイルに形態素解析を行ない，その結果を利用し，各名詞に対して重要度を計算する．なお，名詞が隣接して 2 語以上出現する場合，それらを合わせて複合名詞として扱う．例えば，“プレゼンテーション資料” は形態素解析で “プレゼンテーション” と “資料” に分けられるが，それぞれ名詞であるため，合わせて “プレゼンテーション資料” とし，重要度を計算する．形態素解析には JUMAN[1]を使用した．

名詞の重要度の計算には，以下の式 (1) を用いた．各セクションを 1 つの文書と考え，論文を文書の集合 S と定義する．論文 S において，名詞 t_i の重要度 $W(t_i, S)$ の計算式を以下に示す．

$$W(t_i,S) = (0.5+0.5\times\frac{Tf(t_i,S)}{max_{i=1,\ldots,n}Tf(t_i,S)}) \times(0.5+0.5\times\frac{En(t_i,S)}{max_{i=1,\ldots,n}En(t_i,S)})$$

[1]http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

$$\times \log \frac{|N|}{df(t_i, N)} \times \frac{1}{\log |N|} \quad (1)$$

ただし，

$Tf(t_i, S)$**:** 論文（セクションの集合）S における名詞 t_i の出現頻度.

$$Tf(t_i, S) = \sum_{s \in S} tf(t_i, s) \quad (2)$$

$tf(t_i, s)$**:** セクション $s \in S$ における名詞 t_i の出現頻度.

$En(t_i, S)$**:** 論文 S において，名詞 t_i の出現確率に基づくエントロピー.

$$En(t_i, S) = -\sum_{s \in S} P(t_i, s) \log_2 P(t_i, s) \quad (3)$$

$$P(t_i, s) = \frac{tf(t_i, s)}{Tf(t_i, S)} \quad (4)$$

$df(t_i, N)$**:** コーパス N において，名詞 t_i を含んでいる文書の頻度.

第 1 項では，$Tf(t_i, S)$ によって，論文 S において高い出現頻度を持つ名詞に大きな重要度が割り当てられる．第 2 項では，$En(t_i, S)$ によって，各セクション s にまんべんなく出現する名詞ほど大きな重要度が割り当てられる．つまり，出現頻度が同じでも，あるセクションにのみ出現する名詞と比べ論文全体に出現する名詞のほうが重要な名詞であると考える．第 3 項は，コーパス N における idf 値であり，コーパス中で，その名詞を含む文書の出現頻度が低いほど，大きな重要度が割り当てられる．idf 値を計算するためのコーパス N には 1998 年の毎日新聞のデータを使用した.

3.2.2 スライド枚数の決定

文中に出現する名詞の重要度の和を各文の重要度とする．また，セクション（以下セクション，サブセクション，サブサブセクションを総称してセクションとする）中に出現する文の重要度の和を，各セクションの重要度とする．初期状態で，スライドは各セクションに 1 枚ずつ割り当てる（図や表を含めず，文のみで）.

スライドの枚数は以下の手順で決める.

Step 1: セクションの重要度の平均を計算する.

Step 2: 平均より重要度が小さいセクションは割り当てられた 1 枚のスライドで決定する．大きいセクションは Step 3 へ.

Step 3: 重要度を $1/n(n = 2, 3, 4, ...)$ し，その値が平均より小さくなるときの n をそのセクションに割り当てるスライド枚数とする.

図や表は，それだけで 1 枚のスライドとするため，上記で決定した割り当て枚数に図と表の数を加えた数がスライドの枚数となる．また，式はすべて抜き出すが，文章中に挿入するため，枚数に影響はない.

3.2.3 セクションの分割箇所の決定

スライドが複数枚割り当てられたセクションは，そのセクション内のどこで内容を分割するかを決めなければならない．その方法には，TextTiling 法 [5] を用いた．まず，段落に出現する単語（名詞，動詞，形容詞）を調べて，ベクトル空間法 [6] で段落間の類似度 (sim) を求める．ベクトル空間法では，各段落に対し，その中に出現する単語のベクトルを作成する．本手法では単純に，出現した単語に対応するベクトルの成分の値を 1，出現しない単語に対応するベクトルの成分の値を 0 とした．そして，2 段落間の類似度を，それぞれのベクトルの余弦値によって決定する．これは，段落中に出現する単語が似ているほど類似度が高いという仮定に基づいている.

次に，gap を求める．gap とは，分割箇所を決定する基準であり，以下の式 (5) で定義する.

$$gap(i, i+1) = (sim(i-1, i) - sim(i, i+1)) + (sim(i+1, i+2) - sim(i, i+1)) \quad (5)$$

$gap(i, i+1)$**:** 段落 i と段落 $i+1$ 間の gap

$sim(i, i+1)$**:** 段落 i と段落 $i+1$ 間の類似度

つまり，その段落間の類似度が低く，隣接する段落間の類似度が高い場合，gap は大きくなる．セクション内で gap が大きい順に分割箇所とする.

3.2.4 重要文の抽出

各スライドに割り振られた文の中で，重要度が大きい上位 4 文をそれぞれ抽出する．ただし，上位 4 文以内の文でも「文の重要度の平均 – 文の重要度の標準偏差」より重要度が小さい文は，重要でない文と判断し，抽出しないこととする．また，箇条書きの文は重要度に関係なく，5 文以上であってもすべて抽出することとする．抽出された文が LaTeX コマンドの “ref” を含んでいる場合は，それに対応する図や表に付加して出力する.

スライドに割り当てられた文の重要度の和を各スライドの重要度とする．最後に，重要度が「スライドの重要度の平均 – スライドの重要度の標準偏差」より小さいスライドは重要でないスライドと判断して削除する.

3.3 接続詞を利用した箇条書き生成

スライドでは箇条書き表現が多く使われる．そこで，並列関係の接続詞を利用して，自動的に箇条書きを生成する．接続詞に注目したのは，並列関係をの接続詞を含む文は，その文と対となる文が存在する確率が高いと考えたからである.

3.3.1 論文中の接続詞の調査

箇条書き生成では，EDR 辞書の並列関係を表す接続詞を利用する．その接続詞について，言語処理学会第 11 回年次大会発表論文集の中から 30 論文について，その使用頻度と箇条書き可能であるか否かを調べた．調査結果を表 2 に示す．なお，今回は接続詞が文の先頭に出現した場合のみを考える．

表 2: 接続詞の使用頻度と箇条書きの可否

	また	そして	さらに
出現頻度	94	28	18
接続詞を含む文と箇条書き可能である文が存在	74	5	7
直前の文と箇条書き可能（その中で文末が等しい）	61 (18)	5 (4)	6 (3)
箇条書き可能で文末が等しい	23	4	3

表 2 より，明らかに “また” の使用頻度が高く，箇条書き生成が可能である確率が高いことがわかる．箇条書き生成可能である場合の特徴として，次の 2 つが挙げられる．

- 文末が等しい．
- 出現する単語が似ている．

“また” の場合，この特徴のどちらかが成立すると，直前の文と箇条書き生成可能である確率が高い．また，“そして” と “さらに” に関しても，箇条書き可能である文が存在する確率が低いものの，箇条書き可能である文が存在するとき，直前の文と文末が一致している確率が高いことがわかった．

3.3.2 箇条書き生成手法

調査結果を参考に，箇条書き生成手法を考案した．接続詞 “また” を含む文とその直前の文は，以下の 2 つの条件のどちらかが成立した場合，箇条書きを生成する．

条件 1 文の述部が一致．

条件 2 文間類似度が 0.5 以上（ベクトル空間法で段落間類似度と同様の計算方法で算出）．

条件 1 では，例えば,「～を使用する」という文があるとき,「使用する」が一致した場合となる．また，条件 2 の 0.5 は，例えば，“また” を含む文が 10 単語，その直前の文が 10 単語で，5 単語一致した場合となる．抽出された重要文に “また” を含む文，または，その直前の文のどちらかが抜き出されていた場合，条件が成立すると，もう一方の文を抜き出し，箇条書きを生成する．

条件 1，2 が成立しなかったときは，同じ段落内のさらに前の文について調べ，条件 1 の成立により，箇条書きを生成する．これを条件 3 とする．ここで，対象となる文が，2 文以上見つかった場合は，すべて箇条書きとする．条件 3 に関して，条件 2 を適用しない理由は，同じ段落内であれば，箇条書きできない場合でも，文間類似度が高くなる可能性があるからである．つまり，調査より “また” が直前の文と箇条書きできる確率が高いことが前提である．“そして” と “さらに” については，条件 1 のみ適用する．

4 評価

4.1 評価方法

工学系学生 5 人の論文とそれに対応したスライドを収集し，それぞれ，論文とスライドの対応付けを行ない，正解データを作成した．評価では，次の 3 つの点について，本手法で自動生成したスライドと正解データを比較した．

- スライドの枚数と割り当て方
- 抽出した文
- 箇条書き生成箇所

4.2 評価結果

結果を表 3，4，5 に示す．

表 3: スライドへの割り当ての結果

論文	正解データのスライドの枚数	本手法で生成した枚数	差枚数	タイトルが一致した枚数
A	36	46	10	30
B	31	33	2	25
C	24	19	-5	16
D	26	33	7	23
E	25	10	-15	10

表 4: 抽出した文の結果

論文	論文の文数	正解データの文数	抽出した文数	一致した文数	精度	再現率
A	270	102	97	58	60.0%	56.9%
B	215	74	116	54	46.6%	73.0%
C	75	49	51	38	74.5%	77.6%
D	120	24	67	19	28.4%	79.2%
E	122	43	44	22	50.0%	51.2%

表 5: 箇条書き生成の結果

論文	抽出した文中で箇条書き可能な箇所	箇条書きを生成した箇所	一致した箇所	精度	再現率
A	9	3	3	100%	33.3%
B	3	2	2	100%	66.7%
C	1	0	0	0%	0%
D	3	0	0	0%	0%
E	1	0	0	0%	0%

表 3 より，正解データのスライドの枚数と本手法で生成した枚数にばらつきが見られる．特に，論文 A では本手法で生成したスライドのほうが 10 枚多く，論文 E では 15 枚も少ない．また，表 4 より，論文 C が精度 74.5%，再現率 77.6%とそれぞれよい結果を得られている．一方，論文 D では，再現率は 79.2%と高いが，精度は 28.4%と低い．そして，表 5 より，箇条書きを生成できた箇所は少ないものの，箇条書き生成した箇所はすべて正しいという結果を得られた．

5　考察

まず，スライドの割り当てについて考察する．論文 A では，本手法で生成したスライドのほうが正解データより 10 枚多い．この最も大きな原因として，論文に存在するセクションが，人間が作成したスライドでは省かれている場合があることが挙げられる．プレゼンテーションでは時間が限られるため，発表時間によっては内容を削らなければならない．これを解決するためには，プレゼンテーションでは不要なセクションを選択する機能を付けることが有効であると考えられる．そうすれば，ユーザの希望に沿った内容でスライドを生成することができ，不要なスライドを生成することを防げる．また，図や表に関しても同様で，スライド枚数に影響するため，同様の機能を付けることが有効であると考えられる．

論文 E では，本手法で生成した枚数のほうが 15 枚少ない．この原因は，論文 E は，4 つのセクションと 2 つのサブセクションで構成されており，各セクションの重要度の差が少ないからである．本研究では，初期段階では各セクションに 1 枚のスライドを割り当てるため，細かくセクション分けされている論文ほど枚数は多くなりやすい．また，本手法ではセクションの重要度の平均を閾値とし，それより重要度が大きいセクションの枚数を増やしている．しかし，各セクションの重要度の差が少ないと，各セクションの重要度と閾値の差が小さくなり，いくら情報量が多いセクションでも，スライドは 1 枚の場合が多く，多くても高々2 枚である．これらの理由から，論文 E は生成したスライド枚数が極端に少なくなってしまった．これを解決するために，スライド枚数を増やす条件である閾値にセクションの重要度の平均ではなく，文の重要度の平均を利用することが有効であると考えられる．閾値を「文の重要度の平均 $\times \alpha$」とすれば，各セクションの重要度の差に関係なく，セクションの情報量に基づいた割り当てが実現できる．また，原因の 1 つとして，論文提出からプレゼンテーションまでの間に期間があり，新しい内容が加わっている場合もあった．これに関しては，対処のしようがないため，考慮しないこととする．

次に，抽出した文について考察する．論文 C では，精度 74.5%，再現率 77.6%とそれぞれよい結果が得られた．論文 C は，4 ページで図と表を除けば 75 文程度の短い論文であり，正解データのスライドが論文を忠実に再現していることが，この結果につながった．論文 D の精度が低いのは，論文では文で説明されていた部分が，スライドでは図にして説明している場合が多かったからである．抽出した文の精度，再現率は，各セクションに割り当てるスライドの枚数が大きく影響するため，スライド枚数の調整が，そのまま精度，再現率の向上につながる．

箇条書き生成では高い精度を得ることができたが，再現率の向上が課題である．提案した箇条書き生成の条件に，新たな条件の追加を検討する余地がある．

6　まとめ

本研究では，論文 LaTeX 原稿からのスライド自動生成の手法を提案した．本手法を実装したところ，論文に忠実なスライドの生成には有効であることがわかった．そして，ユーザの希望に沿った内容でのスライド生成への対応と，セクション数の少ない論文への対応が課題であることがわかった．よって，今後は，スライド生成に不要なセクションを選択する機能の付加と，スライドの割り当て枚数を増やす閾値の見直しを行なう予定である．

また，現在のように論文中の文をそのまま抽出するだけでは，スライド作成の参考になっても，そのままスライドとしては利用できない．そこで，長い文を対象として，内容を損なわず短くする，または，2 文に分ける方法 [4] を考える必要がある．

今回は評価として，すでに完成しているスライドを正解データとし，比較を行なった．しかし，スライド生成の結果の正解は個人によって違うため，実際には明確な正解は存在しないといってもよい．そこで，スライドを作成する前に本手法を実装し，どの程度の修正で実際にプレゼンテーションで使用できるかについても評価する必要があると考えている．

謝辞

本研究の一部は文部科学省 21 世紀 COE プログラム「インテリジェント ヒューマンセンシング」および文部科学省研究費特定領域 (B) (2)16092213 の援助により行なわれた．

参考文献

[1] 酒井浩之，増山繁，“ユーザとのインタラクションを導入した複数文書要約システム”，言語処理学会第 10 回年次大会発表論文，pp.285–288，2004.

[2] 安村禎明，武市雅司，新田克己，“論文からのプレゼンテーション資料の作成支援”，人工知能学会論文誌，18 巻 4 号 F，pp.212–220，2003.

[3] 羽山徹彩，難波英嗣，國藤進，“隠れマルコフモデルを用いた論文とプレゼンテーションシート対応付け”，情報処理学会 研究報告，2004-NL-164(2).

[4] 黒橋禎夫，長尾眞，“長い日本語文における並列構造の推定”，情報処理学会論文誌，Vol.33，No.8，pp1022–1031，1992.

[5] M.Hearst, TextTiling: Segmenting Text into Multi Paragraph Subtopic Passages, Computational Linguistics , 23(1),pp.33–64,March 1997.

[6] G. Salton，“Automatic Text Processing,” Addison-Wesley，1988.