プレゼンテーションスライドからの構成抽出

坂本　祥之† 　清水　敏之†† 　吉川　正俊††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †sakamoto@db.soc.i.kyoto-u.ac.jp, ††{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし　近年，学会での発表や講義で使われたプレゼンテーションスライドが蓄積されている．学会発表資料としてプレゼンテーションスライドが用いられており，それが入手可能な場合，論文を理解するための補助として参照することも多い． しかし，プレゼンテーションスライドには文脈があり，一見してその構成を理解するのは難しい場合がある．本研究では，スライド中の語の出現やインデントといった情報から，スライド全体の構成を抽出する手法を提案する． 構成情報を用いることで，スライドの閲覧を容易にすることができると思われる．また，スライドを検索する際に構成情報を利用して部分スライド集合だけを抽出するといった応用も考えられる．

キーワード　プレゼンテーションスライド，構成抽出

1. はじめに

近年，学会の発表や学校の講義で，Microsoft PowerPoint 等のプレゼンテーションスライドが広く使わていれる．プレゼンテーションスライドは，発表，講義等に出席できなかった人や，後で見直したい人のために，Web 上で公開されることも多い．学会発表におけるプレゼンテーションスライドは，論文を読んでいて，理解できない部分があった場合に，その理解の補助として閲覧される場合がある．あるいは，ある事について理解したい場合に，出席していない講義のプレゼンテーションスライドを閲覧する場合がある．しかし，プレゼンテーションスライドには，論文のような明らかな構造がないため，どの話題とどの話題が関係しているかや，どこで話題の転換が行われているかがわかりにくい場合があり，そのような場合，プレゼンテーションスライドの構成を理解することは難しい．

本研究では，プレゼンテーションスライドから構成を抽出する方法を提案する．例えば，研究発表のプレゼンテーションスライドであれば，“背景”，“研究概要”，“提案手法”，“まとめ” といった構成がある．実際には，“提案手法” の中にいくつかの手法があるなど，複雑な構成になっている場合もある．このような構成を抽出し，提示することで，ユーザの理解を促進するようなシステム構築が可能になると考えている．

構成抽出のために，まず，プレゼンテーションスライドから，スライド中に出現する単語と，インデントの情報を抽出する．次に，構成抽出のために重要となる目次スライドの判定を行い，プレゼンテーションスライドを 3 種類に分類する．そして，その種類に応じて，プレゼンテーションスライドを “セグメント” に分ける方法を提案する．セグメントとは，プレゼンテーションスライドの中で，共通の話題について話している部分である．目次スライドの情報から，セグメント分割を行える場合は，それによりセグメント分割を行う．そうでない場合は，隣り合うスライドやセグメント同士について，共通の単語や，そのインデント情報を見ることで，セグメント分割を行う．これにより求められたセグメント同士で単語の比較等を行うことで，プレゼンテーションスライドからの構成抽出ができると考えている．

以下，Microsoft PowerPoint 等によって作られた発表用ファイルを “プレゼンテーションスライド” と呼び，その中の一枚一枚のシートを “スライド” と呼ぶことにする．

2. 関連研究

プレゼンテーションスライドからの情報取得や，検索等についての関連研究として，以下の研究が挙げられる．

羽山ら [1] [2] の研究では，1 枚のスライド内の構造化を行なっている．スライド内の配置の情報とフォントの情報から，スライド内のオブジェクトを，タイトル，本文，図，表，装飾という 5 つに分類し，一枚のスライド内の情報を木構造で表現している．一枚のスライドを木構造で表す点について，我々の手法と似ているが，我々は，タイトルや本文といった情報は，スライド内の配置からではなく，Office Open XML から取得する．また，我々はオブジェクトの分類は行なっておらず，テキストとして取れるものだけを取得している．

さらに，羽山らの別の研究 [3] は, [2] の手法をベースとして，検索要求に関連する箇所だけを抽出し，提示する手法を提案している．

北山ら [4] の研究では，スライド間の関係を求めている．この研究では，スライドの単語と，そのインデントの情報，発表テキストをデータとして利用している．スライド間の関係としては，詳細化，汎化，具体化，付加の 4 種類を定義している．ある基点スライドを決め，そのスライドに出現している一つの単語着目し，基点スライドと関係のあるスライドを求め，関係を分類している．基点スライドが同じでも，着目する単語が違うと，関係が異なる場合がある．我々は発表ビデオは利用せず，

プレゼンテーションスライドのみを使って行っているが，単語やインデント等の利用については，この研究をベースとしている．

王ら [5] の研究では，北山らのような，インデントを利用した表層的構造に加えて，WordNet を利用し，語の IS-A 関係や PART-OF 関係による概念的構造を求め，その両方を利用することにより，スライド間の関係を求めている．

別の王ら [6] の研究では，表層的構造から，スライド間の関係を Detailed と Generalized に分類している．さらに，ユーザが入力するキーワードについて良く知っている場合の DETAIL 検索と，あまり知らない場合の GENERALITY 検索の 2 種類を考え，スライドのランキングを行なっている．我々の研究では，単語の概念的構造は考えていないが，1 枚あたりの記述が少ないスライドにおいて，このような概念構造を利用するのは有用であると考えている．

Vinciarelli ら [7] の研究では，スライドを画像に変換し，文字を識別することにより，スライドからテキストの抽出を行なっている．さらに，スライドは，1 枚に含まれるテキストの量がまちまちであるため，依存しない tf-idf 法を考案し，入力されたキーワードに対して，スライドをランキングする手法を提案している．我々の提案手法の応用として検索システムを構築する場合，利用できる部分がある．

Atapattu ら [8] の研究では，プレゼンテーションスライドに出現する単語の重み付け等を行なっている．プレゼンテーションスライドのテキストとその装飾情報を抽出し，利用している．単語の出現頻度と，その装飾，その単語が長文中に出現しているか短文中に出現しているかという 3 つ情報を使い，単語に重みをつけている．我々の研究では，テキストの装飾情報は利用していないが，装飾の情報は，論文等他の文章には無い特徴である．単語の重み付けをする際に，装飾情報を考慮するのであれば，この研究の手法が利用できるのではないかと考えられる．

田中ら [9] の研究では，プレゼンテーションスライドのテキスト部のみを対象としている従来の検索の手法に対し，図形やイラストの形状や配置を考慮したプレゼンテーションスライドの検索手法を提案している．プレゼンテーションスライドは，論文等に比べて，テキストが少なく，画像や図形が多く使われるため，その形状や配置は重要な情報になると考えられる．また，我々の研究の応用として，プレゼンテーションスライドの検索システムを構築する場合，この田中らの研究の手法は参考にできる．

友安ら [10] の研究では，研究発表において，プレゼンテーションスライドを素材として作られたポスターの閲覧支援システムを構築している．ポスターは，一見しただけではどこから見るべきか分からないため，スライドのポスターの対応付けを行い，スライドの順序にもとづき，ポスターの閲覧順序を決定する．スライドとポスターの対応付けを行う際に，我々の手法と同様に，インデントの情報からそれぞれの構造情報を取得し，テキストの類似度と構造の類似度の両方の算出を行なっている．

Kan [11] の研究では，学会発表のプレゼンテーションスライドと，それに対応する論文の，部分同士の対応付けを行なっている．我々の研究では，プレゼンテーションスライド単体からの構成抽出を考えているが，対応する論文が存在する場合，論文との対応付けをとり，論文の構造を見ることでプレゼンテーションスライドの構造を抽出する手法も考えられる．

山田ら [12] の研究では，技術論文の内容からシナリオを作成し，プレゼンテーションスライドの構成支援をする手法を提案している．我々の研究でも，作成したプレゼンテーションスライドを入力として，スライドが正しく構成されているかをチェックするといったアプリケーションを作成することも可能であると考えられる．例えば，目次スライドと実際の構成が異なっていることを見つけて，作成者に注意を促す等が考えられる．

3. 提 案 手 法

本研究では，プレゼンテーションスライドから，その構成を抽出する方法を提案する．

Microsoft PowerPoint のプレゼンテーションスライドは，Office Open XML というフォーマットが用いられている．このような XML からは，スライドのテキスト情報や，インデントの情報が容易に取得できる．この XML から取り出せる情報を利用して，プレゼンテーションスライドからの構成抽出手法を考案する．

3.1 スライドからの情報取得

スライドが入力された時に，そこから必要な情報だけを取り出さなければならない．今回利用する情報は，スライドに出現する単語と，インデントの情報である．これらの情報は，上述した通り容易に取得可能であり，プレゼンテーションスライドの構成抽出に利用しやすいと考えた．この 2 つの情報を使って，3.2 節で述べるような構成抽出を行う．

まずは，スライド内のテキストを，記号やスペース等の，アルファベット以外の文字で区切り，それぞれを単語とした．さらに，それらをステミングして利用することとした．また，前置詞等は，構成抽出の際に不要なため，以下のように不必要な単語を除外した．

- 1 語の単語は構成抽出には利用しない．“a” 等がこれにあたる．
- 2 語の単語は構成抽出には利用しない．ただし，2 語の単語でも，大文字のみで構成されている語は略語である可能性が高いため，構成抽出に利用する．例えば，“Data Base” を “DB” のように略すことがある．
- ストップワードリストに含まれる単語は構成抽出には利用しない．ストップワードとは，どのような文章にも含まれる一般的な単語のことで，“for”, “you”, “about” 等がこれに含まれる．

単語が抽出できたら，その単語とインデントのレベルを組にして，構成抽出の段階で利用する．インデントについては，以下のように定めた

- タイトルを最上位である 0 とする．
- 本文の最上位を 1 とし，以下インデントされる度に 2,3,… とする．
- インデント情報が含まれていないテキストボックスも存

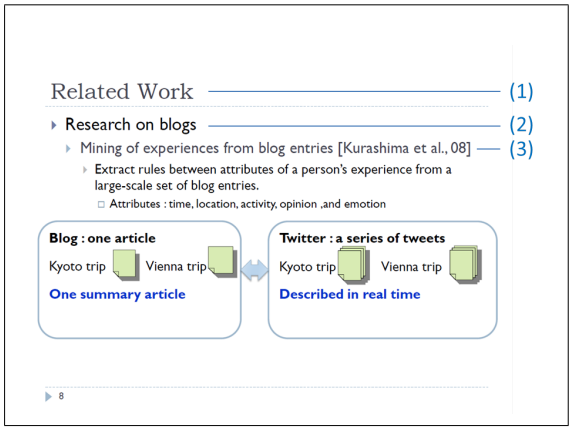

図 1　インデントの例

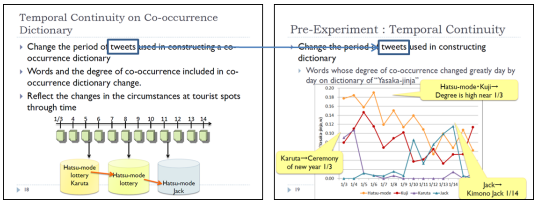

図 2　インデント変化なし

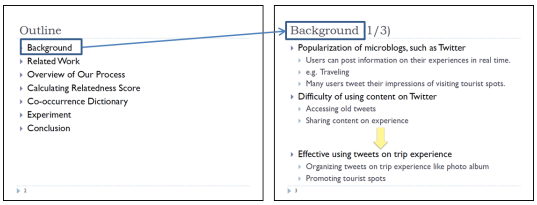

図 3　上　位　化

在する．その場合は，“インデント情報なし” とする．

例えば，図 1 の場合，(1) の部分はインデント 0，(2) の部分はインデント 1，(3) の部分はインデント 2 となる．

3.2　プレゼンテーションスライドからの構成抽出

構成抽出の前の段階として，3.1 節で抽出した単語とインデントの情報を用いることで，プレゼンテーションスライドを，共通の話題について述べたセグメントに分割する．まずは，目次スライドの判定を行う．目次スライドがある場合，それを利用することにより，プレゼンテーションスライドの構成抽出が容易になるためである．次に，2 スライド間の単語とインデントの比較により，セグメント分割を行う．2 スライドで共通の単語がどの程度出現しているか，また，そのような単語があった場合，スライド間での，単語が出現する部分のインデントの違いがどうなっているのかという情報を利用する．

3.2.1 節では，以降の議論の準備として，2 スライド間で共通している単語の出現パターンについて説明する．3.2.2 節で目次スライド判定について説明し，3.2.3 節で，セグメント分割について説明する．

3.2.1　単語の出現パターン

2 枚のスライドに共通の単語があった場合に，その単語のインデントの変化には，以下の 3 つの場合がある．

- インデントが変化なし

2 枚のスライド間に共通の単語が出現した時に，あるスライドに対して，他のスライドの方がインデントが変化しない場合である．図 2 では，左のスライドと右のスライドで，共通の “tweet” という単語が出現しており，ともに本文の最上位なので，インデントが 1 となっている．このような単語が多いスライドは，同じ話題について話しているスライドではないかと考えられる．

- 上位化

2 枚のスライド間に共通の単語が出現した時に，あるスライドに対して，他のスライドの方がインデントが少なくなっている場合，その単語は上位化されていると呼ぶ．例を図 3 に示す．図 3 では，左のスライドにおいて，“Background” という単語が本文の最上位であるインデント 1 の位置に出現しているのに対し，右のスライドでは，インデント 0 であるタイトルに出現している．さらに上位化には，本文の中でインデントが少なくなっている場合と，本文からタイトルに単語が移動している場合の 2 パターンがある．図 3 は，本文からタイトルに上位化されている場合である．スライド A から，他のスライド B,C,D に上位化されている単語が多い場合，スライド A は，スライド B,C,D の概要について述べているスライドではないかと判断できる．

- 下位化

2 枚のスライド間に共通の単語が出現した時に，あるスライドに対して，他のスライドの方がインデントが多くなっている場合，その単語は下位化されていると呼ぶ．上位化とは逆の場合である．

北山ら [4] の研究では，2 スライドと，その間で共通しているある単語に関して，スライド間の関係を，詳細化，具体化，汎化，付加の 4 つに分類している．同じスライド間でも単語ごとに関係が異なる場合もある．また関係なしと判定される場合もある．この関係の分類の際に使われるのが，その単語が上位化されているか，あるいは下位化されているかという情報と，発表ビデオにおいて，その単語の発言回数が増加しているか減少しているかという情報である．

3.2.2　目次スライドの判定

プレゼンテーションスライドに，目次スライドが含まれている場合，それを構成抽出に利用できる．目次スライドとは，例えば，図 4 のようなものである．

図 4 のような目次スライドがあった場合に，後のスライドには，“Background” や，“Related Work” というタイトルのスライドが続くことが多い．

目次スライドの判定により，プレゼンテーションスライドに目次スライドが含まれているかどうかを判定する．プレゼンテーションスライドは，目次スライドの出現パターンから，以下の 3 種類に分類できる．

m1.　**複数目次**

プレゼンテーションスライドの中に，目次スライドが複数含まれている場合である．一つの話題が終わる度に目次スライドをはさみ，今発表している箇所を把握しやすいようにしている場

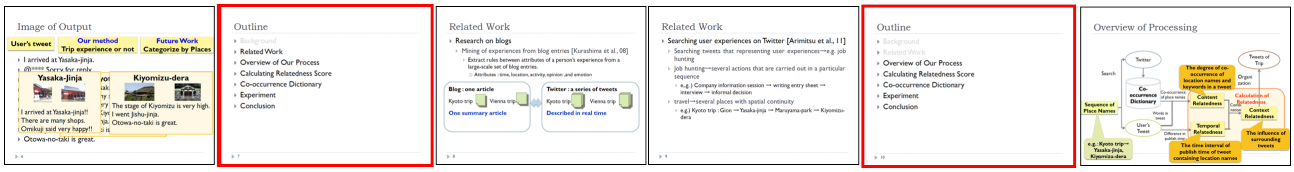

図 5　目次スライドが複数ある例

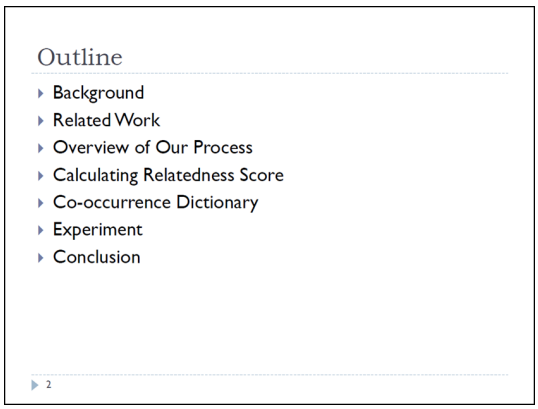

図 4　目次スライドの例

合である．具体例を図 5 に示す．枠で囲ってあるのが目次スライドである．図 5 では，目次スライドの後に，関連研究についてのスライドが 2 枚あり，その次にまた目次スライドを挟んでいる．このように，1 つの話題が終わって目次スライドが出現する，次の話題が終わって目次スライドが出現し，さらに次の話題に... となるようなものがこれにあたる．また，文字の色を変えて，現在の発表の位置を分かりやすく示しているものもある．

m2.　**単一目次**

プレゼンテーションスライドの序盤に 1 枚だけ目次スライドが含まれる場合である．

m3.　**目次なし**

プレゼンテーションスライドに目次スライドが含まれていない場合である．

目次スライドの判定に関しては，スライドのタイトルで判定する方法と，単語の出現パターンから行う方法が考えられる．

タイトルで判定する方法は，“Outline” といった，目次スライドのタイトルとなりやすいような単語を設定しておき，そのタイトルがついたスライドを目次スライドと判定する方法である．しかし，これでは，予め想定して居なかったタイトルを持つ目次スライドには対応できない．そのような場合，単語の出現パターンから目次スライドを判定することが考えられる．

目次スライドの本文には，他のスライドのタイトルが箇条書きで出現していることが多い．このような場合，目次スライドから，その他の複数のスライドに上位化されている単語が多いことになる．よって，他のスライドの上位化されている単語が閾値以上である場合，そのスライドが目次スライドであると判定することができる．

このように目次スライドを判定することで，プレゼンテーションスライドを，上記の m1, m2, m3 に分類することができる．

3.2.3　セグメントの抽出

一つのプレゼンテーションスライドの中では，いくつかの話題が存在する．例えば，研究発表のプレゼンテーションスライドであれば，研究背景があり，関連研究があり，提案手法があり，まとめがある．ある共通の話題について述べているスライド 1 枚以上の集合をセグメントとし，セグメントを抽出する方法を考えた．

セグメント抽出の際は，スライドの連続性を考慮する．隣り合っていない，離れたスライドが，他のセグメントを飛び越えて一つのセグメントにはならないようにした．また，セグメントは階層のような構造を持たず，セグメントを含むセグメント等は定義しないこととした．

セグメントの抽出は，複数目次のプレゼンテーションスライドと，それ以外の単一目次，目次無しのプレゼンテーションスライドでは異なった手法をとる．

複数目次のスライドに関しては，目次スライドをセグメントの切れ目と判定する．目次スライドが複数ある場合は，3.2.2 節で述べたように，一つの話題が終わるごとに目次スライドをはさむのが普通である．よって，目次スライドから目次スライドまでが一つのセグメントであると判定した．

単一目次と，目次無しのプレゼンテーションスライドの場合は，複数目次のような明確な話題の切れ目が無い．この 2 種類に関しては，単語の出現パターンを利用してセグメントを求める．

最初は，スライド 1 枚を 1 つのセグメントととする．これが最小単位のセグメントである．その後，以下の手法に従って，セグメントの抽出を行う．

1.　隣り合っているスライドで，タイトルが完全一致するものは，同じセグメントにする．タイトルが同じであるため，同じ話題について述べている可能性が非常に高い．

2.　隣り合っているスライドで，タイトルに出現する共通の単語が閾値以上である場合，同じセグメントにする．

これに当てはまるものとして，例えば，図 6 のようなものがある．図 6 では，“Calculating Relatedness Score (1/3)”，“Calculating Relatedness Score (2/3)” といったタイトルのスライドがある．このように共通の単語と，数字となっているようなスライドは，タイトルは完全一致していないが，同じ話題について述べていると考えられる．

3.　まだ他のセグメントとまとめられておらず，スライド 1 枚でセグメントとなっているようなセグメントに注目する．

3-a.　注目したセグメントと，隣接するセグメントを比較して，インデントの変化していない共通の単語の数を求める．対象となるセグメントに複数のスライドが含まれる場合は，スラ

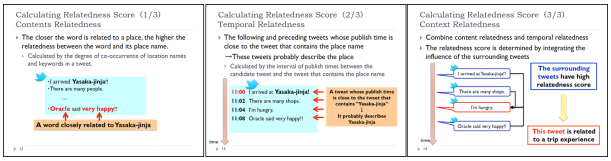

図 6　タイトルに共通の単語が出現する例

イド枚数で割って 1 枚あたりの平均を取る．

3.b.　求めた単語の数 / 注目しているスライドの単語の数を計算し，スコアとする．記述量が少ないスライドほど，1 語当たりの重みを重くするために，注目しているスライドに出現している単語の数で割る

3-c.　まだまとめられていないセグメント全てに対してこの計算を行う．

4.　スコアが一番多かった組み合わせを見る

5.　スコアが閾値以上であれば，2 つのセグメントを 1 つにまとめる

6.　閾値以上のものが存在しなくなるか，スライド 1 枚でセグメントとなっているものがなくなるまで，3〜5 を続ける．

7.　目次スライドについては，このスコアが高くなりがちであるため，他のセグメントとまとまりやすい，しかし，目次スライドはそれ単独で一つのセグメントとするのが正しいと考え，3.2.2 節の手法で目次スライドと判断されたスライドの関しては，単独でセグメントとなるようにした．

以上の手法で，予備実験を行った結果，目次スライドや，結論のスライドでは，スライド全体の内容に言及しいるため，どのスライドとも多くの共通語が出現しており，予期しないセグメントにまとめられてしまうことが多かった．そこで，3.2.2 節で目次スライドと判定されたスライドや結論を述べていると思われるスライドは，それ 1 枚で 1 つのセグメントであるとし，他のセグメントと結合されないようにした．

4.　実　　験

実験に使うデータとしては，VLDB2011 のプレゼンテーションスライドのうち，ウェブ上(注1)から，Microsoft PowerPoint の.pptx ファイルが入手できるもの 46 ファイルを利用した．これらのスライドの言語は英語となる．

4.1　スライドのタイトルに関する予備実験

目次スライドの判別には，スライドのタイトルを利用するため，スライドのタイトルについての予備実験を行った．

スライドのタイトルにどのような単語が出現しているかを調べた．単語は，ステミングとストップワード処理をした後のものを利用した．表 1 はその結果のうち，出現数が 15 以上のものだけをまとめたものである．“outlin” 等，完全な単語の形になっていないものは，ステミング処理によるものである．

出現数については，例えば “outlin” であれば，タイトルに “outlin” という単語を含むスライドが，46 個のプレゼンテーションスライドのうち 24 個に出現したということである．

(注1)：http://www.vldb.org/2011/

表 1　タイトルに出現しやすい単語

出現数	単語
24	outlin
21	data
19	exampl, conclusio, wor
17	result
16	challang, problem, experiment
15	conclus, expero

表 1 に出現している単語には，“outlin” や，“conclusion” など，多くのプレゼンテーションスライドでタイトルになりやすい，一般的な単語である．今回の結果では，2 番目に出現が多い単語として，“data” が現れている．これは，VLBD がデータベースに関する発表を主としているため，出現数が多くなったと考えられる．より広い分野のプレゼンテーションスライドを集めることで，このような単語が上位に現れることはなくなると思われる．

この結果から，目次スライドに使われる単語で，出現数が高いものとして，“outlin” があることが分かった．また，今回対象としたプレゼンテーションスライド集合では，この単語以外に，目次スライドに使われ，かつ出現数が多い単語が見つからなかった．よって，タイトルによる目次スライドの判定は，“outlin” という単語のみを利用することにした．

4.2　セグメントに関する評価実験

単一目次や目次無しのプレゼンテーションスライドについて，本研究の手法で抽出されたセグメントが妥当であるか評価するための実験を行った．

詳細な手法は以下の通りである．

- 目次スライドは，タイトルに “outlin” が出現しているかどうかを判別の条件とした．

これは，4.1 節 の予備実験の結果から，タイトルに出現する特徴語による判別で，十分な目次スライドの判定が行えると判断したためである．

- タイトルに出現する共通の単語については，2 語以上一致していれば同じセグメントにするもとのした．
- 3.2.3 節 のセグメント抽出手法について，インデントの変化しない語によるセグメント抽出における閾値を，0.01 から 0.1 までの間で変化させて実験を行った．閾値が高いほど，セグメントは細かくなる．

この実験には，複数目次のプレゼンテーションスライドから，目次スライドを抜いたものをデータとして用いる．3.2.2 節で述べた通り，複数目次のスライドは，話題が変化する箇所に挟まっていることが普通である．そのため，目次スライドを抜いたものに今回の手法を適用し，本来目次スライドがあった箇所で正しく分割されているかどうかで，提案手法の評価をする．

4.1 節 で述べたように，“outlin” という単語を用いて目次スライドを判定した場合，目次スライドを含むプレゼンテーションスライドは 24 個あり，そのうち複数目次にあたるものは 22 個であった．また，この 22 個のスライドに出現している目次スライドは全部で 93 ページであった．

表 2　セグメント判定の精度

閾値	目次数	切れ目数	正解数	適合率	再現率
0.01	93	253	58	0.229	0.624
0.05	93	253	58	0.229	0.624
0.75	93	302	67	0.222	0.720
0.1	93	314	68	0.217	0.731

これらについて，閾値を変化させながら，適合率と再現率を計算した．目次数とは，プレゼンテーションスライドに本来含まれていた目次の総数である．切れ目数とは，提案手法でセグメントの切れ目と判断された箇所の数であり，正解数は，そのうち本来目次スライドが含まれていた箇所の数である．適合率は，提案手法でセグメントの切れ目と判断された箇所に対して，そこに本来目次スライドが挟まれていたものの割合である．再現率は，本来目次スライドが挟まれていた部分に対して，そこが提案手法で分割されていたものの割合である．

- 適合率 = 正解数 / 切れ目数
- 再現率 = 正解数 / 目次数

となる．

表 2 がその結果である．適合率は 2 割程度となっており，全体として，セグメントが細かく分割されすぎている傾向にある．原因としては，以下のことが考えられる．

目次スライド以外での分割点

例えば，目次スライドが挟まれていなくても，細かく見ると，話題が変わっていることがある．あるいは，話題の切れ目に必ずしも目次スライドが含まれているとは限らない．複数目次スライドが含まれている場合も，その目次スライド数はプレゼンテーションスライドによってまちまちであり，過剰に目次スライドが出現しているのを避けている場合もある．このようなスライドは，今回の実験では適合率が低くなる傾向にある．

発表内容と無関係なスライド

例えば，以下のようなスライドがこれにあたる．

- “Thank You!”, “Questions?” 等と書かれたスライド
- バックアップスライド

バックアップスライドは，発表後想定される質問に対して，あらかじめその解答となるようなスライドである．発表用のスライドの後ろにこのようなスライドが入っている場合がある．このようなスライドは，今回の手法では得に考慮しておらず，目次スライドも挟まれていないため，適合率が下がってしまう．

極端に文字が少ないスライド

同一の話題について話しているスライドの間に，極端に文字の少ないスライド，あるいは，画像のみが含まれており文字が含まれていないスライドが存在することがある．提案手法では，テキストの含まれていないスライドや，他のスライドと共通の単語が無いスライドに関しては，他のスライドとまとめられることが無いため，そのようなスライドが存在した場合，セグメントが小さく分割されすぎてしまう．

5. おわりに

本研究では，目次スライドを抽出し，それとスライド間の単語の出現パターンから，プレゼンテーションスライドをいくつかのセグメントに分割する方法を提案した．今後は，このセグメント単位で，単語の出現パターン等から，論文のような階層構造を求めたり，あるいはセグメント間の関係を求めることで，プレゼンテーションスライドの構成が抽出できると考えている．

今回提案した構成抽出手法の応用として，プレゼンテーションスライドの理解支援アプリケーションや，検索アプリケーションが考えられる．このようなシステムの構築方法も，今後考えていきたい．

5.1 応　用　例

今回の手法の応用例の一つとして，プレゼンテーションスライド検索システムが考えられる．

検索システムの実用化例として，SlideShare (注2) や， Speaker Deck (注3) がある．入力されたキーワードに対して，プレゼンテーションスライドのランキングをし，結果を提示するシステムである．しかし，このようなプレゼンテーションスライドの検索では，スライドの構造化等は行なっておらず，結果のプレゼンテーションスライドは，ただ並べられているものを見るという単純なものである．

我々の手法を，検索のランキングに応用する方法を考える．例えば，入力されたキーワードが，目次スライドに出現している場合は，そのプレゼンテーションスライドでは，キーワードについて詳しく扱っていると考えられるため，上位に表示する．あるいは，他にも，インデントの情報を利用してプレゼンテーションスライドのランキング手法も考えられる．

また，入力されたキーワードに対して，プレゼンテーションスライド全体が対応しているとは限らず，キーワードに関連するのはプレゼンテーションスライドの一部である場合もある．例えば，先ほどのように，目次スライドにキーワードが出現している場合，キーワードに関係するのは，プレゼンテーションスライド全体ではなく，その項目に当たる部分だけという事がある．このような場合，全体を提示するよりも，該当するセグメントのみを抽出して提示することで，ユーザはどこを見るべきか迷わなくてすむ．

5.2 今後の課題

今回の課題として，プレゼンテーションスライドは，スライド 1 枚あたりの記述量が少なく，それが原因で正しくセグメント分割ができない場合があることが分かった．以下に今後の課題をいくつか挙げる．

離れたセグメント間の関係抽出

今回はスライドの連続性を考えたセグメントの抽出を主に行ったが，離れたスライドやセグメントの関係も求めたい．離れたセグメントの関係を求めることは，スライドの理解支援に役に立つため，有用であると考えられる．

目次スライドが正確でない場合の対処

提案手法では，複数目次の場合，目次スライドの出現位置によってセグメント分割を行った．しかし，目次スライドが話題

（注2）：http://www.slideshare.net/
（注3）：https://speakerdeck.com/

の切れ目に挟まっていないこともある．このような場合は，目次スライドの出現位置によるセグメントだけでなく，目次の内容とスライドのタイトルの比較を行う必要がある．複数目次の場合だけでなく，単一目次の場合においても，目次スライドの本文を見ることは，構成抽出に有用であるとかんがえられる．

閾値の設定

今回は閾値を様々に変更させて実験を行ったが，実際には，入力されるプレゼンテーションスライドに応じて閾値を変える必要があると考えられる．入力されるプレゼンテーションスライドは，作者の傾向や，使用される場面に応じて，記述の量がまちまちである．閾値が一定の場合，記述量が多いものはセグメントが大きくなりやすく，記述量が少ないものはセグメントが小さくなりやすい傾向にあるため，入力されたプレゼンテーションスライドの記述量に応じた閾値の設定が重要である．

tf-idf による単語の重み付け

スライドに出現する単語には，そのプレゼンテーションスライド全体に出現しやすい単語や，一部のスライドにしか出現しやすい単語などがある．記述量が少ないスライドになると，1 単語あたりの重みは大きくなるため，あまり重要でない単語の一致により同一のセグメントと判断されてしまう場合がある．よって，スライド 1 枚を一つの文書として見て，tf-idf による重み付けをするといった手法が有効なのではないかと考えられる．

装飾やフォントによる単語の重み付け

if-idf による重み付けとは別に，その単語についている装飾やフォントの情報から重み付けをすることも考えられる．例えば，単語に下線がついていたり，太字になっている，あるいは，色が変わっていたり，フォントサイズが大きくなっているといった情報は，その単語が重要な単語であるということを表している．

類似語等の扱い

記述量の少ないプレゼンテーションスライドでは，単なる語の一致だけではなく，類義語を構成抽出に利用することが考えられる．あるいは，WordNet のようなものを利用し，上位語や下位語が出現していることを構成抽出に利用することも考えられる．

図の取り扱い

スライドの記述量が少なくなる要因として，図が考えられる．プレゼンテーションスライドは，直感的な理解を促すために，他の文書よりも図の割合が多くなっている．極端な場合，1 枚のスライドに図しかなく，文字が一切書かれていないこともある．そのため，プレゼンテーションスライドにおける図の取り扱いについての関連研究もいくつか存在する．提案手法では，図については考慮していなかったが，図が含まれているスライドの前後のスライドが，同じ話題について述べていると思われる場合は，図も含めて 1 つのセグメントとする，といった手法が挙げられる．

バックアップスライド等の取り扱い

提案手法や，今回行った実験では，バックアップスライドについて考慮していなかった．バックアップスライドの判定を行い，それを取り除いて構成抽出する等，構成抽出におけるバックアップスライドの扱いを考える必要がある．

謝　　辞

本研究の一部は，科研費 22700097 の助成を受けたものである．

文　　献

[1] 羽山徹彩, 難波英嗣, 國藤進: “プレゼンテーションスライド情報の構造化”, 電子情報通信学会技術研究報告, 2008, Vol. 70, pp. 45-50

[2] 羽山徹彩, 難波英嗣, 國藤進: “プレゼンテーションスライド情報の構造抽出”, 電子情報通信学会論文誌 D, Vol. J92-D(9), pp. 1483-1494, 2009-09

[3] 羽山徹彩, 國藤進: “プレゼンテーションスライド情報検索のためのスライドページからの要求関連情報抽出”, 日本情報処理学会研究報告, 2010, DD-76(2)

[4] 北山大輔, 大谷亜希子, 角谷和俊: “プレゼンテーションコンテンツのためのシーン意味的関係抽出とその応用”, 情報処理学会論文誌データベース, Vol.2 No.2, pp. 71-85, 2009-06

[5] 王元元, 北山大輔, 角谷和俊: “プレゼンテーションコンテンツのための概念的構造と表層的構造に基づくスライドの関係判定方式”, DEIM Forum, 2010 C1-1

[6] Yuanyuan Wang and Kazutoshi Sumiya: “Semantic Ranking of Lecture Slides based on Conceptual Relationship and Presentational Structure”, 1st Workshop on Recommender Systems for Technology Enhanced Learning, pp. 2801-2810, 2010-09

[7] Alessandro Vinciarelli and Jean-Marc Odobez: “Application of Information Retrieval Technologies to Presentation Slides”, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 8, NO. 5, pp. 981-995, 2006-10

[8] Thushari Atapattu, Katrina Falkner, Nickolas Falkner: “Automated Extraction of Semantic Concepts from Semi-structured Data: Supporting Computer-Based Education through the Analysis of Lecture Notes”, Databace and Expert Systems Applications, pp161-175, 2012,

[9] 田中清太郎, 手塚太郎, 青山敦, 木村文則, 前田亮: “図形の形状と配置に着目したスライド検索手法の提案”, DEIM Forum, 2012 E6-4

[10] 友安航太, 王元元, 角谷和俊: “ポスターとスライドの構造に基づくズーミングを用いたポスター閲覧方式”, 日本情報処理学会研究報告, 2012, DBS-155(1)

[11] Min-yen Kan: “SlideSeer: A digital library of aligned document and presentation pairs”, Joint Conference on Digital Libraries, pp. 81-90, 2007

[12] 山田卓也, 前野真輝, 渡邉豊英, 佐川雄二: “ユーザの意図を反映したシナリオに基づいたプレゼンテーション・スライド構成支援”, 情報処理学会研究報告, 1999, 99(85)