

論文とプレゼンテーションスライドの部分対応付けにおける粒度の検証

川上 優平[†] 清水 敏之^{††} 吉川 正俊^{††}

[†] 京都大学工学部情報学科 〒606-8501 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 京都市左京区吉田本町

E-mail: [†]kawakami@db.soc.i.kyoto-u.ac.jp, ^{††}{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし 近年、学術論文の発表に伴いプレゼンテーションスライドの作成が一般的になっている。ある研究を把握したい際、論文とプレゼンテーションスライドの両者を参照できる場合があるが、それぞれが一長一短の特徴を持っている。本研究では、この二つの媒体を組み合わせて用いることにより効率的に情報を得られるのではないかと考え、各スライドに対して論文の節、段落等の部分文書を対応付けする手法を考えた。特に、短時間で研究の概要を把握したい場合、スライド流し読み時に論文の対応する部分文書を補助的に参照できることは有用であると思われる。既存手法を基に、対応付けする部分文書の粒度を意識したアプローチを考察した。対応付け粒度として、論文の節単位および段落単位の両者を対応付け対象と考える場合と段落単位のみを対応付け対象と考える場合の二通りを考え、また、最初に節単位での絞り込みを行う二段階対応付けを考案した。実験において対応付け精度と実行時間を調査し、二段階対応付けの有効性が確認できた。

キーワード 対応付け, 論文, プレゼンテーションスライド

1. はじめに

1.1 研究背景

近年では、学会などで学術論文が発表される際、プレゼンテーションスライドも作成されることが多くなった。論文やプレゼンテーションスライドは、研究者のホームページや学会のページ等の Web サイトを通して参照できることも多い。また、検索システムを通して参照することも考えられ、論文は Google Scholar^(注1)、Microsoft Academic Search^(注2)、プレゼンテーションスライドは SlideShare^(注3) などを用いることができる。

論文とプレゼンテーションスライドに関して、それぞれの文章量が大きく異なるので、利用者は時間的な制約などの状況やそのテーマへの関心度に応じて、論文やプレゼンテーションスライドの利用方法を選択する。例えば、「手法や関連研究などの詳細な情報を知りたいので論文を参照する」、「研究の要点のみが理解できれば良いのでプレゼンテーションスライドを利用する」といった使い分けが考えられる。これは、それぞれの特徴の違いによって判断されていると考えられる。論文の特徴として、文章量が多く詳細な情報・説明が得られるが、図や表、イラストでの説明が少ないことも多いということが挙げられる。また、プレゼンテーションスライドは、図や表、イラストが多くイメージがしやすい反面、文章を簡潔にまとめて記述することが多いので論文と比べて圧倒的に文章量が少ないという特徴を持っている。論文とプレゼンテーションスライドそれぞれが一長一短の特徴を持っているので、この二つの媒体を組み合わせ

せて使うことによって、お互いを補助することができ、効率的に情報を得られると考えられる。

本研究では、自動的に論文とプレゼンテーションスライドの部分対応付けを行うことを目的とする。対応付けは、特に、論文を読むことに不慣れな人が短時間で内容を把握したい場合などで有用であると考えられる。対応付けにより、スライドを流し読みしながら対応する論文の部分文書を補助的に参照することで情報を補完することができ、理解を助けることができる。また、詳細を把握していない論文の紹介を聴講し、研究内容を把握したい場合など、時間をかけて論文を読むことができない際に、発表しているスライドが論文のどの部分を説明しているのかを参照しながら発表を聞くといった用途に用いることが出来る。補助的な利用法としては、論文を精読する際に論文の読んでいる部分に対応するスライドに図やアニメーションなどが記載されている場合に対応するスライドを参照しながら論文を読むことで理解を深めることができると考えられる。さらに、対応付け技術はプレゼンテーションスライド生成への応用も考えられ、論文発表用のスライドを作成する場合に論文と比べてスライドの記述内容が不足している部分を発見するツールとして用いたり、論文からのプレゼンテーションスライド自動生成においてテストデータの作成補助のためのツールとしても利用価値があると考えられる。

1.2 研究の概要

本研究では、先行研究である西田らの研究 [6] の手法を再検討し、論文とプレゼンテーションスライドの部分対応付けについて手法の改善を行った。さらに、対応付けを行う際の論文の部分文書の粒度に着目し、最適な粒度を決定するために 4 通りの手法を検証し、それぞれ部分対応付けの精度と速度を評価した。

本稿では、プレゼンテーション発表などを目的とした PowerPoint などで作成された 1 つの文書全体をプレゼンテーシ

(注1) : <http://scholar.google.co.jp/>

(注2) : <http://academic.research.microsoft.com/>

(注3) : <http://www.slideshare.net/>

ンスライドを呼び、その中のシート 1 枚 1 枚をそれぞれスライドと呼ぶことにする。

2. 関連研究

Kan の研究 [3] では、論文とそのペアとなるプレゼンテーションスライドを取得する手法と、それらの部分対応付け手法について提案している。この手法では、各スライドと各パラグラフの関連度の top- k の中から始点と終点を選び、全ての組合せで関連度を計算する。コサイン類似度や Jaccard 係数を用いていくつかの手法を検討しており、スコアが高かったパラグラフの組をそのままスライドに対応付けるというシンプルな手法が一番良い結果となっている。論文はパラグラフ粒度での対応付けを行っている。本研究では、パラグラフ単位だけでなくセクション単位での対応付けも考慮に入れることで、セクション全体を説明しているスライドの対応付けを容易にしているところに違いがある。

Bahrani らの研究 [1] では、プレゼンテーションスライドと論文の部分対応付けについて、スライド 1 枚に対して論文をセクション単位で対応付けを行う手法を提案している。tf-idf を用いてスライドと論文のセクション文書の文書類似度を計算し、更に論文とスライドの文書全体における出現位置を考慮して対応付けを行う。プレゼンテーションスライドを画像ファイルとして扱い、ヒストグラムを用いて線形サポートベクターマシンによる分類学習を行っている。スライドを Text, Outline, Drawing, Result に分類して、Result スライドは実験と実験結果のセクションに対応付けを行い、それ以外はそれぞれ文書類似度と文書順序の重みを変えて対応付けを行う手法を取っている。本研究では、論文の対応付け粒度についてセクション単位だけでなくパラグラフ単位でのより細かい粒度も含めて対応付けを行っているところに違いがある。また、文書関連度について tf-idf の XML 向けに改良された関連度計算式を採用することによって文書長による正規化を行っているところにも違いがある。

西田らの研究 [6] では、プレゼンテーションスライドと論文の部分対応付けについて、スライド 1 枚に対して論文のパラグラフ又はセクション粒度の部分文書集合と対応付ける手法を提案している。各スライドについて関連度が非常に高い論文の部分文書を発見して軸スライドと定義し、軸スライドを基準として論文の部分文書とスライド 1 枚 1 枚それぞれの内容関連度に順序によるペナルティを付与する。そして得られた関連度からスライドと対応する論文の部分文書集合を決定し、対応付けを行う。本研究では、この研究を先行研究として改善を行っていく。

羽山らの研究 [5] では、論文からプレゼンテーションスライドを自動生成することを目指して、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スライドの対応付けを行っている。これは Jing が提案した新聞記事を対象とした隠れマルコフモデルを用いた対応付け手法 [2] を論文とプレゼンテーションスライドに適応できるように改善したものである。スライド 1 枚ごとに論文の章または節に対応付ける手法であり、論文やプレゼンテーションスライドの順番の相関関係を考慮して対応付け

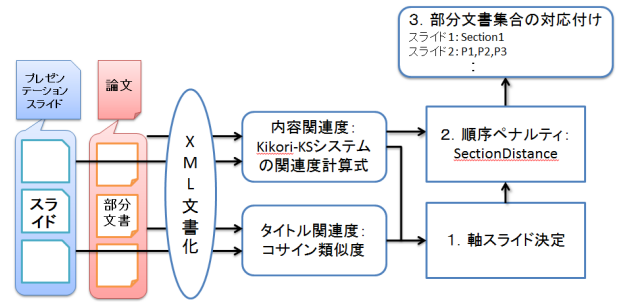


図 1 手法の全体の流れ

範囲を限定したり、手掛かり語集合をもとにしてタイトルによる対応付けを行ったりすることで、精度を向上させている。本研究ではパラグラフ粒度でも対応付けを行っているところの違いがある。

3. 対応付け手法

本研究では、西田らの研究 [6] を先行研究とし、この手法を改善して精度を向上させ、さらに、適切な部分文書の粒度の検証を目的とする。この手法ではスライド 1 枚につき、連続した論文の部分文書集合の対応付けを行う (図 1)。

3.1 事前準備

入力として扱う論文とプレゼンテーションスライドを XML 文書に変換する。先行研究では独自の XML 構造を定義して用いていたが、本研究では論文の XML 文書変換に関して、一般的な規格として DocBook^(注4) に則ったものを使用した。そのうち、論文などの単一の文書のために設計された Simplified DocBook^(注5) を採用した。これは、XML 文書を簡単化かつ一般化するように意識した結果である。プレゼンテーションスライドの XML 文書変換については、スライド 1 枚の単位 (<slide>), スライドのタイトル (<title>), スライドのボディ (<body>) のタグといった簡単なものだけあればよいものと考え、簡易的な独自のスキーマを設計した。

論文では複数の粒度で部分文書を扱い、セクション、サブセクションなどといったセクション粒度と 1 段落毎のパラグラフ粒度で区切られた文章を 1 つの部分文書とする。論文の部分文書の最小単位はパラグラフ粒度である。また、プレゼンテーションスライドは各スライドを 1 単位とする。1 枚のスライドに論文の連続した部分文書集合を対応付ける。これは先行研究も同様である。

3.2 関連度の計算

各スライド中のテキストをクエリとして、各スライドと論文の各部分文書の全ての組合せの関連度を計算する。論文の部分文書において、セクション粒度やパラグラフ粒度で範囲が重複する場合も文字数の違いなどを考慮するため全て計算する必要がある。関連度の計算には、先行研究では tf-idf において文書長を正規化する特徴を持つ OkapiBM25 [4] を用いていたが、本研究では関係データベースを用いた XML 情報検索システムで

(注4) : <http://www.docbook.org/>

(注5) : <http://docbook.org/schemas/5x-custom#simple>

ある Kikori-KS システム [7] で用いられた tf-idf の XML 向け改良版の関連度計算式を用いた。XML 文書の要素を根ノードからの経路ごとにカテゴリ化して ipf(inverse path frequency)として idf の代わりに用いている。また、根ノードからの経路が同じ要素の平均要素長を用いて文書長についての正規化を施している。また、関連度の計算については文字は小文字で統一して扱い、ステミング処理を施し、ストップワードを取り除いている。

3.3 軸スライドの決定

あるスライドが、ある論文の部分文書と対応関係にあることについて確信度が高いとき、そのスライドを軸スライドとして決定する。論文発表用に作られたプレゼンテーションスライドは、論文を基にして作成されることが多いため、その内容が論文に準拠した順序で記述されている可能性が高いと考えられる。したがって、軸スライドに基づいて順序によるペナルティを与えることにより、対応付けの精度の向上が期待できる。軸スライドの決定には、タイトル関連度と内容関連度を用いる。以降、プレゼンテーションスライド中で m 番目のスライドを s_m 、論文中で n 番目に出現した部分文書を d_n と表記することとする。

先行研究では、軸スライドは自身が対応しているタイトル関連度の高いセクションまたは文書関連度の高い部分文書を含むセクションと対応関係にあるとして先に対応付けを決定していた。これは、軸スライドの決定手法の特性上、軸スライドはセクション単位のみと対応関係を調査することが原因である。しかし、確信度の高い軸スライドだからといって必ずセクション単位での対応付けを行うのは不適当であると考え、本研究では軸スライドと対応付けは切り離して考えるものとする。つまり、軸スライドとしてあるスライドがあるセクションと対応関係にあると決定した後、そのセクション中のどの部分文書集合と対応付けるのが適当であるかを再度検討する手法を採用する。また、軸スライド中に順序を逸脱するものが存在しないように、軸スライドも確信度の高いものから矛盾しないように選択して採用するように改善した。

3.3.1 タイトル関連度

スライドのタイトルには“INTRODUCTION”や“CONCLUSION”など頻繁に出現する語がある。これらは論文のセクションタイトルで使われることもあり、スライドのタイトルと論文のセクションタイトルが類似している場合はそのスライドはそのセクションに対応している可能性が高いと考えられる。具体的には、各スライドのスライドタイトルと論文の各セクションのセクションタイトルのテキストから得られたそれぞれの語の集合のコサイン類似度を計算する。あるスライド s_m のスライドタイトルの語の集合を x とし、ある部分文書 d_n のセクションタイトルの語の集合を y としたとき、

$$\cos(x, y) > \alpha \quad (1)$$

上式を満たす場合、スライド s_m を部分文書 d_n が属するセクションと対応関係にある軸スライドとして決定する。ここで、 $\cos$ はコサイン類似度を計算する関数である。 α は閾値であり、 $0 \leq \alpha \leq 1$ を満たすものとする。この手法は先行研究と同様で

ある。また、関連度の計算と同様にタイトルの語集合は小文字で統一して扱い、ステミング処理を施し、ストップワードを取り除いてコサイン類似度を計算している。

3.3.2 内容関連度

次に内容関連度による軸スライドの決定法を述べる。3.2 節で求めた関連度は、スライド内に特徴的な語が多く使われる程、局所的に大きくなる。これを利用して、一番関連度の高い部分文書と次に関連度の高い部分文書の関連度の差が大きければ大きい程、スライドは前者の部分文書と対応関係にあることが確信的であると考えられる。スライド s_m と部分文書 d_n の関連度を $S(s_m, d_n)$ と表記することとする。

$$S(s_k, d_i) - S(s_k, d_j) > \frac{\text{MaxSimAvg}}{\beta} \quad (2)$$

ただし、 d_i は s_k と関連度が一番高い部分文書であり、 d_j は d_i と兄弟関係にも親子関係にもない部分文書の中で一番関連度の高い部分文書であるとする。ここで、MaxSimAvg は各スライドと各部分文書の関連度において最大のものの平均である。 β はパラメータであり、 $\beta > 0$ を満たすものとする。上式を満たす場合、ある特定のセクションとの対応関係が特に強いであるとして、スライド s_k を部分文書 d_i のセクションと対応関係にある軸スライドとして決定する。この手法は先行研究と同様であり、先行研究ではパラメータ $\beta = 2.60$ としている。

3.4 順序ペナルティの付与

3.3 節の手法で決定された軸スライドを用いてそれぞれ関連度にペナルティを与える。論文発表用に作成されたプレゼンテーションスライドは、論文に準拠した流れになっていることが多い。したがって、軸スライドを基にした文書の順序の流れから逸脱したスライドと論文の部分文書の組について関連度にペナルティを与える処理を行う。ここで、先行研究では軸スライドは対応付けが既に決定しているため軸スライドでないスライドの関連度のみにペナルティを与えていたが、本手法では全てのスライドの関連度にペナルティを与える。具体的には、2枚の軸スライド s_i, s_j と軸スライドでないスライド s_x が存在し、スライドの順番は s_i, s_x, s_j であるとする。軸スライド s_i はセクション A、 s_j はセクション B とそれぞれ対応付いているとすると、 s_x はセクション A とセクション B の間のセクションに属する部分文書と対応付く可能性が高いと考えられる。したがって、セクション A より前のセクションやセクション B より後ろのセクションに含まれる部分文書とスライド s_x との関連度にペナルティを与える (図 2)。

ペナルティを付与するための具体的な式としては、先行研究では式 (3) を用いていた。軸スライド以外のスライドとそれぞれの部分文書との関連度 $S(s_x, d_k)$ に対して、以下の式をかけることでペナルティを与える。

$$\{1 - \gamma \cdot \log(1 + \text{SectionDistance})\} \quad (3)$$

ここで、SectionDistance は以下の式で求める。

$$\text{SectionDistance} = \max((\text{sec}_{d_i} - \text{sec}_{d_k}), (\text{sec}_{d_k} - \text{sec}_{d_j}))$$

ただし、軸スライド s_i, s_j と対応関係にあるセクションに相

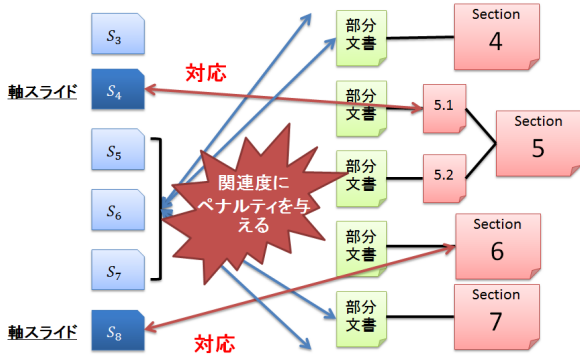


図 2 順序ペナルティの付与

当する論文の部分文書をそれぞれ d_i, d_j とし、部分文書 d_k を含む論文のセクション番号を sec_{d_k} と表すこととする。 γ はペナルティの大きさを決定するパラメータであり、 $0 \leq \gamma \leq 1$ を満たすものとする。先行研究では $\gamma = 0.3$ としている。また、 $SectionDistance < 0$ ならば $SectionDistance = 0$ とする。

本研究では、ペナルティ式が負の値を取り得ることの無いように先行研究の式に正規化を施した。本研究のペナルティとしては以下の式 (4) を採用する。

$$\left\{ 1 - \gamma \cdot \frac{\log(1 + SectionDistance)}{\log(1 + N)} \right\} \quad (4)$$

ただし、 N は論文中のセクション数とする。以上の式と文書関連度の積を順序ペナルティを考慮した最終的な文書関連度として決定する。

3.5 対応部分文書集合の決定

最終的な対応関係を取得する手順について述べる。本研究では、各スライドに対して論文の連続した部分文書集合の対応付けを行う。そこで、各スライドと各粒度の部分文書の関連度にペナルティを与えた最終的な関連度のうち、各スライドについて $top-k$ について調べる。まず、スライド毎に最もスコアの高い部分文書を対応付ける。次に、スコアの高い部分文書から順に対応付け部分文書集合に含めるか判定していく。ここで、その部分文書が既に対応付けされている部分文書集合と、連続した兄弟関係であるか親子関係である場合、対応付け部分文書集合に追加する。既に対応しているどの部分文書とも連続した兄弟関係、もしくは親子関係でない部分文書であった場合、対応付けは行わずそのスライドに対する対応付けの操作を終了する。この作業を、対応付けされない部分文書が出てくるまで、もしくは $top-k$ 全てを操作し終わるまで繰り返し行い、その時点までに対応付けされている部分文書集合を最終的な対応付けとして決定する。図 3 の例では、まずスライドに $top-1$ である P9 を対応付ける。ここでの P はパラグラフの意である。次に $top-2$ である P10 を対応付け部分文書集合に加えるかどうか判定するが、P9 と P10 は連続した兄弟関係にあるので、P10 を対応付けに加える。同様に、Subsection2.1 は P9, P10 と親子関係にあるのでこれも対応付けに加える。 $top-4$ である P2 は Subsection2.1, P9, P10 と連続した兄弟関係にも親子関係にも属さないのに対応付けには加えず、スライドに対する対応付

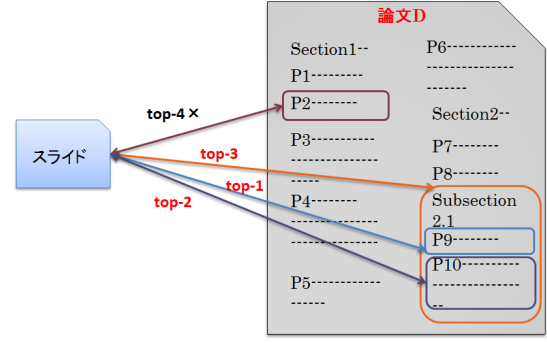


図 3 対応部分集合決定の流れ

けを完了する。

ここで、先行研究では $top-k$ は全て $k = 10$ に固定されて実験が行われていたが、本研究では、各論文とプレゼンテーションスライドの組によって k の値を吟味すべきであると考えて、適切な k の値を検討することにした。具体的には、各スライドについて

$$S(s_m, d_n) > \frac{MaxSimAvg}{\delta} \quad (5)$$

を満たすものを対応付けの候補として、その文書数を k とする。ここで、 $MaxSimAvg$ は各スライドと各部分文書の関連度において最大のものの平均である。 δ はパラメータであり、 $\delta > 0$ を満たすものとする。各論文スライドペア毎に関連度の大きさは様々であり、対応付けるべき部分文書の関連度の最低値は異なっていると考えられる。各論文スライドペアの関連度の最大値の平均値を考慮することで、それぞれに適した $top-k$ を候補として選択できると考えられる。また、“Outline”、“Question”、“Thank you” などスライドの流れを表すスライドなどは、論文との対応関係がないものと考えられる。このようなスライドは式 (5) によって足切を行い $k = 0$ として対応付けを行わないようにするため、パラメータ δ の値を適切に設定する必要がある。

4. 粒度を意識した対応付け

4.1 論文における複数の部分文書粒度

論文とプレゼンテーションスライドの部分対応付けにおいて、論文の部分文書の粒度について、どのように扱うのが最良であるかについて検討を行う。ここでは、部分対応付けの実行結果の精度と実行時間について着目する。部分対応付けに関して、本研究では、論文の部分文書としてパラグラフ粒度の部分文書とセクション粒度の部分文書があると考えた。各スライドについて、パラグラフ粒度のみの連続した部分文書の対応付けと、パラグラフ及びセクション粒度の連続した部分文書の対応付けの、2種類の粒度での対応付けを検討する。パラグラフ粒度のみの対応付けでは、セクション単位の部分文書を考慮しないことでそれぞれの部分文書の重複がなくなり、先行研究より計算量が小さくなると考えられる。各スライドの対応部分文書集合は、親子関係はなくなり、連続した兄弟関係のみで構成される。パラグラフ粒度とセクション粒度の両方を含む部分文書の対応付けは先行研究の手法と同様の粒度での対応付けであり、パラ

グラフ粒度のみでの対応付けを行うよりも1枚のスライドに対応する部分文書集合の範囲が広がることが考えられる。

4.2 二段階対応付け手法

また、対応付けアプローチとして、まず各スライドに対してセクション単位で対応付けを行った後、上述の更に細かい粒度での対応付けを行う、二段階の対応付けを行う手法を提案する。二段階対応付け手法では、まず論文の部分文書はセクション単位のみで対応付けを行う。つまり、それぞれのスライドが、あるセクション全体と対応する部分文書と対応付けられる。そこから、そのセクションに含まれるパラグラフ粒度の部分文書、もしくはパラグラフ粒度とセクション粒度の部分文書との対応関係を再度検討し直す第2段階目の対応付け作業を加える。具体的には、第1段階目で対応付けられたセクション中のパラグラフ粒度、もしくはパラグラフ粒度とセクション粒度の各部分文書と各スライドとの関連度を計算し、それぞれのスライドと対応する部分文書集合を決定する。

二段階対応付けでは、予めセクション単位で範囲を絞りこむことで、更に細かい粒度の部分文書との関連度計算が省力化でき、二段階対応付けを行わない手法よりも計算量が少なくなると考えられる。また操作するデータ数も少なくなり、先行研究よりもプログラムの実行時間が短くなると期待できる。さらに、二段階での対応付けによって、どのように対応付け精度に影響が出てくるのかを調査する。細かい粒度ではなくセクション単位のみで最初に対応付けられることにより、二段階対応付けを行わない手法とは違ったセクションと対応付けられるスライドが出てくることが考えられる。対応付け粒度として、パラグラフのみを対象とする場合とパラグラフ及びセクションの両者を対象とする場合の二種類が考えられ、対応付けアプローチとして、二段階対応付けを行う場合と行わない場合の二種類が考えられる。これらの組合せの以下の4パターンに対して実験を行い、精度と速度の検証を行う。

- 手法 P1：パラグラフ粒度・二段階非適用
- 手法 P2：パラグラフ粒度・二段階適用
- 手法 SP1：セクション及びパラグラフ粒度・二段階非適用
- 手法 SP2：セクション及びパラグラフ粒度・二段階適用

5. 実験

5.1 パラメータの設定

先行研究の手法を改良しているため、タイトル関連度による軸スライドを決定する際のパラメータ α 、内容関連度による軸スライドを決定する際のパラメータ β 、順序ペナルティを付与する際のパラメータ γ を再検討するために実験を行った。

この実験を行うにあたり、手で論文とスライドの部分対応関係の正解データを作成した5つの論文スライドペアを用いた。これらをそれぞれ順にペア1～ペア5とする。各ペアの特徴は以下の通りである。ペア1はスライドの文章量が多く、文書間の関連度から正しい対応関係を得やすいペアであると言える。ただし、スライドが論文の順序通りでなく、5章、6章、4章の順に作成されているという特徴を持つ。また、論文のセクション内のパラグラフが細かく区切られており、パラグラフ

表1 α による軸スライド数の変化

論文・スライド	α	軸スライド数	正解軸スライド数
ペア1 スライド数 35枚	0.70	7	6
	0.75	5	5
	0.80	2	2
ペア3 スライド数 28枚	0.70	11	9
	0.75	7	7
	0.80	3	3
ペア5 スライド数 51枚	0.70	6	3
	0.75	2	2
	0.80	1	1

数が多くなっている。ペア2はスライドの文章量が多く、流れの説明をするタイトルスライドが途中にいくつか存在するので、内容関連度で正しい対応付けを行いやすいだけでなく、タイトル関連度による軸スライドも得やすいと思われる。スライドも論文の順序に沿って作成されている。ペア3は図だけで作られたスライドが多く、文章がほとんど書かれていないスライドが多い。また、スライドの順序も論文の順序に準拠していないので正しい対応付けが難しいと考えられる。ペア4はペア2に傾向が似ており、スライドの文章量が多く、順序も論文に準拠している。ペア5はどのスライドにも図や表が存在し、平均的な文章量が少ない。また、順序を逸脱して1章と対応するスライドがまんべんなく全体に10枚以上存在するので、順序ペナルティによる効果が得にくいと考えられる。また、ペア1と同様にパラグラフ数が多い。

5.1.1 タイトル関連度による軸スライド決定のパラメータ

3.3.1節で述べた式(1)における、スライドのタイトルと論文のセクション等のタイトルの類似度によって軸スライドを決定する際のしきい値 α の決定のための実験を行った。しきい値は小さい方が多くのスライドを軸スライドとして決定するが、それぞれのタイトルの語数が少ない場合コサイン類似度は高い値になることがあるので、しきい値 α は適切な値に設定する必要がある。

各論文スライドペアに対して閾値 α を変化させて得られた軸スライドの数と、その中で人手で判定した軸スライドに一致した数(正解軸スライド数)を計測したところ、表1の結果が得られた。しきい値 α は0.70～1.0の間を0.50ずつ変化させて結果を観察した。ここでの“正解軸スライド”とは、軸スライドとして対応しているセクションが、そのスライドが実際に対応している正解の部分文書集合を含むような軸スライドのことを指す。なお、 $\alpha = 0.85$ 以上については得られた軸スライドは全て正解軸スライドであった。

軸スライドは全てのスライドと部分文書の関連度とペナルティを与える指標となるものであり、誤った軸スライドは許容すべきではない。したがって、軸スライドの中に正解でない軸スライドが含まれないように精度重視で閾値を決定する必要がある。表1の実験結果について考察すると、ペア1、ペア3、ペア5については、 $\alpha = 0.70$ のときに決定された軸スライドの中に1～3枚の正解軸スライドでない軸スライドが存在する。

表 2 β による軸スライド数の変化

論文・スライド	β	軸スライド数	正解軸スライド数
ペア 1 スライド数 35 枚	2.50	0	0
	2.75	2	1
	3.00	3	2
ペア 3 スライド数 28 枚	2.50	3	3
	2.75	3	3
	3.00	4	3
ペア 5 スライド数 51 枚	2.50	6	6
	2.75	10	9
	3.00	11	10

しかし、 $\alpha = 0.75$ 以上のときは軸スライドは全て正解軸スライドである。ペア 2、ペア 4 については、 $\alpha = 0.70$ 以上で決定された軸スライドは全て正解軸スライドであった。また、全てのペアについて、しきい値 α が大きくなればなるほど軸スライドの数は減少していく事が分かる。

以上の結果より、タイトル関連度による軸スライド決定のためのパラメータ α は、 $\alpha = 0.75$ のとき、決定される軸スライドのうちに軸スライドに適さないものを含まず、かつ軸スライドを最大数得ることができると分かる。したがって、今後の実験では $\alpha = 0.75$ とする。

5.1.2 内容関連度による軸スライド決定のパラメータ

3.3.2 節で述べた式 (2) における、スライドと論文の部分文書の内容関連度によって軸スライドを決定する際のパラメータ β の決定のための実験を行った。各スライドについて、一番関連度の高い部分文書を含むセクションが、他のセクションよりも確かにそのスライドと対応付く可能性が高いと判断されたものを軸スライドと決定する。

5.1.1 節と同様に、各論文スライドペアに対して閾値 β を変化させて得られた軸スライドの数と、その中で人手で判定した軸スライドに一致した数 (正解軸スライド数) を計測したところ、表 2 の結果が得られた。 β は 2.5~4.5 の間で 0.25 ずつ変化させて結果を観察した。ただし、 $\beta = 3.25$ 以上の結果は割愛している。

表 2 の結果について考察すると、ペア 1 とペア 5 については β が 2.75 以上の際に誤った軸スライドと決定してしまい、ペア 3 については β が 3.00 以上で誤った軸スライドを決定してしまった。ペア 2 とペア 4 については、 β が 3.0 以下の時に誤った軸スライドを決定することはなかった。 β は式 (2) において、しきい値の分母にあたるパラメータであるので、値が増加するにしたがってそれぞれのペアの軸スライド数は増加していく。

以上の結果より、スライドと論文の部分文書の内容関連度によって軸スライドを決定する際のパラメータ β は、 $\beta = 2.50$ のとき、決定される軸スライドのうちに誤った軸スライドを含まず、かつ軸スライドを最大数決定することができたので、今後の実験では $\beta = 2.50$ とする。

5.1.3 順序ペナルティのパラメータ

3.4 節で述べた式 (4) における、スライドと部分文書の内容関連度に順序によるペナルティを与える際のパラメータ γ の決

表 3 γ による各スライドの最大関連度部分集合の精度の変化

論文・スライド	γ	部分適合スライド	部分適合率	範囲適合スライド	範囲適合率
ペア 1 スライド数 35 枚	0.00	9	0.257	12	0.343
	0.25	11	0.314	14	0.400
	0.50	13	0.371	17	0.486
	0.75	13	0.371	17	0.486
	1.00	13	0.371	17	0.486
ペア 2 スライド数 50 枚	0.00	29	0.580	30	0.600
	0.25	29	0.580	30	0.600
	0.50	32	0.640	33	0.660
	0.75	34	0.680	36	0.720
	1.00	34	0.680	36	0.720
ペア 3 スライド数 28 枚	0.00	7	0.250	8	0.286
	0.25	8	0.286	9	0.321
	0.50	10	0.357	13	0.464
	0.75	12	0.429	15	0.536
	1.00	12	0.429	15	0.536

定のための実験を行った。順序を逸脱したスライドと部分文書の関連度について、上記の式 (4) によるペナルティをかけて、その積を新しい関連度として定義する。 γ はペナルティの影響力を決定するパラメータで、 γ を増加させると順序によるペナルティの強さが大きくなっていく。 $\gamma = 0$ のとき、順序によるペナルティを考えないことになる。

各論文スライドペアに対してパラメータ γ を変化させて部分適合スライド数と範囲適合スライド数にどのような影響を与えるかを調べたところ、表 3 の結果が得られた。 γ は 0.00~1.00 の間で 0.25 ずつ変化させて結果を観察した。ここで、各スライドについて関連度が一番高い部分文書が実際の正解の対応関係にある部分文書集合の一部と重なるようなスライドを部分適合スライドと定義し、部分適合スライドの割合を部分適合率とする。また、同じく各スライドについて関連度が一番高い部分文書が実際の正解の対応関係にある部分文書集合の一部と重なっていても、同一セクションに存在するようなスライドを範囲適合スライドと呼び、範囲適合スライドの割合を範囲適合率とする。

表 3 の結果について考察すると、パラメータ γ を大きくしていくと、部分適合率・範囲適合率ともに増加し、精度が改善されていく傾向が見られた。特に、 $\gamma = 0.25$ から $\gamma = 0.50$ に変化させたときに、全てのペアにおいて部分適合スライドが 2~3 枚、範囲適合スライドが 3~4 枚増加しており、順序ペナルティによる大きな改善が見られる。しかし、ペア 1 では $\gamma = 0.5$ 以上、ペア 2 とペア 3 では $\gamma = 0.75$ 以上において、部分適合スライド数と範囲適合スライド数に変化は見られなかった。

ペア 1 において $\gamma = 0.00$ から $\gamma = 0.25$ に変化させた際、順序を逸脱しているスライドと部分文書の対応関係を元々正しく対応付けていたスライドに、ペナルティを考慮することによって誤った対応付けを行ってしまっていた。このような問題はこの 1 件のみであったが、順序を逸脱していると思われるスライドに関しては、これとは別のペナルティを与えるような工夫が必要であると考えられる。

表 4 各手法による対応付けの精度の平均

手法	適合率	再現率	F 値
手法 P1	0.612	0.314	0.415
手法 P2	0.658	0.360	0.465
手法 SP1	0.282	0.680	0.399
手法 SP2	0.270	0.765	0.399

順序を逸脱するスライドも正しく対応付けるためには出来るだけ γ の値は小さい値であることが望ましいので、以上の結果より、スライドと部分文書の内容関連度に関数によるペナルティを与える際のパラメータ γ は $\gamma = 0.75$ として以後の実験を行う。

5.2 各粒度の性能比較

本節では、4つの手法により得られた論文とプレゼンテーションスライドの部分対応付けについて、その性能を対応付け精度と実行時間について比較する。

5.2.1 精度比較

まず、4つの手法について対応付け精度の観点から比較した。その際に、適合率、再現率を尺度として用いた。適合率と再現率を計算する際、セクションとパラグラフという複数の粒度が混在していてそのままでは正しい評価を算出することができないため、全てパラグラフ粒度に置き換えて評価を行った。つまり、セクション粒度で正しく対応付けを行うことができた場合、そのセクションに含まれるパラグラフ数だけ対応付けが的中したということになる。

実験の結果、5つの論文スライドペアの適合率と再現率と F 値の平均として表 4 が得られた。全体的な傾向として、パラグラフ粒度で対応付けを行った手法 P1 と手法 P2 は適合率が高く再現率が低い結果となり、セクション粒度とパラグラフ粒度を混在して対応付けを行った手法 SP1 と手法 SP2 は適合率が低く再現率が高い結果となった。パラグラフ粒度で対応付けを行った場合、3.5 節の対応付け部分集合決定アルゴリズムに従うと、多くのスライドに関して対応付け範囲が 1 ~ 3 個の部分文書しか得られなかったため、十分な広さの範囲で対応付けが行うことができなかったことが再現率の低さの原因となっていると考えられる。しかし、その関連度の高い数個の部分文書は手法 P2 で平均 0.685 の適合率を得られた。一方、セクション・パラグラフ粒度で対応付けを行った場合、多くのスライドでセクション単位での対応付けが行われ、広い範囲と対応付く結果となり、冗長な対応付けが多く見られた。その結果、適合率が低くなった一方、手法 SP2 で平均 0.765 の再現率を得られた。この結果を考察すると、手法 P1 と手法 P2 では、あるスライドが論文のどの部分文書と最低限対応しているかを高い精度で検出することができ、その部分文書の付近を読めばスライドの内容を詳しく説明している記述を発見できると考えられる。また手法 SP1 と手法 SP2 では、あるスライドがセクション単位でどのセクションと対応関係にあるかを発見することに長けている。

二段階の対応付けを行うことで、全てのペアで手法 P1 から手法 P2 にかけて適合率・再現率ともに精度の改善が見られた。

表 5 スライド単位での部分正解の精度

手法	ペア 1	ペア 2	ペア 3	ペア 4	ペア 5	平均
手法 P1	0.543	0.780	0.429	0.684	0.529	0.593
手法 P2	0.600	0.840	0.500	0.842	0.588	0.674

これは、手法 P2 でまずセクション単位で対応付けを行ってから第二段階目の対応付け集合の決定を行うとき、そのセクションのみの部分文書の関連度から top- k を操作するため、連続した兄弟関係の部分集合を集めることが容易になり対応付け範囲を広げることが出来た結果であると考えられる。また、その広がった対応付け範囲が正解の対応付けであったことが分かる。平均で適合率を 0.046、再現率を 0.046 向上させることができた。特に、ペア 4 における適合率向上が顕著であり、0.13 向上させ適合率 0.857 が得られた。

手法 SP1 から手法 SP2 にかけて、全てのペアで再現率が向上した。また、適合率はペア 1・ペア 2・ペア 5 では低下し、ペア 3 とペア 4 では向上した。手法 P2 と同様に手法 SP2 も二段階の対応付けにより対応付け範囲を拡張したスライドが多く存在し、手法 SP1 であるパラグラフと対応付けられていたスライドが、手法 SP2 でセクション単位で対応付けられることがあった。そのようなスライドは適合率の分母が非常に大きくなり、適合率が低下してしまっただと考えられる。また、対応付け範囲が拡張したため、正解の対応付け範囲をより多く含むスライドが増加し、再現率は大きく向上した。平均で適合率が 0.012 減少し、再現率が 0.085 増加した。適合率が低くなってしまったが、この手法は適合率より再現率に比重を置いているので、再現率を高く出来たことは大きな改善点であり優位性があると考えられる。

手法 P2 と手法 SP2 に共通して、スライド毎に関連度が一番高い部分文書の対応付けを正しくするケースも見られた。3.5 節の対応付け部分集合決定手法では、関連度が一番高い部分文書の範囲を広げて対応付けを行うものである。各スライドの関連度が一番高い部分文書が最も重要になってくる。例えば、セクション 3 に属する部分文書と対応づくスライドが、セクション 5 のあるパラグラフと局所的に関連度が高くなり、関連度の top-1 がセクション 5 の部分文書、top-2 から top-10 まではセクション 3 の部分文書であったとする。そのような場合、セクション全体の関連度はセクション 3 の方がセクション 5 より大きくなる場合があるので、二段階の対応付けによりまずセクション単位で対応付けを行うことで正しい対応付けを行うことが出来た。

また、パラグラフ粒度の対応付けについて、本研究の手法では部分文書単位での再現率が著しく低い値になってしまうため、スライド単位ではどのくらい対応関係を網羅出来ているのかについて評価した。各スライドについて得られた対応関係が、正解部分文書集合と一部が一致するようなスライドを部分正解と定義し、その割合を部分正解率として結果をまとめたものが表 5 である。

この結果より、手法 P1 では約 59 %、手法 P2 では約 67 % のスライドについて、正解部分文書の一部を対応付けることが出

表 6 各手法による対応付けの平均実行時間

手法	実行時間 (秒)	手法	実行時間 (秒)
手法 P1	28.37	手法 SP1	30.35
手法 P2	21.84	手法 SP2	22.69

来ていることが分かる。各スライドの正解部分文書集合は連続した部分文書の集まりであることが多いので、手法 P2 については約 67 % のスライドで対応付けられた部分文書の前後を読むことでスライドの記述と関連のある記述を論文の中から見つけることが出来る。また、二段階対応付けによって部分正解率が 0.081 向上しており、部分正解スライドの数が増加していることが分かる。

5.2.2 速度比較

次に、4 つの手法を実行時間の観点から評価した。対応付けは Java を用いて実装した。実験環境は CPU: Intel Core i5 M 580 2.67 GHz, メモリ: 4.0 GB, OS: Windows 7 Professional である。実験の結果をまとめたものが表 6 である。

手法 P1 と手法 SP1, 手法 P2 と手法 SP2 を比べると、それぞれ平均 1.98 秒, 0.85 秒ずつ前者の方が実行時間が短くなっている。これは、前者がセクション粒度を考慮していないため部分文書の範囲の重複が存在しないので、部分文書数が少なくなっているためである。手法 P1 と手法 P2, 手法 SP1 と手法 SP2 を比べると、それぞれ平均 6.53 秒, 7.66 秒ずつ後者の方が実行時間が短くなっている。これは、後者は二段階の対応付けを行うことでペナルティなどの計算を行う範囲が小さくなり操作するデータ数も少なくなるので、その分の計算量の削減による実行時間の減少が、対応付けの回数の増加に伴う実行時間の増加を上回った結果である。二段階対応付けによって、対応付けの実行時間はわずかではあるが改善されたと考える。

二段階対応付けによって高速化を図ったが、結果として大きな改善は見られなかった。実行時間は論文の部分文書数とスライドの枚数に依存しており、それぞれが大きいほど実行時間は長くなる。実行時間が最も長いものでは二段階対応付けによって約 25.4 秒短縮することが出来たが、論文スライドペアのサイズが小さいものでは範囲を絞り込む利点が小さかったため、実行時間に大きな差は見られなかったと考えられる。

6. おわりに

6.1 まとめ

本研究では、学術論文や論文発表用のプレゼンテーションスライドが Web などで容易に参照できるようになった近年の状況を踏まえて、論文とそのペアとなるプレゼンテーションスライドの部分対応付けを行う手法について考えた。西田らの研究 [6] を基に、その手法の改善を行った。また、部分対応付けを行う際の論文の部分文書の粒度について、パラグラフ粒度、セクション・パラグラフ粒度、またこれらを最初にセクション単位で対応付けてから行う二段階対応付けといった 4 種類の手法を提案し、その性能を対応付け精度と実行時間の観点から比較した。

4 種類の手法の精度を評価する実験では、全体的にパラグラ

フ粒度での対応付けは適合率が高く、セクション・パラグラフ粒度では再現率が高くなった。二段階対応付けによって対応付けの範囲が大きくなり、手法 P2 は手法 P1 よりも適合率・再現率ともに精度を向上することが出来た。手法 SP2 は手法 SP1 よりも再現率が高くなったが、適合率は全体的には低くなった。しかし、これらは再現率を重視した手法であるので、再現率を改善した点で評価できると考えられる。また、手法 P1・手法 SP1 で局所的に関連度が高くなった部分文書を誤って対応付けていたスライドを、二段階対応付けによって最大関連度の部分文書を正しく対応付けることで、上手く対応付けることが出来るようになった。実行時間は、全てのペアにおいて二段階対応付けを行う方が短くなった。

対応付けの精度の観点、実行時間の観点から、二段階対応付け手法は先行研究に対して優位性があると考えられる。

6.2 今後の課題

対応付け手法に関して、2 種類の粒度では違った傾向が得られたので、それぞれの特徴を踏まえた改善が必要となる。また、現在は利用していないプレゼンテーションスライド中の情報を活用することも考えられる。精度の評価の際、セクション粒度をパラグラフ粒度に換算して精度の評価を行ったが、セクションの比重がパラグラフと比べて著しく高くなってしまったので、評価手法を検討する必要がある。更に、論文とプレゼンテーションスライドを同時に閲覧するようなアプリケーションとしての実装を進めていきたい。アプリケーションでは、単純に論文とプレゼンテーションの対応関係を提示するだけではなく、関連度の順位によって関係性の強弱を示したりなど、利用者のニーズに応えた実装を行いたいと考えている。

7. 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] Bamdad Bahrani and Min-Yen Kan. Multimodal alignment of scholarly documents and their presentations. *JCDL*, pp. 281–284, 2013.
- [2] Hongyan Jing. Using hidden markov modeling to decompose human-written summaries. *Computational Linguistic*, Vol. 28(4), pp. 527–544.
- [3] Min-Yen Kan. Slideseer: A digital library of aligned document and presentation pairs. *JCDL*, pp. 81–90, 2007.
- [4] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. *SIGIR*, pp. 232–241, 1994.
- [5] 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の対応付け. 電子情報通信学会言語理解とコミュニケーション (NLC) 研究会, Vol. 104, pp. 7–12, 2004.
- [6] 西田惇人, 清水敏之, 吉川正俊. 論文とスライドの部分対応付け手法. *DEIM*, 2013.
- [7] 清水敏之, 寺田憲正, 吉川正俊. 関係データベースを用いた XML 情報検索システムの開発. 情報処理学会論文誌. データベース 48, Vol. SIG 11(TOD 34), pp. 224–234, 2007.