

論文とスライドの部分対応付け手法

西田 惇人[†] 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: [†]a.nishida@db.soc.i.kyoto-u.ac.jp, ^{††}{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし 本研究では、論文とそれと対になるプレゼンテーションスライドの部分的な対応をとる手法を提案する。部分的な対応関係が分かることは論文やプレゼンテーションスライドの内容を理解する上で有用であり、形態の違いから片方にしか存在しない情報の把握にも利用できる。論文を元に作成されたプレゼンテーションスライドでは論文と共通の箇所が存在し、順序に関してもある程度は論文準拠となると考えられる。このような順序の関係や、論文の構造から得られる異なる粒度の部分文書を考慮して、スライドを問い合わせとして論文の部分文書との関連度を計算する。スライドと論文の部分文書との対応関係の取得は、タイトルの類似度や文書の類似度から確信度の高いスライドを決定し、そのスライドを順序による補正を与えるための軸として、他のスライドの対応関係を取得し、最終的にプレゼンテーションスライドと論文の部分対応関係を取得する。また、本稿では、実際の論文とプレゼンテーションスライドの組を対象に提案手法の評価実験を行い、その精度向上について確認した。

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

研究成果等を発表する様々な会議では、多くの学術論文が発表されており、近年では発表に伴いプレゼンテーションスライドが作成されるようになった。それらの文書の多くは、作者のホームページや Google Scholar^(注1), SlideShare^(注2)等で電子的に利用、閲覧ができる。論文とプレゼンテーションスライドに関し、それぞれの媒体に含まれる文章量は大きく異なり、また本来利用する状況の違いから抽象度や情報量が異なる。それらの特徴を踏まえて、利用者は時間的制約などの状況や、自身の興味の度合いに応じて論文やプレゼンテーションスライドを利用すると考えられる。しかし、例えば論文だけの閲覧ではプレゼンテーションスライドに比べて図や表、イラストでの説明が少ないことも多く抽象的なイメージに乏しかったり、プレゼンテーションスライドだけの閲覧では論文に比べて、圧倒的に情報量が少なく詳細な理解が得られなかったりと、どちらか一方だけの媒体の利用だけでは内容の理解が難しい場合もある。このような場合、内容の理解を深めるためには利用者は両方の媒体を利用すると考えられる。プレゼンテーションスライドを用いて概要を把握し、アルゴリズムの詳細だけ論文を用いて理解するなど、必要に応じて両方の媒体を適切に利用するということが必要になる。2つの媒体を利用する際に、事前に論文とプレゼンテーションスライドの内容の部分対応関係が分かっているれば、片方の媒体からもう一方の媒体における同様の内容の記述箇所を手早く参照でき、効率的に論文やプレゼンテーションスライドの内容を理解することができる。逆に、その両方の媒体に部分対応関係がない部分の中には、プレゼンテーションスライ

ドに記述されていない詳しい説明が論文では記述されていたり、論文にはない実験の図がプレゼンテーションスライドには掲載されていたりと、片方の媒体にしか記述されていない独自の情報を持つ場合もある。そのようなとき2つの媒体同士の部分対応関係が分かっていると、同じ情報を重複して読むことも少なくなり利用者が欲しい情報を効率よく手に入れる事ができる。このように論文とプレゼンテーションスライドを相互補助的に利用し、効率的に情報を得るためには、論文とプレゼンテーションスライドの内容の部分対応関係の取得が重要になる。

本研究では、パワーポイントなどで作成された1つの発表に相当する文書をプレゼンテーションスライドと呼び、その中のシート1枚1枚を単にスライドと呼ぶこととする。そして、論文とプレゼンテーションスライドの相互補助的利用のために、各スライド1枚ずつと論文の部分文書との対応関係を取得することを目的とし、論文の構造やスライドの特徴を考慮した対応付けの手法を提案する。既存の研究には、論文のパラグラフを部分文書としスライドと連続したパラグラフの集合との対応関係を求める研究がある。その研究では、スライドは論文の順序に沿って作成される場合が多いことから、主にスライドの順序を考慮して対応付ける手法を提案している[4]、しかし、その順序を考慮して対応関係の関連度を与えるペナルティは、全てのスライドに対して前後のスライドとの関係から等しく与えられるため、例えば図や表を含むスライドが多いプレゼンテーションスライドの場合や、プレゼンテーションスライドの一部の順序が論文の連続するセクションの順序とは前後するといった少しの変化がある場合に対して、望ましくないペナルティがより多くなると思われる。そこでまず我々は、スライドと部分文書の対応関係に関してより確信度が高いスライドをいくつか選びだす処理を行い、そのスライドを基準とした順序によるペナルティを他のスライドに与える。以後その順序によるペナルティ

(注1) : <http://scholar.google.co.jp/>

(注2) : <http://www.slideshare.net/>

の基準となるスライドを軸スライドと呼ぶ。

このように対応関係の決定を段階的に行うことで、順序によるペナルティを全てのスライドに一樣に与える方法より正確に割り当て、最終的に各スライド1枚と論文の部分文書集合の対応関係を取得する。

2. 関連研究

藤本ら [1] は、構造化がなされていない PDF 形式の論文を、レイアウト情報とテキスト情報をもとにした構造化ルールを用いて論文構造タグを付与し、論文を XML 形式で表現し、論文検索システムを作成している。

羽山ら [2] は、プレゼンテーションスライドの自動生成を目指し、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について述べている。対応付けには Jing の隠れマルコフモデル [3] を用いた対応付け手法をもとにスライド1枚ごとに論文の章、節を対応づける方法を試している。しかし、新聞記事を対象として良い精度が報告されている Jing の手法を、そのまま論文とスライドに適用したのでは、新聞記事と比べ1テキスト中の文数と一文当りの長さが大きく、また本文の他に図表が含まれるため、十分な対応付けができないので、その対応付けの問題を改善するために論文とスライドの対応位置やレイアウト情報を利用した対応付け手法について提案している。

Kan [4] は、論文とその論文をもとに作成されたプレゼンテーションの組の取得、論文とプレゼンテーションスライドの対応付けの手法について提案している。この研究での最終的な対応関係はスライド1枚に対して論文の連続したパラグラフの集合を対応付けることである。 $\cos$ 類似度や Jaccard 係数、さらに *bigram* で $\cos$ 類似度と Jaccard 係数を計算し、各スライドと論文の1つのパラグラフの関連度マトリックスを計算している。その後、単純に高スコアを対応付ける手法、隠れマルコフモデルを用いた手法等4つの手法を比較し、論文とプレゼンテーションの組み合わせに関しては単純に高スコアのパラグラフを対応付ける手法の精度が一番高くなる結果が出ている。Kan の手法ではスライドと各パラグラフの関連度の $\text{top-}k$ (この論文では $k=10$ としている。) から2つ選んでできる組み合わせ全てに関して、その2つのパラグラフを連続パラグラフスパンの始点と終点とし、スライドとそのスパンの関連度を算出した後、各スライドの前後のスライドとの対応関係から順序によるペナルティを考慮し、最終的に1枚のスライドと連続パラグラフスパンの対応関係を決定している。また例外としてスライドの中には対応関係が取れないものの決定もしている。

Kan の手法は対応関係を決定する際の順序によるペナルティに関して、対象となるスライド及び前後のスライドとその関連度の高いパラグラフスパンの順序に基づいて普遍的にペナルティを与えているのに対し、本稿の提案手法は対応関係の確信度の高いスライドを基準として用いて順序によるペナルティを与えている点異なる。また Kan の手法は最終的にスライドと連続したパラグラフスパンとの対応関係を取得するが、本稿の提案手法では、論文のセクションやサブセクションといった

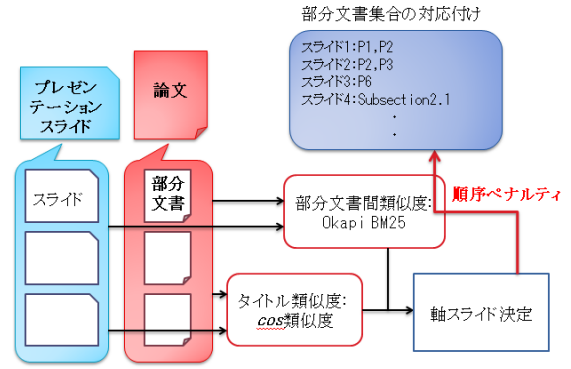


図1 対応関係取得の流れ

XML のノードで分けられる粒度に着目し、スライドとある粒度の連続した部分文書との対応関係を取得する点異なる。

3. 事前準備と手法の概略

我々は、論文とプレゼンテーションスライドの部分的な対応付けを取得するために、論文とプレゼンテーションスライドを変換し XML 文書として扱う。このとき論文は XML のタグに基づいた複数の粒度を考慮し、セクションやサブセクション等で区切られた文章を一つの部分文書として捉え、論文の1段落を1パラグラフとしたものを論文の部分文書の最小単位とする。以後、セクションやパラグラフの単位の区切りをセクション粒度、パラグラフ粒度と呼ぶ。またプレゼンテーションスライドに関しては、各スライドを1単位とし1枚のスライドと論文の部分文書の集合を対応付ける。本研究の対応付け手法の入力は XML 形式の論文とプレゼンテーションスライドとし、その両者の部分文書の対応関係を出力とする。実際のスライドと部分文書の対応イメージは図1のようになる。

また、論文とプレゼンテーションスライドの部分対応関係を取得する手法の概略は以下の通りである。まず、スライドを問合せとし、OkapiBM25 [5] をベースに論文の各粒度の部分文書とスライドの関連度を計算する。そして、セクション粒度などの粗い粒度の関連度やスライドのタイトルと論文の節タイトルの関連度から対応の確信度が高いものを選出し、それらの対応関係をその他のスライドと部分文書の関連度に関して順序によるペナルティを与えるために利用する。この粗い粒度において対応の確信度が高いスライドやスライドタイトルと論文の節タイトルが一致したスライドは、順序によるペナルティを与える基準のスライドとなることから軸スライドと呼ぶ。OkapiBM25 及び順序によるペナルティから計算した関連度から、全てのスライドにおいて一番関連度が高い部分文書をそのスライドと対応関係があるとし、更に各スライドに対して、既に対応関係が決定された部分文書の論文中における位置と他の部分文書の関連度を考慮して、その他の部分文書をスライドに対応付け、最終的に部分文書集合とスライドの対応関係を出力とする。

3.1 OkapiBM25 による関連度計算

スライドの文章をクエリとし、各スライドと論文のそれぞれ

の粒度における部分文書との関連度を計算する。論文はセクション粒度間やパラグラフ粒度間といった同一粒度においても、例えば手法に関して述べたセクションは文章量が多く結論のセクションはそれに比べて少ないなどセクションによって文字数が大きく異なるので、関連度を計算する際には文書長を考慮する必要がある。そこで関連度の計算には文書長を正規化する特徴をもった *OkapiBM25* [5] を使用する。 *OkapiBM25* は以下の式で表される。

$$score(D, Q) = \sum_{q \in Q} s_{BM25}(D, q) \quad (1)$$

$$s_{BM25}(D, q) = IDF(q) \cdot \frac{f(q, D) \cdot (k+1)}{f(q, D) + k \cdot (1-b + b \cdot \frac{|D|}{avgdl})}$$

$$IDF(q) = \log \frac{N - n(q) + 0.5}{n(q) + 0.5} \quad (2)$$

ここで Q はクエリ、 $f(q, D)$ は文書 D における単語 q の出現頻度、 $|D|$ は文書 D の文書長、 $avgdl$ は文書集合の平均文書長、 N は文書の総数、 $n(q)$ は単語 q が出現する文書の数である。 k と b はパラメータで、一般的には $k = 2.0$, $b = 0.75$ とされており、本研究でもこの値を使用する。また、セクション粒度とパラグラフ粒度といった異粒度間においては同一粒度間より顕著に文書長に差があることを考慮して、異粒度間では異なる文書集合として関連度を計算する。この際、Martin [6] らが提案した、*OkapiBM25* を XML 形式の文書にも効率よく適用でき、XML のタグごとに関連度を計算するように改良した以下の式を使用する。

$$score(n, A, Q) = \sum_{q \in Q} s_{BM25}(n, q) \quad (3)$$

$$s_{BM25}(n, q) = \log \frac{N_A - ef_A(q) + 0.5}{ef_A(q) + 0.5} \cdot \frac{(k_1 + 1) \cdot ftf(q, n)}{K + ftf(q, n)}$$

with

$$K = k_1((1-b) + b \cdot \frac{\sum_{q \in \text{full content of } n} ftf(q, n)}{avg_n\{\sum_q ftf(q, \hat{n}) | \hat{n} \text{ with tag } A\}})$$

ここで、 $ftf(q, n)$ はノード n の文書に単語 q が出現する回数、 N_A はタグ A のノードの数、 $ef_A(q)$ はタグ A のノードのうち単語 q を含むノードの数、 k_1 は式 (2) における k である。

4. 軸スライドの決定

ある 1 枚のスライドと論文の部分文書との関連度は、スライド内に特徴的な語が少ない程、各関連度の差は小さくなり、特徴的な語が多く使われる程、関連度は局所的に大きくなる。このことから一番高い関連度を持つ部分文書と次に高い関連度を持つ部分文書を比べた時、その差が大きければ大きいほど、スライドと前者の部分文書との対応関係があることの確信度が高いと言える。またスライドの中にはタイトルに INTRODUCTION や CONCLUSION などの単語を使用することが多く、またこれらは論文のセクションのタイトルにもなっていることも多い。よってスライドのタイトルと論文の節や小節のタイトルが類似している時、そのスライドとその節にあたる部分文書との対応関係があることの確信度が高いと言える。これら、スライドと

部分文書の関連度の差やスライドタイトル等を利用して、まずいくつかのプレゼンテーションスライドと部分文書の対応関係を決定する。このようにして次の小節で述べる順序によるペナルティのための軸となるスライドを選出する。

4.1 タイトルの関連度による軸スライドの決定

スライドの中にはタイトルのみで成るスライドがあり、そのタイトルはそのまま論文のセクション等のタイトルを利用していることが多い。したがってセクション等のタイトルをそのまま利用したスライドは、そのセクション等と対応関係があり、またその対応は確信度が高いと考えられるので、順序によるペナルティの基準となる軸スライドとして利用する。以降、プレゼンテーションスライドにおける m 番目のスライドを s_m 、論文の n 番目に出現した部分文書を d_n で表す。3.1 節で得られた関連度とは別に各スライドと部分文書におけるタイトル同士の間接度を計算する。あるスライド s_k にタイトルが存在すれば、そのタイトルの語の集合 x と論文の i 番目に表れる節や小節におけるタイトルの語の集合 y_i との $\cos$ 類似度を計算し、

$$\cos(x, y_i) > \alpha \quad (4)$$

上の式を満たすスライドを軸スライドとする。スライドや論文の部分文書のタイトルなど語数が少ない場合には $\cos$ 類似度は高い値になることが多く、 α の値が低すぎると誤った対応を行いやすいので、適切に設定してやる必要がある。

4.2 内容関連度による軸スライドの決定

3.1 節で得られた関連度の差から軸スライドを決定する。3.1 節で得られたスライド s_m と部分文書 d_n の関連度を $S(s_m, d_n)$ で表す。ここでスライド s_k と一番関連度が高い部分文書を d_i とし、 d_i と兄弟関係にも子孫関係にもない部分文書の中で s_k と一番関連度の高い部分文書を d_j とした時、

$$\begin{cases} S(s_k, d_i) - S(s_k, d_j) > \frac{MaxSimAvg}{\beta} \\ S(s_k, d_i) > 0 \end{cases} \quad (5)$$

を満たすような s_k を軸スライドとする。また、 d_i はスライド s_k と対応関係があるものとする。ここで $MaxSimAvg$ は各スライドと各部分文書との関連度において最大のものの平均であり、 β はパラメータとする。プレゼンテーションスライドは、それを作成する人の傾向や論文の内容によって文章量が増える。スライドの文章量が多い場合には、スライドと論文の部分文書との関連度は全体的に大きくなるが、逆に文章量が少ない場合には関連度は全体的に小さくなる。したがって、各スライドが軸スライドであるかどうかを判定する基準にある定数を使用すると、文章量が少ないスライドが多い場合には軸スライドが決定されないこととなるので、関連度の数値を相対的に利用し、かつある程度 $S(s_k, d_i)$ と $S(s_k, d_j)$ に差があるものを軸スライドと決定するために、 $MaxSimAvg$ とパラメータ β を利用する。単純に関連度の最大値を使用してしまうと、一部のスライドと部分文書の関連度が突出して高い数値をとるような場合、誤ったスライドを軸スライドとしてしまう可能性が高くなるので、関連度の最大値の平均を用いて、閾値が高い数値になるのを抑制することができる。

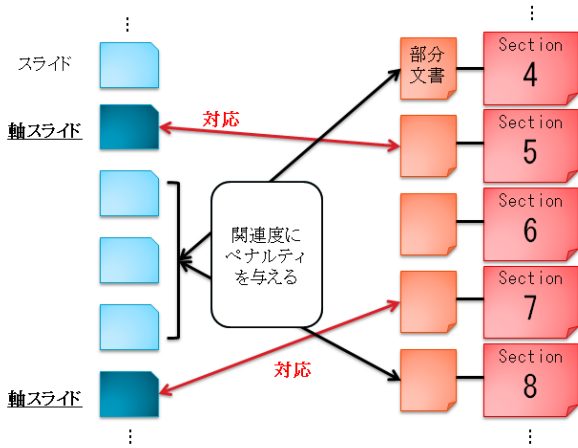


図 2 順序ペナルティの付与

4.3 軸となるスライドの例外

スライドと論文の部分文書の関連度が高いものでも、APPENDIX 等と対応するスライドや、アウトラインを表すスライド、類似度は高いが前後の何枚かのスライドと比べて明らかに対応範囲が異なるものなどは、スライドと論文の単調な順序対応を逸脱したと考えられるので、異常値が高いとして軸スライドには使用しない。

5. 対応関係の取得

本節では、全てのスライドと 1 つの部分文書を対応付けるために、スライドと部分文書の関連度に対して、軸スライドを基準とした順序によるペナルティを与える手法について述べ、そこから最終的な目標である各スライドと論文の部分文書集合とを対応づける手法について述べる。

5.1 順序によるペナルティの付与

4. で選出した軸スライドを利用して、軸スライドではない他のスライドと論文の部分文書との対応関係を決定する。論文を基に作成されたプレゼンテーションスライドでは、論文の流れに沿ってスライドが作成される傾向がある。そのようなプレゼンテーションスライドにおいては、各スライドの順序も論文の内容の順序を踏襲したものになるので、スライドと論文の順序を考慮して、軸ではないスライドの対応関係を決定する。例えば、2 枚の軸スライド s_i , s_j があり軸ではないスライド s_x があると、スライドの順番は前から s_i , s_x , s_j であるとする。そしてスライド s_i はセクション A、スライド s_j はセクション B と対応関係があり、セクション Y をセクション A と B の間のセクションとするとスライド s_x はセクション Y と対応関係がある可能性が高い。このことからセクション 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内の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。イメージは図 2 となる。

具体的には軸スライドではないスライドと部分文書の各関連度 $S(s_x, d_k)$ に対して

$$\{1 - \gamma \cdot \log(1 + \text{SectionDistance})\} \quad (6)$$

をかける。ここで

$$\text{SectionDistance} = \max\{(s_i \text{sec} - d_k \text{sec}), (d_k \text{sec} - s_j \text{sec})\}$$

ただし $\text{SectionDistance} < 0$ ならば $\text{SectionDistance} = 0$

$s_i \text{sec}$ はスライド s_i と対応関係があるセクションの番号、 $d_k \text{sec}$ は部分文書 d_k を含むセクションの番号、 γ はパラメータであり、 γ は後の実験で値を定める。

このペナルティは軸スライドでないスライドの関連度に関して、前後の軸スライドと対応するセクション間の部分文書にはペナルティがかからず、それ以外の部分文書には対応箇所の距離に応じたペナルティがかかるようになっている。このペナルティを全ての軸でないスライドに行った後、各スライドと関連度が一番高い部分文書に対応関係があると決定する。この処理で全てのスライドに対して一つの部分文書との対応関係が決定される。

5.2 対応関係のないスライド

outline, question, thank you という語がタイトルとなっているスライドはスライドの流れを表すスライドであったり、実際にプレゼンテーションを行う際の聴衆に向けての言葉であったりして論文の部分文書とは対応関係がない。図が主なスライドは文書の内容による関連度が低い傾向にあるが、論文の部分文書と対応付けて理解の手助けに利用できると有用である。逆に、図を含まずにかつ語が少なく、関連度が低いスライドは、タイトルの一致が見られなければ、上記のスライドと同様に情報も少なく論文の部分文書と対応関係があるとは考えにくい。したがってそれらのスライドは論文の部分文書と対応関係がないものとして決定する。

5.3 対応部分文書集合の決定

この節では、最終的な目標であるスライドと部分文書集合との対応関係を取得する手法を述べる。1 枚のスライドでは文章量が限られるので、様々な部分文書と対応関係があるというスライドはそれ程多くなく、そういったスライドよりも、1 つの paragraph 粒度の部分文書、もしくは複数の隣り合う部分文書と対応関係があるスライドの方が出現頻度が高いと予想される。したがって後者のスライドの正解の対応を得ることに重点を置き、対応関係のある部分文書集合を求める。

順序ペナルティを既に与えたスライドと各粒度の部分文書の関連度の top- k を用いて、1 枚のスライドと各部分文書との関連度において一番高いものは 5.1 節で対応関係が決定しているので、それ以下の関連度の部分文書も対応関係にあるかどうかを決定する。具体的には top- k の上位の関連度から走査していき、対象の部分文書が既に対応関係が決定している部分文書と兄弟関係であるか子孫関係であった場合、新しくその部分文書をスライドと対応関係があるとして対応付けに加える。この時、対象の部分文書が、既に対応関係が決定しているどの部分文書とも兄弟間にも子孫関係にも無かった場合、そのスライドに対する走査を終了し、その時点で対応づけられている部分文書集合を最終的に対応関係があるものとして決定する。

6. 実験

本節では、提案手法である、内容関連度とタイトル関連度から軸スライドを決め、そのスライドを基準とした順序によるペナルティを与え、最後に部分文書集合の対応関係を取得するという3段階のステップの実験を行う。実験を行うために論文とプレゼンテーションスライドを3組用意し、人手でスライドと論文の部分文書との正確な対応関係を決定し正解データを作成した。合計3つの実験を行い提案手法について考察する。まず2つの実験を行い、軸スライド決定のためのパラメータ β と順序ペナルティのためのパラメータ γ という2つのパラメータの適切な値を求める。そしてその値を定めたパラメータのもとで、複数の尺度で、提案手法により得られた各スライドと論文の部分文書の対応関係の精度と既存研究との精度比較の実験を行う。ここで実験に使用した3組のプレゼンテーションスライドの主な特徴について述べる。論文「Approximate XML Query Answers [7]」は比較的スライドの文章量が多く内容による関連度から対応関係が得られやすいプレゼンテーションスライドとなっている。またスライド単体の順序の大きな逸脱は見受けられないが、セクション5、セクション7、セクション6の順でスライドを作成しているという特徴がある。論文「Deriving Private Information from Randomized Data [8]」はプレゼンテーションスライドに図を含んだスライドが多く、文章量が少ない。したがって内容の関連度は全体的に小さく、正解の対応関係を得ることが比較的困難なプレゼンテーションスライドである。プレゼンテーションスライドの順序はおおよそ論文の順序に準拠している。論文「Data Exchange beyond Complete Data [9]」は、文章量と図の量に関して一つ目の論文とプレゼンテーションスライドの組と似ている。しかし、論文には明確に記述していない内容のスライドがあり、スライド単体での順序の大きな逸脱が見受けられる箇所があるという特徴がある。

6.1 軸スライド決定のためのパラメータ

4.2節で述べた、式(8)

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

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 について、最適な値を決定するために実験を行ったところ、表1の結果が得られた。パラメータ β は2から3.5の値で変化させ、表中の“軸スライド”とは、式(8)によって得られた軸スライドの個数であり、“正解軸スライド”とは、その軸スライドの中で関連度が一番高い部分文書と実際の正解の部分文書とが兄弟関係もしくは子孫関係にあるスライドの数、つまり同一セクション内にあるスライドの数である。

表1の実験結果について考察すると、ペア2に関しては、文章量が少ないスライドが多く $MaxSimAvg$ の値も低くなり、全体的に関連度も小さいので内容の関連度を考慮した軸スライドは30枚のスライドのうち2枚しか得られなかった。ペア1に関しては $MaxSimAvg$ もそれなりに大きく、 $\beta = 2.00$ では軸スライド5枚、 $\beta = 3.50$ では軸スライド14枚が得られ、全

表1 β による軸スライド数の変化

	β	軸スライド	正解軸スライド
論文スライドペア1 スライド数53 $MaxSimAvg = 10.099$	2.00	5	5
	2.25	6	6
	2.50	7	7
	2.75	10	9
	3.00	12	11
	3.25	13	12
	3.50	14	12
論文スライドペア2 スライド数30 $MaxSimAvg = 6.053$	2.00	2	2
	2.25	2	2
	2.50	2	2
	2.75	2	2
	3.00	2	2
	3.25	2	2
	3.50	2	2
論文スライドペア3 スライド数56 $MaxSimAvg = 13.969$	2.00	3	3
	2.25	4	4
	2.50	4	4
	2.75	6	6
	3.00	7	6
	3.25	8	7
	3.50	10	9

体のスライドのおよそ $\frac{1}{4}$ が軸スライドとして得られた。しかし $\beta = 2.75$ を境に、提案手法で得られた軸スライドのうち、最大の関連度を持つ論文の部分文書が実際に対応関係のある部分文書とは同一でも子孫関係でもない、不正解となる順序ペナルティの基準には適さないスライドを1,2枚軸スライドとしてしまう結果となった。ペア3もペア1と同様の結果で、 β を大きくし軸スライドを判定する閾値が小さくなれば、軸スライドの枚数は増えるものの $\beta = 3.00$ を境に軸スライドには適さないスライドを軸スライドとしてしまう結果となった。

この3組の論文とプレゼンテーションでは、ある閾値以上で全ての軸スライドの部分文書との対応関係が実際の正解と一致し、ある閾値以下では軸スライドの部分文書との対応関係が実際の正解と一致しないようなものが得られる。したがってスライドと一番関連度の高い部分文書と他の部分文書との関連度の差が大きいスライドにおいて、一番高い関連度の対応関係が正解である確信度が高く、そのスライドの対応関係を決定し軸スライドとすることは妥当であると考ええる。しかし、実験に用いた論文とプレゼンテーションスライドの組が少なく閾値の幅が大きいので、他に多くの組で実験を行い、閾値のより適切な値や算出方法を模索する必要がある。以後に行う実験においてはペア1とペア3の正解軸スライド、不正解軸スライドの個数を考慮し、また更に実験を行った結果から、 $\beta = 2.60$ と定める。

6.2 順序ペナルティのためのパラメータ

5.1節で述べた、式(9)

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

における、軸スライドを基準としたスライドの順序によるペナルティを付与するためのパラメータ γ について、最適な値を決

表 2 順序ペナルティによる正解スライドの変化

	γ	部分適合 スライド数	部分 適合率	範囲適合 スライド数	範囲 適合率
論文スライド ペア 1 スライド数 53	0.0	32	0.603	46	0.867
	0.1	34	0.641	48	0.905
	0.2	34	0.641	49	0.924
	0.3	35	0.660	49	0.924
	0.4	35	0.660	49	0.924
	0.5	35	0.660	49	0.924
論文スライド ペア 2 スライド数 30	0.0	12	0.400	13	0.433
	0.1	13	0.433	17	0.566
	0.2	14	0.466	18	0.600
	0.3	14	0.466	18	0.600
	0.4	14	0.466	18	0.600
	0.5	14	0.466	18	0.600
論文スライド ペア 3 スライド数 56	0.0	26	0.464	31	0.553
	0.1	28	0.500	35	0.625
	0.2	29	0.517	37	0.660
	0.3	29	0.517	38	0.678
	0.4	29	0.517	39	0.696
	0.5	29	0.517	39	0.696

定するために実験を行ったところ、表 2 の結果が得られた。パラメータ γ は順序によるペナルティがない 0.0 から順序によるペナルティがある 0.5 の値まで変化させた。表中の”部分適合率”とは、式 (9) によって順序によるペナルティをかけた後、関連度が最大の部分文書をスライドと対応関係があるとし、実際の正解の対応関係の一部と一致するようなスライドの数の割合である。”範囲適合率”とは、その得られた対応関係の部分文書が実際の正解の対応関係の部分文書と同一セクション内に存在するようなスライドの数の割合である。

この実験結果について考察すると、ペア 1 に関しては、 $\gamma = 0.0$ の順序によるペナルティを与えないときに比べて $\gamma = 0.3$ としたときは、実際の正解の対応関係と一致するスライドが 3 枚増加した。ペア 2 とペア 3 に関しては、 $\gamma = 0.0$ の順序によるペナルティを与えないときに比べて、それぞれ $\gamma = 0.2$ のときには 2 枚、 $\gamma = 0.4$ の時が 3 枚、実際の正解の対応関係と一致するスライドが増加し、実験に寄って得られた対応関係をもつ部分文書と、正解の対応関係の部分文書とが同一セクション内にあるようなスライドもそれぞれ 3 枚と 2 枚増加している。なお、どの論文とプレゼンテーションスライドの組においても、元々最大の関連度の対応関係のうち正解の対応関係と一致していたものが、順序によるペナルティによって不正解となったスライドはなかった。この実験から順序によるペナルティを与えることで、正解の対応関係と一致するスライドは増え、他のスライドにおいても正解と一致はしないものの、比較的近いパラグラフや同一セクションのパラグラフと対応関係を得ることができることが分かり、比較的論文の順序に基づいて作成されたプレゼンテーションスライドにおいて、軸スライドを基準とした順序によるペナルティを与える手法は妥当であると言える。ただし、スライド単体で順序を大きく逸脱したスライドに関しては、元々対応関係が正しかったものがこの手法によって不正

解になることも考えられるので、順序を大きく逸脱した単体のスライドの対応関係を正しく求めるためには、この順序によるペナルティとは異なった手法を用いて評価する必要があると考える。以後の実験において表 2 の結果よりパラメータ γ は、 $\gamma = 0.3$ と定める。

6.3 既存研究との精度比較

この節では、提案手法により得られた各スライドと論文の部分文書の対応関係の精度を評価し、既存研究によって得られる対応関係の精度と比較する。その際に、適合率、再現率、 F 値、2 種類の部分正解、完全正解という 6 つ尺度で精度を評価する。ここで、2 種類の部分正解とは、スライド s_k において提案手法により得られた対応関係が正解データを少なくとも一つ含むスライド s_k の数の割合とする尺度（部分正解 (a) とする）と、スライド s_k において提案手法により得られた対応関係が正解データを全て含むスライド s_k の数の割合とする尺度（部分正解 (b) とする）である。完全正解は、スライド s_k において提案手法により得られた対応関係が正解データと完全に一致するスライド s_k の数の割合とする。この実験における適合率は、スライドと正解の対応関係がある論文の部分文書は様々な粒度となるが、全てパラグラフ粒度の部分文書に置き換えてその部分文書を正解データの数を使用する。例えばその正解の対応関係がセクション粒度の部分文書ならばその範囲内にあるパラグラフ粒度の部分文書の個数を数える。提案手法により得られた対応関係も同様にパラグラフ粒度の部分文書の数を使用する。つまり、提案手法により得られた対応関係があるパラグラフ粒度の部分文書の個数をスライド枚に数え、その中に正解の部分文書がスライド枚にいくつあるかという値を適合率とする。再現率に関してもこの適合率と同様にパラグラフ粒度の部分文書から算出する。

ここで、我々が行った Kan による既存研究 [4] の実装について簡単に述べる。既存研究の手法は、*bigram* の *Jaccrad* 類似度を用いてスライドと論文のパラグラフ粒度の部分文書との関連度を計算しているが、我々はその関連度の計算に提案手法に使用した *OkapiBM25* を用いた。そして、各スライドに対して部分文書との関連度の top-10 から 2 つの部分文書を選び、その 2 つの部分文書を始点と終点のパラグラフとし、その間全ての部分文書の関連度の平均とその長さによる重みから、新たな連続パラグラフスパンの関連度を計算する。更に、対象のスライド s_i と各連続パラグラフスパンの関連度に対して、前後のスライド s_{i-1} と s_{i+1} と最大の関連度を持つ連続パラグラフスパンとの順序を考慮して順序によるペナルティを与える。そのうえで、最も関連度が高くなる連続パラグラフスパンをスライドと対応付ける。実際には対応関係がないスライドに関しては、その連続パラグラフスパンの長さや関連度の低さ、スライドの語の少なさから、対応関係を持つ部分文書が無いと決定する。

表 3 が提案手法と既存研究の手法によって得られた対応関係の適合率と再現率、 F 値である。ペア 2 においては、文章量が少ないプレゼンテーションスライドなので、Kan の手法は、スライドの対応関係を対応がないとするか、長いスパンの連続パラグラフと対応付けてしまったので、適合率、再現率が共に低

表 3 パラグラフ粒度の部分文書単位での比較

	手法	適合率	再現率	F 値
論文スライドペア 1 スライド数 53	提案手法	0.248	0.517	0.335
	既存研究	0.218	0.282	0.246
論文スライドペア 2 スライド数 30	提案手法	0.234	0.382	0.290
	既存研究	0.102	0.034	0.051
論文スライドペア 3 スライド数 56	提案手法	0.363	0.430	0.394
	既存研究	0.410	0.353	0.380

い値となっている。ペア 1 に関しては、提案手法の軸スライドがプレゼンテーションスライドに対して偏ることなく得られ、軸スライドによる順序ペナルティが多くのスライドに対して適切に作用し、適合率、再現率、F 値全てにおいて提案手法が Kan の手法を上回る結果となった。ペア 3 に関しては、Kan の手法の方が適合率が高くなり、提案手法の方が再現率が高くなった。これは、提案手法のスライドと対応づけられる部分文書がパラグラフ粒度の部分文書に限らないことに起因しており、セクション粒度の部分文書やサブセクション粒度の部分文書と対応付けることが多くなるので、対応関係があるパラグラフ粒度の部分文書の個数が多くなり、ペア 1 でも見られるように提案手法は適合率より再現率が高くなる傾向にあることからだと考えられる。また、Kan の手法は対応付けのパラグラフスパンが長くなるスライドは対応付けがないと判定することが多く、適合率が高くなると考えられる。他の実験からも提案手法は適合率以上に再現率が高くなる傾向にあり、Kan の手法は適合率が高くなる傾向にあることが伺える。しかしながら、ペア 3 の結果に関しては一概にどちらの手法が優れているとは判断できないので、その優劣は別の尺度を用いた表 4 で評価する。

次の、表 4 が対応関係があるスライドの割合の尺度である部分正解 (a)、部分正解 (b) と完全正解による評価実験の結果である。比較的文章量の少ない論文においてもペア 2 から、3 つの尺度すべてで提案手法が Kan の手法を上回る結果が出ている。またペア 1 とペア 3 においては、完全に正解データと対応関係が一致するスライドの数こそ、提案手法も Kan の手法も同じであるが、部分正解 (a) と部分正解 (b) で提案手法が上回る結果となった。上で述べたように提案手法が対応関係の取得の方法により再現度が高くなる傾向にあるので、このような結果になったと考えられるが、例えば実際に論文とプレゼンテーションスライドを相互補助的に利用するインターフェースがあった場合、スライドとまったく関係のない部分文書が対応づけられて利用者に表示されるよりも、多少範囲が広がったとしても正解の対応がある部分文書が含まれる範囲を対応付けて表示させる方が有用だと思われる。したがって、今後実際にアプリケーション等に反映させることを考えた場合、提案手法の対応付けの方が有用であると考えている。

7. おわりに

7.1 ま と め

本研究では、論文とその論文をもとに作成されたプレゼンテーションスライドの部分的な対応付けを行う手法について提

表 4 スライド単位での比較

	手法	部分正解 (a)	部分正解 (b)	完全正解
論文スライドペア 1 スライド数 53	提案手法	0.641	0.547	0.396
	既存研究	0.452	0.415	0.396
論文スライドペア 2 スライド数 30	提案手法	0.466	0.400	0.200
	既存研究	0.133	0.133	0.133
論文スライドペア 3 スライド数 56	提案手法	0.500	0.464	0.375
	既存研究	0.410	0.375	0.375

案した。提案手法では、プレゼンテーションスライドの各スライドによって文章量や図の量、タイトルが異なることから、論文の部分文書との正確な対応付けを得られやすいスライドとそうでないスライドがあることに着目し、正確な対応付けの得られにくいスライドにおいても正確な対応付けを得るための方法として、対応付けが正確であるという確信度が高いスライドを優先的に選びだす方法と、その確信度が高いスライドを基準とした順序によるペナルティを与えて、最終的に他のスライドと論文の部分文書の正確な対応付けを得る方法について提案した。

提案手法と既存研究との精度比較の実験では、対象とした論文とプレゼンテーションスライドの全ての組に対する、スライドと論文の部分文書の正確な対応関係の取得は、既存研究の手法より提案手法の方が、概ね良い精度が得られる結果となった。しかし、1 組の論文とプレゼンテーションスライドのペアに関しては、提案手法が既存研究の手法より再現率こそ上回るものの、適合率が下回る結果となった。これは、適合率の計算の仕方とそれぞれの手法の特徴によるものだと考えられる。提案手法では、スライドと対応付ける部分文書にセクションノード内の文章、パラグラフノード内の文章といった粒度があり、タイトルの一致による対応関係の決定や兄弟や子孫の部分文書を対応関係として追加していく最終的な対応関係の求め方から、各スライドと対応関係がある部分文書の範囲が広くなる傾向にあり、適合率より再現率が低くなりやすいと思われる。また、提案手法は対応関係がある部分文書の範囲が広くなるという傾向から、各スライドが一部でも正解の対応関係をもつようなスライドの割合という異なる尺度の評価では、全ての論文とプレゼンテーションスライドの組において、既存研究の手法より高い数値を示した。

適合率、再現率、F 値の尺度による評価とスライドごとの尺度による評価とを総合的に判断して、提案手法は既存研究よりよい精度が得られると結論付ける。さらに、手法により得られた対応関係を利用して、実際に論文とプレゼンテーションスライドを相互補助的に利用する場合を考えた時、対応関係の再現率が高いと、手法により得られた対応関係に実際の正解があるスライドや論文の部分文書が多くなるので、提案手法は既存研究より有用であると考えている。

7.2 今後の課題

実験から得られた結果をもとに、今後更に提案手法を改善していくための課題について述べる。

提案した手法は、適合率と再現率が絶対的には高くはならな

かったが、前節でも述べたようにこれは、手法により得られる対応関係の範囲が広がる傾向にあることによると思われる。例えば、スライドのタイトルが論文のセクションタイトルと一致するが、そのスライドのタイトル以外の文章量が多い場合や、他のスライドでも同じタイトルを持っている場合などは、実際はパラグラフ粒度の部分文書と対応付けられることも考えられ、提案手法ではこのケースに対処しきれず、広い範囲の対応関係を得て、適合率と再現率の分母が大きくなってしまった。したがって、スライドのタイトルと論文の部分文書のタイトルの関連度から軸スライドを選ぶ処理の他に、そのスライドの文章量の考慮や、タイトルが一致した部分文書の子の粒度におけるそれぞれの部分文書の内容関連度の相対的な考慮をして、最終的な対応関係を決定する必要がある。またスライド同士のタイトルの比較も行い、その関連度が高い場合には、複数枚のスライドにわたって同一タイトルが使用されていることになり、そのスライド全てがセクションなどの粗い粒度の部分文書と対応関係があるとは考えにくいので、より子供の粒度の部分文書と対応付けやすくするといった処理が必要となる。

また、スライドや論文の部分文書のタイトルによる関連度の計算には、タイトルに含まれる語は比較的少ないので、少しの誤差が大きな影響となる。今回の実験でも、タイトルに含まれる語の活用語尾が変化したために望ましい対応関係が取得できないということが生じたので、語幹を利用して関連度を計算することが必要である

提案手法では、より正確な対応関係を得るために、軸スライドとそのスライドを基準とした順序によるペナルティを与える手法について提案したが、この手法はある1つのスライドが順序を大きく逸脱した時にそのスライドと部分文書の正確な対応付けを得ることが難しい。しかし、そのようなスライドは、例の図であったり、最後の章の付録に記述した、詳しい数式や、図であるといった特別な特徴を持つと考えられるので、多くの論文とプレゼンテーションスライドの組に対して実験を行い、正解データの学習をし、どのような語を含むスライドが順序を大幅に逸脱するスライドとなりやすいかを判定する必要がある。また順序の逸脱だけでなく、単純に提案手法の有用性をより詳しく考察するためにも多くの論文とプレゼンテーションスライドの組で実験を行う必要がある。

最終的には手法により得られた対応関係を利用して、論文とプレゼンテーションスライドを同時に閲覧するアプリケーションなどの実装が課題として考えられる。またその際、提案手法の対応関係の範囲が広くなり再現率が高くなりやすい傾向を生かして、対応関係がある部分文書に対応度の強弱を与えることができると、利用の際に役立つと思われる。これを実現するために、提案手法では、最終的にスライドと対応関係のある部分文書集合を求める方法は、*OkapiBM25* と軸スライドを基準とした順序による補正から計算される関連度の *top-k* において、既に対応関係があると決定された部分文書と、兄弟関係にあるか、子孫関係にある部分文書をさらにスライドと対応関係があるとして追加していく方法なので、その関連度の順位によって対応関係に強弱をつけるといった手法が考えられる。

以上のように、スライドタイトルによる対応関係の決定の際にさらに精密な処理を加えてより正確な対応関係を得ることや、より多くの論文とプレゼンテーションスライドの組で実験を行い、順序を逸脱したスライドにおいても正確な対応関係を得ることで、スライドと論文の部分文書の対応付けの精度の向上。また、アプリケーションの実装と、それを利用するにあたって提案手法を生かした機能の実装が課題としてあげられる。

文 献

- [1] 藤本 圭, 杉木 健二, 清水 敏之: XML を用いた論文検索システム, 第 28 回電子図書館ワークショップ (2005)
- [2] 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, No.416, Pp.7 - 12 (2004)
- [3] Hongyan Jing, "Using hidden markov modeling to decompose human-written summaries ", Computational Linguistics, 28(4):pp.527 - 544, 2002
- [4] Min-Yen Kan, SlideSeer: A Digital Library of Aligned Document and Presentation Pairs, JCDL (2007)
- [5] S. E. Robertson, S. Walker: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In SIGIR, pages 232-241 (1994)
- [6] Martin Theobald, Ralf Schenkel, Gerhard Weikum, An Efficient and Versatile Query Engine for TopX Search, VLDB (2005)
- [7] Neoklis Polyzotis, Minos Garofalakis, Yannis Ioannidis: Approximate XML query answers, SIGMOD'04 pages 263-274 (2004)
- [8] Zhengli Huang, Wenliang Du, Biao Chen: Deriving private information from randomized data, SIGMOD'05 pages 37-48 (2005)
- [9] Marcelo Arenas, Jorge Perez, Juan Reutter: Data exchange beyond complete data, PODS'11 pages 83-94 (2011)