講義スライドに基づく課題自動生成のための Linked Data生成手法の提案

石井 佑弥[1] 梶岡 慎輔[1] 山本 大介[1] 高橋 直久[1]

概要：近年，日本だけでなく世界的に新型コロナウイルスが流行しており，感染者数は増加の一途をたどっている．その影響により，リモートワークやソーシャルディスタンシングが実践されるようになったことで，ビデオ会議などのビジネスコラボレーションツールの使用が記録的な増加を示している．そこで学習面で対面授業との比較がされており，オンライン授業の問題点が挙げられている．その問題点として，オンライン授業ではコミュニケーションが不足し，質の高い授業を行うための到達度確認ができない事が挙げられる．一般的に到達度確認のためには小テストが有効だが，一問一答形式の問題は Web 検索で容易に答えを見つけられるといったことや出題する問題が同じだと学生同士で話し合って問題を解答する等の不正行為が容易でなので，これらの問題点を解決する問題形式が必要となる．そこで本研究では，講義スライドのアノテーション付き Linked Data 生成システムを提案する．Linked Data 生成によって直接的に本研究の目的であるオンライン授業の双方向性を確保することにはつながらないが，今後本研究で生成された Linked Data を用いて課題を生成することにより，生徒からの応答を得る仕組みとなり，双方向性を確保できるシステムとなる．また，提案システムに基づいてプロトタイプシステムを実装し，評価実験を行った．その結果，特徴語抽出の精度と特徴語抽出の適切な個数を推定することができた．

A Proposal of Linked Data Generation Method for Automatic Task generation based on Lecture Slides

YUYA ISHII[1] SHINSUKE KAJIOKA[1] DAISUKE YAMAMOTO[1] NAOHISA TAKAHASHI[1]

1. はじめに

近年，日本だけでなく世界的に新型コロナウイルスが流行しており，陽性者数・死亡者数ともに増加の一途をたどっている．また，新型コロナウイルスは 2020 年 3 月 11 日に流行がパンデミック状態であると WHO から発表され，これは 2009 年の新型インフルエンザ以来の事例となる．その影響により，リモートワークやソーシャルディスタンシングが実践されるようになったことで，Zoom や Microsoft Teams といったビデオ会議などのビジネスコラボレーションツールの使用が記録的な増加を示している．これらのツールは，オンラインミーティング・社内研修・オンラインセミナー等に利用されている．この中でも，新型コロナウイルスの感染拡大前と比較した際に，オンライン授業は 95.4%の増加となっており最も増加している．このようにオンライン授業が普及したことにより，新型コロナウイルスの感染リスクを軽減することができるようになったことはオンライン授業の最大の利点として挙げられるが，学習面で対面授業との比較がされており，オンライン授業の問題点がいくつか挙げられている．その問題点として，オンライン授業ではコミュニケーションが不足し，質の高い授業を行うための到達度確認ができない事が挙げられる．一般的に到達度確認のためには小テストが有効だが，一問一答形式の問題は Web 検索で容易に答えを見つけられるといったことや出題する問題が同じだと学生同士で話し合って問題を解答する等の不正行為が容易でなので，これらの問題点を解決する問題形式が必要となる．

また，近年の学習活動において，教科や科目全体を通した俯瞰的な学習が求められている [1]．教科横断的な視点

[1] 名古屋工業大学大学院工学研究科
Graduate School of Engineering, Nagoya Institute of Technology

から教育活動の改善を行っていくことが重要な鍵となり，学習指導計画の作成にあたって配慮すべき事項として，各教科・科目等について相互の連携を図ることが挙げられている．その例として，地理歴史科目の第 3 款である「各科目にわたる指導計画の作成と内容の取扱い」では，教科全体として調和のとれた指導や中学校社会科及び公民科との関連並びに地理歴史科に属する科目相互の関連に留意をすることが記されている．

これらのことから，俯瞰的で，かつ，個々の学生に異なった課題を自動生成することが必要となる．そこで本研究では，講義スライドから Linked Data を生成し，それに基づいて上記の要求条件を満たす課題を自動生成する仕組みを考案した．ただ今回は時間の都合により，前者の仕組みについて検討した．

2. 実現上の課題とアプローチ

提案システムを実現するために，以下の課題が存在する．その課題とそれらに対するアプローチを以下に示す．

2.1 Linked Data の構成

1 つ目の課題として，本研究では講義スライドから Linked Data を生成するということで，Linked Data の構成を独自に定義する必要があるという点が挙げられる．

この課題は，今後俯瞰的な問題を生成するうえで，問題を生成しやすくするために，Linked Data の構成を自ら定義することで解決する．本研究で定義した Linked Data の構成を図 1 に示す．

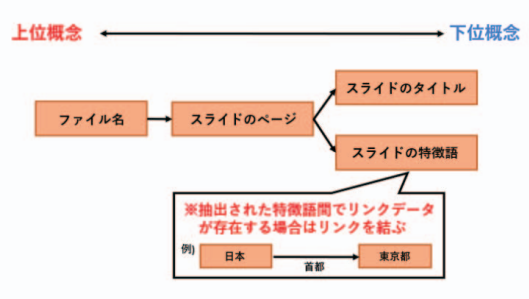

図 1 定義した Linked Data の構成

2.2 Linked Data 生成に必要なノードデータの取得

2 つ目の課題として，Linked Data を生成するのに必要な情報を講義スライドの中から取得する必要があるという点が挙げられる．

この課題は，API を用いて講義スライドをページごとに取得したり，スライドのタイトルや特徴語を抽出することにより解決する．

2.3 Linked Data のリンクデータの取得

3 つ目の課題として，本研究で生成する Linked Data のリンクデータを取得する必要があるという点が挙げられる．

この課題は，ネット上に公開されている Linked Open Data からリンクデータを取得することで解決する．

3. 関連研究

3.1 Linked Open Data

Linked Open Data とは，Tim Berners-Lee によって提唱された 5 star Open Data のひとつで，構造的グラフデータで表され，Web の仕組みを用いてデータ同士を相互にリンクさせたものである．5 star Open Data とは，オープンデータが満たすべき条件を 5 段階のステップとして表現したものであり，図 2 に示す．そして，既に公開・リンクされている Linked Open Data の関係を表したものを Linked Open Data Cloud と呼び，図 3 に示す．

図 2 5 star Open Data

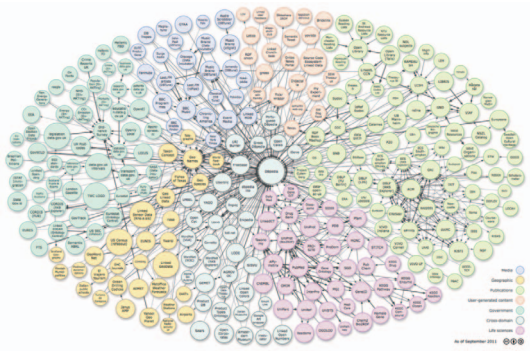

図 3 Linked Open Data Cloud

Linked Open Data などの Web 上のリソースを表現するための形式として RDF(Resource Description Framework) があり，これは「主語 (subject)」「述語 (predicate)」「目的語 (object)」の 3 つの要素で構成され，トリプル構造で表現される．そして，トリプル構造は有向グラフで表現される．その例として，「名古屋工業大学の大学設置年は 1949 年である」という文章があった場合，トリプル構造として「主語」: 名古屋工業大学,「述語」: 大学設置年,「目的語」: 1949 年となり，有向グラフで表したものを図 4 に示す．

図 4　RDF のトリプル構造の例

Linked Open Data は様々な分野で取り組まれており，その中でもよく使われているものとして Wikipedia のデータを Linked Data 形式にした DBpedia があり，日本語版の DBpedia Japanese も存在する．DBpedia 上には「Person」や「Place」など語句の種類としてクラスが定義されており，DBpedia に記載のある語句は特定のクラスに属する．

3.2　多肢選択文法問題の設問形式に関する研究

文献 [2] では，多肢選択文法問題が学習においてどのような思考をもたらすのかを示している．択一式と複数選択式の解答プロセスに焦点を当てており，問題を解く際の思考や，学習者の理解の度合いを算出している．テスト・テイキング・ストラテジーと呼ばれる各問題の正答を導き出す過程で用いられるストラテジーに着目しており，択一式と複数選択式との間にはテスト・テイキング・ストラテジーに大きな差がないことが示されている．

3.3　Linked Open Data からの質問自動生成手法

文献 [3] では，Linked Open Data を利用した知識テスト自動生成手法を提案している．この手法は，エンティティの要約化と RDF の言語化を行うことにより，自然言語による出題を可能にしている．この手法を用いて生成される問題は，一問一答形式で「はい」と「いいえ」の 2 択を問う形式となっている．

3.4　Linked Data から俯瞰的な多肢選択問題を生成する手法

文献 [4] では，Linked Data を用いた俯瞰的な多肢選択式問題自動生成手法を提案している．これは，答えとなる単語を人的に選択したのちに，Web 上に公開されている Linked Open Data から問題を自動生成する手法である．

4.　提案システムの概要

4.1　提案システムの構成

提案システムの構成図を図 5 に示す．提案システムとして，スライド取得機能と特徴語抽出機能，講義スライド外の語句追加機能，アノテーション付与機能，リンクデータ付与機能を実装した．まず，スライド取得機能で講義スライドの本文・タイトルを取得する．そして，スライドの本文から形態素解析を用いて名詞を抽出し，名詞の中から特徴語を抽出する．その後，抽出した特徴語にアノテーションの付与を行い，特徴語間にリンクデータを付与する．そして，これらの機能によって取得したデータを用いて Linked Data を生成する．そして最後に，生成された Linked Data の一部を抜粋して課題を生成する流れとなる．

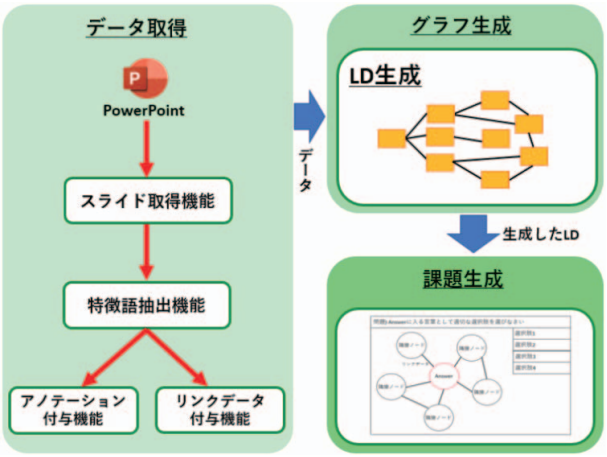

図 5　システム構成図

4.2　提案システムの機能

提案システムの機能を説明する．ここで説明する機能は，今後学習問題を作成していくうえで必要になると予測されるものとなる．

4.2.1　スライド取得機能

スライド取得機能では，講義スライドからスライドのページごとにタイトルと本文を取得する．Apache POI を用いて PowerPoint のタイトルと本文を取得する．その際に，特徴語抽出でスライドのページごとの特徴語を抽出することになるので，スライドのページごとに本文を取得する．

4.2.2　特徴語抽出機能

特徴語抽出機能では，講義スライドから取得したスライド本文から各スライドの特徴を表している語句を抽出する．講義スライドから特徴語を抽出するうえで，以下の 2 つの条件を満たすような方法を用いて実装する．

条件 1　スライドのページごとに特徴語を抽出する．

講義スライドはスライドのページごとにキーワードとなる語句が異なるため，スライドのページごとに特徴語を抽出する．これにより，スライドの各ページに合った問題が生成可能になると考える．また，同じ語句でもスライドのページごとに重要な語句であるかが変わるため，スライドのページごとに特徴語を抽出する．

条件 2　専門用語に多い複合名詞の優先度を高くする．

講義スライドには専門用語が多く含まれている．そこで，本研究では専門用語には複合名詞が多いという前提のもと，特徴語抽出の際に複合名詞の優先度を高くする．

4.2.3 講義スライド外の語句追加機能

講義スライド外の語句追加機能では，特徴語抽出によって得られた語句に加えて必要に応じてスライド外の語句を追加する．講義スライドの中には文章が不足しているものや，図や表が多用されているものがあり十分な Linked Data が生成できない可能性がある．そこで，スライド外の語句を必要最低限で追加することによって，LInked Data の情報不足を補う．

4.2.4 アノテーション付与機能

アノテーション付与機能では，抽出した特徴語に固有表現タグを付与する機能と DBpedia Japanese から特徴語の概要を取得する機能を実装する．特徴語に固有表現タグを付与する機能では，講義スライド本文を NamedEntityAPI に送り，単語ごとに付与された固有表現タグを特徴語抽出で算出された特徴語に付与する．DBpedia Japanese から特徴語の概要を取得する機能では，特徴語抽出によって得られた語句の概要を DBpedia Japanese から取得し語句に付与する．

4.2.5 リンクデータ付与機能

リンクデータ付与機能では，特徴語抽出によって得られた単語間の関係を DBpedia Japanese から取得し，リンクデータとして付与する．DBpedia Japanese には 1 億を超えるトリプル構造が存在する．そのトリプル構造の「主語 (subject)」と「目的語 (object)」に特徴語抽出をして得られた語句を入力し「述語 (predicate)」が得られた場合，その「述語 (predicate)」をリンクデータとして特徴語間に付与する．

5. 提案システムの実現法

5.1 スライド取得機能の実現法

Apache POI を用いて，講義スライドの取得を行う．Apache POI とは，Apache ソフトウェア財団が提供するプロジェクトであり，Excel や Word，PowerPoint など OLE2 複合ドキュメント形式に基づいた様々なファイル様式を Java で取り扱うための Java ライブラリのことである．この Java ライブラリを用いて，講義スライドのタイトルと本文をスライドのページごとに取得する．

5.1.1 スライドのタイトルの取得

Apache POI の Java ライブラリを用いて，スライドのタイトルをスライドのページごとに取得する．PowerPoint の Office テーマでタイトルと定義されている箇所を取得することができる．Office テーマとは，Microsoft Office における文書全体の書式を一括して設定・管理できる機能である．Microsoft Office 2007 で追加された機能であり，PowerPoint 2007 でも利用することができる．Office テーマには具体的に「タイトル　スライド」や「タイトルとコンテンツ」といったテーマがあり，そのテーマのタイトルにあたる文字列を取得することができる．

5.1.2 スライドの本文の取得

スライドの本文をスライドのページごとに取得する．スライド内のテキストボックスのテキスト文字列をすべて取得することができる．よって，グラフや添付された画像の中の文字列は取得することができない．複数のテキストボックスが 1 つのページ内に存在していた場合は改行で区別された形でスライド本文を取得することができ，また，テキストボックス内で改行が使用されていた場合も改行で区別された形でスライド本文を取得することができる．

5.2 特徴語抽出機能の実現法

取得したスライドからスライドの特徴語となる語句を抽出する．特徴語を抽出するまでの手順は以下のようになる．

手順 1　取得したスライドの本文を形態素解析して，名詞を取り出す．

手順 2　取り出した名詞から複合名詞を作成する

手順 3　作成された複合名詞を用いて特徴語抽出を行う．

また，講義スライドから特徴語を抽出するうえで，以下の 2 つの条件を満たすような方法を用いて実装する．

条件 1　スライドのページごとの特徴語を抽出する．

条件 2　専門用語に多い複合名詞の優先度を高くする．

この条件を満たすような特徴語抽出の方法として，Okapi-BM25 と MC-value を使用する．以下に特徴語抽出するまでの手順と特徴語抽出の方法についての詳細を示す．

5.2.1 名詞の抽出

形態素解析を用いて，スライドの本文から名詞を抽出する．本研究では，MeCab を用いて形態素解析を行う．形態素解析とは自然言語を形態素まで分割する技術のことである．取得したスライド本文を形態素解析し，名詞と解析された語句を取り出すことで名詞を抽出する．形態素解析結果の例を図 6 に示す．

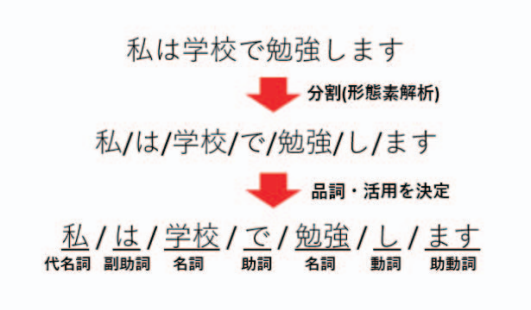

図 6 形態素解析結果の例

5.2.2 複合名詞の作成

形態素解析をした結果からそのまま名詞を取り出してしまうと，本来複合名詞だった名詞も別々の名詞として取得されてしまう．そこで，形態素解析をして得られた結果で，名詞が連続していた場合，名詞の連結を行う．例として，「情報セキュリティ」を形態素解析すると「情報」: 名詞，「セキュリティ」: 名詞といったように別々に形態素解析されてしまうが，隣り合う名詞を連結することにより，「情報セキュリティ」として取得することができる．また，3 つ以上の名詞が連続していた場合は「2 つの名詞を連結した名詞」，「3 つの名詞を連結した名詞」とすべてのパターンで名詞の連結を行い，複合名詞を作成する．

5.2.3 特徴語抽出手法

・Okapi-BM25

Okapi-BM25 は，文書検索において，複数のキーワードで構成される検索クエリに対して，文書のスコアを計算する手法である．文書 D とキーワード q とによる Okapi-BM25 のスコアは，以下の式によって算出される．

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

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

- f(q,D)：文書 D におけるキーワード q の出現頻度
- n(q)：q を含む文書数
- ｜ D ｜：文書 D の長さ
- avgdl：全文書における平均長
- N：全文書数

定数は，k=2.0，b=0.75 とした．本研究では，講義スライドを文書 D とし，キーワード q についてスライドの各ページの特徴語を score(D，q) で求めることにする．Okapi-BM25 を用いることによって，特徴語抽出の条件 1 であるスライドのページごとに特徴語を抽出することが可能になる．

・MC-value

MC-value は，複合名詞の長さや構成要素によってスコアを計算する関数である．複合名詞 a による MC-value のスコアは以下の式によって算出される．

$$MC-value(a) = length(a) \times (n(a) - \frac{t(a)}{c(a)}) \quad (3)$$

- a：複合名詞
- length(a)：a の長さ (構成単名詞数)
- n(a)：コーパスにおける a の出現回数
- t(a)：a を含むより長い複合名詞の出現回数
- c(a)：a を含むより長い複合名詞の異なり数

MC-value を用いることによって，特徴語抽出の条件 2 の専門用語に多い複合名詞の優先度を高くすることが可能になる．

5.3 講義スライド外の語句追加機能の実現法

講義スライド外の語句追加機能では，作成する Linked Data の情報不足を補うために講義スライドにはない語句を追加する．追加する語句の条件としては，講義スライド内の語句 2 つ以上から Linked Open Data のトリプル構造でつながっている場合とした．その語句がトリプル構造の「主語」であるか「目的語」であるかはどちらでも良いものとした．

5.4 アノテーション付与機能の実現法

特徴語抽出によって得られた語句にアノテーションを付与する．アノテーションとはデータに対して関連する情報を注釈として付与することである．そこで，アノテーションとして固有表現タグの付与と DBpedia Japanese から得られる特徴語の概要の付与を行う．

5.4.1 固有表現タグの付与

固有表現タグの付与では，NamedEntityAPI を用いる．NamedEntityAPI とは膨大な量の文章を教師データとし、文章の中の単語・品詞の並び・関係性などを機械学習することで、ユーザが入力した文章内の固有表現を予測する機能を提供する API である．API に取得したスライドの本文を送ることで，語句ごとに固有表現タグが付与される．付与されるタグは表 1 のようになる．

表 1 固有表現タグと例

固有表現タグ	例
ART(固有物名)	Windows10，ノーベル文学賞
LOC(地名)	アメリカ，千葉県
ORG(組織)	自民党，NHK
PSN(人名)	安倍晋三，メルケル
DAT(日付)	1 月 29 日，2020/1/29
TIM(時間)	午後三時，10:45
MNY(金額)	241 円，8 ドル
PNT(割合)	10%，3 割
O(固有表現に該当しないもの)	私，は

5.4.2 DBpedia Japanese から得られる特徴語の概要の付与

Linked Open Data である DBpedia Japanese から特徴

語の概要を取得し，アノテーションとして付与する．RDF 用のデータベースからトリプル構造のデータを検索して表形式でデータを取得する．データの取得の仕方と得られる結果の例を図 7 に示す．図 7 は，トリプル構造の「主語」の部分に「名古屋市」の URI を入力することによって得られた結果の例である．その結果として，都道府県が愛知県であるというデータや，隣接自治体が瀬戸市や豊明市であるというデータが取得できる．

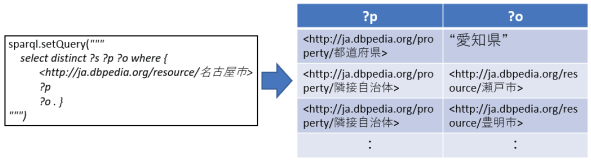

図 7 SPARQL を用いて得られる結果の例

SPARQL を用いて特徴語の概要を取得する場合は，「主語」の部分に概要を取得したい特徴語の URI を入力し，「述語」の部分に「概要」を表す「abstract」の URI を入力することで概要の取得が可能になる．

5.5 リンクデータ付与機能の実現法

リンクデータ付与機能では，SPARQL を用いて抽出した特徴語間にリンクデータを付与する．トリプル構造の「主語」と「目的語」の部分に抽出した特徴語の URI を入力して，DBpedia Japanese にその語句間のリンクデータがあった場合に「述語」の部分をリンクデータとして得る．そこで得られた結果を特徴語間にリンクデータとして付与する．リンクデータとして得られるデータには Wikipedia のリンクがあることを示す「wikiPageWikiLink」や「主語」がどんな項目に属するかを示す「subject」などが存在する．

6. プロトタイプシステム

システムの開発は，Windows10 上で Eclipse の環境の下で行い，プログラミング言語として Java と Python を使用した．3 章で提案した機能で取得したデータを用いて Linked Data を生成する．Linked Data を生成するツールとして，オントロジー開発ツールである法造を使用した．作成した Linked Data の全容を図 8 に示す．法造を用いて概念を表すノードと概念の一般-特殊関係を表す is-a 関係を描画している．

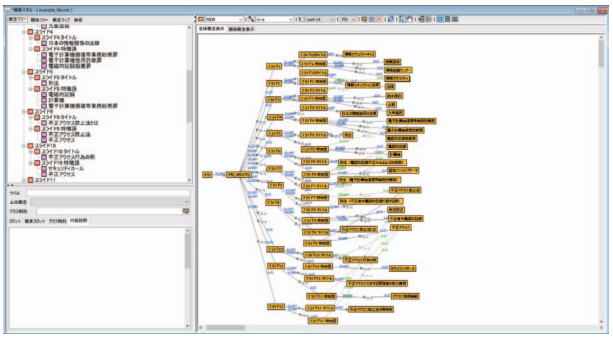

図 8 生成した Linekd Data の全容

7. 評価実験

7.1 評価実験 1(特徴語抽出手法の推定精度)

7.1.1 実験の目的

実験の目的は，提案手法に基づいて実装した特徴語抽出機能の有用性を検証することである．本研究では，「Okapi-BM25 と MC-value を組み合わせた手法」を用いたが，比較する特徴語抽出の手法として「TF-IDF」と「Okapi-BM25」を用いた．

7.1.2 実験方法

スライドの内容をよく理解している被験者 (本学学生 5 人) に講義スライドを見てもらい，スライドの中から重要であると思われる語句を選択してもらったのちに，特徴語抽出を用いて算出された語句と比較し，適合率と再現率を算出する．被験者に語句を選んでもらう際には，スライドのページごとに重要であると思われる語句を選択してもらい，スライドの 1 ページごとに 1 つ以上の語句を選択してもらうこととした．そして，スライドの内容の違いによる比較をするために，講義スライドを 2 つ用意して実験を行った．また，特徴語抽出をして得られたデータから特徴語として取り出す語句の個数は被験者が選んだ語句の個数と同じ個数とする．

再現率と適合率の算出方法を表 2 と式 4，式 5 に示す．再現率は式 4 で表され，偽陽性を低く抑えたい場合に採用する指標である．適合率は式 5 で表され，偽陰性を低く抑えたい場合に採用する指標である．一般的に，再現率と適合率はトレードオフ関係にあるため，式 6 に表されるように，調和平均をとった f 値を指標として用いる．f 値が高ければ，精度の高い特徴語抽出の手法であるといえる．今回の評価実験では被験者 5 人による再現率，適合率，f 値が算出されるので，5 人の再現率，適合率，f 値の平均の値を用いることとする．

表 2 予測と正解の正負による分類

	正解で正	正解で負
予測で正	真陽性 (TP)	偽陽性 (FP)
予測で負	偽陰性 (FN)	真陰性 (TN)

$$\text{再現率} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{適合率} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{f 値} = \frac{2 \cdot \text{適合率} \cdot \text{再現率}}{\text{適合率} + \text{再現率}} \tag{6}$$

7.1.3 結果と考察

評価実験 1 では，特徴語抽出で取り出す語句の個数が被験者が選んだ語句の個数と等しく，再現率と適合率が等しくなるので，f 値のみ算出する．1 つ目のスライドの f 値のグラフを図 9 に示し，2 つ目のスライドの f 値のグラフを図 10 に示す．

スライド 1 とスライド 2 の両方で「TF-IDF」，「Okapi-BM25」，「Okapi-BM25 と MC-value 組み合わせた手法」の順に f 値が高くなっている事がわかる．この結果から本研究で用いた「Okapi-BM25 と MC-value 組み合わせた手法」は有効であることがわかる．また，複合名詞を考慮していない「TF-IDF」と「Okapi-BM25」を比べると「Okapi-BM25」のほうが f 値が大きくなっていることから，スライド 1 ページあたりの語句数が少ないほど語句の重要度が上がる手法が有効であることがわかる．これは，スライドの形式を考慮すると，単語数の少ないスライドは文章ではなく語句だけの場合が多く，語句だけの形式のスライドは重要な語句のまとめといった重要スライドが多いので，「Okapi-BM25」が有効であったと考えられる．

次に「Okapi-BM25 と MC-value を組み合わせた手法」の f 値が最も高くなっていることから，特徴語抽出をする際に定義した条件 2 の「専門用語に多い複合名詞の優先度を高くする」という条件が有効であったと考えられる．ただ，実際に特徴語抽出したデータを見てみると，複合名詞ではない語句でも重要な語句は存在していることから，「MC-value」以外の部分で複合名詞ではないが重要である語句を抽出できる手法が必要になってくると考えられる．

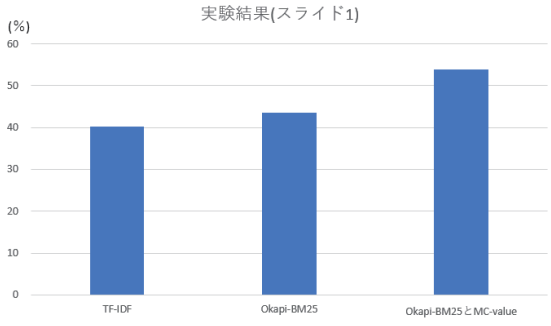

図 9 評価実験 1 の結果 (スライド 1)

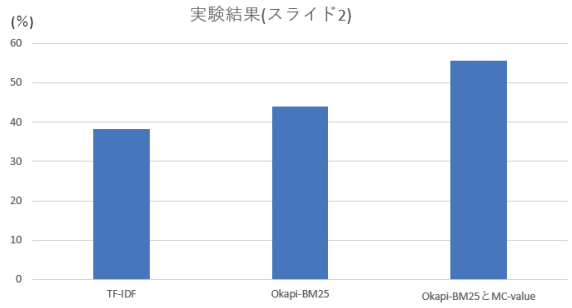

図 10 評価実験 1 の結果 (スライド 2)

7.2 評価実験 2(抽出する特徴語の個数の推定)

7.2.1 実験の目的

評価実験 1 では，特徴語抽出をして得られたデータから特徴語として取り出す語句の個数は被験者が選んだ語句の個数と同じ個数という理想の条件だった．ただ，本来は特徴語抽出によって算出された値を用いて特徴語を抽出しなけらばならない．そこで，閾値を変化させて，どの閾値が適切かを推定する．

7.2.2 実験方法

使用したスライドや被験者によって選ばれた語句は評価実験 1 で用いたデータと同じものを使用する．そのデータを用いて，評価実験 1 で最も有効であると算出された「Okapi-BM25 と MC-value 組み合わせた手法」を用いて特徴語を抽出する個数の推定を行う．閾値は 0．1 から 1 まで 0．1 刻みで増加させていき，閾値以上の語句を特徴語として抽出する．抽出された特徴語と被験者によって選択された語句を用いて再現率，適合率，f 値を算出する．

7.2.3 結果と考察

2 つのスライドを用いて，閾値を変化させた場合の再現率，適合率，f 値をそれぞれ図 11 と図 12 に示す．

スライド 1 では閾値が 0．3 の時に f 値が最大になり，スライド 2 では閾値が 0．2 の時に f 値が最大になった．スライド 1 では，閾値を 0．2〜0．6 まで変化させても f 値の変化が 1%以内に収まっているのに対し，スライド 2 では閾値が 0．2 の場合とそれ以外の場合で f 値の差が 3%以上あることから，今回用いた講義スライドでは閾値として 0．2 が適しているのではないかと考えられる．ただ，理想の個数を特徴語として抽出した場合と比較すると，スライド 1 で f 値が 12%ほど減少しており，スライド 2 では f 値が 26%ほど減少している．このことから，まだ理想に近い数の特徴語を抽出できていないことがわかるので，閾値を固定の数値にするのではなく，算出する方法も検討するべきではないかと考えられる．また，スライド 2 の f 値の減少がスライド 1 の f 値の減少よりも大幅に多いことがわかる．これは，スライド 2 のほうがスライド全体の語句の数が少なく，特徴語抽出する際に情報不足だったのではない

かと考えられる．この結果を踏まえて，スライドの内の語句の数が少なくても精度の良い結果が得られる手法を考える必要があると考えられる．

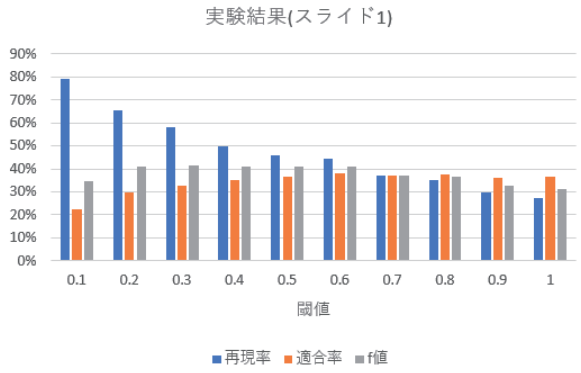

図 11　評価実験 2 の結果 (スライド 1)

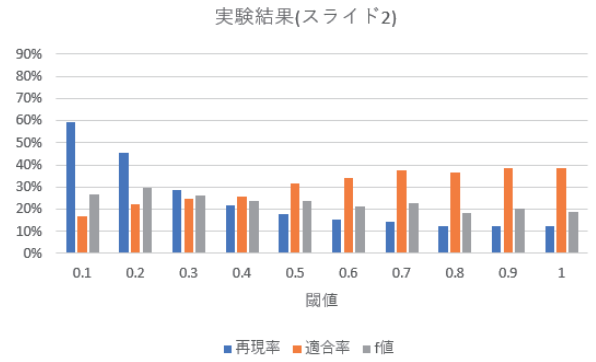

図 12　評価実験 2 の結果 (スライド 2)

8. おわりに

本研究では，俯瞰的な問題を生成するための講義スライドのアノテーション付き Linked Data 生成システムを提案した．その機能はスライド取得，特徴語抽出，アノテーション付与，リンクデータ付与を用いて実現した．また，提案システムに基づいてプロトタイプシステムを実装し，特徴語抽出に関する評価実験を行った．

評価実験 1 では，提案手法に基づいて実装した特徴語抽出機能の有用性を本研究で提案した「Okapi-BM25 と MC-value を組み合わせた手法」と特徴語抽出の方法として一般的に使用されている「TF-IDF」と「Okapi-BM25」との比較を行うことで検証した．その結果，本研究で提案した「Okapi-BM25 と MC-value を組み合わせた手法」が 3 つの手法の中では最も有用であることが分かった．また，複合名詞を考慮する場合としない場合で f 値に 10%以上の差が生まれたことから，複合名詞の優先度を高くすることが特に有用であると分かった．

評価実験 2 では，特徴語抽出をする際の閾値を変化させて，どの閾値が適切かを推定した．その結果，閾値を 0．2 にすることで推定精度が高くなることが分かった．ただ，理想の個数を抽出した場合と比較すると精度に差があり改善が必要となる結果となった．

今後の課題としては，様々なスライド形式の対応が挙げられる．今回用いた講義スライドは文章量の多いものを選んだので，幅広い形式のスライドでも対応できるように改善していく必要がある．また，今後俯瞰的な問題を生成する際に，今回生成した Linked Data では不足している部分が出てくると予測されるので，その際に Linked Data の情報を追加・修正することも今後の課題として挙げられる．

参考文献

[1] 教育再生会議：第七次提言 https://www.kantei.go.jp/jp/singi/kyouikusaisei/pdf/dai7_1.pdf（参照 2021-05-01）

[2] 池上真人:多肢選択文法問題の設問形式に関する研究:択一式と複数選択式の解答プロセスに焦点をあてて，言語文化研究，Vol. 35， No. 1， pp. 55–72， 2015

[3] B ¨ uhmann， L.， Usbeck， R. and Ngomo， A. -C. N. :ASSESS—Automatic Self-Assessment Using LinkedData，International Semantic Web Conference， pp. 76–89， 2015

[4] Linked Data を用いた俯瞰的な多肢選択式問題自動生成手法の提案，情報処理学会論文誌， 60 巻 10 号 pp. 1738–1756， 2015

[5] Jouault， C.， Seta， K. and Hayashi， Y. : Content-Dependent Question Generation using LOD for HistoryLearning in Open Learning Space， New GenerationComputing，Vol. 34，No. 4，pp. 367–394 (2016).

[6] Fionda， V. and Pirr‘o， G. : Meta Structures in KnowledgeGraphs， International Semantic Web Conference，pp. 296–312， Springer (2017).

[7] Maillot，P.，Raimbault，T.，Genest，D. and Loiseau，S. : Targeted Linked-Data Extractor， Proc. 6th InternationalConference on Agents and ArtificialIntelligence—Volume 1: ICAART， INSTICC，pp. 336–341，SciTePress (online)， DOI: 10. 5220/0004758503360341 (2014).