説明の難易度と講義内容との一致度を考慮した講義用語検索

竹村 直規[1]　　小林 伸行[2]　　椎名 広光[3]
[1] 岡山理科大学大学院 総合情報研究科 情報科学専攻
[2] 山陽学園大学 総合人間学部 生活心理学科
[3] 岡山理科大学 総合情報学部 情報科学科
i15im03tn@ous.jp[1], koba_nob@sguc.ac.jp[2], shiina@mis.ous.ac.jp[3]

1　はじめに

近年，日本への外国人留学生は増加傾向にある [1]．海外から日本へ留学するにはある程度の日本語能力が必要となる．この日本語能力の証明として日本語留学試験（EJU）[2]，日本語能力試験（JLPT）[3] といった試験が存在する．このような日本語試験の受講を入学条件としている大学も少なくない．大学教育においても，留学生向けの講義の工夫がなされおり，報告がなされている [4]．しかし，大学における専門的な知識を教授する講義では，日本語試験で出題されるような日常用語よりも，より特殊な用語を用いることが多い．そういった専門的な用語が頻出した場合，留学生によっては講義の理解が難しいと考えられる．

そこで，講義に連動して専門用語を検索する際に，日本語の理解も併せて考えるために母国語ではなく理解がしやすい日本語での検索結果の出力が必要であると考えられる．

一方，日本語の難易度に関連する研究としては，文章の難易度の推定 [5] や語彙の平易化システム [6] については，提案されている．

本研究では，日本語を母語としない学生に向けて講義に則した検索を行う際，Wikipedia[7] や WordNet[8]，Web 検索の結果を，説明文の難易度や講義との適合度を示す指標による順位付けを行うシステムを提案する．

2　使用したデータ

講義の単語を検索する検索データベースと，対象となる講義について述べる．

(1) 検索データベース

検索データベースとして，Wikipedia，Wikipedia Simple English 版 [9]，WordNet の 3 つを使用した．Wikipedia では曖昧性の低減を，WordNet では表記ゆれの補完や網羅性を高める．

(2) 講義データ

講義データは，岡山理科大学総合情報学部情報科学科で行われている講義のうち 1 年次開講科目の “プログラミング基礎” と 2 年次開講科目の “データベース” の VOD 講義の字幕データ及び講義資料を使用した．

3　講義用語検索の概要

検索システムとしては，検索表示機能，リーダビリティの判定，講義との適合度の項目からなっている．

(1) 検索表示機能 (図 1)

(1-1)Wikipedia の検索結果の提示: 同義語の場合は，#転送及び#REDIRECT 先の記事ページを検索結果とする．

(1-2)WordNet の検索結果の提示: WordNet では，単語にリンクされている複数の synset 全てを提示する．

(1-3)Wikipedia の simple english の提示: Wikipedia の simple english ページの記事を取得し，それを Microsoft Translator Text API[10] によって機械翻訳したものを表示する．

(1-4) 検索エンジンの結果: Microsoft が提供する Bing Search API[11] を使用した．本研究では検索結果の上位 5 件を表示する．

(2) 検索表示結果の難易度判定

検索結果については，意味情報の記述がよりやさしいか難しいかといった読みやすさの難易度の判定を決定木及び k 近傍法による機械学習によって行っている．

(3) 講義との適合度

複数の語義を持つ多義性を解消するために単語の意味情報を，講義で使われる単語から TF–IDF を用いた特徴抽出を行うことで，講義の内容に適合した表示

単語	読み	wikipedia	wordnet(synset)	SimpleEnglish
授業	じゅぎょう	授業(じゅぎょう、)は、学校教育などにおいて教科・科目などの教育として行われるもののことである。大学の課程(短期大学の課程、大学院の課程を含む)などにおける授業は、講義、演習、実験、実習、実技に区分される。	指示部隊 0.00115 学校の授業期間 0 技能か知識を分け与える 0 気付かせる 0	学校でのレッスンは、学習のための予定期間です。それは先生によって教えられている 1 つまたは複数の学生を含みます。レッスンは、いずれかの教科書の 1 つのセクションまたは短時間学習のテーマについて教えてまたはアクティビティを実行する方法を教えの中にあります。レッスンは教室で一般的に教えているが、代わりに学習環境で起こるかもしれない。
行動	こうどう	行動(こうどう、)は、人間を含む動物の活動や行い全般を指す言葉である。ただし、日本語の「行動」がもっぱら生物(とくに動物)に適用されるのに対し、英語の「behavio(u)r」は物体・機械など無生物の挙動・振舞いの意味で用いられることがある。	人々が行う、あるいは起こす事 0.00222 何かをすることを決める行為 0.00054 なされた事柄(通常、言われた事柄に対して言う) 0.00016 自分らしく振る舞う、または自分を管理する方法 0.00016 ある刺激に対して、ある生体があらゆる状況で示す反応や動作を総合したもの 0 ある態度で、振る舞う 0 ある方法で行動する 0 その人の任務を赦免する 0 政府または超国家の機関による行為 0 特定の方法で振る舞う 0 行動を実行する、うまく乗り切る、(行動を)遂行する 0	行動(または動作)動物または植物はどういう動作です。意識や無意識の行動があります。遺伝や学習があります。
小数点	しょうすうてん	小数点とは、整数の位と小数の位との境を表す記号である。	小数の左側にある点 0	桁区切り記号は、整数と 10 進数の数字の割合(数学)アル部分の間の境界線をマークするために使用のシンボルです。この記号は、期間をすることができます。またはコンマ、小数点、小数点コンマは小数点区切り文字に共通の名前であります。たとえば 9.5 は、英語圏の国で 9 と 1 つの半分を意味します。
基礎	きそ	基礎(きそ、)とは、構造物からの力を地盤に伝え、構造物を安全に支える機能をもつ構造である。下部構造(かぶこうぞう)とも呼ばれ、それに対して建築物本体を上部構造と呼ぶ。	何かが開始されるか、開発されるか、計算されるか、または説明される根本的な仮定 0.00049 他の真実を導き出すことが出来る原則 0.00039 何かを築く基礎となるもの 0.00025 臓器の連結している箇所に一番近い部位 0.0002 次のさらに高い桁の数と同値の正の整数 0.00019 ある構造の最も下の支えの部分 0.0001 支えまたは基礎 0 知識分野の基本を教育すること、あるいは教えること 0	基礎は、建物の最下部です。建物には、長い時間放置する場合は、強固な基盤が必要があります。
記録	きろく	記録(きろく)は、安定した形で定着・保存された状態にある情報である。会議の議事録が、会議の終わった後にも残るように会議中の発言を文書などの形にして保存するのがこの例である。	認識された業績の蓄積 0.00563 極端な知識 0.0056 詳細に記録する 0.00277 過去の出来事に関する永久的な証拠または情報を提供する(文書、レコード録音、写真のような)もの 0.00277 情報を提供する文書(特に公式自然の情報) 0.00069 何かまたは誰かに関する周知の事実の収集物 0.00064 取引の法的な証拠として機能できるドキュメント 0.00043 記録を作成する 0.0004 記録(特に音声記録)をする行為 0.00031 書面にする 0.00016	記録の意味: 情報又は情報の翻訳記憶媒体またはレコードに保存することをキャプチャします。
その他	そのほか	その他(そのた、そのほか)は、特定の事柄以外のものを一つにまとめて指す語。	追加的な不特定の寄せ集め 0	エトセトラ手段とラテン語で残りの部分。何が通常書き込むことができるより長いリストを続行する英語の言語でよく使用されます。人々 がよく書くなどと cetera.ごく稀に、それはまた書かれて &c ため、アンパサンドまたは &, et と同じです。それはまたのための記号と。一部の人々は、電気ショック療法としてそれを書くが、それは間違っています。

図 1: 講義用検索システムの表示例

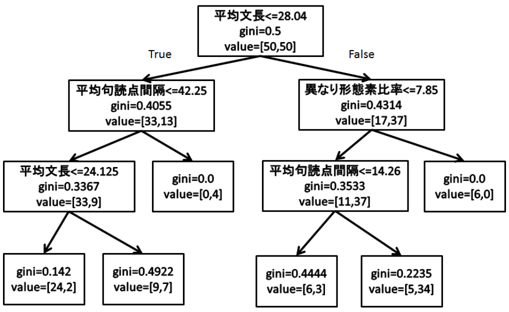

図 2: 決定木

も行う．また講義回ごとの講義との適合度の指標についても示す．

(4) 難易度と適合度を合わせた総合的な指標

WordNet と Web 検索の結果を (2)(3) と合わせて評価することで，総合的な指標を表す．

4 検索結果の難易度評価

検索を対象とする辞書データは，冗長な文，簡潔な文，単語だけの文など様々形態で記述されている．このような辞書データでは，用語の意味を捉えることが難しいと思われる．そこで辞書データの評価をすることで，辞書データが易しいか難しいかの判定を行う．

4.1 決定木学習による難易度判定

本研究では，決定木学習の CART アルゴリズムを利用して，平均文長，平均句読点間隔，平均ひらがなカタカナ比率，異なり形態素比率を項目から決定木を作成した．説明記事やさしいか難しいかの判断のためのに作成した決定木を図 2 に示す．

4.2 k 近傍法による難易度判定

k 近傍法 [12] は学習用のデータに含まれる情報をベクトル空間上にプロットしておき，そこから距離が近い順に任意の k 個を取得し，多数決でデータが属するクラスを推定するモデルである．本研究では，学習用データであるやさしいまたは難しい表現の学習データと講義用語の説明文との類似度に基づき，近傍に存在する学習データの最多となるクラスを難易度判定の結果としている．またやさしい表現の学習データにはやさしい日本語ニュース [13] を，難しい記述の学習データとして読売オンライン [14] のニュース記事を使用した．検索結果の説明記事の判定システムにおける流れを示す．

(1) 学習データを Mecab を用いて形態素解析を行う．

(2) 平均文長，平均句読点間隔，平均ひらがなカタカナ比率，異なり形態素比率，平均日本語能力試験級数，平均学習指標値 [15] を求める．

(3) (2) で求められたデータを教師データとして利用し，判定を行う．

(2) で求めた各学習指標の特徴を示す (図 3)．難易度の判定として特に上手く分類できているのは，平均ひらがなカタカナ比率，平均日本語能力試験級数，平均学習指標値である．今回の学習データである“やさしい日本語ニュース”では難解な漢字を極力使わず，なるべくやさしい語彙を使用しているため，難しい漢字や語彙を多様する“読売オンライン”の記事データとの明確な違いが表れたのだと考えられる．一方で，読みやすさの特徴である平均文長，平均句読点間隔はどちらのニュース記事でも重視されているため，分類が困難になっている．近傍値の正答率の変遷を図 4 に示す．簡易と難解に属する k 個の比のうち簡易の比を簡易クラス予測値とする．

図 3: 学習用データの各特徴量

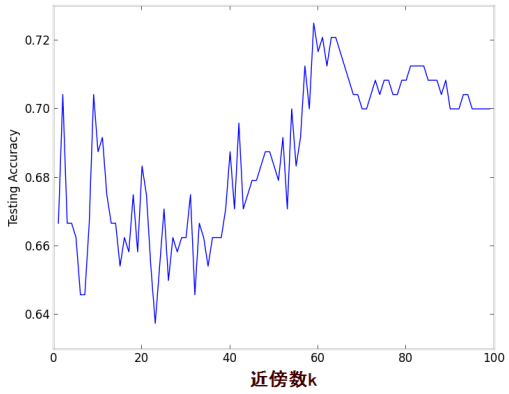

図 4: k 近傍法の近傍数

5 多義性の解消のための適合順の提示

語には多義性があり，Wikipedia，WordNet，検索エンジンを使用した意味情報も多義性に合わせた記述になっている．そこで，講義に連動させた意味の提示を行う．

例えば，データベースやプログラミングの講義でよく使われる単語 “配列” に対して WordNet では，10 個の意味を用意している (表 1)．このうちデータベースやプログラミングの講義では，音楽に関連する意味が含まれていることは不自然であるため，これらをそのまま提示するのは好ましくないと考えられる．そこで，講義に則して，意味の提示順を変えることとした．意味の提示順については，TF–IDF を利用して，適合度の指標を決めている．

5.1 適合度の処理手順

一科目の講義全体や各回の講義での単語の適合度を，TF と IDF から作られる TF–IDF を利用して求める．次にその手順を示す．

表 1: WordNet での単語” 配列”’ の表示結果

単語	記述	一致度
配列	整理や分類のための整った構造。	0.1798
	物事が論理的順番または連続するパターンで続く連続した配置関係。	0.0854
	ある順番に並べる。	0.0496
	順番に、論理的にあるいは、線上のように配列する。	0.0454
	グループの個々の要素を論理的にわかりやすくい配列したもの。	0.0389
	順序正しく置くこと。	0.0248
	人や物の生来の、あるいは後天的な気質、または特徴的な傾向。	0.0093
	物の置き方の空間的性質。	0.0049
	規則正しい取り決め。	0
	音楽作品を編曲する行為。	0

(1) 講義資料をテキスト形式に変換し，形態素に分割する．

(2) 対象文書での単語の比率を出現比率を TF として，ストップワードとしている単語の一部を除外した単語集合に対して計算する．

$$\mathrm{TF}(t,d) = \frac{\text{対象文書での検索語の出現回数}}{\text{対象文書の単語総数}}$$

ストップワードは，“もの”,“する”, “ある”,“こと”,“から”,“この”,“いる”,“ない”,“よう” のようなひらがなで，かつ名詞や動詞を除いている．

(3) 逆文書頻度 (IDF) を下記の式で計算する．

$$\mathrm{IDF(t)} = \log\left(\frac{\text{文書の総数}}{\text{単語 } t \text{ を含む文書数} + 1}\right)$$

(4)TF–IDF を，TF と IDF から求める．

$$\mathrm{TF\text{–}IDF}(t,d) = \mathrm{TF}(t,d) \cdot (\mathrm{IDF}(t) + 1)$$

(5) 適合度の計算

(5-1)WordNet の場合は，検索語の WordNet の同義語リンクに含まれている単語の TF–IDF の合計を，適合度とする． 検索結果としては，適合度の高い順に概念記述の高い順に記述する．

(5-2)Wikipedia，検索エンジンの場合は，記事概要部に現れる単語の TF–IDF の合計を適合度とする．

6 難易度と適合度を合わせた総合指標

検索結果の読み取りやすさと講義との適合度を合わせた総合的な順位付けを行う．総合的な指標には，第 **4** 章の k 近傍法による簡易クラス予測値と第 **5** 章の TF–IDF 値の重み α と β とする線形和を用い，講義で重要となる単語に対して人手で評価順と最も近くなるように重みを設定している．

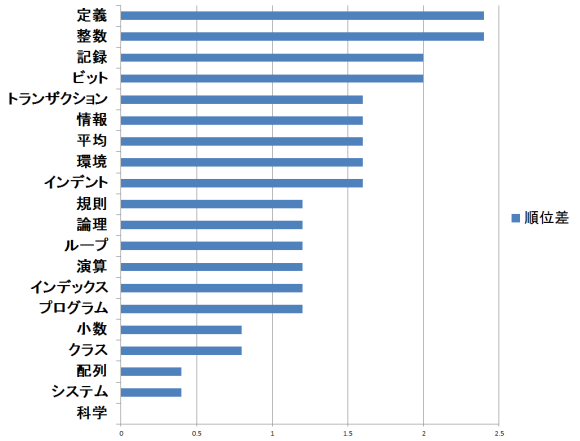

図 5: 検索結果の順位付け評価

$$\text{総合指標} = \alpha \cdot \text{簡易クラス予測値} + \beta \cdot \text{TF-IDF}$$

講義“プログラミング基礎”で使用された単語のうち 20 個を Bing Search による検索結果上位 5 件と人手による評価順位との差を図 5 に示す．講義で使用される単語全体の順位差平均差は 1.264 であった．“配列”，“プログラム”，“ループ”などのプログラミングの講義でよく使用される単語は順位差が小さくなっている．一方で，“定義”，“整数”，“記録”といったプログラミングの講義には出るが，特徴を上手く捉えられない単語は順位差が大きくなっている．今回の実験では，講義との適合度である TF–IDF 値の幅が大きいため，その部分の影響が大きくなっている．全体の順位差は小さくなったが，講義との適合度に大きく影響される結果になったのではないかと考えられる．

7 おわりに

本研究では，説明の難易度及び講義の適合度に即した専門用語の検索システムを開発した．留学生にもわかり易い意味情報を表示する方法として，決定木と k 近傍法を用いた難易度判定を行った．また，意味情報の剪定として，TF–IDF を用いた適合度により，講義の内容に合わせた検索結果の表示を行っている．最後に総合的な順位付けを行った結果，全体の順位差は抑えられたが一部の単語では上手く順位付けすることができなかった．しかし，そういった一部の上手く順位付けできなかった単語郡は講義との一致度が低いため，最終的な単語の表示結果の順番を変える指標になると考えられる．

参考文献

[1] 文部科学省，“留学生 30 万人計画の実現に向けた留学生の住環境支援の在り方に関する検討会報告書”， http://www.mext.go.jp/b_menu/houdou/26/08/__ics Files/afieldfile/2014/08/29/1350840_01_1.pdf

[2] 日本留学試験, http://www.jasso.go.jp/ryugaku/study_j/eju/index.html

[3] 日本語能力試験: http://www.jlpt.jp/index.html

[4] 古本裕子, 苗田 敏美, 松下美知子, 専門教育における留学生の日本語-日本人学生との比較を通した分析-, 金沢大学留学生センター紀要 9, pp.21–33, 2006.

[5] 近藤陽介, 松吉俊, 佐藤理史. 教科書コーパスを用いた日本語テキストの難易度推定. 言語処理学会第 14 回年次大会発表論文集, pp.1113–1116, 2008.

[6] 梶原 智之, 山本 和英, 語釈文を用いた小学生のための語彙平易化, 情報処理学会論文誌, Vol.56, No.3, pp.983–992, 2015.

[7] Wikipedia, https://en.wikipedia.org/wiki/Main_Page

[8] WordNet, https://wordnet.princeton.edu/

[9] Simple English Wikipadia, https://simple.wikipedia.org/wiki/Main_Page

[10] Microsoft Translator Text API, https:/www.microsoft.com/en-us/translator/translatorapi.aspx

[11] Bing Search API, http://azure.microsoft.com/ja-jp/services/cognitive-services/search

[12] Altman, N. S. (1992). “An introduction to kernel and nearest-neighbor nonparametric regression”. The American Statistician. 46 (3): pp. 175–185.

[13] やさしい日本語ニュース, http://www3.nhk.or.jp/news/easy/index.html

[14] 読売オンライン, http://www.yomiuri.co.jp

[15] 徳弘 康代, 日本語学習のためのよく使う漢字 2100, 三省堂, 2010.