2008－NL－184 (18)
2008／3／28

レファレンス記録に対するキーワードの自動付与

Automatic assignment of keywords to the reference records

○原田隆史(慶應義塾大学)，江藤正己(慶應義塾大学大学院)，瀬口真徳(慶應義塾大学)
ushi@slis.keio.ac.jp, eto@slis.keio.ac.jp

本研究の目的は，図書館におけるレファレンス記録の質問文および回答文から，そのレファレンス記録のキーワードとして適切な語句を自動的に抽出することである。自動抽出の手順は(1)候補となる語句を抽出し(2)キーワードとして適切か否かを判定するに大別できるが，本研究では(2)に焦点を当て，機械学習の手法を用いてこれを試みた。具体的には，キーワード候補となる語句の出現箇所，出現回数，語の前後の助詞など 42 個の特徴と人手に基づく正解判定の対を学習させ，自動判定をおこなった。日本史分野のレファレンス記録 507 件中の，人手によってキーワード判定を行った 9,375 語を対象とした実験の結果，精度 56.9%，再現率 48.3%でキーワードを正しく付与することができた。

The purpose of this research is to extract appropriate words automatically as keywords of the reference records from the sentences of questions and answers in the reference records. The automatic extraction method consists of two processes: (1) the process of extracting candidate words from given texts, and (2) the process of judging whether each extracted candidate is appropriate or not as a keyword. We focus on the process of (2) using the machine learning approach. In this method as a training data, we have used 42 features such as where each candidate word was, how many times it appeared in the records, and the postpositional particles surrounding the candidate. We then conducted an evaluation experiment using 9,375 words judged to be keywords manually in 507 reference records of Japanese history. The result showed that the keywords were assigned successfully with precision ratio of 56.9% and recall ratio of 48.3%.

1. 図書館におけるレファレンスサービス

図書館におけるレファレンスサービスとは，何らかの情報(源)を求めている図書館利用者に対し，その必要とする情報ないし情報源を効率よく入手できるように図書館職員が援助するサービスであり[1]，主に所蔵の調査・書誌事項の調査・文献の紹介・事実調査などを行うものである。

このレファレンスサービスは，利用者とのコミュニケーションスキル・資料に対する習熟・探索する能力など図書館員に高度な能力を要求する業務であるとされる。そこで，レファレンスサービスにおいては，図書館員同士で知識と経験を共有するためにレファレンスの記録をとることが有効とされている[2]。また，このようなレファレンス記録を蓄積することは，1)その記録自身がレファレンスツールとなる，2)レファレンスマニュアルの作成に役立てることができる，3)司書の教育目的に使う，4)外部から図書館を評価するための判断材料としても用いることができる，などの効果も期待できる。

レファレンス記録を協同で保存しようとする取り組みが，国立国会図書館の主導によって始まった「レファレンス協同データベース」[3]である。レファレンス協同データベースは，全国の図書館で行われているレファレンスサービスの記録や，そこで蓄積された調べ方などに関する情報をデー

タベース化し，図書館におけるレファレンス業務や，一般の人々の情報検索に役立てることを目的とする協同事業であり[4]，開始直後から多くの図書館の賛同を得ている。2007年10月末で445館が参加，26,382件（うち，一般に公開されているのは15,632件）のレファレンス事例が集まったデータベースとなっている[5]。

レファレンス協同データベースでは，管理番号，質問，回答，回答プロセス，事前調査事項，日本十進分類法(NDC:Nippon Decimal Classification 以下「NDC」とする)の番号，参考資料，キーワード，照会先，寄与者，備考，事例作成日，解決／未解決，調査種別，内容種別，質問者区分，提供館，登録番号，登録日時，最終更新日時の20項目を記録しているが，このうち管理番号，質問，回答，登録番号，登録日時，最終更新日時以外の項目は全て任意の入力項目となっている。これは，各図書館におけるレファレンス記録のフォーマットの違いを考慮したためとされる。

任意入力とされた項目のうち，キーワードの項目は，レファレンス記録を探すための項目として重要であるだけでなく，レファレンス記録を提示する際に類似のものを効果的に表示する際の基本的なデータなどとしても有効である[4]。しかし，レファレンス協同データベースでは約4割のレファレンス記録にはキーワード項目が入力されていないのが現状である。

レファレンス記録に対するキーワードの付与法としては，レファレンス記録中の語句から適切なキーワードを抽出する抽出キーワード法と，レファレンス中には出現しない語句を付与する付与キーワード法がある。

本研究においては，図書館員によって付与されたキーワードもレファレンス記録中に出現する語句を抽出したものが多いことから抽出キーワード法を用いることとした。

2. キーワード付与実験の先行研究

日本語の文章からキーワードを自動抽出する取り組みは数十年前から行われている。もともとキーワード抽出の技術は，全文検索のデータベースが実現不可能だった時代において索引語を付与するためのものであったが，全文検索のデータベースが実現可能となった近年においてもキーワード抽出の手法を用いた各種実験が行われている。

福田らは話し言葉を対象に，その出現位置情報からキーワードを抽出する試みを行っている[6]。句読点など文書構造を示すものがないテキストから，ブロック毎の語句の出現傾向を利用してキーワードを抽出した結果，単純なTF-IDF重み付けの手法より有効にキーワードを抽出している。長町らは文字数が少ない文章に対して，類似した文書を結合することで文書を長く拡張した上でキーワードの抽出を行っている[7]。その結果，キーワード抽出の精度の向上および非一般的な分野に特化したキーワードの抽出に成功している。また，村上らはキーワードの抽出において複合語に含まれる時系列の情報を重み付けの材料として使うことの有用性を説いている[8]。

レファレンス記録を対象とした研究としては，原田らによるレファレンスデータに対するNDCの自動付与がある[9]。この実験では，あらかじめNDCが2桁以上割り振られているレファレンスデータ6,337件から名詞および未知語を抜き出し，それを元に決定木・ナイーブベイズ・SVM(Support Vector Machine)の3種類の機械学習による自動分類を行っている。原田らは，レファレンス記録の主題を示す記号であるNDCの1桁目(類項目)を自動分類する実験と，類項目が付与されたデータを元にして2桁目(綱項目)を付与する実験の2つを実施している。前者の実験においては，判定方法別ではSVMが一番高い精度と再現率を得ることができ，類項目別では7類(芸術.美術)や9類(文学)では6割を超える再現率を出すことに成功している。また，後者の実験ではどの判定方法を用いた場合においても5〜6割の精度と3割前後の再現率を示したとしている。

3. レファレンス記録のキーワード自動付与実験

本研究では，機械学習の手法を用いてキーワードの自動付与実験を行った。すなわち，1)質問文および回答文からキーワードの候補となる語(語

句)をレファレンス記録から抽出した後，2)キーワードとなる語が質問文中や回答文中においてどのような特徴を持っているか学習させ，3)その学習結果を用いてキーワードであるか否かの自動判定を行って，どれだけの語が正しく判定されたかをチェックするという手順である。

本研究は，2007年7月19日時点でレファレンス協同データベース上にて一般公開されていた全てのレファレンス記録14,524件を収集し，そこから日本史分野(NDC210番台)のレファレンス記録で，かつキーワードが1個以上付与されているレファレンス記録507件を実験対象とした。

3.1 キーワードの候補となる語の抽出

まず，レファレンス記録の質問文および回答文からキーワードの候補となる可能性のある語句(以下，「候補語」とする)を抽出する作業を行った。本研究においては東京大学の中川らが開発した専門用語自動抽出システムTermExtract[10]を用いた。TermExtractは形態素解析を行って文章を単語レベルに分割した後に，基本的には名詞が連続で出現した場合は語を統合するなどといった，語の並びと品詞情報を用いて複合語を生成して出力する機能を有している。東京都立図書館が公開しているレファレンス記録のキーワードにも見られるように[11]，レファレンス記録に付与されるキーワードは単語を組み合わせて一つの概念を表すことが多いことから，このような複合語を出力することは本研究に有効であると考えられる。

ただし，TermExtractをそのまま用いた場合には，レファレンス協同データベース中のレファレンス記録に付与されているキーワードのうち，「川中島の戦い」や「生類憐れみの令」などのような漢字とひらがなが混じった語句を抽出できない。また，「群馬県新田郡尾島町長楽寺」のような漢字が連続して続く場合に正しく語を分割できないという問題もある。本研究では，これらの語についてTermExtractの標準辞書に追加して抽出できるようにした。

507件のレファレンス記録を対象として抽出をおこなった結果，候補語としては記録1件当たり平均18.5語，計9,375件が抽出された。

3.2 キーワードの正解判定

のべ9,375個の候補語に対して3人の判定者が人手で当該記録に対するキーワードとして妥当かどうかの正解判定を行った。個人の主観による判定の揺らぎを抑えて客観性を持たせるために，3人中2人以上がキーワードと判定した1,636語をキーワード，残りの7,739語を非キーワードと決定した。レファレンス記録1件あたりのキーワード数は3.23語である(第1表)。

第1表 実験対象とした語とキーワード

レファレンス記録数	のべ候補語数	のべキーワード数
507	9375 (平均18.5語)	1636 (平均3.23語)

3.3 機械学習によるキーワード自動判定

キーワードの自動付与は，レファレンス記録中の候補語の持つ特徴と人手に基づく正解判定の対を学習させ，自動判定をおこなった。機械学習システムは，ワイカト大学で開発されたWekaを使用し[12]，判定定器としては，決定木・ナイーブベイズ(以下，「ベイズ」とする)・SVMの3種類を用いた。

3.4 キーワード判定に用いる候補語の特徴

人がレファレンス記録の中からキーワードを抽出する際は，内容や文章の構造など様々な要素を総合的に捉えてキーワードかの否かの判定を行っている。

そこで本研究においても，レファレンス記録から抽出した候補語に関わる要素として，(1)固有名詞かどうか，NDC中で使用されている語かどうかなど抽出された語の特徴，(2)語がレファレンス記録中の質問文から抽出したものか，回答文から抽出したものかの語の抽出元，(3)語の出現回数，(4)語の前後の助詞などの表現，(5)語が各種の括弧内や前に記述されていたものかどうか，(6)「～を調べたい」や「～が知りたい」などのレファレンス記録に特有の表現，という6つの観点から，各語句に関して第2表に示す42個の特徴を設定し，これらと人手で判定した結果とを機械学習で用いるデータとした。

第2表 特徴語の持つ特徴の種類

要素の観点	属性	要素の観点	属性
[語を抽出した場所]	質問文から抽出	[括弧の有無]	「」の中
	回答文から抽出		『』の中
	質問文の冒頭部		()の中
	回答文の冒頭部		「」の前
[語の特徴]	固有名詞		『』の前
	NDCの語		()の前
[語の出現回数]	2回以上出現	[レファレンス特有の語の使用]	～とある
	3回以上出現		～にある
[語を区切る表現の使用]	「・」の前		～について
	「・」の後		～における
	句点の連続		～に関する
	～に		～を見たい
	～での		～を読み(たい)
	～が		～を調べたい
	～とは		～を知りたい
	～を		～を探し(たい)
	～の		～がある
	～は		～が見たい
	～で		～が読み(たい)
	「に」○○○		～が載っている
	「の」○○○		～の読み方

3. キーワード付与実験の結果

3.1 レファレンス記録全体を対象とした結果

レファレンス記録全体(質問文および回答文の両方)から抽出された9,375個(うち1,636語が人手によりキーワードと判定)の候補語を対象として，キーワードの付与を行った結果を，決定木・ベイズ・SVMの3種類の判定器それぞれについて第3表に示す。

なお，精度，再現率，F値は以下の式で算出した。

精度 = 正しく判定できたキーワード数 ÷ キーワードと判定された語数

再現率 = 正しく判定できたキーワード数 ÷ 人手でキーワードと判定された語数

F値 = 2×(精度×再現率)÷(精度＋再現率)

第3表 キーワード付与実験の結果

	決定木	ベイズ	SVM
語句の抽出場所	全体	全体	全体
付与できたキーワード数	999	1389	893
正しくキーワードと判定した数	634	790	536
誤って判定された数	365	599	357
キーワード付与の精度	63.5%	56.9%	60.0%
キーワード付与の再現率	38.8%	48.3%	32.8%
キーワード付与のF値	0.481	0.522	0.424

精度と再現率の調和平均であるF値は，一般的にトレードオフの関係にある精度と再現率の調和平均であり，F値が高ければ性能が良いことを意味する。

第3表に示すように，レファレンス記録507件全体から抽出された候補語9,375語のうち，判定器として決定木を用いた場合は999個，ベイズを用いた場合は1,369個，SVMを用いた場合は893個の語がキーワードと判定された。そのうち，決定木では634個，ベイズの場合は790個，SVMの場合は536個が正しい判定であった。したがって，精度はそれぞれ63.5%，56.9%，60.0%となる。

また，人手によって付与されたキーワードに対する再現率は，決定木の場合で38.8%，ベイズの場合で48.3%，SVMの場合には32.8%であった。

いずれの判定器の場合においても，精度は60%前後の値であったが，再現率は50%を超えなかった。

この原因のひとつとしては，「資料」「記述」「記録」「意味」「現在」といった頻出語の存在があげられる。これらの頻出語の多くは時制や行動などに関わる一般語であり，キーワードとはならないものが多い。実際に，出現頻度の多い上位34種類(のべ1,174語)は全て人手でキーワードとは判断されない語であった。

3種類の判定器による性能の比較では，再現率ではベイズを用いた場合が最も高い値となるが，精度は決定木を用いた場合が高い結果となった。SVMを用いた場合の値は，精度・再現率ともに決定木の結果を下回った。

精度と再現率のバランスという観点からF値をみると，ベイズ判定器を用いることが最も高い性能を示していると考えることができる。

3.2 質問文と回答文からのキーワード抽出

レファレンス記録は，質問文と回答文という2つの要素を含んでいる記録である。そこで，候補語がどちらから抽出されたかによって，どのような違いが生じるのかを調べるために，質問文から抽出した候補語と，回答文から抽出した候補語の2つに分けて機械学習による実験を行った。その結果を第4表に示す。第4表に示すように，決定木における精度をのぞき，精度，再現率，F値の全てが質問文中から抽出した候補語を対象とした場合が最も良い結果となった。

このように，質問文から抽出した候補語を用いてキーワード付与を行った場合には比較的，精度・再現率ともに高い値を得ることができた。しかし，回答文中の候補語を用いてキーワード付与を行った場合については，精度は大きな低下は見られなかったが，特に再現率において低下する結果が得られた。

質問文と回答文では差が生じた理由のひとつとしては，レファレンス記録における回答文と質問文の性格の違いが考えられる。すなわち，回答文には質問に対する直接的な回答だけではなく，関連する情報や直接回答には結び付かなかった途中経過などが含まれる。そのため，抽出される語句の数と比較してキーワードと判定される語句の数が少ないものが多い。

第4表　質問文および回答文に対するキーワードの付与結果

	決定木		ベイズ		SVM	
語句の抽出場所	質問文	回答文	質問文	回答文	質問文	回答文
抽出された語句の数	2,527	6,642	2,527	6,642	2,527	6,642
人手でキーワードと判定された数	1,068	1,105	1,068	1,105	1,068	1,105
付与できたキーワード数	936	570	1,026	1,046	758	877
正しくキーワードと判定した数	610	383	640	566	510	508
誤って判定された数	326	187	386	480	248	369
キーワード付与の精度	65.2%	67.2%	62.4%	54.1%	67.3%	57.9%
キーワード付与の再現率	57.1%	34.7%	59.9%	51.2%	47.8%	46.0%
キーワード付与のF値	0.474	0.347	0.481	0.422	0.426	0.404

このことは本研究の手法を用いた場合，非キーワードの割合がかなり多いような場合においては，自動判定の性能が低下する可能性があることを示しているとも考えられる。

4. レファレンス記録への効果的なキーワード付与にむけて

本研究では人手によって付与された1,636語のうち，ベイズ判定器を用いた場合に790語(48.3%)のキーワードを正しく付与できた。

しかし，残りの半数以上のキーワードを正しく付与するためには，設定した42の特徴のさらなる分析が必要となろう。具体的にどの特徴がキーワードの抽出を行う上で有効であったのかを分析し，機械学習で用いる特徴についても再検討を行う必要がある。

たとえば，前章でも述べたように出現頻度が高い候補語はキーワードとならないことが明らかとなっており，出現頻度に関してより細かな設定が必要である。

また，候補語を抽出する場所に関して本研究では「質問文や回答文の先頭部に位置する」という特徴を用いた。しかし，本研究で正しくキーワード判定ができなかった候補語の分析を行ったところ，文頭だけではなく，句点の直後の語が多く含まれることも見いだすことができた。単に質問文や回答文の先頭部というだけではなく，文頭に出現する語とするような変更も必要となろう。

さらに，本研究の前段階としてレファレンス記録から候補語を抽出する手法についても考慮していく必要があろう。たとえば，本研究では質問文と回答文からの候補語の選択は同じ基準で行ったが，違う基準を設定することなども検討する必要があろう。さらに，TermExtract自身の学習パラーメタ設定機能の使用なども行うなど，より効果的な候補語の選択を検討していくことが望ましいと考える。

5. 注・引用文献

1) 長澤雅男，レファレンスサービス:図書館における情報サービス．東京，丸善，1995，245p.
2) 斎藤文男，藤村せつ子．実践型レファレンスサービス入門．東京，日本図書館協会，2004，169p.
3) レファレンス協同データベース.[2007-12-8], <http://crd.ndl.go.jp/jp/public/>
4) 依田紀久．レファレンス協同データベース事業に見るデジタルレファレンスサービス．情報の科学と技術．Vol.56, No.3, p.90-95(2006)
5) レファレンス協同データベース事業累積統計(2007年10月末現在). [2007-12-8], <http://crd.ndl.go.jp/jp/library/documents/stats_200710.pdf>
6) 福田雅志，延澤志保，太原育夫．話し言葉における出現位置情報を用いたキーワード抽出．情報処理学会研究報告 音声言語情報処理．Vol. 2005, No. 50, p. 1-6(2005)
7) 長町健太，武田善行，梅村恭司．文書拡張によるキーワード抽出．電子情報通信学会技術研究報告 パターン認識・メディア理解．Vol. 105, No. 301, p. 17-22(2005)
8) 村上明子，渡辺日出雄．時系列情報を利用した複合語キーワードの抽出．情報処理学会研究報告 自然言語処理研究会報告．Vol. 2006, No. 36, p. 1-8(2006)
9) 原田隆史，江藤正己，大西美奈子．レファレンスデータに対するNDCの自動付与．情報知識学会誌．Vol. 17, No. 2, p. 61-64(2007)
10) 専門用語（キーワード）自動抽出用Perlモジュール"TermExtract"の解説. [2007-12-8], <http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html>
11) 東京都立図書館．レファレンス事例100. [2008-02-10], <http://www.library.metro.tokyo.jp/16/16520.html>
12) Weka 3 - Data Mining with Open Source Machine Learning Software in Java. [2007-12-8],<http://www.cs.waikato.ac.nz/ml/weka/>