意味的逆引き辞書『真言(まこと)』

粟飯原俊介† 長尾真‡ 田中久美子†
†九州大学大学院システム情報科学研究院 ‡京都大学名誉教授
{aihara, kumiko}@cl.ait.kyushu-u.ac.jp maknag@fm2.seikyou.ne.jp

1 はじめに

本稿では「意味的逆引き」辞書に関し、実装の第一段階と性能について報告する。

辞書の代表的な形態は、見出し語となる表現とそれに対する意味などの説明文による記述の組みが、一定の順序に並べられたものである。本稿では、意味や意義をユーザが入力すると、大元の見出し語を出力する事を「意味的逆引き」と定義する。たとえば、「天皇が死ぬこと」と入力すると「崩御」を得る事に相当する。

「逆引き」と称して、すでに出版物が多数ある。逆引きの意味はさまざまであり、単語を後方から引くもの [1][2] や、 意味がおおまかに分類されており、分類から大元の表現が得られるようになっているもの [3] などある。このような意味での逆引き場合には、印刷した出版物としての実現が可能である。一方、「意味的逆引き」の場合には、ユーザの入力はやや長めの表現であり、それはユーザが独自に考えるものであって、一意には定まらないことが前提である。このため、静的な出版物としての形態での実現は難しく、電子的なシステムとしての実現が合う。

「意味的逆引き」辞書の意義は、ある意味を的確に表現する短い単語を得る事に役立つ点である。誰しも、ある意味を思いうかべて、肝心の用語を度忘れした経験はあるだろう。たとえば、「磁針がほとんど南北を指す特性を利用し、船舶・航空機などで方位を測定する用具」(広辞苑) を日本語で何というであろうか。これは [4] の 中で、「度忘れ」する単語例として実際に挙げられているが、このようなもどかしい状態を解消する事に「意味的逆引き」は何より役に立つ。このような度忘れ以外にも、知的に短くわかりやすい表現を得る事や、広くは、母語だけではなく、外国語表現としての単語を得る事にも有用であろう。

「意味的逆引き」の実装の第一形態としては、辞書データを前提とすることがまず考えられるであろう。本稿では、そのような実装とその性能と限界を実験的に示した上で、今後「意味的逆引き」がどうあるべきかを議論する。

2 関連研究

「意味的逆引き」辞書は、ある長い表現を的確に表現する短い単語を探す事に相当することから、「言い換え」研究との関係が深い。言い換え研究の全貌は、たとえば [5] に、言い換えの定義、用途、認識、手法にわたり、すばらしくよくまとめられている。言い換え研究は、同じ意味の二表現間の変換に関する広い研究である中で、本研究は、人間のために言い変える研究でありながら、易しく言い換えるよりは、より短い一単語の表現、つまりどちらかというと、難易度の高い難しい表現に言い換える点に特徴があるといえる。

とはいえ、辞書データを前提とする範囲での本報告は、言い換え表現が辞書として所与であるため、言い換え研究との実際の関わりは、ユーザが入力した表現と、辞書項目の説明文との近さを測る程度の処理に留まっている。そして本報告では、この表現の類似性の計測は、検索エンジン技術を応用して行っている。一方で、辞書データを前提とせず、動的に辞書の見出し語相当の表現を獲得する方向で研究を展開するためには、ここに述べる関連研究が大切になる。

関連して、本研究の「短い単語表現を得る」との特徴の観点からは、transliteration 研究との関連もなくはない。ただし、transliteration 研究は異言語間で、ほぼ同音の単語を探し出すのが一般的であるのに対し、本研究は、同一言語内での変換が第一目標となるため、「言い換え」研究の中での位置付けが適切である。transliteration 研究も近年出版された [6] にこれまでの研究の全貌がまとまっており、短い単語表現を探しだす上では、このサーベイ中の技法が参考となる。また、本研究は、多言語への展開も将来的には可能であることから、過去のこれら知見は参考となる。

3　実装

3.1　手法

「意味的逆引き」辞書の第一歩として、本節においては、辞書データを前提とした手法をいくつか述べる。

まず用語の整理をする。既存の辞書は**項目**が決められた順序で並べられているものと捉える。項目には、**見出し語**が一つと、**付加情報**と**説明文** が含まれるものとする。付加情報は主として品詞や文法など見出し語の属性を示すものであり、説明文は主として意味や用例に関する成文での記述である。付加情報と説明文は一項目に複数含まれる場合がある。

辞書データを前提として「意味的逆引き」辞書を実現する上では、ユーザの入力に最も近い説明文を検出し、それに対する見出し語を表示する事が必要となる。この処理は、見出し語をページ、説明文を文書と見なした際の、検索エンジンに相当する。すなわち、各見出し語に対して、説明文に含まれる単語ベクトルを抽出し、検索のための単語文書行列を構成することで実現することができる。そこで、本研究では、単語文書行列に基づく単純なベクトル空間モデルによる検索 [7] をまず実装し、さらに同義語対を用いたクエリ拡張を試すことにした。

「意味的逆引き」では、説明文は数単語しかないものであるので、検索エンジンと比べて、より行列が sparse となることが予想される。しかも、求める対象も文書ではなく短い一表現と条件は厳しい。情報検索の分野ではさまざまに sparseness に取り組む手法が提案されているため、それらの効果を「意味的逆引き」で確かめることが必要となる。そこで、まざまな検索性能向上のための技術の中でも、代表的な LSI[8] を利用することにする。

LSI においてより高い効果を得るためには、辞書の説明文以外に、クエリ拡張と同様に類語辞典から得た見出し語に関する単語群を用いることができる。以下、これらのデータを説明する。

3.2　データ

入手しやすく、また「漢語」が豊富に含まれている三省堂大辞林の「意味的逆引き」の実現を目指した。大辞林は、第三版で大きく改訂された経緯があるが、その内容を、Epwing 形式を経由してテキスト形式で抽出する上では、第二版のみが入手可能であったため、これを用いた。辞書データは、項目に分解し、さらに項目を見出し、付加情報、説明文に分解する。結果、見出し語数 18 万 8086 語に対して、説明文が 24 万 4205 文抽出された。

図 1: 「意味的逆引き」辞書『真言(まこと)』のページと検索例

クエリ拡張ならびに LSI の性能向上には、類語辞典が必要となる。類語辞典としては講談社「類語大辞典」や国立国語研究所「分類語彙表」があるが、電子データでの入手性を鑑み、「分類語彙表」において同一の分類番号が付与されている単語を類語として用いた。クエリ拡張では、クエリ表現中の検索に用いる単語個々を類語辞典により拡張する。また、LSI においては、大辞林の説明文中の検索に用いる単語それぞれの類語を単語文書行列に追加する。

どの検索方式においても、単語文書行列が処理の基本となる。見出し語ならびに説明文中の単語を、形態素解析にかけて正規形を得て、自立語を取り出したところ、10 万 8329 単語の異なり語数を得た。そこから 24 万 4205 文 × 10 万 8329 単語のサイズの単語文書行列を疎行列の形で作成した。単語文書行列内の要素の数値は t df による重みとし、また LSI を用いた実験では、500 次元[1]を用いた。検索クエリに対する類似度はコサイン尺度を用いてランキングを行なった。

3.3　実行例

前提とする辞書データに著作権があるため、現在の実装は未公開であるが、「意味的逆引き」辞書は web ページを経由するインターフェースから利用することができる。一実装としてのシステム『真言(まこと)』を用いた検索例を図 1 に示す。

[1]LSI に用いる次元数としては、100 次元、500 次元、1000 次元、5000 次元での性能の比較を行ったところ、次元数が増加するにしたがって性能も単調増加したが、速度の面から 500 次元を用いている。

4 性能評価

4.1 テストデータ

構築された「意味的逆引き」辞書の性能を評価する上では、テストデータが必要となる。テストデータとしては、見出し語と説明文が対となっているものが望ましい。本システムは、人間が用いる事を前提とするため、そのような対の集合は、本来ユーザから集めるのが妥当である。しかし、ここでは、近似的に別の電子辞書を利用して、大規模な評価を行うことにした。

無論、別辞書を用いた辞書の評価は、辞書間の乖離度合いをもって「意味的逆引き」の性能を測っていることに相当するともいえ、適切な評価となっているかは疑問である。しかし、大規模評価を行う上では別辞書を用いる事がまず思いつくことでもあり、見出し語がそもそも存在しない割合も算出することができるという利点がある。

このため、本稿では Epwing 形式で得られる広辞苑第 6 版から、見出し語と、その第一説明文のペアを得て、これをもとに評価を行うものとした。広辞苑第六版の見出し語の中から、表記ゆれが無く対応が取りやすい語として漢字二字で構成された二漢字語をテストデータとして用いた。 その総数は 85856 語であった。うち、大辞林第二版に存在する見出し語の総数は 67410 語、すなわち全入力の 78.5%が得られるべき状況にある。検索される大辞林の説明文の平均単語数は 10.9 語であるのに対し、テストデータとしてクエリとなる広辞苑の説明文の平均単語数は 10.1 語、これをクエリ拡張した際のクエリの平均単語数は 48.4 語である。

4.2 実験結果

広辞苑の二漢字語の説明文を一つづつ「真言」に入力し、挙がった上位N件の候補に対して、情報検索の文脈上一般的な以下の 3 つの評価値を算出した。

Recall-N: 上位 N 件中に求める見出し語が存在する割合 (図 2)

Precision-N: 見出し語が存在する条件下で、検索結果が上位 N 件中に得られた割合 (図 3)

MRR: 見出し語が存在する条件下での Mean Reciprocal Ranking、すなわち求める見出し語が現れた順位の逆数の平均 (表 1)

表 1: MRR

	単語文書行列	クエリ拡張	LSI	LSI+類語
MRR	0.408	0.381	0.160	0.114

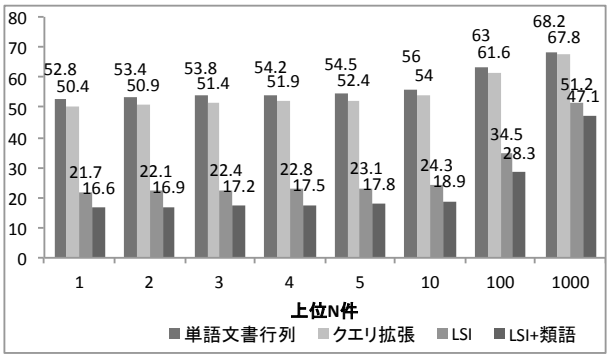

図 2: Recall(単位:%)

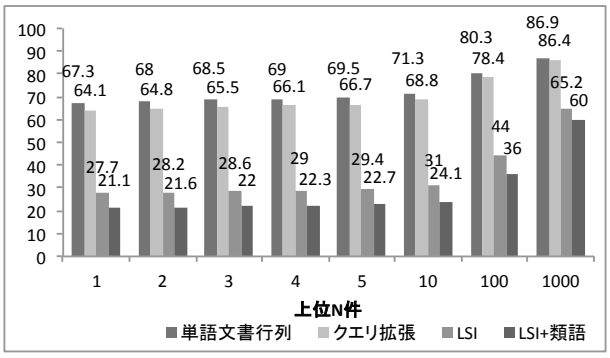

図 3: Precision(単位:%)

また、比較の為、同様のクエリを google で検索を行った場合の結果も報告する。広辞苑の二漢字語すべてに対して実験を行うことは現実的では無いため、広辞苑と大辞林の両方に存在する二漢字語の中でランダムに選択した 500 語を実験に利用した。google の検索結果上位 100 件中のタイトル、またはスニペット中に求める見出し語が存在したのは、全部で 70.8%であった[2]。

4.3 考察

実験結果から以下が読み取れる。

(1) 二漢字語の範囲とはいえ、辞書で辞書を評価した結果として、両方の辞書に存在する単語に限った場合、86.9%が検索された。

(2) MRR の結果を見ると、検索できたものに関しては上位数件以内に入っている。

(3) LSI やクエリ拡張等の工夫は一切しない単なる検索が性能が最も高い。

(1)、(2) については、前述のように、求める見出し語がそもそもない場合は、全入力の 21.5%である中で、全入力の 68.2%が上位 1000 件に入り、その MRR も悪いとは

[2]あくまで求める見出し語が文字列として存在した、ということを評価したものであり、ユーザが求める単語を見つけることが出来るかどうかや、正しく類似する説明文を見つけられるかは評価していない甘い評価である。

いえない状況である。google との結果 (70.8%) と最も良い Precision(86.9%) を比べても単純な実装であるとはいえ本研究の意義をある程度は示していると言えるだろう。

そして、大辞林に見出し語が存在する 78.5%と Recall の差分約 12%の大半は、「求める見出し語に対する入力として広辞苑の説明文が大辞林とまったく異なる説明文を付与している場合」に相当し、たとえば、「悪地」の説明文は、以下のように二つの辞書で異なる。

大辞林 地質や地形が悪く、植物の栽培や住宅の建設・交通などに適さない土地。

広辞苑 植生のとぼしい乾燥地域の軟弱な地層から成る緩傾斜地で、豪雨などの作用で多くの溝を生じ、通行が不便となったところ。バッドランド。

(3) は「意味的逆引き」の本質にかかわる問題である。工夫が効を奏さない具体的理由は以下となる。クエリ拡張の場合には、類語が単なるノイズとなり、Recall を向上させるどころか却って検索時の順位を下げる事が多くなる。また、LSI は本来的には sparseness を緩和する技術ではあるが、「意味的逆引き」では検索に関わる語数が非常に少ないために、特異値分解により不要情報を削減して関連性の高い情報を補完する効果が得られず、むしろ特異値分解を経て次元を削減すること自体が端的に情報の削減につながる結果に終わっている。すなわち、検索における文書相当量が極めて限定される、という特徴を「意味的逆引き」は持っている。

つまり、今後性能向上を目指すには、情報検索上研究されてきた別方式—たとえば、次元削減手法として pLSI を用いる— などの方法では、効果が得られないと予想される。このため、意味や語をよりピンポイントで処理する事を考えなければならない。例えば、単純にクエリ拡張を行うのではなく、類語の処理に構文情報を取り入れた文レベルの類義関係を扱う手法 [9] を用いるなどが考えられる。

このような問題の特性に加え、辞書データを前提とすると、語がそもそも辞書になかったということが 21.5%に上ること、辞書の著作権の問題、辞書とは多分に言葉の規範を示す国語という政治性を帯びた中にある紙媒体出版物であること [10]、などが、実用化に際しては今後妨げとなろう。静的な辞書データを前提としない、語に関する情報をコーパスを用いて動的に獲得するような「動的な辞書」についても検討していくべきであると考えられる。transliteration[6] や数多く提案されている用例検索などは動的辞書の一種と捉えられよう。

「意味的逆引き」は電子形式でなければ実装できない辞書である。辞書データを前提とする今回の方式がその未来形であるかは疑問である。「英辞郎」に対する賛否両論、身近な存在として江戸以前に庶民に用いられた「節用集」としての字引き、そして、語彙の網羅性など、紙媒体出版物としての辞書と対置される電子辞書がどうあるべきであるのか、その可能性の中に新しい形態の辞書を模索したいと考えている。

5 まとめと展望

本稿では、「意味的逆引き」辞書の、辞書データを前提とする一実装を行った。辞書の見出し語を文書、説明文を単語の列と捉え、一般的な情報検索の手法を用いて、性能を調査し、その限界を論じた。性能の向上の余地はまだ大きく、類義表現に関する精緻な処理を用いることで、完成度を高めていくことが出来るだろう。しかしながら、辞書データを前提とする以上、収録されている語彙数の問題や、著作権の問題が等が存在することが本質的な問題となりうる。

「意味的逆引き」辞書には、電子辞書が今後どうあるべきとの問いの中で、特に動的な辞書において大きな広がりが考えられる。数多く行われてきた言い換え研究などを参考に、これから「意味的逆引き」の研究を始めたい。

参考文献

[1] 北原保雄. 日本語逆引き辞典. 大修館書店, 1990.

[2] 郡司俊男. 英語逆引き辞典. 開文社出版, 1968.

[3] 井上宋雄. 例解 慣用句辞典—言いたい内容から逆引きできる. 創拓社, 1992.

[4] J Aitchison. *Words in the Mind.* Blackwell, 1994.

[5] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理, Vol. 11, No. 5, pp. 151–98, 2004.

[6] S. Karimi, F. Scholer, and A. Turpin. Machine transliteration survey. Vol. 43, pp. 17–62, 2011.

[7] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Morgan Kaufmann, 1999.

[8] S. Deerwester, S. Dumais, G. Furnas, K. Landauer, and R. Harshman. *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.

[9] K. Shinzato, T. Shibata, D. Kawahara, and S. Kurohashi. Tsubaki: An open search engine infrastructure for developing information access methodology. *Journal of information processing*, Vol. 20, No. 1, pp. 216–227, 2012.

[10] 安田敏朗. 辞書の政治学. 平凡社, 2006.