B0TB2238

卒業論文

大規模文書集合から学習したベクトル空間モデルによる辞書の意味的逆引き

山田良介

2014年3月31日

東北大学
工学部 情報知能システム総合学科

大規模文書集合から学習したベクトル空間モデルによる辞書の意味的逆引き*

山田良介

内容梗概

近年のブログやマイクロブログの普及に伴い，文書による個人の情報発信が増加している．文書作成において，的確な語を選択することができれば，文書をより短く簡潔に表現することが可能となる．本稿では，個人の文書作成支援を目的とし，意味を与えて単語を検索する「意味的逆引き」を，単語ベクトルを用いたベクトル空間モデルにより，既存辞書への検索として実装した．また既存辞書の情報の少なさを補うべく，大規模文書集合の導入を行った．

キーワード

検索，単語ベクトル，ベクトル空間モデル，大規模文書集合

*東北大学 工学部 情報知能システム総合学科 卒業論文, B0TB2238, 2014年3月31日.

目 次

1　序論

1.1　背景

レポート，報告書，メールなど，日常生活の上で文書を書かなければならない場面は多い．近年ではブログや，Twitter などに代表されるマイクロブログの発展により，個人が文書により情報発信を行う場面はますます増加している．

一方で「若者の語彙力低下」などいうことが叫ばれることがある．実際に若者の使える語彙数が減少しているかはここでは議論しないが，多様な語を使用できれば便利である．例えば

(i) 若者の知っている単語の数が低下

という表現よりも

(ii) 若者の語彙力低下

という表現のほうが，伝えたいことをより簡潔に，端的に示すことができる．これは，文字数の限られるマイクロブログなどでは特に有効であると考えられる．

しかし，このように意味を端的に表す語を誰しもすぐに思い浮かべられるわけではない．意味は分かっているにもかかわらず単語を度忘れすることもあれば，端的に表す単語を知らないこともある．このような時に意味を入力とし，適切な単語を推薦するシステムがあれば，頭の中のもやもやを解消して，簡潔明瞭な文書の作成を支援することができる．このように意味が与えられ，その意味を端的に表す単語を導くことを，本稿では後述する粟飯原ら [1] の手法にならい，意味的逆引きと呼称する．

1.2　実装形態

意味を表す文を入力，単語を出力とすると，意味的逆引きは文から単語への言い換えととらえることができる．言い換えの研究としては梶原ら [2] のような単語平易化があるが，これは単語から単語への変換を行う点で意味的逆引きと異なる．また平易化がより一般的な，使用頻度の高い語への変換を目的とするのに対

し，意味的逆引きは必ずしもそうなるとは限らず，逆により難しい，日常での使用回数が少ない語への変換となる場合もある．例えば (i) の文を (ii) のように書き換える場合を考えると，これは「知っている単語の数」から「語彙」，あるいは「語彙力」へ変換する意味的逆引きに相当する.「語彙」という単語は日常での使用頻度が高くなく，より難しい語への変換となっているといえる．

言い換えと異なる方法としては，既存辞書に対する検索として意味的逆引きを実装することが考えられる．既存辞書には単語の定義文が記述されているため，推薦された単語が本当にユーザーの意図したものなのか，ユーザーが容易に判断可能なことがこの実装形態の利点である．検索対象となる辞書を変更することにより，コンピュータ用語など特定ドメインの単語を推薦したり，日本語文から英単語を推薦したりすることも可能となる．

本稿は 5 章からなる．第 2 章では辞書に対する検索として意味的逆引きを実装した既存研究と，その問題点について述べる．第 3 章では既存研究の問題点を解消する提案手法として，大規模文書集合から学習したベクトル空間モデルについて述べる．第 4 章では評価実験の設定，評価方法，結果及び考察について述べる．第 5 章では本研究によって判明した点と，今後の課題について述べる．

2　関連研究

意味的逆引きは辞書を対象とし，その定義文に対して全文検索を行うことで実装が可能である．全文検索はクエリとして 1 つ以上の単語を与え，与えた単語との一致率が高い文書を検索するものである．意味的逆引きにおいては，辞書の各定義文を検索対象として検索を行う．

例として，次に示す辞書定義文を考える．

齟齬

意見や事柄がくいちがって，合わないこと．くいちがい．

このように定義される単語「齟齬」を検索するには,「事柄 くいちがう」などのクエリで検索を行えばよい．このような実装は Weblio 辞書 [3] などで試すことが可能である．

全文検索の手法を用いて意味的逆引きを行った研究として，粟飯原ら [1] は辞書中の定義文に含まれる自立語をもとに単語-文書行列を作成した．例えば上記の「齟齬」及び次に示す定義文を考える．

外れる

あるべきことや道筋と食い違う．

これらの定義文に対し形態素解析を行い，自立語の基本形を取り出すと次のようになる．

齟齬

意見 / 事柄 / くいちがう / 合う / くいちがい

外れる

ある / 道筋 / 食い違う

さらにこれらに対し，辞書見出し語を行，定義文に含まれる自立語を列とし，単語-文書行列を作成すると表 1 のようになる．

表 1: 単語文書行列の例

	意見	事柄	くいちがう	合う	くいちがい	ある	道筋	くいちがう
齟齬	0.37	0.32	0.56	0.36	0.56	0	0	0
外れる	0	0	0	0	0	0.20	0.60	0.78

行列内の値は tf-idf による重み付けを行い，1 行をベクトルとみなし正規化したものである．単語 i の文書 j における tf-idf は 1 式により定義される．

$$\begin{aligned} tfidf_{i,j} &= tf_{i,j} \cdot idf_i \\ tf_{i,j} &= \frac{n_{i,j}}{N_j} \\ idf_i &= \log(\frac{D}{d_i}) \end{aligned} \tag{1}$$

ただし，$n_{i,j}$ は文書 j 中における単語 i の出現回数，N_j は文書 j 中の総単語数，D は総文書数，d_i は単語 i を含む総文書数である．ここでの文書とは辞書中の各定義文を指す．重み付けを行うことにより，各自立語の重要度を反映することができる．tf-idf による重みづけでは，同じ文書に複数回登場する単語は重要度が高いとみなし，tf 項により重みが大きくなり，多くの文書に共通して出現する単語は重要度が低いとみなし，idf 項により重みが小さくなる．

また，ベクトル $\boldsymbol{v}$ に対する正規化ベクトル $normalize(\boldsymbol{v})$ は 2 式のように表される．

$$normalize(\boldsymbol{v}) = \frac{\boldsymbol{v}}{|\boldsymbol{v}|} \tag{2}$$

単語-文書行列の各行は各見出し語に対するベクトルとみなせるため，以後各行を定義文ベクトルと呼称する．

入力文からは形態素解析により自立語の基本形を取り出すことで検索クエリとする．例えば，

(iii) 意見や事柄がくいちがうこと

という文を与えた場合，形態素解析により

入力文

意見 / 事柄 / くいちがう

と自立語の基本形が取り出され，単語文書行列と同様に表 2 のようにベクトルの形で表現できる．以後このベクトルを入力文ベクトルと呼称する．

表 2: 入力例

	意見	事柄	くいちがう	合う	くいちがい	ある	道筋	くいちがう
入力文	0.49	0.43	0.75	0	0	0	0	0

入力文と定義文に同じ単語が含まれているほど，ベクトルは類似した値をとる．したがって入力文ベクトルと各定義文ベクトルをコサイン類似度により比較することで順位付けを行い，類似度の高い項目を出力する．ベクトル $\boldsymbol{a}, \boldsymbol{b}$ に対するコサイン類似度は次の式で表される．

$$Sim(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} \tag{3}$$

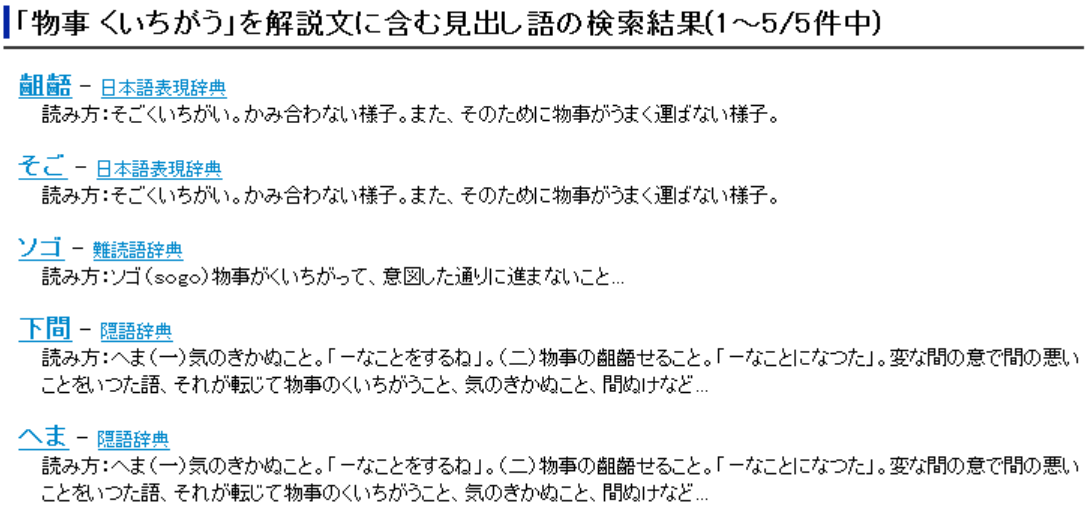
しかしながら，検索エンジン技術による手法では，文に含まれる表記ゆれや類義語の影響を強く受けるという問題がある．例えば，単語「齟齬」を検索することを意図し，以下の 3 通りのクエリを考える．

(iv) 物事 くいちがう

(v) 物事 食い違う

(vi) 事柄 くいちがう

これらのクエリをそれぞれ Weblio 辞書に対して入力した結果，上位 5 件は図 1 のようになった．



「物事 くいちがう」を解説文に含む見出し語の検索結果(1～5/5件中)

齟齬 – 日本語表現辞典
読み方:そごくいちがい。かみ合わない様子。また、そのために物事がうまく運ばない様子。

そご – 日本語表現辞典
読み方:そごくいちがい。かみ合わない様子。また、そのために物事がうまく運ばない様子。

ソゴ – 難読語辞典
読み方:ソゴ(sogo)物事がくいちがって、意図した通りに進まないこと...

下間 – 隠語辞典
読み方:へま(一)気のきかぬこと。「ーなことをするね」。(二)物事の齟齬せること。「ーなことになつた」。変な間の意で間の悪いことをいつた語、それが転じて物事のくいちがうこと、気のきかぬこと、間ぬけなど...

へま – 隠語辞典
読み方:へま(一)気のきかぬこと。「ーなことをするね」。(二)物事の齟齬せること。「ーなことになつた」。変な間の意で間の悪いことをいつた語、それが転じて物事のくいちがうこと、気のきかぬこと、間ぬけなど...

(a) 「物事 くいちがう」の場合

「物事 食い違う」を解説文に含む見出し語の検索結果(1～10/141件中)

齟齬 – Wiktionary日本語版〈日本語カテゴリ〉
出典:『Wiktionary』(2011/05/22 03:19 UTC 版) 名詞齟 齬 (そご) 物事がうまくかみ合わず、食い違ってうまく進まないこと。ゆきちがい。もし然らずしてこの二者の至り及ぶ...

妥協 – 百科事典
妥協(だきょう、英語 compromise)とは、何かの物事を進めるにあたって、関係する双方の意見が食い違い、そのままではそれ以上の進展が望めそうもないときに、いずれか一方が自身の意見を取り下げたり、...

Compromise – 百科事典
妥協(だきょう、英語 compromise)とは、何かの物事を進めるにあたって、関係する双方の意見が食い違い、そのままではそれ以上の進展が望めそうもないときに、いずれか一方が自身の意見を取り下げたり、...

イスカ – 百科事典
イスカイスカ(オス)保全状況評価LEAST CONCERN(IUCN Red List Ver.3.1 (2001))分類界:動物界 Animalia門:脊索動物門 Chordata亜門:脊椎動物亜門...

Red Crossbill – 百科事典
イスカイスカ(オス)保全状況評価LEAST CONCERN(IUCN Red List Ver.3.1 (2001))分類界:動物界 Animalia門:脊索動物門 Chordata亜門:脊椎動物亜門...

(b) 「物事 食い違う」の場合

「事柄 くいちがう」を解説文に含む見出し語の検索結果

Weblio辞書で「事柄 くいちがう」を解説文に含む見出し語は見つかりませんでした。

(c) 「事柄 くいちがう」の場合

図 1: Weblio 辞書に対する検索例

(iv) をクエリとした場合，日本語表現辞典の項目が検索されている．一方「くいちがう」の表記を漢字に変えた (v) の場合，日本語表現辞典の結果は出現せず，Wiktionary 日本語版の結果が検索された.「物事」を類義語である「事柄」に変えた (vi) の場合は，検索結果が 1 つもないという結果となった．このように，検索クエリ中の単語の綴りを変化させた場合，意味としては大きく差が無くとも，検索結果が大きく異なってしまう．

これは単語の綴りのみに着目して検索を行っていることが原因である.「くいちがう」と「食い違う」は同一の意味を表す単語であり，相互に関連がある．しかし既存手法では単語の綴りのみから語を判別するため,「くいちがう」と「食い違う」はまったく別の語とみなされる．したがって，(iv) をクエリとした場合には「くいちがう」を定義文に含む日本語表現辞典の項目が検索されたが，(v) をクエリとした場合には「食い違う」が定義文中に含まれないため，日本語表現辞典の項目は検索されなかった．類義語である「物事」と「事柄」に関しても同様である．

意味的逆引きを行う上で，ユーザーがどのような単語を用いて検索を行うかは事前に予測できないため，表記ゆれや類義語の差異により検索結果が大きく変わる，あるいは検索されない，といった状況は好ましくない．表記ゆれや類義語の差異を吸収し，柔軟な検索を行うためには,「くいちがう」と「食い違う」,「物事」と「事柄」を似た単語として扱うことができる，意味的類似性を考慮した検索を行う必要がある．

表記ゆれ及び類義語の問題に対応するため，粟飯原らは類語辞書を用いたクエリ拡張，Latent Semantic Indexing (LSI)[4] を用いたベクトルの次元削減を試しているが，どちらも精度を下げる結果となっている．特に LSI は精度を大きく下げており，その原因として粟飯原らは検索に関わる語数が少なく，次元の削減が端的に情報の削減に繋がったことを挙げている．

本稿では表記揺れや類義語へ対応するため，単語ベクトルを用いる手法を提案する．単語ベクトルは単語を固定長のベクトルとして表現する手法である．単語を連続的な数値空間の中でとらえることができるため，単語の持つ意味の近さ，遠さを数値で表現することが可能となる．このような性質により意味的類似性を

考慮した検索が可能になると考える．

加えて先に述べたとおり，粟飯原らの研究で，検索に関わる語数が少ないことが問題となることが指摘されている．検索対象となる辞書について考えると，検索に関わる語数が少ないということは，辞書定義文の持つ情報量が少ない，ということができる．この不足している情報量を補うため，大規模文書集合の導入も行う．

3　提案手法

本稿では意味的逆引きを既存辞書への検索として実装するという既存の手法を踏襲しつつ，大規模文書集合から学習した単語ベクトルを用いて意味的逆引きに取り組む．

3.1　単語ベクトル

本稿では文のベクトル表現に単語ベクトルを用いる．単語ベクトルは単語を固定長のベクトルとして表現するものであり，その作成には Mikolov らの手法 [5] を用いた．Mikolov らは単語ベクトルを，表 2 に示す Skip-gram モデルの学習により作成した．

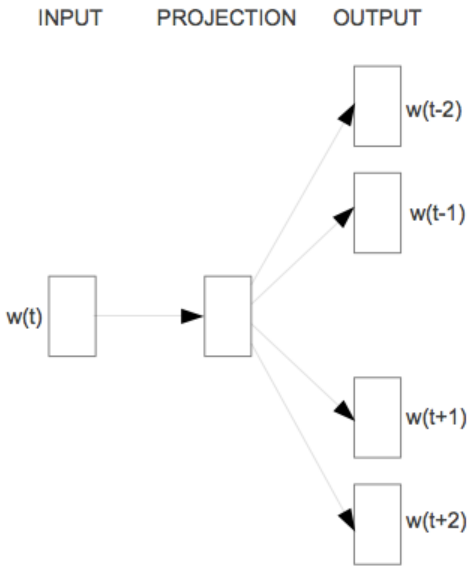


図 2: Skip-gram モデル

Skip-gram モデルは入力単語 $w(t)$ から，j 単語前から j 単語後までの周辺単語 $w(t-j)$〜$w(t+j)$ を予測するモデルである．単語はランダムな値を初期値とする m 次元ベクトルとして表現され，文書集合をもとに，周辺単語をよく予測するよ

うにベクトルを学習する．したがって，周辺単語の分布が似ていれば，ベクトルは似た値をとる．周辺単語，すなわち単語の出現文脈が似ている単語の意味は近いと考えられるため，単語ベクトルの近さが意味の近さを表しているとみなす．

検索エンジン技術による手法では，例えば「くいちがう」と「食い違う」という単語は相互に無関係であった．一方で単語ベクトルを用いた場合,「くいちがう」に対応するベクトルと「食い違う」に対応するベクトルは似た値をとり,「くいちがう」と「事柄」,「食い違う」と「事柄」は離れた値をとる．このように単語ベクトルの適用により，単語間の意味的類似性を連続的な空間の中でとらえることが可能となると考えられる．

本稿では，単語ベクトルの作成に Mikolov らの手法の実装として word2vec[6] を用いた．

3.2 大規模文書集合の利用

単語ベクトルの作成に用いられる文脈情報を得るためには，文書集合が必要である．ユーザーがどのような単語を用いて入力文を与えるか予測できないことから，なるべく多くの単語に対して単語ベクトルを作成したい．したがって，文書集合は多くの単語を含むことが望ましい．また特定種類の文書にのみ多く出現するような単語の影響を避けるため，文書集合は多種の文書を均等に含むことが望ましい．さらに学習精度を向上されるため，規模の大きい文書集合であることが望ましい．

本稿では単語ベクトルの作成に現代日本語書き言葉均衡コーパス (以下，BCCWJ)[7] 2011 年 DVD 版を用いた．BCCWJ は出版サブコーパス，図書館サブコーパス，特定目的サブコーパスからなる文書集合である．各サブコーパスには下記のものが含まれる．

出版サブコーパス 書籍，雑誌，新聞

図書館サブコーパス 書籍

特定目的サブコーパス ベストセラー，白書，検定教科書，広報誌，Web 掲示板，ブログ，韻文，法律，国会会議録

特定目的サブコーパスの一部を除き，BCCWJ に含まれる文書はランダムサンプリングにより収録されており，日本語の全体をバランスよく反映することが目指されている．含まれる文書の偏りが少ないことが期待され，かつ規模が 1 億語規模と大きいことから本稿では BCCWJ を採用した．

なお BCCWJ には品詞情報などのタグが付加されているが，本稿ではタグ情報を削除し，平文に戻してから Mecab[8] により形態素解析を行い，基本形を取り出したものを単語ベクトル学習に用いている．また比較として，辞書中に含まれる全定義文を同様に形態素解析したものから学習した単語ベクトルを用いた実験も行う．

3.3 文への適用

単語ベクトルを入力文及び定義文に適用するにあたっては，最も単純な方法として加算を用いる．以下の定義文

齟齬

意見や事柄がくいちがって，合わないこと．くいちがい．

を例にとると，まず文を形態素解析し，以下のように各形態素の基本形を得る．

齟齬

意見 / や / 事柄 / が / くいちがう / て / 合う / ない / こと / . / くいちがい / .

次に各形態素に対応するベクトルを適用し，加算，正規化を行う．入力文または定義文 s に対する文ベクトル $\boldsymbol{v}(s)$ は次のように表される．

$$\boldsymbol{v}(s) = normalize(\sum_{w \in W_s} \boldsymbol{v}(w)) \tag{4}$$

ただし W_s は文 s に含まれる形態素の基本形，$\boldsymbol{v}(w)$ は単語 w の単語ベクトルである．正規化 $normalize$ は (2) 式により行う．先に述べた単語「齟齬」の定義文について加算，正規化を行うと表 3 のようになる．

表 3: 単語ベクトルの加算

意見	0.105	0.076	0.097	0.227	0.015	...
や	0.009	0.206	0.75	-0.121	0.064	...
…						
.	0.056	-0.050	0.034	-0.010	-0.090	...
計	-0.046	0.605	0.264	0.090	0.066	...
正規化	-0.006	0.075	0.033	0.011	0.008	...

(4) 式による加算を行った場合，意味を持たない記号のような形態素についてもベクトル加算が行われる．また，各形態素の重要度は考慮されていない．よって加算の際には，全形態素に対するベクトルを加算した場合と，形態素解析において記号とされたものを除外し, かつ tf-idf による重みを付加した場合の 2 通りで実験を行う．記号の除外と重み付けを行う場合の文ベクトルは次のように表される．

$$\boldsymbol{v}(s) = normalize(\sum_{w \in T_s} \boldsymbol{v}(w) \cdot tfidf_{w,s}) \tag{5}$$

ただし，T_s は文 s に含まれる形態素のうち，記号を除外したものである．

3.4 実装

以上をまとめたものを図 3 に示す.

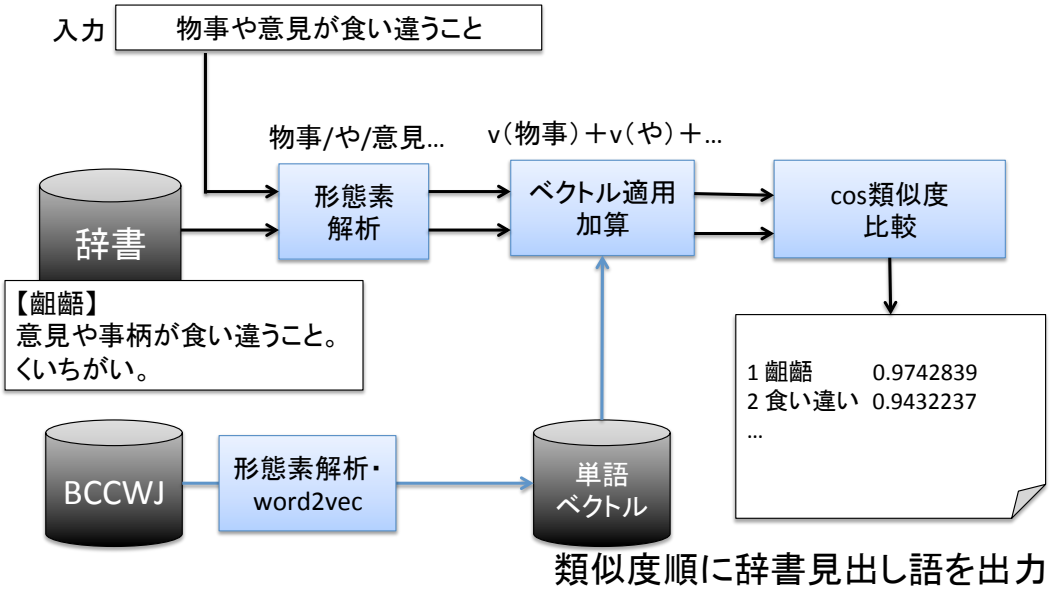

入力
物事や意見が食い違うこと
物事/や/意見…
v(物事)＋v(や)＋…
形態素
解析
ベクトル適用
加算
cos類似度
比較
辞書
【齟齬】
意見や事柄が食い違うこと。
くいちがい。
1 齟齬 0.9742839
2 食い違い 0.9432237
…
BCCWJ
形態素解析・
word2vec
単語
ベクトル
類似度順に辞書見出し語を出力

図 3: 実装図

4　実験

4.1　実験設定

実験において検索対象となる辞書データは，言語資源協会より公開されている「岩波国語辞典第五版タグ付きコーパス 2004」を用いた．本データ中は岩波国語辞典第五版に対して形態素，統語構造，照応，共参照などのタグを付加したものであり，見出し語 60321 語，85351 語義が含まれる．1 つの見出し語に対し複数語義が含まれる場合があるため，語数と語義数は一致しない．語義の区別は予め付与されている語義タグにより行った．本データにおいて定義文はあらかじめ形態素解析が行われ，品詞タグが付与されているが，入力文と形態素解析の方法を統一するため，本実験ではタグ情報を削除したのちに改めて Mecab による形態素解析を行っている．検索は 1 語義を 1 項目とし，入力文との類似度が高い順に見出し語をリストとして出力する．よって，複数語義を持つ見出し語は出力リスト中に複数回出現する．

入力データには言語資源協会より公開されている「計算機用日本語基本辞書ＩＰＡＬ －動詞・形容詞・名詞－」(以下 IPAL 辞書,1214 語，2923 語義) を用いる．各語義に対応する見出し語を正解とし，定義文を入力として与える．正解の判定は見出し語のひらがな表記により行っているため,「かく (書く)」と「かく (核)」など，同じ読みの語の語を正解として誤判定する場合がある．岩波国語辞典と IPAL 辞書の間で漢字表記が異なる場合があり，漢字を用いた判定が難しいため，本稿ではひらがなによる判定を用いている．なお，IPAL 辞書に含まれる全ての見出し語について，岩波国語辞典に一致する項目が存在していることをあらかじめ確認した．

実験設定は以下の 4 通りで比較を行う．

1. 粟飯原らの手法
2. 提案手法，岩波国語辞典により単語ベクトル学習
3. 提案手法，BCCWJ による単語ベクトル学習

4. 提案手法，BCCWJ による単語ベクトル学習，文中の記号を除外，tf-idf による重み付け

提案手法において，単語ベクトルの次元数はすべて 200 次元を用いた．

4.2 評価方法

各入力に対して正解が出現した順位を記録し，上位 N 位中での正解出現頻度及び MRR により評価する．MRR は全入力項目を N，入力項目 n の正解が出現した順位を $rank(n)$ として次のように表される．

$$MRR = \frac{1}{|N|} \sum_{n \in N} \frac{1}{rank(n)} \tag{6}$$

正解が出力中に複数回出現する場合は，最上位のものを評価に用いた．

4.3 実験結果

各実験設定の下，上位 N 位中での正解出現頻度を表 4 に，MRR を表 5 に示す．

表 4: 上位 N 位中の正解出現頻度 (単位:%)

	順位						
実験設定	1	2	3	4	5	10	100
既存手法	8.2	11.0	12.8	14.4	15.5	20.3	43.3
提案手法 (辞書より学習)	1.9	2.9	3.5	3.5	4.1	5.2	12.3
提案手法 (BCCWJ より学習)	8.8	11.9	14.0	15.4	16.5	20.9	37.5
提案手法 (BCCWJ より学習，重み付け)	9.7	13.6	15.9	17.3	18.3	22.9	42.7

表 5: 実験結果 (MRR)

実験設定	MRR
既存手法	0.124
提案手法 (辞書より学習)	0.031
提案手法 (BCCWJ より学習)	0.128
提案手法 (BCCWJ より学習，重み付け)	0.144**

単語ベクトルの学習に岩波国語辞典を用いた場合，既存手法に比べて精度が大きく劣っていることがわかる．

学習に BCCWJ を用いた場合は精度が大きく向上し，既存手法に比べて上位 10 位以内における正解出現頻度，及び MRR が上回った．ここから，大規模文書集合の適用が有効であることがわかる．上位 100 位中での正解出現頻度は既存手法が上回っているが，上位 100 位の単語推薦は現実的ではなく，無視しても構わないと考える．

さらに文ベクトル作成時に記号の除外処理と tf-idf 重み付けの処理を追加すると精度が向上し，MRR において既存研究と比較し，$P < 0.01$ での有意差が得られた．

4.4 考察

4.4.1 単語ベクトルの精度

提案手法において，辞書より学習した単語ベクトルを用いた場合に精度が低くなった．これは，辞書に十分な文脈情報が含まれておらず，単語の意味的類似性をうまくとらえられなかったことが原因であると考えられる．表 6 に，辞書，BCCWJ それぞれで学習した場合での単語「食い違う」とベクトルが類似する単語上位 5 語を示す．

表 6: 「食い違う」の類似単語

辞書から学習	BCCWJ から学習
るまた	似通う
内輪	こじれる
赤身	かみ合う
入れかえる	食い違い
乗りかえる	込み入る

辞書から単語ベクトルを学習した場合，ベクトルの類似性が意味的類似性に繋がっていないことがわかる．BCCWJ から学習した場合は「こじれる」「食い違い」などの単語に意味的な類似性が見られる．よって辞書のみでは意味的な類似性をとらえることは難しく，他のデータによる補完が必要であると考えられる．

BCCWJ から学習した場合において，類義語だけではなく対義語が含まれる点には注意が必要である．対義語は逆の意味を表す単語ではあるが，意味的な類似度は高い．例えば「食い違う」と「かみ合う」は対義語であるが，どちらも何らかのかみ合いについて言及する単語である．出現文脈を考えても，

(vii) 議論が**食い違う**

(viii) 議論が**かみ合う**

のように，似た文脈で使用されると考えられる．したがって対義語の類似度が高くなることは，今回用いた単語ベクトルの作成方法では回避できないと考えられる．類似度の高い単語に対義語が含まれる場合，本来意図した単語とは逆の意味を持つ単語を推薦してしまう可能性があり，精度を下げる要因となっている．

4.4.2 解決できた問題

提案手法により，表記ゆれ及び類義語の差異を吸収することが可能になった．既存手法及び重み付けを行った提案手法について，単語「齟齬」を検索することを意図し，以下の 3 通りの入力

(ix) 物事や意見がくいちがうこと

(x) 物事や意見が食い違うこと

(xi) 事柄や意見がくいちがうこと

を与えた場合の検索結果上位5件は，表7のようになった．なお，今回検索対象として用いた岩波国語辞典において,「齟齬」の定義文は次のように表される．

齟齬

意見や事柄がくいちがって，合わないこと．くいちがい．

表 7: 精度が上がった例

入力	順位	既存手法	提案手法
物事や意見がくいちがうこと	1	あちこち	**齟齬**
	2	**齟齬**	かけちがう
	3	ちぐはぐ	見解
	4	かけちがう	観念
	5	発言	切り盛り
物事や意見が食い違うこと	1	外れる	かけちがう
	2	誤	次第
	3	発言	見解
	4	同意	外れる
	5	建議	**齟齬**
事柄や意見がくいちがうこと	1	**齟齬**	**齟齬**
	2	あちこち	儀
	3	ちぐはぐ	言い分
	4	所論	談話
	5	儀	所論

既存手法と比較して，目的の単語がより上位になっていることがわかる．

4.4.3 悪化した問題

定義文と比較して単語数が少ない文を与えた場合，既存手法と比較して精度が悪化する場合があった．例えば，単語「齟齬」を検索することを意図し，先の例 (ix)〜(xi) の入力から「意見が」の部分を削除し，

(xii) 物事がくいちがうこと

(xiii) 物事が食い違うこと

(xiv) 事柄がくいちがうこと

の 3 通りの入力を与えた場合，結果は表 8 のようになった．

表 8: 精度が下がった例

入力	順位	既存手法	提案手法
物事がくいちがうこと	1	あちこち	切り盛り
	2	ちぐはぐ	過つ
	3	かけちがう	過ち
	4	**齟齬**	太郎
	5	ずれる	終決
物事が食い違うこと	1	外れる	切り盛り
	2	誤	かけちがう
	3	嚆矢	滞留
	4	藪の中	外れる
	5	先駆け	終決
事柄がくいちがうこと	1	あちこち	儀
	2	**齟齬**	作り事
	3	ちぐはぐ	案件
	4	儀	**齟齬**
	5	かけちがう	能事

既存手法と比較し，正解順位が低くなっていることがわかる．

4.4.4 解決できなかった問題

既存手法及び提案手法のどちらも，まったく異なる単語によって説明される単語には対応できない．例えば単語「鋭い」に対し，岩波国語辞典と IPAL 辞書では下記のように定義が異なる．

鋭い

岩波国語辞典 感覚・頭脳などがすばやく動き，すぐれている．
IPAL 辞書 的を得ている

IPAL 辞書では慣用句を用いて説明がなされているため，岩波国語辞典の定義に対し単語レベルでの意味的類似性が存在しない．既存手法，提案手法ともに文を単語に分解してしまうため，このように句として新たな意味を持つものに対してはうまく意味を取ることができない．慣用句は句本来の意味とは違う意味を持つことから，句中の単語から構成的に意味をとらえることが難しい．したがって当該句を 1 つのまとまりとして扱うことで，慣用句の意味をとらえることが可能となると予想される．

ただし，次に示す「勇ましい」の定義文のように，慣用句を用いない場合においても，類似した意味の単語が含まれない場合がある．

勇ましい

岩波国語辞典 勢いが強く，積極的に向かっていく様子だ．
IPAL 辞書 人が危険や困難を恐れない

このような事例においては，慣用句に対する対策では解決が不可能である．

5　まとめ

本稿では単語ベクトルを用いたベクトル空間モデルにより，既存辞書への検索として意味的逆引きを実装し，検索エンジン技術を用いた既存手法との比較を行った．単語ベクトルの学習に大規模文書集合を用いることで精度を向上させることができ，既存手法を上回る精度を得ることができた．文中に含まれる記号の除外，tf-idf による重み付けを行うことで，さらなる精度向上が確認できた．一方で実用を考えた場合にはさらなる精度向上が必要であることもわかった．

本稿では入力文と定義文をそれぞれベクトルに変換して比較することで検索を行ったが，入力文ベクトルと見出し語ベクトルの比較によっても検索が可能である．この比較により精度がどう変わるかについて実験を行いたいと考えている．

本稿の評価実験では入力として既存辞書を与えたが，実用を考えた場合にはユーザーが辞書定義文と同様の入力をするとは限らない．よってユーザーに実際に使用してもらう，ユーザーに入力文を考えてもらうなど，実用に即した評価を行う必要がある．

またユーザーが意図する単語はカテゴリ，品詞などでドメインを制限できる場合がある．ドメイン制限により精度がどう変化するかも評価を行いたいと考えている．

謝辞

本稿を進めるにあたり，ご指導を頂いた乾健太郎教授，岡崎直観准教授，渡邉陽太郎助教に感謝いたします．また，日常の議論や研究会での議論を通じて多くのアドバイスを頂いた乾・岡崎研究室の皆様に感謝いたします．

参考文献

[1] 粟飯原俊介, 長尾真, 田中久美子. 意味的逆引き辞書 『真言』. 言語処理学会 第 19 回年次大会 発表論文集, pp. 406–409, 2013.

[2] 梶原智之, 山本和英. 小学生の読解支援に向けた複数の換言知識を併用した語彙平易化と評価. 言語処理学会 第 19 回年次大会 発表論文集, pp. 272–275, 2013.

[3] 辞典・百科事典の検索サービス - weblio 辞書. http://www.weblio.jp/.

[4] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, Vol. 41, No. 6, pp. 391–407, 1990.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.

[6] Google code. https://code.google.com/p/word2vec/.

[7] 概要 現代日本語書き言葉均衡コーパス （bccwj）. http://www.ninjal.ac.jp/corpus_center/bccwj/index.html.

[8] Mecab: Yet another part-of-speech and morphological analyzer. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.