F-005

索引語ペアを用いた多言語シソーラスの自動構築
Automatic Construction of Multilingual Thesaurus using Index Term Pairs

萩原 正人† 小川 泰弘† 外山 勝彦†
Masato Hagiwara Yasuhiro Ogawa Katsuhiko Toyama

1. はじめに

近年、国際化とインターネットの多言語化が進む中、ある言語で記述された質問文によって、別の言語で記述された情報を検索するクロスリンガル情報検索の必要性が高まっている。クロスリンガル情報検索においては、「言語の壁を越える」ために、辞書に代表される知識源が必要となるが、その中でも重要な役割を果たすのが、多言語シソーラスである。多言語シソーラスは複数の言語の語彙を含むシソーラスであり、クロスリンガル情報検索の性能向上に利用したり、その中間言語として用いたりすることができる。しかし、多言語シソーラスを人手で構築することには、分類基準の言語間での統一や、構築コストなどの問題が常に存在する。

それに対してこれまで、多言語シソーラスを自動構築する研究が行われてきた。特に、文対応のある対訳コーパス中の語の共起情報を利用して、類似する語を得る手法が提案されている [8]。しかしこの手法には、対訳コーパスが簡単には得られない、同じ言語内における語の関係に関する情報が直接は得られないなどの問題がある。

そこで本稿では、言語を問わず任意の語の間で類似度を求めることのできる日英2言語シソーラスを、より入手が容易な知識源である辞書を用いて自動構築する手法を提案する。これまでに、英英辞典などの辞書の語義文をもとに、ベクトル空間モデルを用いて単言語のシソーラスを自動構築する手法はさかんに研究されてきた。本手法はその多言語への拡張であり、未知語であっても語義が記述できれば対応でき、また、辞書を変更することで専門用語も扱うことができる、といった特徴を持つ。

本手法では、日本語と英語の語を、いずれも同一の特徴量を用いてベクトル表現する。そのために、索引語ペアを用いた。索引語ペアとは、例えば (“政府”, “government”) のように、日本語と英語の索引語のうち、対訳として適切なものを組にしたものであり、日本語・英語共通の索引語として扱うことができる。なお、索引語ペアは、日本語と英語の索引語の集合から、和英辞典と英和辞典の語義文を用いて、対訳として適切なペアを選択するという方法で自動生成した。ベクトルを生成した後は、LSI(Latent Semantic Indexing)[2] を用いて主要な意味を抽出し、シソーラスの性能向上を図った。

本稿は以下、2節で基本となるベクトル空間モデルとLSIの概要を述べた後、3節で本手法の特徴となる索引語ペアを導入し、それらを自動生成する方法について詳説する。続く4節では多言語シソーラスを自動構築した実験について述べる。5節では生成されたシソーラスに対する性能評価を行う。

†名古屋大学 大学院 情報科学研究科

2. ベクトル空間モデルと LSI

2.1 ベクトル空間モデル

本手法では、ベクトル空間モデルを用いて、シソーラスの各見出し語 $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ を、ベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ に対応させる。各ベクトルは、特徴づけに用いる索引語 $\boldsymbol{t}_1, ..., \boldsymbol{t}_m$ の語義文中での出現回数を用いて、以下のように作成する[1]。

$$\boldsymbol{x}_j = (x_j^1, x_j^2, ..., x_j^m)^T \quad (1)$$

ここで、x_j^i は、見出し語 $\boldsymbol{w}_j$ の語義文における索引語 $\boldsymbol{t}_i$ の出現回数である。

このようにベクトル表現すると、語 $\boldsymbol{w}_1$ と語 $\boldsymbol{w}_2$ の類似度 $\mathrm{sim}(\boldsymbol{w}_1, \boldsymbol{w}_2)$ は、対応するベクトルの余弦として以下のように求められる。

$$\mathrm{sim}(\boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{\boldsymbol{x}_1 \cdot \boldsymbol{x}_2}{||\boldsymbol{x}_1|| \cdot ||\boldsymbol{x}_2||} \quad (2)$$

2.2 LSI

ここで、複数の索引語が類似した意味を持つ場合 (同義性の問題) や、単一の語が複数の意味を持つような場合 (多義性の問題) などがあるため、必ずしもそれぞれの索引語が見出し語の本質的な意味を表しているとは限らない。そこで、主成分分析の一種である LSI を用いて潜在的な意味を抽出し、シソーラスの性能向上を図る。

LSI では、特異値分解と呼ばれる行列分解を利用し、主要な成分を用いることで、ベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を再構成する。ここで、重要なものから順に $k(< r)$ 個の成分を選ぶことで、各ベクトルの次元を r から k に縮小し、主要な成分のみで各見出し語を表現することができる。

なお、縮小後の次元である k について、4節で述べる実験では、この値を変化させて性能がどのように変化するかを調べた。

3. 索引語ペアによる見出し語のベクトル表現

3.1 索引語ペアの導入

本手法では、日本語と英語の共通の特徴量として「索引語ペア」を用い、日本語と英語の見出し語を同一のベクトル空間上に配置する。索引語ペアは、例えば (“政府”,“government”) のように、対訳として適切な日本語と英語の索引語を対にしたものであり、これは、日本語と英語のどちらの語義文に対しても使用できる共通の索引語であると見なせる。ここで、索引語ペアが語義文中に「出現する」とは、索引語ペアを構成する語のどちらかがその語義文中に出現する場合とする。なお、同一の語を含む複数の索引語ペアが存在する可能性がある。例えば、(“

[1]実際には、各要素に対して重み (定義語ペアの適切さと idf) を乗じる。詳細については 4 節で述べる。

政府”,“government”）と（“行政”,“government”）という2つの索引語ペアが存在し、語義文中に語 “government” が出現する場合である。この場合は、両方の索引語ペアが語義文中に出現したものとして扱う。

例：英英辞典の見出し語 “kitchen” に対応する語義文
“kitchen: the room where you prepare and cook food.”
と、国語辞典の見出し語 “台所” に対応する語義文
“台所: 食物を調理し、煮炊きする部屋。”
を用いて、この2語のベクトル表現と類似度を求める。

ここでは、索引語ペアとして、

$$\begin{aligned} t_0 &= (\text{“部屋”},\text{“room”}) \\ t_1 &= (\text{“準備”},\ \text{“prepare”}) \\ t_2 &= (\text{“調理”},\ \text{“cook”}) \\ t_3 &= (\text{“食物”},\ \text{“food”}) \\ t_4 &= (\text{“煮炊き”},\ \text{“boil”}) \end{aligned}$$

が利用できるとする。これらから、ベクトル $\boldsymbol{x}_{\text{kitchen}}, \boldsymbol{x}_{\text{台所}}$ は、それぞれ

$$\begin{aligned} \boldsymbol{x}_{\text{kitchen}} &= (1,1,1,1,0)^T \\ \boldsymbol{x}_{\text{台所}} &= (1,0,1,1,1)^T \end{aligned}$$

と表現できる。ここで、両語の類似度を式 (2) を用いて求めると、

$$\text{sim}(\text{“kitchen”},\ \text{“台所”}) = \frac{3}{2\cdot 2} = 0.75$$

となる。

3.2 索引語の選択

本手法では、索引語ペアを構成する索引語が、辞書の語義文中に出現するかどうかにより、語のベクトル表現を与える。そのため、索引語は、辞書の語義文中に出現する語をなるべく多く採用するよう選択した。

まず、英語の語義文は、英英辞典 LDOCE[4] のものを使用した。その語義文は、原則として Longman Defining Vocabulary(LDV) と呼ばれる約 2000 語の語彙によって記述されている[2]。したがって、英語に関しては LDV を索引語として選択することにした。

一方、日本語の語義文に関しては、大辞林 [5] のものを使用した。しかし、その語義文は LDV のような語彙制限が行われていないため、索引語は、出現頻度の高い語を使用することとした。ここでは、最終的に多言語シソーラスに含める日本語の見出し語 (詳細は 4 節で述べる) に対応する語義文だけを考え、そこに出現する頻度の高い 6000 語を選んだ。

3.3 索引語ペア自動生成の方法

日英両言語の索引語から、索引語ペアを構成するために、以下の手順を用いた：

日本語の索引語 $\boldsymbol{j}$ と英語の索引語 $\boldsymbol{e}$ について、もし、
(1) 和英辞典中の語 $\boldsymbol{j}$ の語義文中に語 $\boldsymbol{e}$ が出現し、かつ
(2) 英和辞典中の語 $\boldsymbol{e}$ の語義文中に語 $\boldsymbol{j}$ が出現する
ならば、$(\boldsymbol{j},\boldsymbol{e})$ を索引語ペアとする。

ここで、和英辞典と英和辞典の両方を用いた強い条件を与えたのは、対訳として不適切な索引語ペアがなるべく生成されないようにするためである。

[2]派生語など、その原則に沿わない語については、今回は使用しなかった。

表 1: 自動生成された索引語ペアの例

索引語ペア	適切さ	索引語ペア	適切さ
(省略, abbreviation)	1.00	(受ける, accept)	1.00
(能力, ability)	1.00	(応じる, accept)	0.23
(才能, ability)	1.00	(引き受ける, accept)	0.58
(腕前, ability)	0.75	(受け取る, accept)	0.38
(できる, able)	0.75	(事, accident)	0.18
(才能, able)	0.75	(偶然, accident)	1.00
(およそ, about)	1.00	(事故, accident)	0.60
(上, above)	0.75	(災難, accident)	0.25
(以上, above)	0.62	(偶然, accidental)	1.00
(超越, above)	0.50	(話, account)	0.27
(前述, above)	0.67	(理由, account)	0.16
(外国, abroad)	0.62	(責任, account)	0.62

なお、各索引語ペアには、その対訳としての「適切さ」a を持たせる。このとき、「語義文では対訳として適切な語から順に記述されている」という仮定に基づき、語義文中における語 $\boldsymbol{j}$,$\boldsymbol{e}$ の出現順位が高いほど適切さ a が大きくなるように決定する。例えば、英和辞典の見出し語 “government” に対する語義文として、“統治, 政治. 政体. 政府.” とあった場合には、 “政治” よりも “統治” の方が “government” の訳語としてより適切と判断する。

具体的には、語 $\boldsymbol{j}$ の語義文中における語 $\boldsymbol{e}$ の出現順位を p_{JE} (p_{JE}は 1 以上の整数)、語 $\boldsymbol{e}$ の語義文中における語 $\boldsymbol{j}$ の出現順位を p_{EJ} (p_{EJ}は 1 以上の整数) とすると、索引語ペア $(\boldsymbol{j},\boldsymbol{e})$ の適切さ a は、両者の逆数の平均、すなわち、

$$a = \frac{p_{\text{JE}} + p_{\text{EJ}}}{2\cdot p_{\text{JE}}\cdot p_{\text{EJ}}} \tag{3}$$

とした。ただし、ひとつの見出し語に対応して複数の語義文が存在する場合、ある語の出現順位はその語の出現する語義文の先頭から数えるものとする。この「適切さ」は、ベクトル生成の際の重みとして利用する。その詳細は 4 節で述べる。

本実験では、和英辞典 [6] および英和辞典 [7] を使用した。ただし、和英辞典中の複合語、英和辞典中の成句・慣用句に関する記述は語義文として使用しなかった。これは、成句・慣用句では、語が本来の意味とは異なる意味で用いられている場合が多いため、索引語ペアの質の低下につながると判断したためである。なお、英和辞典の語義文は茶筌を用いて形態素に分割し、和英辞典の語義文は原形を求める処理を施してそれぞれ使用した。後者の処理は、[1] の活用情報を利用して、活用形と原形の対応表を作成して行った。

3.4 索引語ペア自動生成の結果

3.3 で述べた方法により索引語ペアの自動生成を行った結果、4787 個の索引語ペアが生成された。その一部を表 1 に示す。生成された索引語ペアはすべて、見出し語の特徴づけの際に使用する。

4. 多言語シソーラス自動構築実験

本節で述べる実験では、日本語と英語の見出し語を、前節で生成した索引語ペアを用いてそれぞれベクトル表現する。これにより各見出し語間の類似度を求めることが可能となり、日英 2 言語シソーラスが自動生成できる。

シソーラスに含める見出し語と、ベクトル生成の際に用いる語義文は、以下のようにして選択した。

4.1 シソーラス見出し語の選択

本実験では、シソーラスに含める見出し語として、以下のものを使用した。

- **日本語**: 教育基本語彙 [3](約 6000 語) のうち、「より基本的な語」とされている語で、感嘆詞と連語を除いたもの (2039 語)(以下、この見出し語の集合を JBV(Japanese Basic Vocabulary) と呼ぶ。)
- **英語**: LDOCE[4] の見出し語のうちで、「書き言葉における頻度が上位 3000 位以内」と「話し言葉における頻度が上位 3000 位以内」の頻度情報のうち少なくとも一方が付加されており、かつ、その語義文中の 85 %以上の語に 3 節で求めた索引語ペアが存在するもの[3](2479 語)(以下、この見出し語の集合を LFE(Longman Frequent Entries) と呼ぶ。)

4.2 使用する辞書と語義文の選択

見出し語に対応するベクトル生成の際に用いた語義文は、以下のとおりである。

- **日本語**: 大辞林 [5] の語義文。JBV と大辞林の見出し語との対応づけの際には、JBV に付加されている漢字表記を参照し、読みが同じで、かつ共通の漢字表記を 1 つでも持てば、対応する語として扱うようにした。
- **英語**: LDOCE[4] の語義文。

なお、両辞書とも、用例や成句・慣用句に関する記述は、語義文としては使用しなかった。また、LDOCE の語義文中、“a”, “an”, “the”, “something”, “etc”, “ ’s”(所有の’s), “sth”(something の略), “sb”(somebody の略) の 7 語に関しては、ストップワードとして除外した。英英辞典は和英辞典と同様に語の原形を求める処理を、また、国語辞典は英和辞典と同様に形態素解析を、それぞれ語義文に施して使用した。

4.3 重み付けの詳細

見出し語に対応するベクトルの各要素は、索引語ペアが語義文中に出現した回数であり、これに各索引語ペアの適切さと idf(inverse document frequency) の 2 つを用いて重みづけをする。本手法では、idf の値を計算する際は、日本語と英語で別々に行うことにした。これは、国語辞典と英英辞典で、索引語ペアの出現する「特徴」が異なってしまう事態を避けるためである。例えば、英英辞典の語義文中には、“especially”, “something” といった語が、また、国語辞典の語義文中には “こと”、“もの” といった語が頻繁に出現する。したがって、その語義文を用いて生成したベクトルも、必然的にこのような違いの影響を受けたものになる。実際、idf による重みづけを行わずに予備実験を行ったところ、日本語の見出し語間および英語の見出し語間での類似度が相対的に高くなり、日本語と英語の間で類似する語が得にくくなった。

[3]日本語に対してはこの条件を付加しなかった。これは、日本語の見出し語でこの条件を満たすものが英語と比較して少なく、見出し語の数における公平性が保たれなくなるためである。

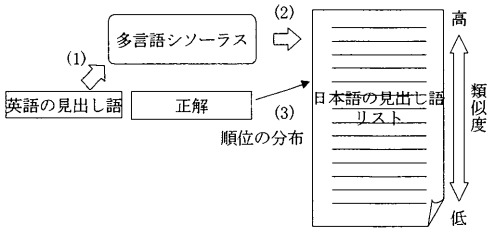

図 1: 性能評価方法の概念図

ある索引語ペア $\boldsymbol{t}_i$ に対応する日本語と英語の語義文のそれぞれにおける idf の値 $\mathrm{idf}^{\mathrm{J}}{}_i$、$\mathrm{idf}^{\mathrm{E}}{}_i$ は、その索引語ペアが出現する日本語と英語のそれぞれの見出し語数 $\mathrm{df}^{\mathrm{J}}{}_i, \mathrm{df}^{\mathrm{E}}{}_i$ を用いて、以下のように与える。ただし、idf 値は、$0 \leq \mathrm{idf}_i \leq 1$ となるように正規化した。

$$\mathrm{idf}^{\mathrm{J}}{}_i = \frac{1}{\log n^{\mathrm{J}}} \log \frac{n^{\mathrm{J}}}{\mathrm{df}_i} = 1 - \log_{n^{\mathrm{J}}} \mathrm{df}^{\mathrm{J}}{}_i \tag{4}$$

$$\mathrm{idf}^{\mathrm{E}}{}_i = \frac{1}{\log n^{\mathrm{E}}} \log \frac{n^{\mathrm{E}}}{\mathrm{df}_i} = 1 - \log_{n^{\mathrm{E}}} \mathrm{df}^{\mathrm{E}}{}_i \tag{5}$$

なお、ここで $n^{\mathrm{J}}, n^{\mathrm{E}}$ はそれぞれ、日本語と英語の見出し語の総数である。

以上より、シソーラスの見出し語 $\boldsymbol{w}_j$ に対応するベクトル $\boldsymbol{x}_j$ は、各索引語ペア $\boldsymbol{t}_i$ の適切さ a_i と idf 値である idf_i、さらに、その索引語ペアが見出し語 $\boldsymbol{w}_j$ の語義文中に出現する回数 tf^i_j を用いて以下のように求められる。

$$\boldsymbol{x}_j = (x^1_j, x^2_j, \ldots, x^m_j)^T, \tag{6}$$

$$x^i_j = \begin{cases} a_i \cdot \mathrm{idf}^{\mathrm{J}}{}_i \cdot \mathrm{tf}^i_j & (\boldsymbol{w}_j \in \mathrm{JBV}) \\ a_i \cdot \mathrm{idf}^{\mathrm{E}}{}_i \cdot \mathrm{tf}^i_j & (\boldsymbol{w}_j \in \mathrm{LFE}) \end{cases} \tag{7}$$

5. 多言語シソーラスの性能評価

本節では、生成されたシソーラスに対する性能評価の方法と、その結果についての考察を述べる。

5.1 性能評価の方法

生成された多言語シソーラスに対しての性能評価は、その英和辞典としての働きに着目し、「ある英語の見出し語を与えたときに、その語と類似する日本語の見出し語が正しく得られるか」という観点から行った。

ある英語の見出し語を与えたときに、類似する語として求められる日本語の見出し語、すなわち「正解」は、3 節で述べた索引語ペアの自動生成と同じ手法を用いて「正解ペア」を生成することにより与えた。

このように生成した正解ペアを用いて、多言語シソーラスの性能評価を次のようにして行った。

(1) 自動構築された多言語シソーラスに対して、英語の見出し語を入力する。この見出し語には上記の手法によって求めた正解が対応づけられている。

(2) その見出し語と類似する日本語の見出し語のリストを、多言語シソーラスを用いて求める。

(3) 正解がそのリスト中に占める順位を調べる。

この性能評価方法の概念図を図 1 に示す。

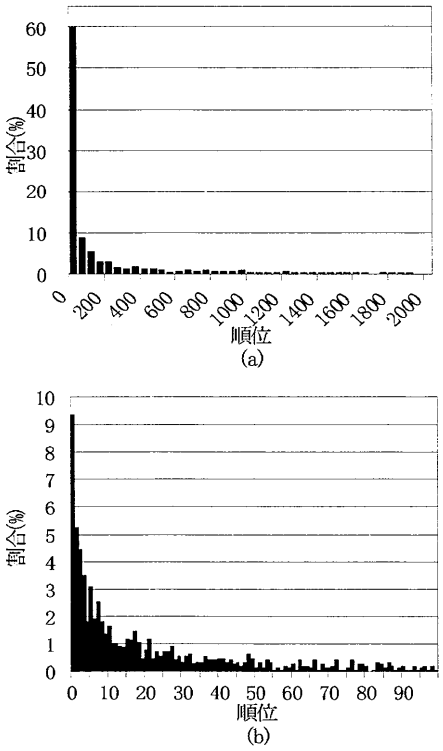

図 2: 正解の順位の (a) 分布の全体と (b) その一部の詳細

5.2 性能評価の結果

まず、LSI による縮小後の次元 k を決定するために、予備実験を行った。その結果より、以下では $k = 500$ と固定することとした。

このときの正解の順位分布を図 2 に示す。(a) のヒストグラムは分布の全体を示している。ここで、横軸の順位は 1 から日本語の見出し語数である 2039 までの値を取る。(b) は、(a) のうち 1 位から 100 位までの分布を拡大して示したものである。なお、“今晩”、“light(名詞)”の 2 語について、生成されたシソーラスを用いて類似する語を求めた結果を表 2 に示す。

5.3 考察

図 2 より、英語の見出し語 2479 語のうち約 6 割に関しては、正解の語の順位が 50 位まで、すなわち全体の上位 1/40 に位置することがわかり、類似する語がおおよそ正しく求められた。しかし、残りの約 4 割の語については、正解がその範囲に存在しておらず、類似していない対訳が求められた。

その原因として考えられるのは、本手法では、語の語義文内における出現のみを扱っており、係り受けや構文の情報は利用していないことが挙げられる。このため、例えば語義文中に否定語などが現れた場合、その情報を正しく利用することができない。

また、辞書では本来の意味 (“青い” に対して “よく晴れた空の色”) のほかに、慣用的な意味 (“未熟である”) なども網羅して記述されているという点も性能低下の原因として挙げられる。ただ、このように慣用的な意味に注目した類義語関係 (“blue” と “nervous” など) を求めたいこともあり得る。この実験結果については、他にも原因があると思われるので、さらに検討を続けていく。

表 2: 求められた類似する見出し語の例

“今晩”		“light(名)”	
見出し語	類似度	見出し語	類似度
晩	0.77	light(動)	0.68
night	0.68	light(形)	0.63
毎晩	0.64	sunshine	0.49
今夜	0.62	contrast(名)	0.46
夕方	0.61	electricity	0.46
夜中	0.61	pale(形)	0.42
朝	0.60	shadow(名)	0.42
明日	0.43	candle	0.41
evening(名)	0.41	dark(形)	0.41
夕べ	0.40	plastic(名)	0.40
goodnight	0.34	rose(名)	0.40
dinner	0.31	bright	0.38
dark(形)	0.29	照らす	0.38
sleep(動)	0.29	electric	0.37
delay(動)	0.29	lamp	0.37

6. おわりに

本稿では、辞書の記述を用いて、任意の語の間で類似度を求めることのできる日英 2 言語シソーラスを自動構築する手法を提案した。その際、日本語と英語共通の特徴量として用いるために索引語ペアを使用し、見出し語を同一ベクトル空間上にベクトル表現した。

生成されたシソーラスに対する性能評価の結果、約 6 割の語に対して、類似する語がおおよそ正しく求められることを確認した。しかし、生成された多言語シソーラスの性能は十分なものではなく、今後の性能向上が課題である。このためには、語義文の係り受け、構文情報の利用などが有効であると考える。

また、本手法を用いて、3 言語以上の語彙を含む多言語シソーラスを構築することは可能であると考えられ、引き続き検討する。

参考文献

[1] Collins Cobuild Major New Edition CD-ROM, HarperCollins Publishers, 2002.

[2] Scott Deerwester, et al.; Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 41(6), pp.391-407, 1990.

[3] 国立国語研究所: 教育基本語彙の基本的研究, 2001.

[4] Longman Group: Longman Dictionary of Contemporary English 3rd edition, 1995.

[5] 松村 明 (編): 大辞林 第二版, 三省堂, 1999. (http://www.sanseido.net/sup/ash/AsahiTop.asp)

[6] 三省堂: エクシード和英辞典, 1998. (http://www.sanseido.net/sup/ash/AsahiTop.asp)

[7] 三省堂: エクシード英和辞典, 1998. (http://www.sanseido.net/sup/ash/AsahiTop.asp)

[8] 辻慶太, 影浦峡, 芳鐘冬樹: 対訳コーパスからの訳語対抽出：多言語シソーラス自動構築に向けて, 1999 年度日本図書館情報学会春季研究集会, pp.25-28, 1999.