学術文献の和英著者キーワードを用いた多言語クラスタ生成の試み

相澤 彰子　　影浦 峡

学術情報センター

{akiko, kyo}@rd.nacsis.ac.jp

1 はじめに

学術文献を特徴づけるキーワードとなる用語は，一般語と比較して2カ国語間の対応がとりやすく，あらかじめ分野を限定すると多義語の影響も少ない．また，これらの用語を学術文献から自動抽出することは一般には容易ではないが，研究者が自ら提示する著者キーワードに注目すると，これらは検索に有用な用語である可能性が高く，標準辞書による専門用語と比較して，対象とする文献に固有の用法や最新の話題を反映しているという利点がある．

そこで本稿では，学術文献の学会発表データベースから著者キーワードを取り出して，和英混在の用語クラスタを作成する試みについて述べる．具体的には，和英キーワード間の対訳関係を利用して同義関係にある用語を抽出してクラスタとするが，この際に，著者による希少な訳例をなるべく保存しつつ，誤訳やアラインメント誤りを効率よく検出するための手法を検討する．このようにして作成した用語クラスタは，多言語検索における検索語拡張や自動索引づけのための同義語辞書として用いることを前提としている．

以下，まず **2.** で，対訳辞書としての和英著者キーワードの特性を調べ，次に **3.** で，多言語用語クラスタの生成に従来の類似度尺度に基づく分類手法を適用することの問題点を述べる．**4.** では，従来手法に代わるものとして用語グラフに基づくクラスタ化手法を提案し，実際に用語クラスタを作成した結果を示す．最後に **5.** で今後の課題を述べる．

2 和英著者キーワードの分析

2.1 和英対訳用語コーパスの作成

文献が和英双方のキーワードを持つ場合でも，著者は各言語について独立にキーワードを選ぶことができるため，両者の対応関係は必ずしも保証されていない．そこで，著者による和英キーワードの対訳用語コーパスとしての有用性を判断するため，学術情報センターの学会発表データベースに登録された情報処理および人工知能分野に関連する文献のうち，和英著者キーワードを持つ23,540件について，リストの先頭から出現順に和英用語の対応をとった場合のアラインメント（対応順序）誤りの頻度を調べた．

まず，和英キーワードの数が異なる文献は23,540件のうち1,339件（6%）存在した．これらの中には，いずれかの言語のキーワードが1つ欠けているものから，単一のキーワードを複数に分割して登録してしまったものまで多様なケースが含まれ，和英の自動的な対応付けは困難であることが予想される．一方，キーワード数が等しい文献については，ほとんどの場合，出現順に和英キーワードが対応しており，23,540件の文献から100件を無作為抽出して人手により調べたところ，合計452個の和英キーワード対のうちアラインメント誤りは1カ所（2対）のみであった．

以上の分析結果に基づき，和英キーワードの数が等しいものにつき単純に出現順に対応をとることにし，異なる60,186対の和英用語を取り出して，情報処理および人工知能分野に関する対訳用語コーパスとした．これより抜粋した対訳例を表1に示す．

表1: 和英著者キーワードによる対訳の例

和用語	英用語	出現数
キーワード	information retrieval	1
キーワード	keyword	39
テキスト検索	information retrieval	1
テキスト検索	text retrieval	6
テキスト検索	text search	3
検索指示語	keyword	1
広域情報検索	information retrieval	1
情報検索	information gathering	4
情報検索	information retreival	1
情報検索	information retrieval	320
情報検索	information search	5
情報収集	information gathering	6
情報収集	information retrieval	1
文献検索	bibliographich search	1
文献検索	document retrieval	11
文書検索	document retrieval	19
文書検索	text retrieval	1

2.2 標準辞書との比較

上記で作成した対訳用語コーパスと既存の対訳専門用語辞書との共通性を調べるため,「人工知能大辞典（丸善）項目目次元」「人工知能大辞典（丸善）索引データ

An Approach to Automatic Generation of Multilingual Clusters of Technical Terms based on Japanese-English Keywords given by Authors of Academic Papers
Akiko AIZAWA, Kyo KAGEURA
National Center for Science Information Systems
3-29-1 Otsuka, Bunkyoku, Tokyo 112-8640, Japan

元」「人工知能ハンドブック（オーム社）索引データ元」「コンピュータ大百科（朝倉）索引」「情報処理用語大辞典（オーム社）索引」の5つの辞書から得た異なる22,690対の対訳データを用いて比較を行った．その結果を表2に示す．

表2: 既存の専門用語辞書と対訳用語コーパスの比較

	既存辞書	対訳コーパス	共通
和用語の数	20,636	37,170	3,966
英用語の数	19,562	49,918	2,814
和英対訳の数	22,690	60,186	2,066
平均訳語数(和)	1.10	1.62	−−
平均訳語数(英)	1.16	1.21	−−
最大訳語数(和)	7	86	−−
最大訳語数(英)	6	29	−−

表2より，既存の専門用語辞書と対訳用語コーパスでは共通する用語や対訳の数は比較的少ないことがわかる．すなわち，著者がキーワードとしてあげる用語の多くは，既存の辞書には登録されていない．一方，用語あたりの訳語数を比較すると，コーパスによる対訳の方が多い．これは，専門用語辞書が訳語の統一性を意識して編集されるのに対して，著者キーワードには研究者の独自の見解が反映されており，文献に固有の用法や最新の用語を含むことによると考えられる．

ただし，著者キーワードによる対訳の中には，対応順序の誤りに起因する誤訳，関連語だが訳語としては適切でないもの，表記の揺れ，入力誤りなども多く含まれている．特に，本稿で目的とするような文献検索への適用を考慮すると，

- 精度の観点からは，誤訳や不適切な訳語を検出して取り除く操作が重要であり，
- 再現率の観点からは，表記揺れや入力誤りをそのまま保持することが有利である

と考えられる．また，文献検索のためには，同一言語内でも同義関係にある用語を参照できることが望ましい．そこで以下では，対訳用語コーパスから誤訳や不適切な訳語を検出して，和英用語が混在する同義語クラスタを生成する手法を検討する．

3 既存の統計的手法の適用と問題点

3.1 相互情報量に基づく対訳関係の自動抽出

対訳コーパスからの対訳関係の自動抽出技術としては，統計的類似度尺度に基づく方法が一般的である．これらの手法は基本的に，対訳コーパス中の対応可能な用語（単語列）について共起頻度を求め，その対応のもっともらしさを相互情報量などの数値で表現するものである．しかし，これらの手法の目的は「正しい」すなわち相対頻度の高い対訳関係の選択にあり，相対的に頻度の低い誤訳，希少な訳例，表記揺れなどはすべて一様に無視されてしまうことになる．また当然のことながら，直接的に共起しない和用語と和用語，英用語と英用語のような同一言語内での比較は行えない．

3.2 LSIの適用による用語間類似度の計算

一方，文献検索における自動索引技術としては，用語の出現頻度に基づき文献間の類似度を計算する方法が一般的である．特に，文献 [1] による LSI (Latent Semantic Indexing) では，特異値分解と呼ばれる手法を用いて出現頻度行列を変形し，文献−文献間，用語−用語間，文献−用語間それぞれについて，ベクトル空間上での類似度計算を可能にしている．

用語の対訳関係の抽出にあたっても同様にして，コーパス中での対訳関係の出現頻度行列を作成して特異値分解を適用すると，用語間の類似度を求めることができる．図1に，表2の例題から和用語 − 英用語の 8×9 頻度行列を作成して特異値分解を適用した結果を示す．図中で各軸の括弧内の数字は各成分に与えられる重みを表しており，類似度計算における第1,2,3成分の寄与率が97%であることを示している．

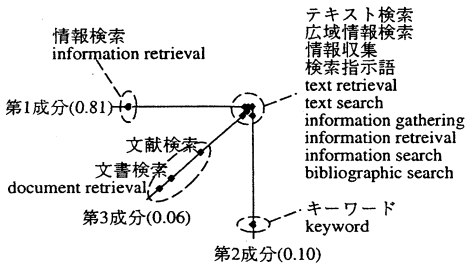

図 1: 特異値分解による用語クラスタ生成の例

このように特異値分解を適用すると，対訳用語コーパス中での出現頻度が高い用語に関しては，見通しよくクラスタを生成することが可能である．しかし，この場合にも用語間の類似度が相対的な共起頻度に依存することに変わりはなく，ベクトル空間上での類似度の定義やクラスタ化手順を工夫しても，ともに共起頻度の低い誤訳や標記揺れを数値的に区別することは困難である．また，類似度の高いものからボトムアップにクラスタ化を行うために計算量が多くなり，大規模なデータに対応できないことも問題点としてあげられる．

4 用語グラフに基づく多言語用語クラスタの生成

4.1 対訳関係に基づく用語グラフの生成

上記で述べた問題を解決する手段として，用語のグラフ表現に基づくクラスタ化手法を検討する．具体的

には，和英用語をノード，対訳関係をリンクとみなして大規模な用語グラフを生成し，対訳誤りの検出をグラフ理論の最小カット問題に帰着させることにより，効率のよい用語クラスタの生成を目指す．

このためにまず，コーパス中に含まれるすべての対訳関係を用いて，コーパス全体を1つの用語グラフで表現する．図2に表1の例題に基づき作成した用語グラフを示す．次に，出現頻度によらず対訳リンクによって連結された用語をすべて同義語であると定義し，用語グラフ上での連結成分を1つの用語クラスタに対応させる．図2の例ではすべてのノードが互いに連結していることから，全用語が同一のクラスタに所属することになる．以下の誤り検出処理は，このようにして作成した用語クラスタを単位として行うものとする．

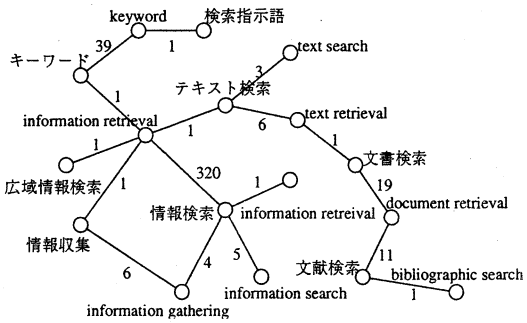

図 2: 例題に基づき作成した用語グラフ

4.2 誤り候補となる対訳リンクの検出

2.1 で作成したコーパスの 60,186 個の対訳関係を用いて用語グラフを作成すると，全対訳の約 34%にあたる 20,659 個が1つのクラスタに含まれてしまうことがわかる．このように大きな用語クラスタが生成される原因となるのが，図2における <キーワード，information retrieval >のように，本来対応しない用語どうしを連結する対訳誤りの存在である．本稿ではこの点に着目し，

- 連結する用語クラスタを分割するような対訳リンクの集合，すなわちグラフ理論における辺カットが対訳誤りの候補となる

ことを利用して誤り候補の検出を行う．グラフ理論ではこのように連結グラフを切断する辺カットのうち，その容量（対訳の出現頻度に対応）の和が最小のものを最小辺カットと呼ぶ．最小辺カットは，複数の対訳誤りに対応する場合もある．たとえば以下の図3に示す例では，用語クラスタを分割するために <情報検索，text retrieval>，<テキスト検索，information retrieval>の2つの対訳リンクを同時に削除しなければならない．

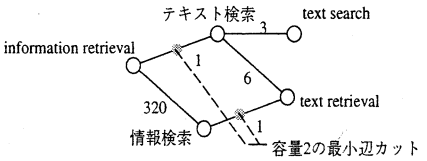

図 3: 最小辺カット数が2となる用語クラスタの例

最小辺カットを求める問題はグラフ理論の中で最も基本的な問題の1つであり，数多くの効率的なアルゴリズムが研究されている [2]．本稿で用いる単純な Dinits の方法でも，ノード数 n，リンク数 m に対して $O(n^2m)$ で実行可能であり，実用上十分な処理速度を得ることができる．

4.3 対訳の正誤判定による用語クラスタの分割

さて，上記の手順に基づき得られる最小辺カットがつねに対訳誤りであるとは限らない．正しい対訳関係を不必要に削除してしまうと，本来同義関係にあるべきクラスタが過剰に分割されることになる．しかし一方で，対訳語と，関連性はあるが対訳とはみなせない語との区別は利用目的や状況にも依存しており，正誤の判定は専門家でも難しい．たとえば，< テキスト検索，information retrieval > という対訳は，専門用語を定義する立場からは不適切であるとしても，検索者の立場からは必ずしも誤りではない．

本稿では利用目的を文献検索に限定していることから，生成する用語クラスタの大きさ，すなわちそのクラスタ中の用語を著者キーワードとして持つ文献の数をなるべく平均化するように，対訳の出現頻度を手がかりとして以下の正誤判定を行う．

まず，次の3つのいずれかにあてはまる対訳リンクは削除の対象から除外する．

[対訳リンクの削除可能性に関する条件]

(a) コーパス中での出現頻度が N_α 回以上
(b) 和英用語が同一表記
(c) 和英いずれかの用語について，対訳関係にある用語がただ1つ（その対訳リンク自身）だけ

ここで (a) は対訳がつねに正しいとみなすための最小支持数であり，たとえば $N_\alpha = 5$ のとき全体の約 5% にあたる 2,723 個の対訳リンクがこれに相当する．(b) は人名，システム名，略記などであり，全体の約 2%に

あたる 1,336 個がこれに相当する．(c) は用語クラスタが必ず和英両方の用語を持つようにするための条件で，全体の約 82% にあたる 49,233 個がこれに相当する．

また，用語クラスタが分割可能であるのは，次の条件をともに満足する場合だけに限る．

[用語クラスタの分割可能性に関する条件]

(d) 最小辺カットである対訳リンクの出現頻度の和が N_ϵ 以下

(e) 最小辺カットを用いて用語クラスタを分割した場合に，得られるすべてのクラスタについて，出現頻度が N_β 以上の用語が 1 つ以上存在する

用語クラスタの分割は再帰的に行うものとし，条件 (d)(e) を満足するクラスタがなくなるまで処理を続ける．

以上をまとめると，クラスタ分割手順においてあらかじめ指定するパラメタは N_ϵ，N_α，N_β の 3 つである．現在は単純に $N_\epsilon = N_\alpha = N_\beta$ とし，すべての用語クラスタについて共通の値を用いている．

4.4 用語クラスタ生成の結果

図 4に，$N_\epsilon = N_\alpha = N_\beta = 5$ として図 2の用語クラスタを分割した結果を示す．対訳誤りの検出を行った結果，<キーワード，information retrieval> <テキスト検索，information retrieval> <文書検索，text retreival> の 3 つが削除され，A,B,C,D の 4 つの用語クラスタが新たに生成されている．ここで図中太線で示した対訳リンクは，前述の条件 (c) により削除の対象にはならない．したがって，<情報検索，information retreival>（綴り誤り），<検索指示語，keyword>（希少訳例），<広域情報検索，information retrieval>（類義関係）などはそのままの形で残されることになる．

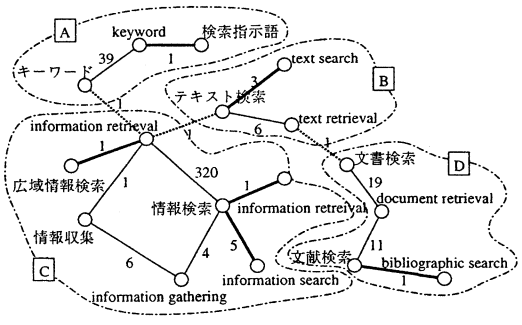

図 4: 用語クラスタの分割の例

ここで $N_\beta = 10$ とすると，条件 (e) からグループ B は用語クラスタとして独立することができず，グループ C またはグループ D と連結したままの状態で残される．いずれのグループに連結するかは，現在のところ，各グループに含まれる用語の文字ヒストグラムの類似度から判定することにしているが，和英用語の文法上の構成単位の意味的対応づけなど，より高度な手がかりの利用も有効であると考えられる．

図 4の例と同様にして，**2.1** で作成したコーパスに本手法を適用したところ，60,186 個の対訳関係のうち 1,469 個が誤りとして削除され，最終的に 27,918 個の用語クラスタが生成された．実行時間は Ultra2(200MHz) 上で 4 分 30 秒程度であった．また，最大の用語クラスタに含まれる対訳の数は，当初の 20,659 個から 159 個（用語数では和用語 43，英用語 72）になった．この中に含まれる用語のうち頻度が 3 以上であるものを以下に示す．

和 :並列処理 (740), 並列 (62), 並列化 (56), 並列計算 (29), 並行処理 (19), 並列性 (10), 並列システム (6), 多重処理 (6), 並行システム (5), マルチプロセス (5), 並列プロセス (4), 並列度 (3), 多重プロセス (3), マルチプロセシング (3), パラレル処理 (3)

英 :parallel processing(672), parallel(74), parallelization(44), concurrent processing(20), parallel computing(18), parallel computation(18), parallelism(14), parallel process(8), multi process(8), concurrent system(7), parallel system(6), parallel processings(6), multi processing(6), multiprocessing(5), multiprocess(5), concurrent(5), future(4), parallelize(4), parallel processes(4), parallel operation(4), paralell processing(4)

5 今後の課題

上記のように著者キーワードに基づき作成した用語クラスタは標準辞書にはない多様な表記のバリエーションを含んでいる．これらは検索語の入力としても想定されるものであり，今後は現在作成中の NACSIS 版文献検索テストコレクション [3] を用いて有効性を評価する予定である．

謝辞

本研究は学術振興会の未来開拓学術研究推進事業による「高度分散情報資源活用のためのユービキタス情報システムに関する研究」のもとで行われた．

参考文献

[1] Deerwester,S., Dumais, S.T., Furnas, G.W., Landauer T.K., and Harshman, R. : "Indexing by Latent Semantic Analysis," Journal of the American Society of Information Science, Vol,41, No.6, pp.391-407 (1990).

[2] 永持仁 :「グラフの最小カット」，離散構造とアルゴリズム II, 第 4 章, 藤重悟編，近代科学社 (1993).

[3] Kageura, K., Koyama, T., Yoshioka, M., Takasu, A., Nozue, T. and Tsuji, K. "NACSIS Corpus Project for IR and Terminological Research," *Natural Language Processing Pacific Rim Symposium 1997*, p.493-496 (1997).