学術文献の和英著者キーワードを用いた類義語クラスタの自動生成

相澤彰子[†] 影浦峡[†]

横断的検索における検索語翻訳や検索語拡張においては，意味的に類似する用語をまとめた類義語辞書が重要な役割を果たすが，専門性の高い学術分野においては，このような語彙的資源を構築することは容易ではない．そこで本論文では，学術文献データベースの著者キーワードを利用して，和英が混在する学術用語の類義語クラスタを自動生成する試みについて述べる．具体的には，和英著者キーワード間の対訳関係を利用して，共通の訳語を持つ用語どうしを類義語と見なすことによりクラスタを生成する．この際に，表記揺れや希少な訳例をなるべく保存しつつ，意味的に誤った対応を検出することが重要である．提案手法では専門用語をノード，和英の対応関係をリンクに対応させて大規模な「用語グラフ」を生成し，誤り候補の検出をグラフの最小カット問題に帰着することによって効率良く誤りを検出して用語クラスタを生成する．本論文中ではまた，提案手法を実際の学術文献データベースの和英著者キーワードに適用して類義語クラスタを生成し，検出した対応誤りを分析して有効性を評価した結果を報告する．

Automatic Generation of Clusters of Synonyms Utilizing Japanese-English Keyword Lists of Academic Papers

Akiko Aizawa[†] and Kyo Kageura[†]

Multilingual keyword clusters with similar meanings are recognized to be crucial linguistic resources in query translation or query expansion in information retrieval across language or resource boundary. However, the construction of such lexical resources easily becomes a bottleneck in information retrieval in academic research fields where most of the keywords are domain-specific. Based on this, we report in this paper a method for automatically generating Japanese and English keyword clusters using the keyword lists assigned to academic papers by the authors. Here, similarity is roughly defined as the correspondences between keyword lists from both languages, but special attention is paid for the treatment of low-frequent pairs since these pairs contain both semantically incorrect pairs, which need to be removed, and expressional variations, which should preferably be maintained. In our approach, we first generate a keyword graph representing keywords as nodes and their translation pairs as links, and then effectively search for erroneous links using a minimum-cut detection algorithm in graph theory to generate final clustering results. The proposed method is applied to a set of Japanese and English keywords extracted from real-scale academic conference papers to examine the effectiveness of the generated clusters through the analysis of the detected errors.

1. はじめに

複数の言語あるいは情報源をまたがる横断的な情報検索においては，類似する意味の用語をまとめた類義語クラスタが重要な役割を果たす．学術分野においては学問領域に固有の用語が多く変遷も激しいため，このような語彙的資源を人手によらず自動構築することが望まれる．

ここで，学術用語は一般語と比較して言語間の意味的対応づけが明確である場合が多く，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，学術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，論文の和英著者キーワードに注目すると，大量のデータが既存のデータベースに登録されており，処理コストのかかる形態素解析やアラインメント処理を行わなくても，精度の良い対訳データを簡単に取り出すことができる．さらに，研究者が自ら提示する著者キーワード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もある．しかしながら，著者キーワードを対訳コーパスとして積極的に利用しようという試みはこれまでほとんどなされておらず，コーパスとしての性質やそ

† 学術情報センター研究開発部
Research & Development Department, National Center for Science Information Systems

れにあわせた対訳抽出法についても検討は行われていない．

以上の背景に基づき本論文では，学術文献データベースの著者キーワードを利用して，和英混在の学術用語クラスタを作成する試みについて述べる[1)]．具体的には，和英キーワード間の対訳関係を利用して，共通の訳語を持つ用語どうしを類義語と見なすことによりクラスタを生成するが，この際に，著者による希少な訳例をなるべく保存しつつ，対応誤りを効率良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，横断的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 章で，対訳用語コーパスとしての和英著者キーワードの特性を調べる．これに基づき 3 章では，和英著者キーワードからの対訳関係の抽出において，相互情報量や LSI などの従来手法の適用が困難である理由を述べる．次に 4 章で，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，5 章で，実際の学術文献データベースから抽出した情報に基づき作成した用語クラスタを分析して評価を行う．最後に 6 章で今後の課題を述べる．

2. コーパスとしての和英著者キーワード

2.1 対訳関係の精度

一般に，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ない．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[2)]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☆28,122 件のうち，和英両方の著者キーワードを持つ 27,339 件（約 97.5%）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対訳関係の精度を調べた．

まず，和英キーワードの数が異なる文献は 27,339 件のうち 1,339 件（約 5%）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たため，和英の自動的な対応づけは困難であると判断した．キーワード数が等しい残りの文献について，和英それぞれのキーワードリストの先頭から出現順に対応をとって得られた 112,364 個の用語対について，無作為に 1,000 個を抽出して，人手により対応関係を以下の 5 通りに分類した．

表 1 和英キーワードの対応関係の分析結果
Table 1 Result of classification of Japanese-English keyword pairs.

対応関係の種類	1,000 サンプル中の出現数 N				
	$f=1$	$f=2$	$f=3$	$f\geq 4$	計
(1) 正しい対応	240	146	101	370	856
(2) 表記誤り	22	1	1	1	25
(3) 希少表現	30	7	2	8	47
(4) 関連語	39	3	2	8	53
(5) 誤対応	19	0	0	0	19
計	350	157	106	387	1,000

(1) 訳語として正しく対応がとれたもの
例.〈情報検索, *information retrieval*〉

(2) 訳語としての意味的な対応は正しいが，スペル誤りや誤字など表記上の誤りを含むもの
例.〈情報検索, *information retreival*〉

(3) 訳語としての意味的な対応は正しいが，一般的でない表現が用いられているもの
例.〈情報検索, *information retrieving*〉

(4) 意味的に関連はあるが，訳語としては適切でないもの
例.〈情報検索システム, *information retrieval*〉

(5) 明らかな対応誤りと見なされるもの
例.〈キーワード, *information retrieval*〉

分類の結果を**表 1** に示す．表の中では，用語対全体の中での出現頻度を f として，$f=1,2,3$，$f\geq 4$ について，(1)〜(5) の各カテゴリが出現した回数 N およびその合計値を示している．

表 1 の結果から，(1)，(2)，(3) をあわせると約 93%の和英用語対が意味的に対応しており，一方，(5) の明らかな対応誤りは約 2%と少ないこと，$f=1$ の用語対は全体の約 35%を占め，その中で約 70%が正しい意味関係にあることから，頻度の低い用語対にも多くの有用な情報が含まれていることが分かる．また，(5) の場合については f の値がすべて 1 であることから，同じ対応誤りが再現される確率はきわめて低いといえる．

ここで，(2) の表記誤りや (3) の希少表現は，専門用語として標準的な辞書に登録されることはないが，現実には，利用者からの検索要求や OCR で読み込まれた文書など多くの電子テキストに普遍的に存在するものである．本論文で目的とする情報検索への適用を考慮すると，訳語として正しい (1) に加え，意味的な対応づけが正しい (2)，(3) についても，再現率の観点から保持することが有用である．一方，意味的に関

☆ 情報処理学会および人工知能学会で発表された文献.

表 2 既存の専門用語辞書と対訳用語コーパスの比較
Table 2 Comparison of the technical dictionaries and the bilingual keyword data.

	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	−
和英同一表記数	57	1,336	18

連のない対応である (5) については，適合率の観点から検出して取り除く操作が重要であると考えられる．また，意味的な関連性を持つ (4) については，対象分野に応じて適不適を判定する必要があると考えられる☆．

2.2 既存の専門用語辞書との比較

上記により和英キーワードの数が等しい文献について単純に出現順に和英の対応をとれば，高い精度の対訳関係が得られることが確認できた．以下本論文では，このようにして得られた 112,364 個の用語対（異なりでは 60,186 個）を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」として用いることとし，以下，これに基づく類義語クラスタの抽出法を検討する．

ここで，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，「人工知能大辞典項目」[3] 目次および索引データ，「人工知能ハンドブック」[4] 索引データ，「コンピュータ大百科」[5] 索引，「情報処理用語大辞典」[6] 索引の 4 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 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が比較的少ないことが分かる．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，(i) データベースに登録された著者キーワードには入力誤りに起因する表記誤り（すなわち表 1 における (2) のタイプの誤り）が含まれること，(ii) 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，著者固有の用法や最新の用語を含むこと，の両者の影響であると考えられる．また，著者キーワードでは和英が同一表記であるものが多いことも特徴である．これは日本語表記が存在しない固有名詞などを多く含むためであると考えられる．

なお，表中の数値は異なり数であるが，のべ数で換算すると，正規化処理前でコーパス全体の約 10%，処理後で約 30%が標準辞書と共通の用語対となっている．これは頻度の高い用語対ほど標準辞書に含まれる確率が高いという傾向に対応している．

2.3 スパース性

本論文で扱う対訳用語コーパスでは，対訳として共起する語の数が，和用語に対して平均 1.62 語，英用語に対して平均 1.21 語と少なく，相互情報量や LSI といった共起頻度に基づく従来の分析手法の適用が困難であることが予想される．実際に，対訳用語コーパスの作成に用いたのと同じ文献を対象として，和文表題，和文著者キーワード，和文抄録よりなるテキスト領域中で共起する語の平均数を比較すると 378 語となっており[7]☆☆，総語数に対する共起語数の比率を見ても 2 桁の違いがあるなど，対訳用語コーパスがきわめてスパースであることが分かる．

3. 既存の統計的手法の適用における問題点

3.1 共起スコア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，あらかじめ対応づけされた領域内での共起スコア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語対の共起頻度を求め，その対応のもっともらしさを ϕ^2 統計量[9]，相互情報量[10),11)]，t スコア[11),12)] Dice 係数[13),14)]，などの統計的類似度尺度で評価するものである．

これら共起スコアに基づくアプローチの目的は，文レベルなどでゆるく対応づけされた対訳コーパスから「正しい」すなわち相対的に頻度の高い対訳関係を識別することにある．しかしながら表 1 の結果が示すように，本論文で前提とする対訳用語コーパスの対応づけの大半は正しいものであり，同様の手法を適用すると，相対的に頻度の低い対応については誤訳だけではなく，正しい対訳関係にあるものや希少な訳例を含めてすべて一様に無視されてしまうことになる．

さらに当然のことながら，対訳用語対の中での共起だけに注目すると，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類義性は

☆ 特に (4) において頻度が高く現れている用語対に関しては，〈計算量〉を〈*complexity*〉に対応させるなど，分野を特定すると必ずしも不適切とはいえない場合が多い．

☆☆ 具体的には，テキストを形態素解析ツール（CHASEN Ver1.5[8]）を用いて解析したのちに簡単な複合語処理を行い，抽出されたすべての出現語について同一文献中で共起する語の数の平均値を求めた．

判定できないという問題がある．

3.2 ベクトル空間表現に基づく類義関係の抽出

一方，情報検索の分野で従来より用いられてきた自動索引抽出の手法を適用して，語を，その語と共起する語集合を用いて特徴付け，ベクトル空間上に配置する方法がある．文献中に出現する語の頻度ベクトルに基づく文献間の類似度計算を，テキスト領域中で共起する語の頻度ベクトルに基づく用語間の類似度計算に置き換えたもので，近接する語が似ている語は類似するという仮定のもとで類義関係を抽出するものである．

特に近年，LSI（Latent Semantic Indexing）[15]を対訳コーパスに適用することによって多言語混在のベクトル空間を生成する試みが報告されている[16),17)]．LSI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．直接用語を共有しない文献間でも類似度計算が可能であり，対訳コーパスに適用する場合には，コーパス中での出現頻度が高い用語に関しては，異なる言語間だけではなく，同一言語内での類義関係を含めて，見通し良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標記揺れや誤訳を数値的に区別することは困難である．

3.3 辞書定義文を利用した関連度計算

コーパスからの対訳および類義関係の自動抽出とは視点が異なるが，本論文の提案手法と関連を持つものとして，人手により構築された辞書やシソーラスに基づく関連度計算がある．語をノード，語の間の関係をリンクに対応させたネットワークあるいはグラフ的な表現を用いるもので，たとえば，EDR 辞書[18),19)]，LDOCE[20)]，Collins English Dictionary[21)]，WordNet[22)]を利用した例がある．具体的な関連度の計算法は，シソーラス上で定義された階層関係に基づく手法[18),19)]，辞書から自動抽出した意味ネットワーク上での活性伝搬に基づく手法[20),21)]など多様である．コーパス中での頻度情報をあわせて用いる場合もある[22)]．対訳辞書に適用した例としては市販の和英/英和辞典や EDR 対訳辞書を利用した試みがあり[23)]，和英双方の言語における見出し語と語義の対応関係を手がかりに，精度良く語義対応が抽出されたことが報告されている．

これらのアプローチはコーパスに基づく方法と対比させて論じられることが多いが[19),21)]，本論文で想定する対訳用語コーパスは，精度が高くスパースであるという点で辞書に近い性格を持っており，類似のグラフ的な視点が有用であると考えられる．

表 3 和英著者キーワードによる対訳の例
Table 3 Example of keyword pairs extracted from bilingual keyword lists.

和用語	英用語	出現頻度
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

4. 用語グラフに基づく多言語用語クラスタの生成

4.1 対訳関係に基づく初期用語グラフの生成

以上の背景に基づき，用語のグラク表現に基づくクラスタ化手法について検討する．このためにまず，コーパス中に含まれる和英用語をノード，対訳関係をリンクと見なして，コーパス全体を 1 つの「用語グラフ」で表現する．用語グラフ上の各リンクの容量は，対応する対訳関係のコーパス内での出現頻度とする．定義より，用語グラフ上のリンクは必ず和用語ノードと英用語ノードを連結する形となり，同一言語の用語ノード間にリンクは存在しない．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て，用語グラフ上で互いに連結された部分グラフを 1 つの「用語クラスタ」に対応させる．

表 3 に，本節における説明のために対訳用語コーパスから抜粋した対訳の例を，**図 1** に表 3 の対訳に基づき作成した初期用語グラフを示す．こ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用語クラスタはただ 1 つであり，初期段階ではすべての用語が類義であると見なされることになる．

4.2 誤り候補となる対訳リンクの検出

2 章で述べた対訳用語コーパスの 60,186 個の和英用語対を用いて上記の用語グラフを作成すると，全対

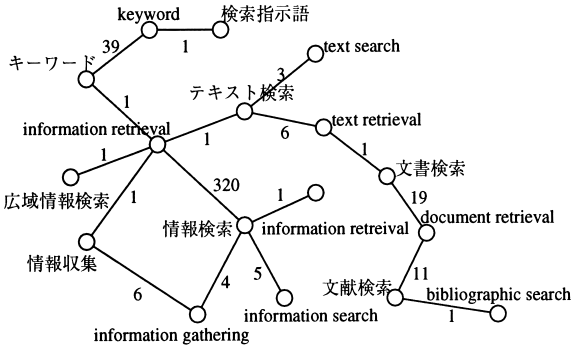


図 1 例題に基づき作成した初期用語グラフ
Fig. 1 Example of initial keyword graph.

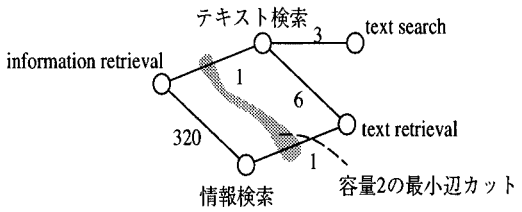


図 2 最小カット数が 2 となる用語クラスタの例
Fig. 2 Example of a keyword cluster with two simultaneous translation errors.

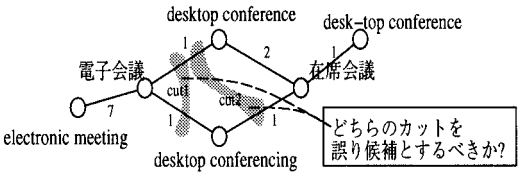


図 3 複数の最小辺カットを持つクラスタの例
Fig. 3 Example of a cluster with conflicting two minimum edge cuts.

訳の約 34%にあたる 20,659 対が 1 つのクラスタに含まれてしまうことが分かる．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 1 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．本論文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」

ことを仮定して誤り候補の検出を行う．グラフ上の 2 点間を結ぶ辺カットは，2 点間を結ぶ重複しない経路の容量の総和に等しく，用語グラフ上では，2 つの語間の対訳関係による結び付きの強さを表現していると見なせる．

さらにグラフ理論では，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，すべての辺カットのうちで最も容量が少ないもの，すなわち与えられた用語グラフ上で，最も対訳関係の結び付きが弱いリンク集合に対応している．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[24]．これらのアルゴリズムは基本的に，自己ループを持たない有向グラフ（リンクに方向が与えられたグラフ）を想定したものである．本論文における用語グラフは無向グラフであるが，無向グラフについても，各リンクを互いに逆向きの 2 本の有向リンクに置き換えることで,「有向グラフ」と同様に最小辺カットが検出できることが知られている．また，本論文における用語グラフは，和英が同一表記である場合にも和用語と英用語は異なるノードで表現していることから「自己ループを持たない」という条件を満足している．

上記に基づき現在の実装システムでは，グラフ上の 2 点間で，つねに最短の経路を優先しつつ辺カットを含む経路集合を構成する Edmonds&Karp の方法を適用し，得られた経路集合から最小辺カットを定めている．この単純な方法でもノード数 n，リンク数 m に対して $O(nm^2)$ で実行可能であり，実用上十分な処理速度が得られている．また現在の実装では，可能なすべての最小辺カットを求めることはしていないが，このための効率的なグラフアルゴリズムの存在も知られている[24]．

最小辺カットは任意個の辺集合で与えられるため，複数の対訳誤りに対応する場合もある．たとえば**図 2** に示すクラスタが得られたとすると，これをさらに分割するためには ⟨ テキスト検索, *information retrieval* ⟩ ⟨ 情報検索, *text 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．

また，最小辺カットは唯一に定まるとは限らない．特に，あいまい性の解消が必要になるのは，**図 3** に例を示すように，リンクを共有する複数の最小辺カットが存在し，いずれかの 1 つを削除の候補として選ばなければならない場合である．このような場合に現在の実装では，単純に文字列の類似度に基づき判定を行っている．たとえば図 3 において，文字列 'desktop conferencing' は 'electronic meeting' よりも 'desktop conference' に近いことから，⟨ 在席会議, *desktop conferencing* ⟩ ではなく ⟨ 電子会議, *desktop conferencing* ⟩ を削除候補として選択する．今後の拡張として，形態素レベルでの情報を利用することも検討している．

4.3 用語クラスタの分割手順

最小辺カットを取り除くことにより，用語クラスタを複数個のクラスタ集合に分割することができる．提案手法では，このような分割は再帰的に適用し，すべてのクラスタについてそれ以上分割が行えなくなるまで，グラフアルゴリズムの適用による削除リンクの検出とクラスタ分割の手順を繰り返す．

ここで，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見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，そ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．このような点を考慮して提案手法では，コーパス内での出現頻度を手がかりに，以下の手順に従ってクラスタ分割の条件を設定する．

(1) **削除の対象とならない対訳リンクのマーク**

次のいずれかにあてはまる対訳リンクは正しいと仮定して削除の対象から除外する．

(a) 和英用語が同一表記
(b) コーパス中の出現頻度が N_α より大きい
(c) 和英いずれかの用語について，その対訳リンク自身が唯一の対訳関係にある

(a) は人名やシステム名などの固有名詞，略記などであり，全体の約 2%の対訳リンクがこれに相当する．(b) は出現頻度を手がかりに対訳の正しさを判定するものである．表 1 においてすべての対応誤りと約 85%の関連語対応の出現頻度が 3 以下であることから $N_\alpha = 3$ に設定すると，全体の約 8%の対訳リンクがこれに相当することになる．(c) は和英すべての用語が必ず 1 つは対訳を持つための条件で全体の約 81%がこれに相当する．本論文で例題として用いる対訳用語コーパスでは，最終的にこれらを除外した全体の約 9%が削除の対象となる対訳リンクとなる．また，用語グラフ上のリンクは必ず和英の用語ノードを端点とすることから，(c) により，分割後もすべてのクラスタは，最低 1 つの和用語と英用語を含むことになる．

(2) **重要語のマーク**

以下の条件を満足する語は，その分野で専門語として一般的に用いられる重要語としてマークする．

(d) 頻度が N_β よりも大きい対訳を少なくとも 1 つ持つ

クラスタ分割においては，分割の結果得られるすべてのクラスタについて，上記の条件を満たす重要語が 1 つ以上存在することを分割の条件とする．重要語をあらかじめマークしておくことによって，最小辺カットの検出アルゴリズムの適用回数を減らすことができる．すなわち，最小辺カットの検出では，クラスタ内から選んだ任意の 2 つのノード間を切断する最小辺カットを求める手順を繰り返すため，あらかじめクラスタの核となりうる N 個の重要語をチェックしておけば，実装上は $(N-1)$ 個の重要語の組合せについて最小辺カット検出アルゴリズムを適用すれば十分である．

(3) **最小辺カットの検出**

最後に，検出した最小辺カットが以下を満足する場合に削除の対象とする．

(e) 最小辺カットを構成する対訳リンクの出現頻度の和が N_ϵ 以下

すなわち，最小辺カットの容量が N_ϵ よりも大きい場合は，クラスタ内の重要語どうしの結び付きが強いことから，これらの語は類義であると見なす．

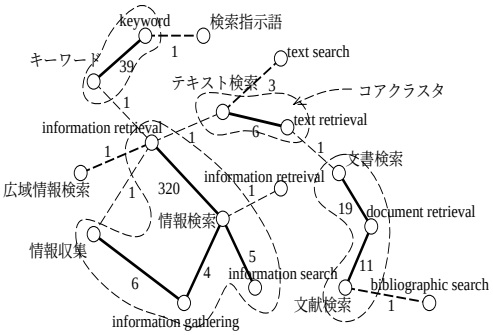
以上をまとめると，クラスタの分割において，あらかじめ指定するパラメータは N_ϵ，N_α，N_β の 3 つである．これらのパラメータによって指定される分割条件に従って，すべてのクラスタについてそれ以上分割が行えなくなるまで再帰的にクラスタ分割を繰り返す．N_α の値はコーパスの性質に応じてクラスタによらず定まると考えられることから，現在は $N_\alpha = 3$ に固定している．N_β，N_ϵ の効果については次節でより詳細に調べるが，定義より N_β については値が小さいほど，N_ϵ については値が大きいほど，クラスタの分割が進むことが予想される．

4.4 クラスタ分割の例

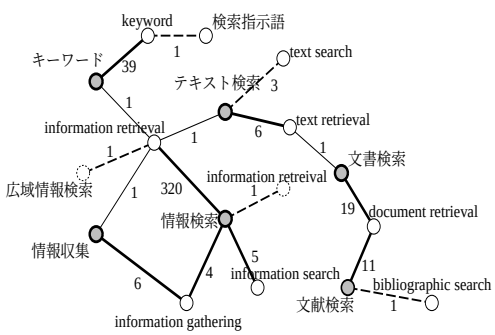
図 4 に，$N_\epsilon = N_\alpha = N_\beta = 3$ として図 1 の用語クラスタを分割した結果を示す．

まずステップ (1) では，上記の条件 (a)，(b)，(c) のいずれかを満足する対訳リンクを削除の対象から除外する．条件 (b) から太い実線で示した⟨情報検索, *information retrieval*⟩などの対訳リンクを，条件 (c) から太い破線で示した⟨検索指示語, *keyword*⟩などの対訳リンクをマークしている．直感的には，ステップ (1) においてマークした削除不能な対訳リンクによって，破線で囲んだような核クラスタが生成され，これらのクラスタをまたがるような対訳リンクだけが後のステップで誤りとして検出されることになる．

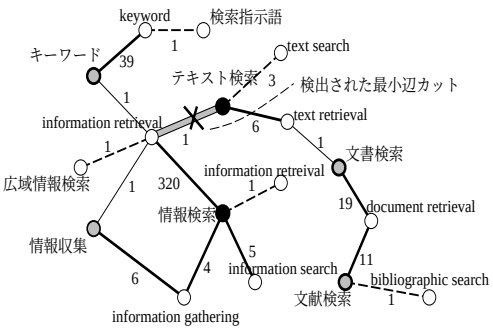
ステップ (2) では，条件 (d) を満足する⟨キーワード⟩や⟨情報検索⟩などの語をマークしている．ステップ (3) では，⟨情報検索⟩と⟨テキスト検索⟩の間で最小辺カットの検出アルゴリズムを適用し，その結果



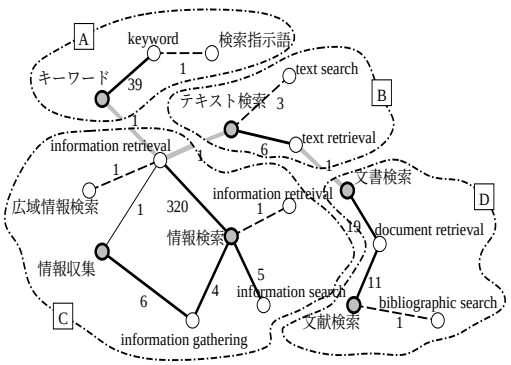
(a) ステップ (1)：削除不能な対訳リンクのチェック



(b) ステップ (2)：重要語のチェック（図中では和用語のみ強調表示）



(c) ステップ (3)：2 つの重要語の間の最小辺カットの検出



(d) 最終的なクラスタ分割結果

図 4 クラスタ分割手順
Fig. 4 Cluster partitioning procedure.

⟨ テキスト検索, *information retrieval* ⟩ を削除するべき誤りとして検出している．

以上の繰返しによって，⟨ 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る．⟨ 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．

ここで $N_\beta = 10$ とすると，グループ B は重要語を 1 つも含まないことから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，4.2 節で述べたように，各グループに含まれる用語の文字列類似度から判定することになる．

4.5 用語クラスタの出力例と情報検索への適用例

学会発表データベースから取り出した 60,186 個の和英用語対に対して，$N_\alpha = N_\beta = N_\epsilon = 3$ の条件で提案手法を適用したところ，合計 28,077 個の用語クラスタが得られた．これらのクラスタに，表記揺れや希少表現がどのような形で含まれるかを具体的に例示するため，「発想支援」という語を含むクラスタを取り上げ，クラスタ内に含まれる語を以下に列挙する．括弧内の数値は各用語のコーパス中での出現頻度である．用語の統一がされていない比較的新しい専門用語についても，意味的に関連する用語を 1 つのクラスタに集めることが可能であることが分かる．

例 1： 「発想支援」を含む和英用語クラスタ
和： 発想支援 (81)，思考支援 (24)，発想法 (14)，創造性支援 (4)，創造的思考 (4)，ユーザの興味 (3)，コンセプトメイキング (1)，アイデア生成 (1)，考えの筋道 (1)，忠考過程 (1)，コンセプトメイキング (1)，視覚的思考法 (1)，知的生産 (1)，アイディアプロセッシング

(1)，発想的思考 (1)，概念形成支援 (1)
英： idea generation (17), creativity support (12), idea generation support (10), creative thinking (8), creative thinking support (7), idea processing (7), thinking support (6), computer aided thinking (4), concept making (3), enhancing thought (3), creation support (3), idea generation method (3), support of concept formation (2), visual thinking (2), the way of thinking (2), support for thinking (2), user's interests (2), conception support (2), the ways of thinking (2), idea promoting (2), thinking aids (2), concept formation support tool (1), idea promotion (1), idea promoting system (1), abduction support (1), creative work support (1), enhancement thinking process (1), computer-aided thinking (1), idea processing support (1), idea stimulation (1), knowledge emergence (1), support for idea generating (1), computer aided abduction (1), abduction-method (1), idea creation support (1), new-idea generation support (1), idea generating support (1), idea accelarator (1), computer aided motin acquisition (1), thinking supporting (1), thought assist (1), aiding creative concept formation (1), computer aided creation (1), creative concept formation aid (1), idea acceleration (1), idea process suppont (1), idea jeneration (1), thinking enhancement (1), intelligence amplifier (1), support for creativity thinking (1), new idea generation support (1), interests of users (1), user interests (1), support for creating conception (1), computer-aided creativity (1), a support system for good ideas (1), thinking aid(1), augmenting human intelligence(1), support for generating ideas.(1), thinking method (1)

生成したクラスタには平均で約 2.14 個の和英用語対が含まれており，3 種類以上の用語対を含むクラスタが全体の約 14%存在した．また，ただ 1 つの用語対からなるクラスタが全体の約 69%存在し，2.1 節の分析により，その大半は正しい対応づけであると予想される．これより，生成した用語クラスタは，高頻度語に関して例 1 に示したような多様なバリエーションを保持するとともに，標準辞書にはない多くの低頻度語に対しても訳例を提供していることが分かる．これらは検索語の入力としても想定されるものであり，単一言語検索あるいは言語横断検索における検索語拡張への適用が期待される．

文献 25) では，上記で生成した 28,077 個の用語クラスタを，本論文で対象としたものと同じ学会発表データベースを用いた「NACSIS 版情報検索テストコレクションの試験バージョン」[26]に適用した結果について報告している．ここでは，題目や抄録などの情報が各文献ごとに和文，英文の両方で登録されていることを利用して，(1) 日本語の検索文に対してデータベースの和文登録情報を参照して検索を行う「単一言語検索」，および，(2) 日本語の検索文に対してデータベースの英文登録情報を参照して検索を行う「言語横断検索」，の 2 つのタスクについて，再現率に対する適合率の 11 ポイント平均によって検索性能を評価している．

表 4 クラスタサイズに対するクラスタ化パラメータの影響
Table 4 Change of cluster sizes against different clustering parameters.

N_ϵ	N_β	クラスタ総数	最大クラスタ	削除リンク数
1	1	28,067	5,011	1,415
1	3	27,637	6,067	887
1	10	27,409	7,708	612
3	1	29,103	332	2,567
3	3	28,077	499	1,437
3	10	27,738	539	1,035
10	1	28,988	161	2,837
10	3	27,991	192	1,672
10	10	27,670	404	1,223
削除なし		26,983	15,409	0
すべて削除		30,991	131	6,393

文献 25) によると，得られた用語クラスタ内のすべての語を新たに検索語として加える場合，単一言語検索については平均 9.7 語，言語横断検索については平均 5.8 語の検索語が新規に追加された．検索性能の評価では，単一言語検索について，検索語の拡張を行わない場合と比較して約 10%の性能改善が得られた．また言語横断検索については，同じ検索文により同じ文献の（英文ではなく対応する）和文登録情報を参照して検索を行う単一言語検索と比較して 50%程度の性能が達成された．検索方式が適合性フィードバックを用いない単純なものであることを考慮すると，この数値は，生成した用語クラスタの情報検索における有効性を示すに十分な性能であるといえる．具体的な検索手法や評価の詳細については文献中で詳しく報告されている．

5. 結果の分析

5.1 生成されるクラスタの大きさ

クラスタの分割における N_β，N_ϵ の影響を調べるため，これらのパラメータの値をそれぞれ 1，3，10 の 3 通りに設定した場合について，(a) 最終的に得られる和英用語クラスタの総数，(b) 最大クラスタに含まれる用語の数，(c) 削除された対訳リンクの数を比較した．その結果を**表 4** にまとめる．参考のため，対訳リンクの削除を行わなかった場合（「削除なし」），削除対象となる 6,393 個の対訳リンクをすべて削除した場合（「すべて削除」）についても数値を示している．

表 4 の結果より，N_β の値が小さく N_ϵ の値が大きいほど削除される対訳リンク数が多く，クラスタが細

分化されることが分かる．また，削除の候補となるリンクの数は全体の約 9%とそれほど多くはないが，最大クラスタの大きさの違いに注目すると，その存在の有無はクラスタリングの結果に強い影響を持つことが分かる．

なお，これらのクラスタ分割の処理に要する時間は，Ultra2 ワークステーション（200 MHz）で 3 分から 4 分程度であり，グラフアルゴリズムの適用により高速な処理が実現されていることが分かる．

5.2 誤り率と検出率

提案手法では，削除するべき対訳リンクを正しく選択することが，最終的に得られる用語クラスタの質を左右する主要因となる．

ここで問題となる判定誤りとしては，(a) 意味的に正しい対訳関係を削除してしまう場合，および，(b) 意味的に誤りである対応関係の検出に失敗する場合の 2 通りが考えられる．正しいにもかかわらずリンクを削除すると，類似の意味を持つ複数のクラスタが生成されてしまうことになる．一方，検出されない誤りリンクの存在は，異なる意味の単語が同一クラスタに分類されていることを意味する．

実際のクラスタ分割における判定誤りの程度および分割パラメータの影響を調べるため，削除の対象となりうる 6,393 個の対訳リンクのうち，和英用語がいずれも重要語である 2,110 個を人手により調べ，以下に示す 3 つのクラスに分類した．

（1） 意味的に正しい対応 ··· 758 個（36%）
（2） 関連が深い語どうしの対応 ··· 781 個（37%）
（3） 意味的な対応誤り ··· 569 個（27%）

ここで第 1 のカテゴリは表記誤りおよび希少表現を含むものである．

次に $N_\beta, N_\epsilon = 1, 3, 10$ なる組合せについて，(a) これら 2,110 個のリンクのうち削除されたものの総数（「合計」），(b) その種類の内訳（「正しい」「関連」「誤り」）を求め，これに基づき，(c) 正しいにもかかわらず誤りとして削除してしまったリンク数の削除総数に占める割合（「エラー率」），および (d) 誤りとして検出したリンク数の誤り総数の中に占める割合（「検出率」）を計算した．その結果を **表 5** にまとめる．

いずれの組合せでも，検出した誤りのおよそ 80%から 90%が明らかな誤りまたは関連語誤りとなっており，正しい対訳関係についてはコーパス全体では 85%を占めるにもかかわらず，検出された誤りの中では $N_\beta > 1$ とすれば数%程度におさえられることが分かる．また，誤りリンクの検出率は 36%から 90%とばらつきが大きく．特に $N_\beta = 1$ とした場合に高い検出率を示すことが分かる．

表 5 エラー率と検出率によるクラスタ化のパラメータの比較
Table 5 Change of error and detection rates against different clustering parameters.

N_ϵ	N_β	合計	正しい	関連	誤り	エラー率	検出率
1	1	640	46	246	348	7.2%	61.2%
1	3	454	21	159	274	4.6%	48.2%
1	10	319	15	98	206	4.7%	36.2%
3	1	1,126	206	522	498	18.3%	87.5%
3	3	775	47	327	401	6.1%	70.5%
3	10	557	29	209	319	5.2%	56.1%
10	1	1,357	229	615	513	16.9%	90.2%
10	3	916	61	436	419	6.7%	73.7%
10	10	663	37	292	334	5.6%	58.7%

5.3 正しい対応関係の削除

これらの判断誤りの原因や影響をさらに分析するため，まず，正しい対訳関係を削除してしまう場合について，9 通りのパラメータの組合せの少なくとも 1 つで削除された 228 個の正しい対訳関係について調べたところ，以下のような内訳が得られた．

（1） 異表記によるもの ··· 73 個 (32%)
（2） 多義語によるもの ··· 64 個 (28%)
　（a） 略語の多義 ··· 27 個
　（b） 一般語の多義 ··· 35 個
　（c） 専門語の多義 ··· 2 個
（3） コーパス誤りによるもの ··· 91 個 (40%)
　（a） 頻度 1 の誤り ··· 87 個
　（b） 頻度 2 以上の誤り ··· 4 個

異表記の多くは，英和専門用語を対応させる際に，語尾の変化や語順の対応づけが慣例としてあいまいになっていることに起因するもので，具体例として〈処理〉を〈*process*〉と〈*processing*〉の両方に対応させる場合などがあげられる．一方，多義語に対する対訳のいずれかが誤りと判定される場合についてみると，略語および一般的な単語に起因するものが大半を占め，具体例としてそれぞれ，〈*TP*〉が〈トランスポートプロトコル〉と〈トランザクション処理〉の両方に対応する場合，〈評価〉や〈*performance*〉などの一般的な単語に複数の意味が対応する場合などが観察されている．複合的な語の同義による削除は 2 例だけであった．いずれの場合についても，多義的な対応の一方を削除することで異なる意味を持つ用語集合が異なるクラスタに分割され，多義性の問題が解決されている．さらに，コーパス自体の誤りに関して，4 件の例外を除くすべてが $N_\beta = 1$ としたときに得られるものであった．この中には例 2 に示すように，リンク削除の結果として，誤りや一般的でない対訳をただ 1 つ含む

クラスタが分離される場合が多く観察された．

例 2：〈遺伝子アルゴリズム，*genetic algorithm*〉の削除により生成されるクラスタ
クラスタ 1：遺伝子アルゴリズム (1)（希少表現），gentic algorithm(1)（綴り誤り）
クラスタ 2：遺伝的アルゴリズム (265)，遺伝アルゴリズム (39)，genetic algorithm(185)，genetic algorithms(99)，GA(14)，（以下，頻度 1 の用語は略）

以上の分析より，判定誤りのうち正しい対応関係を誤りとしてしまう場合については，(i) 略語の処理，(ii) 広義の一般語の処理，(iii) 綴り誤りや表記揺れの処理，の 3 つが有効であると考えられる．(i) について，専門用語の場合には特に，英文字の略語が和文の中でも一般的に用いられることから英文字略語と和用語の同義性の判定は重要であり，クラスタ内の代表的な英用語の頭文字を調べるなど言語的な知識に基づく処理が必要となる．また，(ii) については一般的な標準辞書と対応をとることで，(iii) については単語文字列の類似度を計算することで，大量の情報に対してもある程度機械的に対処できる．さらに，形態素情報を手がかりに語彙的な対応づけをとることも考えられるが，この場合に処理コストの観点から，あらかじめ候補となる組合せを選別したうえで，形態素解析を適用することが重要である．

5.4 誤った対応関係検出の失敗

次に，誤りであるにもかかわらず削除されなかった場合について，9 通りのパラメータの組合せのいずれでも削除されなかった 54 個の対訳関係について調べたところ，以下のような内訳が得られた．

(1) 頻度が高い誤りに起因 · · · 3 個（6%）
(2) 正しい対応の頻度の低さに起因 · · · 51 個（94%）

頻度が高い誤り対応に起因するケースは，用語数が多く総頻度が高いクラスタにおいて生じている．具体例として〈類似〉を〈*analogy*〉および〈*similarity*〉に，〈類似度〉を〈*similarity*〉に対応させるリンクの存在から，〈類似度，*analogy*〉が正しいと認識されてしまう場合などがあげられる．一方，頻度の低い小さなクラスタにおいては例 3 に示すように，正しい対応と誤った対応がともに頻度 1 で出現するために判別ができない場合が多く観察された．

例 3：〈視界，*perspective projection*〉の誤りリンクを含むクラスタ
和用語：透視投影 (3)，透視変換 (3)，中心投影 (2)，視界 (2)，透視写影 (1)
英用語：perspective projection (7)，fieldofview (1)，perspective transform (1)，projective preposition (1)，perspective transformation (1)

主要な原因が頻度情報の不足であることを確認するため，該当する 51 個の誤りについて，用いる文献の数を約 10 倍の 30 万件に増やしてクラスタの生成を行ったところ，半数以上の 27 個が誤りと正しく判別され削除された．

頻度情報の不足を補うための拡張としては，(i) 形態素情報の利用による複合語の意味的類似性の推察，(ii) 同一文献中での共起性を利用するなど他のコーパスより自動抽出した類似性に基づく判定などが考えられる．(i) については，生成したクラスタの整合性のチェック等に用いることが考えられ，今後の課題となっている．また現在，(ii) について検討を行っており，その一段階として，著者キーワード内での共起情報に基づく用語間の距離計算を，テキスト自動分類問題を用いて評価した結果を報告している[7]．

また，以上の分析から，クラスタの総頻度が低い場合と高い場合では判定誤りの性質が異なることが予想される．N_β, N_ϵ の値について，論文中ではすべてのクラスタに共通であると仮定したが，クラスタの総頻度に応じて適応的に定めることが有利であると考えられ，今後の課題となっている．

5.5 関連語の扱い

表 5 の結果から，$N_\beta = 1$ ではエラー率が高くなり，一方，$N_\epsilon = 1$ では検出率が低くなることが分かる．また，検出率は $N_\beta = 10$ とした場合にも大幅に減少する．これより両者に有効なパラメータの値は，$N_\beta = 3$ かつ $N_\epsilon \geq 3$ であると判断できる．

さらに，N_ϵ の影響を調べるために $N_\beta = 3$ で固定したうえで，N_ϵ の値を 3 から 10 まで変化させて，削除リンクの中での「正しい」「関連」「誤り」それぞれのカテゴリの数を比較した．**図 5** に示すように，N_ϵ の値が大きくなっても,「正しい」および「誤り」リンクの削除数はそれほど変化しないが,「関連」リンクの

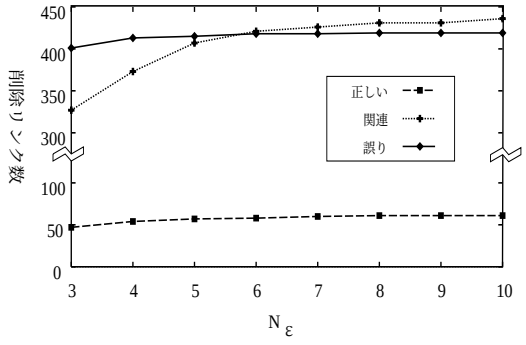


図 5 削除リンクに対する最小カット数の影響
Fig. 5 Change of deleted links against based on minimum cut size.

占める割合が増加し，結果として意味的に関連を持つ語集合が細分化されることが確認できる．N_ϵ の最適値は分析の結果からは定めることができないが，4.5節で述べた情報検索テストコレクションによる評価では，$N_\epsilon = 3$ として良好な結果を得ている．

6. おわりに

本論文では，学術文献の和英著者キーワードを利用した多言語類義語クラスタの自動生成について述べた．和英著者キーワードは，文単位でゆるく対応づけされた通常の対訳テキストコーパスと比較して和英の対応づけが良いことから，統計処理を用いて最適な対訳候補を選ぶ従来手法よりも，類義関係の抽出に悪影響を及ぼす誤対応を取り除く処理が有効であると考え，グラフアルゴリズムを適用して誤り候補を効率的に検出する手法を新たに提案した．

提案手法は従来の統計的手法と対立するものではなく，たとえば統計的手法により取り出した精度の高い対訳関係に提案手法を適用するなど，処理の対象とするコーパスの性質に応じて相補的に使い分けることで，広範囲の適用が期待できる．また，計算時間オーダーが理論的に保証されたグラフアルゴリズムの適用により比較的大規模なデータの処理を可能にしており，形態素レベルでの和英の対応づけなど，処理コストが高い自然言語処理の前処理として用いることも考えられる．

提案手法の利点として，特に頻出語について，豊富な類義例を提供できることがあげられるが，これをそのまま情報検索に適用すると，過剰な検索語拡張により精度の観点から不利になることが実際の評価において観察されている．この問題点に対処する方法として，頻度や最小辺カット容量に基づきクラスタ内の代表語を定め，選択的に検索語拡張を行うことが考えられる．また，情報検索への適用における予備的評価において，あらかじめファイルに出力した類義語クラスタを参照するのではなく，与えられた検索語とクラスタ内の語の間で最小辺カット（すなわち対訳関係の結び付きの強さ）を求めて拡張語を動的に選択する方式が最も高い性能をあげたことから，対話的な検索への拡張も有効であると考えられる．さらに，情報検索における性能向上という観点からは，和英著者キーワードに加えて，和英題目や和英抄録の共起情報を併用することで高い効果が期待できる．

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) Aizawa, A. and Kageura, K.: An Approach to the Automatic Generation of Multilingual Keyword Clusters, *Proc. 1st Workshop for Computational Terminology* (1998).
2) NACSIS: *Introduction to the National Center for Science Information Systems*, Tokyo (1997).
3) Shapiro, S. (Ed.): *Encyclopedia of Artificial Intelligence*, John Wiley, New York (1987). 大須賀節雄（訳編）：人工知能大辞典，丸善 (1991).
4) 人工知能学会（編）：人工知能ハンドブック，オーム社 (1990).
5) Ralston, A. (Ed.): *Encyclopedia of Computer Science and Engineering*, Van Nostrand Reinhold, Amsterdam (1983). 棟上昭男（訳編）：コンピュータ大百科，朝倉書店 (1987).
6) 相磯秀夫（編）：情報処理用語大辞典，オーム社 (1993).
7) 相澤彰子，影浦　峡：学術文献の著者キーワードに基づく専門用語間の関連度計算とその応用，情報処理学会研究会報告 NL-131, pp.55–62 (1999).
8) 松本裕治，北内　啓，山下達雄，平野善隆，今一修，今村友明：日本語形態素解析システム CHASEN Vresion 1.5 使用説明書，NAIST Technical Report NAIST-IS-TR97007, 奈良先端科学技術大学院大学 (1997).
9) Gale, W.A. and Church, K.W.: Identifying Word Correspondences in Parallel Texts, *Proc. DARPA Speech and Natural Language Workshop*, pp.152–157 (1991).
10) Eijk, v.d.P.: Automating the Acquisition of Bilingual Terminology, *EACL'93*, pp.113–119 (1993).
11) Fung, P.: A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora, *ACL'95*, pp.233–236 (1995).
12) Ahrenberg, L., Andersson, M. and Merkel, M.: A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts, *COLING-ACL'98*, pp.29–35 (1998).
13) Smadja, F., McKeown, K.R. and Hatzivassiloglou, V.: Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics*, Vol.22, No.1, pp.1–38 (1996).
14) 北村美穂子，松本裕治：対訳コーパスを利用した対訳表現の自動抽出，情報処理学会論文誌，Vol.38, No.4, pp.727–736 (1997).
15) 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).
16) Schütze, H. and Pedersen, J.O.: A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval, *Information Processing & Management*, Vol.33, No.3, pp.307–318 (1997).
17) Kikui, G.: Term-list Translation using Monolingual Word Co-occurrence Vectors, *COLING-ACL'98*, pp.670–674 (1998).
18) 湯浅夏樹，上田 徹，外川文雄：大量文書データ中の単語共起を利用した文書分類，情報処理学会論文誌，Vol.36, No.8, pp.1819–1827 (1995).
19) 福本文代，鈴木良弥，福本淳一：辞書の語義文を用いた文書の自動分類，情報処理学会論文誌，Vol.37, No.10, pp.1789–1799 (1996).
20) Kozima, H. and Furugori, T.: Similarity between Words Computed by Spreading Activation on an English Dictionary, *EACL'93*, pp.232–239 (1993).
21) Niwa, Y. and Nitta, Y.: Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries, *COLING'94*, pp.304–309 (1994).
22) Jiang, J. and Conrath, D.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, *COLING'97* (1997).
23) Tokunaga, T. and Tanaka, H.: The Automatic Extraction of Conceptual Items from Bilingual Dictionaries, *PRICAI'90*, pp.304–309 (1990).
24) 永持 仁：グラフの最小カット，離散構造とアルゴリズム II，藤重 悟（編），chapter 4, pp.371–381, 近代科学社 (1993).
25) Kando, N. and Aizawa, A.: Cross-lingual Information Retrieval Using Automatically Generated Multilingual Keyword Clusters, *IRAL'98* (1998).
26) Kando, N., Koyama, T., Oyama, K., Kageura, K., Yoshioka, M., Nozue, T., Matsumura, A. and Kuriyama, K.: NTCIR : NACSIS Test Collection Project [Poster], *The 20th Annual BCS-IRSG Collquium on Information Retrieval Research* (1998).

(平成 11 年 6 月 11 日受付)
(平成 12 年 1 月 6 日採録)

相澤 彰子（正会員）

1985 年東京大学工学部電子学科卒業．1990 年同大学大学院電気工学専攻博士課程修了．工学博士．1990 年から 1992 年，イリノイ大学アーバナ・シャンペイン校客員研究員．現在，学術情報センター助教授．情報システム，遺伝的アルゴリズム等の研究に従事．

影浦 峡

1986 年東京大学教育学部卒業．1993 年マンチェスター大学 Ph.D. 1996 年シェフィールド大学客員研究員．現在学術情報センター助教授．専門用語の研究に従事．