技術文書を対象とした言語横断情報検索のための複合語翻訳

藤井 敦† 石川徹也†

近年，検索質問と異なる言語の文書を検索する言語横断情報検索（CLIR）の研究が情報検索や自然言語処理の分野でさかんに行われている．本論文は，専門用語翻訳を用いた日本語と英語の CLIR システムを提案する．専門用語は複合語が多く，既存の語基の組合せによって漸進的に作られるため，対訳を網羅的に翻訳辞書に記述することが困難である．我々は語基単位の対訳を組み合わせて翻訳し，統計的に訳語曖昧性を解消する．学術情報センターが作成した NACSIS テストコレクションを用いた比較実験の結果，本 CLIR システムは単言語検索の性能とほぼ等しいことが確認された．

Translating Compound Words in the Cross-Language Information Retrieval of Technical Documents

Atsushi Fujii† and Tetsuya Ishikawa†

Cross-language information retrieval (CLIR), where queries and retrieved documents are in different languages, has of late become one of the major topics within the information retrieval and natural language processing communities. This paper proposes a Japanese/English CLIR system focusing mainly on the translation of technical terms. Since technical terms, most of which are compound words, can be progressively generated simply by combining existing base words, it is not entirely satisfactory or feasible to exhaustively enumerate newly emerging terms in translation dictionaries. To counter this problem, we use a bilingual dictionary of base words and collocational statistics to resolve translation ambiguity. Experiments using the NACSIS test collection showed that our system is quite comparable with a monolingual IR system in performance.

1. はじめに

言語横断情報検索（Cross-Language Information Retrieval: CLIR）は，検索質問と異なる言語の文書を検索する処理であり，情報検索における翻訳技術の応用と捉えることができる．CLIR の歴史は 1960 年代の文献検索システム[17]や 1970 年代の Salton の実験[20]にまで遡る．近年は，コンピュータネットワークを通じて外国語文書を入手できる機会が増えており，CLIR の研究はますますさかんになっている．ACM SIGIR[1]などの会議や TREC（Text REtrieval Conference）[22]などのコンテストにおいても CLIR は主要なテーマの 1 つである．日本における CLIR のコンテストとして，学術情報センター（NACSIS）が主催する「NACSIS コレクション・ワークショップ」の言語横断検索タスクがある．当コンテストは，技術論文の抄録で構成される「NACSIS コレクション」[13]を検索対象文書として用いる☆．そこで，新聞記事の検索に比べると，専門用語の翻訳が検索性能を大きく左右する．

本論文の目的は，専門用語翻訳に基づく技術論文用 CLIR システムを構築し，NACSIS コレクションを用いて，その有効性を示すことにある．Pirkola[19]は，一般語の対訳辞書と専門用語対訳辞書を組み合わせて検索質問を翻訳し，専門的な文書（TREC コレクションの「health」トピック）の検索性能を向上させた．これは専門用語翻訳の有効性を示す 1 つの例である．しかし，専門用語の多くは複合語であり，既存の語基（形態素）を組み合わせて漸進的に作られるため，対訳を網羅的に辞書に記述することは困難である．そこで，対訳辞書に定義されている語基を適宜組み合わせる必要がある．そのためには以下の 2 つの課題を解決しなければならない．

† 図書館情報大学
University of Library and Information Science

☆ http://www.rd.nacsis.ac.jp/~ntcadm/workshop/work-ja.html

● **課題 (1)**：語基辞書の作成

既存の辞書は語基単位の対訳を必ずしも網羅的に定義していない．たとえば，情報処理関連の専門用語 12 万語を収録した「EDR 日英専門用語対訳辞書」[29]には，「知識抽出（knowledge extraction）」や「特徴抽出アルゴリズム（feature extraction algorithm）」などの複合語は定義されていても「抽出（extraction）」という語基は定義されていない．そこで，「情報抽出（information extraction）」のような新たな語基の組合せを翻訳できない．

● **課題 (2)**：訳語曖昧性の解消

複合語を語基単位に翻訳すると訳語曖昧性が多くなる．訳語曖昧性を解消しないと検索結果に不必要な文書が含まれて検索性能の低下につながる[6),31)]．他方において，1 つの専門用語に複数の訳語が対応することがあるので（たとえば「文書検索」に対して「document retrieval」や「text retrieval」などが対応する），訳語候補に優先度を与えて，優先度が高い複数の候補を検索に用いることが好ましい．

以下，2 章で従来の CLIR の研究について検討し，3 章と 4 章で我々の CLIR システムと複合語翻訳法についてそれぞれ説明する．5 章で評価実験を通して複合語翻訳法の有効性を示し，今後の研究課題について考察する．

2. 先行研究の検討と本研究の立場

本章は，CLIR の先行研究を翻訳方式（2.1 節），検索結果の提示方法（2.2 節），評価方法（2.3 節）の観点から検討して，本研究の立場を明確にする（2.4 節）．

2.1 翻訳方式

従来提案されている CLIR の翻訳方式を，大きく以下の (1)〜(3) に分類する．

(1) **検索質問翻訳方式**

検索質問を検索対象文書の言語に翻訳して検索を行う方式である．検索質問を翻訳した後は，単言語の情報検索と同一の処理を行う．そこで，既存の検索システムを利用でき，実装コストが低いという利点がある．この方式は，翻訳に用いる資源によって以下の 3 つの手法 (a)〜(c) に細分類できる．

(a) 対訳辞書中の訳語をすべて検索に利用する手法[10)]

(b) 2 言語コーパス☆から対訳を抽出して検索に利用する手法[4)]

(c) コーパスを用いて対訳辞書の訳語曖昧性を解消する手法[2),6),31)]

(2) **対象文書翻訳方式**

検索質問ではなく，検索対象文書を翻訳する方式である．翻訳には機械翻訳システムが利用されることが多い[18),27)]．Oard [18)]は，対象文書翻訳方式が検索質問翻訳方式よりも検索性能が良いことを実験によって示している．しかし，データベース中の全文書を翻訳するため，コストが高いことが問題である．

(3) **中間言語表現方式**

検索質問と検索対象文書の両方を中間言語表現に変換して言語の表層的な違いを吸収する方式である．中間言語として 2 言語シソーラスの意味クラスを用いる手法[9),14),17),20),23)]がある．また，単言語検索で用いられるベクトル空間法[21)]を拡張して，言語に依存しない軸でベクトル空間を構成する手法[4),7)]がある．これらの手法は，2 言語シソーラスやコーパスの対応付けを必要とするので，実装コストが高いという問題がある．

2.2 検索結果の提示方法

CLIR の検索文書はユーザの母国語以外で記述されていることがあるため☆☆，効果的な検索結果の提示は単言語検索以上に重要な課題である．鈴木ら[26)]は，検索結果の提示方法を変えながら被験者の検索効率（検索時間，検索結果から正しい文書を選択する割合）を測定する実験を行った．彼らの実験結果をまとめると，以下の 4 種類の提示方法の中で手法 (c) が検索効率を最も向上させることが確認された．

(a) キーワードを翻訳せずに提示する手法

(b) 辞書中の最初の訳語を用いてキーワードを翻訳して提示する手法

(c) キーワードを対訳辞書とコーパスに基づいて翻訳して提示する手法

(d) 検索文書の要約（既存の要約ソフトを利用）を人手で翻訳して提示する手法

ここで「キーワード」は検索文書中の頻出語を指す．手法 (c) の翻訳法は 2.1 節の手法 (1–c) と同じである．

2.3 評価方法

CLIR の検索性能の評価方法は，単言語検索の評価とほぼ同じである．すなわち，あらかじめ用意された検索質問を用いて文書検索を行い，その結果に対して適合率と再現率で評価する．実験データには CLIR 用のテストコレクションを用いることが理想的である．

☆ 従来の CLIR のほとんどが 2 言語を対象にしているので，本論文でも「多言語」ではなく「2 言語」という言葉を用いる．しかし，CLIR は 3 言語以上の検索も含む点に注意する必要がある．

☆☆ 外国語キーワードを用いて母国語文書を検索する場合，この問題は発生しない．

そのようなデータがない場合は，単言語検索用テストコレクションの検索質問を人手で翻訳して用いる場合[27),31)]と検索対象文書を機械翻訳システムで翻訳して用いる場合[27)]がある．しかし，既存のコレクションを翻訳する場合は，翻訳の恣意性が生じることがある．酒井ら[27)]は，複数の人間が（個別に）翻訳した検索質問を用いて実験を行い，翻訳の質が実験結果に影響することを報告している．

検索結果の提示方法については，被験者を用いて検索効率を測定する評価方法（2.2 節）がある．

2.4 本研究の立場

2.1〜2.3 節の議論に基づいて，本研究の立場を明確にする．

まず，翻訳方式については実装コストの低さから，検索質問翻訳方式を採用する．具体的には，適切な訳語曖昧性解消の必要性から，方式 (1–c) を採用する☆．

次に，検索結果の提示方法については，検索文書中のキーワードを翻訳してユーザに提示する．論文などのようにあらかじめ著者キーワードが付与されている場合には，キーワードの特定は容易である．そうでない場合には，検索文書内の頻出語をキーワードとして利用する．

最後に，本研究で提案する CLIR システムの検索性能は，NACSIS コレクションを用いて評価する．このコレクションには，日本語検索質問と正解の英語文書が用意されているので，検索質問や検索対象文書を改めて翻訳しなくても日英 CLIR の検索性能を評価できる．なお，英日 CLIR および検索結果の提示方法の評価は本論文の範囲外であり，今後の研究課題である．

3. CLIR システムの構成

本研究で提案する CLIR システムの構成を**図 1** に示す．システムは主に次の 2 つのモジュールからなる．

3.1 複合語翻訳

「複合語翻訳」モジュールは，与えられた専門用語を対訳辞書を用いて語基に分割しながら翻訳し，語の共起情報を用いて訳語曖昧性を解消する．なお，本モジュールは日本語と英語の双方向の翻訳を行うことができる．共起情報は文書コレクションから抽出する．複合語翻訳には 2 つの役割がある．1 つはユーザが入力した検索質問中の専門用語を文書言語に翻訳することであり，もう 1 つは検索文書中のキーワードをユーザ言語に翻訳することである．

検索質問が文や句で記述されている場合は，専門用語を抽出しなければならない．日本語の検索質問に対しては,「茶筌」[32)]を用いて形態素解析を行い，品詞情報に基づいて内容語（名詞，動詞など）のみを抽出し，連続する内容語を専門用語としてまとめる☆☆．たとえば,「クラスタリングにおける特徴次元リダクション」という検索質問からは,「クラスタリング」と「特徴次元リダクション」を抽出する．英語検索質問に対しては，まず WordNet[16)] を用いて不要語（stopword）を削除する．次に，残った内容語を原形に直して，連続する内容語を専門用語として抽出する．

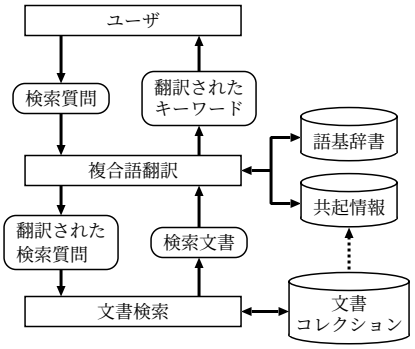


図 1 CLIR システムの構成
Fig. 1 The overall design of our CLIR system.

3.2 文書検索

「文書検索」モジュールは翻訳された検索質問を用いてベクトル空間法[21)]に基づいて関連文書を検索する．すなわち，検索質問と各文書をターム（索引語）の重みベクトルとして表現して，2 つのベクトルがなす角の余弦（cosine）によって検索質問と文書の類似度を計算し，類似度が大きい順に検索文書をソートする．タームの重みは TF・IDF を用いて計算する．TF（Term Frequency）は検索質問や各文書におけるタームの出現頻度であり，IDF（Inverse Document Frequency）はタームが出現する文書数に対する全文書数の比の対数である．

インデクシング（索引付け）は日本語/英語ともに単語単位で行う．英語文書に対しては，WordNet を用いて不要語の削除と屈折語の復元を行う．日本語文書は,「茶筌」を用いて形態素解析を行い内容語を抽出する．そこで，検索時には元の複合語における語基の語順は考慮しない．たとえば，検索質問「特徴次元リダクション」が「feature dimension reduction」に翻訳されると,「feature」「dimension」「reduction」の

☆ 本論文では他の翻訳方式との一般的な優劣は議論しない．

☆☆ 過去の複数の研究や我々の予備実験によれば,「茶筌」は 90%以上の精度で分かち書きや品詞付与を行う．

いずれかが任意の位置に現れる文書はすべて検索される．今後は，複合語に基づく高度なインデクシングが必要である．

4. 複合語翻訳法

我々の複合語翻訳法は，語基の語順を保持したまま☆，対訳辞書を用いて語基の翻訳候補を導出し，共起情報を用いて語基の訳語曖昧性を解消する．

以下，4.1 節と 4.2 節で語基辞書の作成と訳語曖昧性の解消法について説明する．これらは，1 章の課題 (1) と (2) の解決にそれぞれ対応している．

4.1 語基辞書の作成

我々の語基辞書は，EDR 日英専門用語対訳辞書[29]を用いて作成した．日本語には語基の区切りがないため，英語語基との対応付けが困難である．しかも，日本語分割の難しさは語基数の増加とともに顕著になる．そこで，2 語基から構成される英単語とその日本語対訳 59,533 対のみを抽出し，ヒューリスティックスによって日本語を 2 語基に分割して（語順を保持したまま）英語語基との対応付けを行った．語基数は英語見出し語によって容易に特定することができる．我々のヒューリスティックスを以下に示す．

- 字種の切れ目（複数ある場合は最左の位置）で分割する．
- 同一の字種だけで構成される場合は，中央（奇数文字列の場合は中央文字の左側）で分割する．
- 単語の先頭に現れない文字（「ァ」「ン」「ー」など）の直前では分割しない．このような場合は，分割位置を右にずらす．

その結果，日本語語基数 24,439，英語語基数 7,910 を含む語基辞書（日英/英日の両方）を作成した．**表 1** に日本語分割後の EDR 専門用語対訳辞書の一部を示す．**表 2** は，表 1 から作成した英日語基辞書である．

4.2 訳語曖昧性の解消

訳語曖昧性の解消には，品詞付け[5]や機械翻訳[3]で用いられている統計的手法を用いた．原言語の複合語 S と，その翻訳候補の 1 つ T を次のように定義する．

$$S = s_1, s_2, \ldots, s_n; \quad T = t_1, t_2, \ldots, t_n$$

ここで s_i と t_i は i 番目の語基を表す．原言語が日本語の場合は，語基辞書を引きながら可能な分割を調べ，語基数最小の分割（複数ある場合はすべて）を考慮する．語基辞書に未登録の語基がある場合は，先頭から分割できたところまでを部分的に翻訳する．日本語の専門用語にアルファベット列（英単語や略語）が含まれる場合，その部分はそのまま出力する．

訳語曖昧性の解消は，$P(T|S)$ を最大化する T を選択することであり，ベイズの定理によって式 (1) のように変形できる．複数の翻訳候補を許容する場合は，$P(T|S)$ の値の大きい T から順に選択する．

$$\arg\max_T P(T|S) = \arg\max_T P(S|T)\cdot P(T) \quad (1)$$

さらに，$P(S|T)$ と $P(T)$ を式 (2) で近似する．

$$\begin{aligned} P(S|T) &\approx \prod_{i=1}^{n} P(s_i|t_i) \\ P(T) &\approx P(t_1)\cdot\prod_{i=1}^{n-1} P(t_{i+1}|t_i) \end{aligned} \quad (2)$$

翻訳対象が複合語でない単純語の場合は，訳語曖昧性解消のための情報が得られないので，語基辞書に定義されている訳語をすべて検索に利用する．

次に，式 (2) の各項の推定方法について説明する．$P(s_i|t_i)$ は表 1 の日本語分割後の EDR 専門用語対訳辞書を用いて推定する．表 1 では，たとえば「相関」は「associative」と 1 回,「correlation」と 2 回対応しているので，式 (3) が成り立つ．

表 1 EDR 専門用語対訳辞書の例
Table 1 A fragment of the EDR technical terminology dictionary.

英語複合語	日本語複合語
CCD memory	CCD メモリー
IC memory	IC メモリ
associative learning	相関 学習
associative memory	連想 メモリ
associative record	結合 レコード
correlation function	相関 関数
factor correlation	因子 相関
hybrid IC	ハイブリッド 集積回路

表 2 英日語基辞書の例
Table 2 A fragment of an English-Japanese base word dictionary.

英語語基	日本語語基
CCD	CCD
IC	IC, 集積回路
associative	相関，連想，結合
correlation	相関
factor	因子
function	関数
hybrid	ハイブリッド
learning	学習
memory	メモリ，メモリー
record	レコード

☆ 既存の対訳辞書[8]に定義されている複合語の約 95%は原言語と目的言語で語基数と語基の語順が一致する．

表 3 NACSIS コレクション検索課題の例
Table 3 A fragment of the NACSIS collection query.

課題番号	0005
タイトル	特徴次元リダクション
検索要求	クラスタリングにおける特徴次元リダクション
検索要求説明	（省略）... 画像処理などの実験の操作の一部として特徴次元リダクションを用いているだけでは要求を満たさない.
概念	特徴選択，主成分分析，グラフ理論，情報の粒度，幾何クラスタリング
分野	電子・情報・制御

$$\begin{aligned} P(\text{associative} \mid \text{相関}) &= 1/3 \\ P(\text{correlation} \mid \text{相関}) &= 2/3 \end{aligned} \quad (3)$$

本来ならば，語基の翻訳確率 $P(s_i|t_i)$ は，単語単位の対応付けがなされた 2 言語コーパスを用いて推定することが好ましい．しかし，このような言語資源は現在非常に高価である．また，専門用語辞書ではその分野で使われやすい単語が繰り返し語基として使用される傾向があるため，本推定法は分野依存の統計頻度をある程度反映していると考える．

$P(t_{i+1}|t_i)$ は，文書コレクションから抽出した目的言語における語の共起情報を用いて推定する（図 1 参照）．インデクシングと同様に，英語文書に対しては WordNet を，日本語文書に対しては「茶筌」を用いて内容語を抽出する．$P(t_1)$ は，共起情報から単語単位の頻度を抽出して計算する．2 言語コーパスを利用する手法[4),31)]とは異なり，我々の翻訳法は目的言語コーパスの共起情報だけを利用するため，実装コストが低いという利点がある．

5. 評価実験と考察

5.1 NACSIS コレクションの概要と利用法

NACSIS コレクション[13)]は，65 学会の論文から収録した約 33 万件の抄録（文書），日本語の検索課題，各検索課題に対する正解文書リストからなる．文書には,「文書番号」「論文タイトル」「著者名」「出典」「抄録」「著者キーワード」「学会名」のフィールドがある．これらのうち,「文書番号」を除くフィールドは日本語か英語，あるいはその両方で書かれている．検索課題には NACSIS コレクション・ワークショップの予備試験用に配布された 21 件を用いた．検索課題には，**表 3** に示すようなフィールドがある．予備試験用の検索課題は主に「電子・情報・制御」の分野に関するものである．正解文書判定には「プーリング法」が用いられている．すなわち，複数の異なる検索システムの検索結果を集めて正解文書候補とする．次に，正解候補に対して，関連する (A)，部分的に関連する (B)，関連しない (C) という 3 つのランクに基づいてコレクション作成者が最終的な判定を行う．1 つの検索課題について，プーリング法による正解候補数は平均約 4,400，関連文書数と部分的な関連文書数はそれぞれ 144 と 13 である．

NACSIS ワークショップでは，検索要求フィールドのみを用いた検索が必須課題であり，我々の実験でもそれに従った．文書についても NACSIS ワークショップのルールに準拠し，抄録が日本語と英語の両方で書かれている文書（約 19 万件）を検索対象とした．そこで，日英 CLIR の検索性能を日本語単言語検索の性能と比較することができる．従来の実験でも，単言語検索の性能は CLIR の検索性能の上限として使われている．文書のインデクシングには「論文タイトル」「抄録」「著者キーワード」を用いた．また，A と B ランクの両方を正解文書と見なした（A ランクのみを正解と見なしても実験結果はほとんど同じだった）.

5.2 実　　験

評価実験は，NACSIS コレクション・ワークショップのルールに準拠して行った．すなわち，日本語の検索要求を用いて英語文書の検索と順位付けを行い，上位 1,000 文書に対して，再現率–適合率曲線と 11 点補間なし平均適合率を用いて検索性能を評価した．比較対象とした検索法を以下に示す.

(1) 日本語単言語検索（CLIR の検索性能の上限）
(2) EDR 専門用語対訳辞書に定義されている訳語をすべて検索に用いる．辞書に定義されていない複合語は，EDR 辞書のエントリを組み合わせて翻訳した.
(3) 語基辞書（4.1 節）を用いて，全訳語候補を検索に用いる.
(4) 語基辞書の訳語候補からランダムに N 個の訳語を選択して検索に用いる.
(5) 本論文で提案した翻訳法（4 章）を用いて，上位 N 個の訳語を検索に用いる.

ここで，N は複合語単位の訳語数である．N の値を変えて予備実験を繰り返した結果，$N = 3$ のときに手法 (5) の性能が最も良かった．そこで，手法 (4) と (5) では $N = 3$ としている．検索課題 21 件につい

表 4 11 点補間なし平均適合率の比較
Table 4 Comparison of 11-pt non-interpolated average precisions.

手法	ターム数	平均適合率	手法 (1) に対する比
(1)	83	0.204	—
(2)	72	0.130	0.637
(3)	424	0.171	0.838
(4)	191	0.116	0.569
(5)	164	0.193	0.946

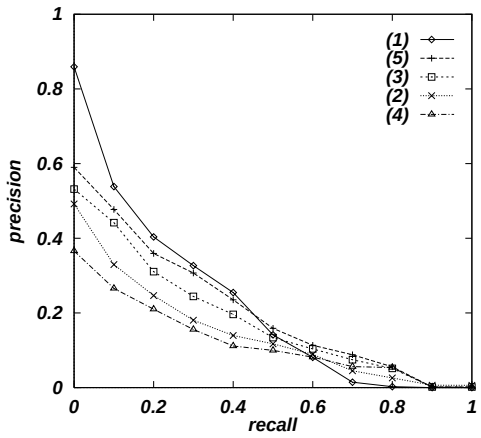


図 2 再現率–適合率曲線の比較
Fig. 2 Comparison of recall-precision curves.

て，検索に使われたターム（本研究では単語に相当する）の総数を各手法ごとに**表 4** に示す．手法 (2) は訳語曖昧性がほとんどなく，しかも翻訳できない語基があったので，手法 (1) のターム数よりも少ない．

各手法の再現率–適合率曲線と平均適合率を**図 2** と表 4 に示す．この結果から得られる結論を以下にまとめる．まず，手法 (2) と (3) を比較すると，手法 (3) の方が一般に高い検索性能を示している．両者の違いは語基辞書の差にあるので，ここから我々の語基辞書の有効性が分かり，1 章の課題 (1) を解決できた．次に，手法 (3)〜(5) を比較すると，手法 (3) が最も高い検索性能を示している．これは適切な訳語選択によるものであり，ここから統計的な訳語曖昧性解消の有効性が確認され，1 章の課題 (2) を解決できた．最後に，手法 (1) と (5) を比較すると，我々の CLIR システムの平均適合率は単言語検索の約 95%である．従来の実験では，CLIR の検索性能は単言語検索のおよそ 50〜75%という報告がある[22]．実験環境が異なるので直接的な比較はできないものの，本実験結果は良好である．

ただし，手法 (1) と (5) の違いは日本語と英語の文書検索の違いが影響している可能性がある．NACSIS コレクションの検索要求を英訳すれば，英語単言語検索との比較も可能である．しかし，検索要求の訳質が実験結果に影響する可能性がある（2.3 節）．また，NACSIS コレクション・ワークショップでも日本語単言語検索を日英 CLIR の上限としており，今回の実験設定は妥当なものである考える．

5.3 関連研究との比較

さらなる評価のために，本実験結果を NACSIS コレクションを用いた他の実験結果と比較する．Kando ら[12]は，我々とほぼ同じ環境で CLIR の評価実験を行っている．彼らの CLIR は，NACSIS コレクションの日英著者キーワード対を用いて対訳辞書を作成し，検索質問を翻訳する．実験環境における差異は，文書検索モジュールの違いによって彼らの日本語単言語検索の検索性能が我々の結果よりも良い点である．しかし，彼らの CLIR の平均適合率は，(彼らの) 単言語検索の約 50〜60%であり，その絶対値は我々の CLIR とほぼ同じである☆．以上より，我々の検索質問翻訳法は文書検索モジュールの性能の差異を補完していることが分かる．

5.4 今後の研究課題

検索要求の翻訳精度を下げている主な原因として，語基辞書の不備がある．この問題の解決法として，語基辞書作成における複合語分割法（4.1 節）の洗練がある．Tsuji ら[25]は HMM を用いて複合語を語基に分割する手法を提案している．彼らの手法は，2 語基以上の複合語を平均 80〜90%程度の精度で分割でき，我々の研究にも役立つものと期待できる．また，2 言語コーパスから対訳を抽出する手法[4),11),24)]によって辞書を拡張する対処法がある．

他方において，複合語分割法の洗練や対訳辞書の拡張だけでは対処が困難な問題もある．

専門用語は外来語のカタカナ表記が多く，単語の移入にともなって漸進的に語基が作られる．このような新カタカナ語は，音韻情報に基づく翻字[15),28),30)]を行う必要がある．今回の実験に用いた検索要求には，辞書未登録のカタカナ語基として「コロケーション（collocation）」と「マイニング（mining）」があった☆☆．これらの単語は EDR 専門用語対訳辞書に単純語としても複合語としても定義されておらず，複合語分割法を洗練したとしても翻訳できない．なお,「コロ

☆ Kando と Aizawa は，再現率–適合率曲線と単言語検索との相対的な平均適合率しか示していないため，平均適合率の絶対値を厳密に比較することはできない．ここでの議論は，再現率–適合率曲線の視覚的な比較に基づいている．

☆☆「コロケーション」は単純語として,「マイニング」は「データマイニング（data mining）」という複合語として現れた．

ケーション」と「マイニング」に対して人手で正しい訳語を与えたところ，5.2 節の手法 (5) の平均適合率が 0.212 に向上し，日本語単言語検索を若干上回った．

また,「新聞記事」が翻訳できず，一般語の対訳辞書も必要であることが分かった．ただし，一般/専門辞書の組合せ方について十分に検討する必要がある[19]．

英語の略語が日本語として使われる問題もある．実験に用いた日本語検索要求には「LFG（lexical functional grammar）」が存在した.「LFG」のままでも英語文書は検索できるものの，原形も用いることで再現率を向上できる可能性がある．括弧表現などを手掛かりにして，略語と原形を検索文書からあらかじめ抽出するなどの対処法[28]がある☆．

6. おわりに

本論文は，技術文書のための日本語と英語の言語横断検索システムを提案した．本システムの特長は，検索質問中の専門用語や検索文書中のキーワードを語基の対訳辞書と語の共起情報を用いて翻訳する点にある．NACSIS コレクションを用いた検索実験の結果，本システムの検索性能は単言語検索とほぼ同じであることが確認された．考察を通して特定した研究課題への対処については，別稿にて報告する予定である．

謝辞 NACSIS コレクションは，学術情報センターの許諾を得て使用させていただきました．この場を借りて深謝いたします．

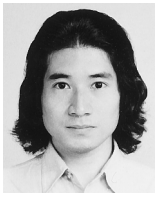
参考文献

1) ACM SIGIR: *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1996–1998).
2) Ballesteros, L. and Croft, W.B.: Resolving Ambiguity for Cross-language Retrieval, *Proc. 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.64–71 (1998).
3) Brown, P.F., Pietra, S.A.D., Pietra, V.J.D. and Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol.19, No.2, pp.263–311 (1993).
4) Carbonell, J.G., Yang, Y., Frederking, R.E., Brown, R.D., Geng, Y. and Lee, D.: Translingual Information Retrieval: A Comparative Evaluation, *Proc.15th International Joint Conference on Artificial Intelligence*, pp.708–714 (1997).
5) Church, K.W. and Mercer, R.L.: Introduction to the Special Issue on Computational Linguistics Using Large Corpora, *Computational Linguistics*, Vol.19, No.1, pp.1–24 (1993).
6) Davis, M.W. and Ogden, W.C.: QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System, *Proc. 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.92–98 (1997).
7) Dumais, S.T., Landauer, T.K. and Littman, M.L.: Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing, *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval* (1996).
8) Ferber, G.: *English-Japanese, Japanese-English Dictionary of Computer and Data-Processing Terms*, MIT Press (1989).
9) Gilarranz, J., Gonzalo, J. and Verdejo, F.: An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database, *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval* (1997).
10) Hull, D.A. and Grefenstette, G.: Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval, *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.49–57 (1996).
11) Kaji, H. and Aizono, T.: Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information, *Proc. 16th International Conference on Computational Linguistics*, pp.23–28 (1996).
12) Kando, N. and Aizawa, A.: Cross-Lingual Information Retrieval using Automatically Generated Multilingual Keyword Clusters, *Proc. 3rd International Workshop on Information Retrieval with Asian Languages*, pp.86–94 (1998).
13) Kando, N., Kuriyama, K. and Nozue, T.: NACSIS Test Collection Workshop (NTCIR-1), *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.299–300 (1999).
14) Kawate, F. and Ishikawa, T.: A Mutual Retrieval System for Japan/China-MARC using NDC and CLC, *Proc. 2nd International Conference on Terminology, Standardization and Technology Transfer*, pp.516–523 (1997).
15) Knight, K. and Graehl, J.: Machine Translit-

☆ 今回の実験では，一般語と略語の問題に対して人手で正しい訳語を与えても，我々の CLIR の結果はほとんど向上しなかった．

eration, *Computational Linguistics*, Vol.24, No.4, pp.599–612 (1998).
16) Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. and Tengi, R.: Five papers on WordNet, Technical Report CLS-Rep-43, Cognitive Science Laboratory, Princeton University (1993).
17) Mongar, P.: International Co-operation in Abstracting Services for Road Engineering, *The Information Scientist*, Vol.3, pp.51–62 (1969).
18) Oard, D.W.: A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval, *Proc. 3rd Conference of the Association for Machine Translation in the Americas*, pp.472–483 (1998).
19) Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-language Information Retrieval, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.55–63 (1998).
20) Salton, G.: Automatic Processing of Foreign Language Documents, *Journal of the American Society for Information Science*, Vol.21, No.3, pp.187–194 (1970).
21) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
22) Schäuble, P. and Sheridan, P.: Cross-Language Information Retrieval (CLIR) Track Overview, *The 6th Text Retrieval Conference* (1997).
23) Sheridan, P. and Ballerini, J.P.: Experiments in Multilingual Information Retrieval using the SPIDER system, *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.58–65 (1996).
24) 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).
25) Tsuji, K. and Kageura, K.: An HMM-based Method for Segmenting Japanese Terms and Keywords based on Domain-Specific Bilingual Corpora, *Proc. 4th Natural Language Processing Pacific Rim Symposium*, pp.557–560 (1997).
26) 鈴木雅実，井ノ上直己，橋本和夫：翻訳情報の提示によるクロスリンガル情報検索結果からの文書選択，言語処理学会第 5 回年次大会発表論文集，pp.371–374 (1999).
27) 酒井哲也，梶浦正浩，住田一男：Cross-language 情報検索のための BMIR-J2 を用いた一考察，情報処理学会自然言語処理研究会報告，Vol.99, No.2, pp.41–48 (1999).
28) 藤井　敦，石川徹也：言語横断検索システム Quest，言語処理学会第 5 回年次大会発表論文集，pp.353–356 (1999).
29) 日本電子化辞書研究所：専門用語辞書（情報処理）(1995).
30) 熊野　明：カタカナ表記からの英訳推定による専門用語辞書作成，言語処理学会第 1 回年次大会発表論文集，pp.221–224 (1995).
31) 奥村明俊，石川　開，佐藤研治：コンパラブルコーパスと対訳辞書による日英クロス言語検索，自然言語処理，Vol.5, No.4, pp.77–93 (1998).
32) 松本裕治，北内　啓，山下達雄，今一　修，今村友明：日本語形態素解析システム『茶筌』version 1.5 使用説明書，技術報告 NAIST-IS-TR97007，奈良先端科学技術大学院大学 (1997).

(平成 11 年 4 月 15 日受付)
(平成 12 年 2 月 4 日採録)



藤井　敦（正会員）

1993 年 3 月東京工業大学工学部情報工学科卒業．1998 年 3 月同大学大学院博士課程修了．1998 年図書館情報大学助手，現在に至る．博士（工学）．自然言語処理，情報検索の研究に従事．人工知能学会，言語処理学会，Association for Computational Linguistics 各会員．

石川　徹也（正会員）

1977 年 3 月慶応義塾大学大学院修士課程（図書館情報学専攻）修了．富士写真フイルム（株）足柄研究所入社，図書館短期大学等を経て現在，図書館情報大学教授．工学博士．情報管理システムの高度化の研究に従事．人工知能学会，言語処理学会，ACM 等会員．