日英言語横断情報検索のための翻訳知識の獲得

阿玉泰宗† 橋本泰一†
徳永健伸† 田中穂積†

言語横断情報検索 (CLIR) は，クエリと異なる言語で記述された文書集合に対しておこなう情報検索である．このタスクでは，文書が記述されている言語にクエリを翻訳する手法が広く用いられているが，この手法には辞書の語彙不足や翻訳曖昧性の問題が生じる．本論文は，日英言語横断情報検索において，これらの問題を解決し，検索性能を向上させる手法を提案する．語彙不足を解決するために，藤井らの提案した語基の学習手法を改良し，対訳との共起を用いる手法を提案する．また，複数の形態素解析結果で場合分けして翻字をおこない，結果を Bigram でリランキングする手法を提案する．実験により，これらの手法はいずれも従来手法より高い精度を示すことを確認した．また，翻訳曖昧性を解消するために局所的な共起と大域的な共起の情報を組み合わせる手法を提案する．これらの提案手法をシステムに組み，NTCIR コレクションを用いた評価実験をおこなった．その結果，提案手法を組み込むことにより，従来よりも検索性能が向上することを確認した．

Acquisition of Translation Knowledge for Japanese-English Cross-lingual Information Retrieval

Yasumune Adama,† Taiichi Hashimoto,† Takenobu Tokunaga† and Hozumi Tanaka†

In cross lingual information retrieval (CLIR), languages used to describe queries and documents are different. To bridge this language gap, translating queries into a language describing documents is widely adopted. This approach requires wide coverage translation dictionaries, but it is generally difficult to build such kind of dictionaries. In addition, as in ordinary machine translation systems, translation ambiguities should be resolved when translating queries. To improve the coverage of the dictionary, this paper proposes a method to build a "base word dictionary" from bilingual dictionaries and a document collection. In addition, an improved transliteration method is introduced to deal with *Katakana* words not being in the dictionary. In order to solve translation ambiguity, the paper proposes a disambiguation method by using both local and global concurrence information. The proposed methods were integrated into an experimental Japanese-English CLIR system, and the system was evaluated by using the NTCIR test collection to show the effectiveness of the proposed methods.

1. はじめに

近年，WWW に代表される電子文書は急速に増大しているが，ユーザの要求に適合する文書が母国語で記述されているとは限らない．この言語ギャップを解消するひとつの手段として，母国語のクエリによって他言語の文書を検索する言語横断情報検索 (CLIR) への関心が高まっている．対訳辞書を利用してクエリを翻訳する手法は CLIR において多く用いられているが，この手法には辞書の語彙不足や翻訳曖昧性の問題が伴う．

語彙不足の問題を解決するため，藤井らは，専門語辞書の句の対訳関係からその要素語の対訳関係を学習して「語基辞書」を作成する手法を提案している[3]．また，未知語の多くを占めるカタカナ語の翻訳手法として，発音情報を利用する翻字があり，Brill らは，文脈を考慮した翻字のモデルを提案している[4]．一方，翻訳曖昧性を解消するための手法として，前田らの相互情報量に基づく大域的な共起情報を用いる手法[6] や，藤井らの Bigram に基づく局所的な共起情報を用いる手法[3] などがこれまでに提案されている．

しかし，これらの手法にはいくつかの問題点が考えられる．語基辞書の作成は，見出し語のみに注目しており，対訳の情報は用いていない．また，日本語と英

† 東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

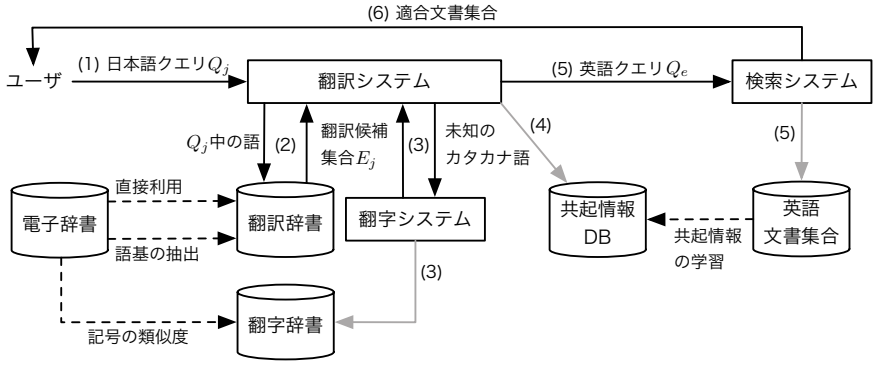

図 1　CLIR システムの概要
Fig. 1　Overview of a CLIR System

語などの言語間での翻字には，入力の形態素数と出力の単語数の不一致の問題がある．翻訳曖昧性の解消に関しては，大域的な共起を用いると計算量が多くなり，局所的な共起を用いると使用できる文脈が制限されるという問題がある．

本論文では，日英の CLIR を対象とし，語彙不足を解決するために英単語との共起情報を用いた語基辞書の作成手法，形態素数で場合分けした翻字のモデルを提案する．また，翻訳曖昧性の解消のために Bigram と相互情報量を併用する手法を提案し，NTCIR[10] コレクションを用いて精度の評価をおこなう．

本論文で提案するシステムの構成を，図 **1** に示す．図中の黒い実線はデータの流れを，灰色の実線はデータの参照を，破線は前処理の段階でデータが生成されていることを表わす．システムの処理の概要は以下のとおりである．各項目の番号は図中の番号と対応している．

(1)　ユーザがシステムに日本語クエリ Q_j を与える．
(2)　翻訳システムが Q_j 中の語を語単位で辞書引きし，翻訳語の候補語集合 E_i を得る．
(3)　辞書引きに失敗したカタカナ語を翻字する．
(4)　任意の候補語の組み合わせについて，共起情報から語の関連性を計算し，翻訳曖昧性を解消する．
(5)　翻訳曖昧性解消によって得られた英語クエリ Q_e を用いて検索をおこなう．
(6)　検索結果をユーザに提示する．

以下，2，3 節では，本論文で提案する語基辞書の構築手法，翻字手法についてそれぞれ説明した後，これらの手法単独の評価実験についても述べる．4 節では翻訳曖昧性の解消手法について述べるとともに，提案手法を組み込んで実現したシステムを NTCIR テストコレクションを用いて評価した結果について述べる．

2. 語基辞書の構築

2.1 先 行 研 究

藤井らは，EDR 専門語辞書の対訳情報のうち，対訳が英語 2 単語である見出し語に着目し，見出し語を対訳の各単語に対応する部分に分割することで，より細かい対訳関係を学習する手法を提案している[3]．この手法では，辞書で直接定義されているよりも細かい単位 (語基) の対訳関係が得られるため，語基を組み合わせることによって，多くの語が翻訳可能になるという特徴がある．彼らは，この対訳関係を収集したものを「語基辞書」と呼んでおり，見出し語の文字種の異なりを利用して，下の 2 種類の辞書の作成手法を提案している[2]．

(藤井 1)：すべての文字種の異なりを利用する．
(藤井 2)：十分に信頼できる文字種の異なりから，基本の対応を得る．

(藤井 1) の手法は，見出し語で最も先頭に近い文字種の異なりで見出し語を分割する．見出し語が単一の文字種からなる場合は，語の中央で分割をおこなう．

(藤井 2) の手法は，分割位置として十分に信頼度の高い文字種の異なりとして以下のパターンにマッチする見出し語のみを分割して “seed” を学習する．

- C+K+ → C+/K+
- C+A+ → C+/A+
- A+K+ → A+/K+
- K+A+ → K+/A+
- CCCC → CC/CC

なお，A,C,K はそれぞれアルファベット，漢字，カタカナを示し，“X+” は X が一つ以上連続して出現することを示す．また，パターンにマッチしない見出し

語については，得られる語基の種類を最小化するように分割をおこなう．

これらの手法では，対訳の英単語の情報を語基数の特定にのみ用いているが，本論文では，語基数の情報に加えて日英の対訳辞書から抽出した対訳の共起情報を利用して語基辞書を作成する．

2.2 提案手法

2.2.1 共起確率の学習

対訳の共起確率は，EDR 専門語和英辞書[8] の見出しのうち，対訳が英語 2 語以上からなるエントリから学習する．

まず，隠れマルコフモデル (HMM) を用いて日本語の見出しを形態素解析する[9]．すなわち，見出し語が形態素 $s_0, s_1, \ldots, s_n (n \geq 1)$ に分割でき，その品詞が $p_0, p_1, \ldots, p_n$ であるとき，

$$\arg \max_{s_0, s_1, \ldots, s_n} \prod_{i=0}^{n} P(s_i|p_i) \cdot P(p_i|p_{i-1}) \qquad (1)$$

を最大化する形態素列 $s_0, s_1, \ldots, s_n$ を見出し語の形態素列と見なす．ここで，品詞の連接確率 $P(p_i|p_{i-1})$ と単語の生起確率 $P(s_i|p_i)$ は，NTCIR-2[10] の日本語コーパスを形態素解析プログラム茶筅[11] で解析した結果から学習した．ただし，対訳が 2 語以上のエントリの見出しを解析しているので，形態素数が 1 となる解析結果は棄却する．以下，この形態素解析の手法を「HMM による形態素解析」と呼ぶが，形態素数 1 の解析結果を棄却している点では一般の HMM による形態素解析とは異なる点に注意されたい．

茶筅のような一般的な形態素解析プログラムで見出し語を直接解析しない理由は，これらのプログラムがより長い文の単位の形態素解析を想定して設計されており，見出し語のような短い単位の解析では，その性能を十分発揮できないことと，専門語が複合語の場合でもそれが辞書に含まれていれば 1 形態素として解析されてしまうためである．専門語が 1 形態素として解析されてしまうと，語基を取り出すという本論文の目的には適さない．上述の方法で形態素解析をおこなうと，形態素数 1 の結果は捨てるので，必ず複数形態素からなる解析結果が得られる．

なお，形態素解析が失敗した場合は，見出し語の各文字を 1 形態素と見なして分割をおこなう．

次に，得られた形態素列に対してすべての可能な連接操作をおこない，見出し語の部分文字列の集合を生成する．各文字を 1 形態素とした場合は，見出し語のすべての部分文字列を生成することになる．

ここで，日本語の見出し語 J に n 個の対訳 $E_1, E_2, \ldots, E_n$ が与えられており，E_i が $m(m > 2)$ 語からなるとき，見出し語 J から生成した見出し語の部分文字列集合 $\{j_1, j_2, \ldots, j_k\}$ 中の要素と E_i 中の単語集合 $\{e_{i1}, e_{i2}, \ldots, e_{im}\}$ 中の語の共起頻度として $1/(k \cdot m \cdot n)$ を与える．

たとえば，見出し語「16 進数」が “hexadecimal numeral” と “hexadecimal number” の 2 つの対訳を持つ場合 $(n = 2)$，いずれの対訳も 2 単語からなる $(m = 2)$．また，この見出し語が形態素解析によって「16」「進」,「数」の 3 つに分割されたとすると部分文字列集合は {16, 進, 数, 16 進, 進数, 16 進数 } $(k = 6)$ となる．したがって，J の部分文字列集合中の j_i に対して $freq(j_i, \text{“number”}) = 1/24$ の共起頻度を，また対訳に 2 回出現する *“hexadecimal”* との共起頻度として $freq(j_i, \text{“hexadecimal”}) = 2/24$ を与える．

すべての見出し語から共起頻度を計算した後，すべての英単語 e と，それと共起する見出し語の部分文字列 j について，

$$P(j|e) = \frac{freq(j, e)}{\sum_{j_i} freq(j_i, e)} \qquad (2)$$

を計算する．

表 **1** 形態素解析の精度
Table 1 Precision of morphological analysis

手法	正解	不正解	精度 (%)
茶筅	1,679	162 (147/15)	91.2
HMM	1,824	17 (10/7)	99.1

2.3 の評価実験で用いる 1,841 見出し語を茶筅と HMM で解析した精度を**表 1** に示す．正解は人手によって作成し，正しい区切り位置が形態素の境界となっているときに正解と判定した．不正解の列の括弧中の数値は，解析結果の形態素数の内訳である (1 形態素/複数形態素)．表 1 からわかるとおり，茶筅では 1 形態素として解析されてしまう失敗例が多い．一方，HMM を用いた手法では原理的に見出し語を複数形態素に分割するので，HMM における 1 形態素による不正解数は，解析に失敗した数を表わす．

2.2.2 語基の学習

語基の学習には，EDR 専門語和英辞書[8] 中の，対訳が 2 つの英単語からなる見出し語 57,023 語を用いる．まず，EDR 英和辞書を利用して対応付けをおこなう．対訳の英単語をそれぞれ辞書引きし，得られた日本語訳を組み合わせて，元の和英辞書の日本語の見出し語が得られるならば，辞書から得られた対応は正しいと考える．このような単純な辞書引きによって 57,023 見出し語中，約 52%に相当する 29,490 見出し語の対

応付けが可能であった．以下の処理では対応付けできなかった残りの 27,713 の見出し語について，共起を用いた対応付けをおこなう．

まず，2.2.1 で述べた共起確率の学習と同様に HMM を用いて見出し語の形態素解析をおこない，付属語や接尾辞の直前を除く任意の形態素の境界を分割の候補位置とする．形態素解析に失敗した (分割ができなかった) 場合は，任意の文字の間を候補位置とする．ただし，ここでは，2 分割の場合しか考えない点が共起確率の学習とは異なる．

こうして得られる見出し語の 2 分割の組を (j_0, j_1) とする．一方，見出し語の対訳が e_0, e_1 の 2 語から成るとする．見出し語を j_0, j_1 に分割するスコア sc を式 (3) のように定義し，(j_0, j_1) の分割および (j_x, e_y) の対応について sc を最大化する組み合わせをそれぞれの語基の対訳関係として抽出する．ここで $P(j|e)$ は式 (2) によって計算する．

$$sc(j_0, j_1, e_0, e_1) = \max(P(j_0|e_0)P(j_1|e_1), P(j_1|e_0)P(j_0|e_1)) \quad (3)$$

2.3 評価実験

EDR 英和辞書の辞書引きによって対応付けできなかった 27,712 語の見出し語からランダムに抽出した 1,841 見出し語を語基に分割し，対訳関係を抽出した結果を表 **2** に示す．この 1,814 例について，正解は人手によって作成した．(藤井 1)，(藤井 2) は 2.1 で説明した藤井らの手法である．

表 **2** 語基の対訳関係の抽出精度
Table 2 Precision of extracting base word translations

手法	正解数	精度 (%)
(藤井 1)	1,320	71.7
(藤井 2)	1,647	89.5
提案手法	1,775	96.4

表 **3** 語基辞書の比較
Table 3 Comparison of extracted base-word dictionaries

	既知	未知		合計
		見出し有	見出し無	
(藤井 1)	6,121	9,980	19,624	35,725
	(17.1%)	(27.9%)	(55.0%)	
(藤井 2)	6,156	9,729	9,567	25,452
	(24.2%)	(38.2%)	(37.6%)	
提案手法	6,534	8,802	8,200	23,536
	(27.8%)	(37.4%)	(34.9%)	

学習対象とした EDR 専門和英辞書の 57,023 見出しから得られた語基辞書の統計情報を表 **3** に示す．「既知」，「未知」はそれぞれ得られた語基の対訳関係が学習に使用した対訳辞書にすでに含まれている場合，含まれていない場合をそれぞれ表わす．対訳関係が辞書に存在しなかったもののうち，「見出し有」は日本語の語基が見出し語として存在したが，学習できた対訳は含まれていない場合，「見出し無」はそもそも日本語の語基が見出し語に含まれていない場合を表わす．

すべての語基の対訳について正解を作成しているわけではないので，表 3 からは間接的な推定しかできないが，辞書に含まれる対訳関係の比率が大きく，辞書に含まれない見出し語 (不適切な語を多く含むと考えられる) が少なくなっている事実から，表 2 の結果と合わせて考えると，提案手法によって多くの分割が正しい位置でおこなわれ，適切な対訳関係が抽出できていると推定できる．

3. 翻　　字

カタカナ語は専門分野で多く用いられ，種類が豊富である上に新語が生成されやすい．また，表記の揺れもあり辞書式に列挙するのは困難である．こうした特徴を持つカタカナ語を翻訳する手法として，Knight らはカタカナ語が表音文字であり，外来語の表記に用いられる点に注目して「カタカナ語を発音的に等価，または類似した英単語列に翻訳する」という，翻字の手法を提案している[5]．しかし，Knight らの手法は音韻辞書に登録された語しか出力できないという制限がある．藤井らは EDR 辞書から抽出したカタカナ語から人手によって作成した経験的知識を用いて翻字辞書を構築し，この辞書を使って翻字をおこなう手法を提案している[3]．一方，Brill らはスペル訂正のモデルを用い，文脈を用いた Grapheme (表記記号) レベルの翻字手法を提案している[4]．以上のような背景から本節では，人手の介入を必要としない Brill の手法を基礎として，これに，形態素分割と英語 Bigram によるリランキングを導入して拡張した翻字手法を提案する．

3.1 Brill の翻字モデル

Brill らは，辞書に含まれない文字列 s から辞書に含まれる単語 w へのスペル訂正を，式 (4) で定義している[1]．

$$\arg\max_w P(w|s) = \arg\max_w P(s|w) \cdot P(w)$$
$$P(s|w) = \sum_{R \in Part(w)} P(R|w) \sum_{\substack{T \in Part(s) \\ |T|=|R|}} \prod_{i=1}^{|R|} P(T_i|R_i) \quad (4)$$

ここで，R_i は辞書中に含まれる正しい単語 w を分割するパタン，T_i は辞書に含まれない文字列 (ミススペ

ル) s を分割するパタン，$Part(w)$ は文字列 w のあらゆる可能な分割の集合である．Brill らは，式 (4) を，最適な部分列への分割の場合のみを考慮した式 (5) で近似している．

$$P(s|w) = \max_{\substack{R \in Part(w) \\ T \in Part(s)}} P(R|w) \prod_{i=1}^{|R|} P(T_i|R_i) \quad (5)$$

$P(T_i|R_i)$ は，人手で作成した綴り間違いと正解の綴りの対の集合から学習する．また，単語 w を部分列 $R = R_1, R_2, \ldots, R_n$ に分割する確率 $P(R|w)$ は推定できないため，この項を無視した式 (6) を用いている．

$$P(s|w) = \max_{\substack{R \in Part(w) \\ T \in Part(s)}} \prod_{i=1}^{|R|} P(T_i|R_i) \quad (6)$$

この式を用いると，求める確率 $P(w|s)$ は式 (7) で得られる．

$$P(w|s) = P(w) \cdot \max_{\substack{R \in Part(w) \\ T \in Part(s)}} \prod_{i=1}^{|R|} P(T_i|R_i) \quad (7)$$

Brill らは，式 (7) を用いた実験をおこない，その有効性を確認している[1]．また，部分列の単語中での位置情報の利用によって精度が向上することも報告している．

この他に，Web 検索エンジンのログからカタカナ語と英語の対を収集する実験において式 (7) が翻字に有効であるという報告もある[4]．

それまでのモデルが 1 文字ごとの遷移確率を用いて確率 $P(s|w)$ を計算するのに対し，Brill のモデルは注目する文字の前後の文字を文脈として利用することで，よりもっともらしい確率計算を可能にしている．しかし，このモデルの問題として以下の 2 つが考えられる．

- 確率 $P(T_i|R_i)$ の学習に重み無しの編集距離を用いているが，翻字はスペル訂正に比べ置換すべき文字列が多いので，重み無しの編集距離では，同一コストの対応づけが多数得られることが予想される．
- 翻字では一形態素が一英単語を出力するという関係が常には成り立たないので，位置情報が利用できない場合がある．たとえば，「アイパターン」が “eye pattern” に翻字されるとき，カタカナ語の 1 形態素が 2 語からなる対訳に対応するので，「パ」が “pattern” の先頭に対応するといった，単語中の位置情報が利用できない．

本節では，この 2 つの問題に注目し，それを改善する手法を提案する．

3.2 提 案 手 法

3.2.1 編集距離の重みの導入

まず，モデルに編集距離の重みを導入する．以下の手順で反復学習をおこない，編集距離の重み考慮した文字間の対応付けを学習する．学習に使用したデータは EDR 和英辞書，EDR 専門語和英辞書に含まれるカタカナ語のうち，1 英単語からなる対訳をただひとつ持つカタカナ語 14,022 見出しである．学習時には，カタカナをローマ字に変換した．

(1) カタカナ語のローマ字表記の 1 文字 r と英単語の 1 文字 e の間の編集操作 $r \to e (r \neq e)$ の重みを 1，$r \to e (r = e)$ の重みを 0 に初期化．
(2) 対応付けの集合 S を空集合で初期化．
(3) すべての見出し語と訳語の対で，編集距離を最小化する操作列を求める．(編集距離)/(操作列長) が閾値以下☆の場合，この対応付けを S に加える．
(4) 対応付けの集合 S が直前のサイクルと一致した場合，終了．
(5) 任意の r, e について，S 中での頻度から $P(e|r)$ を学習し，操作 $r \to e$ の重みを $1 - P(e|r)$ に更新し，(2) に戻る．

例として，「ステンレス (*“stainless”*)」と「クライシス (*“crisis”*)」の 1 回目のサイクル，最後のサイクルのステップ (3) で得られる対応付けを図 **2** に示す．ここで下線 (_) は空文字を表わす．

	初期状態	終了状態
カタカナ語	s u t e n r e s u	s u t _ e n r e s u
対訳語	s t a i n l e s s	s _ t a i n l e s s
カタカナ語	k u r a i s i s u	k u r a i s i s u
対訳語	_ c r _ i s i s _	c _ r _ i s i s _

図 **2** 反復による文字対応付けの変化
Fig. 2 Result of iterative learning of chracter mapping

「ステンレス」では，反復により尤もらしい対応が得られているが，この対応付けは重み無し距離では最小にならない．また，「クライシス」の例では，1 回目に得られた対応は，平均コストが閾値を上回るため，学習に用いる集合 S には加えられない．しかし，k→c や u→ ε などの操作の重みが低下し，最終的には閾値を下回り，もっともらしい対応を学習できた．

次に，このようにして得られた対応付けを基礎として文脈を考慮したモデルを学習する．文脈としては，前後それぞれ 2 文字までを文脈とみなし，隣接する対応付けを結合し，頻度を計算した．たとえば，「ステン

☆ 以降の実験では経験値として 0.6 を閾値として用いた．

レス」の “r→l” の対応付けで文脈を考慮すると，以下の 9 種類の文脈付きの対応が得られる．

r → l
nr → nl
nr → inl
re → le
nre → nle
nre → inle
res → les
nres → nles
nres → inles

すべての対応付けからこのような文脈付きの対応付けを生成し，各対応付けの頻度から文字列単位の遷移確率 $P(r'|e')$ を学習する．ここで r', e' は文字ではなく，文脈付きの文字 (文字列) である．

さらに，確率 $P(r'|e')$ が 0.01 未満の規則を雑音として除去した結果，9,440 見出しから 66,815 種類の規則を得た．

3.2.2 形態素数による場合分け

翻字において，対訳関係にあるカタカナとその対訳の語 (形態素) 数が必ずしも一致しない問題に対処するために，あらかじめカタカナ語を形態素分割する処理を導入する．Brill の手法がカタカナ語 1 に対して複数の英単語を対応させながら翻字するのに対し，提案手法では，まずカタカナ語を形態素分割し，各形態素と英単語を 1 対 1 に対応付けて翻字する．これによって特に長いカタカナ語における性能の改善が期待できる．

翻字の対象となるカタカナ語 K は，語基辞書の学習と同様に HMM を用いて $s = 1 \sim 5$ 形態素に分割し，形態素の列 K_s を得た後，形態素数によって場合分けして適切な候補を選択する．また，個々の形態素から，式 (7) を大きくする上位 10 件の候補語を出力した．それぞれの語の生成を独立と考えると，カタカナ形態素列 $K = k_1, k_2, \ldots, k_s$ が英単語列 $E = e_1, e_2, \ldots, e_s$ に翻字される確率は，式 (8) で計算できる．

$$P(E|K) = \prod_{i=1}^{s} P(k_i|e_i) \cdot P(e_i) \qquad (8)$$

しかし，連続する単語の生成確率は独立とは考えにくい．そこで，得られる単語の組について，式 (9) を用いてリランキングをおこなった．

$$P(E|K) = \prod_{i=1}^{s} P(k_i|e_i) \cdot P(e_i|e_{i-1}) \qquad (9)$$

ただし，$P(e_1|e_0) = P(e_1)$ である．

3.3 評価実験

以下の 3 種のテストセットを用いて，翻字の精度比較をおこなった．

- **NTCIR2**：NTCIR-2 検索課題に出現するカタカナ語 185 語
- **EDR/1**：EDR の和英辞書および専門語和英辞書から得た，1 単語の対訳を複数持つカタカナ語 192 語
- **EDR/m**：EDR の和英辞書および専門語和英辞書から得た，複数単語の対訳を持つカタカナ語 214 語

候補語としては，NTCIR-1，NTCIR-2 の英語文書に出現する語を用いた．この際，TreeTagger[7] を用いて文書を単語に区切り，式 (4) の単語生成確率 $P(w)$ (=式 (8) の $P(e_i)$)，および式 (9) の $P(e_i|e_{i-1})$ を学習した．ただし，以下の条件に合致するものは綴り間違いとして除去した．

- 頻度 1 で，未知語のタグを与えられたもの
  (例) “aaplication” (“*application*” の間違い)
- 頻度 10 未満で，単一の編集操作によって十分に (自身の 100 倍以上) 頻度の高い語を得るもの
  (例) “anaysis” (“*analysis*” の間違い)
- 頻度 10 未満で，複数の既知の単語に分割可能であり，Bigram で計算した単語列の生成頻度が自身より十分大きいもの
  (例) “inarchitectural” (“*in architectural*” の間違い)

これにより，文書に出現した 494,478 単語のうち，147,951 単語を候補語として得た．なお，評価データはいずれも正解の単語がすべてこの集合に含まれる語である．

まず，編集距離の重みの効果を調べるために，あらかじめ形態素分割をおこなわない従来のモデルを用い，重み無しの編集距離で学習した遷移確率 (重み無し) と，反復学習により学習した遷移確率 (重み付き) のそれぞれを用いて精度の比較をおこなった．実験の結果を表 **4** に示す．いずれも編集距離に重みを導入することによって精度が向上した．特に，EDR/m では，効果が顕著であった．これは，EDR/m の入力が比較的長いため，重みを考慮していない初期状態では，不適切な規則が適用される可能性がより高くなったためだと考えられる．

次に，形態素分割の効果を調べるために，編集距離の重みを考慮して反復学習をした上で以下のモデルの比較をおこなった．

- 分割なし：形態素分割なし
- 分割あり **(式 (8))**：形態素分割あり，式 (8) を使用．
- 分割あり **(式 (9))**：形態素分割あり，式 (9) を使用．

表 4　重み付き編集距離の翻字精度への効果
Table 4　Effects of weighted edit distances on transliteration

		精度 (%)		
		NTCIR2	EDR/1	EDR/m
重み無し	Top 1	70.3	62.0	46.3
	Top 10	89.7	87.0	65.9
重み付き	Top 1	79.5	65.6	54.7
	Top 10	92.4	88.0	72.0

実験結果を表 **5** に示す．

表 5　形態素分割による翻字精度への効果
Table 5　Effects of morphologial segmentation on transliteration

		精度 (%)		
		NTCIR2	EDR/1	EDR/m
分割なし	Top 1	79.5	65.6	54.7
	Top 10	92.4	88.0	72.0
分割あり	Top 1	83.2	67.7	62.1
(式 (8))	Top 10	96.2	89.0	90.1
分割あり	Top 1	89.7	67.7	68.2
(式 (9))	Top 10	97.3	88.5	92.1

式 (9) によって連接関係を考慮することによって，EDR/m の精度が改善された．これは，出力単語の間の関係を考慮して，よりもっともらしい単語列を出力できたことを示す．しかし，式 (9) では単語の生成確率を条件付き確率 $P(e_i|e_{i-1})$ で計算するため，その確率は一般に式 (8) における生成確率 $P(e_i)$ より大きくなる．そのため，式 (9) を用いたモデルでは複数単語からなる解が式 (8) を用いたモデルより上位に現れやすくなると言える．そのため，EDR/1 では連接関係を考慮することによりわずかに精度の低下が生じている．

4. 翻訳曖昧性の解消

辞書を用いてクエリを翻訳すると，一般にそれぞれの語に複数の訳語が存在し，曖昧性が生じる．クエリ翻訳手法を用いた CLIR でこの曖昧性を解消するために，文書集合の言語コーパスから学習した共起確率を用いるのが一般的である．同一クエリ中に出現する単語同士は，関連性があると考えられる．一方，関連性のある語同士は文書中でも共起する確率が高いと予測できるからである．本節では局所的な共起を用いる手法，大域的な共起を用いる手法のそれぞれについて説明した後，これら局所的な共起と大域的な共起を組み合わせた手法を提案する．

4.1 局所的共起の利用

藤井らは，日本語の専門分野における複合語の多くが，その要素語の訳を並べることで翻訳可能であることを指摘し，語基辞書を用いた CLIR の曖昧性解消に，Bigram を利用している[3]．これは，注目する単語の直前のみの，言わば局所的な共起情報を用いて曖昧性を解消する手法といえる．この手法で，日本語の複合語 $S = s_1, s_2, \ldots, s_n$ が英語の複合語 $T = t_1, t_2, \ldots, t_n$ に翻訳される確率は，Bigram を用いて以下のように計算できる．

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

$$P(S|T) = \prod_{i=1}^{n} P(s_i|t_i) \qquad (10)$$

$$P(T) = \prod_{i=1}^{n} P(t_i|t_{i-1})$$

ここで，$P(s_i|t_i)$ は語基辞書から得る．すなわち，2 節で述べた手法により学習した語基の翻訳対において，

$$P(s_i|t_i) = \frac{freq(s_i, t_i)}{\sum_s freq(s, t_i)}$$

である．しかし，この手法では直前の単語以外の文脈は考慮していない．藤井らは NTCIR-1 のデータを用いた実験をおこなっているが，表 **6** に示すように NTCIR-2 は NTCIR-1 と比べてクエリが長く，句の数が多いので，Bigram による曖昧性解消では，より多くの情報の損失が生じ，適切な訳が得られない可能性がある．

表 6　NTCIR クエリ (Description) の平均形態素数と句数
Table 6　Average number of morphemes and phrases in NTCIR topics (Description)

クエリ	形態素数	自立語数	句数
NTCIR-1	8.81	5.62	2.62
NTCIR-2	14.5	9.18	3.98

4.2 大域的共起の利用

前田らは，Web 全体をコーパスとして利用し，検索エンジンに翻訳候補語の組を与えることで検索されたページ数から単語間の結束性を計算する手法を提案している[6]．彼らは相互情報量や χ^2 検定など複数の尺度を用いて結束性を計算し，結束性の総和を最大化する対を選択することで曖昧性解消をおこない，日本語・英語間の CLIR 実験において，いずれの尺度も同程度の効果を示すことを報告している．

しかし，この手法ではクエリに含まれる語から得られる翻訳候補語のすべての対について計算をおこなう必要があり，計算量が膨大になる．また，藤井らが指摘しているように，複合語では局所的な共起を用いた方が適切な翻訳が可能である．そこで，本論文では，

式 (10) を用いて句の翻訳候補の絞り込みをおこなった後，χ^2 検定を用いて句の翻訳候補の間の結束性を求めて曖昧性解消をおこなう手法を提案する．

4.3 局所的共起と大域的共起の併用

日本語クエリから得た句の集合 $J = \{j_1, j_2, \ldots, j_m\}$ を英語句の集合 $E = \{e_1, e_2, \ldots, e_m\}$ に翻訳するとき，χ^2 検定を用いたスコア $sc(E)$ を式 (11) で定義する．

$$sc(E) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \log(CHI(e_i, e_j)) \qquad (11)$$

$CHI(e_i, e_j)$ は，前田らと同様に式 (12) で計算する．

$$CHI(e_i, e_j) = \begin{cases} \frac{N(|\chi(e_i,e_j)|-\frac{N}{2})^2}{n(e_i)n(\bar{e_i})n(e_j)n(\bar{e_j})}, \text{ if } \min(n(*,*)) < 5 \\ \frac{N(\chi(e_i,e_j))^2}{n(e_i)n(\bar{e_i})n(e_j)n(\bar{e_j})}, \text{ otherwise} \end{cases} \qquad (12)$$

$$\chi(e_i, e_j) = (n(e_i, e_j)n(\bar{e_i}, \bar{e_j})) - (n(e_i, \bar{e_j})n(\bar{e_i}, e_j))$$

ここで，N は検索対象の文書数，$n(e)$ は単語 e が出現する文書数，$n(\bar{e})$ は e が出現しない文書数を，$n(e_i, e_j)$ は e_i と e_j が同時に出現する文書数を表わす．また，式 (11) に式 (10) で求めた $P(e_i|j_i)$ を加味したスコア $scw(J, E)$ を，式 (13) で定義する．

$$scw(J, E) = \exp(sc(J, E)) \cdot \prod_{i=1}^{m} P(e_i|j_i) \quad (13)$$

4.4 クエリ翻訳の手順

クエリの翻訳処理では，まず茶筅による形態素解析結果から翻訳すべき語の抽出をおこなう．本論文では，「自立名詞」,「未知語」,「接頭辞」,「接尾辞」,「アルファベット」のいずれかが品詞として付与された語を抽出した．

次に，抽出した語が連続して出現する部分を句と見なし，句ごとに翻訳対象語を決定する．なお，連続するカタカナ語はひとつの形態素にまとめる．これは，カタカナ語は未知語が多く，不適切な解析結果が得られる場合が多いためである．

句の中の翻訳対象語は，辞書を参照して句を最長一致法によって分割することによって決定する．ここで用いる辞書は，2 節で構築した語基辞書と，後述するように EDR 和英辞書から対訳が 1 英単語だけからなる対訳関係を抽出して構築した一般語辞書の 2 つである．

たとえば NTCIR-2 クエリ (ID=0138) 中には「安定三重項カルベン」という句があり，これは，形態素解析により「安定/三/重/項/カルベン」と分割される．

表 7 実験データの諸元
Table 7 Statistics of experimental data

データ	クエリ数	文書数	正解文書数
NTCIR-1	21	187,080	1,756
NTCIR-2	49	322,058	1,410

ここで，辞書に「三重項 (*triplet*)」という訳が存在すれば，それを翻訳対象語とし,「三」や「項」の翻訳は考慮しない．

さらに接尾辞を含む語については以下の処理をする．まず，接尾辞はそれ自身単独では翻訳対象としない．たとえば「受容体」(ID=0106) が「受容/体」と形態素分割され,「受容体」が辞書にない場合,「受容」は翻訳対象とするが,「体」は接尾辞なので翻訳対象としない．逆に「受容体」が辞書にある場合は,「受容体」を翻訳対象とすることは上述のとおりだが,「受容」も翻訳対象に加える．これは，語基の学習源が専門語辞書の見出し語に現れる複合語であり，接尾辞が省略されることがあるためである．また，接尾辞を省略しても概念的な相違は少ないため，より文書に適切な訳が得られる可能性がある．

翻訳候補語は語基辞書，一般語辞書，翻字の順の逐次検索により取得する．すなわち，語基辞書に記載された語については一般語辞書の検索をおこなわない．

続いて単語ごとの翻訳候補語から得られる任意の組み合わせについて式 (10) を用いて翻訳確率を計算し，上位 10 件を得る．

最後に，句ごとの翻訳候補語から得られる任意の組み合わせについて式 (11) あるいは式 (13) を用いて大域的な共起度を計算し 1 位の組み合わせを出力する．

4.5 評 価 実 験

評価実験として，NTCIR コレクションを用いた実験をおこなった．NTCIR コレクションは検索課題，文書，正解判定を含む．本論文ではそこから日本語検索課題と英語文書を用いた．クエリは，NTCIR-1 の検索課題 (ID=0001-0030) と NTCIR-2 の検索課題 (ID=0101-0149) から Description フィールドを抽出して利用した．正解判定は，部分適合を不正解と見なす rigid 判定を用いた．文書やクエリに関する情報を表 **7** に示す．

語基辞書は専門語の対訳関係から学習したものであり，一般的な語彙が不足している．そのため，EDR 和英辞書から対訳が 1 英単語からなるの対訳関係 373,265 対 (173,249 見出し) を抽出して式 (10) の $P(s|t)$ を算出し，一般語辞書を作成した．得られた辞書は見出し数 105,238，英単語 40,693 語，対訳関係 283,282 対を含む．

クエリと文書の類似度は TF·IDF で重み付けしたベクトル空間モデルで計算し，上位 1,000 件を出力した．クエリ Q と文書 D の類似度は式 (14) で計算する．

$$sim(Q,D) = \sum_{t \in Q} tf(t,D) \cdot idf(t) \qquad idf(t) = \log(\frac{N}{df(t)}) \tag{14}$$

$tf(t,D)$ は文書 D 中の語 t の頻度，N は文書の総数，$df(t)$ は語 t の出現する文書数である．

翻訳曖昧性の解消には，以下の 4 つの手法を用いた．

- **bigram1:** 式 (10) を使い，句の内部の Bigram を最大化する候補を選択する．
- **bigram2:** 句単位への分割を考慮せず，クエリ全体を一つの句とみなし，全体の Bigram を最大化する候補を選択する．
- **chi:** 式 (11) を最大化する候補を選択する．
- **combi:** 式 (13) を最大化する候補を選択する．

また，ベースラインとして，以下のような機械翻訳システムにより日本語クエリから翻訳した英語クエリ，人手で作成した英語クエリを用いた．

- **MT1:** NTCIR-1 日本語検索課題を Web 翻訳エンジンで翻訳
- **MT2:** NTCIR-2 日本語検索課題を Web 翻訳エンジンで翻訳
- **MN2:**NTCIR-2 英語検索課題の Description

翻訳エンジンとしては,「Excite 翻訳☆」を用いた．この際，カタカナ語の翻訳に失敗し，ローマ字表記が出力されたものについては正しく置き換えた．ベースラインの精度を表 **8** に示す．

表 **8** 機械翻訳・人手翻訳による英語クエリの結果
Table 8 Results of English queries translated by a MT system and human

手法	再現率 (%)	R 精度 (%)
MT1	44.2	21.5
MT2	43.3	8.73
MN2	42.0	9.20

まず，語基辞書と翻字による検索性能の変化を調べた．再現率・精度曲線を図 **3** に，再現率と平均精度を表 **9** に示す．なお，曖昧性解消には式 (13) を用いる手法 (combi) を利用した．図，表中で「語基辞書」は 2 節で述べた手法により構築した語基辞書を用いていることを表わす．(藤井 1)，(藤井 2) は藤井らによるにより構築した語基辞書を用いていることを表わす．「翻字」は 3 節で述べた手法による翻字を用いていることを表わす．

☆ http://www.excite.co.jp/world/text/

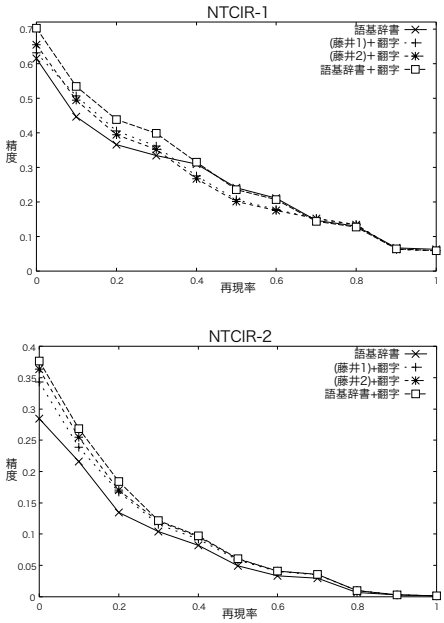

図 **3** 語基辞書，翻字による再現率・精度曲線
Fig. 3 Recall-precision curve with base-word dictionary and transliteration

表 **9** 語基辞書，翻字による検索性能 (%)
Table 9 Effect of base-word dictionary and transliteration

手法	NTCIR-1		NTCIR-2	
	再現率	平均精度	再現率	平均精度
語基辞書	32.4	23.7	31.4	7.72
(藤井 1)+翻字	49.2	24.3	38.2	8.99
(藤井 2)+翻字	49.3	23.9	38.7	9.33
語基辞書+翻字	47.8	26.2	39.2	9.69

次に曖昧性解消の手法による検索性能の変化を調べた．再現率・精度曲線を図 **4** に，再現率と平均精度を表 **10** に示す．語基辞書には 2 節で提案した手法の辞書を用い，未知のカタカナ語は 3 節で提案した手法によって翻字をおこなった．また，ベースラインとして Web 翻訳を用いた場合も併せて示した．

4.6 考　　察

2.3，3.3 の実験で示したとおり，提案した語基辞書の学習，翻字はいずれも単独で従来手法より高い精度が得られた．これらを CLIR システムに組み込んでシステムの性能を評価した結果，検索性能を改善できることが確認できた．しかし，NTCIR-1 のデータを用いた場合，再現率が低下している．

このひとつの理由として,「故障診断システム」とい

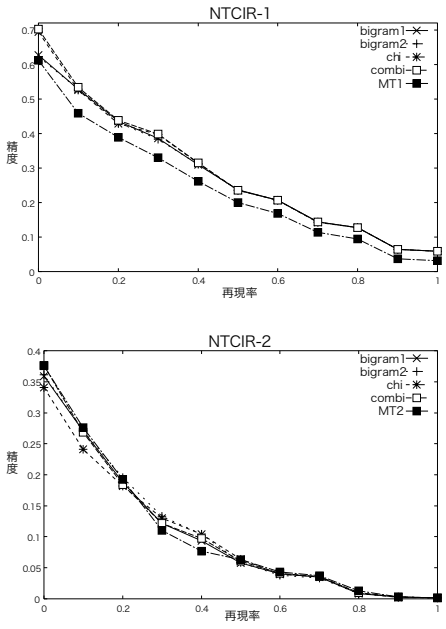

図 **4** 曖昧性解消手法による再現率・精度曲線
Fig. 4 Recall-precision curve with translation disambiguation

表 **10** 曖昧性解消手法による検索性能
Table 10 Effect of translation disambiguation

手法	NTCIR-1		NTCIR-2	
	再現率	平均精度	再現率	平均精度
bigram1	45.6	25.8	38.7	9.59
bigram2	44.8	25.7	<u>43.8</u>	<u>9.82</u>
chi	47.5	25.9	36.7	9.27
combi	<u>47.8</u>	<u>26.2</u>	39.2	9.69

うクエリ (ID=0014) に対して，提案手法の辞書を用いた場合に “*diagnosis system*” という訳を出力していたことが考えられる．提案手法では「遠隔故障診断 (*“remote diagnosis”*)」というエントリから,「故障診断」と “*diagnosis*” という対応を学習していた．この訳は，他の語基辞書で出力された訳 “*fault diagnosis system*” の一部であり，“*fault*” という概念が欠損したものと言える．そのため，より一般的な診断システム，例えば医療に関する文書などが上位に現れやすくなった．その結果，正解 151 文書のうち (藤井 1) や (藤井 2) 手法では 127 文書得られたものが，提案手法では 99 文書に減少した．

NTCIR-1，2 のクエリ中には 19 のカタカナ語が含まれており (NTCIR-1：6 語，NTCIR-2：13 語)，これらはいずれも提案手法によって正しく翻字できた．これらのカタカナ語のうち，クエリを翻訳する段階で正しく形態素分割できなかった例を表 **11** に挙げる．

表 **11** 翻字結果の例
Table 11 Example of transliteration

解析結果	翻字結果
キノ／ロン	quinolone
ビデオストリーミング	video streaming
ラベルスイッチルータ	label switch router
バイオ／フィルム	biofilm
インスタント／ン	instanton
プレストレストコンクリート	prestressed concrete

表 11 より，複数の語で形態素数が英単語数より多くなっており，連続するカタカナをまとめて翻字処理する必要性が確認できる．また,「ラベルスイッチルータ」と「プレストレストコンクリート」は，3.3 の実験においてベースラインでは正解を得られなかった．このことから，提案する翻字モデルの有効性が確認できる．

曖昧性の解消については，NTCIR-1 と NTCIR-2 で異なった結果が得られ，提案手法の優位性を確認することはできなかった．提案手法による翻訳結果を bigram2 手法と比較すると，表 **12** のような相違があった．括弧中の数値は，それぞれ再現率と平均精度である．

表 12 から，同じような文脈で使われやすい語を翻訳する際には大域的な共起情報を考慮することによって悪影響を受けることがあることがわかる．たとえば,「電磁特性」の例は,「電磁特性」が「磁気特性」と同じような文脈で使われることが多く,「磁気特性」と “*magnetic property*” の対訳関係の頻度が高かったことと，クエリが短かくこれらを十分に弁別する手がかりが他になかったために，誤って翻訳されたと考えられる．この問題は，閾値を用いて複数の翻訳候補を得るなどすれば影響を軽減できる可能性がある．

また，精度にはあまり影響がなかったが，不適切な訳として「宇宙定数問題 (*space dissipation problem*, ID=0129)」のような例があった．これは，クエリや個々の句が比較的長く，正しい訳語 “*space constant problem*” と共起しない句が存在したためだと考えられる．より大きなコーパスから共起情報を学習したり，文書中での要素語の位置を考慮しながら句の出現頻度を求めるなど，より正確に結束性を求める手法が必要であろう．

5. 結　　論

本論文ではクエリ翻訳による日英言語横断情報検索において，翻訳辞書の語彙不足を補うために，新しい語基辞書の獲得手法と翻字手法を提案した．実験により語基辞書の学習，翻字ともに従来手法より高い精度

表 **12** 大域的共起の利用による翻訳結果
Table 12 Effect of translation disambiguation with global coocurrence

句 (クエリ ID)	bigram2 (再現率/平均精度 (%))	combi (再現率/平均精度 (%))
手法，理論 (0028)	approach, theory (25.8/6.81)	method, algorithm (30.8/11.2)
重力波 (0113)	gravitational wave (71.4/38.4)	gravity wave (57.1/18.1)
構築 (0146)	architecture (100/15.7)	construction (100/44.6)
電磁特性 (0147)	electromagnetic properties (45.5/11.6)	magnetic properties (5.51/0.8)

が得られることを確認した．さらに翻訳曖昧性を解消するために局所的な単語共起と大域的な単語共起を併用する手法を提案し，これらの手法を実験システムに組み込み NTCIR テストコレクションを用いて評価をおこなった．実験の結果，システム全体の精度も向上することを確認した．しかし，局所的な単語共起と大域的な単語共起を併用した曖昧性解消は精度の向上には貢献しなかった．この理由の大きなものは，句単位での共起を学習するにはコーパスの規模が十分でなかったことが考えられる．また，単語間の距離などを考慮して単純な同一文書中での共起より精密な結束性の推定が出来ればさらに精度を改善できる可能性がある．さらに，本論文では語単位の翻訳をおこなっているが，句単位の翻訳の利用も今後の課題である．

謝　　辞

NTCIR コレクションは国立情報学研究所 (NII) の許可を得て使用させて頂きました．この場を借りて深謝いたします．

参 考 文 献

1) Brill, E. and Moore, R. C.: An Improved Error Model for Noisy Channel Spelling Correction, *Proceedings of the 38th Annual Meeting of the Association for Computarional Linguistics (ACL 2000)*, Tokyo, Japan, pp.286–293 (2000).
2) Fujii, A. and Ishikawa, T.: Cross-Lauguage Information Retrieval ad ULIS, *Proceedings of th 1st NTCIR Workshop* (1999).
3) Fujii, A. and Ishikawa, T.: Japanese/English Cross-language Information Retrieval: Exploration of Query Translation and Transliteration, *Computers and the Humanities*, Vol. 35, No. 4, pp. 389–420 (2001).
4) Gary, E.B.: Automatically Harvesting Katakana English Term Pairs from Search Engine Query Logs, *Proceedings of the 6th NLPRS*, pp. 393–399 (2001).
5) Knight, K. and Graehl, J.: Machine Transliteration, *Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the EACL* (Cohen, P. R. and Wahlster, W.(eds.)), Somerset, New Jersey, pp. 128–135 (1997).
6) Maeda, A., Sadat, F., Yoshikawa, M. and Uemura, S.: Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine, *Proceedings of IRAL2000*, pp. 25–32 (2000).
7) Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proceedings of International Conference on New Methods in Language Processing*, pp. 44–49 (1994).
8) (株) 日本電子化辞書研究所: EDR 電子化辞書仕様説明書 (1993).
9) 北研二: 言語と計算–4 確率的言語モデル, 東京大学出版会, chapter 7.1 (1999).
10) 国立情報学研究所: `http://research.nii.ac.jp/ntcir/index-ja.html`.
11) 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸: 形態素解析システム『茶筌』Version 2.3.3 使用説明書.