自然言語処理 122−11
(1997. 11. 20)

多言語情報検索における利用者支援について
−主要キーワードの対訳付与に関する検討−

鈴木 雅実　井ノ上 直己　橋本 和夫

KDD 研究所

〒356 上福岡市大原 2-1-15

E-mail: {msuzuki, inoue, kh}@lab.kdd.co.jp

最近の WWW による情報発信および情報閲覧環境の目覚しい発展は，多様な言語による情報源の飛躍的な増大をもたらし，言語間にまたがるクロスリンガル情報検索も現実の課題となろうとしている。クロスリンガル情報検索では，当然のことながらある言語による検索要求に対して検索対象として指定された異なる言語のコンテンツが検索されることが必須条件である。一方，検索結果に対しても検索要求言語による概要が提供されれば，利用者にとって有益な検索支援情報となる。本報告では，この目的を実現するための最初の段階として，検索結果の文書中の主要キーワードを検索要求言語に対訳する試みを紹介し，試作中の「多言語情報サーバ」における実際の検討課題を述べる。

Browsing Support in Cross-Language Information Retrieval

Masami SUZUKI　Naomi INOUE　Kazuo HASHIMOTO

KDD Research and Development Laboratories
(2-1-15, Ohara, Kamifukuoka-shi, 356 JAPAN)

E-mail: {msuzuki, inoue, kh}@lab.kdd.co.jp

Cross-language information retrieval (CLIR) has become a realistic target in the current WWW environment, where a lot of multilingual information sources are available for users to browse with multilingual fonts supports. A cross-lingual search engine should enable users to input queries in their native (or familiar) language for retrieving certain target language(s). As well as this obvious purpose, we emphasize that useful information in the query language should be attached in the retrieval result in an original (target) language. As an approach to this goal, we are trying to give suitable translation for major keywords in the retrieved documents in a search result list. In this article, we describe the related ideas for supporting cross-language text retrieval that we are examining in our prototype of "Cross-Lingual Information Links Server".

1 はじめに

最近のWWWによる情報発信および情報閲覧環境の目覚しい発展は，多様な言語による情報源の飛躍的な増大をもたらし，言語間にまたがるクロスリンガル情報検索 (cross-language information retrieval) [1] も現実の課題となろうとしている [6]。実際，従来理論面での考察や実験報告が主であったのに対し，実践面での具体的なアプリケーション事例が見られるようになってきた。

理論的な手法の比較を行なった最近の例として，Carbonell らの論文 [1] があるが，一般に情報検索手法やシステムに関して議論される性能項目は，多くの場合検索要求入力に対する再現率／正解率の値であり，その他の機能項目については付随的な色彩が強い。しかし，クロスリンガル情報検索の場合は，検索要求に合致したコンテンツを多く含む検索結果を導くシステム側の検索効率ばかりでなく，検索結果一覧の中から利用者の検索意図に近いコンテンツを比較的容易に選択可能とする支援的な側面が，モノリンガルな情報検索と比較して，より重要であると考えられる。

すなわち，利用者が検索要求を行なった言語 (利用者の母国語あるいは馴染の言語) とは異なる言語の検索結果に対し，検索要求と同じ言語による付加情報が与えられることにより，検索結果の中から必要な情報を取り出すことが容易になる可能性が高い。筆者らは，このような考察の下にクロスリンガル情報検索を支援する「多言語情報サーバ」CLINKS (Cross-Lingual Information Links Server) のプロトタイプを試作している。以下では2章でクロスリンガル情報検索支援における問題点を分析し，3章でCLINKSで試行中の検索結果一覧提示の際の各文書へのキーワード対訳情報付加機能について述べる。さらに4章において手法の評価等の検討課題について考察する。

2 クロスリンガル情報検索支援の問題点

冒頭に述べたように，インターネット上でクロスリンガル情報検索を実用指向のサービス機能として実現するための環境が整いつつある。最近のWWW情報源の飛躍的な増大に加えて，一般の利用者向けの多言語表示環境の改善により，多くの利用者が多様な言語のテキスト情報に接する機会が増えている。ところが，現在提供されている主要な検索サービスのほとんどは，単一言語内 (モノリンガル) の検索であり，多くの言語が検索可能となっている検索エンジン (たとえばALTAVISTA) でも，検索処理対象となる各言語毎のDBを選択するような仕組みを提供しているに過ぎない。すなわち，この種の多言語対応ではあるがクロスリンガル対応ではない検索機構は，検索対象言語に対して十分な知識を持たなければ使用は困難である。

これに対し，検索対象言語に関する知識が十分かどうかによらず，次のようにクロスリンガル情報検索支援への要求が生ずる場合が想定される。

1. 検索対象言語に関する読解知識が十分な場合

   このケースでも，原語によるキーワード入力が困難であったり (能力および端末環境の問題) 正確な用語が分からない等の状況下では，利用者が最も利用し易い言語による検索要求が受け入れられることが望ましい。

2. 検索対象言語に関する知識が一応はある場合

   辞書を引きながら読むことは可能である等種々のレベルが想定されるが，ある対象言語の文書の探索を行なう際の的確な用語の指示は1の場合よりもさらに困難であろう。また，検索結果が得られたとしてもその閲覧には時間を要する。

3. 検索対象言語に関する知識が殆んどない場合

   このケースでは，検索対象言語で検索要求を出すことはほぼ不可能であり，クロスリンガル情報検索支援が最も必要となる状況である。このような条件での検索ニーズとしては，検索結果としてテキストではなく画像を求めている場合等が考えられる。

いずれの場合にも，検索要求とは異なる検索対象言語のコンテンツから検索結果を導く処理が実行されることになるが，上記の2および3のケースの利用者にとっては，検索結果が原語のまま表示されるだけでは，検索機能としては十分とは言えない。そこで，検索内容に関して何らかの利用者の (検索要求) 言語による付加情報が求められる。たとえばTITAN([3])では，検索結果一覧中に英

[1] 最近の英語表記では "cross-language" が主流のように思われる．日本語では「言語間」という用語も考えられるが定着度に決め手がなく，ここでは文献 [4] に倣った．

日逐語翻訳されたHTML文書のタイトルが表示される。しかし，HTML文書のタイトルは文書の内容を代表するには非常に一般的であったり(欠落している場合もある)，便宜的に付与されることも多く，翻訳が成功したとしてもコンテンツの探索に十分有効とは限らない。また，QUILT([2])[2]では，検索結果一覧中から選択された文書については，原文書(スペイン語)とその逐語訳文書(英語)が提示される。このように，検索結果の中からの探索に関する限り，現在の利用者支援方式はまだ十分であるとは言えない。

理想的には内容の的確な短い要約を生成することが文書探索時の支援情報の提供として有益であると考えられるが，この課題は長期的な研究を要すると思われる。そこで，筆者らは「コンテンツ中から重要な語句(キーワード)を抽出しそれに適切な対訳を与えること」をそれに至る中間目標として設定した。重要な語句を抽出すること自体も研究的には大きな課題であり，手法の評価も難しいと思われるので，当初は比較的利用し易い語の頻度情報に着目して試行を開始した。

3 CLINKSにおけるクロスリンガル検索支援

前章の議論を踏まえ，筆者らの試作システムCLINKS(Cross-Lingual Information Links Server)では，まず第1段階として，検索のために生成するインデックスから導かれる，文書中の出現頻度の高い語(機能語は除く)に対して対訳を提供することとし，

・検索結果一覧表示時に，対象文書内の主要キーワードを利用者の言語で表示
・検索結果一覧から選択された文書の閲覧時に，文書内の重要語を対訳表示

することにより，クロスリンガル情報検索の有用性の向上を図ろうとしている。

3.1 検索支援機能の概略

上記の訳語付与の精度を向上させる手法の提案について述べる前に，その前提となるCLINKSにおけるクロスリンガル情報検索機能の概略を，本システムに特有な機能／処理を中心に記述する。

3.1.1 コンテンツのプロファイル作成機能

CLINKSで検索対象としているWWWコンテンツは，東アジア地域の主要な国々および地域に関する情報発信サイトから，コンテンツ収集ソフトウェア(ロボット)[3]により収集したHTML文書である。1997年9月現在で約22,000ページに相当する。取得したコンテンツは，各文書毎に言語／コードの推定を行ない，WAIS Indexerと形態素解析(日本語・中国語)を用いて検索インデックスを作成し，英語／日本語／中国語のそれぞれのディレクトリに保存する。また，この際に主要キーワード・リストの生成と検索結果一覧に提示する対訳情報の付与等の処理を行なう。

3.1.2 クロスリンガル情報検索機能

CLINKSの初期画面では，キーワードの入力・検索条件の選択(ANDまたはOR)，検索対象言語の選択(日・英・中)，検索対象ドメインの選択(日本・米国・中国・香港・台湾・その他)を行なうための入力／指示が可能となっている。入力文字列から切り出された(スペースに従って分割)キーワード候補は，対訳辞書とマッチングを行ない，対訳辞書に記載されている対訳候補の論理和を，検索用キーワード集合として用いる(現在の初版での処理)。この後，WAIS[4]の検索モジュールが起動され，各言語毎の検索結果が導かれる。

3.1.3 クロスリンガル情報検索支援機能

(1) 検索結果一覧表示における主要キーワードの表示

CLINKSでは，検索結果は検索対象言語(日・英・中)毎に個別にスコア順にソートされた一覧を表示している。表示項目は次の通りである。

・文書のタイトル(原語表示)
・文書のURL
・文書の存在するドメイン(国等に対応)
・文書のサイズ(Kbyte)
・文書中の主要キーワード・リスト(現在6個)

[2]辞書と対訳コーパスを組み合わせたアプローチで注目に値する研究成果が報告されている.

[3]KDD-Explorer

[4]free WAIS 0.5の改良版を使用

このうち，タイトルと主要キーワード・リストの表示部分が，利用者にとって検索結果一覧中から意図に近い文書を発見するための支援情報として提供されるものである。CLINKSではタイトルは原語のまま表示し[5]，主要キーワード(文書内の頻度に準拠)については，利用者の検索要求言語(日本語または英語)に翻訳したものを提供することを試行している。この際の問題は，次節で述べるように，複数の訳語候補から適切なものを選択することである。

(2) 文書閲覧時における重要キーワードの対訳表示

検索結果一覧から，ある文書のタイトル表示部分が選択(クリック)された場合，CLINKSは該当文書そのものの閲覧に加えて，文書中の重要キーワード[6]と，その対訳(検索要求言語)をブラウザ画面上の下部に生成される「支援フレーム」内に表示する機能を持つ。

3.2 キーワード対訳精度の向上手法

現在試行中の主要キーワードの対訳情報の提供については，各キーワードに対して対訳辞書中からデフォールトの対訳語候補を与えるレベルから，対象となる主要キーワードの組合せに対し最適な対訳語の組み合わせを選択するレベルまで，様々な段階が存在する。

最適な訳語の組合せを与えるためには，多くの学習データや文書の解析が必要となり，現実には達成が困難である。そこで，すでに蓄積した文書コーパスから，テキスト中の2単語の共起分布に関するデータを作成し，2単語の組合せの出現頻度の違いから，対象となる(6個ないし適当個数の)キーワードの対訳を推定することを検討している。方法は以下の通りである。

(1) 基本的なアイディア

検索要求キーワードを対訳辞書を用いて対象言語に変換する場合(対訳キーワードを一意に決定することは困難)とは異なり，検索結果文書中の主要キーワードの対訳を求める場合は，それらのキーワードが「同一文書中に共起している」ことを利用して訳語の曖昧性を解消することが比較的容易であると考えられる。もちろん，このためには語の共起に関する何らかの知識ないし学習データを必要とする。

検索支援の目的からすれば，検索対象となる文書のコーパスの特徴を反映した共起情報の参照が望ましいが，ここで問題は原語(たとえば英語)の主要キーワードの共起を直接利用することが一般に困難な点である。もし，Carbonellらのように([1])，大量の対訳コーパスから用例翻訳の手法を用いてキーワードの訳語を学習させることが可能であれば，対訳を提供する言語(たとえば日本語)の側でも，より正確な対訳語の組合せを選択することが容易になるはずである。

しかし，CLINKSで検索対象としているような多言語のWWW文書では，検索対象となるすべての言語による対訳コーパスを用意して訳語候補を予め絞っておくことはあまり現実的ではない。そこで，対訳辞書に用意したあるキーワードの対訳候補の中からより適切な訳語を選択するため，訳語生成側の言語でのコーパス内の語の共起データを参照することとする。すなわち，複数の訳語候補同士の同一文書内での共起の度合を比較することにより，近似的に尤もらしい訳語の組合せを決定する。

(2) 実装上の工夫

上記の問題は，次のように記述できる。言語の主要キーワード(英語と仮定)のリストを

$$(W_{E1}, W_{E2}, ..., W_{Em})$$

とする。(日本語)訳語候補が各 W_E に対して数個づつ存在するとすれば，たとえば

$$((W_{J11}, W_{J12}), (W_{J21} \sim W_{J23}), ..., (W_{Jm1}))$$

の各 W_J の中から適切な組合せを選択する問題となる。コーパス中から m 個(6個程度を想定)の語の組合せの共起頻度を求めることはデータ量の問題等から実際的でないので，2つの語の共起頻度を利用する。すなわち，検索対象コーパスのうち対訳生成言語(日本語)の文書コーパスに対して，相対的に出現する頻度(unigram)の高い語同士の同一文書内の出現頻度(bigram)データを予め作成し，これを参照して各 W_Jのどの訳語候補を選択すべきかを決定する。現在，実用上十分な共起データの量，および計算コストの削減手法等について検討を進めている。

[5]利用者の利用環境を考慮し，たとえば中国語検索結果に関しては中国語の文書タイトルはイメージで表示している。

[6]現在の版では，参照する対訳辞書内容に依存している．

4 検討課題

4.1 キーワード対訳情報提供機能の効果測定方法

クロスリンガル情報検索結果一覧に対して，一定個数の主要キーワードの対訳情報が示されることがどの程度の支援効果を持つかを知るための，評価方法について試案を述べる。比較対象は次のような条件の違いである (表 1 に実験条件の違いを例示)

1. 各キーワードを原語のまま表示
2. 各キーワードに対して対訳辞書中の最初 (デフォールト) の対訳語候補を表示
3. 対象となる主要キーワードの組合せに対し，共起分布を考慮した対訳語の組合せを表示
4. 人手で選んだ同個数のキーワードに対し適切な訳語を表示

これらの条件を比較するため，評価用の文書セットを用意し，被験者に次のような課題を課す。

- 検索要求言語 (日本語とする) で書かれた文書の内容 (検索目標) を一定の短時間で把握する。
- 検索対象言語 (英語とする) の検索結果一覧を探索し，主要キーワードリスト等を参照しながら検索目標に合致した文書を同定する。

上記の探索課題に費やされた時間，操作回数等を比較することにより，各条件の支援効果の差を測定する。被験者による個人性の問題を極力回避するため，各被験者に対して複数行なう試行毎の支援情報提示条件の順序等について注意深く実験管理を行なう必要がある。また，このような評価方法以外の比較方法についても検討を試みるべきであろう。

表 1 支援情報の提示内容の違いの例

実験条件	主要キーワード
1	(home,artist,station,TV,organ)
2	(家, 芸術家, 駅, TV, 臓器)
3	(ホーム, アーティスト, 局, TV, オルガン)

4.2 主要キーワードの抽出方法

2 章でも述べた通り，対訳を提供すべき主要キーワードをどのように決定するかも重要な問題である。究極的にはクロスリンガルな要約生成の問題にも関連する重要語の抽出には，頻度情報ばかりでなく次ように様々なレベルでの手がかり情報が考えられる。

1. 語彙レベルでの特徴

   複合語や固有名詞はキーワードとしての重要性が一般に高く，主要キーワードの候補になり易い。

2. 文レベルでの特徴抽出

   特定の修辞語句や定型表現等を含む文を重要文と見なし，その中に含まれる自立語の重要度を加点する方法が考えられる。

3. 文書構造レベルでの特徴抽出

   表題や見出し的な部分，文書中のパラグラフの位置等の構造的な情報を考慮して重要度を計算する。

これらは，いずれも HTML のタグ情報等が利用可能であれば，さらに手がかり情報が増す。また，文書／コンテンツのスタイルに応じた重要度の定義を使用することも不可能ではないが，より多くのテキスト処理を行なう必要がある。さらに，利用者の目的や注目情報等に合致させた重要箇所の特定を行なう等のカスタマイズが求められることになろう。

一方，クロスリンガル情報検索を支援する視点で考えると，対象言語や文化に関する利用者の知識の程度や検索の利用目的を考慮した，より効果的な重要度の判断や訳語提供方式の提案が有り得るであろう。そのためには，CLINKS による今後のフィールド試験等の実践を通じて，望ましい利用者支援とそれを実現するため技術課題を明らかにする必要がある。

5 おわりに

本稿では，クロスリンガル情報検索における支援情報提供の重要性を述べ，この点に配慮した機能の実現案を示した。試作中のシステムにおける検索支援情報提供内容は，検索結果一覧における各文書中の主要キーワードの対訳であり，適切な訳語の選択をサーバが持つ検索対象コンテンツに基づいて実行する点に特徴がある。今後は提案手法を実装した上で，期待し得る性能上の限界を踏まえつつ，4章の検討課題で述べたような評価試験により，支援効果の比較を行なう予定である。

謝辞

本研究の遂行に際してご指導，ご助言を頂いたKDD研究所の村上所長・鈴木審議役に感謝いたします。また，有益なコメントを寄せられた関連研究グループの諸氏に深謝いたします。さらに，システム試作の過程および，方式検討等に際してご協力頂いた社内外の方々に，この場で厚くお礼申し上げます。

参考文献

[1] J. G. Carbonell et. al. : "Translingual Information Retrieval: A Comparative Evaluation", *Proceedings of IJCAI'97*, pp. 708-714 , Nagoya, 1997.

[2] M. W. Davis and W. C. Ogden: "Implementing Cross-Language Text Retrieval Systems for Large-scale Text Collections and the World Wide Web", *AAAI Spring Symposiumon Cross-Language Text and Speech Retrieval Electronic Working Notes*, 1997.

[3] 菊井 玄一郎, 他: "インターネット情報ナビゲーションにおける多言語機能", 自然言語処理の応用に関するシンポジウム, 情報処理学会, pp.97-106, 1995.

[4] 菊井 玄一郎: "インターネットと多言語情報処理", 情報処理, Vol.38 No.1, pp.1-8, 1997.

[5] M. Suzuki and K. Hashimoto: "Enhancing Source Text for WWW Distribution", *Proc. of Workshop on Information Retrieval with Oriental Languages (IROL'96)*, pp.51-56, Taejon, 1996.

[6] 鈴木 雅実, 井ノ上 直己, 橋本 和夫: "実用指向の言語間情報検索に関する一考察", 情報処理学会第54回全国大会, 1997.