実用指向の言語間情報検索に関する一考察

4K－6

鈴木　雅実　　橋本　和夫

ＫＤＤ研究所

1　はじめに

異なる言語間にまたがる情報検索 (Cross-lingual Information Retrieval) ないし，２言語以上の多様な言語に及ぶ情報源についての検索 (Multi-lingual Information Retrieval) については，種々の理論的な考察や手法の効果測定のための実験報告が見られるものの，具体的なアプリケーションの視点からの研究事例はまだ少ない。ところが，最近のインターネット (特に WWW) の普及に伴って，実際に多言語の情報源に接する機会が飛躍的に増加しており，その中から言語に依らず所望の要求を満たす文書を探索したり，新たな知識を発見することが現実のものとなろうとしている。本稿では，このような潜在的なニーズに応える，実用指向の言語間情報検索の課題について考察する。

2　多言語情報検索の現状

一般に多言語情報検索とは，検索要求言語と検索対象テキストの記述言語 (文書中に複数言語が混在する場合も含めて複数) が独立した関係にある情報検索と定義されている。これまで各種の関連研究が報告されているが，ネットワーク上での一般の利用を視野に入れた研究事例は比較的最近のものである。多言語情報検索を研究的な側面に焦点を当てたサーベイに文献 [3] がある。この報告では，代表的な２つのアプローチとして，多言語シソーラス (辞書) に基づく方法とコーパスに基づく方法を挙げ，その利害得失を述べている。要約すると，概念検索 (concept retrieval) 的な言語間にまたがる類義語検索に相当する前者のアプローチは，限定された領域では威力を発揮するが，拡大は容易ではない。一方，後者では学習用の対訳コーパスとは別の対象コーパスについて得られる検索性能の評価方法が問題である。結論としては，シソーラスとコーパスベースの技法を組み合わせるのが現実的な解であり，双方の特長を捉えることができそうであるが，両者の橋渡しを行なうための鍵とも言える，シソーラスの自動生成に関する研究等が焦点になると思われる。

この種の基礎的な研究の重要性は言うまでもないが，多言語情報検索の典型例として，現実のネットワーク上の情報源を対象とした実用指向の言語間情報検索を考える場合には，関連する種々の環境条件を考慮することが必須である。そこで，次章では，この観点から言語間情報検索に関する課題を整理し，実現すべき機能等について検討する。

3　実用指向の言語間情報検索の課題

3.1 ユーザモデルと情報フィルターの問題

言語間情報検索では，単言語内の場合と比較して多義性がより深刻な問題となる。また収集可能な情報源すべてを検索対象とすることは，処理の効率にも影響するので，基本的な検索機構に加えて情報フィルターの機能が必要となるが，大きく分けると

(1) 検索要求獲得支援のための情報フィルター

(2) 検索結果閲覧支援のための情報フィルター

が考えられる (図１参照)。この両面で，言語間情報検索の利用者のユーザモデルが深く関わってくる。

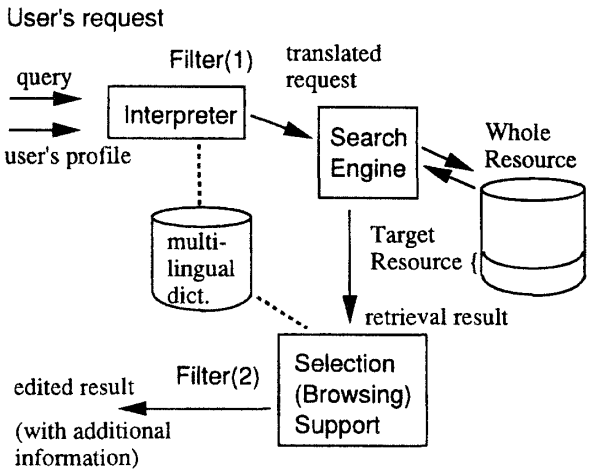

図 1: 言語間情報検索における情報フィルター

3.1.1 検索要求の獲得支援

Towards Practical Cross-lingual Information Retrieval
Masami Suzuki and Kazuo Hashimoto
KDD R&D Laboratories, 2-1-15 Ohara, Kamifukuoka-shi, Saitama 356 JAPAN　(msuzuki@lab.kdd.co.jp)

情報フィルター (1) は利用者の検索要求を明確化し，検索範囲を絞り込んだり検索語を限定することに相当する。検索対象を特に限定しない場合は入力される検索要求 (キーワードや問い合わせフレーズ) のみでは情報が不足するため，それを補う各種の属性情報からなるユーザプロファイルの獲得方法が課題となる。たとえば，検索対象地域・言語やジャンルを予め用意した項目から選択させることが一つの方法である。さらに，これに加えて，探索したい大まかなネットワーク上のドメイン (サイト) が明確に限定されている利用者の場合は，検索の出発点となるページあるいはドメイン情報の入力を基に検索範囲を絞り込む方法が考えられるが，様々な局面で利用者に多くの質問を課すのは望ましくない。従って，最小限のユーザ情報から，利用者モデル (プロファイル) を推定する方法の検討が必要である。

3.1.2 検索結果の閲覧支援

情報フィルター (2) は，検索結果の閲覧時に，利用者が必要な文書を選択する行為を支援することにより実現される。言語別の一覧表示や，検索結果の一覧表示時における読解支援として，部分的な翻訳・オンライン辞書引き機能や，文書中の検索要求入力以外の主要キーワードの対訳表示等が有効と思われる [1] [4]。この種の付加情報の提供が，文書の取捨選択に役立つほか，再検索を行なう上で有効と考えられる。また，利用者が検索対象言語に関してどのような能力を持っているかにより，提供情報内容を切替えることも考えられる。この点についても，前項で述べたようなユーザ情報 (モデル) の獲得とそれに基づく閲覧支援機能のカスタマイズが鍵となる。

3.2 多言語表示および言語識別の問題

テキスト情報検索のためのインタフェースとしての多言語検索機構には，検索対象となる (収集した) 文書が含む言語を識別し，適切な検索結果を導くばかりでなく，適切なフォントの選択等，表示その他の文書利用に必要な情報を補完することも求められよう。この点に関連する，文字コードの国際標準を巡る様々な動向や多言語同時表示，さらに検索対象となる文書が含む言語の識別等の問題については，文献 [1] に参照文献を含めた解説として詳しく述べられている通り，各言語文化を尊重する観点から注目すべき問題である。

さらに，不特定多数の言語間で情報検索を行なう場合，情報閲覧のためには多くの言語リソースが必要となるが，利用者が自己の文書閲覧のために必要最小限のリソース (端的な例ではフォント等) だけをダウンロードできるような機構の提案等も現実的な視点である [2]。

3.3 言語間の情報発信支援機能

言語間の情報交換を促進する観点からは，多言語情報検索に加えて，言語間にまたがる情報発信を支援することも，今後重要な点の一つである [1]。視点を変えれば，言語種別や主要なキーワードの対訳情報等，前に述べた検索結果に対する閲覧支援情報を，発信対象文書に自動的に付加すること (たとえば英語等他の言語によるキーワードの添付等) が可能であれば，その文書の付加価値が増し，元の文書とは別の言語による検索が容易となることが期待できる (図 2 参照) [4]。

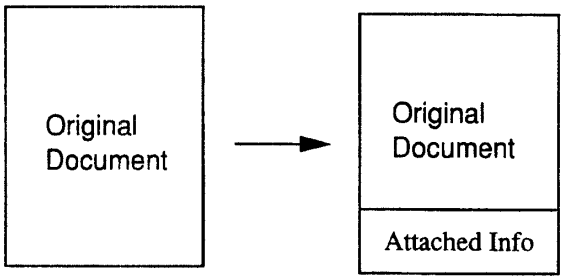

図 2: 言語間の情報発信支援

4 おわりに

言語間にまたがる (多言語) 情報検索，あるいはこれと対をなす情報発信の支援について，現況を踏まえ実用指向の観点から考察を行なった。検討課題として挙げた問題のうち，ユーザモデルの同定から得られる情報の追加が，言語間情報検索には特に有効と考えられる。今後はこの分野の試行がさらに拡大されるものと予想されるが，3 章で述べたような情報フィルターを伴うインタラクティブな検索機能の評価方法や，ネットワーク上での資源共有の有り方についての議論が深まることを期待する。

参考文献

[1] 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.

[2] 前田 亮, 他: “組み込みフォントを必要としない WWW のための多言語ブラウザ”, 図書館情報大学, 1995.

[3] D. W. Oard and B. J. Dorr: “A Survey of Multilingual Text Retrieval”, UMIACS-TR-96-19 CS-TR-3615, Univ. of Maryland, 1996.

[4] M. Suzuki and K. Hashimoto: “Enhancing Source Text for WWW Distribution”, *Proc. of IROL-96*, pp.51-56, Taejon, 1996.