クロスリンガル情報検索における閲覧支援機能について

4U−2

鈴木 雅実　井ノ上 直己　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，すなわちクロスリンガル情報検索 (Cross-Language Information Retrieval) に関する研究事例が近年増加しつつあり ([1]など)，インターネット上の情報検索における一分野としても動向が注目される。英語以外の言語による情報発信量の増加や，必ずしも英語能力の高くない利用者の検索要求等を考慮すると，今後クロスリンガル情報検索へのニーズは拡大するものと思われる。本稿では，クロスリンガル情報検索において，検索結果一覧から検索意図に近い情報源を特定するための参考情報を提供する支援機能の重要性を述べ，その第一段階として主なキーワードの対訳情報を与える手法を提案する。

2 クロスリンガル情報検索における閲覧支援

ここでは，ネットワークを通じたクロスリンガル情報検索を支援する機能について検討する。まず前提として，検索要求が利用者の母国語あるいは得意とする言語でなされるものと仮定する。その入力に対して特定の目標言語の文書が検索された後，次のような各段階での閲覧支援情報が提供されることが望ましい。

(1) 検索結果からの文書選択における支援

検索結果一覧中にリストアップされた各文書に関する抄訳情報等の提供。

(2) 対象文書の閲覧時における支援

対象文書に対応する翻訳結果の提供や，文書の理解を助ける用語説明など。

これらの関係を図1に示す。このうち，(1) の検索結果からの文書選択の支援は，数ある情報源からの検索意図に合致する可能性のある文書の特定を短時間で効率良く行なう上で極めて重要である。検索目的／対象にもよるが，通常，断片的な検索要求に対してノイズ (不適切な文書) を含まない検索結果を示すことは困難であることから，検索結果の中から実際に閲覧すべき対象文書を選択しなければならないのが普通である。また，明らかに，モノリンガルの場合に比較してクロスリンガル情報検索の場合は，利用者にとってより負荷を与える作業となる。さらに，検索結果を見て利用者が検索要求にフィードバックを与える枠組 (Relevence Feedback) を提供する上でも，検索結果を的確に判断することが不可欠である。従って，検索結果一覧中に利用者が文書の取捨選択を行なうために必要十分な情報を提示することは大きな意義を持つ。一方，(2) の実際の文書内容閲覧の際にも，質の高い全文翻訳が得られることが期待されるとしても，それが不可能な場合にも文書中の重要語句の対訳説明があればコンテンツの再利用 (たとえば画像とそのキャプション情報等) が図れる場面も想定できる [4]。

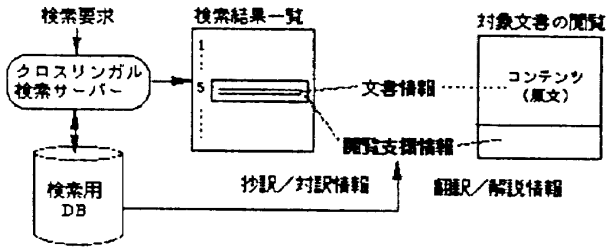

図 1: クロスリンガル情報検索における閲覧支援

3 閲覧支援としての対訳情報の提供

前章で述べたように，クロスリンガル情報検索において重要と思われる閲覧支援情報の提供を段階的に向上させることを目的として，本研究では，検索結果一覧中に各文書のコンテンツの特徴を反映した情報を利用者の検索要求言語で提供する方法について検討している。究極的には，各文書の抄訳 (的確な要約翻訳文) を提示することが望ましいと考えられるが，技術的にはまだ困難であり，その中間段階での目標を「文書中の主要キーワードの対訳を一覧中に表示すること」と設定した。これは，文書に付与された主要なキーワードの一覧が，情報検索において情報源の性格を判断する材料として有効であるとの仮定に基づいており，前

A Browsing Support Method for Cross-Language Information Retrieval
Masami Suzuki, Naomi Inoue and Kazuo Hashimoto
KDD R&D Laboratories, 2-1-15 Ohara, Kamifukuoka-shi, Saitama 356 JAPAN (msuzuki@lab.kdd.co.jp)

章で述べたコンテンツの取捨選択にどの程度貢献し得るかについては，後述するように評価が必要である。

現在の試行では，文書中の頻度の多いキーワード(10語以内の程度)についての訳語を文書タイトル(当面は原語で表示)とともに表示することとしている(表1の例を参照)。この際に，対訳辞書中に複数の訳語候補がある場合には，各文書中のキーワード分布を考慮した尤もらしい訳語を選択することが課題である。次章では，このような対訳情報を得るための手法の概略と問題点を述べる。

4 検索対象を反映した適切な訳語選択の手法

4.1 対訳生成言語コーパス中のKW共起情報の利用

Carbonell [2] は，大量の対訳コーパスを用いて学習させた訳語候補に基づくクロスリンガル情報検索の精度が，一般の対訳辞書を参照した場合と比較して格段に優れていることを，種々の統計的なモデルの適用結果とともに示している。このことから，検索対象コーパスと同等の性質を持つ対訳コーパスが存在すれば，同一文書中の一定個数のキーワード集合に対する，妥当な対訳キーワード集合を計算することが可能と考えられる。しかし，実際には検索対象とする文書の対訳コーパスを(多言語で)用意しておくことは非常に困難である。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(検索対象とほぼ同分野と仮定)内の語の共起データを参照することとする。すなわち，共起し易い訳語同士の組合せを優先することにより，近似的に尤もらしい訳語を決定する(この基本的なアイディアについては文献 [3] を参照のこと)。現在，インターネット上から収集した約5千ページの日本語WWW文書から抽出した総出現頻度上位2,000語間の共起データを基に訳語の選択を行なう実験を行なっている。表1にその例を示す。

表 1: 検索結果への訳語付与の例

検索語 = 放送 AND 衛星 AND 政策
検索例(経済構造改革に関する英文記事)
原語キーワード： (development,telecommunications, system,measures,examination,reform)
対訳表示： (開発, 電器通信, 制度, 措置, 検討, 改革)

4.2 問題点と今後の課題

表1の例に示したような主要なキーワードの対訳情報が，閲覧すべきコンテンツの選択にどの程度有効に利用可能であるか，すなわち支援情報としての参照価値を確認するため，評価実験を予定している。これは，被験者に対し，検索目的を予め検索要求言語(日本語)で示した場合に，検索結果の一覧表示における対象言語(英語)の文書中のキーワードの対訳提示が，目的に近い文書の選択にどの程度有効であるかを測定するものである。

また，本研究の提案と4.1に記した訳語選択の方式に関する現状の問題点としては，次のような項目が挙げられる。

(1) 主要なキーワードとして適切なものを選択する問題

現在は便宜的に文書中で頻度の多いキーワード(の対訳)を支援情報としているが，閲覧上重要な語句は文書の内容的な性格や，構造上の特徴，さらに利用者の関心等の様々な要因に応じて決定されるべきものであろう。文書の要約手法とも関連する大きな研究課題と考えられる。

(2) 訳語生成側の言語内の語の共起のみを参照していることによる性能劣化の問題

対訳生成言語側のコーパス中の共起データから訳語の組合せを決定する以上，原キーワードと訳語の多対多の対応関係により不適切な訳語が生成される可能性がある。この劣化の程度と実用上の影響の推定，および何らかの手段による回避策の検討が今後の課題である。

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機能として，検索結果一覧中の抄訳情報の重要性を指摘し，各文書中の主要キーワードの対訳を直接の検索対象とは別の対訳生成言語側のコンテンツに基づいて選択する手法について述べた。今後は提案手法の評価および，訳語を付与すべき重要語の決定方法や辞書の更新等の問題への取り組みを行なう予定である。

参考文献

[1] *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval Electronic Working Notes*, 1997.

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

[3] 鈴木 雅実, 井ノ上 直己, 橋本 和夫: "多言語情報検索における利用者支援について", 情報処理学会研究報告 97-NL-122, 1997.

[4] 井ノ上 直己, 鈴木 雅実, 橋本 和夫: "クロスリンガル情報発信のためのキャプション翻訳処理", 情報処理学会第56回全国大会, 1998.