

クロスリンガル情報検索結果の選択閲覧を支援する

4 L-6

翻訳情報提示方法の比較

鈴木 雅実 井ノ上 直己 橋本 和夫

KDD研究所

1はじめに

クロスリンガル情報検索において、検索結果一覧からの文書の取捨選択を容易にするための閲覧支援情報として、各文書中の主要キーワードを利用者の言語に翻訳して提供する枠組を考案した[2][3]。本稿では、原語でキーワードを提示した場合、提供する対訳情報の品質(生成方法)を変えた場合等の支援効果の比較について報告する。

2検索結果一覧からの選択閲覧支援

WWWコンテンツ等の情報検索においては、情報検索結果としてどのような文書集合を導くかという、検索性能自体の問題のほかに、その検索結果をどのように利用者に提示すれば実際の情報獲得に役立つかという視点が重要である。すなわち、個々の(多くの場合断片的な)検索要求に対して、所望の文書集合のみを検索結果として導くことはほとんど不可能であり、必然的に利用者自身が検索結果を眺めること(browsing)によって、閲覧すべき対象の取捨選択が行なわれることとなる。モノリンガルな情報検索結果の一覧表示では、コンテンツの一部のテキスト等の表示がしばしば用いられるが、クロスリンガル情報検索においては、検索対象言語に精通した利用者でない限り、コンテンツの一部が原語で表示されたとしても、文書の取捨選択に直接役立てることは相当に難しいか、多くの時間を要すると考えられる。従って、文書の選択閲覧を支援する情報は利用者の言語に翻訳されて提供されることが望ましい。

ここで問題は、どのような翻訳情報を提示することが、検索結果から迅速に閲覧対象を判断するために有効かという点である。表示対象としては、

- (1)HTML文書のタイトル(代表例はTITAN[1])
- (2)主要なキーワードのリスト(一例はCLINKS[3])
- (3)テキストの要約(クロスリンガルでの事例は不明)

等が考えられる。いずれも、情報量や翻訳の質が問題となるが、これらの提示内容/方法の比較については

未検証であると言える。Resnik[4]は、WWWコンテンツのうち1件が数行程度のイエローページ¹を対象として、名詞部分の辞書引き結果(可能性のある訳語を列記したもの)を提示した“gisting”を参照して、被験者が各イエローページ項目を予め指定された6種類のジャンルに分類するというタスクを実行した結果を示し、利用者の“decision making”に十分役立つと結論づけている。しかし、検索対象がより長いテキストである場合については、支援情報の種類/内容により、どの程度の効果があるかを調べる必要がある。

そこで、次章で述べる実験報告では、まず、筆者らの多言語検索サーバCLINKSにおける、出現頻度を反映した主要キーワードのリストを翻訳する枠組の有効性を検証するための評価を行ない、その結果に基づいて、さらに他の手法(要約テキストの提示等)との比較方法を検討して行くこととした。

3閲覧支援機能の評価実験

ここで述べる評価実験は、クロスリンガル情報検索の利用者が、検索結果の一覧表示の段階で、種々の形式で与えられる各文書の主要キーワード・リストを参照して、探索課題に該当する文書をどの程度正しく選択できるかを比較しようとするものである。比較条件は、キーワードの表示言語(英語/日本語)、翻訳(日本語)キーワードの場合の生成方法(訳語品質)の違い、キーワード表示個数、被験者の対象言語能力等である。以下、実験方法とその結果について順に記す。

3.1 実験内容

この実験では、被験者として日本人の大学生を主体とする19~32歳の男女64名の協力を得た。評価実験用に用意したテキスト・コーパスは、中国の経済分野の新聞記事(原文は中国語)で、これを英語および日本語に翻訳したものである²。記事数は224件で、これを用いて8組の探索課題と、それに対する仮想的な検索結果を作成した。各検索課題は、同コーパスの日本語版の記事タイトルから選んだ。すなわち、被験者に対して、「中国の小売業の現状」のような探索テーマを与

¹一種のリンク見出しのようなもので、企業名とその所在地。サービス内容を簡潔に記した内容

²英語版の1記事当たりの平均語数は380語程度である。

えるとともに、該当記事 1 件を含む 10 件の記事検索結果を用意した。

主要キーワードの提示条件

被験者が探索課題における該当文書の選択のために参照する、各文書の主要キーワードの提示条件は次の通りである。

条件 A：原語（英語）キーワード

条件 B：翻訳（日本語）キーワード

英語キーワード群に対して、各々の英日対訳辞書中の最初の訳語を当てたもの。

条件 C：翻訳（日本語）キーワード

英語キーワード群に対して、各々の英日対訳辞書中の訳語候補の中から、日本語コーパス（インターネット上から収集した WWW コンテンツ）内の共起分布を参照して、尤度の高い組み合わせを選択したもの。

また表示個数は、3 個、6 個、9 個、12 個のいずれかである。

3.2 実験結果

以下の記述において、「正解率」とは、各探索課題において被験者が選択した文書番号（最大 3 値まで）が、実際の探索テーマが示す記事と一致した度合を指すものとする。キーワードの表示内容の提示条件 A, B, C の各場合に、被験者全体の正解率として得られた値を図 1 に示す。ここで、各課題毎に正解文書の選択のし易さは異なることが予想されるが、図 1 では、正解率の数値に従って課題全体を 2 分した場合（各 4 課題）の条件別の正解率を棒グラフで示し、全体の平均値を折れ線で表示している。さらに、被験者が各探索課題への回答に費やした平均の時間について、条件 A, B, C を比較した結果を表 1 に示す。

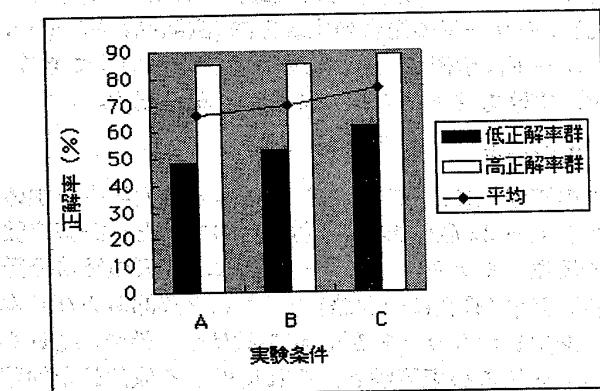


図 1 キーワード表示条件と正解率

主要キーワードの表示内容 / 個数による支援効果の相違

被験者にとっての外国語である英語で表示された場合に比して、被験者の母国語である日本語で表示された場合の方が、検索結果一覧の段階で目的に合致するコンテンツを選択することが容易であり、判断に要した時間も短縮された。また、同様に日本語で表示された

場合でも、英語のキーワードを基に辞書引きを行ない最初の訳語を用いる方法と、翻訳生成側の日本語コーパス内の訳語候補同士の共起分布を考慮して訳語を生成する方法とでは、後者の方が、コンテンツの選択をより容易にした。さらに、以上のような支援効果の程度は、コンテンツ選択の難度が高い場合により顕著であった。さらに、1 探索課題における回答までの平均所要時間は、キーワードが英語で表示される場合と、日本語で表示される場合とでは前者が後者の約 1.4 倍程度の長さとなっており、大きな差が見られた。一方、キーワードの表示個数と支援効果との間には、明確な相関関係は見られなかった。

表 1 探索課題の回答所要時間

	条件 A	条件 B	条件 C
所要時間	145 sec.	104 sec.	105 sec.

被験者の個人属性と正解率との関係

被験者に対して、実験の前に英語能力を判定するための筆記試験³を実施した。また、実験後に被験者のアンケートを取り、インターネット／情報検索等への親近性を聴取した。しかし、これらと正解率の関係データからは、（一面的な）英語能力、およびインターネット等の経験度の相違が、実験結果に影響を与える可能性についての確証は得られなかった。

5 おわりに

本稿では、クロスリンガル情報検索における検索結果の選択閲覧支援の重要性をふまえ、検索結果一覧表示の段階で表示する翻訳情報の提示が文書内容の推定にどの程度役立つかを、主要キーワードの表示条件の間で比較した実験結果を報告した。今後は、これに統いて、テキストの要約の翻訳を提示した場合との比較を行なうとともに、検索対象に応じた選択閲覧を効率良く支援するための方法をさらに追求する予定である。

参考文献

- [1] 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.
- [2] 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 - 主要キーワードの対訳付与に関する検討 -”, IPSJ-NL122-11, 1997.
- [3] 鈴木, 井ノ上, 橋本: “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”, 情報処理学会第 56 回全国大会, 4U-2, 1998.
- [4] P. Resnik: “Evaluating Multilingual Gisting of Web Pages”, AAAI Spring Symposium Cross-Language Text and Speech Retrieval Electronic Working Notes, 1997.

³英語能力に関してある一面を測定するもので、大規模な能力試験とは異なる。