言語横断検索システム Quest*

藤井 敦　石川 徹也
図書館情報大学
{fujii,ishikawa}@ulis.ac.jp

1 はじめに

言語横断情報検索 (Cross-Language Information Retrieval: CLIR) は、検索質問と異なる言語の文書を検索する処理であり、ACM SIGIR や ACL などの学会、TREC や NTCIR [9] などの情報検索コンテストにおいて近年主要なテーマのひとつである。CLIR は検索質問や対象文書を翻訳する点が単言語検索と異なる。本論文は、日英／英日 (双方向) CLIR システム「Quest」を提案し、要素技術の説明と評価実験の報告を行う。

Quest の特長は、検索質問 (キーワード) として専門用語を対象とする点にある。機械翻訳と同様に、CLIR においても、専門用語や固有名詞の翻訳は依然困難な課題である。Pirkola [16] は、一般辞書と領域固有の専門用語辞書を用いて、専門性の高い文書の検索性能を向上させた。しかし、専門用語を網羅的に辞書に記述することは困難であるため、(一般／領域固有を問わず) 対訳辞書に基づく翻訳だけでは本質的な解決にならない。専門用語翻訳の難しさを以下にまとめる。

(1) 専門用語の多くは複合語であり、既存の形態素 (語基) の組み合わせによって漸進的に作られる。
(2) 専門用語の語基は外来語をカタカナ表記したものが多く、単語の輸入に伴って漸進的に増加する。
(3) 専門用語はしばしば略記され (「CLIR」など)、略語が日本語として使われる。

(1) について、我々は既に統計的な複合語翻訳法を提案し、インターネット上の英日 CLIR に応用している [6]。(2) の解消法として、翻字 (transliteration) がある。日本語 (カタカナ語) から英語への翻字の研究例として、人間の知識や規則に依存した手法 [1, 24] や、音韻特性に基づく統計的手法 [10] がある。本論文では、カタカナとアルファベットの表層的な対応関係に基づく (音韻情報に依存しない) 統計的手法を提案し、Quest に応用する。(3) に対しては、略語がしばしば括弧表現を伴う点に着目して、コーパスから略語辞書を作成した。

検索システムの課題のひとつに、検索文書の効率的な提示 (閲覧支援) がある。特に、CLIR ではユーザが母国語以外の検索文書を (素早く) 理解できるとは限らない。そこで、文書要約とりわけユーザの関心に基づく要約 [11, 14, 19] が必要である。ただし本論文では、あらゆる種類の関心に柔軟に対応するのではなく、固定的な関心のみを対象とする。Quest システムのユーザが持つ主な関心として、「検索キーワードの意味 (定義) を知る」ことを想定した。そこで、Quest は検索文書から専門用語の定義を抽出し、ユーザに提示する機能を持つ。定義文の抽出には、百科事典から収集した定義表現パターンと、HTML 文書のタグ情報を利用する。

*Quest: A Cross-Language Information Retrieval System

2 Quest システム構成

既存の CLIR には以下の 3 通りの検索方式がある。

- 検索質問翻訳方式 [2, 3, 4, 6, 8, 15]
- 対象文書翻訳方式 [7, 13, 20]
- 中間言語表現方式 [3, 17, 18]

Quest は検索質問翻訳方式を採用しており、翻訳と検索モジュールは独立している。そこで、既存の検索エンジンを (用途に応じて) 適宜使い分けることができる。Quest のシステム構成を図 1 に示す。システムは以下の 3 つのモジュールからなる。

- 検索質問翻訳 (3 節): ユーザの検索質問に対して一つ以上の訳語を出力する。検索質問は専門用語「ターム i」の列であり、「訳語 i,j」はターム i の j 番目の訳語候補である。複数の訳語候補は尤度 (確率スコア) に基づいてソートされており、上位 k 個の訳語候補が実際の検索に用いられる (k はユーザが設定可能)。また、ユーザが検索に使用する訳語を選択することもできる。
- 文書検索: 翻訳された検索質問を用いて、データベースから文書を検索する。現在、インターネット上の検索エンジン「goo[1]」と「Altavista[2]」を (個別に) 利用している。
- 定義文抽出 (4 節): 検索文書を走査し、抽出規則を用いて各ターム i の定義文を抽出、ユーザに提示する[3]。より詳しい内容を知りたい場合に備えて、元文書に対するハイパーリンクも閲覧画面に並置する。尚、本機能はオプションであり、検索結果をそのまま閲覧することもできる。

[1] http://www.goo.ne.jp/
[2] http://www.altavista.com/
[3] 現在、日本語検索文書にのみ対応している。

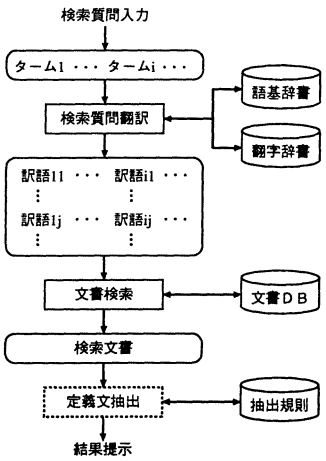


図 1: Quest システムの構成

3 検索質問翻訳

3.1 複合語翻訳

我々が提案した複合語翻訳法 [6] は、入力された複合語に対して、語基ごとの訳語候補を導出し、確率モデルを用いて訳語曖昧性を解消する。本論文ではさらに、辞書未登録の語基を翻字 (3.2 節参照) する機能を追加した[4]。日本語が原言語の場合は、可能な複合語分割パターンを全て考慮する。語基辞書は、EDR 日英専門用語対訳辞書 [22] の 2 語基複合語のうち、複合語分割法 [21] で高い確信度で分割できた 15,067 語から自動的に作成した (エントリ数は、日英辞書 6,097、英日辞書 4,531)。ここで、原言語の検索質問 S と訳語候補の一つ T を以下のように定義する (s_i と t_i は i 番目の語基) [5]。

$$S = s_1, s_2, \ldots, s_n; \quad T = t_1, t_2, \ldots, t_n$$

訳語選択は、$P(T|S)$ を最大化する T を選択することであり、ベイズの定理によって式 (1) で表現できる。

$$\arg\max_T P(T|S) = \arg\max_T P(S|T) \cdot P(T) \quad (1)$$

さらに、式 (2) の近似を用いる。

$$P(S|T) \cdot P(T) \approx \prod_{i=1}^{n} P(s_i|t_i) \cdot \prod_{i=1}^{n-1} P(t_{i+1}|t_i) \quad (2)$$

ここで、$P(s_i|t_i)$ は語基辞書における s_i と t_i の対応頻度に基づいて計算する[6]。$P(t_{i+1}|t_i)$ は、「NACSIS コレクション」[9] (65 学会の論文から収録した英語と日本語の抄録約 33 万件) から収集した語の共起情報 (bi-gram) を用いて計算した[7]。

[4] 日英翻訳の場合、翻字の対象はカタカナ列に限定される。

[5] 英語の「A of B」は、あらかじめ「B A」に変形する。

[6] s_i を翻字した場合は、$P(s_i|t_i)$ としてデフォルト値を与える。

[7] 英語抄録に対しては WordNet [12] を用いて stopword の削除と stemming を行い、日本語抄録に対しては形態素解析システム「茶筌」[25] を用いて内容語を抽出して、共起情報を収集した。

3.2 翻字処理

カタカナとアルファベット文字 (列) の対応関係を記述した辞書 (翻字辞書) と、文字列の共起情報が与えられれば、複合語翻訳法と同様の考え (3.1 節の式 (1) と (2)) に基づいて、日英／英日の統計的翻字ができる。

カタカナをローマ字表記したものは、元の英語綴りと多くのアルファベットを共有しやすい。例えば、「システム (si-su-te-mu)」と「system」からは、共通するアルファベットを基準として「シ-sy」「ス-s」「テ-te」「ム-m」のような文字列対応を抽出できる。しかし、中には「L」と「R」、「C」と「K」のように、日本語の発音では類似する組も存在する。そこで、アルファベット間の類似度 (同じなら 3、日本語の発音が同じなら 2、子音どうしなら 1、それ以外なら 0) を設定した[8]。文字列対応の特定は、図 2 のようなマトリクスから、類似度最大のパスを探索する問題に還元できる。図 2 は「テキスト (te-ki-su-to)」と「text」の対応の例であり、先頭の文字から終端記号 ($) までの矢印が類似度最大パスを示す。すなわち、「テ-te」「キス-x」「ト-t」のような文字列対応を抽出する[9]。語基辞書 (3.1 節) 中のカタカナ語と英訳 3,490 対を用い、翻字辞書と文字列共起情報を収集した。

また、不適切な単語の生成を避けるために、日本語／英語の「単言語」辞書の登録語のみを翻字出力候補とする。英語辞書には WordNet、日本語辞書には NACSIS コレクションから抽出したカタカナ語リストを用いた。

	テ te	キ ki	ス su	ト to	$
t	(3)	1	2	3	0
e	0	0	0	0	0
x	1	(2)	1	1	0
t	3	1	2	(3)	0
$	0	0	0	0	(3)

図 2: 日英文字列の類似度マトリクスの例

3.3 略語の処理

略語を使う場合、元の語形 (原形) と略語を括弧表現で並置することがある。そこで、コーパス中に「... information retrieval (IR) ...」のような表現があれば、括弧に先行する単語のうち「i」と「r」を頭文字とする単語列を原形として抽出する[10]。この方法を用いて NACSIS コーパスの英語抄録から略語–原形 7,307 対を収集し、略語辞書を作成した。現在、略語処理は複合語翻訳／翻字とは独立している。しかし、「IR システム」を「information retrieval system」と翻訳するような拡張も可能である。

[8] 日本語発音が同じアルファベット約 20 組を用意した。

[9] 類似度最大パスは、グラフ問題における探索アルゴリズム (Dijkstra のアルゴリズム [5] など) によって効率的に探索できる。

[10] 原形が括弧内に書かれている場合も同じアイデアで抽出する。

4 定義文抽出

Questでは、2つの異なる種類の定義文抽出規則を用いている。ひとつは言語的な定義表現(「○○とは△△である」など)であり、もうひとつはHTML文書のタグ情報に関する規則である。

定義表現が頻出する言語資源として辞典(事典)がある。我々が対象とするのは主に専門用語であるため、国語辞典ではなく百科事典を利用した。まず、「CD-ROM世界大百科事典」[23]の中から、EDR専門用語辞書の見出し語に関する解説のみを抽出し[11]、「茶筌」を用いて形態素解析を行い、文節に区切る[12]。次に、文節に関するbi-gramを作り、頻度が高く少なくとも片方の文節が見出し語を含むものを収集する。このとき、異なる見出し語は全て共通なシンボルに変換しておく。また、同一文中に現れる文節は、(隣接していなくても)共起するものと見なした。最後に、収集した文節bi-gramを人手で確認あるいは修正して、定義表現パターンを作成した。

言語表現以外の情報として、HTMLのハイパーリンクを用いた。HTML文書には、文中の用語をリンク先の文書で解説しているものがある。そこで、例えば「情報検索」というキーワードで検索された文書中に「<A HREF=...>情報検索</A>」という記述があれば、そのリンク先の文書も提示の対象とする。

5 Questシステムの評価と考察

5.1 検索質問翻訳の評価

EDR専門用語辞書中のエントリを検索質問として、日英/英日CLIRの評価実験を行った。検索質問として、語基辞書作成に用いなかった2語基複合語の中からランダムに1,000語を選んだ。検索エンジンにはgooを用い、日英検索は「海外サイト」、英日検索は「日本語サイト」を検索した。正しい訳語で検索された文書を正解と見なし、適合率/再現率によって翻訳法を比較評価した(ただし、CLIRではノイズを含まない少数文書を提示することが好ましいので、我々は再現率よりも適合率を重視する)。比較した翻訳法は、(1)訳語の曖昧性解消なし、(2)検索文書数の多い順に訳語をソートする手法、(3)確率的曖昧性解消(3.1節)である。さらに手法(2)、(3)に翻字処理を追加したものを手法(4)、(5)とする。日英、英日の実験結果を表1と表2にそれぞれ示す[13]。kは検索に利用した訳語候補数である。概して、日英CLIRにおいて、確率的曖昧性解消と翻字の効果が顕著であった。翻訳結果を分析した結果、日本語の異表記に関する誤りが多かった。実験データに含まれた異表記の例を以下に示す。

「データ/データー」、「データベース/データ・ベース」、「ソフトウェア/ソフトウエア」、「インターフェース/インターフェイス/インタフェース」、「割り込み/割込み」

表1: 日英CLIRの適合率/再現率の比較

手法	適合率/再現率 (%)			
	$k=1$	$k=2$	$k=3$	$k=4$
(1)	42.6/36.5			
(2)	46.2/36.0	43.2/36.3	42.9/36.4	42.9/36.5
(3)	70.9/22.1	70.3/36.4	65.1/36.4	64.6/36.5
(4)	23.0/43.7	19.8/45.4	18.8/45.5	21.6/54.8
(5)	77.9/40.2	62.1/54.6	39.6/54.7	36.3/54.8

表2: 英日CLIRの適合率/再現率の比較

手法	適合率/再現率 (%)			
	$k=1$	$k=2$	$k=3$	$k=4$
(1)	10.8/28.5			
(2)	16.5/21.9	12.9/24.0	12.5/26.1	12.0/27.1
(3)	47.9/20.8	35.7/25.5	26.8/26.5	24.2/26.7
(4)	16.6/22.1	13.0/24.2	12.6/26.4	12.1/27.3
(5)	47.6/21.0	35.5/25.6	26.8/26.6	24.2/26.9

5.2 定義文抽出に関する考察

図3は、検索キーワード「データマイニング」に対する定義文の提示画面である。タイトル(「DW-Gaiyou」など)から元文書へのリンクが張られており、抽出された定義文がその下に提示されている。

少数のキーワードを用いた予備実験によって得られた知見、今後の研究課題を以下に示す。

- 抽出された文は概して、直接的な定義や、元文書が定義を含んでいると予測できるものであった。
- 一つのキーワードに対して、(文書作成者の観点の違いによる)複数の異なる定義文が抽出される。
- 複数行の文章や数式を伴う定義に対しては、文よりも大きな単位の抽出が必要である。今後は、指示語(「このような」など)がある場合に直前の文も提示する方法や、文書構造に関するタグ情報(「パラグラフ<P>」など)を利用する手法が必要である。
- 定義文抽出の時間効率を改善する必要がある。

6 おわりに

本論文は、専門用語キーワードを対象とした日英/英日言語横断検索システム「Quest」を提案し、検索質問翻訳の評価と定義文抽出に関する考察を行った。定義文抽出の定量的評価は今後の研究課題である。

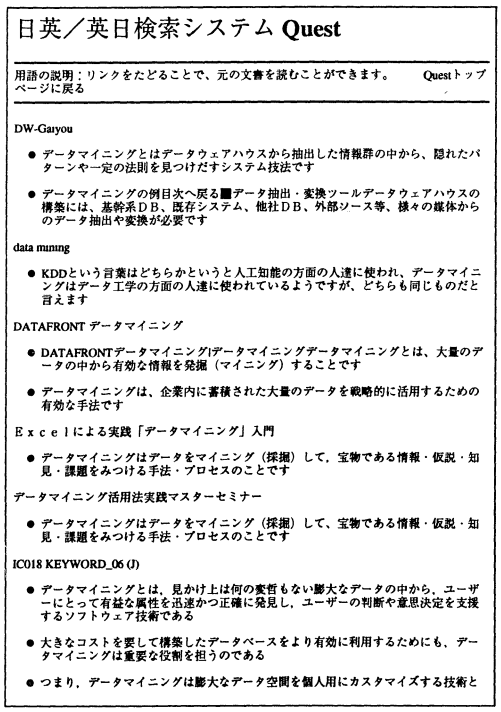
謝辞

「CD-ROM世界大百科事典」は日立デジタル平凡社の許諾を、「NACSISコレクション」は学術情報センターの許諾を得て使用させて頂きました。

[11] 人名や地名などの解説には、専門用語とは異なる言語表現が用いられることが多い。

[12] 茶筌の品詞情報と文節に関する規則を人手で作成して利用した。

[13] 以前の我々の実験[6]では、EDR辞書の2語基複合語ほぼ全て(約5万語)を用いて語基辞書を作成した。それに対して、今回は確信を持って分割できた複合語のみを用いて語基辞書を作成した。そのため語基辞書のカバレッジが小さく、全体的に検索の再現率が低い。


日英／英日検索システム **Quest**

用語の説明；リンクをたどることで、元の文書を読むことができます。　Questトップページに戻る

DW-Gaiyou

- データマイニングとはデータウェアハウスから抽出した情報群の中から、隠れたパターンや一定の法則を見つけだすシステム技法です
- データマイニングの例目次へ戻る■データ抽出・変換ツールデータウェアハウスの構築には、基幹系ＤＢ、既存システム、他社ＤＢ、外部ソース等、様々の媒体からのデータ抽出や変換が必要です

data mining

- KDDという言葉はどちらかというと人工知能の方面の人達に使われ、データマイニングはデータ工学の方面の人達に使われているようですが、どちらも同じものだと言えます

DATAFRONT データマイニング

- DATAFRONTデータマイニング|データマイニングデータマイニングとは、大量のデータの中から有効な情報を発掘（マイニング）することです
- データマイニングは、企業内に蓄積された大量のデータを戦略的に活用するための有効な手法です

Ｅｘｃｅｌによる実践「データマイニング」入門

- データマイニングはデータをマイニング（採掘）して，宝物である情報・仮説・知見・課題をみつける手法・プロセスのことです

データマイニング活用法実践マスターセミナー

- データマイニングはデータをマイニング（採掘）して、宝物である情報・仮説・知見・課題をみつける手法・プロセスのことです

IC018 KEYWORD_06 (J)

- データマイニングとは，見かけ上は何の変哲もない膨大なデータの中から，ユーザーにとって有益な属性を迅速かつ正確に発見し，ユーザーの判断や意思決定を支援するソフトウェア技術である
- 大きなコストを要して構築したデータベースをより有効に利用するためにも、データマイニングは重要な役割を担うのである
- つまり，データマイニングは膨大なデータ空間を個人用にカスタマイズする技術と

図 3:「データマイニング」の定義文抽出例

参考文献

[1] Chinatsu Aone, Nicholas Charocopos, and James Gorlinsky. An intelligent multilingual information browsing and retrieval system using information extraction. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 332–339, 1997.

[2] Lisa Ballesteros and W. Bruce Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 64–71, 1998.

[3] Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th International Joint Conference on Artficial Intelligence*, pp. 708–714, 1997.

[4] Mark W. Davis and William C. Ogden. QUILT: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 92–98, 1997.

[5] Edsgar W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, Vol. 1, pp. 269–271, 1959.

[6] Atsushi Fujii and Tetsuya Ishikawa. Cross-language information retrieval using compound word translation. In *Proceedings of the 18th International Conference on Computer Processing of Oriental Languages*, 1999. (To appear).

[7] Denis A. Gachot, Elke Lange, and Jin Yang. The SYSTRAN NLP browser: An application of machine translation technology in multilingual information retrieval. In *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996.

[8] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–57, 1996.

[9] Noriko Kando, Teruo Koyama, Keizo Oyama, Kyo Kageura, Masaharu Yoshioka, Toshihiko Nozue, Atsushi Matsumura, and Kazuko Kuriyama. NTCIR: NACSIS test collection project. In *the 20th Annual BCS-IRSG Colloquium on Information Retrieval Research*, 1998.

[10] Kevin Knight and Jonathan Graehl. Machine transliteration. *Computational Linguistics*, Vol. 24, No. 4, pp. 599–612, 1998.

[11] Inderjeet Mani and Eric Bloedorn. Machine learning of generic and user-focused summarization. In *Proceedings of AAAI/IAAI-98*, pp. 821–826, 1998.

[12] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller, and Randee Tengi. Five papers on WordNet. Technical Report CLS-Rep-43, Cognitive Science Laboratory, Princeton University, 1993.

[13] Douglas W. Oard and Paul Hackett. Document translation for cross-language text retrieval at the University of Maryland. In *The 6th Text Retrieval Evaluation Conference (TREC-6)*, 1997.

[14] Ryo Ochitani, Yoshio Nakao, and Fumihito Nishino. Goal-directed approach for text summarization. In *Proceedings of the ACL-EACL Workshop on Intelligent Scalable Text Summarization.*

[15] Akitoshi Okumura, Kai Ishikawa, and Kenji Satoh. Translingual information retrieval by a bilingual dictionary and comparable corpus. In *LREC workshop on translingual information management: current levels and future abilities*, 1998.

[16] Ari Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 55–63, 1998.

[17] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, Vol. 21, No. 3, pp. 187–194, 1970.

[18] Páraic Sheridan and Jean Paul Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 58–65, 1996.

[19] Anastasios Tombros and Mark Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2–10, 1998.

[20] 酒井哲也, 梶浦正浩, 住田一男. Cross-language 情報検索のための BMIR-J2 を用いた一考察. 情報処理学会 自然言語処理研究会, Vol. 99, No. 2, pp. 41–48, 1999.

[21] 藤井敦, 石川徹也. 日本語複合語の自動分割と日英語基対訳辞書の作成. 情報処理学会 自然言語処理研究会, Vol. 98, No. 99, pp. 67–72, 1998.

[22] 日本電子化辞書研究所. 専門用語辞書 (情報処理), 1995.

[23] 日立デジタル平凡社. CD-ROM 世界大百科事典プロフェッショナル版, 1998.

[24] 熊野明. カタカナ表記からの英訳推定による専門用語辞書作成. 言語処理学会第 1 回年次大会発表論文集, pp. 221–224, 1995.

[25] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明. 日本語形態素解析システム『茶筌』version 1.0 使用説明書. Technical Report NAIST-IS-TR97007, 奈良先端科学技術大学院大学, 1997.