語彙拡張を利用した言語横断情報検索に関する研究

阿玉 泰宗[†] 橋本 泰一[†] 徳永 健伸[†] 田中 穂積[†]

[†] 東京工業大学大学院 情報理工学研究科 計算工学専攻

{adama, taiichi, take, tanaka}@cl.cs.titech.ac.jp

1 はじめに

近年、WWW に代表される電子文書は急速に増大している。しかし、ユーザの要求に適合する文書が母国語で記述されているとは限らない。そのため、母国語の検索要求によって他の言語の文書を検索する「言語横断情報検索 (CLIR)」への要求が高まっている。CLIR において多く用いられる手法は対訳辞書を利用して検索要求を翻訳する手法であるが、この手法には辞書の語彙不足や翻訳曖昧性の問題が伴う。語彙不足の問題を解決するため、藤井らは、専門語辞書の句の対訳関係からその要素語の対訳関係を学習して「語基辞書」を作成する手法を提案した [2]。また、未知語の多くを占めるカタカナ語の翻訳手法として、発音情報を利用する「翻字」があり、Brill らは、文脈を考慮した翻字のモデルを提案した [3]。翻訳曖昧性の解消には、相互情報量を用いた前田らのように大域的な共起情報を用いる手法 [5] と、Bigram を用いた藤井らのように局所的な共起情報を用いる手法 [2] がある。

しかし、上述の手法にはいくつかの問題点が考えられる。語基辞書の作成は、見出し語のみに注目しており、対訳の情報は用いていない。また、日本語と英語などの言語間での翻字には、入力の形態素数と出力の単語数の不一致の問題がある。曖昧性解消の手法では、大域的な共起では計算量が多くなり、局所的な共起では使用できる文脈が制限されるという問題がある。

本研究では、日英の CLIR を対象とし、語彙不足の解決手法として、英単語との共起情報を用いた語基辞書の作成手法、形態素数で場合分けした翻字のモデルを提案する。また、曖昧性解消のため、Bigram と相互情報量を併用する手法を提案し、NTCIR コレクションを用いて精度の評価を行なう。

2 システムの概要

本研究で構築したシステムを、図 1 に示す。処理は以下のように進行する。

1. ユーザがシステムに日本語クエリ Q_j を与える
2. 翻訳システムが Q_j 中の語 q_i を辞書引きし、候補語集合 E_i を得る
3. 辞書引きに失敗したカタカナ語を翻字する
4. 任意の候補語の組み合わせについて、共起情報から語の関連性を計算し、曖昧性を解消する
5. 曖昧性解消によって得られた英語クエリ Q_e を用いて検索を行なう
6. 検索結果の上位をユーザに提示する

3 語基辞書の学習

藤井らは、見出し語の文字種の異なりに着目し、下の 2 種類の語基辞書の作成手法を提案している [1]。

- 全ての文字種の異なりを利用 (heuristic)
- 十分に信頼できる文字種の異なりから、基本の対応を得る (algorithm)

これらの手法では、対訳の英単語の情報は語基数の特定にのみ用いられるが、本研究では、「翻訳辞書において英単語と共起しやすい文字列は、その英単語と関連性が強い、すなわち英単語の対訳である」という仮説に基づき、対訳の単語との共起確率を用いて語基辞書を作成する。

3.1 共起確率の学習

共起確率は、EDR 専門語和英辞書の 118,236 見出しから学習する。見出し語に n 個の対訳 $E_1, E_2, ..., E_n$ が与えられており、E_i が m 単語からなるとき、見出し語から得られる形態素列の集合 $J = \{J_1, J_2, ..., J_k\}$ 中の要素と E_i 中の単語の共起頻度として $\frac{1}{k \cdot m \cdot n}$ を加える。見出し語「16 進数 (hexadecimal numeral//hexadecimal number)」の場合、対訳数 n は 2、いずれの対訳も 2 単語からなる ($m = 2$)。得られる形態素列は **{16, 進, 数, 16 進, 進数, 16 進数 }**($k = 6$) なので、$J_i \in J$ に対して、$freq(J_i, hexadecimal) = 2/24$、$freq(J_i, number) = 1/24$ のように共起頻度が得られる。

全ての見出し語から共起頻度を学習した後、全ての英単語 E と、それと共起する形態素列 J について、$P(J|E) = \frac{(freq(J,E)}{(freq(*,E)}$ を計算する。

見出し語の形態素解析には、隠れマルコフモデルを利用した [6]。解析に用いた品詞の連接確率と単語の生起確率は、NTCIR-2 の日本語コーパスを ChaSen[7] で解析した結果から学習した。

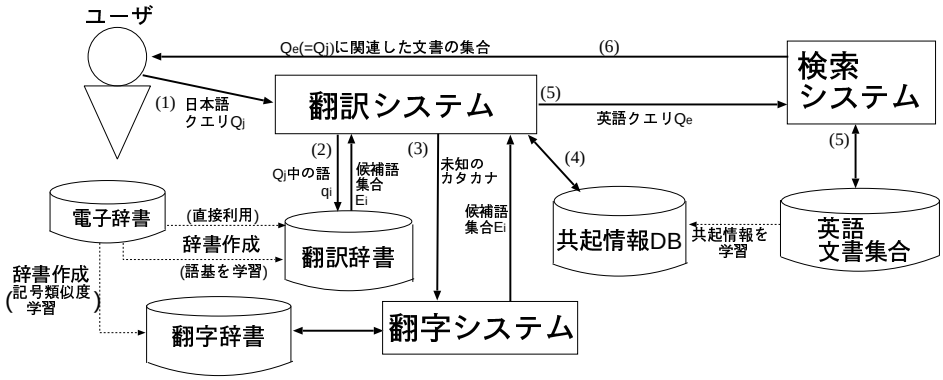

図 1: CLIR システムの概要

3.2 語基の学習

語基の学習には、EDR 専門語和英辞書 [4] 中の、対訳 2 単語の見出し語 57,023 語を用いる。まず、EDR 英和辞書を利用して対応付けを行なう。対訳の英単語を辞書引きし、得られた訳を組み合わせて日本語の見出し語が得られるならば、辞書から得られた対応は正しい。この処理で対応付けに失敗した 27,713 見出しについて、共起確率を用いた対応付けを行なう。

対訳が e_0, e_1 であるとき、見出し語を j_0, j_1 に分割するスコア sc を以下のように定義し、sc を最大化する対応を語基の対訳関係とする。

$$sc = max\big((P(j_0|e_0)\cdot P(j_1|e_1)), (P(j_1|e_0)\cdot P(j_0|e_1))\big) \quad (1)$$

3.3 評価実験

EDR 英和辞書で対応付けできなかった見出し語からランダムに抽出した 1,841 見出し語を語基に分割した結果は、表 1 の通りとなった。なお、以降では提案手法を “probabilitic” 手法と呼ぶ。

表 1: 評価データの対応付け精度

手法	正解数	精度
heuristic	1,320	71.7%
algorithm	1,647	89.5%
probabilistic	1,775	96.4%

学習対象の 57,023 見出しから得られた語基辞書の統計情報は、表 2 のようになった。

4 翻字

4.1 先行研究

Brill らは、辞書に含まれない文字列 s から辞書に含まれる単語 w への Spelling Correction を、

$$\arg\max_w P(w|s) = \arg\max_w P(s|w)\cdot P(w)$$

表 2: 得られた語基辞書の比較

	heuristic	algorithm	probabilistic
対応の種類	35,725	25,452	23,536
辞書に含まれる	6,121 (17.1%)	6,156 (24.2%)	6,534 (27.8%)
辞書に含まれない対応			
見出し語含まれる	9,980 (27.9%)	9,729 (38.2%)	8,802 (37.4%)
見出し語含まない	19,624 (54.9%)	9,567 (37.6%)	8,200 (34.9%)

$$P(s|w) = max_{R\in Part(w), T\in Part(s)} \prod_{i=1}^{|R|} P(T_i|R_i) \quad (2)$$

で計算しており、式 (2) を用いた翻字が有効に機能することを報告している [3]。なお、$Part(w)$ は文字列 w のあらゆる可能な分割の集合である。

彼らは確率 $P(T_i|R_i)$ の学習に重み無しの編集距離を用いている。また、部分列の単語中での位置情報も用いている。しかし、翻字では Spelling Correction より置換すべき文字列が多いので、重み無しの編集距離では、同一コストの対応づけが多数得られる。また、一形態素が一英単語を出力するという関係が常には成り立たないので、位置情報が利用できない場合がある。

4.2 提案手法

第一の問題の解決のため、以下の手順で反復学習を行なう。使用したデータは EDR 和英、専門語和英辞書に含まれるカタカナ語のうち、一単語の対訳をただ一つ持った 14,022 見出し (ローマ字に変換) である。

1. ローマ字 r と英文字 e の間の操作 $r \to e(r \neq e)$ の重みを 1、$r \to e(r = e)$ の重みを 0 に初期化
2. 全ての見出し語と訳語の対で、編集距離を最小化する操作列を求める

3. 全ての対で得られた操作列が直前のサイクルと一致した場合、終了
4. 任意の r, e について $P(e|r)$ を学習し、操作 $r \rightarrow e$ の重みを $1 - P(e|r)$ に更新し、2. に戻る

得られた対応付けに、前後 2 つ以内の対応付けを結合して拡張し、文脈を考慮したモデルを学習する。

また、第二の問題の解決には、語基辞書の学習と同様の形態素解析により、カタカナ語 K を $s = 1$〜5 形態素に分割した結果 K_s を得、一形態素が一単語を出力するようにモデルを変更した。この際、個々の形態素から、式 (2) を大きくする上位 10 件の候補語を出力し、Bigram によってリランキングする。カタカナ語列 $K = k_1, k_2, ..., k_s$ が英単語列 $E = e_1, e_2, ..., e_s$ に翻字される確率は、式 (3) で計算される。

$$P(E|K) = \prod_{i=1}^{s} P(k_i|e_i) \cdot P(e_i|e_{i-1}) \qquad (3)$$

ただし、$P(e_1|e_0) = P(e_1)$ とする。

4.3 評価実験

反復学習によって得られた確率を用い、翻字の実験を行なった。候補語としては、NTCIR-1,NTCIR-2 の英語文書に出現する語を用い、Bigram と Unigram も同文書から学習した。結果を表 3 に示す。

表 3: Transliteration の実験結果

データセット		jquery	oneword	multiword
w/o seg	Top 1	79.5%	65.6%	54.7%
	Top 10	92.4%	88.0%	72.0%
unigram	Top 1	83.2%	67.7%	62.1%
	Top 10	96.2%	89.0%	90.1%
bigram	Top 1	89.7%	67.7%	68,2%
	Top 10	97.3%	88.5%	92.1%

w/o seg 従来の、カタカナ語から複数単語を出力する手法
unigram 形態素解析を行ない、リランキングを行わない
bigram 形態素解析を行ない、リランキングを行う
jquery NTCIR-2 の検索課題に出現するカタカナ語 185 語
oneword EDR 辞書に含まれる、一単語の訳語を複数持つ 192 語
multiword EDR 辞書に含まれる、複数単語の訳語を持つ 214 語

5 曖昧性解消

5.1 Bigram を用いた曖昧性解消

藤井らは、日本語の専門分野での複合語の多くが、その要素語の訳を並べることで翻訳可能であることを指摘し、語基辞書を用いた CLIR の曖昧性解消に、Bigram を利用している [2]。日本語の複合語 $S = s_1, s_2, ..., s_n$ が英語の複合語 $T = t_1, t_2, ..., t_n$ に翻訳される確率は、Bigram を用いて以下のように計算できる。

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

$$P(S|T) = \prod_{i=1}^{n} P(s_i|t_i) \quad P(T) = \prod_{i=1}^{n} P(t_i|t_{i-1}) \quad (4)$$

ここで、$P(s_i|t_i)$ は辞書の対訳関係における共起確率を用いる。しかし、この手法では直前の単語以外の文脈は考慮しておらず、特に短い句が多く現れるクエリでは適切な訳が得られない可能性がある。そこで本研究では、式 (4) を用いて句の翻訳候補の絞り込みを行なった後、句の翻訳候補の間で相互情報量を求めて曖昧性解消を行なう。

5.2 Bigram と相互情報量の併用

日本語クエリから得た句の集合 $J = j_1, j_2, ..., j_m$ を英語句の集合 $E = e_1, e_2, ..., e_m$ に翻訳するとき、相互情報量を用いたスコア $sc(J, E)$ を式 (5) で定義する。

$$sc(J, E) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} MI(e_i, e_j) \qquad (5)$$

また、式 (5) に式 (4) で求めた $P(e_i|j_i)$ を加味したスコア $scw(J, E)$ を、式 (6) で定義する。

$$scw(J, E) = 2^{sc(J,E)} \cdot \prod_{i=1}^{m} P(e_i|j_i) \qquad (6)$$

5.3 評価実験

評価実験として、NTCIR-1 のクエリ (ID=0001-0030) と NTCIR-2 のクエリ (ID=0101-0149) を用いた CLIR の実験を行なった。クエリを ChaSen で形態素解析し、最長一致を用いて語基数を最小化し、語基ごとに翻訳候補語の取得を行なう。この際には、語基辞書→一般語辞書→翻字の順の逐次検索を行なった。類似度は TF·IDF で計算し、上位 1,000 件を出力した。翻訳曖昧性の解消には、

- 句の内部の Bigram を最大化する候補を選択 (big_ph)
- クエリ全体を一つの句とみなし、全体の Bigram を最大化する候補を選択 (big)
- 式 (5) を最大化する候補を選択 (flat)
- 式 (6) を最大化する候補を選択 (weighted)

の 4 種を用い、比較を行なった。語基辞書と翻字による 11 点再現率・精度の相違を図 2 に、再現率と R 精度の相違を表 4 に示す。 また、probabilistic の語基辞書を用い、翻字を行なった場合の、曖昧性解消手法による 11 点再現率・精度の相違を図 3 に、再現率と R 精度の相違を表 5 に示す。

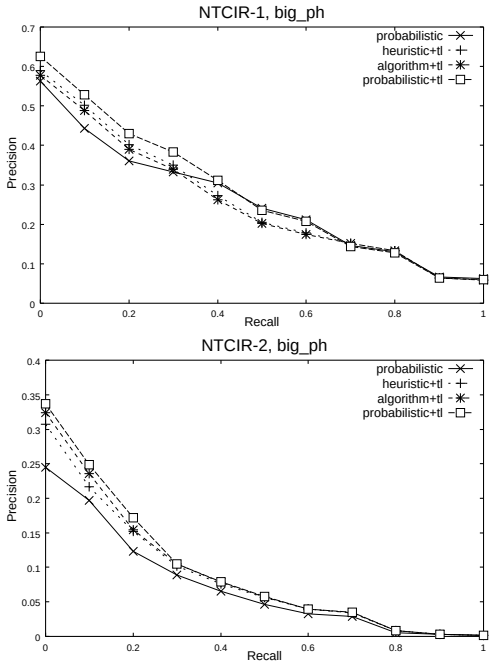

図 2: 語基辞書、翻字による 11 点再現率・精度の相違

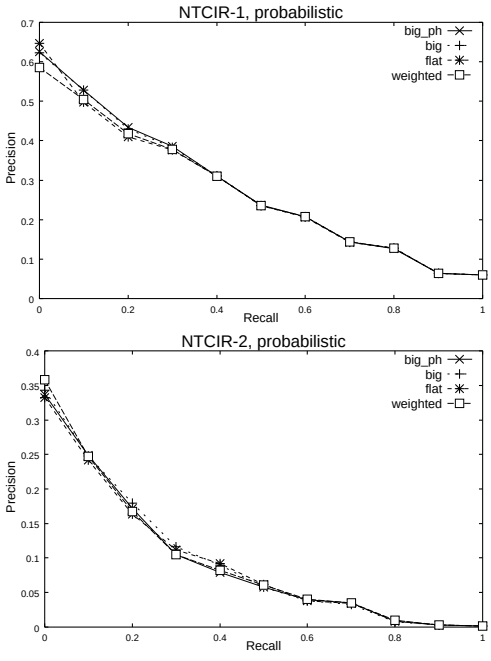

図 3: 曖昧性解消手法による 11 点再現率・精度の相違

表 4: 語基辞書、翻字による精度の相違

手法	NTCIR-1		NTCIR-2	
	再現率	R 精度	再現率	R 精度
prob	0.3183	0.2588	0.3065	0.0927
heu+tl	**0.4687**	0.2776	**0.3837**	0.1208
algo+tl	0.4681	0.2661	0.3830	**0.1233**
prob+tl	0.4533	**0.2943**	**0.3837**	**0.1233**

表 5: 曖昧性解消手法による精度の相違

手法	NTCIR-1		NTCIR-2	
	再現率	R 精度	再現率	R 精度
big_ph	**0.4533**	0.2943	0.3837	0.1233
big	0.4453	**0.2952**	**0.4362**	**0.1310**
flat	0.4510	0.2839	0.3766	0.1173
weighted	0.4459	0.2855	0.3908	0.1196

6 結論

語基辞書の学習、翻字ともに従来手法より高い精度を確認でき、Bigram を用いた曖昧性解消においては CLIR の精度を向上させることが確認できた。しかし、相互情報量を用いた曖昧性解消は精度の向上には繋がらなかった。この理由の大きなものは、句単位での共起を学習するにはコーパスが十分でなかったことが考えられる。また、単語間の距離などを考慮して単純な同一文書中での共起より精密な関連性の推定が出来れば精度の向上に繋がると予想される。加えて、本研究では Word-by-Word の翻訳がなされているが、句の翻訳の利用法も今後の課題として残る。

謝辞

NTCIR コレクションは国立情報学研究所 (NII) の許可を得て使用させて頂きました。この場を借りて深謝いたします。

参考文献

[1] Atsushi Fujii and Tetsuya Ishikawa. Cross-Lauguage Information Retrieval ad ULIS. In *Proceedings of th 1st NTCIR Workshop*, 1999.

[2] Atsushi Fujii and Tetsuya Ishikawa. Japanese/English Cross-language Information Retrieval: Exploration of Query Translation and Transliteration. *Computers and the Humanities*, Vol. Vol.35, No. No.4, pp. 389–420, Nov 2001.

[3] Eric Brill Gary. Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs. In *Proceedings of the 6th NLPRS*, pp. 393–399, Nov 2001.

[4] Communications Research Laboratory. EDR 電子化辞書仕様説明書.

[5] Akira Maeda, Fatiha Sadat, Masatoshi Yoshikawa, and Shunsuke Uemura. Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine. In *Proceedings of IRAL2000*, pp. 25–32, Sep 2000.

[6] 北研二. 言語と計算–4 確率的言語モデル, 第 7.1 章. 東京大学出版会, 1999.

[7] 松本裕次, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 形態素解析システム『茶筌』 Version 2.3.3 使用説明書.