社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－NL－165 (12)
2005/1/12

ウェブ文書資源からの中日対訳推定における文脈窓幅の役割

植野 研　　出羽 達也　　熊野 明

(株)東芝 研究開発センター 知識メディアラボラトリー 〒212-8582 川崎市幸区小向東芝町 1
E-mail:　{ken.ueno, tatsuya.izuha, akira.kumano}@toshiba.co.jp

概要

本研究では，文脈語による言語横断ウェブ検索ならびに中日翻訳エンジンを利用した対訳知識獲得システム **KATrans**（***K**nowledge **A**cquisition on **Trans**lation*）を提案し，文脈窓幅と訳語推定精度の関係性を明らかにする．実験では，中国語の計算機専門用語において，文脈窓幅を変化させたときの日本語の対訳推定精度の変化を調べた. 実験では, 中国語の計算機専門用語 10 語において, 文脈窓幅を 5 語, 21 語，35 語に変化させたときの日本語の対訳推定精度(log*WMRR*)の変化を調べた．実験結果から，見出し語がウェブ上でどのように使われているかの特性により，最適な文脈窓幅が異なることが明らかになった．また，文脈窓幅は文書検索ベクトルの構成要素に大きく影響することが分かった.

キーワード　対訳知識獲得，対訳推定，言語横断検索，文脈，中日翻訳

The Role of Controling Context Window Size on Estimation of Chinese-Japanese Translation on the Web

Ken UENO　Tatsuya IZUHA　and　Akira KUMANO

Knowledge Media Laboratory, Corporate Research and Development Center, TOSHIBA CORPORATION
1　Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582　Japan
E-mail:　{ ken.ueno, tatsuya.izuha, akira.kumano }@toshiba.co.jp

Abstract

In Chinese-Japanese estimation using CLIR (Cross-Language Information Retrieval) on the Web with context vectors, we investigated the relationship between context window size and estimation accuracy.　In this research we propose the bilingual knowledge acquisition system **KATrans** (***K**nowledge **A**cquisition on **Trans**lation*) which collects a pair of non-parallel web documents using Chinese-Japanese machine translation engine and generating search word vectors.　The estimation accuracy (log*WMRR*) is calculated on ten Chinese Computer terms, changing the context window size in 5 words, 21 words, and 35 words.　Our experimental result shows that the best context window size is probably determined by the characteristics how the terms are used in the web documents.　Furthermore the best context wondow size depends on the Japanese search word vector automatically generated by the KATrans.

Keyword　Knowledge Acquisition, Target Word Estimation, Cross Lingual Information Retrieval, Context, Chinese-Japanese Translation

1. 研究の背景

機械翻訳システムの精度は辞書知識の質と量に大きく依存する．しかし，従来のような人手作業中心のやり方では，辞書知識の開発に多大な時間と労力がかかる．次々と生まれる新語に対する訳語を追加するにも時間と労力がかかることも予想される．このような背景から，パラレル文書から統計的に辞書を構築する方法が数多く使われてきた.

しかしながら，パラレル文書が希少な言語対の場合には，この種の手法は利用することが困難であるため，さまざまな工夫が必要となる．たとえば，中日翻訳辞書構築においては，既存の中英辞書と英日辞書の訳語を２段階に連結させる方法などが考えられる．しかしながら，連結させる言語方向によっては，多義性の問題が発生し，適切な訳語を絞り込むことが困難な場合がある.

このような限界から，中日翻訳知識の専門家が訳語をつけていく方法，中英日の訳語連結方式に加えて，中英辞書知識と現状の中日翻訳エンジンを用いて新たな対訳知識を機械

的に求める方法を提案する．本方法により，希少言語対における対訳辞書構築の時間的コストや労力の減少が期待できる．

2. 目的

本研究では，文脈語に着目し，ウェブ文書から対訳を推定する対訳発見方法を提案し，語の文脈幅が対訳推定精度に与えている影響を考察する．本方法では，文脈語の文脈幅が対訳推定精度に依存することが予備実験により明らかとなっていた．見出し語の種別に応じた最適な語の文脈幅を決める手がかりが得られれば，本手法の有効性を十分に引き出すことにつながると考えられる．

3. 従来法のまとめ

コーパスから対訳を推定する研究は 1990 年代に入って盛んになっている． [熊野 94, Izuha01]や[Utsuro02,Hasan01]などの研究では，パラレル文書から文章の塊をユニットとして断片化し，統計処理で対訳知識を抽出する（**図１**）．とくに[Utsuro02]の方法は効果的に言語横断検索（CLIR）を利用しており，対訳発見手法として強力な方法である．しかしながらこれらの方法は精度が高いがニュースサイト等に知識源が限定されるため，本タスクに直接応用できない．

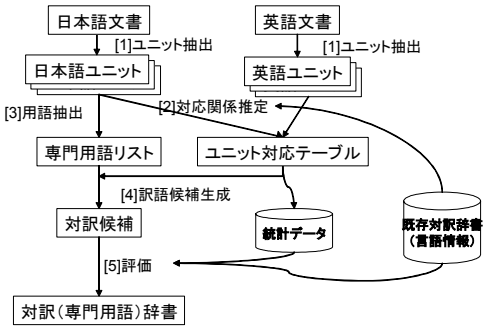

図１ ユニット統計処理を用いる方法

文脈ベクタの類似度を用いる[Rapp95,99]や[Fung98, 00]らの方法は，ノンパラレルコーパスから対訳知識を抽出できるため汎用である(**図２**)．しかしながら，コーパス中に出現するすべての語について総当りで類似度を求めるため，多義語やノイズの影響を受けやすいと考えられる．

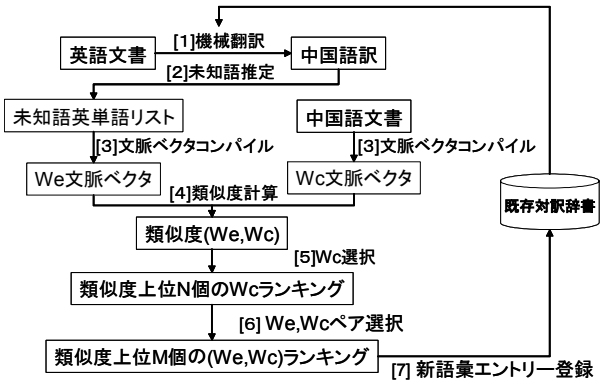

図２ 文脈ベクトル対を用いる方法

[佐藤 02]らの方法は，ウェブから専門用語の関連語を収集するのに強力な方法であるが，直接的に対訳を推定する手法ではない点，システムへのパラメータが多く，経験に基づく調整が必要である点から，そのまま本タスクに適用させるのは困難であると考えられる．

4. KATrans 対訳知識獲得システム

以上に述べた，従来法における本タスクへの適用困難性を鑑み，ここでは，サーチエンジンと中日翻訳エンジンを併用する方法を考案する．ここでは，ウェブなどのオープンな環境においても汎用な対訳知識が抽出可能な対訳知識獲得システム **KATrans** (***K**nowledge **A**cquisition on **Trans**lation*)を提案する．これは，[Fung00]の方式を改良し，言語横断検索を利用することで中国語周辺の文脈語から動的に日本語関連文書を収集し，これらの文脈ベクトル対から対訳を推定する方法である．従来の方法では，ソース言語とターゲット言語から独立にそれぞれ文脈ベクタを生成し，類似度をとることで対訳知識を抽出していた．しかしながら，この場合，文脈語の種類によっては，類似度が低いターゲット言語文脈ベクトルが収集されてしまう場合がある．そこで，KATrans では，[Utsuro02]に基づき，未知語を出発点として，ノンパラレルな文書から関連度の高い文書をまずウェブから選択し，これらの文書から対訳知識を抽出する．これにより見出し語に強く関連した密度の濃い言語資源から対訳知識が抽出できると考えられる．

以下，KATrans の処理過程の概要を述べる（**図３**）．ウェブ資源からの中日対訳知識獲得では日本語訳が不明な中国語の見出し語を入力とし，日本語の訳語を推定する．実際には

訳語候補をソートして尤もらしい順に訳語を出力する．

1．中国語見出し語 *TC* を検索語としてウェブ検索しウェブ文書を取得する．この中国語ウェブ文書集合を *CD* とする．
2．中国語のウェブ文書集合 *CD* を，中日翻訳エンジンを用いて日本語に機械翻訳する．中国語見出し語部分は翻訳せず，未知語マスクをかける．ここで得られたものを日訳ウェブ文書集合 *CJD* とする．
3．日訳ウェブ文書集合 *CJD* を形態素解析し未知語周辺 *m* 語中に名詞を含む日訳文脈ベクトル *VCJ* とする．
4．3で得られたすべての日訳文脈ベクトル *VCJ* を選定し，日本語ウェブ文書をウェブ検索するための日本語文書検索語ベクトル *SWJ* を以下の方法で絞込む．
   （1）日訳文脈ベクトル *VCJ* を，*LTI* 基準（後述）を用いてまず数十個に選定し，日本語文書検索語ベクトル *SWJ* とする．
   （2）検索語ベクトルでウェブ検索し，検索された文書数が 100 を超えたところで *SWJ* の絞込みを完了させる．もしここで 100 を超えなければ（3）に進む．ただし，最低一つは日本語のキーワードを残すこととする．
   （3）日本語文書検索語ベクトル *SWJ* を，語の種別特性を用いて絞り込む．あらかじめ指定された絞込み条件にしたがって日本語文書検索語ベクトル *SWJ* 中の順位が低い語句から一つ除外する．

ここでは語の種別特性を利用した日本語文書検索語ベクトルの絞込みを行う．語の種別特性として,「漢字カタカナ混在」「漢字のみ」「英語大文字のみ」「カタカナのみ」などが考えられる．これらの優先順位をユーザが分野に応じて変更することで検索語の絞込み方法を容易に切替ることができる．

たとえば,「漢字カタカナ混在」＜「漢字のみ」＜「カタカナのみ」＜「英語大文字のみ」をベクトル絞り込み条件として指定した場合,「漢字カタカナ混在」,「漢字のみ」,「カタカナのみ」,「英語」の順に，順位の低い語句から除外語を探していく．あるだけの「漢字カタカナ混在」を除外しつくしてしまったら「漢字のみ」で構成されている語句を順位の低い順に取り除く．ただしいずれの過程においても，必ず検索語ベクトルには日本語がひとつ以上含まれていなければならない．日本語の語句がひとつになってしまった場合は「英語」の語句を順位の低いものから除外していく．

5．絞り込まれた検索語ベクトル *SWJ* を用いてウェブ検索して文書を収集し，日本語ウェブ文書集合 *JD* とする．
6．日本語ウェブ文書集合 *JD* を形態素解析して窓幅 *m* 語中に含まれる名詞を取り出し日本語文脈ベクトル *VJ* とする．
7．日訳文脈ベクトル *VCJ* と日本語文脈ベクトル *VJ* の類似度を計算し，あらかじめ与えられた閾値を超える文脈ベクトル対を選定する．選定した文脈ベクトル対を用いて日本語文脈ベクトルの語句から訳語を抽出する．*RSIM* 基準（後述）により対訳語を順位付けし，上位の対において *VCJ* には含まれない語を *VJ* から取り出して訳語候補 *TJ* とする．

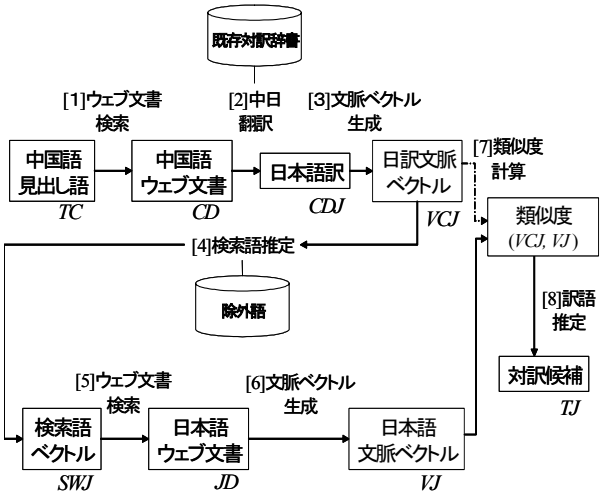

図３ KATrans の処理過程

なお，1KB 未満のウェブ文書は文書がないことを示す文のみを含む場合が多いため除外した．つぎに本手法の特徴について述べる．

4.1. 日本語文書検索語ベクトルの生成方法

日本語文書を検索するための検索語ベクトル *SWJ* を *VCJ* より選定する．選定基準として，KATrans では，予備実験に基づき，logTF・IDF

に基づく尺度 ***LTI*** を定義して用いた.

$$LTI_i = \sum_{j=1}^{n} \log f_{ij} \cdot \log(\frac{n}{n_i})$$

ただし，nは総文書数，n_iは i 番目の語を含む文書の数である．これにより，どの文脈ベクトルにも表れるような語は選定されにくくなる．逆にいうと，各語を特徴付けるような代表語を文脈ベクトルから選定することができる．*LTI* が降順になるように代表語をソートし, 上位 10 個を日本語文書検索語ベクトル *SWJ* とした. また *VCJ* 中に明らかに無関係な用語を除外語として取り除いた．除外語リストは毎日新聞 1 年間分のテキストから *TF* 値が上位 500 位までの語を抽出して作成した.

4.1 日本語文書の収集と訳語推定

（1）単一言語内の類似文脈ベクトルの排除

単一言語内に類似文脈ベクトルが多く含まれると，重なる部分に多くの重み付けがされて文脈がゆがんでしまう問題がある. そこで, あまりにも類似している文脈ベクトルは削除する処理をほどこした．例えば，文脈ベクトル α と β がこの順番で出現するとする. β が α に酷似していれば β を削除する．まず削除の閾値を ξ_{cut}= 0.35 としておく．つぎに α と β の重なり γ を求め, $\xi=|\gamma| / (|\alpha|+|\beta|-|\gamma|)$を計算する. $|\cdot|$はベクトル要素の数とする．$\xi<\xi_{cut}$ であるかを判定しこの条件を満たしていないならば β を文脈ベクトル集合から削除する.

（2）対訳推定基準の改良

絞り込まれた文脈では，より文脈が一致しているほうが訳語が含まれている可能性が高いという観察から，中国語文書の文脈ベクトルと日本語文書の文脈ベクトルの類似度を以下に示す基準 *RSIM* を提案する． 以下，対訳候補集合 *JT* の計算手順を以下に示す.

[1] 日訳文脈ベクトルの生成

あらかじめ決めておいた文脈窓幅 m に従い，見出し語マスクを１つだけ含むような日訳文脈ベクトルを生成する.

$$VCJ = \begin{pmatrix} cw_{11} & \cdots & ew_1 & \cdots & cw_{1m} \\ \vdots & \cdots & ew_i & \cdots & \vdots \\ cw_{n1} & \cdots & ew_n & \cdots & cw_{n\,m} \end{pmatrix}$$

ただし，n は文脈ベクトルの本数，ew_i は文脈ベクトル i における見出し語マスクを，cw_{ij} は文脈語を指す（$1 \leqq i \leqq n$，$1 \leqq j \leqq m$）．つぎに，*VCJ* から見出し語マスクを取り除き純粋文脈ベクトル *VTCJ* を生成する．純粋文脈ベクトル *VTCJ* を以下に示す.

$$VTCJ = \begin{pmatrix} cw_{11} & \cdots & cw_{1\,m-1} \\ \vdots & \ddots & \vdots \\ cw_{n1} & \cdots & cw_{n\,m-1} \end{pmatrix}$$

[2] 日本語文書文脈ベクトルの生成

つぎに，日本語文書文脈ベクトルを生成する．日本語文書文脈ベクトル *VJ* は以下のようになる. ただし r は日本語文書文脈ベクトルの数であり, jw_{rm} は文脈語である．また各文脈語を $jw_{s\,t}$ とする．$1 \leqq s \leqq r$，$1 \leqq t \leqq m$である整数とする.

$$VJ = \begin{pmatrix} jw_{11} & \cdots & jw_{1m} \\ \vdots & \ddots & \vdots \\ jw_{r1} & \cdots & jw_{r\,m} \end{pmatrix}$$

[3] 類似度の計算

各行同士における文脈類似度 *RSIM* を定義する.

$$RSIM_{is} = \frac{\sum_j \sum_t \lambda(cw_{ij}, jw_{st})}{j \cdot t}$$

λ は文脈語同士のマッチング関数で，一致したら１，一致しなければ 0 をとるような関数である．以上の計算で，文脈ベクトルの類似度 $RSIM_{is}$ が高い順にソートし，あらかじめ決めておいた類似度の閾値 θcut を超える文脈ベクトル対を取出す.

[4] 対訳の推定

[3]において類似度の閾値 θcut を超えた文脈ベクトル対を考える．対訳候補は，両文脈に共通して現れるもの以外の語であると考えられる．したがって，日訳文脈語集合から *JW* と *CW* の共通集合を取り除いたものが対訳候補集合 *TW* となる.

$$CW = \bigcup_j cw_{ij},\ JW = \bigcup_t jw_{st},$$

$$TW = JW \setminus (CW \cap JW)$$

ただし “＼”は集合差を表す． このように計算すると，$i \cdot s$ 個の *TW* を得ることが出来る．この中から，語句の出現頻度を計算し，頻度の高い淳に並べ直して *TW* を出力する.

5. 実験

5.1 実験設定

本実験では，中国語の計算機専門用語 10 語（**図４**）を対象とし，日本語の訳語を推定する．文脈窓幅は 5, 21, 35 語の３条件とした．検索語ベクトルの絞込み条件は「漢字カタカナ混在」＜「漢字のみ」＜「カタカナのみ」＜「英語のみ」とした．文書 100 以上を成立条件とした．$\theta cut = 0.75$ とした．

5.2 評価方法

検索精度の評価尺度には，*MRR*（*Mean Reciprocal Rank*）が一般的に良く用いられていることが知られている．これは，検索結果が正解か不正解かの２値である場合には対訳推定評価にそのまま利用できる．しかしながら，本タスクにおいては下位語など，間違ってはいない語句も評価対象に入れる必要があるため, そのままでは *MRR* を使うことができない．そこで，この *MRR* を改良し，重みを付加した log*WMRR*（*Weighted Mean Reciprocal Rank*）で対訳推定精度を評価した．推定された対訳が完全に正解であると考えられる場合には重みを w=1.0 に，関連の深い語や構成語である場合には w=0.5 に設定した．ただし N は正解語語句の個数であり，r は順位を示す．

$$\log WMRR = -1/\log\left(\frac{1}{N}\sum_i w_i \frac{1}{r_i}\right)$$

[Fung00]らの評価方法では，40 位を一つの区切りとして評価得点を調整している．しかしながら，この評価方法ではかなり大雑把な精度計算となるため，log*WMRR* を用いた．

5.3 実験結果

上位 200 位までの訳語推定結果を**図４**に示す．これらの課題語において，すべて 40 位以内で正解を推定できた．また，明らかに文脈幅を狭めたほうが対訳推定精度が向上するものと，逆に文脈幅を広げたほうが精度が良いものとに大別できることが分かった．

6. 訳語推定における文脈窓幅の役割

上位 200 位までの訳語推定結果を図４に示す．本実験結果から文脈窓幅と対訳精度との関係を考察する．まず文脈幅を狭めたほうが対訳推定精度が向上するものと，逆に文脈幅を広げたほうが精度が良いものとに種別を大別することが出来る．これらの見出し語の日本語文書を分析すると日本語文書を検索する語の中心に対訳語が挟まっているものは文脈語を狭くとるほうがよい．しかしながら，日本語文書中で検索語が対訳を挟まない形で散在している場合は文脈幅を広げたほうが対訳推定精度が向上すると考えられる．ベクトル中の検索語に冗長な語が適切に取り除かれると推定精度は向上する（**図５**）．

	No	中国語見出し	正解例（参考）	推定訳語（窓幅5語）	推定訳語（21語）	推定訳語（35語）
「启动盘」タイプ	1	启动盘	起動ディスク	起動ディスク(128)	起動ディスク(7)，セットアップディスク(170)	**起動ディスク(1)**，セットアップディスク(67)，ブータブルCD(87)
	2	计算机病毒	コンピュータウィルス	ウィルス(138)	ウィルス(5)，ワーム(8)，トロイの木馬(34)，ウィルス(77)	**ウィルス(5)**，ワーム(7)，トロイの木馬(27)，ウィルス(92)
	3	以太网	イーサネット	N/A	イーサネット(24)，ギガビット・イーサネット(145)	**イーサネット(11)**，ギガビット・イーサネット(94)
	4	液晶显示器	液晶モニタ	液晶モニタ(181)	TFT液晶(8)，多結晶Si-TFT(24)，有機EL(40)，液晶ディスプレイ(67)，液晶パネル(103)，液晶モニター(125)	TFT液晶(3)，有機EL(9)，**液晶モニター(27)**，液晶パネル(33)，液晶モニタ(40)，多結晶Si-TFT(85)，STN液晶(100)，液晶ディスプレイ(108)
「硬盘驱动器」タイプ	5	硬盘驱动器	ハードディスクドライブ	**ハードディスク(1)**，ストレージ(7)，内蔵ハードディスク(19)，ハードディスクドライブ(58)	ハードディスク(134)	ハードディスク(6)，ストレージ(152)
	6	无线局域网	無線LAN	**無線LAN(2)**	無線LAN(22)	無線LAN(21)
	7	任务栏	タスクバー	**タスクバー(10)**	タスクバー(106)	タスクバー(71)
	8	调制解调器	モデム	**モデム(16)**	モデム(23)，ADSLモデム(85)，ケーブル・モデム(173)	モデム(36)，ADSLモデム(84)
	9	虚拟内存	仮想メモリ	**仮想記憶(18)**	N/A	N/A
	10	条码扫描器	バーコードリーダ	**バーコードリーダー(31)**	N/A	バーコードスキャナ(113)

図４ 推定された訳語の例と推定順位（上位 200 位まで，正解語は下線，最高順位の語句は太字）

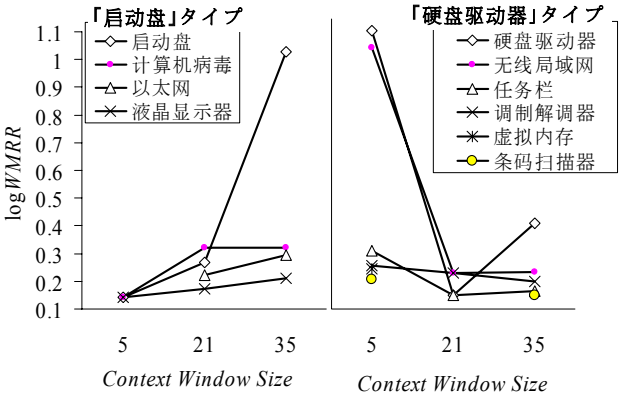

図５ 文脈窓幅と訳語推定精度の関係

つぎに，**図６**に，検索語ベクトルの例を示す．「启动盘」では m=35 が最も logWMRR 値が高く，「硬盘驱动器」では m=5 が最も logWMRR 値が高かった．検索された文書から考えると，「启动盘」は数少ないある特定の文脈で使用されるのに対して，「硬盘驱动器」は様々な文脈で使用されるので検索語ベクトルの数に大きな違いが出たものと考えられる．

启动盘

m=5: Win，ドライブ，CD-ROM，DOS

m=21: Win，ドライブ，CD-ROM，sys

m=35: Win，ドライブ，CD-ROM，Command

硬盘驱动器

m= 5: Ultra，システム，ネットワーク，データ，ハードディスク，SCSI

m=21: ドライブ，Ultra，ネットワーク，ハードディスク，システム，ATA，ネット，SCSI，パラメータ，データ，モデル，プロセス

m=35: ドライブ，Ultra，システム，ハードディスク，ネットワーク，SCSI，ATA，データ，ネット，インタフェース，プロセス，ユーザー，メモリ，コンピュータ

図６ 検索語ベクトルの例

また，検索語ベクトル数と log *WMRR* の関係は，「启动盘」タイプの語では 4 語周辺ならびに 10 語周辺の２箇所にあるのに対して，「硬盘驱动器」タイプでは 4 語から 6 語周辺に精度の良くなるピークがあることが分かった（**図７**）．この結果から，特定の狭い文脈で使用される語と幅広い文脈で使われる語とでは，検索語数と精度の関係に違いが現れることが明らかとなった．

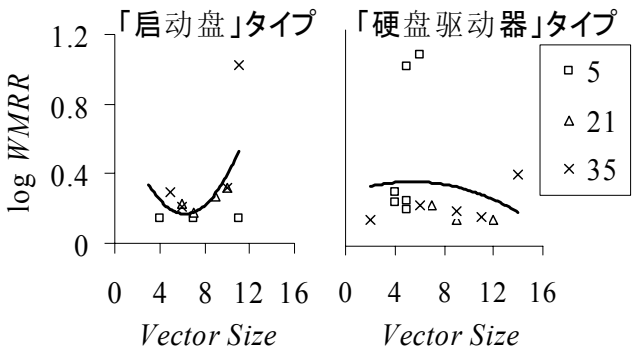

図７：検索語ベクトル数と精度の関係
（各マーカは文脈窓幅，グラフ上には
２次曲線あてはめ結果を付加）

7. まとめ

中日翻訳エンジンならびに文脈語による言語横断ウェブ検索を利用した中日対訳推定システム **KATrans** を提案，試作し，対訳推定実験の結果から，文脈窓幅と対訳訳語推定精度の関係を明らかにした．また，検索語ベクトル数と対訳精度 log *WMRR* の関係を明らかにした．現在は限定された用語のみが推定可能であるが，今後は確率的手法や機械学習手法を導入して精度向上に努める予定である．

参考文献

[Fung 00] Fung, P.: A statistical view on bilingual lexicon extraction, Parallel Text Processing, pp.219-236, 2000.

[Hasan01] Hasan, M. M., and Matsumoto, Y.: Multilingual Document Alignment A Study with Chinese and Japanese, Proc. 6th NLPRS, pp. 617-623, 2001.

[Izuha 01] Izuha, T.: Machine Translation Using Bilingual Term Entries Extracted from Parallel Texts. Congreso sobre traducción automática, vol. 8, 2001.

[熊野 94] 熊野 明, 平川 秀樹: 対訳文書からの機械翻訳専門用語辞書作成, 情報処理学会論文誌, Vol.35, No.11, pp.2283-2290, 1994.

[Rapp95] R. Rapp: Identifying Word Translations in Non-parallel Texts, Proc of the 33rd ACL, pp.320-322,1995.

[Rapp99] R. Rapp: Automatic Identification of Word Translations from Unrelated English and German Corpora, Proc of the 37th ACL, pp.519-526, 1999.

[佐藤 02] 佐藤 理史，佐々木 靖弘: ウェブを利用した関連用語の自動収集，自然言語処理 153-8, 2003.

[Utsuro02] Utsuro, T.: Translation Knowledge Acquisition from Cross-Lingually Relevant News Articles, Proc. of the 2nd China-Japan Natural Language Processing Joint Research Promotion Conference, pp.123-134, 2002.