自然言語処理 112－4
（1996．3．14）

辞書と統計を用いた対訳アライメント

春野 雅彦　　山崎 毅文
NTT コミュニケーション科学研究所

本稿では構造的に異なる言語間の文対応付けを高精度で行なう方法を提案する．日英のように構造が大きく異なる言語間では統計的に得られる訳語対の量に限界があり，統計的手法だけを用いて文対応付けを行なうのは難しい．これは主に機能語の体系が異なることと，1つの単語が様々な品詞の単語に訳出されることによる．逆に既存の単語対訳辞書には分野固有のキーワードが含まれていないことが多く単語対訳辞書だけを用いて汎用の文対応付けシステムを作ることも難しい．我々の手法は既存の単語対訳辞書と統計的に得られた訳語対を同時に利用する．そのため十分な量の訳語対を利用でき，統計的手法を利用することで分野固有のキーワードも得られるため，テキストの分野や長さの変化に対してロバストである．実験の結果様々な種類の日英対訳コーパスに対して既存の手法を上回る良好な文対応付け精度が得られた．

Bilingual Text Alignment Using Statistical and Dictionary Information

Masahiko Haruno　　Takefumi Yamazaki
NTT Communication Science Laboratories

This paper describes an accurate and robust text alignment system for structurally different languages. Among structurally different languages such as Japanese and English, there is a limitation on the amount of word correspondences that can be statistically acquired. The main reason for this is the systems of functional (closed) words are quite different in the two languages. The proposed method makes use of two kinds of word correspondences in aligning bilingual texts. One is a bilingual dictionary of general use. The other is the word correspondences that are statistically acquired in the alignment process. Our method gradually determines sentence pairs (anchors) that correspond to each other by relaxing parameters. The method, by combining two kinds of word correspondences, achieves adequate word correspondences for complete alignment. As a result, texts of various length and of various genres in structurally different languages can be aligned with high precision. Experimental results show our system outperforms conventional methods for various kinds of Japanese-English texts.

1 はじめに

テキストの電子化が進むと共に、対訳コーパスを利用した語彙知識獲得 [7, 12, 9]、機械翻訳 [3, 11]、情報検索 [10] などの研究が盛んになっている。こうした研究では大量の文対応付きコーパスを仮定しているため、対訳コーパスの文対応付けを高精度で自動的に行なうことが重要な課題となっている。

これまでに行なわれた文対応付けの研究は大きく2つの流れに分けることが出来る。1つは文に含まれる単語数 [2] や文字数 [5] などの簡単な属性に基づく方法、他方は2言語の訳語対を利用する語彙に基づく方法である。前者は動的計画法を利用した効率的な実装が可能なこともあり広く利用されているが、その対象となるのは英語 - フランス語などの構造的に類似した言語間の硬い翻訳に限られる。日本語 - 英語などの構造が異なる言語では文字体系、文法体系の違いによってこれらの手法の精度は著しく落ちる。例えば日本語と英語を比較すると日本語が3種類の文字を持つの対して、英語はアルファベットのみであり、各々の文字が持つ情報量は異なる。また日本語と英語では文法、特に機能語の体系が異なり、ヨーロッパ語には存在する冠詞や前置詞が日本語には存在しないため、同じ内容を表現する文でも長さは大きく異なる。従って簡単な属性に基づく方法を構造が大きく異なる言語間に適用することは困難である。

一方、語彙に基づく方法は2言語間の訳語対を利用して文対応付けを行なうもので、[6] は統計的手法を用いた訳語対の獲得と文対応付けを徐々に行なう反復緩和法を提案している。この手法はヨーロッパ言語間では高い精度を達成するが、構造的に異なる言語ではうまく機能しない。日本語 - 英語などでは統計的に得られる訳語対の数が十分ではなく、部分的な対応付けしか得られない。これは主に文法の体系が異なるためである。ヨーロッパ言語間では冠詞、前置詞、否定辞などの機能語や副詞を文対応付けの手がかりとして利用できるが、構造の異なる日本語 - 英語間ではそうはいかない。つまり統計的に獲得された訳語対だけでは十分な文対応付け精度は達成出来ない。[13] は既存の対訳辞書と統計的手法を併用した文対応付け手法を提案している。まず対訳辞書のみによる動的計画法で文対応付けを行ない、統計的訳語対を獲得する。続いて、対訳辞書と統計的訳語対の両者を利用してもう一度動的計画法を適用する。この手法の問題点は動的計画法で利用する評価関数の意味が不明確であることと、精度が対訳辞書の内容に大きく左右されることである。

本稿では上記の問題点を解決するため、統計的手法と既存の対訳辞書を同時に用いる対訳文対応付けシステムを提案する。今日では CD-ROM の普及により多くの対訳辞書が計算機上で利用可能となっており、日 - 英、日 - 仏、日 - 中、日 - 独、英 - 西、英 - 独などの辞書が利用可能である。従ってこれらを対訳文対応付けに利用することは今後有望な手法である。以下に統計に基づく手法と対訳辞書による手法の長短所を示す。これらから両者が文対応付けにおいて、お互いに相補的な役割を果たすことが分かる。

統計の長所 文脈に即した訳語対を獲得出来る。形態素解析の誤りに対してもロバストである。

統計の短所 統計で得られる訳語対の数には限界がある。

対訳辞書の長所 コーパスに1度しか出現しない単語の情報を利用できる。

対訳辞書の短所 専門的な用語を含んでいない。使われる訳語がコーパス中のもの異なっていることがある。また形態素解析の誤りに弱い。

以下に我々のシステムの特徴を示す。

- 反復緩和法を用いるため信頼度の高いものから順に文対応が決定される。
- 対訳辞書と統計的手法の両者を利用するため、分野、長さの異なる様々なテキストに対して高い精度で文対応付けを行なえる。
- 一般の対訳辞書のみを仮定するため経済的である。コストの掛かる専門用語辞書の構築を必要としない。
- 統計的に得られた訳語対をユーザー辞書に登録することで辞書の拡張を容易に行なえる。

本稿の構成は以下の通りである。まず次章で本システムの概要を述べ、3章で文対応付けアルゴリズムを詳細に説明する。4章では新聞の社説、科学論文などに対

して行なった実験結果について報告する。5章ではユーザーが容易に文対応付け結果を確認し、修正できる統合的環境BACCSについて述べる。最後に6章で本稿をまとめる。

2 システム構成

図2に文対応付けシステムの概要を示す。お互いが対訳となっている日英のテキストがシステムの入力である。パターンマッチによって文の境界を確定した後、両言語の形態素解析を行う。[1] 英語の単語に関しては約100の規則を用いて原形を推定している。次にこうして得られた単語列から名詞、動詞、形容詞、副詞、未知語(日本語のみ)を抽出する。これは機能語を残しておくと文法構造の違いのために文対応精度が落ちるためである。

文対応付けアルゴリズムの初期状態は、既に対応している文のペア(アンカー)の集合である。これらのアンカーは記事や章の境界から決定され、最も一般的な場合にはテキストの最初と最後の文のみが初期アンカーを構成する(図2参照)。次に文対応可能なペアをこれらのアンカーから決定する。直観的に言えば、可能な文対応の範囲はアンカーの近くで小さく、アンカーから離れるほど大きくなる(図2参照)。ここで重要なことは正しい文対応が可能な文対応候補の中に含まれていることである。

続いて文対応付けアルゴリズムは2種類の訳語対(統計的手法と対訳辞書から得られたもの)を用いて、文対応可能候補から新しいアンカーを見つける。決められた閾値以上の訳語対を含む文のペアを新たにアンカーとするのである。これら新しく見つかったアンカーによって、統計的訳語推定がより正確になる。この操作をパラメータの緩和を行ないながら繰り返すことで、信頼性の高いものから低いものへと順にアンカーが決定される。緩和法を用いることでアルゴリズムの終盤で起こる対応付け誤りで全体の対応付け精度に及ぼす影響を少なくすることが可能となる。

アルゴリズムの最終的な出力は文対応済みコーパス(アンカーの列)と統計的に得られる単語対応である。ここで述べたアルゴリズムの全体は統合的文対応付け環境BACCS(Bilingual Aligned Corpus Construction System)として実装されている。BACCSを利用することでユーザーは容易に文対応付け結果を確認し修正出来る。また、統計的に得られた訳語対をユーザー辞書として登録することも出来る。

[1]日本語、英語の形態素解析には各々 JUMAN システム [8]、transformation-based tagger [1] を使用した。これらのツールを提供して頂いた方々に感謝します。

3 文対応付けアルゴリズム

3.1 統計的類似度

この節では訳語対を決定するために用いる統計手法について説明する。様々な類似度基準の中から、我々は(自己)相互情報量と$t-$スコアを採用した。これは他の類似度基準に比べ、パラメータの緩和過程を細かく制御出来るからである。相互情報量は出現分布の類似度を示し、t-スコアはその類似度の信頼性を意味する。この2つのパラメータを用いることで1つのパラメータを用いる従来法[6]より細かい緩和が可能となる。

我々が用いる基本的なデータ構造は**文対応可能行列(ASM)**と**アンカー行列(AM)**の2つの行列である。ASMは対応可能な文のペアを表し、0と1から構成される。ASMの$i-j$要素が1であるのは日本文iと英文jが対応可能である場合であり、0であるのは対応不可能な場合である。これに対してAMのi-j要素は日本文iと英文jの文ペアが含む訳語対の'のべ総数'である。すなわち日本文iと英文jのペアがいくつの訳語対によって支持されたかを表す。文対応付けが進むにつれてASM中の1の個数は少なくなり、AM中の要素の値は大きくなる。

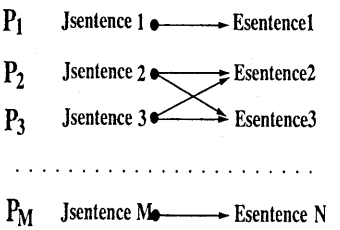

図3: 可能な文対応

図3に示すようにp_iを日本文iとその対応可能な英文から成る集合であるとする。例えばp_2は$Jsentence_2$、

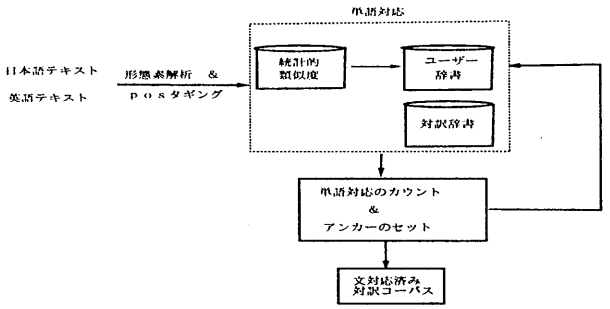

図 1: 文対応付けシステムの概要

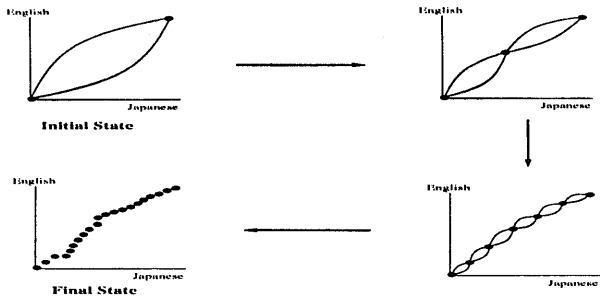

図 2: 文対応付けの過程

$Esentence_2$、$Esentence_3$から成る集合であり、$Jsentence_2$ は $Esentence_2$、$Esentence_3$ と対応可能であることを意味する。

ここで日本語単語 w_{jpn} と英単語 w_{eng} の出現に関する類似度を評価するために表 1に示すような分割表[4]を考える。分割表は以下の４つの頻度分布から構成される: (a) w_{jpn} と w_{eng} の両方が現われた p_i の数、(b) w_{eng} のみが現われた p_i の数、(c) w_{jpn} のみが現われた p_i の数、(d) w_{jpn} と w_{eng} のどちらも現われない p_i の数。ただし１つの w_{eng} は(a)で高々１回しか数えないようにする。

	w_{jpn}	
w_{eng}	a	b
	c	d

表 1: Contingency Matrix

もし w_{jpn} と w_{eng} が正しい訳語対であればａが大きく、ｂとｃが小さくなるはずである。逆に w_{jpn} と w_{eng} が訳語対でなければａが小さく、ｂとｃが大きくなる。これらを定量的に評価するために以下の(自己)相互情報量を導入する。

$$\log \frac{prob(w_{jpn}, w_{eng})}{prob(w_{jpn})prob(w_{eng})}$$

ここで各々の確率は以下のように与えられる。

$$prob(w_{jpn}) = \frac{a+c}{a+b+c+d} = \frac{a+c}{M}$$

$$prob(w_{eng}) = \frac{a+b}{a+b+c+d} = \frac{a+b}{M}$$

$$prob(w_{jpn}, w_{eng}) = \frac{a}{a+b+c+d} = \frac{a}{M}$$

良く知られているように相互情報量は出現頻度の低い事象に対して大きくなる性質がある。そこでその信頼性を評価するために以下の t- スコアを導入する。信頼性の低い相互情報量の値は t- スコアに閾値を設定す

ることで排除する。例えば1.65以上のt-スコアは$p > 0.95$の信頼性レベルである。

$$t \approx \frac{prob(w_{jpn}, w_{eng}) - prob(w_{jpn})prob(w_{eng})}{\frac{1}{M}prob(w_{jpn}, w_{eng})}$$

3.2 対応付けアルゴリズム

我々の基本的なアルゴリズムは文対応可能行列(ASM)を用いてアンカー行列(AM)を更新する操作をパラメータの緩和を行ないながら繰り返すことである。ASMが与えられると、可能な文対応の中の全ての単語対に対して相互情報量とt-スコアを計算する。この2つの値が予め決められた閾値より大きい場合に、その単語対は訳語対であると判断される。続いて新しいアンカーを見つけるために、この統計的に得られた訳語対と対訳辞書から得られた訳語対を利用してAMの要素の値を更新する。AMのi-j要素がある閾値を越えた場合には、i-jのペアは新しいアンカーであると見なされる。以下でアルゴリズムの詳細を説明する。

文対応可能行列(ASM)の初期化 この段階で初期的なASMを構成する。テキストがM文の日本文とN文の英文から成る場合、ASMは$M \times N$行列である。まず記事や章の境界を用いて初期的なアンカー集合を決定する。最も一般的な場合には図2のようにテキストの最初と最後の文同士のみがアンカーに設定される。

続いて、このアンカー集合を用いて可能な文対応の範囲を決定する。直観的には文対応はアンカーを結ぶ直線に近く分布するはずである。我々は2つのアンカーの中央辺りでは1文に対して$O(\sqrt[3]{L})$(Lは2つのアンカーの間に存在する文数)の文を対応付ける関数を用いる。これは統計学的に最大の分散が$O(\sqrt[3]{L})$でモデル化出来るからである[6]。この初期的なASMは正しい文対応を含んでいる限り、文対応の精度には余り影響を及ぼさない。

アンカー行列(AM)の構成 この段階では与えられた文対応行列(AM)と対訳辞書を用いてアンカー行列(AM)を更新する。ここでt_{high}、t_{low}、I_{high}、I_{low}をt-スコアと相互情報量に対する各々2つの閾値であるとする。また、ANCを文ペアがアンカーと認定されるために必要な最小の訳語対の数であるとする。

まず、ASMの対応可能文ペア中に含まれる全ての単語対に対してt-スコアと相互情報量を計算する。これ以降はt_{low}とI_{low}を越えるt-スコアと相互情報量を持つ単語対だけを考察の対象とする。ASM中の全ての対応可能ペア$Jsentence_i$と$Esentence_j$に対して以下の操作を行なう。

1. 次の3つの条件が成り立てばAMのi-j要素に3を加える。(1) $Jsentence_i$と$Esentence_j$が対訳辞書の訳語対(w_{jpn}とw_{eng})を含んでいる。(2)w_{eng}が$Jsentence_i$と対応可能な他の英文に出現しない。(3) $Jsentence_i$と$Esentence_j$が既に閾値ANCを越えている他のアンカーと交差しない。
2. 次の3つの条件が成り立てばAMのi-j要素に3を加える。(1) $Jsentence_i$と$Esentence_j$がt_{high}、I_{high}を越える統計的訳語対(w_{jpn}とw_{eng})を含んでいる。(2)w_{eng}が$Jsentence_i$と対応可能な他の英文に出現しない。(3) $Jsentence_i$と$Esentence_j$が既に閾値ANCを越えている他のアンカーと交差しない。
3. 次の3つの条件が成り立てばAMのi-j要素に1を加える。(1) $Jsentence_i$と$Esentence_j$がt_{low}、I_{low}を越え、t_{high}、I_{high}以下の統計的訳語対(w_{jpn}とw_{eng})を含んでいる。(2)w_{eng}が$Jsentence_i$と対応可能な他の英文に出現しない。(3) $Jsentence_i$と$Esentence_j$が既に閾値ANCを越えている他のアンカーと交差しない。

最初の手続きは対訳辞書の訳語対を扱うためのものであるに対して、2番目の手続きは信頼性の高い統計的訳語対を扱うためのものであり、多くの場合文脈に依存したキーワードを含んでいる。これらの訳語対に対してはAMに等しく3を付加している。3番目の訳語対は1、2番目の訳語対だけでは文対応付けに十分な数が得られないため導入される。これらの訳語対に対してはAMに1を付加する。この手続きで採用される訳語対は訳語としては誤ってい

ることもあるが、文対応付けアルゴリズムの終盤で特に重要な働きをする。

ASMの更新 この段階では上記の手続きで構成したAMを用いてASMを更新する。AM中でANC個以上の訳語対を持つ文ペアはアンカーと見なされる。このアンカー集合から初期的なASMを構成したのと全く同様にして新しいASMが構成される。

我々のアルゴリズムはt_{low}、I_{low}、ANCを徐々に下げることによって1種の反復緩和法を実現している。これによって信頼性の高いものから順に文対応が得られるためダイナミックプログラミングに基づく手法よりも高い精度を実現出来る。

4 実験並びに評価

この章では様々な対訳コーパスに対して行なった文対応付け実験について、文対応の精度と統計的に得られた訳語対の観点から報告する。実験に用いたテキストの種類と性質を表2にまとめた。テキスト1とテキスト2は読売新聞の社説記事であり、WWWサーバーから電子的に得られる[2]。テキスト3は経済評論、テキスト4はサイエンスから取得した科学記事である。表中にはそれぞれのテキストの文数と正しい文対応付け結果の分布を示してある。この表から各々のテキストに対する文対応付け難度の概略を知ることが出来る。

ここで最も長いテキスト4を例にして実験に要した計算時間について簡単に触れておく。全ての前処理に25秒要した後、最初のループが終了するまでに25秒掛かった(t_{low} = 1.55 I_{low} =1.8)。その後残り6回のループ全てを終了するのに20秒を要した。この結果はSpaarc Station 20上で得られたものである。この結果から我々の手法は大規模な対訳コーパスに対しても現実的な速度で動作すると言える。

4.1 文対応付けの精度

表3に表2のテキストに対して行なった文対応付けの精度を示す。統計、対訳辞書はそれぞれ統計的手法、対訳辞書によって得られた訳語対だけを用いて文対応付けを行なった結果である。ここで対訳辞書としては講談社和英辞典を利用している。表中でPRECISIONは自動的に付けられた文対応のうち正しいものの割合いを示し、RECALLは人手で付けた正解のうち対応付けプログラムで再現出来たものの割合いを示している。ただし、この評価中では既存の方法と異なり、誤りを文対文のレベルで評価した。例えば日本文3文が英文1文に対応する場合に既存の評価では高々1つしか誤りが生じないが、我々の評価ではRECALLにおいて最大3つの誤りが生じ得る。従って表3に示した値は既存の評価よりかなり厳しいものである。

テキスト1、テキスト2の短いテキストに対しては統計的手法は著しく成績が悪く、テキスト3、テキスト4に対しては統計的手法の方が対訳辞書のみの手法より成績が良い。これはテキストがある程度の長さを越えると統計によって文脈に依存した訳語対を抽出することが出来るためであると考えられる。ここで注意すべ

日本語	英語	相互情報量
記録	recording	3.58
リアルタイム	real	3.51
ニューロン	neuron	3.51
フィルム	film	3.51
グルコース	glucose	3.51
増加	increase	3.51
磁	MEG	3.51
解像度	resolution	3.43
電気	electrical	3.43
グループ	group	3.39
電気	recording	3.39
記録	electrical	3.39
言う	generate	3.33
提供	provide	3.33
図	MEG	3.33
NMR	NMR	3.17
ファンクショナル	functional	3.17
機器	equipment	3.17
臓器	organ	3.15
注射	compound	3.10
水	water	3.10
標識	radioactive	3.10
PET	PET	3.10
解像度	spatial	3.10
そのようだ	such	3.10
代謝	metabolism	3.06
言う	verb	3.04
科学者	scientist	2.95
地図	mapping	2.92
大学	university	2.92
思考	thought	2.90
化合物	compound	2.82
標識	label	2.82
オートラジオ	radioactivity	2.77
視覚	visual	2.77
信号	signal	2.77
リアルタイム	time	2.69
オートラジオ	autoradiography	2.67
能力	ability	2.63
CT	CT	2.63
聴覚	auditory	2.15
心	mental	2.05
MRI	MRI	1.8

表4: 統計的に得られた単語対応

[2]データの使用を許可頂いた読売新聞社に感謝致します。

No.	テキスト名	日本語	英語	1-1	1-2	2-1	3-1
1	*Root out guns at all costs*	26	28	24	2	0	0
2	*Economy facing last hurdle*	35	42	25	7	2	0
3	*Pacific Asia in the Post-Cold-War World*	134	123	114	0	10	0
4	*Visualizing the Mind*	223	212	186	6	15	1

表 2: 評価に用いたテキスト

	提案手法		統計		対訳辞書	
Text	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL
1	96.4%	96.3%	65.0%	48.5%	89.3%	88.9%
2	95.3%	93.1%	61.3%	49.6%	87.2%	75.1%
3	96.5%	97.1%	87.3%	85.1%	86.3%	88.2%
4	91.6%	93.8%	82.2%	79.3%	74.3%	63.8%

表 3: 文対応付けの結果

きことは全てのテキストに対して提案手法が最も良い成績を収めていることである。このことから提案手法が様々な分野、長さのテキストに対して有効に働くことが分かる。

4.2 統計的に獲得された訳語対

この節では脳の非侵襲測定に関する解説であるテキスト4を例にして、どのような訳語対が統計的に得られるかを示す。表 4にテキスト4から得られた訳語対のうち相互情報量の値が大きいものを示す。表中の NMR、MEG、PET、CT、MRI、Functional MRI はどれも脳を外部から測定するための装置であり、このテキストのキーワードである。ところがこれらの専門用語は我々が用いた対訳辞書には含まれていなかった。

また MEG の正しい日本語訳は脳磁図であるが、形態素解析システムがこれらを脳、磁、図に分解してしまっている。ところが表から統計的手法によって磁や図と MEG の対応を正しく捉えられていることが分かる。このように統計を利用することで分野依存の訳語対を獲得し、文対応付けの精度を大きく向上させることが可能となる。

5 統合的文対応付け環境 BACCS

この節では統合的文対応付け環境 BACCS(Bilingual Aligned Corpus Construction System) について述べる。精度の高い対訳コーパスを大量に作るにはユーザが素早く対応付け結果を確認し、必要があれば修正を行なわなければならない。このためのツールとして我々は BACCS を作成した。ユーザは次の操作手順で対訳コーパスを作成する。

1. 自動文対応付けプログラムを走らせて結果を画面に表示する。
2. プログラムによって出された信頼度(色で区別)を参考にして、結果を確認し、必要があれば修正する。
3. 統計的に得られた訳語対の中から適切なものをユーザー辞書に登録する。

図 5に BACCS の画面を示す。

6 結論

我々は日本語 - 英語のような構造の異なる言語間の対訳コーパスに対して高い精度で文対応付けを行なうための手法を提案した。本手法は統計的手法と対訳辞書

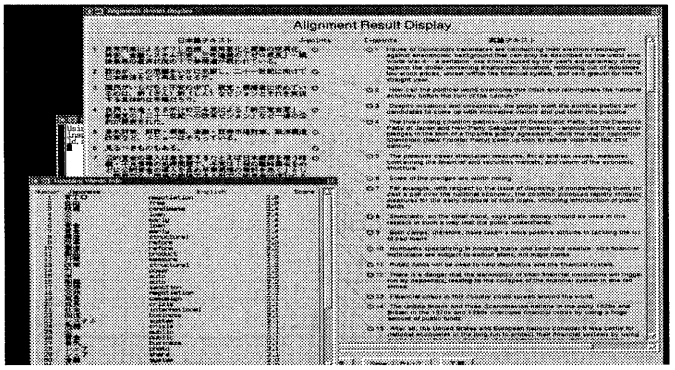

図 4: 統合的文対応付け環境 BACCS

から得られる訳語対を同時に利用することにより、分野や長さの異なる対訳コーパスに対して安定して高い精度の文対応付けを行なえた。

謝辞 本研究に関して辞書検索プログラムを提供、ならびに有益な議論をして頂いた宇津呂 武仁氏に感謝致します。

参考文献

[1] Eric Brill. A simple rule-based part of speech tagger. In *Proc. Third Conference on Apolied Natural Language Processing*, pages 152–155, 1992.

[2] P F Brown et al. Aligning sentences in parallel corpora. In *the 29th Annual Meeting of ACL*, pages 169–176, 1991.

[3] P F Brown et al. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311, June 1993.

[4] Pascale Fung and K W Church. K-vec: A new approach for aligning parallel texts. In *Proc. 15th COLING*, pages 1096–1102, 1994.

[5] W A Gale and K W Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, March 1993.

[6] Martin Kay and Martin Roscheisen. Text-translation alignment. *Computational Linguistics*, 19(1):121–142, March 1993.

[7] Julian Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *the 31th Annual Meeting of ACL*, pages 17–22, 1993.

[8] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of Japanese morphological analyzer juman. In *Proc. International Workshop on Sharable Natural Language Resources*, pages 22–28, 1994.

[9] Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. Structural matching of parallel texts. In *the 31th Annual Meeting of ACL*, pages 23–30, 1993.

[10] Satoshi Sato. CTM: an example-based translation aid system. In *Proc. 14th COLING*, pages 1259–1263, 1992.

[11] Satoshi Sato and Makoto Nagao. Toward memory-based translation. In *Proc. 13th COLING*, pages 247–252, 1990.

[12] Frank Smadja and Kathleen McKeown. Translating collocations for use in bilingual lexicons. In *ARPA Human Language Technology Workshop 94*, pages 152–156, 1994.

[13] Takehito Utsuro, Hiroshi Ikeda Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. Bilingual text matching using bilingual dictionary and statistics. In *Proc. 15th COLING*, pages 1076–1082, 1994.