用例ベース翻訳のための日英アライメント確信度語類似度を用いた訳語選択

荒牧　英治[†,††] 黒橋　禎夫[†,††]
柏岡　秀紀[††] 田中　英輝[††]

本稿では，内容レベルで対応のとれている対訳記事コーパスを用いて，用例ベース翻訳を実現する手法を提案する．まず，対訳コーパスの文・句アライメントを行い，確信度の高いものを抽出し，翻訳用例データベースに登録する．次に，与えられた入力文と類似しており，かつ，アライメント確信度の高い翻訳用例をデータベースから選択し，翻訳文を生成する．訳語選択という観点からおこなった実験は 82 %の精度であり，用例ベース翻訳が可能であることを実証的に示す．

キーワード:　訳語選択，用例ベース機械翻訳，アライメント

Word Selection based on Source Language Similarity and Parallel Alignment Confidence

Eiji Aramaki[†,††] , Sadao Kurohashi[†,††] , Hideki Kashioka[††]
and Hideki Tanaka[††]

We propose a method of constructing an example-based machine translation (EBMT) system that exploits a content-aligned bilingual corpus. First, the sentences and phrases in the corpus are aligned across the two languages, and the pairs with high translation confidence are selected and stored in the translation example database. Then, for a given input sentences, the system searches for fitting examples based on both the monolingual similarity and the translation confidence of the pair, and the obtained results are then combined to generate the translation. Our experiments on translation selection showed the accuracy of 82 % demonstrating the basic feasibility of our approach.

KeyWords:　*Word Selection, Example-based Machine Translation, Alignment*

1　はじめに

用例ベース翻訳 (Nagao 1984) とは，入力文と類似した用例をデータベースから探しだし，その用例を組み合わせて翻訳を行う手法である．この方式で実用的なシステムを動かすためには，構造的情報を持った翻訳用例を大量に構築することが必要である．このためには，大規模な対訳コーパスや高精度な構文解析が必要である．

近年，高精度な構文解析は徐々に利用可能となってきたが，質のよい対訳コーパス (パラレ

† 東京大学大学院情報理工学系研究科, Graduate School of Information Science and Technology, The University of Tokyo
†† ATR 音声言語コミュニケーション研究所, ATR Spoken Language Translation Reserch Laboratories

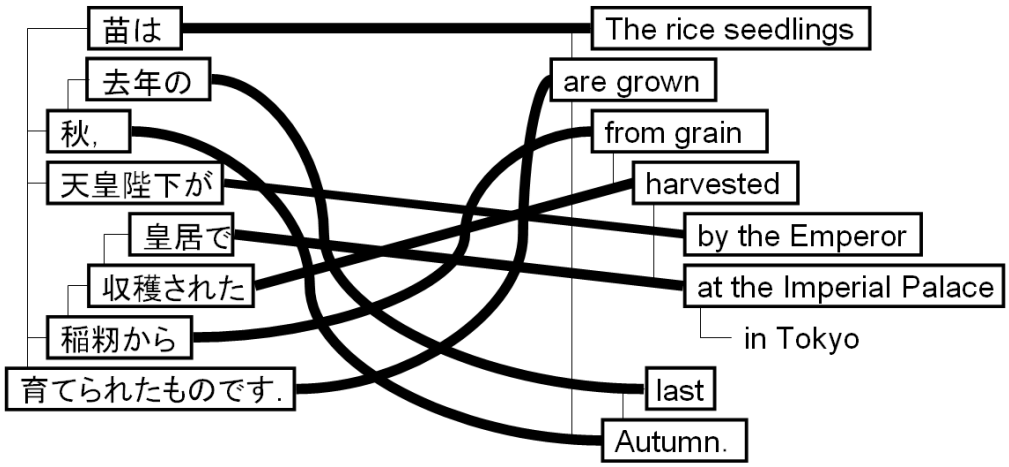

図 1　翻訳用例 (TE) の例

ルコーパス) の量は期待していたほど増加していない．一方，新聞記事や放送原稿などの両言語が内容レベルで一致している対訳コーパスの量が増えているのが現実である．このような対訳コーパスをここでは**コンテントアラインコーパス**とよぶ．本稿では，コンテントアラインコーパスを用いて用例ベース翻訳を実現する手法について述べる．

本手法では，まず，対訳コーパスの文・句アライメントを自動的に行い，その中の確信度の高い部分だけを対訳用例データベースとする．次に，入力文に対して，入力文と用例の類似度，および用例のアライメントの確信度に基づき翻訳に使える用例を選択する．最後に，用例を組み合わせることによって翻訳文を生成する．提案手法は対象とする言語ペアを特定しないが，日英対訳コーパスを用いたため，本稿は日英翻訳を中心にして述べる．

本稿の構成は以下のとおりである．2 章において，対訳コーパスから対応関係の確信度の高い部分だけ抜き出して用例を構築する方法を述べる．3 章で，構築した用例ベース翻訳システムについて，用例の選択手法を中心に述べる．4 章では，訳語選択という観点から行なった評価実験について述べ，5 章に関連研究，6 章でまとめを述べる．

2　用例の作成

用例ベース翻訳では，非常に短い文や，ドメイン依存のパターンで説明されるような文を対象とする特殊な場合を除いて，入力文が一つの用例だけで翻訳されるということは考えにくい．そこで，複数の用例を利用し，それらを組み合わせて翻訳文を作り出さなければならない．そのためには，入力文の構文解析が必要となるのは当然として，用例の両言語側の文も十分に構造化され，用例の言語間が適切な単位で対応付けられている必要がある．

表 1　NHK 対訳記事コーパスの例 “田植えフェスティバル”

田植えフェスティバル**石川県**輪島市で外国の大使や一般の参加者など千人あまりが急な斜面の棚田で田植えを体験する催しが行われました。輪島市**白米町には（しろよねまち）**千枚田と呼ばれる（せんまいだ）**大小**二千百枚の棚田が急な斜面から海に向かって拡がっています。田植え体験は**農作業を通して米作りの意義などを考えていこうという**地球環境平和財団の呼び掛けで開かれたもので、**海外三十四ヵ**国の大使や書記官、それに一般の参加者ら合わせておよそ千人が集まりました。田植えに使われた苗は去年の秋、天皇陛下が皇居で収穫された稲籾から育てたものです。参加者たちは裸足になって水田に足を踏み入れ地元に伝わる田植え歌に合わせて**慣れない手つきで**苗を植えていました。**きょうの輪島市は雲が広がったもののまずまずの天気となり、出席された高円宮さまも海からの風に吹かれながら田植えに加わっていました。地球環境平和財団では今年の夏休みに全国の子どもたちを対象に草刈りや生きものの観察会を開く他、秋には稲刈体験を行なう予定にしています。**
Ambassadors and diplomats from **37** countries took part in a rice planting festival **on Sunday** in small paddies on steep hillsides in Wajima, **central Japan**. About one-thousand people gathered at the hill, where some two-thousand 100 miniature paddies, called Senmaida, stretch toward the Sea of Japan. The event was organized by the private Foundation for Global Peace and Environment. The rice seedlings are grown from grain harvested by the Emperor at the Imperial Palace **in Tokyo** last autumn. Barefoot participants waded into the paddies to plant the seedlings by hand while singing a local folk song about the practice of rice planting.

本論文では，句を最小単位として適切に対応付けられた対訳文を翻訳用例 (Translation Example，**TE**) とよぶことにする (図 1)．本章では，対訳コーパスを構造化，対応付けすることにより，このような TE のデータベースを構築する方法を述べる．

本研究で用いる対訳コーパスは直訳的なものではなく，自動的な対応付けには誤りが含まれるため，コーパス全体の自動解析結果を用例として用いることには問題がある．そこで，解析精度が比較的高いものだけを収集し，TE データベースとする．

2.1　NHK 対訳記事コーパス

本研究では，対訳コーパスとして NHK 対訳記事コーパスを用いた．これは，日本語記事がまずあり，それをもとに英語記事が書かれたもので，日英 4 万記事ペア (5 年間分) からなる．

表 2　NHK 対訳記事コーパスの例 “女子バレーボールオリンピック出場権獲得”

バレーボールの女子チームのアトランタオリンピック出場が決まりました。女子バレーボールのアトランタオリンピック出場権をかけた世界最終予選で、今夜日本はクロアチアをセットカウント三対〇のストレートで破って三位以内を確定し、アトランタオリンピックの出場を決めました。日本の女子は、オリンピックにバレーボールが採用された一九六四年の東京大会以来、連続で出場権を獲得しました。
Japan, **the Netherlands and Ukraine** have qualified for the women's volleyball event in the **Summer** Olympic Games in Atlanta. The three countries secured the top three spots in the final qualifying tournament **in Osaka on Saturday**. Japan defeated Croatia three to zero, and together with the Netherlands and Ukraine had a record of five wins and one loss in the eight-country round-robin tournament. The Japanese women's team has qualified for every Olympic volleyball competition since the sport became an Olympic event in 1964. **The Japanese men's team failed to qualify for the Atlanta Olympics.**

1 記事の平均日本語文数は 5.2，英語文数は 7.4 である．表 1，2に対訳記事の例を示す．NHK の対訳記事は，内容全体としては対応しているものの，直訳されて作成されているわけではなく，相手側言語に対応する表現がない場合がある．このような情報の過不足を，表 1，2中では，**ボールド書体**で示している．この情報の過不足は，主に次の 4 つの要因で発生している．

(1)　**外国人向けの情報追加**

土地の所在地や歴史的な知識など，日本では常識的であるが，外国人には説明を要することがある．例えば，表 1では，“皇居で at the Imperial Palace **in Tokyo**” のように，皇居が東京にあるという情報を追加している．

外国人向けの情報削除

比較的重要性が低い箇所は英語記事では削除されることがある．例えば，表 1では，日本語記事の来年の予定に関する記述が，英語記事では省略されている．

(2)　**視点の差異**

日本語原稿においては原則的に日本を中心とした原稿作成が行われるのに対して，英語原稿では日本を重要視しない場合がある．例えば，表 2では，日本語記事では，日本以外の国がオリンピック出場権を獲得したかどうかは明らかにされないが，英語記事では “Japan, the Netherlands and Ukraine have qualified for ...” と，3 カ国の中の 1 国として日本が表現されている．

(3) **報道時間のずれ**

時差や他の報道内容の都合上，英語原稿が放送される日付は，日本語放送のそれと異なる場合がある．このため，英語報道ではしばしば日付情報が付加される．例えば，表 1では，英語側で “**on Sunday**” と日付情報が追加されている．また，日本語原稿が報道されてから，事態が進展した場合などは，英語原稿に追加ニュースが加わることもある．表 2では，日本語記事は女子バレーボールの試合結果を伝えているのに対して，英語記事は男子バレーボールの試合結果が追加されている．

2.2 文アライメント

まず，DP マッチングにより対訳文の文対応を求める．文対応の基本単位としては，日本語文と英語文の文数が 1:1，1:2，1:3，2:1，2:2 の文対応を考える (後述する実験で用いた評価コーパスで，これらの文対応の割合が全体の 84 %を占めた)．

文対応のスコアは，その中の内容語で翻訳関係にあるものの割合，内容語対応率 (WCR) を用いた．

$$WCR = \frac{W_d}{W_j + W_e} \tag{1}$$

ここで，W_j は日本語内容語数，W_e は英語内容語数，W_d は翻訳辞書によって翻訳関係にあると判断された (両言語の) 内容語数とする．

利用した翻訳辞書は，EDR 日英対訳辞書，EDICT(一般的な日英対訳辞書)，ENAMDICT(固有名詞の日英対訳辞書)，アンカー日英対訳辞書，英辞郎である．これらの辞書にはのべ約 200 万語 (句) の対応が記載されている．

評価コーパスを用いて記事対応付けの精度を評価したところ，次の式で計算される適合率の値は 60.7 %であった．

$$\text{適合率} = \frac{\text{正しく推定された文対応数}}{\text{推定された文対応数}} \tag{2}$$

これを文対応の種類ごとにみると，1:1 対応として対応付けられたものの適合率が最も高く，1:1 対応だけについては適合率は 77.5 %であった．以降の処理においても 1:1 対応が最も扱いやすいものであるため，1:1 対応として求まった対訳文ペアだけを以降の処理対象とすることとした．

2.3 句アライメント

次に，前節で求まった 1:1 の対訳文に対して，(Aramaki, Kurohashi, Sato and Watanabe 2001) に基づく手法によって句アライメントをとる．句アライメントは次の 3 つのステップからなる．

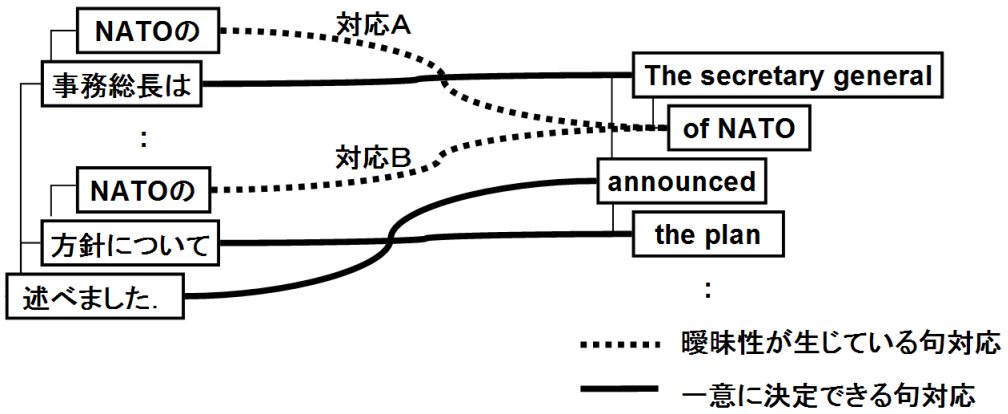

図 2　辞書引きの曖昧性解消の例

ステップ 1:　句を単位とした依存構造への変換

対訳の日本語文を KNP(Kurohashi and Nagao 1994) によって，英語文を nl-parser(Charniak 2000) によって統語解析する．KNP の出力は文節単位の依存構造であり，基本的にそれをそのまま用いる．nl-parser の出力は単語単位の句構造であるので，次の基準によって句にまとめ，さらに主辞を持ち上げていくことにより，やはり句を単位とする依存構造に変換する．

(1)　機能語を後続する内容語にまとめる．

(2)　複合名詞を構成する名詞は一つの句にまとめる．

(3)　助動詞を主動詞にまとめる．

ステップ 2:　翻訳辞書に基づく句対応関係の推定

翻訳辞書を用いて日英の句間に対応をつける．ある語の訳語が相手側言語に複数存在すれば，1 対多や多対多という曖昧性が生じる．この曖昧性はある対応の周辺に他の対応が多くあればあるほどよいというヒューリスティックスを用いて解消する．例えば，図 2では，“NATO” について曖昧性が生じており，対応 A と対応 B という 2 つの句対応の可能性が考えられる．この場合，対応 A は周辺に “事務総長 / The secretary general” という対応をもつため，提案手法は対応 A を採用することになる．

ステップ 3:　未対応句の処理

最後に，辞書によって対応付けられなかった句 (未対応句) に対して，周辺の対応との整合性を考慮して，既存の対応への併合や新規の対応付けを行う．この操作は両言語間で句の依存関係は保存されるという原則に基づいた規則によって行われる．図 3にこのような規則の一例と，新たに発見される対応の例を示す．

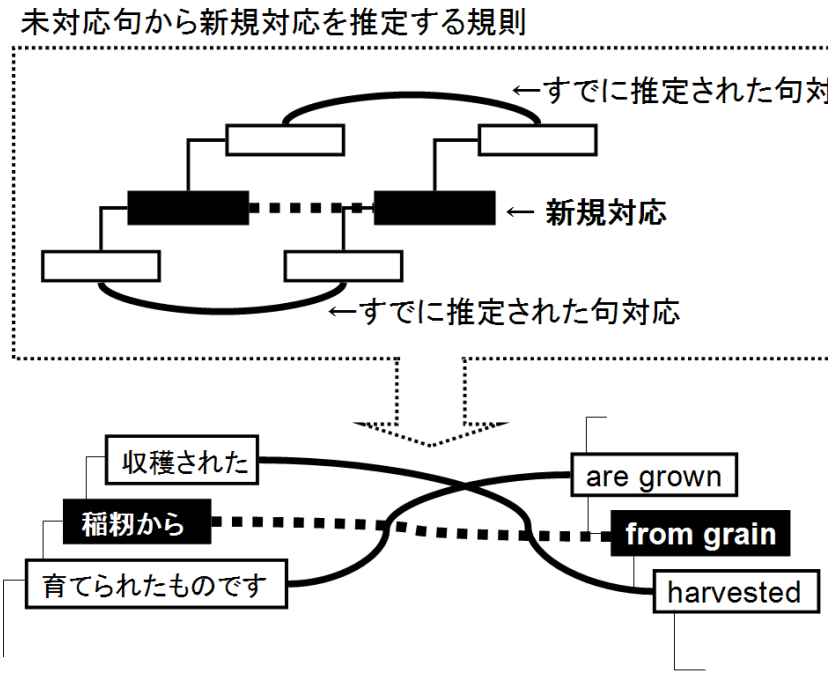

図 3　新規対応を推定する規則と推定された対応の例

以上の処理によって，図 1に示すような句アライメントを行うことができる．

句アライメントの評価のため，前節で文アライメント行なった評価データのうち 1:1 文対応 (145 文ペア) に対して，適切な内容語対応を人手で付与した．評価は，本手法で得られた句アライメントが，1 つ以上の人手による対応を過不足なく含んでいる場合を正解，それ以外を不正解とした．

この評価では，句アライメントの適合率は 50 %であった．この適合率と，対訳文における内容語対応率 (WCR) の関係を調べたところ，図 4に示すように，WCR と適合率の相関関係が明かとなった．これに加えて，WCR と文アライメントの適合率についても相関関係がある．そこで，両方の適合率がともにほぼ上限に達する WCR 0.3 以上の対訳文を TE とすることとした．

2.4　TE データベースの構築

前節で述べたように，NHK の日英対訳記事コーパスについては，自動的に文アライメント，句アライメントを行ったもののうち，1:1 の文対応で，WCR 0.3 以上のものを TE として用いることとした．表 3に WCR ごとの TE 数を示す．

これに加えて，科学技術白書，経済白書の日英対訳コーパス，および SENSEVAL2(Kurohashi 2001) の用例を利用する．これらは，はじめから文対応がとれたものであり，両言語で対応する

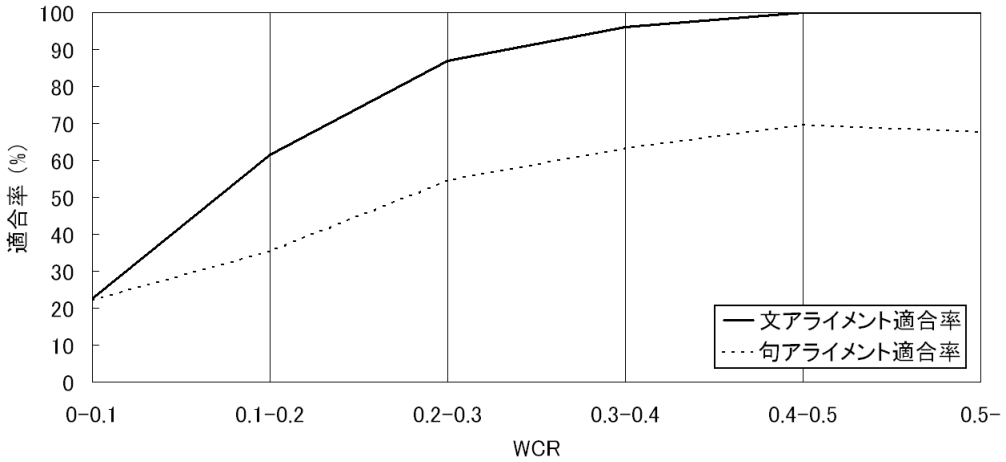

図 4　WCR(内容語対応率) と適合率

表 3　コーパスと TE の数

コーパス	WCR	TE 数
NHK 対訳記事コーパス	0.3〜0.4	18290
	0.4〜0.5	6975
	0.5〜	2314
SENSEVAL	-	6920
白書	-	2225

度合いが高く，本手法による句アライメントの精度も全体として高い (70 %以上) ので，全体を TE として用いることとした (表 3).

3　用例ベース翻訳システム

本研究の翻訳システムは，日本語文を入力とし，その英語翻訳文を出力する．翻訳対象の入力文は，まず構文解析を行い，句単位の依存構造に変換する．次に，構築した TE データベースを用いて，各句 (およびその周辺) を翻訳するのに最も適切な用例を選択し，その英語部分を適切に結合・表層化することにより翻訳文を生成する (図 5).

本節では，適切な用例の選択方法を中心に翻訳システムについて述べる．

3.1　TE 片選択の考え方

TE とは，句 (またはそれ以上) の単位で対応付けされた対訳文全体をさす．TE の中の一部分で，入力文中の句 (またはそれ以上の大きさの部分) の翻訳に直接利用できるものを TE 片と

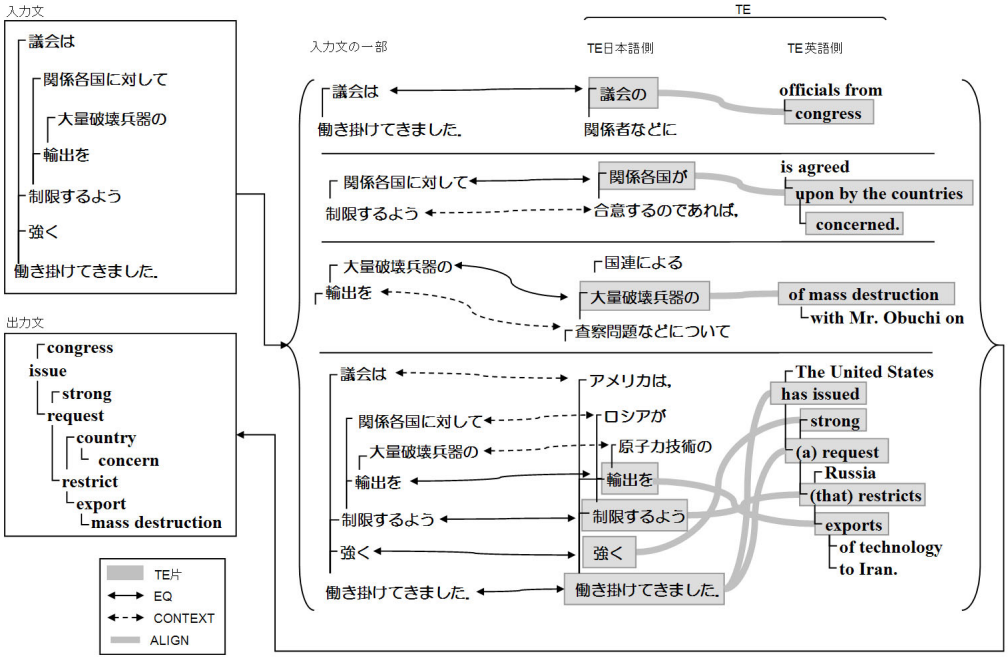

図 5　手法のながれ

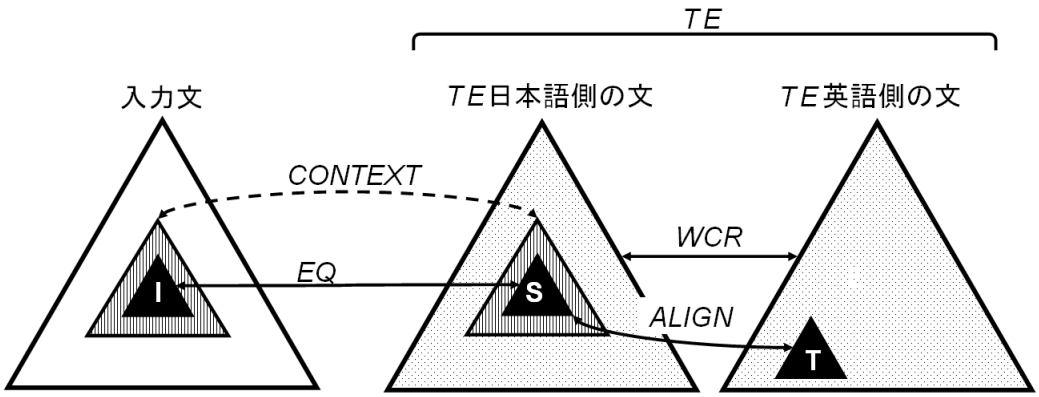

図 6　TE の選択

よぶことにする．この時，この TE 片によって翻訳される日本語部分を I，TE 片の日本語部分を S，TE 片の英語部分を T と表すことにする (図 6)．このような I，S，T の間には，翻訳に用いるための当然のこととして，次の条件がかせられる．以下，これを **TE 片制約**とよぶ．

(1)　　I，S，T は，それぞれ依存構造上で連続している．

(2)　I, Sは，その一番外側の付属表現を除いて，完全に一致する．

(3)　Sの句と Tの句はすべて過不足なく対応している (句アライメントされている)．

ある Iに対して，TE データベース中にこのような条件をみたす TE 片 (S-T) が複数存在することが考えられる．本研究では，できるだけよい S-Tを選ぶために，I-S-Tの総合的な関係を次のように考慮することとした．

(1)　I(=S) の大きさと一致度 (付属表現は一致していない場合もある)

(2)　Iの周りと，Sの周りの類似度．

(3)　Sと Tのアライメントの確信度．

以下の節でそれぞれの具体的計算方法を示す．なお，説明の都合上，Iと Sの句対応の集合を EQ，Iと Sの周り (隣接する句) の類似句の対応の集合を $CONTEXT$，Sと Tの句対応の集合を $ALIGN$と表すことにする．

3.2　日本語対応の類似度

Iと Sの一致度は，その中の句対応 (EQ) それぞれについて次の式によって一致度を計算し，その総和をとった $\sum_{i \in EQ} EQUALITY_i$ とする．

$$EQUALITY_i = \frac{\sum S_{cont} \times 2}{\#_{cont}} + 0.2 \times \frac{\sum S_{func} \times 2}{\#_{func}} \tag{3}$$

ここで，$\#_{cont}$ は句対応に含まれる内容語数，$\#_{func}$ は機能語数，S_{cont} は内容語間の一致度，S_{func} は機能語間の一致度である．S_{cont} と S_{func} の計算方法は表 4にまとめる．

これによって計算される全体の一致度は，基本的に Iの句数 (EQ の句対応数) となるが，活用語，付属語などによって若干変化する．

一方，Iの周りと Sの周りの類似度は，その部分の句対応 ($CONTEXT$) それぞれについて次の式によって類似度を計算し，その総和をとった $\sum_{i \in CONTEXT} SIMILARITY_i$ に定める．

$$SIMILARITY_i = (\frac{S_{cont} \times 2}{\#_{cont}} + 0.2 \times \frac{S_{func} \times 2}{\#_{func}}) \times S_{connect} \tag{4}$$

上式は，基本的には $EQUALITY$と同じ計算を行うが，$S_{connect}$ によって類似部分と一致部分の接続関係を考慮している．ここでいう接続関係とは，KNP の出力する格解析結果と句に含まれる機能語と考え，格解析結果または機能語が同一である場合には類似部分の寄与が大きいと考え $S_{connect} = 1.0$ とし，格解析結果と機能語の両方が異なる場合は $S_{connect} = 0.5$ とする．

例えば，“選挙までの (道のり) / 学校までの (道のり)” は，機能語 “までの” が一致しているため，$S_{connect} = 1.0$ と考える．また，“日本の (努力を) / アメリカが (努力する)” は，機能語は異なるが格解析結果が両方ともガ格となるため，やはり，$S_{connect} = 1.0$ と考える．

最終的に，日本語対応の類似度 ($SIM_{I,S}$) は，上記の 2 つの尺度をたしあせたもので次のように定める．

$$SIM_{I,S} = \sum_{i \in EQ} EQUALITY_i + \sum_{i \in CONTEXT} SIMILARITY_i \quad (5)$$

3.3 日英対応の確信度

日英対応 (*S-T*) の確信度は，次の 3 つの要素から計算される．

まず，*S-T* に含まれる日英対応が，どの程度翻訳辞書で対応しているかを調べる．この情報は，日英対応の内部の整合性ということから，本稿では内的整合性とよぶ．*ALIGN* を構成する日英対応 i の内的整合性の値 (内的整合性$_i$) は，表 4で示される値とする．*S-T* 全体の内的整合性は，その平均をとった次の値とする．

$$\text{内的整合性}_{S,T} = \frac{\sum_{i \in ALIGN} \text{内的整合性}_i}{\#ALIGN} \quad (6)$$

次に，日英対応の周辺に他の日英対応が存在している場合は，その対応はより確かだと考えられる．これは，対応の外部の情報という意味で，外的整合性とよび，次の式で計算する．

$$\text{外的整合性}_i = \frac{\text{日英対応 } i \text{ に隣接し，かつ，他の日英対応に含まれる句数}}{\text{日英対応 } i \text{ に隣接する句数}} \quad (7)$$

S-T 全体の外的整合性は，その平均をとった次の値とする．

$$\text{外的整合性}_{S,T} = \frac{\sum_{i \in ALIGN} \text{外的整合性}_i}{\#ALIGN} \quad (8)$$

最後に，*S-T* が含まれる，TE 全体の確信度を考える．これは，対訳文全体の内容語対応率 (WCR) の値を用いる．

以上の 3 つの要素を考慮して，日英句アライメントの確信度 *CONF(S,T)* を，次のように定義する．

$$CONF_{S,T} = \left\{ w \times \text{内的整合性}_{S,T} + (1-w) \times \text{外的整合性}_{S,T} \right\} \times WCR \quad (9)$$

w は内的整合性と外的整合性のどちらを重視するかの重みである．実験を行なった結果，w=0.8 とした場合が良好な精度を示した．

以上の式 (5) の日本語対応類似度と，式 (9) の日英句アライメント確信度を総合して，最終的に，*I-S-T* のよさを次の式で評価することとする．

$$SIM_{I,S} \times CONF_{S,T} \quad (10)$$

表 4　類似度と確信度の計算パラメータ

S_{cont}	1.1	活用も含めて一致する場合
	1.0	原型が一致する場合
	$0.5 \times S_{ntt} + 0.3$	類似度 S_{ntt} が得られる場合
	0.3	品詞が一致する場合
	0	その他
S_{ntt}	0 - 1	日本語語彙大系 (NTT 1997) を用いて計算する類似度. 単語 w_1, w_2 の類似度を $\frac{2\times(w_1\text{と }w_2\text{のシソーラス上で一致している階層の深さ})}{w_1\text{のシソーラスの根からの階層の深さ}+w_2\text{のシソーラスの根からの階層の深さ}}$ として，次の式によって計算する. $\frac{2\times(\text{入力文と }TE\text{ 日本語側の句に含まれる単語の類似度の総和})}{\text{入力文と }TE\text{ 日本語側の句に含まれる単語数}}$
S_{func}	1.1	活用も含めて一致する場合
	1.0	原型が一致する場合
	0	その他
内的整合性$_i$	1.0	日英対応 i のすべての内容語が翻訳辞書で対応している場合
	0.5	日英対応 i の一部の内容語が翻訳辞書で対応している場合
	0	日英対応 i に翻訳辞書で対応がとれる内容語がない場合

3.4　TE 片の探索アルゴリズム

入力文の各句に対して (その句を P とする)，最もスコアの高い TE 片を選択する．この探索は，次のアルゴリズムを用いる．

(1)　P と一致し，さらにその先の日英句アライメントのとれている，(すなわち 3.1節の TE 片制約を満たす)TE 片を TE データベースから取り出す.

(2)　P に対して，多くの TE 片がデータベースから発見される場合 (現在，6 つ以上の場合としている) は，高速化のための枝狩りを行う．これは，P と TE 片の前後 3 文節の語を比較し，次の枝狩り用スコアの最高値の 0.7 倍を閾値として行う.

$$\text{枝狩り用スコア} = (\text{一致する内容語数}) + (\text{一致する機能語数}) \times 0.2$$

(3)　得られた各 TE 片について，P の周りに一致する句があるか，それらが日英対応のとれたものであるかを順に調べ，TE 片制約を満たす最大の I の範囲を得る．そして，この I-S-T のスコアを計算し，スコア最大の TE 片を P に対する TE 片とする.

(4) Pに対してTE片が得られない場合には，その内容語列を翻訳辞書で辞書引きし，得られた語をTEと同様に扱う．句の中に内容語が複数ある場合，辞書引きは句の前方から最長一致法で行う．また，翻訳辞書に複数の訳語がある場合は，ニュースコーパスでの頻度の高い語／句を採用する．

このように入力文の各句に対してTE片を選択した上で，次に入力文全体をカバーするTE片の集合を選択する．本来的には，S-Tの関係も含め全体の整合性を考えて最適解を探索する必要があるが，現時点では貪欲法を用いている．すなわち，スコアの高いTE片から順に採用する．後から採用されるTE片について，それに対応するIの一部がすでに別のTE片によってカバーされている場合には，その部分は翻訳には利用しない．

図5中央に，このようなアルゴリズムによって選択されたTEとTE片の例をあげる[1]．

3.5 翻訳文の生成

選択されたTE片の英語句同士を結合して英語文の依存構造を作成する．このとき，TE片内の句の依存構造は保存し，TE片間は，対応先の入力文句の親子関係にもとづいて結合する．図5左下にTE片を結合した依存構造の例を示す．

依存構造を表層化する際には適切な語順とする必要がある．この処理は，TE片内の語順は保存し，TE片間の順序は抽象化された規則によって決定する．このモジュールは，現在の翻訳システムではまだ試験的なものである．

活用，冠詞，単数-複数の制御などは，現在の翻訳システムではまだインプリメントされていない．

4 実験と考察

評価のために，30文ペアを試験文とし，TEデータベース構築の際には利用せず，その日本語側を入力文として翻訳を行い，英語翻訳文を参考にして人手で評価を行なった．評価は，入力文の句単位で，適切な訳語／句が得られているかどうかを判定した．この際，接続詞や日付・数字表現 (“三日”，“150人が” など) は評価の対象外とした．この結果，本手法の精度は，82.7％であった．

用例選択のそれぞれの要素がどの程度の効果をもっているか調べるため，次の4つの手法を比較した．

(1) Eq_Context_Align: 提案手法

(2) Eq_Align: 提案手法で類似部分を考慮しないスコアを用いた場合．TE片は，式(10)の代わりに次の式で評価する．

1 図5の一番下のTE片のように，日英間で構造が異なる表現の場合でも，それが一つのTE片によって扱われることにより構造を変換する翻訳が自然に実現されることになる．

表 5 実験結果

	正解	不正解	精度
EQ_CONTEXT_ALIGN	134	28	82.7%
	(121)	(23)	(84.0%)
EQ_ALIGN	132	30	81.4%
	(119)	(15)	(82.6%)
EQ_CONTEXT	112	50	69.1%
	(99)	(45)	(68.7%)
DIC_ONLY	117	45	72.2%

* 括弧内の値は，TE 片が見つかった場合のみの値である (TE 片が見つからなかった場合は翻訳辞書が使用される)

$$\sum_{i \in EQ} EQUALITY_i \times CONF_{S,T} \tag{11}$$

(3) EQ_CONTEXT: 提案手法で日英アライメントの確信度を考慮しないスコアを用いた場合．すなわち，TE 片は式 (5) の $SIM_{I,S}$ だけで評価する．

$$\sum_{i \in EQ} EQUALITY_i + \sum_{i \in CONTEXT} SIMILARITY_i \tag{12}$$

(4) DIC_ONLY: 翻訳辞書で得られた語／句のうち，コーパス中でもっとも頻度の高いものを選んだ場合

この比較実験の結果を表 5に示す．提案手法である EQ_CONTEXT_ALIGN が他の手法よりも高い精度であることから，提案手法が有効であることが分かる．また，日英対応の確信度を考慮しない場合 (EQ_CONTEXT) の精度は低く，翻訳辞書による手法 (DIC_ONLY) と同程度であることから，NHK ニュースコーパスのようなパラレリズムの低いコーパスを扱う際には，日英対応の確信度が重要な尺度となっていることが分かる．

提案手法 (EQ_CONTEXT_ALIGN) と翻訳辞書による手法 (DIC_ONLY) を比較した結果を図 7 に示す．提案手法は，日本語対応の類似度と日英対応の確信度の両方を考慮しているため，自然な訳語選択が可能となっている．例えば，図 7の最上部の例では，日本語対応の類似度が高いため，“(... で) 開かれる” ⇔ “to be held (in ...)” など自然な訳語選択が行えている．

提案手法の不正解を分析してみると，TE 片がデータベース中に少量しか見つからない場合が多い．この場合，見つかった TE 片のスコアが低くても，これらの中から TE 片が採用されることになり，不正解の原因となる．

この問題を解決するためには，日英対応の推定精度を向上させ，誤ったアライメントの数を減らすことが必要であり，今後の課題とする．また，誤った日英対応を含む TE が存在した場合も，それらの TE 片スコアが低いことを考えれば，TE の量が増加すれば自然と精度は向上す

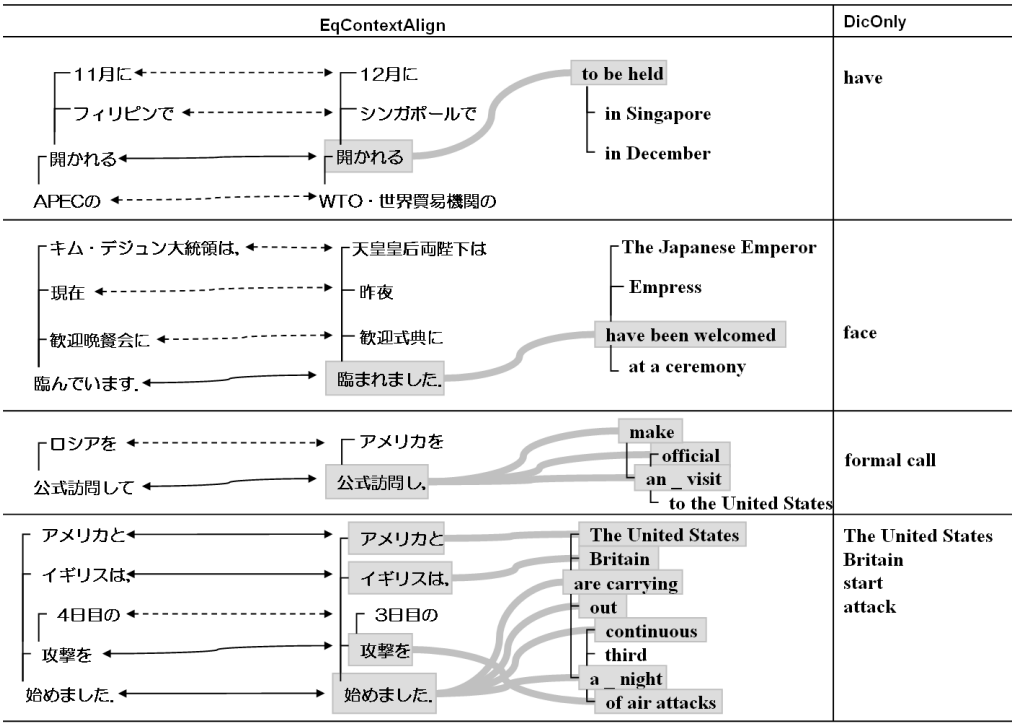

図 7　Word Selection of Dic_Only and Eq_Context_Align

ると考えられる．本研究で扱ったようなコンテントアラインコーパスの量は年々増加しているため，今後，問題の解決は容易になると期待される．

5　関連研究

用例ベース翻訳 (Nagao 1984) が最初に提案されてからしばらくの間は，実験的な翻訳システムによって用例ベース翻訳の可能性を示すという研究が多数行われた (Sato and Nagao 1990; Sadler and Vendelmans 1990; Maruyama and Watanabe 1992; Furuse and Iida 1994)．

最近の研究では，用例ベース翻訳を実用化することに焦点があてられ，制限されたドメインでの翻訳システムの構築が行なわれている．例えば，(Richardson, Dolan, Menezes and Corston-Oliver 2001; Menezes and Richardson 2001) は，マニュアルドメインでの翻訳を行い，(Imamura 2002) は，旅行会話の翻訳を行なっている．これらの研究では，翻訳は入力文と用例の入力文側との間の類似度だけを考慮して，用例を選択している．本手法のように，日英対応の確信度を考慮せずに用例を選択できるのは，これらの研究で扱っている対訳コーパスが高いパラレリズムを持っているからだと考えられる．

一方，本稿ではコンテントアラインコーパスを扱っている．コンテントアラインコーパスは，パラレルコーパスに比べて入手しやすい対訳コーパスであるが，比較的自由な翻訳がなされるため，低いパラレリズムしかもたない．このようなコーパスを扱う場合は，本稿の実験が示すように，日英翻訳の確信度を含めた総合的な尺度を考えることが重要である．

6　結論

本稿では，内容レベルで対応のとれている対訳記事コーパスを用いて，用例ベース翻訳を実現する手法を提案した．この過程で重要な問題のひとつは，適切な翻訳用例を選択することである．本稿では，入力文との類似度と，日英対応の確信度を用いた新しい翻訳用例の選択手法を提案した．

謝辞

本研究は，通信・放送機構の研究受託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．

参考文献

Aramaki, E., Kurohashi, S., Sato, S., and Watanabe, H. (2001). “Finding Translation Correspondences from Parallel Parsed Corpus for Example-based Translation.” In *Proceedings of MT Summit VIII*, pp. 27–32.

Charniak, E. (2000). “A maximum-entropy-inspired parser.” In *Proceedings of NAACL 2000*, pp. 132–139.

Furuse, O. and Iida, H. (1994). “Constituent Boundary Parsing for Example-Based Machine Translation.” In *Proceedings of the 15th COLING*, pp. 105–111.

Imamura, K. (2002). “Application of Translation Knowledgeacquired by Hierarchical Phrase Alignment for Pattern-based MT.” In *Proceedings of TMI-2002*, pp. 74–84.

Kurohashi, S. and Nagao, M. (1994). “A Syntactic Analysis Method of Long Japanese Sentences based on the Detection of Conjunctive Structures.” *Computational Linguistics*, **20** (4).

Kurohashi, S. (2001). “SENSEVAL2 Japanese Translation Task.” In *Proceedings of SENSEVAL2*, pp. 37–40.

Maruyama, H. and Watanabe, H. (1992). “The Cover Search Algorithm for Example-based Translation.” In *Proceedings of TMI-1992*, pp. 173–184.

Menezes, A. and Richardson, S. D. (2001). “A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora.” In *Proceedings of the ACL 2001*

Workshop on Data-Driven Methods in Machine Translation, pp. 39–46.

Nagao, M. (1984). “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle.” Elithorn, A. and Banerji, R. (eds.)： Artificial and Human Intelligence, pp. 173–180.

NTT コミュニケーション科学研究所 (1997). 日本語語彙大系. 岩波書店.

Richardson, S. D., Dolan, W. B., Menezes, A., and Corston-Oliver, M. (2001). “Overcoming the customization bottleneck using example-based MT.” In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pp. 9–16.

Sadler, V. and Vendelmans, R. (1990). “Pilot Implementation of a Bilingual Knowledge Bank.” In *Proeedings of the 13th COLING*, pp. 449–451.

Sato, S. and Nagao, M. (1990). “Toward Memory-based Translation.” In *Proceedings of the 13th COLING*, pp. 247–252.

略歴

荒牧 英治： 1998 年 京都大学総合人間学部基礎科学科卒業．2002 年 京都大学情報学研究科修士課程修了．現在，東京大学大学院情報理工学系研究科博士課程在学中．機械翻訳の研究に従事.

黒橋 禎夫: 1989 年 京都大学工学部電気工学第二学科卒業．1994 年 同大学院博士課程修了．京都大学工学部助手，京都大学情報学研究科講師を経て，2001 年東京大学大学院情報理工学系研究科助教授，現在に至る．自然言語処理，知識情報処理の研究に従事.

柏岡 秀紀: 1993 年 大阪大学大学院基礎工学研究科博士後期課程修了．博士 (工学)．同年 ATR 音声翻訳通信研究所入社．1998 年 同研究所主任研究員 (現 ATR 音声言語コミュニケーション研究所)．1999 年 奈良先端科学技術大学院大学情報学研究科客員助教授．主に自然言語処理，機械翻訳の研究に従事.

田中 英輝: 1982 年 九州大学工学部電子工学科卒業．1984 年 同大学院，修士課程修了．NHK 放送技術研究所，ATR 音声翻訳通信研究所，ATR 音声言語通信研究所を経て，2001 年より ATR 音声言語コミュニケーション研究所，第四研究室室長，現在にいたる．機械翻訳，情報検索，音声認識などの自然言語処理の研究に従事.

(2003 年 7 月 6 日 受付)
(2003 年 10 月 12 日 再受付)
(2003 年 10 月 30 日 採録)