用例ベース翻訳のための対訳文の句アライメント

荒牧　英治[†]　黒橋　禎夫[†]
佐藤　理史[††]　渡辺日出雄[†††]

用例ベース翻訳を実現するためには，大量の用例が必要である．本研究は，対訳文を用例として利用できるようにするために，対訳文に対して句アライメントを行なう手法を提案する．従来の句アライメントでは，語アライメントを得てから，その情報をもとに句アライメントに拡張する手法が方式が多かった．本手法では基本句という文節に相当する単位を導入して，基本句間のアライメントを行なう．実験を行なった結果，良好な結果を得た．

キーワード:　句アライメント，用例ベース機械翻訳，翻訳知識獲得

Phrase Alignment for Example-Based Machine Translation

Eiji Aramaki[†] , Sadao Kurohashi[†] , Satoshi Sato[††]
and Hideo Watanabe[†††]

Example-based machine translation requires a large set of translation patterns. In this paper, we propose a phrase alignment method that aims to acquire translation patterns from bilingual sentense pairs. Most of previous methods employ word alingnment for phrase alingnment. This method uses the basic-phrase as the unit of phrase alingnment, and estimates alignment between basic-phrases. The experimental results show that this method performs well.

KeyWords: *Phrase Alignment, Example-based Machine Translation, Translation Knowledge Acquisition*

1　はじめに

インターネットが急速に広まり，その社会における重要性が急速に高まりつつある現在，他言語のウェブ情報を閲覧したり，多言語で情報を発信するなど，機械翻訳の需要は一層高まっている．これまで，機械翻訳の様々な手法が提案されてきたが，大量のコーパスが利用可能となってきたことにともない用例ベース翻訳 (Nagao 1984) や統計ベース翻訳 (Brown et al. 1990) が主な研究対象となってきている．本稿は前者の用例ベース翻訳に注目する．

用例ベース翻訳とは，翻訳すべき入力文に対して，それと類似した翻訳用例をもとに翻訳

† 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
†† 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University
††† 日本 IBM 株式会社東京基礎研究所, Tokyo Research Laboratory, IBM Research

を行なう方式である．経験豊かな人間が翻訳を行う場合でも用例を利用して翻訳を行っており，この方式は他の手法よりも自然な翻訳文の生成が可能だと考えられる．また，用例の追加により容易にシステムを改善可能である．以上のような利点を持つものの，用例ベース方式は翻訳対象領域をマニュアルや旅行会話などに限定して研究されている段階であり，ウェブドキュメント等を翻訳できるような一般的な翻訳システムは実現されていない．

その実現が困難な理由の一つに，用例の不足が挙げられる．用例ベース翻訳は入力文とできるだけ近い文脈をもつ用例を使うため，用例は対訳辞書のように独立した翻訳ペアではなく，まわりに文脈を持つことが必要である．つまり，用例中のある句が相手側言語のある句と対応するというような対応関係が必要となる．用例ベース翻訳を実現するためには大量の用例が必要だが，人手でこのような用例を作成するのは大量のコストがかかる．そこで，対訳文に対して句アライメントを行い用例として利用できるように変換する研究が 90 年代初頭から行われてきた．

当初は，依存構造や句構造を用いた研究が中心であったが (Sadler and Vendelmans 1990; Matsumoto et al. 1993; Kaji et al. 1992)，構文解析の精度が低いために実証的な成果が上がらなかった．その後には，構造を用いず用例を単なる語列として扱った統計的手法が研究の中心となっている (北村，松本 1997; Sato and Saito 2002)．統計的手法によって対応関係を高精度に得ることは可能だが，そのためには大量の対訳コーパスが必要となる．

近年は構文解析の精度が日英両言語で飛躍的に向上し，再び構造的な対応付けが試みられている．Menezes 等 (Menezes and Richardson 2001) は，マニュアルというドメインで依存構造上の句アライメントを行なっている．今村 (今村 2002) は，旅行会話というドメインで句構造的上の句アライメントを行なっている．これらの先行研究は，限定されたドメインのパラレリズムが高いコーパスを扱っており，一般的なコーパスが用いられていない．本稿はコーパスに依存しない対応付けを実現するために依存構造上の位置関係を一般的に扱い，対応全体の整合性を考慮することにより対応関係を推定する．これは，(Watanabe et al. 2000) を基本句の概念を導入して発展させたものである．

本稿の構成は以下のとおりである．2 章で提案手法について述べる．3 章で実験と考察を述べ，4 章にまとめを付す．

2　提案手法

提案手法は，対訳文を入力とし，両言語に含まれている**基本句** (次節にて定義) 間の対応関係を推定する．提案手法の本質的な部分は対象とする言語ペアに依存しないが，実験は日英間で行なったため，以下の章では日英を対象として手法を説明する．提案手法の大まかな流れは次のようになる．

(1) **基本句を単位とした依存構造への変換**：日英両言語の文を構文解析し，語をまとめることにより，基本句を単位とした依存構造を得る．

(2) **辞書の情報による対応関係の推定**：日英対訳辞書 (以下，**辞書**とよぶ) を利用し基本句対応を推定する．

(3) **未対応句の処理**：辞書の情報では対応のつかなかった基本句 (以下，このような句を**未対応句**とよぶ) を含んだ基本句対応を推定する．

2.1 基本句

日本語では，語 (形態素) という単位の基準が曖昧であり，高精度に自動検出することができる文節が統語解析の単位として一般的に用いられてきた．一方，英語においては，語の基準が明確であるため，文節に相当する単位は通常用いられない．これら両言語の言語の構造を照合する際には，従来，語を単位とした構造照合が行われてきた (Matsumoto et al. 1993; Kaji et al. 1992)．

提案手法では，文節に相当する単位を英語に導入し，両言語の構造を文節に相当する単位で照合する．英語には，文節に相当する単位は存在しないため，本稿では，この単位を**基本句**とよぶ．基本句を対応の単位とすることのメリットは次のようにまとめられる．

(1) 複合名詞などにおける基本句の内部の結びつきは強い結び付きであるので，対応の整合性を調べる際の強い手がかりとできる．

(2) 機能語は各言語固有の振る舞いがあり，それをバラバラに扱うと問題が複雑になってしまう．基本句を単位とすることにより内容語中心の取り扱いとなり，問題が単純化される．

(3) ある対応が周辺と整合的であるかどうかを計算する場合に，まわりとの距離の尺度が必要であるが，単語を単位として考えるよりも，基本句を単位とする方が妥当な尺度となる．

2.2 基本句を単位とした依存構造への変換

まず，対訳文中の両言語の文を統語解析し，その結果を基本句を単位とした依存構造に変換する．これは日英それぞれ次のように行なう．

日本語の文については，KNP(Kurohashi and Nagao 1994) を用いて統語解析を行なう．その結果得られる「(接頭辞＊)(内容語＋)(機能語＊)」という構造を持つ文節を基本句とする (＊は 0 回以上の繰り返し，＋は 1 回以上の繰り返し)．文節と異なる点は,「(〜に) ついて」や「(〜に) おいて」など，機能的表現となっている内容語を直前の文節にまとめることである．ここでいう機能的表現となる文節は，人手で登録した文節パターン約 30 に当てはまる文節と，KNP

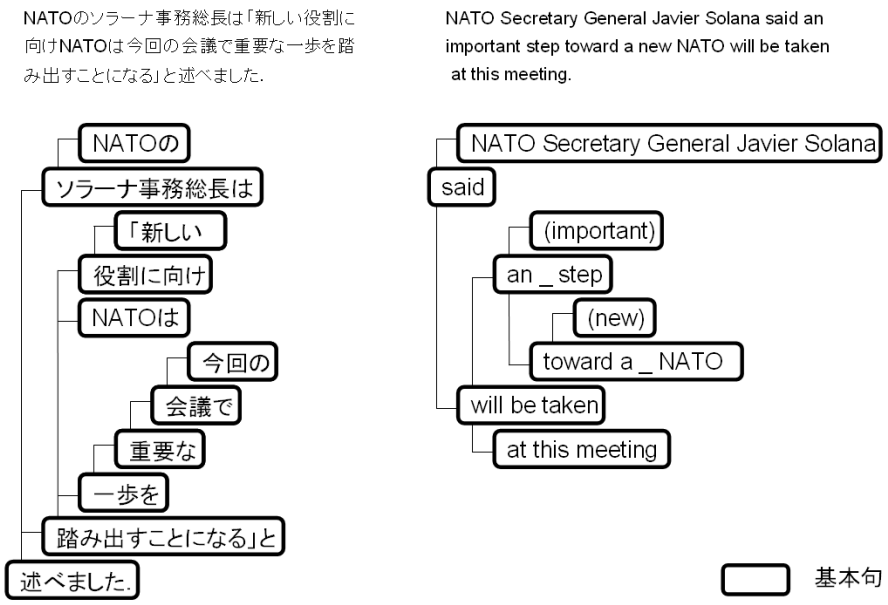

図 1　基本句を単位とした依存構造

の解析結果が＜複合辞＞となる文節とする．

英語の文については，Charniak の統語解析システム (Charniak 2000) を用いて解析を行なう．これは句構造を出力するので，各句構造規則に主辞を定義することにより依存構造に変換する．次に内容語を中心にして，その前後の機能語 (前置詞・冠詞・不定詞など) をまとめる．これには次の規則を用いた．

(1)　複合名詞をまとめる (例：the oil crises)
(2)　機能語を内容語とまとめる (例：at this meeting)
(3)　助動詞や助動詞的表現を主動詞とまとめる (例：had better study)

図 1に，両言語の文を基本句を単位とした依存構造に変換した例を示す．日本語側の「役割に向け」の「向け」や,「踏み出すことになる」と」の「ことになる」が機能的表現であるために直前の文節にまとめられて基本句となっている．

英語側では，(1)〜(3) の規則により，図 1右のように語がまとめられる．日本語側との大きな相違点は，基本句をまとめる際に表層の文字列上でギャップが発生する場合があることである．例えば，“〜 toward a new NATO 〜” は，“toward” と “a” をその係り先である “NATO” にまとめた結果，“toward a NATO’ と “new” に分けて基本句となる．このように基本句でもとの語順が保存されない場合は，本稿中では，“an __ step”, “(important)” のように表記する．

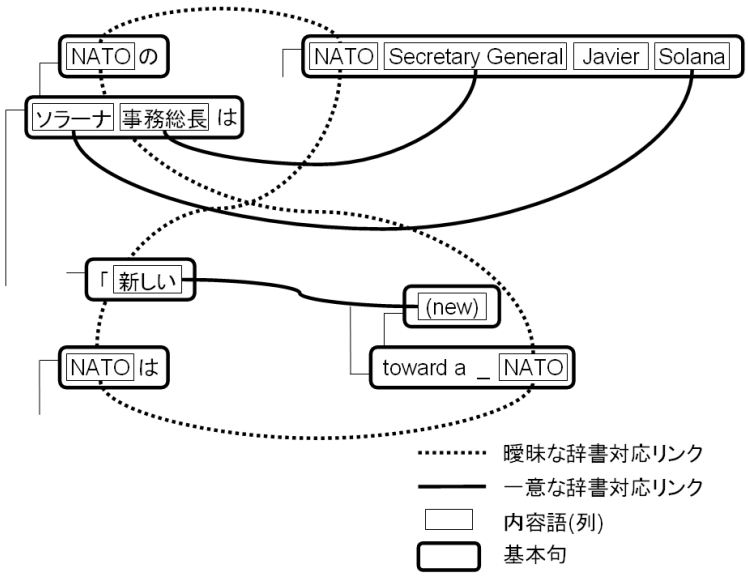

図 2　辞書対応リンク

2.3　辞書の情報による対応関係の推定

2.3.1　辞書対応リンクの付与

辞書 (日本語の見出し語:約 14 万語) を用いて日英の内容語間に対応をつける．辞書には単語間あるいは単語列間 (例:「寝る」⇔“go to bed”，「事務総長」⇔“secretary general” など) の対応関係が記述されており，対訳文中の両言語の語 (列) と一致するものがあれば，対訳文中の両言語の語 (列) 同士をリンク付ける (以下，このリンクを**辞書対応リンク**とよぶ)．両言語とも内容語を含まない対応関係 (例:「で」⇔“at” など) や，片側の言語で内容語を含まない対応関係 (例:「進行中の」⇔“on” など) については，曖昧性が高いために辞書対応リンクを付与しない．

ある語 (列) の訳語が相手側言語に複数存在すれば，語 (列) は複数の辞書対応リンクを持つ．以降，このような辞書対応リンクを**曖昧な**辞書対応リンクよぶ．一方，辞書対応リンクが 1 本だけ付与されている単語 (列) において，その辞書対応リンクを**一意な**辞書対応リンクよぶ．図 2は先の図 1の対訳文に張られる辞書対応リンクの例である．

2.3.2　辞書の情報による基本句対応の推定

次に辞書対応リンクを用いて，基本句の対応 (**基本句対応**) を推定する．これは次のような手続きで行なう．まず，基本句が辞書対応リンクで接続されており，かつ，日英それぞれで依

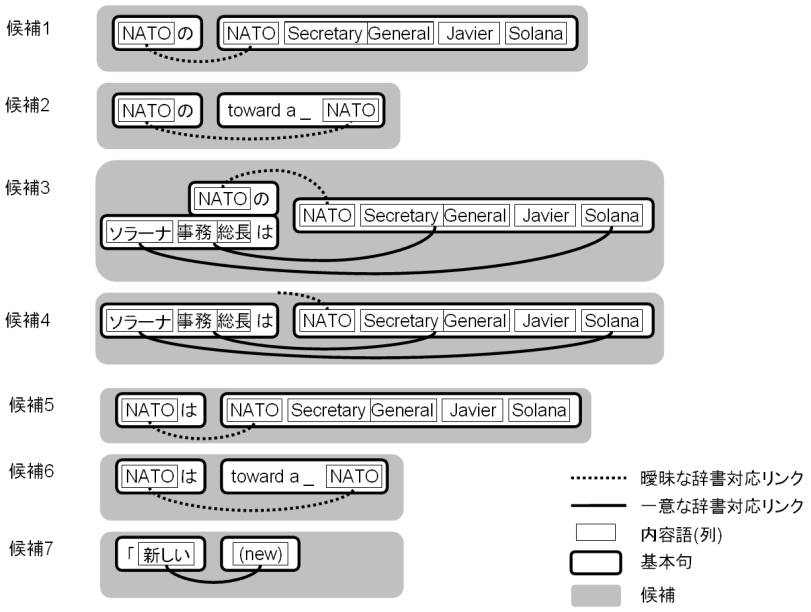

図 3　辞書対応リンクと基本句対応の候補

存関係で接続されていることを条件に，対訳文中からあらゆる基本句対応の候補 (以降，**候補**とよぶ) を生成する．例えば，図 2の例では，7 つの候補を生成する (図 3の候補 1〜7)．

ここで，候補 7 に含まれている語 (列) には，曖昧な辞書対応リンクを持ったものがない．このような候補 (以降，**一意な候補**とよぶ) を採用して基本句対応とする．候補 1〜6 には，曖昧な辞書対応リンクをもった語 (列) が存在する．このような候補 (以降，**曖昧な候補**とよぶ) については次の手順で採用する候補を判定する．

step1: 曖昧な候補のスコア計算
step2: 最高スコアを持つ候補を採用
step3: 採用した候補と重複する候補は削除．候補が残っているなら step1 へ，残っていないなら step4 へ．
step4: 基本句対応の棄却判定

step1 では曖昧な候補にスコアを付与する．スコアは次の 2 つの整合性を用いて以下のように定義する．

内的整合性：対応内部に含まれる辞書対応リンクの整合性
外的整合性：近傍のすでに採用された基本句対応の整合性

$$(\text{基本句対応候補のスコア}) = (\text{内的整合性}) \times C + (\text{外的整合性}) \times (1 - C)$$

C は定数であり，どちらの情報をより重視するかのパラメータである．

内的整合性

基本句対応の候補内に辞書対応リンクを多く持っている場合，その候補は信頼性が高い．そこで，候補内の辞書対応リンクの情報を内的整合性として，以下のように定義する．

$$(\text{内的整合性}) = \frac{D_j + D_e}{W_j + W_e} \times \log(D_j + D_e)$$

ただし，W_j は候補内の日本語の内容語の数，W_e は英語の内容語の数である．D_j は辞書対応リンクを付与されている日本語の内容語の数，D_e は辞書対応リンクを付与されている英語の内容語の数である．

$\frac{D_j+D_e}{W_j+W_e}$ は，候補内の辞書対応リンクの充足の度合いを示しており，候補内のすべての内容語が辞書対応リンクを付与されている場合に最大値である 1 をとる．ここで，すべての内容語が辞書対応リンクで接続されていても，内容語数が 1:1 の候補と，2:2 の候補では，後者の候補の方が候補内の辞書対応リンクがお互いに支持しあっており，信頼性が高いと考えられる．そこで後者を優先するために，$\log(D_j + D_e)$ により重み付けをしている．

例えば，図 3の候補 3 は内容語数が 9 つ (日本語側 4 つ，英語側 5 つ) であり，そのうちの 8 つが辞書対応リンクをもつ．よって，内的整合性は 1.84 ($=\frac{4+4}{4+5} \times \log(4+4)$) となり，候補 1〜6 の中でもっとも高い値となる．

外的整合性

候補の近傍に基本句対応が多く存在すればするほど，その候補は他の基本句対応に支持されており確かだと考えられる．本稿ではこの支持を外的整合性とよぶ．外的整合性は，候補の近傍の基本句のうち，基本句対応に含まれるものの割合を用いて以下のように定義する．

$$(\text{外的整合性}) = \frac{\sum_i (C_i\text{による候補への支持})}{\#(N_j) + \#(N_e)}$$

$$(C_i\text{による候補への支持}) = \begin{cases} \#(C_{ij} \cap N_j) + \#(C_{ie} \cap N_e) & \text{if } \#(C_{ij} \cap N_j) > 0 \text{ かつ} \#(C_{ie} \cap N_e) > 0 \\ 0 & \text{otherwise} \end{cases}$$

ただし，N_j は候補の日本語側の近傍に存在する基本句の集合，N_e は候補の英語側の近傍に存在する基本句の集合，C_{ij} は基本句対応 C_i の日本語側の基本句の集合，C_{ie} は基本句対応

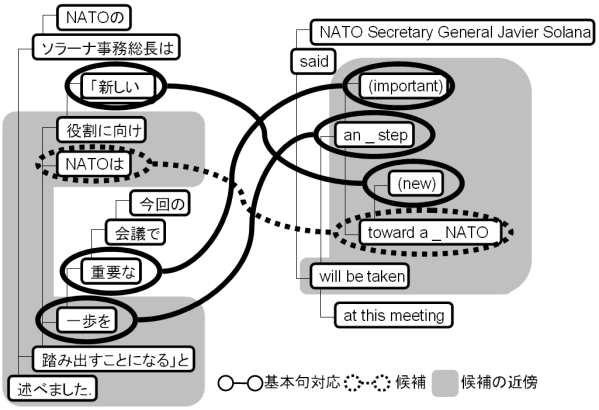

図 4　外的整合性

C_i の英語側の基本句の集合である．#は集合の要素数を示す．**近傍**とは，依存構造上で基本句間の距離が 2 以内に含まれる基本句の集合と定義する[1]．

例えば，図 4で「NATO は」⇔“toward a __ NATO” という候補に注目してみと，日本語側の「NATO は」近傍には 4 つの基本句があり，英語側の “toward a __ NATO” 近傍にも 4 つの基本句がある．この近傍の基本句対応は「一歩を」⇔“an __ step” だけなので，外的整合性は，0.25 $(=\frac{2}{4+4})$ となる．

step2 では，曖昧な候補のうち最も高いスコアを持つものを採用し，候補を基本句対応とする．この際，基本句対応を次の 4 つに分類しておく．この分類は後の処理で利用する．

(1)　**充足対応**: 内的整合性が 1 である基本句対応．この基本句対応は対応内のすべての内容語が辞書対応リンクを付与されている．

(2)　**日本語過剰対応**: 基本句対応内の英語側の内容語はすべて辞書対応リンクが付与されており，日本語側の内容語の一部が辞書対応リンクを付与されていない基本句対応．

(3)　**英語過剰対応**: 基本句対応内の日本語側の内容語はすべて辞書対応リンクが付与されており，英語側の内容語の一部が辞書対応リンクを付与されていない基本句対応．

(4)　**不安定対応**: 基本句対応内の日英いずれの言語側にも辞書対応リンクを付与されていない内容語が含まれている基本句対応．

step3 では，基本句対応が持つ基本句と候補が持つ基本句が重複していれば，その候補を削除する．例えば図 2の下例では候補 3 を採用すると，候補 1〜2,4〜6 は削除する．まだ候補が残っているならば，次の候補を採用するために step1 に戻る．候補が存在しないならば，step4

1 近傍を依存構造上で基本句間の距離 1〜3 と変化させて実験を行ったが，距離 2 が高い精度を示した．よって，ここでは依存構造上で基本句間の距離 2 を近傍とした．どの範囲を近傍と扱うのがよいかは対訳コーパスに依存し，本稿では詳細に取り扱わなかった．

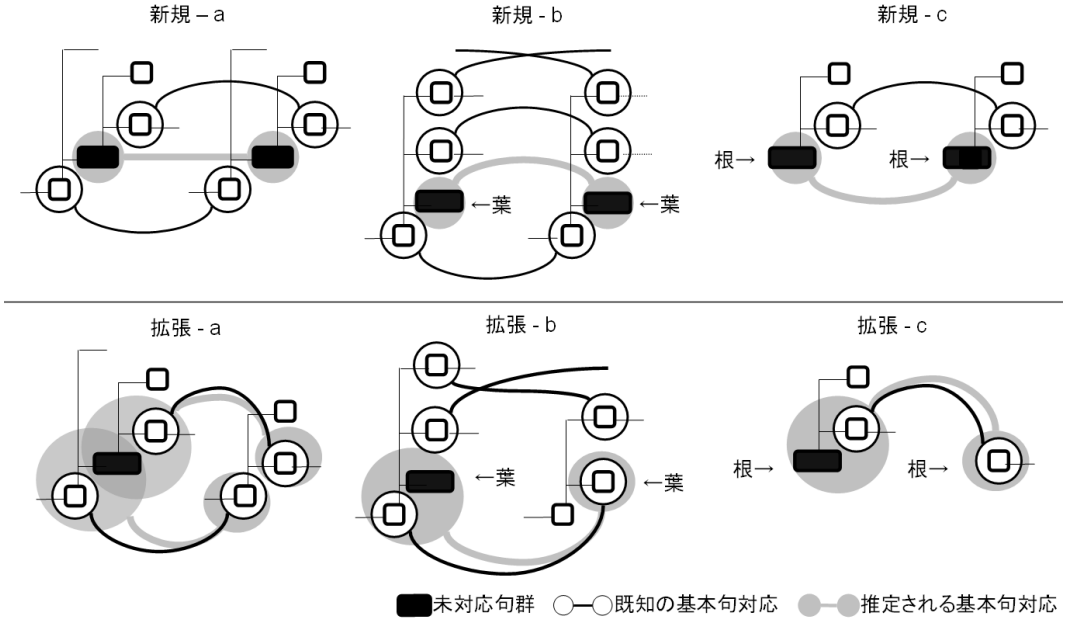

図 5　未対応句の推定規則

に移る．

step4 では，すべての基本句対応について外的整合性を計算し，外的整合性が 0 であれば，その基本句対応を棄却する．この処理を行なう理由は次のようになる．

先に一意な候補は外的整合性がなくても無条件に採用した．また，曖昧な候補についても，内的整合性が高ければ外的整合性がなくても採用されうる．しかし，外的整合性が小さい基本句対応は誤っている可能性が高く，ここで外的整合性の有無を手がかりとして基本句対応の棄却を行なう．

2.4　未対応句の処理

先の処理で推定された基本句対応だけでは，対応付けられていない基本句が残る場合がある．このような未対応句を対応付ける場合，次の 2 通りの可能性がある．

(1)　未対応句同士を基本句対応とする (以降，**新規対応**とよぶ)

(2)　未対応句をすでに推定された基本句対応に含めて新たな基本句対応とする (以降，**拡張対応**とよぶ)

そこで，コーパスを調査して図 5のような新規対応と拡張対応を推定する規則を作成した．未対応句の対応付けは規則に適合した場合に行なう．

新規対応の規則は，日本語側の n 個の未対応句 (以降，**未対応句群**とよぶ) と英語側の m 個

の未対応句群が，それぞれの親方向，子方向が対応付けられている，あるいは端点となっている場合に，それらを基本句対応とする．ただし，未対応句群に含まれる未対応句は依存構造上で連続していることを条件とする．

新規-a: 両言語の未対応句群の親同士が対応付けられていおり，子同士が対応付けられている場合．ただし，子が複数ある場合は，いずれかの子同士が対応付けられていればよいものとする．

新規-b: 両言語の未対応句群がともに依存構造上の葉である場合，両言語の未対応句群の親同士が対応付けられており，かつ，兄弟がすべて基本句対応に含まれている場合．ただし，兄弟同士が対応付けられている必要はない．

新規-c: 両言語の未対応句群がともに依存構造上の根であり，子同士が対応付けられている場合．ただし，子が複数ある場合はいずれかの子同士が対応付けられていればよいものとする．

拡張対応も新規対応と同様に，未対応句群の親方向，子方向が対応付けられている，あるいは端点となっている場合に未対応句群の対応付けを行なうが，未対応句群が片方の言語側だけに存在する点が異なる．ここで，基本句対応を不適切に拡張対応としてしまわないように，拡張対応は次の 2 つの条件のいずれかを満たすものとする．条件のいずれかを満たした上で以下の規則に適合すれば拡張対応を推定する．

条件 1: 英語過剰対応の日本語側に未対応句群が拡張される場合．または，日本語過剰対応の英語側に未対応句群が拡張される場合

条件 2: 不安定対応に未対応句群を拡張する場合

拡張-a: 2 つの基本句対応が，片方の言語側で依存構造で接続されており，他方の言語側では未対応句群を挟んで接続されている場合に，未対応句群をいずれかの基本句対応に加える．どちらの基本句対応を拡張対応とするかの判定は，条件 1 の基本句対応を条件 2 の基本句対応よりも優先する．また，条件 1 同士や，条件 2 同士の判定は，内的整合性の低い基本句対応を優先する．

拡張-b: 未対応句群が依存構造上の葉である場合，親と兄弟すべてが基本句対応に含まれており，親の対応先が葉であれば，未対応句群を親の基本句対応に加える．

拡張-c: 未対応句群が依存構造上の根である場合，子が基本句対応に含まれており，かつ，子の対応先が根であれば，未対応句群を子の基本句対応に加える．

規則に適合する未対応句群が存在しなくなると，未対応句の処理は終了する．

	翻訳用例コーパス	辞書用例コーパス	白書コーパス
日本語の平均文字数	8.18 文字	12.4 文字	51.7 文字
英語の平均語数	4.98 語	6.0 語	21.4 語

表 1　各コーパスの文の長さ

3　実験と考察

3.1　コーパスと実験環境

実験には以下の 3 種類の対訳コーパスを使用した．

(2) **翻訳用例コーパス**: SENSEVAL2 の translation task(Kurohashi 2001) にて作成されたもので，文以下のサイズの対訳表現からなる．
例:「私はそれについて多くを知らない.」 / “I do not know much about it.”

(1) **辞書用例コーパス**: 短文であり，平易な表現が多い．
例:「危なくて仕方がない」 / “to be nothing but danger”

(3) **白書コーパス**: 科学技術庁及び経済の白書．文長が長く，専門用語が多く含まれている．
例:「年 1 回，過去 12 回開催され，我が国は第 6 回より参加している.」 / “ The conference has been held annually for 12 years, and Japan has participated since the 6th meeting.”

それぞれのコーパスの平均文長を表 1に示す．実験にあたっては 3 つのコーパスから，それぞれ 100 対訳文ずつ計 300 対訳文を無作為抽出した．使用した統語解析システムは日本語においては KNP(Kurohashi and Nagao 1994)，英語においては Charniak の nl-parser(Charniak 2000) である．内的整合性と外的整合性の比であるパラメータ C は 0.2 とした[2]．未対応句の処理で未対応句群を扱う個数は日英とも 1 とした．

3.2　基本句対応の評価

システムが推定した基本句対応を評価するために，正しい対応関係を内容語単位で作成した (以降，この対応を**内容語対応**とよぶ)．これは，内容語を基本句にまとめる規則が変化した場合も評価を可能にするためである．具体的な記述例を以下に示す．

対訳文 1「主要国の科学技術政策動向」 / “Trends among the major countries”
内容語対応 (1) 主要 ⇔ major, (2) 国 ⇔ countries, (3) 動向 ⇔ Trends

対訳文 2「可能性 は 限りなく ゼロ に 近い」 / “It is almost impossible”

2 実験の結果，この値がもっともよい精度を示した．

対応のサイズ	翻訳用例コーパス	辞書用例コーパス	白書コーパス
1:1	205 (81 %)	303 (84 %)	637 (80 %)
2:1	23 (9.1 %)	23 (6.4 %)	79 (9.9 %)
2:2	3 (1.1 %)	10 (2.7 %)	34 (4.2 %)
それ以上	21 (8.3 %)	23 (6.4 %)	42 (5.3 %)
合計	252	359	792

表 2　内容語対応のサイズ

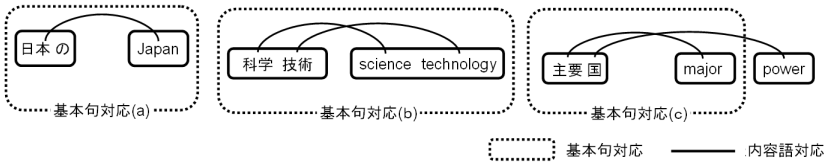

図 6　内容語対応による評価

内容語対応 (1) 可能性 は 限りなく ゼロ に 近い ⇔ It is almost impossible

対訳文 2 のように個々の内容語のレベルでは対応がつかない場合は，n 語:m 語の内容語対応を記述した ($n \geq 1$, $m \geq 1$)．作成された内容語対応のサイズを表 2に示す．各コーパスで 8 割以上が 1 語:1 語の対応となった．

評価は内容語対応を用いて情報検索と同様に適合率と再現率の 2 つの尺度で行なった．ただし，出力は基本句対応であるのに対して正解は内容語対応なので，適合率は基本句対応の適合率，再現率は内容語対応の再現率とした．

(基本句対応) 適合率は以下のように定義した．

$$\text{(基本句対応) 適合率} = \frac{\text{(内容語対応を完全に含んでいる基本句対応の数)}}{\text{(システムが推定した基本句対応の数)}}$$

例えば，図 6の 3 つの基本句対応のうち基本句対応 (a) と (b) は内容語対応を完全に含んでいるが，基本句対応 (c) は 1 つの内容語対応を含んでいない．よって，3 つの基本句対応の適合率は 0.66(=2/3) となる．この定義では大きなサイズの基本句対応を推定すれば，適合率が高くなる．しかし，表 6の出力例が示すように提案手法は不当に大きなサイズの対応を推定する性質を持っていない．

(内容語対応) 再現率は以下のように定義した．

$$\text{(内容語対応) 再現率} = \frac{\text{(基本句対応に完全に含まれている内容語対応の数)}}{\text{(内容語対応の数)}}$$

	翻訳用例コーパス	辞書用例コーパス	白書コーパス
(基本句対応) 適合率	82.2％ (134/163)	90.6％ (253/279)	92.8％ (454/489)
(内容語対応) 再現率	81.7％ (206/252)	86.3％ (310/359)	76.7％ (608/792)

表 3　コーパスと精度

	翻訳用例コーパス	辞書用例コーパス	白書コーパス
辞書対応	91.5％ (108/118)	95.7％ (199/208)	94.6％ (421/445)
拡張対応	76.2％ (48/63)	76.2％ (32/42)	80.0％ (40/50)
新規対応	66.7％ (20/30)	73.8％ (31/42)	72.7％ (16/22)

表 4　基本句対応の分類と適合率

例えば図 6では，5 つの内容語対応のうち 4 つだけが基本句対応に含まれており，再現率は 0.8(=4/5) となる．

提案手法の (基本句対応) 適合率と (内容語対応) 再現率は表 3に示す．また，見つかった基本句対応を次の 3 つに分類し，それぞれの適合率を調べた結果を表 4に示す．

辞書対応: 未対応句が含まれない基本句対応．

拡張対応: 辞書対応を未対応句によって拡張した基本句対応．

新規対応: 未対応句同士の基本句対応．または，それを拡張したもの．

拡張対応と新規対応の精度は辞書対応に比べて低いが，再現率をあげるために重要である．対訳コーパスは一言語のコーパスと比べて量が少なく貴重であるため，再現率の高さは重要である．

3.3　基本句対応についての考察

コーパスには，しばしば対応すべき内容の表現が異なっていたり，依存構造が異なっている対訳文が含まれる．例えば，白書コーパスでは，図 7のように日本語側の「わが国」が英語側で “Japan” と訳されている．このように，対応すべき表現の対応関係が辞書で得られないような現象を，ここでは表現の異なりとよぶ．また，この対訳文では，日本語側が「わが国を取り巻く国際的状況は 〜 問題をはらんでいる」と対応する部分が，英語側で “Japan is confronting 〜 in the international arena.” となっており,「わが国」と “Japan” の係り先が異なる．このような対応すべき表現の係り先が異なる現象を構造の異なりとよぶ．

表現の異なりと依存構造の異なりのいずれか一方だけが起こっている場合には，提案手法は対応関係を正しく推定できる．例えば，図 8の対訳文では構造が一致しているため，表現が異

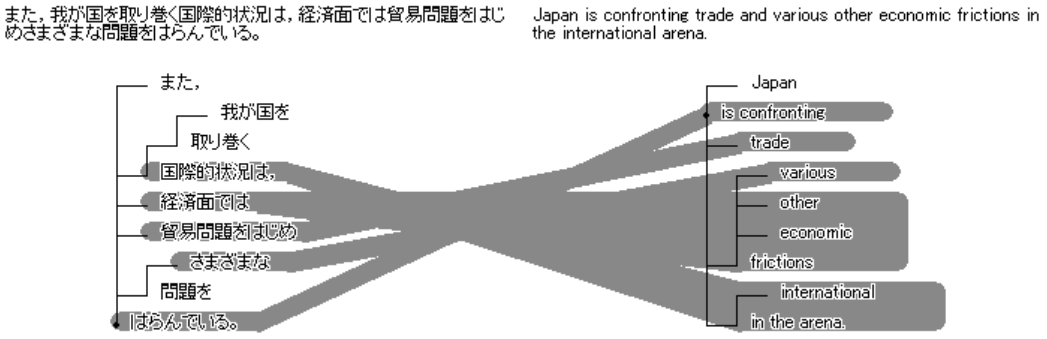

図 7　構造の異なりと表現の異なり

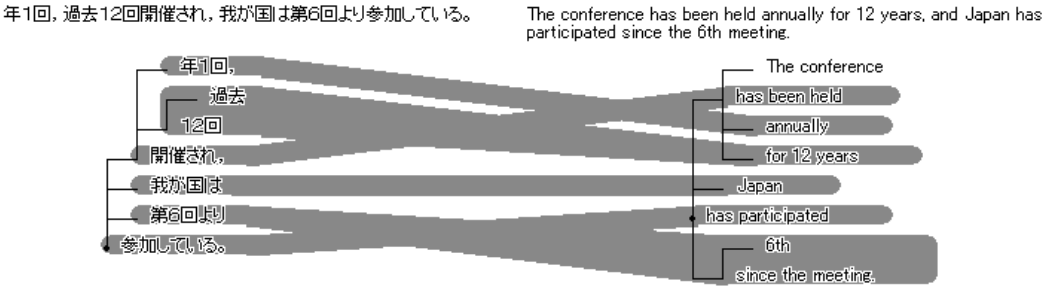

図 8　表現の異なり

なる「わが国」と “Japan” の対応関係を推定できている．

一方，表現の異なりと構造の異なりが同時に起こっている場合は，対応関係を正しく推定できず，誤りの主要な原因となっている．これは，提案手法では，表現の異なる部分 (未対応句) の対応関係を構造を手がかりとした規則で処理しているからである．提案手法だけではこの問題の解決は困難であるが，白書コーパスでは「わが国」が “Japan” と訳される頻度は多く，提案手法に重み付きダイス係数などの統計量を用いることで，ドメイン特有の表現の異なりをある程度吸収できると考えられる．また，構造の異なりが統語解析結果の誤りによって引き起こされる場合がある．この場合は文献 (Matsumoto et al. 1993) にて両言語の解析結果を照合し，適切な統語解析を選択するという手法が提案されており，提案手法に導入することで精度の向上が期待できる．

	翻訳用例コーパス	辞書用例コーパス	白書コーパス
提案手法	100％ (5/5)	70.5％ (12/17)	94.4％(170/180)
ベースライン	50.0％(2/4)	76.9％(10/13)％	89.9％(143/159)

表 5　辞書対応リンクの精度

3.4　辞書対応リンクの曖昧性解消の評価

本節では，2.3 節で述べた辞書対応リンクの曖昧性を解消する部分のみの精度を調べ，考察した結果を述べる．精度は，基本句の単位を導入した場合 (提案手法) と，語を単位とした場合 (ベースライン) の精度を比較した．

ベースラインでは，語を単位とした依存構造上で近傍 (4 語以内) に存在するすでに採用された他の辞書対応リンクの多いものを採用することによって，辞書対応リンクの曖昧性の解消を行なう．近傍を 4 語以内としたのは，提案手法とほぼ同じ範囲の情報を用いるためである（提案手法は近傍の 2 基本句の情報を用いており，1 つの基本句は約 2 語から構成されている）．また，提案手法が，外的整合性が 0 となる辞書対応リンク (近傍 2 基本句以内に他の辞書対応リンクが存在しない辞書対応リンク) を採用しないように (2.3 節の step4)，ベースラインも近傍 4 語以内に他の辞書対応リンクが存在しない場合は採用しないものとする．

評価は，採用した辞書対応リンクのうち曖昧性のあるものが，人手による内容語対応が一致していれば正解とし，そうでない場合は不正解とした．実験の結果は表 5のようになった．

翻訳用例コーパスや辞書用例コーパスでは，文長が短いため曖昧な辞書対応リンクは少数であり，精度の違いについて有意な議論はできない．一方，文長の長い白書コーパスでは曖昧な辞書対応リンクの数は多く，曖昧性解消が重要な問題となっている．この白書コーパスにおいて，提案手法の精度は 94.4％であり，ベースラインの精度の 89.9％よりも高い．ベースラインと提案手法は依存構造のほぼ同じ範囲の対応情報を利用しているため，提案手法の精度が高い理由は基本句という単位を導入した効果と考えられる．

4　まとめ

本稿は，句アライメントの推定において基本句の概念と辞書を用いた新しい手法を提案した．本手法が解決すべき問題は次の 2 つにまとめられる．

(1) 辞書による対応に曖昧性があった場合に，その曖昧性を解消する問題

(2) 辞書で対応つかない場合に，対応関係を推定する問題

(1) に関しては，基本句の導入で高い精度を得ることができ，また，精度も高いことから問題は解決したと考えられる．

(2) に関しては，十分な精度は得れなかった．しかし，人手による修正が可能な範囲の精度

辞書対応	
示さ, れて, いる	is,indicated
使命, である	is,an,mission
科学, 技術, 局, が	the,Office,of,Science,and,Technology
政府, の, 施策, の	government,policy
世界, 規模, で	on,a,global,scale
踏まえ,、	Based,on
国際, 協力, へ, の	toward,international,cooperation
先進国, 間, の	among,advanced,countries
政策, 担当, 者, を	The,policy,makers
新規対応	
全, 要素, 生産性, が	of,TFP
先進, 7 , カ国, の	of,the,G7,nations
策定, に, 加え,、	In,addition,to,the,formulation
発足, した	came,into,power
上げる, こと, だけ, で, なく,、	is,not,only,to,improve
終わった	are,over
ただちに	lost,no,time
かって, もらった	have,cut
拡張対応	
冷戦, 終結, 後, の, 世界, に, おいて, は	in,the,post,Cold,War,years
雇用, 創出, に, おいて	and,job,creation
有形, 固定, 資産, 購入, 費, の	of,expenditures,on,tangible,fixed,assets
転換, 期, に, おける	during,the,period,of,transition
グローバル, 化, の, 進展, の, 中, で,、	Amid,globalization
輸送, 用, 機械, 工業, の, 出超, は	The,surplus,in,the,transport,equipment,industry
勉強, し, さえ, すれば, よい	have,to,study
それ, を, して, しまう, でしょう	will,have,done,it
健康です	in,good,health
まだ, 有効だ	still,holds,good

表 6　基本句対応の具体例

であり，また，本手法に統計量等の手がかりを導入することで，今後精度を上げることが可能だと考えられる．

参考文献

Brown, P. F., Cocke, J., Pietra, S. A. D., cent J. Della Pietra, V., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). “A Statistical Approach to Machine Translation.” *Computational Linguistics*, 16 (2).

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

今村賢治 (2002). “構文解析と融合した階層的句アライメント.” 自然言語処理, 9 (5).

Kaji, H., Kida, Y., and Morimoto, Y. (1992). “Learning Translation Templates from Bilingual Texts.” In *Proceedings of the 14th COLING*, pp. 672–678.

北村美穂子，松本裕治 (1997). “対訳コーパスを利用した対訳表現の自動抽出.” 情報処理学会論文誌, 38 (4).

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

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).

Matsumoto, Y., Ishimoto, H., and Utsuro, T. (1993). “Structural Matching of Parallel Texts.” In *Proceedings of the ACL 93*, pp. 23–30.

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.” In *In Artificial and Human Intelligence*, pp. 173–180.

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

Sato, K. and Saito, H. (2002). “Extracting Word Sequence Correspondences with Support Vector Machine.” In *Proceedings of the 19th COLING*, pp. 870–876.

Watanabe, H., Kurohashi, S., and Aramaki, E. (2000). “Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation.” In *Proceedings of the 18th COLING*, pp. 906–912.

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．

佐藤 理史: 1983 年京都大学工学部電気工学第二学科卒業．1992 年同大学院修士課程修了．現在，京都大学大学院情報学研究科知能情報学専攻助教授．

渡辺 日出雄: 1984 年京都大学工学部電気工学第二学科卒業．1986 年同大学院

修士課程修了. 京都大学工学博士. 1986 年日本アイ・ビー・エム株式会社に入社，現在同社東京基礎研究所にて専任研究員及びグループリーダー. 機械翻訳や自動要約などの自然言語処理研究に従事.

(2003 年 1 月 15 日 受付)
(2003 年 3 月 13 日 再受付)
(2003 年 4 月 30 日 採録)