日英結合価パターン対辞書の一般利用者向け作成支援処理

白井 諭[*1] 横尾 昭男[*2] 奥山 信輔[*3] 河村美砂子[*3] 池原 悟[*4]

[*1]NTT コミュニケーション科学研究所 [*2]ATR 音声翻訳通信研究所
[*3]NTT ソフトウェア [*4]鳥取大学 工学部

1 はじめに

機械翻訳システムを現実の文書に適用するには，翻訳対象に合った利用者辞書が必要である。特に高品質翻訳を狙ったシステムでは，各単語に2,000以上に分類された意味属性の付与が必要であるとされ，一般の利用者による情報付与は困難であった。筆者らは，先に，日本語単語と英語訳語を対にして与えれば，システム辞書を利用して意味属性を推定する方法を提案し，人手による辞書作成の場合と同等の品質となることを報告した[池原 95]。本稿では，結合価パターン対の作成に対して同様の方法を実現するため，簡単な対訳情報の入力により利用者パターン対辞書を作成する方法を提案する。また，形態素解析や日英単語変換で用いる辞書との連携について検討する。

2 日英翻訳システム ALT-J/E の辞書体系

システム辞書は，単語意味辞書，構文意味辞書，日英対照辞書，英語辞書から構成される。単語意味辞書は，単語に文法属性や意味属性を与えるだけでなく，標準表記を設けることにより表記の揺らぎを吸収する役割も担う。以降の辞書の日本語エントリは，すべて標準表記を用いて記述することにより，表記の揺らぎの影響を受けなくて済むという利点が生じるからである。構文意味辞書は，「誰が何をどうする」のような日本語の単文に対する英語の基本構文を与える。日本語の「誰が」や「何を」に具体的にどのような名詞が来うるかは，一般名詞意味属性により指定される。結合価表現による日本語の文型パターンと英語の文型パターンとを対にして保持しているところから，結合価パターン対辞書と

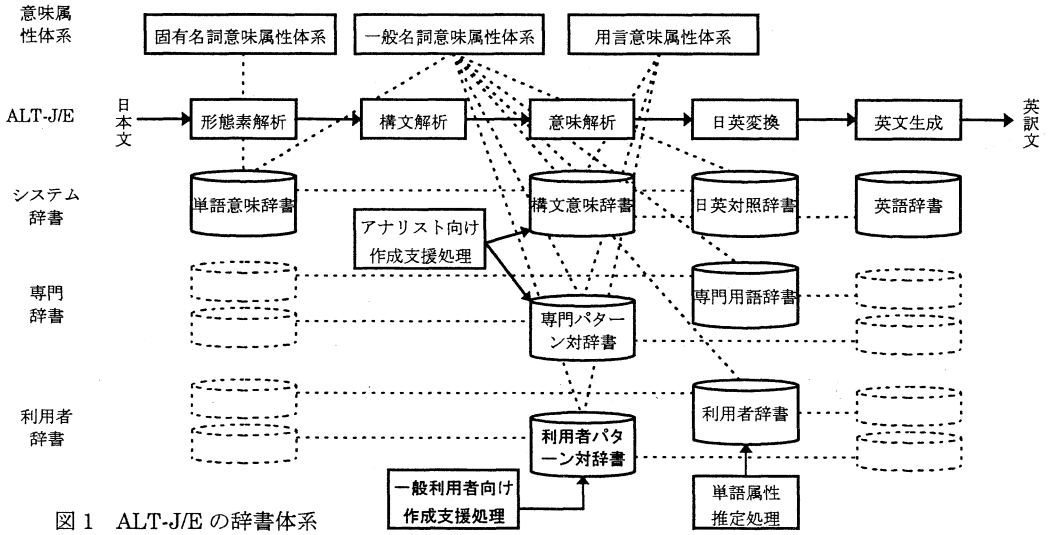

図1 ALT-J/E の辞書体系

も言う。日英対照辞書は，日本語の単語と意味属性の対に対して，英語の訳語表現を与える。英語辞書は，英語の語形変化の情報を与える。

専門辞書は，論理的には構文意味辞書相当の専門パターン対辞書または日英対照辞書相当の専門用語辞書に見せているが，物理的には単語意味辞書相当と英語辞書相当の部分を分けて使用される。

利用者辞書も，専門辞書と同様に，物理的には単語意味辞書と英語辞書を分けて使用される。

これらの辞書を併用すると，利用者辞書，専門辞書，システム辞書の順に優先して適用される。

ALT-J/E で使用される一般名詞意味属性体系は，名詞を約 2,700 種類に分類体系化されている[池原 97]。これを一般利用者が使いこなすのは難しく，単に利用者辞書に訳語対を登録しても却って訳文品質が低下する場合があった。これは既存単語との整合性が崩れるのが主な原因で，一般利用者にも適切な意味属性を容易に付与できるような環境を用意する必要があった。そこで，日英の訳語対を入力にして，システム辞書の類似表現を参考にして意味属性を推定する処理を実現したところ，アナリストとほぼ同等の精度で意味属性が付与されるようになり，利用者の大幅な負担軽減が可能になっている[池原 95]。

このように，利用者による単語レベルの翻訳品質の改善は可能になったが，適切な構文が選定されない場合には無力である。そこで本稿では，利用者による構文レベルの翻訳品質の改善を可能にすることを狙いとして，結合価パターン対の作成支援処理を提案する。

3 作成支援の考え方

3.1 対訳用例からの作成支援処理

1 つの構文には，格要素，副詞要素，用言，様相表現が含まれ，それぞれが日本語パターンおよび英語パターンを形成しているため，単語レベルの辞書に比べ多種の情報記述が要求される。さらに，格要素は名詞と助詞から構成されるが，名詞に対しては一般名詞意味属性体系を用いて条件指定が行なわれる。

システム用の構文意味辞書として必要な規模は 25,000 パターン程度と予想され[白井 95]，1997 年末現在で 16,600 パターンを収集している。初期は和英辞書を参照してパターン対を作成したが，それだけでは機械翻訳に十分なパターン対が確保できないことがわかったため，動詞や形容詞に対する日英対訳例文を網羅的に収集する[白井 97]とともに，収集した対訳文に対する作成支援処理[白井 96]を開発することにより効率化を図っている。

この作成支援処理は，アナリストの使用を前提としたものである。アナリストによるパターン対の作成は次のようなステップに分けることができる。

①パターン対の作成の必要性の分析
②パターン対作成用の対訳用例の適正化
③日英パターン対の記述
　③-1 日英の文構造の分析と文要素間の対応
　③-2 パターン対に記述すべき文要素の選択
　③-3 格要素の名詞の意味属性による抽象化
　③-4 既存のパターン対との優先関係の調整
④記述したパターン対の適用性の確認
⑤辞書全体としての整合性の検証

①と②は高度な知的判断を伴う作業であるため自動化は難しい。対訳例文を収集し，対訳例文の日本語に対する日英翻訳システムの翻訳結果と対訳例文の英語を比較することにより，①の判断はかなり軽減することができた。③-1 と③-2 は ALT-J/E の解析処理と連携させることにより半自動化した。③-3 と③-4 および④は支援処理により効率化された。なお，⑤は辞書全体の情況を見渡す必要があるため，現時点では効率的な支援方法は分かっておらず，アナリストの経験に依存している。

3.2 問題点

次に，前節の作成支援処理を一般利用者が利用す

る場合の問題点について検討する。

①は「訳せない」が動機となるので，必要性は容易に認識されると思われる。しかし，パターン対を追加するか，既存のパターン対を修正するかを，アナリストのように的確に判断するのは難しいため，余計なパターン対を作る恐れが生じる。

②は，トレーニングデータにより事前準備をしている場合を除いて，基本的に難しい。

③-1と③-2は半自動化されているとはいえ，最終的な判断には言語学的なセンスが要求される。③-3と③-4には支援処理があるといっても，アナリストの豊富な経験を前提にしたものに過ぎない。一般利用者が使いこなすのはいずれも困難である。

このように，一般利用者によるパターン対の記述内容自体に多くの問題が予想されることから，④や⑤は辞書の整合性を保つ手段として期待できない。

3.3 一般利用者向け作成支援の基本方針

前節の検討から，一般利用者がパターン対を容易に作る仕掛けを提供するだけでなく，それが翻訳結果に適切に反映されるとともに，他への悪影響が出ないように翻訳システムの処理と辞書の連携を図ることが必要である。

そこで，一般利用者には対訳表現を極めて単純な形式で入力してもらい，あまり抽象化を行なわないでパターンを作り，そのパターン対を最優先で適用することを基本方針として考える。すなわち，対訳表現を入力してもらうことにより①と②を回避し，その入力が単純な形式であることから③-1と③-2が解消し，あまり抽象化しないことにより③-3を不要にするほか，思わぬ表現に適用されてしまう副作用を防止し，作られたパターン対を優先適用することにより③-4以降を軽減する。

4 一般利用者向け作成支援処理

単純な形式の入力として，一般利用者には「誰が／何を／どこで／どうする」と“GA do/did/done WO at DE”のような表現対を指定してもらう。ここで，GAやWOやDEは日本語と英語の格要素を結びつけるためのもので，処理を簡略化する効果があり，利用者にもそれほど負担にはならない。また，do/did/done は英語の不規則変化を示すためのもので，通常は不要である。以下では，このような入力に対するパターン対の自動作成方法を検討する。

4.1 品詞推定

日本語表現の末尾により，述語の種類を判定し，単語意味辞書相当の辞書への登録情報を生成する。この情報が既存の単語意味辞書に未登録の場合，利用者パターン対辞書に対応した単語意味辞書を作成する。

表1　日本語の品詞の判定

日本語末尾	品詞判定	単語意味辞書への登録
する	サ変動詞	スルを除いた部分
だ	形容動詞	語幹部分
い	形容詞	語幹部分
ウ段	動詞（詳細は略）	語幹部分

4.2 日本語パターンの格要素

ガ，ヲ，ニ，カラ，ヨリ，ヘ，マデ，ト，デのほか，ノを検出することにより，格要素を認定し，それぞれの名詞と助詞を分離する。ノは「何の何がどうだ」のような場合に有効に働く。

また，一般利用者が「何はどうだ」のようにハ格を指定した場合，また「何は何がどうだ」のようにハ格とガ格を同時に指定した場合は表2に従う。

表2　ハ格とガ格

利用者の指定	登録する格
ハ格とガ格	それぞれをガ格とヲ格とする
ガ格とハ格	それぞれをヲ格とガ格とする
ハ格（他にガ格なし）	ガ格を登録する
（以外）	対象外とする

4.3 名詞の抽象化

前節における格要素の分析により得られた名詞を対象に単語意味辞書を検索し，一般名詞意味属性を取り出す。取り出した意味属性を，一般名詞意味

属性体系の上から3段目（部分的には4段目）までの70属性に一律に抽象化する。これは適切な抽象化ではない場合もしばしばあるが，名詞句解析はこのレベルで行なわれているので，ある程度の整合性は期待される。なお，作成したパターン対が選択されない場合は，この抽象化に問題がある場合が多いので，名詞を字面指定に変更することもできる。

4.4 英語パターン

英語パターンとして600種類ほど使用されるが，70%までが9種類，80%までが18種類，90%が51種類でカバーされている[池原 97]。ここでは90%をカバーする51種類と，専門パターン対における英語パターンの偏り[横尾 94]を考慮し，表3の50種類を認定対象として選定した。

利用者が指定した英語表現をこれらと照合することにより，英語表現に含まれる単語の品詞を推定し，それらのうち既存の英語辞書に未登録の場合，利用者パターン対辞書に対応した英語辞書を作成する。

表3　英語パターン一覧

there be N pr N	N vi adv	N vt N pr adj noun
N be adj	N vi adv pr N	N vt adv N pr N
N be adj pr N	N vi pr noun	N vt N N
N be adv adj	N vi pr noun pr N	N vt N adj
N be adj adv	N vi pr adj noun	N vt adv noun
N be adv adj pr N	N vi N	N vt adj noun pr N
N be adj adv pr N	N vi adj	N vt adj noun
N be adj pr N pr N	N vi adj pr N	N vt N adv pr N
N be adj pr noun	N vi noun pr N	N vt adj N
N be vpp	N vt N	N vi adv pr N pr N
N be vpp pr N	N vt N pr N	N vt N(subj)'s N
N be vpp pr N pr N	N vt N adv	N vt noun_idiom
N be vpp pr noun	N vt adv N	N vt that_cl
N be vpp adv pr N	N vt N pr noun	N vt pr N that_cl
N vi	N vt noun pr N	N vt N that-節
N vi pr N	N vt noun	N vt N to-不定詞
N vi pr N pr N	N vt noun pr N pr N	

N	名詞の変数	vi	自動詞	vpp	過去完了
noun	名詞の字面	vt	他動詞	be	be 動詞
adj	形容詞	adv	副詞	pr	前置詞
noun_idiom	名詞性の慣用表現				

4.5 利用者パターン対辞書の作成

利用者パターン対辞書は最優先で適用されるため，システム辞書や専門辞書との重複は考慮しない。ただし，利用者パターン対辞書内での重複は検出し，利用者に問い合わせることにより，整合性が崩れるのをある程度防止する。

5 おわりに

日英機械翻訳では，一般に構文レベルの変換失敗は単語レベルの変換失敗に比べ致命傷となりがちである。構文レベルの表現には，単語レベルの場合に比べると，遥かに多種多様の情報を付与する必要があるため，一般利用者が適切な辞書登録を行なうのは容易ではなかった。本稿では，単純な対訳表現の入力を仮定することにより，一般利用者にも可能な結合価パターン対の作成方法を提案した。なお，この方法に基づいた作成支援処理は構築済みで，現在評価中である。

アナリスト向け作成支援処理では，対訳用例からの作成を前提としているが，各種バリエーションを同時に作成する必要がしばしば生じ，その場合には本稿のような作成支援が有効であると考えられる。今後は，アナリスト向け作成支援処理との連携を行ない，さらに操作性を向上させるための検討を並行して進める予定である。

参考文献

[池原 95] 池原,白井,横尾,Bond,小見: 日英機械翻訳における利用者登録語の意味属性の自動推定，自然言語処理, Vol.2, No.1, pp.3-17 (1995)

[池原 97] 池原,宮崎,白井,横尾,中岩,小倉,大山,林: 日本語語彙大系，岩波書店 (1997)

[白井 95] S. Shirai, S. Ikehara, A. Yokoo and H. Inoue: The quantity of valency pattern pairs required for Japanese to English machine translation and their compilation, NLPRS 95, pp. 443-448 (1995)

[白井 96] 白井,上田,兵藤,横尾,池原: 日英機械翻訳のための結合価パターン対の作成支援処理，電子情報通信学会技術研究報告, NLC96-34, pp.25-30 (1996)

[白井 97] 白井,池原,相澤,鳴海,横尾: 結合価パターン対作成のための日英対訳用例文の収集，情報処理学会研究報告, 97-NL-122-1, pp.1-6 (1997)

[横尾 94] 横尾,中岩,白井,池原: 日英機械翻訳用スケルトンーフレッシュ型構文意味辞書の構成，情報処理学会第48回全国大会, 6Q-8, pp.3－139-140