専門分野対応の日英機械翻訳用構文意味辞書の構築

白井 諭† 井上浩子‡ 井田倉紀子‡ 池原 悟† 横尾昭男†

†NTTコミュニケーション科学研究所 ‡NTTアドバンステクノロジ(株)

1 はじめに

解析ベースの機械翻訳では，動詞と名詞の意味的共起を結合価パターンとして記述し，それを原言語と目的言語とで対にしたパターン対の使用が翻訳精度の向上に有効であることが知られている。この方法には，パターン記述精度とパターン対収集の2つの問題がある。

記述精度の問題に関しては，日英機械翻訳では名詞の意味属性を2,000種類以上に分類すれば日本語の用言の訳し分けに必要な精度が確保される[池原93]。収集の問題に関しては，和英辞書や収集可能な用例文のバリエーションの検討などから，パターン対の収集目標は一般的パターン対と慣用的パターン対を併せて25,000件程度収集する必要があること，現状では作成支援処理を強化すれば訓練されたアナリストの人手作業により十分収集可能であることを報告した[白井95b]。以上に基づき，これまでに一般的パターン対12,000件と慣用的パターン対3,000件を収集し，構文意味辞書の完成を目指している。

ところで，実際の文書に機械翻訳を適用するには、文書固有の表現に対応する必要がある。機械翻訳は類似した文書が継続的に出現する分野に適用するのが有利であると考えられるので，利用者は対訳用例を利用しながら必要な表現集を整備すると予想される。従って，この表現集に対して機械翻訳で利用するために必要な情報をどれだけ適切に付与するかが機械翻訳の精度に直結した問題となる。これに関連して，体言性の専門語とその訳語のペアに対して，翻訳システムが持つシステム辞書の情報を参照して単語情報を自動的に付与する方法が提案され，訓練されたアナリストが単語属性を付与する場合に比べ遜色のない精度が確保できることが知られている[池原95]。

しかし，この方法は格要素の共起関係を考慮すべき用言性の専門語には適用できない。訓練されたアナリストも，必ずしも対象とする専門分野に精通しているわけではないので，専門的パターン対の収集は容易とはいえない。用言の訳出が適切でなければ，文の骨格構造が正しくなくなるため，訳文全体の品質が大幅に低下する恐れがある。本稿では，用言性の専門語の登録を容易にするための第1ステップとして，筆者らになじみの薄い専門分野を対象に専門的パターン対の収集実験を行ない，実現可能な手順について考察する。

2 専門的パターン対の収集方法

機械翻訳は類似した文書が継続的に出現する分野への適用するのが現実的であるので，ある程度の対訳データが収集可能であることを前提とした。また，専門的パターン対を収集するための課題を明らかにするため，筆者らになじみの薄い分野を選定した。

これらの条件を満たす分野として，本稿では日本経済新聞社のデータベースから抽出した市況速報文を対象とした。この市況速報文には経済関係や株式関係の専門用語が多数使用され，文の表現にも特殊なものが多く含まれる。日本文記事は日経テレコンBIZに，英文記事はNIKKEI Telecom Japan News & Retrievalに収録され，ダウンロードして使用することが許容されている。

具体的な収集方法として，①対訳用例文を収集し，②専門的パターン対の候補を発見し，③専門的パターン対を作成する，という3ステップを仮定し，以下では各ステップについて検討する。

2.1 対訳用例文の収集

テレコンデータベースに収録されている市況速報文の日本語記事と英語記事とでは，内容的に対応するものは見つかるが，文の対応という点では意訳・省略・統合・分割・転置が多いほか，株価が異なることもあり（記事の発信時刻の違いに起因？），対応する記事を対にして取り出すのは難

しい。そこで，一定の期間の記事を取り出しておき，人手により，主として固有名詞に着目しながら別掲の用語集を参照することにより，記事の対応付けと文の対応付けを行なった。

収集したのは1995年6月第4週の市況速報文で，記事数は274件，日本文は1,834文，英文は865文で，日本文と英文で主述的表現単位が対応するのは1,598箇所，一部でも対応箇所を持つ日本文は異なりで1,495文である。対応する記事の例を図1に示す。

95/06/26/17:15
東京外為17時・円、小反発－－26銭高84円21－24銭
　円相場は薄商いの中、小反発。前週末比26銭円高・ドル安の1ドル＝84円21－24銭で大方の取引を終えた。ドル安・マルク高を受けて小口の円買いが先行したが、日米自動車問題での橋本通産相とカンター米通商代表部（ＵＳＴＲ）代表との会談を日本時間27日未明に控えて模様眺めムードが強まり、円は84円前半で小動きに推移した。市場では「米国に対日制裁を決める28日までは動きにくい」との声も聞かれた。
　朝方は値ざや稼ぎ狙いの小口の円買いが先行し円は一時84円13銭まで上昇した。その後中値決済でドル買い需要が多かったため円は84円38銭まで押し戻されたが、「84円半ば近辺には輸出企業のドル売り注文が控えている」としてドル買いも続かず、狭い範囲での値動きに終始した。
　ドルは対マルクで反落し1ドル＝1.3863－66マルクでほぼ取引を終えた。先週末発表のドイツ3州の消費者物価指数が予想以上に高かったことから独連銀による早期利下げ観測がやや後退しマルク買いが優勢になった。ただ、5月末に日米欧がドル買い協調介入を実施した水準近辺にドルが下落しているので介入警戒感も強く、東京市場では一段とドルを売り込む動きはみられなかった。円は対マルクで1マルク＝60円後半に反落した。

Tokyo Forex 5 PM: Dollar at 84.21−84.24 yen
　The dollar remained weak against the yen Monday mid−afternoon, but was boxed in a very tight band throughout the day because of the Japan−U.S. auto issue. It inched down on small−lot yen buying in line with the dollar's setback against the German mark.
　Forex market trading was extremely quiet ahead of further auto talks between Japan and the U.S., slated for early dawn Tuesday.
　The dollar stood 0.26 yen lower at 84.21−84.24 at 5 p.m.
　The U.S. currency was quoted at 1.361−1.3863 German marks at 5:15 p.m.

図1 対応する市況速報記事の例

2.2 専門的パターン対候補の発見

機械翻訳で使用する辞書を対訳テキストから自動的に作成する方法が提案されている[北村95]。この方法は2回以上現れた表現に適用可能であるが，前節で作成した対訳集は意訳や省略などのほか類似の日本文が異なる英文に訳されていることが多く，代表的な訳を選別する必要があるため，本稿では人手による作業を前提とした。

そこで，対訳コーパスの原言語表現を翻訳システムにかけ，翻訳結果と対訳コーパスの目的言語表現との比較に基づいて，利用者辞書等を整備することにする。本稿では，専門的パターン対の作成上の問題点の明確化を図るため，次の2段階に作業を分離した。まず対訳コーパスの日本文を日英翻訳システムにかけて，[池原95]の方法により体言性の専門語の辞書登録を行ない，再び対訳コーパスの日本文を日英翻訳システムにかけて，専門的パターン対候補の発見を試みた。

機械訳と対訳コーパスの英文との比較により，専門的パターン対の候補が発見されると期待される。本稿では見通しをよくするため，対訳コーパスの日本文と英文そして機械訳を組にしたものを日本語用言に着目してソートすることにより，専門的パターン対の候補を抽出した。このようにして取り出した「上げる」に対する対訳コーパスを図2に示す。

各用言に対する対訳コーパスから，用言に対する訳語は必ずしも一定しないが，代表的な訳語が決定可能であることがわかった。そのほか，着目した用言の対義語はほぼ同じ格要素をとること（図3），また他の用言性の専門語との比較から，

日米自動車協議の合意を受け山川工業が[上げ]、ＪＡＦＣＯがしっかり。 Yamakawa Industrial gained following agreement in the Japan−U.S. auto talks.
前場は一時前日比38円高の721円まで[上げ]、その後も堅調に推移している。 The stock gained 38 Yen to hit 721 early morning and traded at above 710 early afternoon.
半面、オキシ、モトローラが[上げ]、アップル、Ｄベンツ、マクドナルドが小じっかり。 Meanwhile, Occidental Petroleum and Motorola were up.
東電が商いを伴って[上げ]、関西電、日石、丸井もしっかり。 Tokyo Electric Power (No. 1) rose on high volume. Kansai Electric Power (No. 3), Nippon Oil (No. 6), Marui (No. 7) gained.
ソニーが[上げ]、東ガス、大同特鋼が堅調。 Sony (No. 4) was up. Tokyo Gas (No. 3) and Daido Steel (No. 3) maintained their gains.
ソフトバンクが反発、メモレックス、新日無が[上げ]、南野建設、サワコーがしっかり。 Softbank, Memorex Telex Japan and New Japan Ratio firmed.

図2 「上げる」に対する対訳コーパス

前場に前日比70円安の1510円まで[下げ]、3月24日の安値1570円を下回り、年初来安値を更新した。 Sumitomo Bank remained weak Thursday midafternoon after marking a low for the year of 1,510 Yen, down 70, in the morning.
ＰＨモリス、Ｐ＆Ｇが[下げ]、ＢＰ、アップル、ＮＢＣが軟調。 Philip Morris and Procter & Gamble suffered losses.
日経平均7月物はコールが[下げ]、プットが上げた。 July Nikkei 225 call options declined Tuesday morning, while puts advanced.
半面、アコム、プロミスが[下げ]、新光電工、岡野バ、三信電が軟調だった。 Meanwhile, Acom and Promise fell. Sanshin Electronics was weak.

図3 「下げる」に対する対訳コーパス

少なくとも市況分野に関しては，格要素として特定の条件が多用されることも明らかになった。

2.3 専門的パターン対の作成

日英翻訳システム上で，システムの内部情報を利用して動作する一般的パターン対の作成支援方法が提案され，パターン対に記述すべき情報を段階的に投入する場合に比べ6倍の効率化が達成されている[白井 95a]。本稿ではこの支援方法をベースに，2.2節の検討から得られた専門的パターン対の特徴を考慮して次の機能を追加した。

Ⓐ対義語を指定すれば，日本語用言と英語動詞のみを差し替えた別のパターン対を作る。

Ⓑよく現れる格要素の条件指定をメニュとして登録可能とし，条件設定の簡略化を図る。

具体的には，次の手順で作成した。

①対訳コーパスの日本文を日英翻訳システムに投入する。（図2の第1文を仮定する）
②解析情報に基づいて日本語パターンを生成する。（X(企業名 "前場" ...)(が)／上がる）
③必要に応じて，[企業名]（意味属性），「前場」などをグループ化してメニュに登録する。
④対訳コーパスの英文の解析し，英語パターンを自動的に生成する。（X(SUBJ)／gain(VP)）
⑤対訳コーパス上で日本語用言に対する対義語を見つけ，対応する英語動詞を決定する。
⑥日本語用言と英語動詞を変更したパターン対を作成する。（上がる➡下がる，gain ➡ decline）

手順中，②④⑥が支援処理により自動的に実行される。一般的パターン対を作成する場合に比べると，専門的パターン対で使用される文型はかなり限られていること[横尾 94]，⑥のステップの追加によりパターン対2件が一度に作成可能なこと，が特徴的である。このため，一般的パターン対を作成する場合に比べ，パターン対1件当たりの作成時間は2割程度短くてすんだ。

3 専門的パターン対の収集結果

前節の方法により収集された専門的パターン対は213件で，前節の対訳コーパスの日本文1,834文では延べ1,273回使用されている。多く使用された専門的パターン対を表1に示す。このほか，この日本文には一般的パターン対が436件（延べ1,507回），「拍車をかける」「顔を出す」といった慣用的パターン対が7件（同7回）使用されている。以上のパターン対の出現の様子を図4に示す。

表1 使用回数の多い専門的パターン対

度数	日本語パターン	英語パターン
52	X("物" "一角" "関連" "セクター" 企業 元素 札・券 文書類 権利 経済 数量)(が)／Y("付近" "水準" "近辺" "台" 貨幣 金銭 先後 数量)(に へ まで)／上げる	X rise to Y
50	X("物" 企業 元素 札・券 貨幣 文書類 権利 株式 金銭 元利 相場 数量)(が)／Y("利食い売り" "利食い買い" 売買 取引)(に で)／下げる	X fall on Y
43	X("物" 企業 元素 札・券 権利 経済 数量)(が)／Y("水準" "台" 金銭 貨幣 先後 数量)(で)／寄り付く	X open at Y
38	X("感" "空気" 感情 方法 表情 測定 推量 状態)(が)／強い	X prevail
30	X("物" 企業 元素 札・券 権利 経済 数量)(が)／続伸する	X continue rising
29	X("物" 企業 元素 札・券 株式 権利 文書類 数量)(が)／軟調	X be lower
28	X(主体)(が)／Y("物" 企業 元素 札・券 貨幣 文書類 権利 金銭 株式 元利)(を)／買う	X buy Y
28	X(事 取引 抽象的関係)(が)／見る+られる	there be X
27	X("物" 企業 元素 札・券 文書類 権利 経済 取引 数量)(が)／Y("小反落" "小反発" 値上げ・値下げ 上がり・下がり 先後 数量 状態)(で)／始まる	X open with Y
26	X("物" 企業 元素 札・券 権利 経済)(が)／高い	X be higher
24	X(*)(が)／しっかりだ	X edge up
21	X("物" 企業 元素 札・券 権利 経済 数量)(が)／安い	X be lower
20	X(*)(が)／Y(抽象)(に)／転じる	X shift to Y
20	X(*)(が)／Y(*)(を)／好感する	X be encouraged by Y

(注)日本語の格の条件：ｸｫｰﾄは表記指定，他は意味属性指定

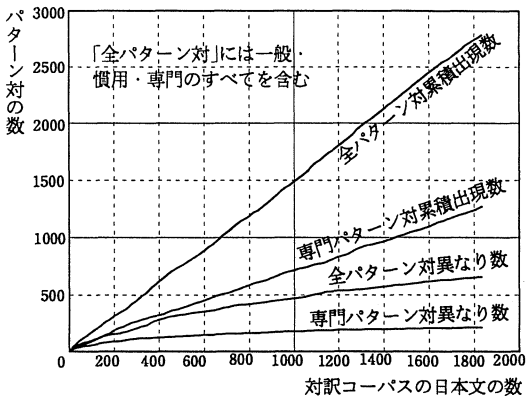

図4 パターン対の出現情況

4 専門的パターン対の収集に関する考察

専門的パターン対の作成という観点では，対訳コーパスがあれば，なじみのない専門分野であっても，それほど作業効率には影響しないことがわかった。すなわち，①翻訳システムの訳文と対訳コーパスの目的言語の比較により問題箇所は比較的容易に指摘可能であること，②問題箇所別のソートによりどのようなパターン対を作成すべきかがリストアップ可能であること，③対訳コーパスを利用したパターン対の作成支援方法によりパターン対の記述自体は容易であること，などの条件が整っているからであると考えられる。

しかし，本稿で述べた方法を実施した際，最も工数を必要としたのは対訳コーパスの収集であった。経済用語に馴れていないせいもあるが，それよりもむしろ，2.1 節で述べたように，日本語と英語はそれぞれ別のデータベースから収集したものであり，英語記事は日本語記事の一部を省略したような内容であるため，記事の対応を発見するのが容易ではないというのが直接の原因であるといえる。また，文の対応付けでは，意訳・統合・分割などが多いため，対応している箇所を最終的に決定するのに細かな検討を要した。

このように辞書をゼロから作る場合には，対訳コーパスをどれだけ効率よく収集できるかが重要である。記事の対応付けに関しては[高橋 96]，文の対応付けに関しては[春野 96]といった方法が提案されており，これらを応用した対訳コーパス支援環境を整えることが必要である。

5 おわりに

本稿では，特定の専門分野を対象としたパターン対辞書を構築するには対訳コーパスの利用が有効であること，専門的パターン対の格要素には特定の条件が多用されること，対義語のパターン対を並行して作成することにより作業効率が向上すること，その結果，約 1,500 文の市況速報の対訳コーパスから専門的パターン対 213 件が収集されたことなどを報告した。その一方で，辞書をゼロから構築するには対訳コーパスをどれだけ効率よく収集できるかが課題となることが，実際の作業から明らかになった。

今後は，他の専門分野に関して専門的パターン対の収集を試みることにより，パターン対の収集に関する一般的課題と専門分野ごとの特徴的事項を分離するよう検討を進める予定である。また，対訳コーパスの収集に関しても検討を加える予定である。これらにより，一般利用者にもパターン対の作成が可能な環境の構築を進めたい。

謝辞

ご討論くださったFrancis Bond 氏を始めとするＮＴＴの機械翻訳グループの各位，並びに，対訳用例集，関連する一般パターン対の整備を担当された小出ひとみ氏，パターン対の各種属性の付与を担当された渡邊いづみ氏と関加代氏，作成支援処理を担当された兵藤富子氏と上田洋美ら，ＮＴＴアドバンステクノロジの各位に感謝する。

参考文献

[春野 96] 春野，山崎：辞書と統計を用いた対訳アライメント環境，言語処理学会第 2 回年次大会 B3−1

[池原 93] 池原，宮崎，横尾：日英機械翻訳のための意味解析用の知識とその分解能，情報処理学会論文誌 Vol.34 No.8

[池原 95] 池原，白井，横尾，F. 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

[金田 94] 金田，秋葉，石井，H. Almualim ：事例に基づく英語動詞選択ルールの修正型学習方式，「自然言語処理における学習」シンポジウム論文集，電子情報通信学会 & ソフトウェア科学会

[北村 95] 北村，松本：対訳テキストからの翻訳知識の獲得と機械翻訳システムへの適用，言語処理学会第 1 回年次大会 C2−7

[白井 95a] 白井，兵藤，上田，横尾，池原：日英機械翻訳用構文意味辞書の作成支援，平成 7 年電気関係学会関西支部連合大会論文集 G14−3

[白井 95b] 白井，池原，横尾，井上：日英機械翻訳に必要な結合価パターン対の数とその収集方法，情報処理学会研究報告 95−NL−110−7

[高橋 96] 高橋，白井，藤波，池原，上田，松島：DB から抽出した日英新聞記事の自動対応付け，言語処理学会第 2 回年次大会 B3−3

[横尾 94] 横尾，中岩，白井，池原：日英機械翻訳用スケルトン−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

経済用語に関する参考資料

(1) M. Robinson, I. Etzkorn, and T. Tunstall : "The Wall Street Journal" Guide to Understanding Money & Markets, Joint Venture: ACCESS PRESS and Siegel & Gale, Inc. (1990)

(2) 日本経済新聞社編：金融証券英語辞典，日経文庫，(1990 年 6 月)

(3) 日本経済新聞社編：経済新語辞典 1995 年版 (1994 年 9 月)，同 1996 年版(1995 年 9 月)