用例利用型翻訳に適した対訳用例の自動作成について

足立 貴行　高橋 大和　内野 一　古瀬 蔵
ＮＴＴサイバーソリューション研究所

1．はじめに

用例利用型翻訳[1]では，事前に原言語文と目的言語文の対である対訳用例を収集しておき，入力文に類似した原言語文に対する目的言語文を模倣して翻訳を行う．

用例利用型翻訳で高品質の翻訳結果を得るには，対訳用例を大量に集める必要がある．現在，計算機上で利用可能な文書は単言語で記述されたものがほとんどであるが，２言語の単言語コーパス間で文対応付けして対訳コーパスを作成する研究[2, 3]が行われており，大量の対訳用例が使用できるようになりつつある．しかし，①自動の文対応付けの精度は完全でない，②人手で行うにしても，全てが１対１の文対応とはならない，③対応が正しくても，２言語の表す情報の量が同じではない，などの理由により，対訳コーパスの全ての対訳用例が用例利用型翻訳に適しているわけではない．高精度な翻訳を行うには，対訳用例の原言語文全体と目的言語文全体の表す情報が同じである必要がある．

本稿では文対応付けされた対訳コーパスから対訳用例を絞り込み，用例利用型翻訳に適した対訳用例を自動作成する手法を提案する．また，市況速報文を翻訳対象とする用例利用型日英機械翻訳における提案手法の有効性を示す．

2．コーパスの文対応付け

筆者らは市況速報分野の日英機械翻訳を対象として用例利用型翻訳の研究を行っている．市況速報は，①継続的に記事が配信されるので大量に対訳用例を収集できる，②定型的な文が多数含まれている，③省略や特殊な言い回しが多用されるため要素合成的な翻訳手法では高精度の翻訳結果を得ることが難しい，という特徴を持ち，用例利用型翻訳が有効な分野と考えられる．現在は，市況速報に関する日本語と英語の単言語コーパスを人手で文対応付けした対訳コーパスを用いているが，同じ内容について２言語の単言語コーパス間で自動記事対応付け[4]と自動文対応付け[5]を行って得られた対訳コーパスを用いることも検討している．

用例利用型翻訳に適した対訳用例の例を図１に示す．図１の J01 と E01，J02 と E02 の組はそれぞれ日本語と英語の情報が等価である対訳用例であり，用例利用型翻訳で入力文の類似用例として検索されれば高品質の訳文を出力することが期待できる．また，J02 と E02 の組は市況速報分野で頻出する定型的な組であり，用例利用型翻訳でも高頻度な利用が期待できる．

● 日本語文の用例
J01:TOPIX 10月 物、日経 300 の 10月物は閑散。
J02:64.537・△0.045。

● 英語文の用例
E01:Trading in Oct. TOPIX and Nikkei 300 options was light.
E02:The Nikkei World Commodity Index rose 0.045 point to 64.537 Wednesday.

● 文対応情報（日本語文:英語文）
J01:E01, J02:E02

図１：用例利用型翻訳に適した対訳用例の例

しかし，実際の対訳コーパスにおける対訳用例はこのような理想的なものばかりとは限らない．本研究で使用した日英の対訳コーパスは以下のような特徴がある．

(I) 同じ内容の文でも表現方法が異なる（言語の違いによる特徴）
日本語文：CD の３カ月物は 0.50-0.53%の気配で、前週末とほぼ同水準。
英語文　：Three-month CDs were bid at 0.53% and offered at 0.50% , nearly flat from Friday.

(II) 文対応付けの誤りを含む（コーパスの特徴）
日本語文：全般に商いは低調。
英語文　：Aug. Nikkei 225 call options mostly fell and many puts firmed Monday morning, amid sluggish trading.

(III) 日本語文と英語文で示される情報の量が異なる（コーパスの特徴）
日本語文：先物の中心限月である 12月物の前場終値は前日比 16銭高の 121円 15銭。
英語文　：The key No. 174 bond settled at 114.05 yen, up 0.24, yielding 2.660%, down 3 basis points. The benchmark Dec. contract

settled at 121.15 yen, up 0.16.

(III)の特徴が現れる理由は，(a)英語文は日本語文の直訳ではない，(b)M対Nで対応付けられているコーパスを1対N対応として利用している，ためである．

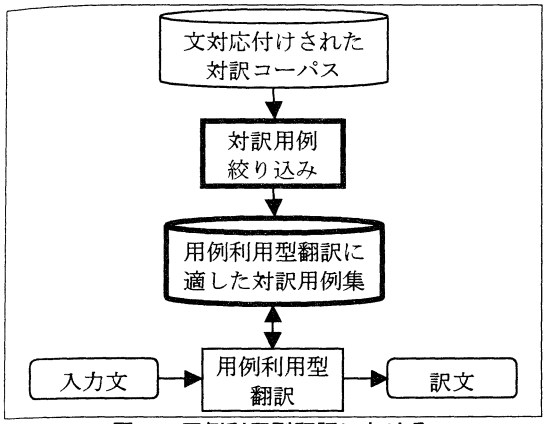

図2：用例利用型翻訳における対訳用例の絞り込み

3．用例利用型翻訳に適した対訳用例の絞り込み

対訳コーパスが2節の特徴を持つため，対訳用例をそのまま用例利用型翻訳に使用すると正しい訳が得られない場合がある．例えば，以下のような対訳用例を用いて翻訳を行うと，入力文が対訳用例の日本語文と全く同じものであっても，日本語文と英語文で情報の量が異なっているので高品質な訳文が得られない．

日本語文：日経平均先物12月物は前週末比270円高の1万8180円で前場の取引を終えた。
英語文　：Dec. Nikkei 225 futures ended higher Monday morning.

このような例を取り除くため，対訳コーパスから対訳用例の絞り込み（図2）を行う必要がある．

本稿では対訳用例の絞り込みの方針を，(1)典型的な対訳用例を用いる，(2)語句対応の比率の高い対訳用例を用いるとして，それぞれの基準に従った2種類の絞り込み手法を提案する．以下，各絞り込みの説明を3.1節，3.2節で，各手法の効果を4.1節，4.2節で述べる．

3.1．頻出表現に関する典型的な対訳用例の絞り込み

対訳コーパスに頻出する同型の日本語文が常に同じ英語文と対応していれば，そのまま対訳用例として用例利用型翻訳に用いることができる．しかし，頻出する同型の日本語文に対する英語文に

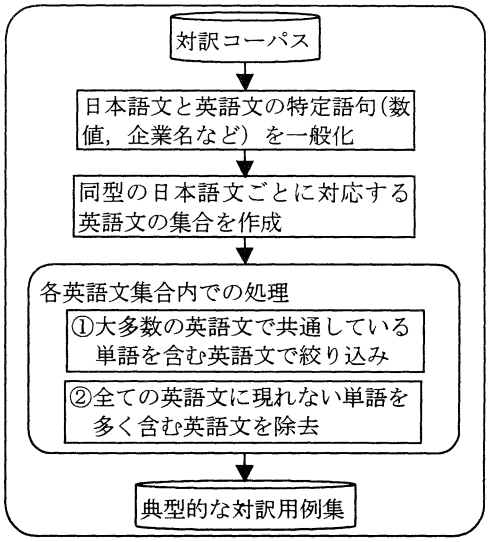

図3：典型的な対訳用例抽出のアルゴリズム

● 日本語文（□は一般化した部分）
J11: 9 日の日経国際商品指数は 65・457 で、前日比 0・024 ポイント下落した。
J12: 16 日の日経国際商品指数は 65・368 で、前日比 0・186 ポイント下落した。
J13: 5 日の日経国際商品指数は 63・766 で、前日比 0・048 ポイント下落した。
J14: 12 日の日経国際商品指数は 65・376 で、前日比 0・034 ポイント下落した。

● 英語文（□は一般化した部分，下線は大多数の英文で共通している語，太字イタリック文字は全ての文では共通して現れない語）
E11: ***The*** Nikkei World Commodity Index ***shed*** 0.024 ***of a*** point to 65.457 Thursday.
E12: ***The*** Nikkei World Commodity Index ***fell*** 0.186 point to 65.368 Thursday.
E13: Nikkei World Commodity Index ***fell*** 0.048 point to 63.766 Thursday.
E14: Index fell 0.034 point to 65.376 Thursday.

● 文対応情報（日本語文:英語文）
J11:E11, J12:E12, J13:E13, J14:E14

● 絞り込まれた対訳用例
(J12:E12), (J13:E13)

図4：典型的な対訳用例の絞り込み例

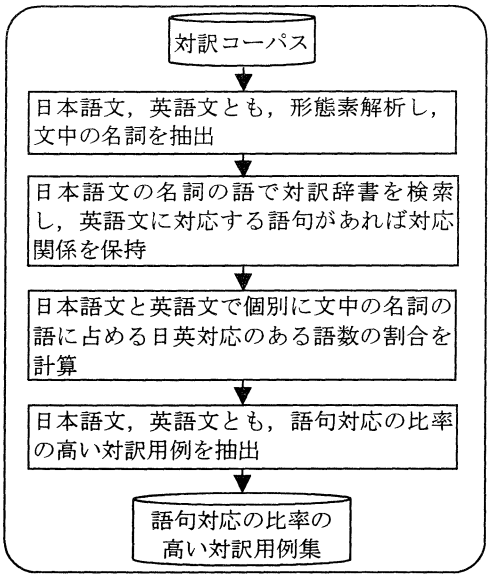

図５：語句対応の比率の高い対訳用例の絞り込みのアルゴリズム

● 日本語文（下線は名詞，□は対応のある名詞）

J2: TOPIX先物 12月 物 も 同様の 動き で 18 ポイント 高 の 1433 ポイント 、 日経 300 先物 12 月 物 は 同 2.0 ポイント 高 の 266.0 ポイント で 引け た 。

● 英語文（下線は名詞，□は対応のある名詞）

E21: Dec. TOPIX futures settled 18 yen higher at 1,433 yen.

E22: Dec. Nikkei 300 futures finished 2.0 up at 266.0 yen.

● 文対応情報（日本語文:英語文）

J2:E21,E22（1対2対応）

● 名詞語句の対応割合

（語句対応のある名詞数／文中の名詞数）

日本語文の対応割合(J2)　：9 / 12 = 0.75

英語文の対応割合(E21+E22)：(5+6) / (7+7) = 0.78

図６：語句対応の比率の高い対訳用例の例

は部分的に様々な言い回しが使われていることが多く，そのままでは翻訳に利用できるか分からない．そこで，同型の日本語文に対応する様々な英語文は類似していると仮定し，図３に示すように，その中で最も典型的な英文を選ぶことで，対訳用例の絞り込みを行う．

図３のアルゴリズムに従って，図４の典型的な対訳用例の絞り込み例を説明する．まず，数詞などを一般化すると同型である日本語文 J11，J12，J13，J14 に対応する英語文の集合{E11, E12, E13, E14}を作成する．次に①の処理において，全文には含まれてないが，多数の文に含まれている共通語（下線）がない E14 を除き，英語文集合{E11, E12, E13}に変更する．さらに②の処理において，全ての文では現れない語（太字イタリック文字）を多数含んでいる E11 を除き，英語文集合{E12, E13}に変更する．結果，絞り込まれた英語文 E12，E13 を含む対訳用例が典型的な対訳用例として作成される．

３．２．　語句対応の比率の高い対訳用例の絞り込み

対訳用例について，対訳辞書を用いて文中で語句対応が取れている語の割合を調べ，その割合を元に抽出を行う．今回は，企業名のような日本語文と英語文で対応が取りやすく，情報として重要なものを含んでいる名詞（数詞，時詞を含む）に限定して処理を行った．形態素解析には日本語は ALT-J/E の形態素解析部[6]を，英語は Brill の Tagger[7]を使用し，その結果をそのまま用いた．図５のアルゴリズムに従い，図６の対訳用例を用いて，対訳用例の絞り込み処理を説明をする．

まず，図６の日本語文 J2，英語文 E21，E22 を形態素解析すると，下線の語が名詞であることが分かる．次に日本語文の名詞の語で対訳辞書を検索し，対応する語句が英語文に存在するか調べると，四角の囲みの語が該当する．さらに日本語文と英語文で個別に名詞語句の対応割合を調べる．この例では，日本語文の名詞の語数中の語句対応割合は 9/12=0.75，同様に英語文では (5+6) /(7+7)=0.78 となる．最終的に，日本語文と英語文の語句対応の割合が高い対訳用例を抽出する．

４．対訳用例絞り込みの効果

４．１．　典型的な対訳用例の絞り込み

日本経済新聞社が配信する日英の市況速報記事(1995 年 8 月～1996 年 3 月分)のうち人手で文対応付けして抽出された対訳用例 29,157 ペアについて，数詞，企業名を一般化したのち，同型の日本語文である対訳用例 9,288 ペア（日本語文異なり数 1,702 文）を対象とした．本アルゴリズムによる絞り込みの結果，対訳用例 9,288 ペアから 3,973 ペアに絞り込まれた．

３．１節の絞り込み方法により用例利用型翻訳の訳質が向上する例を図７に示す．図７の E32 は E31，E33 と比べて共通しない語句を多く含む対訳用例であるので３．１節の方法で除かれる．また，E32 は J32 とは異なる情報を多く含んでいるため，他とは異なった語句が多い．もし，用例利用型翻訳

● 日本語文
J31: 全般に商いは低調。
J32: 全般に商いは低調。
J33: 全般に商いは低調。

● 英語文（[ ]は一般化したもの）
E31:Trading was generally muted.
E32:Aug. Nikkei [225] call options mostly fell and many puts firmed Monday morning, amid sluggish trading.
E33:Overall turnover was weak.

● 文対応情報（日本語文:英語文）
J31:E31, J32:E32, J33:E33

● 絞り込まれた対訳用例
(J31:E31), (J33:E33)

図７：典型的な対訳用例の絞り込みで翻訳の訳質が向上する対訳用例の例

● 日本語文（下線は名詞，[ ]は対応のある名詞）
J4: 情報 通信 関連 を [中心] と し た 好業績銘柄 の 物色 の 流れ は 続く だろ う 。

● 日本語文（下線は名詞，[ ]は対応のある名詞）
E4: Market participants will focus on issues with sound fundamentals and cause the Nikkei Stock Average to start climbing toward 18,500 , he said.

● 文対応情報（日本語文:英語文）
J4:E4

● 名詞語句の対応割合
（語句対応のある名詞数／文中の名詞数）
日本語文の対応割合(J4)：1 / 7 = 0.14
英語文の対応割合(E4)　：0 / 9 = 0

図８：語句対応の比率の低い対訳用例の例

において３．１節の絞り込みを行わずに入力文がJ32と同じもので翻訳を行ったならば，E32が入力文の情報とは異なった訳として出力されてしまう．絞り込みを行えば，このような場合でも入力文に対し典型的な訳であるE31またはE33が出力される．

４．２．　語句対応の比率の高い対訳用例の絞り込み

現在，この手法については評価中であるが，以下のような例で有効であることが確認されている．

図６は多数の語句対応が取れる例である．図６の日本語文と英語文で多くの対応が取れるのは，市況速報分野によく現れる数詞，企業名の語が多く含まれており，これらの語が対訳辞書に記載されているからである．また，日本語文，英語文とも語句対応の比率が高いので対訳用例の日本語文と英語文の情報が等価である．図６の対訳用例を用例利用型翻訳で用いた場合，入力文として図６のJ2と同じであれば，E21，E22の英文を適切な訳として出力できる．

図８は日本語文と英語文で語句対応が付いていないので，３．２節の対訳用例の絞り込みによって除かれた対訳用例である．この絞り込みを行ってから用例利用型翻訳を行うと，入力文がJ4と同じ文であってもE4は出力されず，他にJ4に類似した日本語文で，日本語文と英語文の情報が等価な対訳用例があれば翻訳される．

５．おわりに

用例利用型翻訳の訳出の品質を上げるために，文対応付けされた対訳コーパスから用例利用型翻訳に適した対訳用例を自動作成する方法について提案した．

今後は，提案方法を実際のシステムに適用して評価を行いながら，改良を加えていく．また，典型的な対訳用例と，語句対応の比率の高い対訳用例の２つの絞り込み手法の組合せを検討する．さらに，市況速報分野以外にも提案手法の適用を試み，汎用的な高精度の用例利用型翻訳の実現を目指す．

参考文献

[1] 足立他: 市況速報文を対象とする用例利用型日英機械翻訳, 情報処理学会第 57 回全国大会講演論文集(2), pp.263-264, 1998.
[2] 春野雅彦: 辞書と統計を用いた対訳アライメント, 情報処理学会論文誌, Vol.38, No.4, pp.719-726, 1997.
[3] 宇津呂他: 対訳辞書および統計情報を用いた二言語対訳テキスト照合, コンピュータソフトウェア, Vol.12, No.5, pp.414-423, 1995.
[4] 高橋他: 日英新聞記事の記事対応コーパス自動作成, 言語処理学会第 3 回年次大会発表論文集, pp.127-130, 1997.
[5] 白井他: 新聞記事日英対訳コーパスの構築(3)－記事の特徴分析と文の対応関係の検討—, 平成７年度電気関係学会九州支部連合会大会論文集, p.857, 1996.
[6] 白井他: 日英翻訳のための日本語解析技術, NTT R&D, Vol.46, No.12, 1997.
[7] Brill, E: A Simple Rule-based Part of Speech Tagger, Proc. 3rd Conference on Applied National Language Processing, pp.152-155, 1992.