対訳コーパス中の規則獲得不適文対の自動認定

5R－8

中岩浩巳

NTT　コミュニケーション科学研究所

1．はじめに

近年，自然言語処理では，解析や翻訳，分類，検索，生成等の処理で必要となる各種規則を，人手作業なしに効率的に作成するために，コーパスを自動分析し，その分析結果に基づいて，各種自然言語処理のための規則を自動獲得する研究が盛んに行われている[1][2][3]．これらの研究では，適用する自然言語処理に応じて，用いるコーパスの種類が異なってくる．例えば，機械翻訳の翻訳規則のように対訳表現対が規則として必要になる場合[1]，省略要素への補完規則のように補完対象言語ではほぼ常に省略される要素に対する規則獲得がその言語のコーパスだけでは困難な場合[2]，また，構文解析規則の獲得のように解析対象言語のコーパスだけでは構文解析の曖昧性が解消困難な場合[3]等には，対訳コーパスが利用されることが多い．

対訳関係にある文対の集合である対訳コーパスから各種規則を抽出する場合，そのコーパス中の文の集合から対訳関係にある文対を抽出する技術，対訳関係にある文対から対訳関係にある表現対を抽出する技術，得られた対訳表現対を元に各種規則を抽出する技術が重要となり，従来から様々な手法が提案されてきている．しかし，これらの技術により認定された対訳関係にある文対は，ある言語の文中の表現がすべて1対1で別の言語表現に忠実に訳された文からなるものばかりとは限らず，意訳されていたり，要約されていたり，1部の表現のみが翻訳されていたりするものも多い．よって，これらの文対では，対訳関係にある各文のすべての表現が，規則作成のための情報源として利用できるわけではない．また，対訳表現対を元に規則を抽出する技術では，対訳関係にある文対を解析して，その解析結果を元に規則を自動獲得する手法が多いが，文の解析に失敗した場合には，適切で正確な規則を自動獲得するのは困難である．以上のことから，認定された文対のすべてが，規則獲得に適切なものである保証はなく，場合によっては，不適切な文対から獲得することにより規則の品質が低下することも予想される．

本稿ではこのような問題を回避するために，対訳関係にある日英の対訳文対集から，規則の自動獲得に不適切な対訳文対を自動認定する手法を提案する．

2．規則獲得不適文対

本節では，規則の自動獲得に不適切だと判断できる対訳文対の傾向について述べる．前節でも述べた通り，規則獲得に不適切な文対は，大きく以下の2種類に分類することが出来る．

● **意訳・要約され忠実な対訳文対ではないもの**

これは，文対としては，対訳関係にはあるが，要約されたり，意訳されたりして，文より細かな表現単位では対訳関係にある表現対の認定が困難である場合である．例えば，

(1) J: 国民は歓呼して彼を国王に迎えた
E: The people acclaimed his king.

では，日本語文は2用言からなるが，英語文では1用言を用いて意訳されている．これらの文対は，対訳表現対の情報を利用して規則獲得する手法においては，規則獲得が困難な不適切な文対となる．

● **文の解析に失敗したもの**

これは，対訳文対の文の解析において，解析処理系が誤った解析構造を生成してしまったために，その誤った解析構造情報を元に，規則を作成してしまう場合である．例えば，

(2) J: ぜったいにそうではない
E: Certainly not.

では，日本語文には，「そうではない」の1カ所しか用言がないが，日本語解析の際に誤って，本来副詞である「ぜったい」も用言として認定してしまい，不適切な日本語構造が生成されている．

このような，規則獲得には不適切な対訳文対を詳細に検討したところ，下記のような傾向にあることが分かった．

● **対訳文同士で単文の数が違うものが多い**

これは，意訳や要約されることにより，対訳文間の両者の単文の数が異なる場合には，両者の構文構造が大幅に異なっており，その対訳文対からの規則の獲得が困難であることを示している．また，意訳や要約されていない場合でも，解析失敗により単文の数が異なる場合には，致命的な解析失敗が起こっている可能性が高いことを示している．

● **完全に解析が出来なかったものが多い**

未知語や解析規則不足等により，完全な解析構造が得られなかったり，機械翻訳システムの解析系で翻訳出来ず原言語表現のまま残ってしまう等，完全な解析が得られなかった場合には，不適切な構造である可能性が高いことを示している．

3．規則獲得不適文対の自動認定法

2章の考察をもとに，日英対訳コーパスから規則獲得に不適切な文対を自動認定する手法について述べる．提案手法の構成を図1に示す．図の通り，入力された日英対訳コーパス中の対訳関係にある日本語文と英語文を解析し，その文対から対訳関係にあ

Automatic Identification of Unsuitable Aligned Sentence Paris for Rule Extraction from Bilingual Corpus.
Hiromi Nakaiwa, NTT Communication Science Labs.

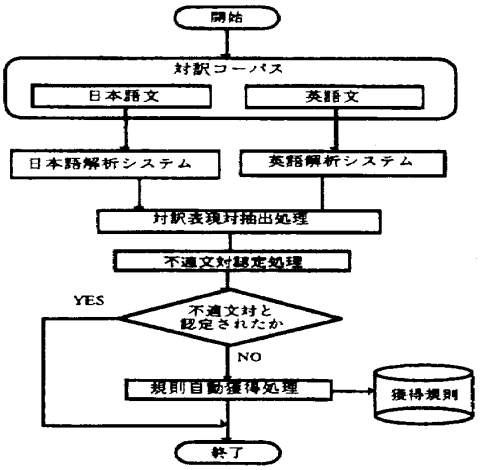

図1　規則獲得不適文対の自動認定処理の構成

る表現対を抽出する[4]．今回の実装では，日本語解析システムとして，日英機械翻訳システムの日本語解析系を利用し，英語解析システムとして，汎用的な英語構文解析系を利用した．次に，日本語文，英語文の構文構造及び日英対訳表現対候補の情報を用いて，規則の自動獲得に不適切であると予想される日英対訳文対を下記の条件により認定する．

- ●日本語動詞が英語名詞と対訳関係にあると認定された日本語文の単文部分を除いて，日本語文の単文の数と英語文の単文の数が異なる場合．
- ●日本語文の日英機械翻訳システムによる機械翻訳の結果，機械訳文中に翻訳できない日本語表現が残った場合．
- ●英語文の構文解析系による解析の結果，解析できない部分が残った場合．

以上の条件により規則獲得不適文対と認定された日英対訳文対に対しては以降の規則獲得のための処理は行なわれない．

4．評価

4．1　評価方法

本論文で提案した規則獲得不適文対の自動認定手法により，日英対訳コーパスから規則獲得不適文対を自動的に抽出し，その結果を調査して提案手法を評価した．評価条件の詳細は以下のとおりである．

日本語文・英語文の解析：日英機械翻訳システムALT-J/E[5]を日本語解析系に利用して日本語文の構文意味構造を生成し，Brill の英語 tagger[6]とLink Grammar Parser[7]を利用して英語文の構文構造を生成した．

不適文対：対訳コーパスからのゼロ代名詞照応解析規則の自動獲得法[2]において，そのまま活用すると不適切な規則が獲得される文対を不適文対として人手で判断して評価した．

獲得対象：学習研究社の英和辞書電子データ中の日英対訳例文対（約4000文対）中で，ALT-J/Eの日本語解析の結果，構文意味構造にゼロ代名詞が存在した文対（759文対）から無作為に選択した200文対を用いた．

評価条件：個々の認定結果に対する日本語解析成功の可否，英語文が直訳か意訳かを人手で調査し集計した．

4．2　評価結果

規則獲得不適文対の認定処理の結果を表1に示す．この表によると，不適文対と認定された82文対(41%)の内67文対(33.5%)は，本処理で排除したい日本語解析に失敗したもので，14文対(7%)は，日本語解析は成功したが照応解析規則の獲得が困難な英語文が意訳されたものであった．よって，この82文対(41%)の内81文対(40.5%)は，本手法により照応解析規則の獲得に不適切な文対として正しく認定できた．また，適切文対と認定された118文対(59%)の内86文対(43%)は，有効な照応解析規則の獲得が期待できる日本語解析が成功した文対であった．さらに，適切文対と認定されたが日本語解析に失敗した32文対(16%)の内17文対(10%)は，英語文が直訳で日英ともに1単文からなる単純な構造の文対であった．これらは，日英の構造を比較することで，正しい日本語構造に容易に修正することができるので，正しい日本語構造に修正した後，有効な規則の獲得が期待出来る．よって，この118文対(59%)の内，103文対(51.5%)は，適切な照応解析獲得が期待できる適切な文対であった．以上の結果から，3章に示した比較的単純な規則だけでもかなりの割合の不適切な日英対訳文対が正確に排除できることが分った．

表1　規則獲得不適切文対の自動認定処理の結果

不適対訳文対認定結果	日本語解析成功		日本語解析失敗		小計
	英語文直訳	英語文意訳	英語文直訳	英語文意訳	
適切文対	37% (74文対)	6% (12文対)	14.5% (29文対)	1.5% (3文対)	59% (118文対)
不適文対	0.5% (1文対)	7% (14文対)	28.5% (57文対)	5% (10文対)	41% (82文対)
小計	37.5% (75文対)	13% (28文対)	43% (86文対)	6.5% (13文対)	100% (200文対)

5．まとめ

本稿では，対訳コーパスからの規則自動獲得において，適当でない対訳文対を自動認定する手法を提案した．今後は，具体的な規則自動獲得手法と提案手法を結合して，提案手法の規則自動獲得での有効性を詳細に検討したい．

参考文献

[1] Kaji, H. et al. : Learning translation templates from bilingual text, Proc. of COLING-92, pp.672-678, (1992).
[2] 中岩: 日英対訳コーパス中のゼロ代名詞とその指示対象の自動認定,情報処理学会研究報告, Vol.NL-123, No.5, pp.33-40 (1998).
[3] Dagan, I. et al. : Two languages are more informative than one, Proc. of 29th ACL, pp. 130-137, (1991).
[4] Yamada, S. et al. : A New Method of Automatically Aligning Expressions within Aligned Sentence Pairs, Proc. of NeMLaP2, pp. 56-65 (1996).
[5] 池原他: 言語における話者の認識と多段変換方式, 情報処理学会論文誌, Vol.28,No.12, pp.1269-1279 (1987).
[6] Brill, E. : A simple rule-based part of speech tagger,Proc. of ANLP'92 pp.152-155 (1992).
[7] Sleator, D. and Temperley, D. : Parsing English with a Link Grammar,Carnegie Mellon University Computer Science technical report, CMU-CS-91-196 (1991).