用例ベース翻訳における用言句の簡潔な翻訳の実現

荒牧 英治 †‡　　黒橋 禎夫 †‡　　柏岡 秀紀 ‡　　田中 英輝*
† 東京大学大学院情報理工学系研究科
‡ATR 音声言語コミュニケーション研究所
＊ NHK 放送技術研究所

{aramaki, kuro}@kc.t.u-tokyo.ac.jp
hideki.kashioka@atr.co.jp, tanaka.h-ja@nhk.or.jp

1　はじめに

インターネットの急速な発展とともに利用可能な電子化テキストの量が増加しつづける現状にともない，用例ベース翻訳 [6] や統計ベース翻訳 [3] など大量の対訳テキストを用いた機械翻訳 (データドリブン MT) に関する研究が盛んに行われている．

こうしたデータドリブンの手法は，まず，対訳テキスト間の対応関係を推定するが，この際，ある言語に出現する表現が他方の言語で表現されていない場合があり，対応の推定を困難とする一因となっている．このような言語間で表現がずれる例を次に示す．

J: カナダで 開かれた 通商会議で...
E: At a trade conference in Canada...

E は下線部の表現 “開かれた” を明示的に表現していないが，この表現は前後のコンテキストから推測可能である．本稿では，このような相手側言語で表現されなくても推測できる表現を**推論可能表現**と呼ぶことにする．この例のように，用言はしばしば推論可能表現となるが，従来のデータドリブン MT では，このような現象について十分に議論されてこなかった．

そこで，用言の翻訳のされ方を調べるために，対訳文に対して (1) どの句が用言であるか，(2) 用言句は相手側言語のどの句に対応しているか，の 2 つの情報をアノテートした**用言対応コーパス**を作成した．そして，その観察結果から，推論可能表現が出現するパターンに関する知見を得た．

本稿の議論は対象とする言語ペアおよび翻訳方向に依存しないが，本稿では日本語の推論可能表現について述べる．

2　用言対応コーパス

2.1　用言対応コーパスの作成

言語間の表現のずれを調べるためには直訳された対訳コーパスよりも，それぞれの言語において自然な翻訳がなされた対訳コーパスが必要となる．そこで，NHK ニュースの対訳記事を用いて用言対応コーパスを作成した．NHK ニュースの対訳記事は，日本語原稿がまずあり，それをもとに英語原稿が英語のニュースとして自然となるように翻訳されている．用言対応コーパスの作成に際しては，この対訳記事に対して，まず自動で対訳文の抽出と対応付けを行い，その結果をもとに人手で用言対応に注目して修正を行うという方法で作成した．この処理は次の 4 つのステップからなる．

ステップ 1: 文アライメントの推定
まず，翻訳辞書 (約 200 万語対) を用いた DP マッチングによる手法で対訳記事の文アライメントを推定する [1]．次に，アライメント結果が 1 文:1 文対応と推定された対応のみを抽出する．

ステップ 2: 句を単位とした依存構造に自動変換
次に，対訳文の両言語をパーサを用いて句を単位とした依存構造に変換する [2]．英語パーサ [4] は語を単位とした句構造を出力するので，以下の規則により語をまとめて句にし，ヘッドを決定することによって句を単位とした依存構造とした．

1. 機能語を後続する内容語にまとめる．
2. 複合名詞を構成する名詞は一つの句にまとめる．
3. 助動詞を主動詞にまとめる．

日本語パーサ KNP[5] は，句を単位とした依存構造を出力するので，これをそのまま用いる．

ステップ 3: 句対応の自動推定
文献 [2] の手法を用い用言句/非用言句の自動推定，および，句対応の自動推定を行う．ここでいう用言句とは，(1) 動詞を含む句，および (2) 格要素を持つ形容詞

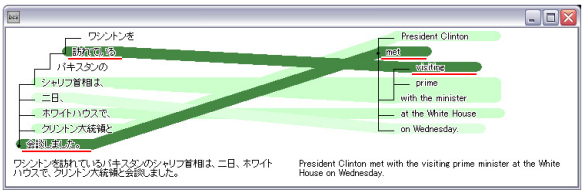


図 1: 用言対応コーパス

表 1: 用言対応の分類と数

用言対応の分類 (日本語:英語)	対応数
用言句-用言句	9779
用言句-ϕ	6831
用言句-前置詞句 または 用言句-名詞句	*710*
その他	*316*

* 用言かどうかの区別しかアノテートしていないため，*Italic* で示される値については自動判定した値を示した．

を含む句とする．ただし，自動推定には誤りも含まれるため，次のステップで人手により修正する．

ステップ 4: アノテーション
作業者が用言句/非用言句の自動推定結果を修正する．同時に，用言句に対しては，その対応先を修正する．ある言語側の用言句は必ずしも相手側言語の用言句に対応しているとは限らないため，用言句以外の対応先も作業者に許す．また，対応先となる表現が存在しない場合は，対応先が存在しないという情報 (**用言句**-ϕ) をアノテートする．

以上の手順で 5,500 対訳文の対応付けの作業を行った．作成した例を図 1 に示す．用言句 (下線) に対して，人手で対応関係が付与されている．

2.2 用言対応コーパスの分析

用言対応コーパスを観察すると，日本語の用言句が英語側の用言句と対応しない場合が数多く観察される．日本語の用言句が英語側でどのように表現されているかを集計した結果 (表 1)，**用言句**-ϕ の割合が多いことが分かる．**用言句**-ϕ が起こるのは，次の 2 つの原因によるものである: (1) 文アライメントの失敗のために，そもそも対応する用言が対訳文中になく別の文にある場合，(2) 推論可能表現であるため訳出されない場合．

(2) の例としては，1 節の “開かれた” のような場合が挙げられ，以下のような構造となる:

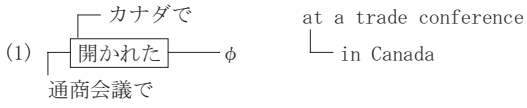


表 2: CAP の判定

判定	分類	#
推論可能	P-依存	21
	C-依存	16
	BOTH-依存	19
	計	56
推論不可能	統語解析のエラー	3
	アライメントのエラー	11
	句のチャンキングエラー	1
	その他	9
	計	24

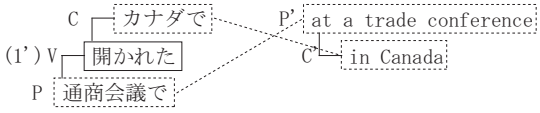
* 推論可能に分類された中での区分 (P, C, BOTH-依存) は，次のページで述べる．

この場合，すでに述べたように，このコンテキストをともなう “開かれた” は推論可能で訳出されていない．

3 推論可能表現の学習

3.1 推論可能表現の出現するパターン

本研究では，**用言句**-ϕ の数が多いことから，**用言句**-ϕ を取り扱うことにする．**用言句**-ϕ の周辺に注目すると，その周辺の表現は両言語間で対応している．例えば，前節の例では周辺の句が次のように対応している:



この形は，日本語用言句 (以降,V) の親 (parent, 以降 P) と子 (child, 以降 C) の両方の対応先 (P', C') が英語側でも親子関係になっていると捉えられる．本稿では，この形にあてはまる日本語 3 句 (以上)，英語 2 句 (以上) のペアを Condensed Alignment Pattern (以降，**CAP**) と呼ぶことにする*.

ここで，もし，CAP の形で出現する用言 (V) は推論可能表現であり，V を訳出する必要がないならば，CAP を収集し，これを用例として用いることで，V の省略を実現する翻訳が可能となる．そこで，この仮定を調べるために句アライメントが自動推定された対訳文から CAP を収集し，V が推論可能であるかどうかを調査した (表 2)．

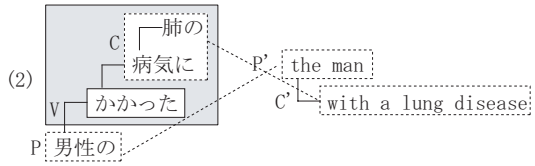
その結果，表 2 に示されるように，アライメント失

*ここでいう CAP とは逆に，日本語 2 句-英語 3 句の CAP も存在するが，提案手法は日英翻訳方向を扱っているため，本稿では取り扱わなかった．

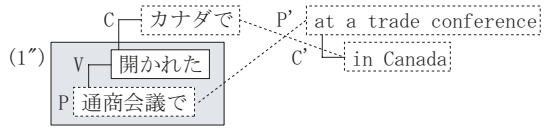
敗や統語解析失敗である場合をのぞいて，CAP 中の用言 (V) の省略は推論可能であった．

しかし，この日本語 3 句からなる CAP 全体をそのまま翻訳用例として使う場合，入力文と用例との間で 3 句が一致する必要があり，その利用機会は少ないと考えられる．そこで，CAP 中で汎化できる場所はあるかどうか，また，それはどこかという観点から CAP を次の 3 つに細分類した．

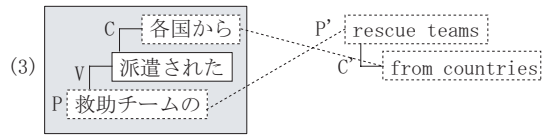
1. **C-依存**(**C** によって V が推論されるタイプ)
   次の例では，C“肺の病気” により，V“かかる” が推論されるため，V が訳出されていない．この場合，P を汎化できる．



2. **P-依存**(**P** によって V が推論されるタイプ)
   前章の (1) がこれにあたり，P“通商会議” により，V“開かれる” が推論される．この場合，C を汎化できる．



3. **BOTH-依存**(P と C の両方によって V が推論されるタイプ)
   次の例では，C“各国” と P“救助チーム” の両方がそろって V“派遣” を連想させる．このような場合は，汎化できずに CAP 全体を用例と考えるしかない．



3.2 頻度による学習

前節の分類は場合によっては主観的なものであり，すべての CAP について，分類を行うことは困難である．しかし，P が明らかに V を推測させる場合は，P と V を含んだ CAP が多数存在するはずである．そこで，以下の手法により，CAP を (P,V,P') と (V,C,C') の 2 つの CAP の断片に分けて集計し，CAP の自動分類を行うことを考える．

ステップ 1: P,C が名詞句である場合は主辞の名詞に，動詞句である場合は主動詞に汎化する．

ステップ 2: CAP を 2 つの CAP の断片 (P,V,P') と (V,C,C') に分割し，用例全体から集計を行う．ここで，前者の出現頻度を $freq(P)$，後者の出現頻度を $freq(C)$ とする．

ステップ 3: 集計の結果，$freq(P) > freq(C) \times 2$ ならば，P-依存とする．逆に，$freq(C) > freq(P) \times 2$ ならば，C-依存とする．それ以外は，BOTH-依存とする．

表 3: 自動分類された CAP

	# of CAPs
P-依存	1120
C-依存	297
BOTH-依存	2802

表 4: BLEU スコア

	テストセット [240 文]	サブセット [104 文]	サブセット [14 文]
BASELINE	24.6	24.7	26.3
CAPMT	24.8 (+0.8%)	-	29.0 (+10.2%)
CAPMT+	25.0 (+1.6%)	25.7 (+4.0%)	-

* () 内の数字はベースラインと比べての比率である．

4 実験

提案手法の有効性を確かめるため，次の 2 つの観点: (1) どれくらいの CAP が得られるか (2) 翻訳精度をどれだけ向上させるか，から実験を行った．

4.1 得られた CAP の評価

まず，どれくらいの数の CAP が用例から得られるのか調べてみた．句アラメントが自動推定された 52,749 対訳文から CAP を抽出したところ，ここから延べ 4,219 個の CAP を収集できた．これらの自動分類結果は表 3 に示されるように，BOTH-依存と判定されたものが多い．しかし，このうちのほとんど (2,272 個) は 1 回しか出現しない CAP であった．よって，より多くの CAP を集めると，現在，BOTH-依存に含まれているものは P,C-依存のいずれかに分類されると考えられる．

4.2 翻訳文の評価

最後に，翻訳文全体の評価で提案手法の妥当性を検証した．これは BLEU スコア [7] を用いて行った．BLEU スコアは翻訳結果で正解に出現する N-gram の幾何平均である．実験では N=3 とし，NHK ニュース記事の先頭文 240 文 (正解例数 4) をテストセットとして，次の 3 つのシステムを比較した．

1. **BASELINE**: CAP を用例として登録しない用例ベース翻訳システム [1].
2. **CAPMT**: BASELINE の用例に加えて，汎化をせずに CAP 全体を用例として利用したシステム．
3. **CAPMT+**: BASELINE の用例に加えて，自動分類結果から汎化した CAP を用例として利用したシステム．

表 5: 翻訳例

入力文	... アフガニスタン北東部で 起きた 地震の被災地では ...
正解例	... quake struck areas along northeastern Afghanistan ...
BASELINE	... disaster area of the earthquake occurred in afghanistan northeast ...
CAPMT+	... disaster area of the earthquake in afghanistan northeast ...
入力文	アメリカのメリーランド州で十四日に 行われた 航空ショーで...
正解例	An air show in the US state of Maryland on the 14th ...
BASELINE	Air show was held in maryland of the united states on the 14th ...
CAPMT+	Air show in maryland of the united states on the 14th ...
入力文	... 二十五日に 行われる 日韓首脳会談に...
正解例	... summit due to be held on the 25th.
BASELINE	... summit meeting conducted on 25th.
CAPMT+	... summit meeting on 25th.

* 提案手法により省略される用言を下線部で示した.

ただし，テストセットの中にはCAPMTやCAPMT+によって省略が実現しない文を含まれている．そこで，これらの手法により省略が実現された場合だけをサブセットとし，この精度も比較した．

実験の結果，CAPMTでは240文中14文で省略が行われ，CAPMT+では240文中104文で省略が行われた．その精度を表4に示す．CAPMT+では240文のうち104文で省略が行われ，サブセットでの精度はベースラインよりも4.0%向上している．これは，CAPMT+により翻訳文の一部しか変化がおきないことを考慮すれば，手法の適切さを示唆する向上だと考えられる．これに対して，CAPMTでは240文のうちわずか14文しか省略が行わず，その精度の差は有意な差とはいえない．

翻訳結果の具体例を表5に示す．表5が示すように，提案手法(CAPMT+)は推論可能表現を適切に省略している．ただし，表5最下例のように，必ずしも正解例が推論可能表現を省略しない場合も存在し，逆に精度を下げる場合もあった．提案手法は推定された推論可能表現を必ず省略するが，今後の課題として，省略するかどうかの判定を行うことが考えられる．

5 おわりに

本稿では，用言対応コーパスの作成法を述べた．また，用言対応コーパスを観察した結果から，推論可能表現の出現パターン(CAP)を発見・分類し，さらに自動分類する手法について述べた．実験結果は提案手法の適切さを示しており，今後，対訳コーパスの量が増え，より多くのCAPを収集できれば，さらに機械翻訳の扱う表現の幅が広がると考えられる．

参考文献

[1] Eiji Aramaki, Sadao Kurohashi, Hideki Kashioka, and Hideki Tanaka. Word selection for ebmt based on monolingual similarity and translation confidence. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 57–64, 2003.

[2] Eiji Aramaki, Sadao Kurohashi, Satoshi Sato, and Hideo Watanabe. Finding translation correspondences from parallel parsed corpus for example-based translation. In *Proceedings of MT Summit VIII*, pp. 27–32, 2001.

[3] Peter F. Brown, Stephen A. Della Pietra, Vi cent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, No. 2, 1993.

[4] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of NAACL 2000*, pp. 132–139, 2000.

[5] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, Vol. 20, No. 4, 1994.

[6] Makoto Nagao. A framework of a mechanical translation between Japanese and english by analogy principle. In *Artificial and Human Intelligence*, pp. 173–180., 1984.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pp. 311–318, 2002.