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

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] などの大量のデータを用いた機械翻訳 (MT) に関する研究が盛んに行われている．これまでの研究はおもに旅行会話 [9, 10] やマニュアル [8] などのドメインを扱って部分的に成功を収めてきた．しかし，現在の MT 研究で翻訳における表現のずれを扱ったものは少ない．次に入力文 S の人間の翻訳結果 T_{human} と機械翻訳結果 T_{mt} を示す．

S: カナダで 開かれた 通商会議で...
T_{human}: At a trade conference in Canada...
T_{mt}: At a trade conference held in Canada...

T_{mt} のように機械翻訳システムは逐語的に翻訳する傾向がある．それに対して，人間は “開かれた” 表現を明示的に訳出していない．この理由は，人間が冗長な表現を避け，コンパクトな訳を好むからだと考えられる．本稿ではこのように推論可能で訳出されない表現を**推論可能表現**と呼ぶことにする．先の例のように，用言はしばしば推論可能表現となることに加えて，用言の句アライメント推定の困難さは従来から指摘されてきた [2].

そこで我々は，用言の翻訳のされ方を調べるために，対訳文に対して，対訳文に対して次の 2 つの情報 (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，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，用言の省略に関する知見を得たので報告する．

2 用言対応コーパス

用言対応コーパスとは，まず自動で対訳文の構造化と対応付けを行い，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．

2.1 NHK ニュースコーパス

用言対応コーパスを作るためには，比較的自由な翻訳がなされたコーパスが必要となる．NHK 対訳ニュースコーパスから抽出した対訳文を用いた．NHK 対訳ニュースコーパスとは，日本語記事がまずあり，それをもとに英語記事が自然な翻訳により作成されたもので，日英 4 万記事ペア (5 年間分) からなる．

2.2 アノテート 作業

用言対応コーパスのアノテーションは，次の 4 つのステップからなる．

STEP 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．

STEP 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．

STEP 3: 句のアノテーション
作業者が用言句をアノテートする．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を含む句とする．したがって，動名詞 (gerund) を含んでいれば，通常，PP と扱われている句も用言句と考える．

STEP 4: 対応関係のアノテーション
作業者が日本語側の用言に対応する相手言語の個所をアノテートする．日本語の用言句は必ずしも英語

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．

の用言句に対応しているとは限らないため，我々は用言句以外の対応先もアノテーターに許す．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．同様に，英語側の用言に対しても，対応する日本語の個所をアノテートする．

以上の手順で 5500 対訳文の対応対応付けの作業を行った．

2.3 用言対応コーパスの分析

用言対応コーパスでは，日本語の用言句が必ずしも英語側の用言句と対応しない．日本語用言句が英語側でどのように表現されているかという観点から，分類と集計を行った (表 1)．その結果，両言語の用言同士の対応以外の対応の数が 40%以上あり，無視できない現象であることが分かる．また，その中でも特に用言句-ϕの割合が多い．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) 省略可能なコンテキストで出現した用言である場合．

後者の省略可能な用言とは，1 章での “開かれた” のような場合で，以下のように示される:

.

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．

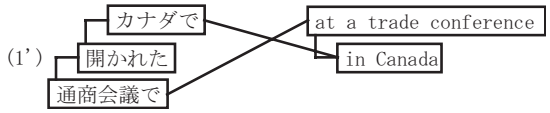
3 省略される用言の学習

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，前章の例では周辺の句が次のように対応している:

表 2: CAP の妥当性の判定

判定	分類	#
good	P-CONTEXT	21
	C-CONTEXT	16
	BOTH-CONTEXT	19
	計	56
bad	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

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



.

この形は，下のように日本語用言句の親 (parent, 以降 P) と子 (parent, 以降 C) の両方の対応先が英語側でも，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) の句のペアを Condensed Alignment Pattern (以降，**CAP**) と呼ぶことにする[1]．

ここで，もし，CAP のコンテキストで出現する用言は省略してもよいならば，CAP を収集し用例として持ちいれば，用言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用例ベース翻訳システム [1] で用いている句アライメント済みの対訳文から CAP を収集し，表現されていない用言が妥当かどうかを調査した．

その結果，表 2 に示されるように，アライメント失敗や統語解析失敗である場合をのぞいて，CAP における用言の省略は適切であることが分かる．よって，CAP を用例として用いれば用言の省略が実現できると考えられる．

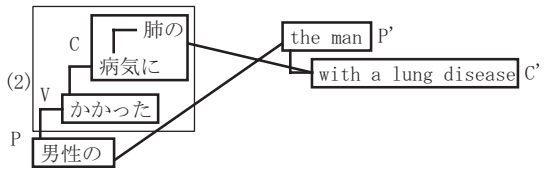
しかし，この CAP 全体をそのまま翻訳用例として使うのでは，用例として利用できる機会は少ない．そこで，CAP の頻度を集計し，P か C のいずれか

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

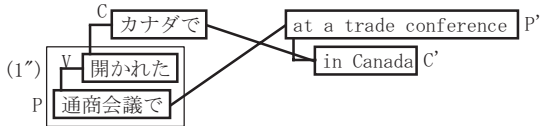
のコンテキストが強くはたらいて用言が省略されているのかという観点から CAP をさらに分類した．

3.1 CAP の分類

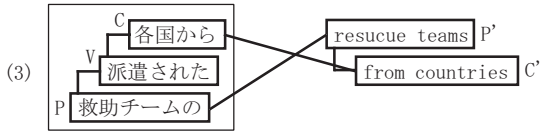
1. **C-CONTEXT**: C のみがコンテキストとなる場合．
   次の例では，C“肺の病気” により，V“かかる” が推論されるため，V が訳出されていない．



2. **P-CONTEXT**: P のみがコンテキストとなる場合．
   先とは逆に，P によって V が推論される場合がある．例えば，前の章の (1) がそうであり，次のように示される．



3. **BOTH-CONTEXT**: P と C の両方がコンテキストである場合．
   P と C の両方がそろってはじめて V が推論される場合がある．このような場合は，CAP 全体を用例と考えるしかない．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．



3.2 頻度による学習

前節の分類はグレイスケールで主観的なものであり，すべての CAP について，そのコンテキストの所在を明確にすることは困難である．しかし，P が明らかにコンテキストである場合は，P と V を含んだ CAP が多数存在するはずである．そこで，CAP を P,V,P’ と V,C,C’ の 2 つの CAP の断片に分けて集計すれば，省略される用言にとって必要なコンテキストを推定できると考える．

そこで，次の手続きで CAP のコンテキストを推定した．

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．
2. CAP を 2 つの CAP の断片，(C,V,C’), (V,P,P’) 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．
3. 集計の結果，$freq(P) > freq(C) \times 2$ ならば，P-CONTEXT と考える．逆に，$freq(C) > freq(P) \times 2$ ならば，C-CONTEXT と考える．それ以外は，BOTH-CONTEXT と考える．

表 3: 推定された CAP のコンテキスト

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

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．句アラメントされた 52749 対訳文から CAP を抽出したところ，ここから延べ 4219 個の CAP を収集できた．すなわち，12 対訳文に 1 つから CAP を取り出せることが分かる．

そのうちコンテキストの推定された割合は表 3 のようになり，BOTH-CONTEXT と判定された数が多い．しかし，このうちのほとんど (2272 個) は 1 回しか出現しない CAP であった．よって，今後，より多くの CAP を集めると，現在 BOTH-CONTEXT に含まれているものは P,C-CONTEXT に分類されると思われる．

表 4: BLEU スコア

	Testset [240]	Subset [104]	Subset [14]
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．[] 内は文数を示す.

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考えてみる．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．我々は N=3 を用い，NHK ニュース記事の先頭文 240 文を無作為抽出し，そのそれぞれに対して 4 人が正解を作成した．この正解は実際に NHK で翻訳を行っているプロの翻訳者よって行われた．

実験は，3 つの異なる用例を用いたシステムを用いて行った．

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，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，コンテキストが推定された CAP を用例として利用したシステム.

また，実験セットの中には 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．まず，240 文全体では CAP,CAP+とも BLEU スコアは大きく上昇しない．しかし，CAP+では 240 文のうち 104 文で省略が行われ，この精度はベースラインよりも 4.0%向上している．これに対して，CAPMT では 240 文のうちわずか 14 文しか省略が行われない．その精度は，10%以上向上しているが数が少ないため有意な差とはいえない．

5 おわりに

CAP が適応された場合の精度が向上することから，本手法は理論的側面では有効である．実験ではテストセット全体の翻訳精度を大きく向上することはできなかったが，これは対訳コーパスから得られる CAP が少なかったのが大きな原因と考えられる．しかし，ニュース原稿は毎日増加し続けているため，ますます問題の解決は容易になると考えられる．

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.

[8] Stephen D. Richardson, William B. Dolan, Arul Menezes, and Monica Corston-Oliver. Overcoming the customization bottleneck using example-based mt. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pp. 9–16, 2001.

[9] Eiichiro Sumita. Example-based machine translation using dp-matching between word sequences. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pp. 1–8, 2001.

[10] Wolfgang Wahlster. Verbmobil: Translation of face to face dialogs. In *Proceedings of MT Summit IV*, pp. 127–135, 1993.