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

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 [3] や統計ベース翻訳 [2] 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．しかし，これまでの 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．例えば，S を翻訳する際に次の 2 つの翻訳結果 T_1 と T_2 が考えられる．

S: カナダで 開かれた 通商会議で...
T_1: At a trade conference held in Canada...
T_2: At a trade conference in Canada...

T_2 は “開かれた” 表現を明示的に訳出していないが，前後のコンテキストから理解可能な翻訳となっている．本稿ではこのような表現を**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．

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

2.1 用言対応コーパスの作成

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．作成した例を図 1 に示す．用言句 (赤下線) に対して，人手で対応関係が付与されている (濃い緑)．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) をアノテートした．このようにして，5500 対訳文の対応対応付けの作業を行った．

2.2 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集計を行った (表 1)．その結果，用言句-ϕ の割合が多いことが分かる．用言句-ϕ 対応は，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，英語 2 句の句のペアを Condensed Alignment Pattern (以降，**CAP**) と呼ぶことにし，V が推論可能となるコンテキストの位置から，次の 3 つタイプに分類した．

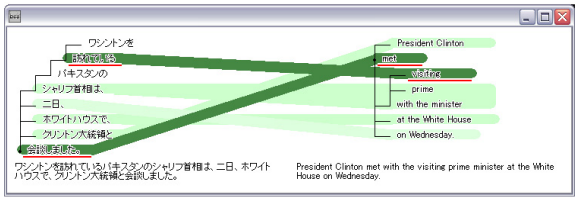


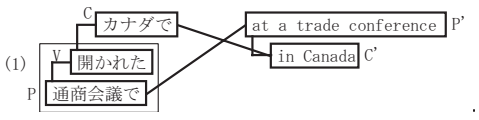
図 1: 用言対応コーパス

表 1: 用言対応の分類と数

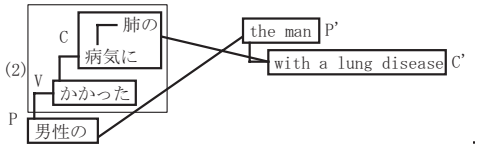
用言対応の分類 (日本語:英語)	対応数
用言句-用言句	9779
用言句-ϕ	6831
用言句-前置詞句 または 用言句-名詞句	*710*
その他	*316*

* 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．

1. **P-CONTEXT**: P によって V が推論されるタイプ．例えば，前の章の例がそうであり，次のように示される．



2. **C-CONTEXT**: 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” が推論されている．



3. **BOTH-CONTEXT**: P と C の両方がそろってはじめて 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．

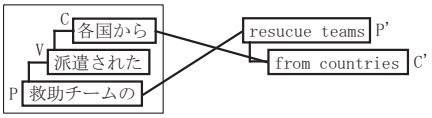
表 2: BLEU スコア

	Testset [240]	Subset [104]	Subset [14]
BASELINE	24.6	24.7	26.3
CAPMT	24.8	-	29.0
CAPMT+	25.0	25.7	-

* [] 内は文数を示す.

表 3: 抽出された CAP と推定されたコンテキスト

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

(3)


.

3 用言の簡潔な翻訳の実現

3.1 提案手法

CAP を翻訳用例として用いることで，用言の簡潔な翻訳を実現できる．しかし，CAP 全体をそのまま用例として用いると，日本語 3 句が入力文と一致する必要があるため，その利用機会は少ない．そこで，CAP のコンテキストを推定し，コンテキストでない部分を汎化した用例とすること手法を考える．これは，P が明らかにコンテキストである場合は，P と V を含んだ対訳文が多数存在するはずという仮定にもとづき，次の手続きによって行った．

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 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．

3.2 実験と考察

提案手法の有効性を確かめるため，翻訳用例 (52,749 対訳文) 中から CAP を抽出・自動分類 (表 3) を行い，CAP を翻訳用例として用いた場合の翻訳精度を調べた．これは，次の 3 つのシステムの出力する翻訳結果 (240 文) を BLEU スコア (N=3)[4] によって評価することによって行った．

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．

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，省略が実現された (翻訳結果が異なった) 場合だけの精度も比較した．

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．それぞれの精度を表 2 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 1.0 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ず，有意な結果とはいえない．

4 おわりに

用言の簡潔な翻訳を実現する手法を提案した．これは，CAP を翻訳用例として用いる手法と，CAP のコンテキストを推定する手法からなる．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，対訳コーパスは日々増加し続けているため，今後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] 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.

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

[4] 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.