DP を使用した重文複文の日英翻訳の精度

岡田 敏　村上 仁一　徳久 雅人　池原 悟
鳥取大学大学院工学研究科
{sokada,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

機械翻訳の方式として, 従来要素合成法を基本とした研究開発が行われてきた. 例えばトランスファー方式を基本とするシステムでは, 入力文の各要素を翻訳する過程と入力文の構文構造を目的言語の構文構造へ変換する過程を持ち, 両者を合成することにより目的言語の表現を作成する. これは, 構文構造と表現の意味が別々に処理されており, 言語表現全体の意味が構成要素に分解される過程で失われ, 目的言語を合成する過程で復元されない問題がある. 近年, この問題を克復することを狙い動的計画法を用いた用例翻訳方式 [1] やパターンを利用する翻訳方式 [7] が研究されている.

しかし, 一般的な動的計画法を用いた用例翻訳方式では英文を作成する際に日英文の構成要素の対応付けが複雑になる問題がある. また, パターン翻訳方式では変数化された全ての要素を翻訳しなければならないため翻訳精度が低くなる可能性がある.

そこで本稿では動的計画法を用いた用例翻訳と文型パターン翻訳の特性を合わせ持つ翻訳方式を提案する. 両者を併用することで, 用例翻訳方式における英文作成の難しさとパターン翻訳方式における変数の問題を克服する. 具体的には, まず動的計画法で入力文に似た類似文を検索する. 次に入力文と類似文の相違箇所を単語辞書で翻訳する. 最後に類似文の文型パターンを用いて英文を生成する.

実験は要素合成法では翻訳が難しいとされる重文複文を対象に提案手法の日英翻訳精度を検証する.

2　対訳用例文集と文型パターン

2.1　対訳用例文集

用例翻訳方式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. 近年電子辞書から対訳用例を収集することができるようになった [2]. これを利用して複文・重文を主とする大規模対訳用例文集を作成した. 本稿ではこのうち重文・複文を主に収集した約 12 万文の対訳用例文集を使用する.

2.2　文型パターン

提案手法では英文生成の際に類似文の英語文型パターンを使用する. 文型パターンとは, 対訳用例文集の日英言語の線形要素を変数記号に書き換えた表現である [3]. ここで線形要素とは言語表現を構成要素と同等の意味を持つ他の要素に置き換えても元の言語表現全体の概念が変化しない要素である.

対訳用例文集は表 1 の構造をしており, 対訳用例全てに文型パターンが用意されている. 文型パターンの変数を介することで日英言語の翻訳対において構成要素の対応付けをすることができる.
表 1 では線形要素 (私,I) と (おば,my aunt) を"N1","N2"に置換している.

表 1: 対訳用例文集の構造

わたしは**おば**を頼って上京した
N1 は/**N2** を/頼って/上京した
I came to Tokyo from the country
counting on **my aunt's** help.
N1 came to Tokyo from the country
counting on **N2 ' s** help.

3　提案する翻訳方式

3.1　本稿で用いる翻訳方式

本稿では動的計画法による用例翻訳と文型パターン翻訳の特性を合わせ持つ翻訳方式を提案する. 具体的には, まず動的計画法で入力文に似た類似文を検索する. 次に入力文と類似文の相違箇所を単語辞書で翻訳する. 最後に類似文の文型パターンを用いて英文を生成する.

以下に動的計画法による用例翻訳方式やパターン翻訳方式の違いを述べる.

3.2　動的計画法による用例翻訳方式との違い

動的計画法による用例翻訳方式 [1] では入力文と類似文の相違箇所を単語辞書で翻訳し, 類似文の翻訳例の適切な箇所と置換して翻訳する. 英文生成の際, 文型パターンを使用しない. しかしそのため, 文法等による表現の変化により構成要素の対応付けが複雑になる.

提案手法では英文生成に文型パターンを利用することで比較的容易に構成要素の対応付けができる. 表 1

では” わたし” は”N1” と,“N1” は”I” と対応している. よって” わたし” と”I” は対応付けることができる. 同様に” おば” は”my aunt” と対応付けることができる.

3.3 パターン翻訳方式との違い

パターン翻訳方式では入力文と文型パターンをパーサを用いて照合し, 変数部分を翻訳することで英文生成を行う [7]. しかし, 文型パターンは線形要素を全て変数化している. よってパターン翻訳方式では全ての変数を翻訳しなければならない. 一方, 提案手法では入力文と類似文の相違箇所のみを翻訳すればよい.

翻訳する箇所が増えると翻訳精度が減少するため, 提案手法はパターン翻訳方式よりも翻訳精度が高まる可能性がある.

4 翻訳の概要

4.1 動的計画法による類似文検索

提案手法では入力文を類似文の文型パターンを用いて翻訳する. よって入力文と類似文との類似度が高いほど高品質な翻訳が作成できる.

本稿では類似度を, 構成する単語の相違数で決定する. 単語相違数が少なければ類似文の類似度は高くなる. 単語相違数の計算には動的計画法を用い計算機で求める. 相違には三種類あり, 入力文の単語が余剰である削除誤り, 類似文の単語が余剰である挿入誤り, お互いの単語を交換することで相違箇所を解消できる置換誤りがある. 各誤りの合計が少ない例文を類似文とする.

本稿では類似文が複数個検索された場合, 最も良い英文が生成できる文型パターンを持つ類似文を人手で選択する.

4.2 文型パターンを利用した英文生成

本稿では単語置換誤り箇所を人手で翻訳し, 類似文の文型パターンを用いて英文を生成する.

4.2.1 単語置換誤り箇所の対応

各誤り箇所は計算機で対応付ける. 表 2 では (私, わたし),(彼, おば) が対応する単語置換誤り箇所である.

表 2: 単語置換誤り箇所の対応付け

入力文: **私**/は/ **彼** /を/頼っ/て/上京し/た
類似文: **わたし**/は/**おば**/を/頼っ/て/上京し/た
私 ↔ **わたし**
彼 ↔ **おば**

4.2.2 日英文間の単語対応付け

文型パターンの変数を介することで日英文間の単語対応付けを行う. 表 3 では (私,N1),(おば,N2) が対応している.

表 3: 類似文と翻訳例での単語の対応付け

わたし/は/**おば**/を/頼っ/て/上京し/た
N1 は/**N2** を/頼って/ 上京した
I came to Tokyo from the country
counting on my aunt’s help.
N1 came to Tokyo from the country
counting on **N2 ’ s** help.
わたし ↔**N1**
おば ↔**N2**

4.2.3 置換誤り箇所の単語辞書引き

入力文と類似文間で対応する単語置換誤り箇所を人手で辞書引きする. 辞書引きした英単語候補のうち, 最も適切な単語を人手で選び翻訳に使用する. 単語置換謝り箇所が動詞の場合は形態素解析情報から単語の原型を調べ辞書引きする. 表 2 で置換誤り箇所の” 私”,” 彼” を辞書引きする.

4.2.4 対応箇所の置換

辞書引きした単語を類似文の英語文型パターンの対応箇所と置換し英文を生成する. なお置換の際, 文型パターンに沿った形に単語を変形させる.

表 4 では” 私”,” 彼” の翻訳 “I”,”he” を英語パターンの”N1”,”N2” と置換する. “he” を置換する際, 英語文型パターンの “N2’s” の形に合わせて”his” と変形する. 表 5 に作成した英文を示す.

表 4: 英文生成

私 ↔**I**↔**N1**
彼 ↔**he**↔**N2**
N1 は/**N2** を/頼って/上京した
N1 came to Tokyo from the country
counting on **N2 ’ s** help.

表 5: 出力英文

I came to Tokyo from the country
counting on his help.

5 評価実験

提案手法の翻訳精度を検証するために 4 章で述べた手順で翻訳実験をする. 入力文セットとして重文・複文の日本文 100 文を使用する.

表 6: 日英対訳用例文集

日英対訳用例文集		
文数	総単語数 (日本文)	平均単語数 (日本文)
127813	1564515	12

表 7: 入力文セット

入力文セット		
文数	総単語数 (日本文)	平均単語数 (日本文)
100	1216	12

5.1 実験条件

入力文 100 文を 4 章で述べた方法で翻訳する. 形態素解析には形態素解析プログラム ALTJAWS の結果を人手で修正したデータを使用する. 類似文検索には動的計画法プログラムである HTK[4] の HResult を使

用する．類似文の判定は単語相違数 4 までとする．辞書はランダムハウス和英辞典を使用する．

5.2 評価方法

5.2.1 人手による評価

作成した翻訳を人手により 5 段階に評価する．評価基準は [1] で使用された基準を使用し, さらに評価 E を加えて評価する．評価者は 1 名で評価する．

A:出力英文が入力文の翻訳として使用可能

B:簡単な文法間違いや任意要素の欠如があるが簡単に修正可能

C:入力文を部分的に翻訳

D:出力英文が入力文の翻訳として使用不可能

E:文型パターンの変数化問題で翻訳に失敗

評価 E については考察で詳しく述べる．

5.2.2 機械による評価

翻訳の質を機械で自動的に評価する方法として BLEU アルゴリズムと NIST アルゴリズムがある．

BLEU は米国の IBM が提案したアルゴリズム [5] で,n(n ≧ 1) 文の模範訳の 4-gram 単語共起情報を用いて出力英文を評価する．BLEU によるスコアは 0〜1 であり, 模範訳と出力英文が同じ文であれば 1 になる．

NIST は米国 NIST が提案したアルゴリズム [6] で,BLEU と同じく n 文の模範訳の 5-gram 単語共起情報を用いて出力英文を評価する．

本稿では NIST MT evaluation kit version 9[6] を使用する．また, 模範訳は 1 文で評価する．

6 実験結果

日本語入力文 100 文中 38 文において単語相違数 4 以下の類似文が対訳用例文集から発見できた．入力文 1 文あたりの平均類似文数は 50 文である．類似文検索ができた 38 文に対して各単語相違数と平均類似文数の関係を表 8 に, 翻訳した結果を表 9 に示す．

提案手法では翻訳に有効な類似文を検索できる反面, 類似文が検索できた入力文の数は少なく, 結果として翻訳成功数は少なかった．

表 8: 各単語相違数と平均類似文数の関係

相違数	文数	平均類似文数
0	2	1.0
1	5	1.0
2	7	2.9
3	13	10
4	33	53

表 9: 提案手法による翻訳結果

調査数 (文)	A	B	C	D	E
38	32% (12)	24% (9)	13% (5)	29% (11)	2.7 (1)

調査数 (文)	BLUE	NIST
38	0.30	4.60

6.1 各評価における出力英文の例

評価 A から D の出力英文を表 10 に示す．表中のパターンは英文生成に使用した類似文の英語文型パターンである．

表 10: 各評価における出力英文の例

評価 A	
入力文	**彼にはその任務を果たせるだけの能力がなかった**
	He was not equal to the task.
パターン	N1 had little N5 to V4 ♯2[N1.poss] N3.
出力英文	**He had little faculty to perform the task.**

評価 B	
入力文	**雄弁は政治家に必要な属性だ**
	Eloquence is a necessary attribute in a politician.
パターン	N1 be AJ4 N5 ♯2[for N3].
出力英文	**Eloquence is necessary attribute for politician.**

評価 C	
入力文	**家を出たとたん空に稲妻が走った**
翻訳例	The moment I stepped out of my house, there was a flash of lightning in the sky.
パターン	The minute (N1|I) left N2 , N3 V4.past
出力英文	**The minute I left home, lightning runed.**

評価 D	
入力文	**雄弁は政治家に必要な属性だ**
翻訳例	Eloquence is a necessary attribute in a politician.
パターン	N1 be a useful N2 to have.
出力英文	**Eloquence is a useful politician to have.**

7 考察

7.1 パターン翻訳方式との比較

文献 [7] では本稿と同じ対訳用例文集と入力文セットを使用してパターン翻訳実験をしている．翻訳実験の結果を表 11 に示す．

パターン翻訳の結果と比較すると, 提案手法は B が多く D が少ない結果となった．また BLUE,NIST の値は提案手法が高い値を示した．ただし, 使用するパターンの決定は人手で行っているため公平な評価にはならない．

表 11: パターン翻訳方式の実験結果

調査数 (文)	A	B	C	D
48	32% (16)	10% (5)	15% (7)	43% (21)

調査数 (文)	BLUE	NIST
48	0.19	3.3

7.2 動的計画法を用いる用例翻訳方式との比較実験

動的計画法を用いる用例翻訳方式では入力文と類似文の相違箇所を翻訳し, 類似文の翻訳例の適切な箇所と置換して英文を生成する．そのため文型パターンを用いない．

動的計画法を用いる用例翻訳方式で英文作成を行った結果を表 12 に示す. 表 9 と比較すると正解率は同じだが, 提案手法は C が増え D が減少した. よって, 提案手法は文型パターンを利用することで動的計画法を用いる用例翻訳方式より良い結果が得られていることが示された.

表 12: 一般的動的計画法を用いる翻訳実験結果

調査数 (文)	A	B	C	D
38	37% (14)	18% (7)	11% (4)	34% (13)

調査数 (文)	BLUE	NIST
38	0.29	3.81

7.3 名詞置換による実験

6 章では日本文の単語字面情報を使用して類似文検索をしたが, 単語相違数 4 以下の類似文検索に成功する入力文は少なかった. そこで, 名詞を N と置換することで類似文の検索精度を向上させることを試みた.

用例文集の日本文全ての名詞を形態素解析情報を用いて N へと変換する. 入力文セットも名詞を N へと変換し類似文検索をする. 翻訳実験手順は単語単位の実験と同様に行う.

7.3.1 名詞変換による類似文検索

名詞変換した類似文検索で, 日本語入力文 100 文中 62 文が単語相違数 4 以下の類似文が対訳用例文集から発見できた. また, 入力文 1 文あたりの平均類似文数は 686 文に増加した. 類似文検索ができた 62 文に対して単語相違数と平均類似文数の関係を表 13 に示す.

7.3.2 名詞変換による類似文検索結果からの英文生成

名詞変換による類似文検索結果を利用して英文作成をした結果を表 14 に示す. 単語字面情報を使用した実験よりも類似文検索精度が向上し, 結果として翻訳作成できた入力文の数が増加した.

表 13: 各単語相違数と平均類似文数の関係

相違数	文数	平均類似文数
0	5	3.4
1	11	8.6
2	32	42
3	50	148
4	60	561

表 14: 名詞変換による翻訳実験結果

調査数 (文)	A	B	C	D	E
62	26% (16)	24% (15)	15% (9)	31% (19)	6.5 (3)

調査数 (文)	BLUE	NIST
62	0.19	3.54

類似文検索に失敗した入力文を見ると, 動詞の置換や名詞連続複合語を 1 つの N と置換することでさらに翻訳精度が向上する可能性がある.

7.4 評価 E の問題

本稿で提案した翻訳方式では, 文型パターンで変数化されていない箇所が相違箇所として対応している場合, 辞書引きした単語を類似文の英語パターンのどの箇所と置換すればいいのか判断できないため翻訳に失敗することがある.

表 15 では (夏祭り, あなた),(たいへんな, ずいぶんな),(人出, 問題) が対応する相違箇所である. 類似文の文型パターンでは" 問題" が変数化されていない. よって" 人出" の訳"turnout" が"trouble" の位置と置換できず, 翻訳に失敗する.

評価 E の問題を回避するためには文型パターンの変更が考えられる. 本稿で使用した文型パターンは文を構成する要素の線形性に着目し, 文を分類する目的で作られている. しかし, 英文生成の観点から考え, 日英言語の翻訳対で対応する要素全てを変数化したパターンを利用することで評価 E の問題を回避できる.

表 15: 評価 E の問題

入力文:**夏祭り**/に/は/**たいへんな**/**人出**/です/ Lots of people turn out for the summer festival.
類似文:**あなた**/に/は/**ずいぶんな**/**問題**/です/ That's a lot of trouble for you. N1 /に は/ ずいぶんな/問題です That ' s a lot of trouble for N1.

8 おわりに

本稿では動的計画法による用例翻訳と文型パターンによる翻訳の特性を合わせ持つ翻訳方式を提案し, 翻訳実験を行った. 翻訳対象に従来の要素合成法を基本とする翻訳システムでは難しいとされている複文・重文を用いた. 実験の結果 38%のカバレージ, 正解率 56%で入力文の翻訳を得ることができた. よって提案手法が複雑な文に適応できる可能性が示された.

今後はデータベースの拡充をするとともに, 翻訳に有効な類似文の検索手法を検討する必要がある.

参考文献

[1]Sumita Eiichiro:Example-based machine translation using DP-matching between word sequences,DDMT workshop of 39th ACL,2001

[2] 村上 池原 徳久:日本語英語の文対応の対訳データベースの作成,「言語, 認識, 表現」第 7 回年次研究会,2002 年 11 月

[3] 池原:非線型な言語表現と文型パターンによる意味の記述, 情報処理学会, 自然言語処理研究会,2004-NL-159,pp.139-146,2004-1

[4] Hidden Markov Model Toolkit(HTK) http://htk.eng.cam.ac.uk/

[5]Kishore Papineni,Salom Roukos,Todd Ward,Wei-Jing Zhu:BLEU:a Method for Automatic Evaluation of Machine Translation,IBM Research Report,September17,2001

[6]NIST:Automatic Evaluation of Machine Translation Quality Using N-gramCo-Occurrence Statistics,http://www.nist.gov/speech/tests/mt/mt2001/resource/,2002

[7] 前田 村上 徳久:パターンを使用した重文複文の日英翻訳の精度, 言語処理学会第 10 回年次大会,2004 年 3 月発表予定