主語補完及び文法構造変換を前処理に用いた日英統計翻訳

古市将仁 村上仁一 徳久雅人 村田真樹
鳥取大学大学院　工学研究科 情報エレクトロニクス専攻
{s072047, murakami, tokuhisa, murata} @ ike.tottori-u.ac.jp

1　はじめに

近年の機械翻訳では，統計翻訳が主流となっている．

機械翻訳の問題点の 1 つとして，日本語において省略格要素が存在する文における翻訳精度は低いことが挙げられる．このような問題を解決するために，中岩氏らは，ゼロ代名詞 (日本語において省略されている格要素) の解析を行い，ゼロ代名詞照応解析の方法を提案した [1][2][3][4]．さらに先行研究では，日英統計翻訳において，主語が省略されている日本語文 (以後，主語省略文とする) に対して，主語を付与 (以後，主語補完とする) し，翻訳精度の向上を報告した [5].

また，統計翻訳において，日本語-英語など，文法構造が異なる言語間では，翻訳精度が低い．この問題を解決するために，統計翻訳の前処理として，翻訳前の言語の文法構造を翻訳後の言語の文法構造に近づける (以後，文法構造変換とする) 研究が行われている [6][7][8].

そこで本研究では，主語補完と文法構造変換を組み合わせることで，翻訳精度の向上を目指す．まず，主語省略文に対し，主語補完を行う．次に，日本語の文法構造を英語の文法構造に近づけるために，品詞情報と文節情報に基づいた単純なルールを用い，文法構造変換を行う．最後に，句に基づく統計翻訳を行い，翻訳精度の変化を調査する．

2　提案手法

本研究では，日本語文に対して，主語補完と文法構造変換を併用し，翻訳精度を向上させる．まず，主語省略文に対し，主語補完を行う．次に，品詞情報と文節情報に基づいた単純なルールを用い，文法構造変換を行う．最後に，統計翻訳を行い，翻訳精度の変化を調査する．提案手法の学習部と翻訳部の手順を，それぞれ以下に示す．

1) 学習部

提案手法の学習における手順を以下に示す．

手順 1　学習データの主語省略文に対して，対訳英語文を参照し，主語を補完する．
手順 2　学習データの日本語文に対し，S(主語)V(動詞)O(目的語) の形式に文法構造変換を行う．
手順 3　上記で得たデータより，翻訳モデルと言語モデルを学習する．

2) 翻訳部

提案手法の翻訳における手順を以下に示す．

手順 1　テストデータの主語省略文に対して，文頭に “私は”を補完する．
手順 2　テストデータに対し，S(主語)V(動詞)O(目的語) の形式に文法構造変換を行う．
手順 3　上記で得たデータに対して，統計翻訳を行う．

主語補完については 3 章，文法構造変換については 4 章で詳細を示す．

3　主語補完

3.1　主語省略文の判断条件

日本語文において，主語省略文が存在する．表 1 に主語省略文の例を示す．

表 1　主語省略文の例

昼食をたっぷり取った。
サッカーをした。

また，日本語文において “しろ”や，“しなさい”のような動詞を含む命令文には，主語が無いことが多い．

そこで本研究では，主語と主語省略文の判断条件の定義を行った．主語の定義を表 2 に示す．また，主語省略文の判断条件を表 3 に示す．

表 2　主語の定義

定義 1	名詞の後に “は”がある文節
定義 2	名詞の後に “が”がある文節
定義 3	名詞の後に “も”がある文節

表 3　主語省略文の判断条件

判断条件 1	助詞 “は”，“が”および “も”が含まれない文
判断条件 2	動詞が命令形ではない文

3.2　テストデータに対する主語補完

テストデータに対しては，主語省略文を抽出し，文頭に “私は”を補完する．以下に，テストデータに対する主語補完の手順を示す．

手順 1　日本語文に対し，形態素解析を行う．
手順 2　形態素解析を行ったデータを参照し，表 3 で定義した判断条件 1 と判断条件 2 を満たす主語省略文を抽出する．以下に抽出される文と抽出されない文の例を示す．

表 4　日本語補完抽出文例

抽出される文 (主語省略文)	昼食 を たっぷり 取っ た 。
抽出されない文	彼 は 山 へ 行く 。

手順 3　手順 2 で抽出した主語省略文に対し，文頭に “私は”を補完する．以下に例を示す．

表 5　主語補完例

主語補完前	昼食をたっぷり取った。
主語補完後	私は昼食をたっぷり取った。

3.3　学習データに対する主語補完

学習データの主語省略文に対する主語補完は，対訳英語文を参照する．対訳英語文を参照することで，精度の高い主語補完が可能である．学習データに対する主語補完の手順を以下に示す．

手順 1　日本語データに対し，形態素解析を行う．
手順 2　形態素解析を行ったデータを参照し，表 3 で定義した判断条件 1 を満たす文のみを抽出する．抽出する文の例を表 6 に示す．

表 6 抽出文例

日本語文	昼食をたっぷり取った。
対訳英語文	He had a big lunch .
日本語文	声を低くしなさい。
対訳英語文	Keep your voice down .

手順 3 手順 2 で抽出した文に対する対訳英語文の文頭単語を抽出する．文頭単語の抽出の例を図 1 と図 2 に示す．

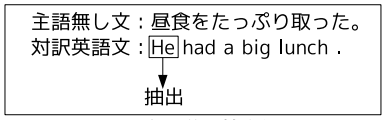

図 1 文頭単語抽出例 1

図 2 文頭単語抽出例 2

手順 4 抽出した対訳英語文の文頭単語を，変換規則に従って日本語に変換する．変換規則を表 7 に示す．なお，図 2 のように，抽出した対訳英語文の文頭単語が表 7 の変換規則に適応しなかった場合，主語補完は行わない．

表 7 文頭単語の日本語への変換規則

対訳英語文の文頭単語	変換する日本語
I	私は
He	彼は
She	彼女は
We	私たちは
You	あなたは
They	彼らは
It	それは
Someone	誰かが
Anyone	誰かが
Somebody	誰かが
Anybody	誰かが

手順 5 変換した日本語を，手順 2 で抽出した文の文頭に補完する．例を表 8 に示す．

表 8 主語補完例

主語補完前	昼食をたっぷり取った。
主語補完後	彼は昼食をたっぷり取った。

4 文法構造変換

本研究では，日本語の文法構造を S(主語)V(動詞)O(目的語) の形式に変換するために，文法構造変換を行う．手順を以下に示す．

変換手順

手順 1 形態素解析を行い，文節区切りにする．図 3 に例を示す．

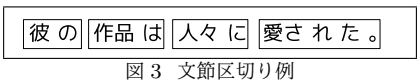

図 3 文節区切り例

手順 2 以下の条件を満たす文節を S とする．

- 表 2 の条件を満たす文節 (主語である文節)
- 主語の前に位置し，助詞 “で”，“に”，“ほど”を含まない文節

手順 3 品詞が動詞である語を含んでいる文節を V とする．

手順 4 S と V の条件に当てはまらない文節を O とする．例を図 4 に示す．

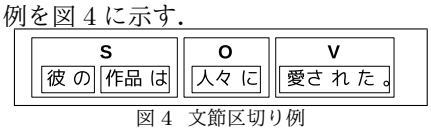

図 4 文節区切り例

手順 5 以上で定めた S，V，O を “SVO”の順に出力する．

手順 6 出力された文中に “。”が現れた場合，“、”に変換する．

手順 7 文の末尾に “。”がない場合，文の末尾に “。”を付与する．例を図 5 に示す．

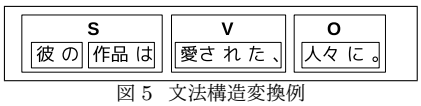

図 5 文法構造変換例

表 9 に，文法構造変換を行った文の例を示す．

表 9 文法構造変換文例

変換前	彼 の 作品 は 人々 に 愛さ れ た 。
文法構造変換	彼 の 作品 は 愛さ れ た 、 人々 に 。
対訳英語文	His works are loved by everyone .
変換前	船 は 沖合い へ 乗り出し た 。
文法構造変換	船 は 乗り出し た 、 沖合い へ 。
対訳英語文	The ship stood off to sea .
変換前	壁 に 絵 が 掛かっ て いる 。
文法構造変換	絵 が 掛かっ て いる 、 壁 に 。
対訳英語文	The picture is hanging on the wall .

5 実験環境

5.1 実験データ

実験データには，辞書の例文より抽出した単文コーパス 181,988 文 [9] から，学習データとして 100,000 文，テストデータとして 10,000 文を用いる．統計翻訳の前処理として，日本語文に対して，MeCab[10] を用いて分かち書きを行う．また，英語文に対して，Moses[11] 付属の tokenizer[11] を用いて，分かち書きを行う．表 10 に日英対訳文の例を示す．

表 10 単文コーパス例文

日本語文	昼食 を たっぷり 取っ た 。
英語文	I had a big lunch .
日本語文	私 は 猫 を 1 匹 飼っ て いる 。
英語文	I have a cat .

また，4 章に示す文法構造変換では，CaboCha[12] を用いて文節区切りにする．

5.2 実験の種類

本研究では，Moses を用いた句に基づく統計翻訳をベースラインとする．本研究では，4 つの実験を行う．表 11 に，実験の種類を示す．

表 11 実験の種類

1) ベースライン	Moses を用いた翻訳
2) 主語補完	主語補完のみを行った翻訳
3) 文法構造変換	文法構造変換のみを行った翻訳
4) 提案手法 (主語補完 + 文法構造変換)	主語補完と文法構造変換を行った翻訳

5.3 デコーダ

デコーダには，Moses を用いる．また本研究では，パラメータチューニング [13] は行わない．

5.4 評価方法

5.4.1 自動評価

本研究では，出力文の自動評価に，BLEU[14]，METEOR[15]，NIST[16] を使用する．

5.4.2 人手評価

本研究では，人手評価としてベースラインと提案手法の対比較評価を行う．対比較評価は，提案手法の出力文 10,000 文よりランダムに 100 文抽出し，ベースラインの出力文と比較し，評価する．以下に，評価基準を示す．

a) 提案手法○ 提案手法の翻訳品質がベースラインの翻訳品質より優れている場合

b) ベースライン○ ベースラインの翻訳品質が提案手法の翻訳品質より優れている場合

c) 差なし 提案手法の翻訳品質とベースラインの翻訳品質に差がない場合

d) 同一出力 提案手法の翻訳結果とベースラインの翻訳結果が同じ場合

6 実験結果

6.1 主語補完の文数

テストデータ 10,000 文と，学習データ 100,000 文に対して主語補完を行った文数をそれぞれ表 12 に示す．

表 12 主語補完文数

	主語補完文数
テストデータ	1,760/10,000
学習データ	10,844/100,000

表 12 より，テストデータの約 17%，学習データの約 11% に対して主語補完を行ったことが確認できる．

6.2 自動評価結果

ベースライン，主語補完，文法構造変換，提案手法の自動評価の結果を表 13 に示す．

表 13 自動評価結果

	BLEU	METEOR	NIST
1) ベースライン	0.1163	0.4098	4.091
2) 主語補完	0.1171	0.4095	4.138
3) 文法構造変換	0.1235	0.4281	4.428
4) 提案手法 (主語補完+文法構造変換)	0.1264	0.4296	4.491

表 13 より，提案手法の有効性が確認できる．

6.3 人手評価結果

ベースラインと提案手法の対比較評価結果を表 14 に示す．

表 14 対比較評価結果

ベースライン○	提案手法○	差なし	同一出力
7	14	72	7

表 14 の人手評価結果より，提案手法の有効性が確認できる．

対比較評価において，提案手法○の例を表 15 に，ベースライン○の例を表 16 に示す．

表 15 提案手法○の例

ベースライン入力文 1	人間 に は 5つ の 感覚 が ある 。
提案手法入力文 1	人間 に は ある 、 5つ の 感覚 が 。
正解文 1	Human beings have five senses .
ベースライン出力文 1	Man five sense .
提案手法出力文 1	Man has five sense .
ベースライン入力文 2	無事 成田空港 に 着い た 。
提案手法入力文 2	私 は 着い た 、 無事 成田空港 に 。
正解文 2	I arrived at Narita Airport safely .
ベースライン出力文 2	We 成田空港 .
提案手法出力文 2	I arrived safely at Narita Airport .

表 16 ベースライン○の例

ベースライン入力文 3	漢字 は もう 3 0 0 字 習い まし た 。
提案手法入力文 3	漢字 は 習い まし た 、 もう 3 0 0 字 。
正解文 3	I’ve already learned 300 Chinese characters .
ベースライン出力文 3	I have learned kanji three hundred characters .
提案手法出力文 3	I take three hundred characters kanji .
ベースライン入力文 4	軍隊 は 3 0 キロ 前進 し た 。
提案手法入力文 4	軍隊 は 前進 し た 、 3 0 キロ 。
正解文 4	The troops advanced 30 kilometers .
ベースライン出力文 4	The army marched 30 kilometers .
提案手法出力文 4	The troops advanced to the .

6.4 実験結果のまとめ

自動評価において，提案手法は，ベースラインと比較し BLEU 値が 1.01% 向上した．また人手評価において，提案手法は，ベースラインと比較し翻訳精度が優れていることが確認できた．以上の結果より，提案手法の有効性が確認できた．

7 考察

7.1 提案手法とベースラインの比較

主語補完が有効である例を表 17 に示す．また，表 17 での，ベースラインの入力文と出力文におけるフレーズ対応 (以後，日英フレーズ対応とする)，提案手法の日英フレーズ対応を表 18 に示す．

表 17 主語補完が有効である例

ベースライン入力文	無事 成田空港 に 着い た 。
提案手法入力文	私 は 着い た 、 無事 成田空港 に 。
正解文	I arrived at Narita Airport safely .
ベースライン出力文	We 成田空港 .
提案手法出力文	I arrived safely at Narita Airport .

表 18 表 17 における日英フレーズ対応

ベースライン	提案手法
無事 ||| We	私 は 着い た 、 ||| I arrived
成田空港 ||| 成田空港	無事 ||| safely
に 着い た 。 ||| .	成田空港 に 。 ||| at Narita Airport .

表 18 では，“私は”を “着いた”の前に補完することで，日本語フレーズ “私 は 着い た 、”と，英語フレーズ “I arrived”が対応している．このように提案手法は，主語補完と文法構造変換を併用することで，主語と動詞が出力され，翻訳精度が向上したと考えられる．

7.2 主語補完の効果

表 14 に示した人手評価 100 文において，主語補完を行った文は 16 文であった．そこで，16 文の人手評価の内訳を調査した．表 19 に示す．

表 19 主語補完における評価内訳

提案手法○	ベースライン○	差なし
2	1	13

表 19 より，主語補完で翻訳精度が向上した例はわずかである．これは，主語が省略されている場合でも，出力文に主語が出力される場合が多いためである．

7.3 パラメータチューニングを行った実験

6 章において，提案手法が有効であることが確認できた．しかし，6 章の実験では，パラメータチューニングを行っていない．そこで，本節では，パラメータチューニングを用いた翻訳実験を行う．本研究では，development データとして単文 1,000 文を用いる．

7.3.1 自動評価結果

ベースライン，主語補完，文法構造変換，提案手法の自動評価の結果を表 20 に示す．

表 20 自動評価結果

	BLEU	METEOR	NIST
1) ベースライン	0.1534	0.4764	5.107
2) 主語補完	0.1550	0.4651	5.056
3) 文法構造変換	0.1502	0.4746	5.047
4) 提案手法 (主語補完 + 文法構造変換)	0.1520	0.4727	4.991

表 20 では，提案手法は，ベースラインと比較し，BLEU 値が 0.14% 低下している．

7.3.2 人手評価結果

ベースラインと提案手法の人手評価結果を表 21 に示す． 表 21 の人手評価結果より，ベースラインと提案手法の翻訳品質にほとんど差がないことが確認できる．

表 21 人手評価結果

ベースライン○	提案手法○	差なし	同一出力
8	7	77	8

7.3.3 パラメータチューニングを行った実験のまとめ

パラメータチューニングを行った実験において，提案手法はベースラインと比較し，翻訳精度にほとんど差がないことが確認できた．この結果は，6.4 節に示す結果と異なる．

7.3.4 パラメータチューニングを行った実験への考察

出力文の例を表 22 に示す．また，表 22 における日英フレーズ対応を表 23 に示す．

主語補完について

表 23 において，ベースラインでは，“を 見 た”と “I saw the”が対応している．このように，主語省略文において，日本語の “動詞”と英語の “主語と動詞”が対応している．そして，出力文の主語を出力している．つまり，主語補完の効果が翻訳精度に影響しない．このような例が多い．

表 22 翻訳結果例

ベースライン入力文	昨夜 テレビ で 古い 映画 を 見 た 。
提案手法入力文	私 は 見 た 、 昨夜 テレビ で 古い 映画 を 。
正解文	I saw an old movie on television last night .
ベースライン出力文	I saw the old movie on television last night .
提案手法出力文	I saw a on television last night old movie .

表 23 表 22 における日英フレーズ対応

ベースライン	提案手法
昨夜 ||| last night テレビ で ||| television 古い 映画 ||| old movieon を 見 た ||| I saw the 。 ||| .	私 は ||| I 見 た ||| saw 、 ||| a 昨夜 ||| last night テレビ で ||| on television 古い ||| old 映画 を 。 ||| movie .

文法構造変換について

表 23 において，提案手法では，日英フレーズ対応は正しく行われている．しかし表 22 では，提案手法は翻訳精度が低い．この原因は，入力文の目的語における単語列の語順であると考える．

日本語において，目的語における単語数は多い．しかし本研究では，目的語における単語列に対して，文法構造変換を行っていない．よって，翻訳精度が向上しなかったと考える．今後，目的語の単語列を英語の語順に近づけることで，翻訳精度が向上すると考えている．

8 おわりに

本研究では，日英統計翻訳において，主語補完と文法構造変換を併用し，翻訳精度の向上を目指した．まず，主語が省略されている日本語文に対し，主語補完を行った．次に，単純なルールに基づき，日本語文を SVO の形式に変換した．最後に，日英統計翻訳を行い，翻訳精度を調査した．その結果，翻訳精度の向上が確認できた．

しかし，パラメータチューニングを行った場合，翻訳精度は，ベースラインとほとんど変わらなかった．今後，目的語の単語列を英語の語順に近づけることで，翻訳精度の向上を目指す．

参考文献

[1] 中岩浩巳, 池原悟,“日英翻訳システムにおける用言意味属性を用いたゼロ代名詞照応解析”, 情報処理学会論文誌 34(8),pp1705-1715,1993.
[2] 中岩浩巳, 池原悟,“語用論的・意味論的制約を用いた日本語ゼロ代名詞の文内照応解析”, 自然言語処理,Vol.3, No.4,pp50-65,1996.
[3] 中岩浩巳,“日英対訳コーパスを用いたゼロ代名詞とその指示対象の自動認定” 自然言語処理研究会 (NL-123),pp33-40,1998.
[4] 中岩浩巳 “日英機械翻訳におけるゼロ代名詞照応解析に関する研究”,2002.
[5] 古市将仁, 村上仁一, 徳久雅人, 村田真樹,“日英統計翻訳における主語補完の効果”, 言語処理学会第 17 回年次大会,pp163-166,2011.
[6] Jason Katz-Brown,Michael Collins,“Syntactic Reordering in Preprocessing for Japanese → English Translation : MIT System Description for NTCIR-7 Patent Translation Task”, Proceedings of NTCIR-7 Workshop Meeting, pp.409-414, 2008.
[7] 岡崎良太, 村上仁一, 徳久雅人, 池原悟,“日本語文法構造の変換による日英統計翻訳”, 言語処理学会 2008 年度次大会,pp240-243,2009.
[8] Katsuhito Sudoh,Kevin Duh,Hajime Tsukada,Masaaki Nagata,Xianchao Wu,Takuya Matsuzaki,jun'ichi Tsujii,“NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT”, Proceedings of NTCIR-9 Workshop Meeting,2011.
[9] 村上仁一,“日英対訳データベースの作成のための 1 考察”, 言語処理学会第 17 回年次大会,pp979-982,2011.
[10] MeCab http://mecab.sourceforge.net/
[11] Moses,moses.2007-05-29.tgz http://www.statmt.org/moses/
[12] CaboCha : Yet Another Japanese Dependency Structure Analizer http://chasen.org/˜taku/software/cabocha
[13] Franz Josef Och,“Minimum error rate training for statistical machine translation”,Proceedings of the ACL, 2003.
[14] BLEU, NIST Open MT Scoring http://www.itl.nist.gov/iad/894.01/tests/mt/2008/ scoring.html
[15] METEOR, The METEOR Automatic Machine Translation Evaluation System http://www-2.cs.cmu.edu/˜alavie/METEOR/
[16] NIST,Automatic Evaluation of Machine Translation Quality Ising N-gram Co-Occurence Statistics http://www.itl.nist.gov/iad/mig/tests/mt/2008/ scoring.html