日本語文法構造の変換による日英統計翻訳

岡崎弘樹　村上仁一　徳久雅人　池原悟
鳥取大学 工学部 知能情報工学科
{s052018,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

現在，機械翻訳において，統計翻訳の研究が注目されている．しかし，日英統計翻訳は，異なる文法構造間の翻訳であるため，翻訳精度を向上することが困難である．その問題を解決するために，単語を並び替えてから，統計翻訳を行なう研究が行われている [1]．この研究は特許文を対象としている．しかし，特許文は文法構造が複雑であるため，単語の並び替えの効果を判断することが難しい．

そこで，本研究では，単純な文法構造である単文もしくは重文複文を用いて，単語の並び替えを行なった後，統計翻訳を行なう．そして，翻訳精度を調査する．

2　日英統計翻訳システム

2.1　概要

日英統計翻訳は，日本語文 j が与えられた場合，翻訳モデルと言語モデルの組合せから確率値が最大となる英語文ê を探索することにより翻訳を行なう．

$$\begin{aligned}\hat{e} &= argmax_e P(e|j) \\ &\simeq argmax_e P(j|e)P(e)\end{aligned}$$

$P(j|e)$ は翻訳モデル，$P(e)$ は言語モデルである．

2.2　翻訳モデル

翻訳モデルは，日本語の句から英語の句へと確率的に翻訳を行なうためのモデルである．句に基づく翻訳モデルは，表 1 に示すフレーズテーブルという表により管理される．

表 1　フレーズテーブルの例

5 5 を |||55 is|||(0)(0)(1)|||(0,1)(2)||| 0.0149254 0 1 0
” L ”|||” l ”|||(0)(1)(2)|||(0)(1)(2)||| 0.170732 0 1 0

2.3　言語モデル

日英統計翻訳において，言語モデルは，翻訳文候補から英語として自然な文を選出する．言語モデルは，N-gram モデルが代表的である．

3　日本語文法構造の変換による翻訳

3.1　従来の翻訳手法

従来の日英統計翻訳は，まず日本語文のフレーズを翻訳する．そして，翻訳されたフレーズを並び替えて，翻訳候補を生成する．本研究では，従来の翻訳手法をベースラインと呼ぶ．例を図 1 に示す．

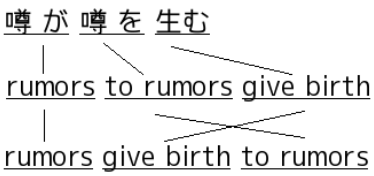

図 1　ベースラインの翻訳

言語モデルである N-gram モデルは局所的な情報である．そのため，N-gram モデルを用いて長い文章を翻訳した場合，翻訳精度が低下する．

3.2　提案手法

提案手法は，入力文をできるだけ翻訳文にあわせるために，日本語文の文法構造を変換する．具体的には日本語文の動詞を主語の後に並び替える．その後，従来の統計翻訳を行なう．例を図 2 に示す．

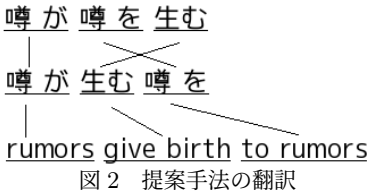

図 2　提案手法の翻訳

4　提案手法の手順

日本語文の文法構造の変更手順を以下に示す．

1. 日本語文に対し，Mecab[2] を用いて形態素解析を行う．

例 1

形態素解析前:
これは書くためのものである
形態素解析後:
これ は 書く ため の もの で ある

2. 最後に現れる「が」または「は」の位置を記憶する．例 2 では，「これ は」の「は」の位置を記憶する．

例 2

これ は 書く ため の もの で ある

3. 「が」または「は」がない場合，仮の主語である「< 任意主語 > は」を文頭につける．

例 3

仮の主語付与前:
一時 的 に 実験 を 中止 し た
仮の主語付与後:
< 任意主語 > は 一時 的 に 実験 を 中止 し た

4. 「が」または「は」に付属するフレーズを主語とする．例 4 では，「これ は」が主語となる．

例 4

これ は 書く ため の もの で ある

5. 文末にあるフレーズを動詞とする．例 5 では「もので ある」が動詞となる．

例 5

これ は 書く ため の もの で ある

6. 動詞のフレーズ位置を主語のフレーズの後ろに並び替える．以下に例 6 を示す．

例 6

並び替え前:
これ は 書く ため の もの で ある
並び替え後:
これ は もの で ある 書く ため の

5 実験環境

5.1 実験データ

5.1.1 ベースラインに用いるデータ

辞書の例文から抽出した対訳データを，単文と重文複文に分類したコーパスを用いる．この分類は，日本語文のみ着目して行なう．単文コーパスは 182,899 文 [3]，重文複文コーパスは，122,719 文である [4]．

翻訳の前処理として，各コーパスの日本語文に対しては，chasen[5] を用いて形態素解析を行なった後，日本語文の句読点を削除する．また，英語文に対しては，tokenizer.perl を用いて形態素解析を行なう．また，英語の大文字の小文字化を行なう．単文コーパスと重文複文コーパスの対訳文の例を表 2 に示す．

5.1.2 提案手法に用いるデータ

5.1.1 節の対訳データに対し，提案手法により，日本語文の動詞を主語の後に変更する．その後，日本語文に対し chasen で形態素解析を行なう．単文コーパスと重文コーパスの対訳文の例を表 3 に示す．

5.2 翻訳モデルの学習

翻訳モデルの学習は，“train-factored-phrase-model.perl” を用いて行なう．このプログラムは，IBM model1 ～5 に基づく GIZA++[6] を用いる．

表 2 ベースラインにおける対訳データ

単文コーパス	
日本語文 1	彼 の 商売 は なかなか 繁盛 し た
英語原文 1	his business has prospered .
日本語文 2	あひる は ガーガー 鳴く
英語原文 2	a duck quacks .
重文複文コーパス	
日本語文 1	彼 は 大 惨事 が 起こる と 予言 し た
英語原文 1	he prophesied of disasters to come .
日本語文 2	これら の プログラム プロダクト は この システム を 最大限 に 利用 できる よう に する ため の ライセンス プログラム で ある
英語原文 2	these program products are licensed programs that help you make maximum use of the system .

表 3 提案手法における対訳データ

単文コーパス	
日本語文 1	1 彼 の 商売 は 繁盛 し た なかなか
英語原文 1	his business has prospered .
日本語文 2	あひる は 鳴く ガーガー
英語原文 2	a duck quacks .
重文複文コーパス	
日本語文 1	彼 は 予言 し た 大 惨事 が 起こる と
英語原文 1	he prophesied of disasters to come .
日本語文 2	これら の プログラム プロダクト は ライセンス プログラム で ある する ため の この システム を 最大限 に 利用 できる よう に
英語原文 2	these program products are licensed programs that help you make maximum use of the system .

5.3 言語モデルの学習

本研究において，言語モデルは N-gram モデルを用いる．N-gram モデルの学習には，SRILM[7] の ngram-count を用いる．なお，本研究では，N-gram モデルは，過去の研究より，5-gram とする．

5.4 デコーダーに関するパラメータ

デコーダーは，moses[8] を用いる．本研究では，パラメータチューニングを行わない．ただし，“weight-t” の値は “0.5 0.0 0.5 0.0 0.0” とする．また，日本語から英語への翻訳は，動詞の位置が大きく変化する．そこで，“distortion-weight” は “0.2”，“distortion-limit” は “-1” とする．

6 提案手法の効果

6.1 実験データ

単文の実験データとして，単文コーパス 182,899 文からトレーニングデータ 100,000 文，テストデータ 1,000 文を抽出して用いる．また，重文複文コーパス 122,719 文からトレーニングデータ 100,000 文，テストデータ 1,000 文を抽出して用いる．単文の翻訳実験には単文の

トレーニングデータとテストデータを，重文複文の翻訳実験には重文複文のトレーニングデータとテストデータを用いる．

6.2 翻訳精度

ベースラインと提案手法の翻訳精度を表 4 に示す．表 4 から，提案手法の翻訳精度はベースラインと比較して，向上しないことがわかる．

表 4 ベースラインと提案手法の翻訳精度の比較

	ベースライン	提案手法
単文	0 .1177	0.1170
重文複文	0.0965	0.0939

6.3 提案手法の翻訳例

提案手法により，翻訳精度が向上した例として，単文の例を表 5 に，重文複文の例を表 6 に示す．また翻訳精度が低下した例として，単文の例を表 7 に，重文複文の例を表 8 に示す．

表 5 翻訳精度が向上した単文の例

単文例 1	
日本語文	彼 は 自分 の 欠点 が わから ない
提案手法	彼 は わから ない 自分 の 欠点 が
正解文	he is blind to his own defects .
ベースライン	i don't understand his own faults .
提案手法	he does not know his own defects .
単文例 2	
日本語文	彼 は 手 を 拍 って 笑っ た
提案手法	彼 は 笑っ た 手 を 拍 って
正解文	he clapped his hands for joy .
ベースライン	he is he chapped his hands for joy .
提案手法	he chapped his hands for joy .

表 6 翻訳精度が向上した重文複文の例

重文複文例 1	
日本語文	彼 は むすこ に あと を つい て 来る よう に いっ た
提案手法	彼 は いっ た あと を つい て 来る よう に むすこ に
正解文	he told his son to follow him .
ベースライン	he went to his son . , as follows .
提案手法	he told me to come up with after the son .
重文複文例 2	
日本語文	親 の 中 に は 教育 について 独特 の 観念 を もっ て いる 人 が 多い
提案手法	親 の 中 に は もっ て いる 人 が 多い 独特 の 観念 を 教育 について
正解文	many parents have their own peculiar ideas about education .
ベースライン	his parents in his ideas about 独特 with many people of education .
提案手法	there are many people who have a ideas about the education in his parents .

表 7 翻訳精度が低下した単文の例

単文例 1	
日本語文	夜 を 勉強 に 充て て いる
提案手法	＜ 任意主語 ＞ は 充て て いる 夜 を 勉強 に
正解文	i allot evening hours to study .
ベースライン	he is studying for the night .
提案手法	the night is given over to study .
単文例 2	
日本語文	注文 書 に 記入 し た 上 で 返送 し て ください
提案手法	＜ 任意主語 ＞ は 返送 し て ください 注 文書 に 記入 し た 上 で
正解文	please complete the order form and mail to us .
ベースライン	please return and was on the order .
提案手法	i was in order to please return on the matter .

表 8 翻訳精度が低下した重文複文の例

重文複文例 1	
日本語文	彼 が 苦情 を 申し入れ た か どう か の 問題 で ある
提案手法	彼 が 問題 で ある 苦情 を 申し入れ た か どう か の
正解文	it is a question of whether he lodged a complaint .
ベースライン	it is a question whether he to lodge complaints .
提案手法	the question is whether he is a 申し入れ complaints .
重文複文例 2	
日本語文	その 酔っ払い は 身動き も でき ず 横 に なっ て い た
提案手法	その 酔っ払い は 身動き も なっ て い た でき ず 横 に
正解文	i can not lay down the move was in some drunk .
ベースライン	it is a question whether he to lodge complaints .
提案手法	the some drunk , i can not move . had been side .

7 考察

本研究では，提案手法において，翻訳精度が低下した原因として，文法構造の変更プログラムに問題があると考える．以下にプログラムの問題点を記載する．

7.1 複合動詞の問題点

提案手法では，品詞を特定するために mecab を用いた．しかし，mecab は「気になる」といった複合動詞に対応しておらず，「気」＋「に」＋「なる」の 3 つの形態素に分解される．このため，提案手法において，複合動詞の一部のみが並び替えされる．例を表 9 に示す．

表 9 複合動詞の一部のみが並び替えされる例

日本語文	彼 は 世間 の 毀誉 褒貶 は 気 に し ない
提案手法	彼 は し ない 世間 の 毀誉 褒貶 は 気 に

このため，mecab で形態素解析を行った後，複合動詞

を再構成する必要がある．

7.2 日本語文の主語と動詞の対応における問題点

提案手法では，文末の動詞を主語の後に並び替える．しかし，「目が肥えているからあの奥さんは着物の注文がなかなか満足しない」といった，主語と動詞が二つある文に対応していない．このため，提案手法において，「目が」の後ろに「満足しない」の動詞が並び替えされる．このため，主語と動詞が二つある場合，主語と動詞の対応関係を考える必要がある．

8 おわりに

本研究では，単文及び重文複文の日本語文法構造の変換を行い，翻訳精度の変化を調べた．実験の結果，提案手法では，翻訳精度を向上することできなかった．その理由として，文法構造のプログラムに問題があったことがあげられる．今後，提案手法の翻訳精度を向上するために，文法構造を変更するときに複合動詞に対応する必要がある．また，主語と述語の対応関係を考える必要がある．

参考文献

[1] 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

[2] Mecab,http://mecab.sourceforge.net/

[3] 西山七絵, 村上仁一, 徳久雅人, 池原悟,“単文文型パターン辞書の構築”, 言語処理学会第 11 回年次大会,pp.372-375,2005

[4] 村上仁一, 池原悟, 徳久雅人,“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

[5] ChaSen,http://chasen-legacy.sourceforge.jp/

[6] SRILM,The SRI Language Modeling Toolkit /http://www.speech.sri.com/projects/srilm

[7] GIZA++,http://www.fjoch.com/GIZA++

[8] Moses,moses.2007-05-29.tgz http://www.statmt.org/moses/