

概要

近年では、自動的に翻訳システムを構築できる統計翻訳が注目されている。統計翻訳では、文全体の構造は解析しないため、言語間の文法構造が相違しているときには翻訳精度が低い傾向があり、言語間の文法構造が類似しているときには翻訳精度が高い傾向がある。

そこで本研究では、言語間の文法構造が類似しているときには翻訳精度が高くなる統計翻訳の長所を効果的に用いる。そのために、まず前処理として、文法構造を大きく変換する必要がある文に対しても、英語の文法構造に類似した文を出力することができるルールベース翻訳で日英翻訳を行なう。次に、英英統計翻訳を行なうことで翻訳精度の向上を試みた。結果として、BLEU スコアと METEOR スコアが向上し、本研究の有効性が確認できた。しかしながら、重文複文において、ルールベース翻訳のみの出力文と提案手法の出力文に対して、対比較評価したときに、提案手法の方が悪い結果となった。これは前処理において、英語の文法構造に類似させることができなかったことが原因であると考えている。

今後は、重文複文に対しても、英語の文法構造に類似させることができる前処理の方法を調査することを考えている。

目次

1	はじめに	1
2	日英統計翻訳システム	2
2.1	概要	2
2.2	翻訳モデル	3
2.3	IBM 翻訳モデル	3
2.3.1	Model1	4
2.3.2	Model2	5
2.3.3	Model3	6
2.3.4	Model4	7
2.3.5	Model5	7
2.4	GIZA++	8
2.5	言語モデル	8
2.5.1	概要	8
2.5.2	N -gram モデル	9
2.6	デコーダー	9
2.7	パラメータチューニング	10
3	ルールベース翻訳と統計翻訳の組合せ手法	11
3.1	過去の研究	11
3.2	提案手法	11
3.2.1	概要	11
3.2.2	提案手法の手順	12
4	実験環境	14
4.1	言語モデルの学習	14
4.2	翻訳モデルの学習	14
4.3	デコーダのパラメータ	14
4.4	実験データ	14
4.4.1	単文コーパス	14
4.4.2	重文複文コーパス	15

4.5	ルールベース翻訳	15
4.6	評価方法	15
5	翻訳実験	16
5.1	自動評価	16
5.2	対比較評価	17
6	考察	22
6.1	自動評価と対比較評価の不整合	22
6.2	未知語の影響	25
6.3	distortion-limit の影響	27
7	おわりに	30

目 次

1	日英統計翻訳の枠組	2
2	デコーダーの動作例	9
3	提案手法の枠組	12

表目次

1	単文における評価	16
2	重文複文における評価	16
3	提案手法 の例 (単文)	18
4	提案手法 の例 (重文複文)	18
5	提案手法 × の例 (単文)	19
6	提案手法 × の例 (重文複文)	19
7	提案手法 の例 (単文)	20
8	提案手法 の例 (重文複文)	20
9	提案手法=の例 (単文)	21
10	提案手法=の例 (重文複文)	21
11	対比較評価	21
12	提案手法 の例 (単文)	22
13	提案手法 の例 (重文複文)	22
14	提案手法 × の例 (単文)	23
15	提案手法 × の例 (重文複文)	23
16	対比較評価	24
17	未知語の減少例	25
18	未知語数の比較	26
19	distortion-limit の違いによる出力例 (単文)	27
20	distortion-limit の違いによる出力例 (重文複文)	28
21	ベースラインと提案手法の評価の比較	28

1 はじめに

従来の機械翻訳は、人手によって文法のルールを厳密に定義するルールベース翻訳であった。しかしながら、ルールベース翻訳システムを構築するためには、言語の専門家の知識を必要とし、莫大な時間や労力を必要とする短所がある。そこで近年では、自動的に翻訳システムを構築できる統計翻訳が注目されている。この翻訳手法が注目されたことに対して、近年の計算機性能の大幅な向上によって、処理時間が大幅に短縮されたことが影響している。また、統計翻訳の長所として、翻訳言語対から自動的にモデルを作成するため、様々な言語への拡張が容易に行なえることが挙げられる。

統計翻訳において、言語間の文法構造が相違している日本語-英語間の翻訳と比較して、言語間の文法構造が類似しているイタリア語-英語間の翻訳では翻訳精度が高くなる傾向がある [1]。

そこで本研究では、言語間の文法構造が類似しているときには翻訳精度が高くなる統計翻訳の長所を効果的に用いる。そのために、まず前処理として、文法構造を大きく変換する必要がある文に対しても、英語の文法構造に類似した文を出力することができるルールベース翻訳で日英翻訳を行なう。次に、英英統計翻訳を行なうことで翻訳精度の向上を試みる。

2章において、日英統計翻訳の概要についての説明を行なう。3章において、提案手法の手順や類似研究についての説明を行なう。4章において、実験環境についての説明を行なう。5章において、提案手法の効果を示す。6章において、考察を行なう。7章において、結論を述べる。

2 日英統計翻訳システム

2.1 概要

日英統計翻訳システムの枠組を図1に示す。

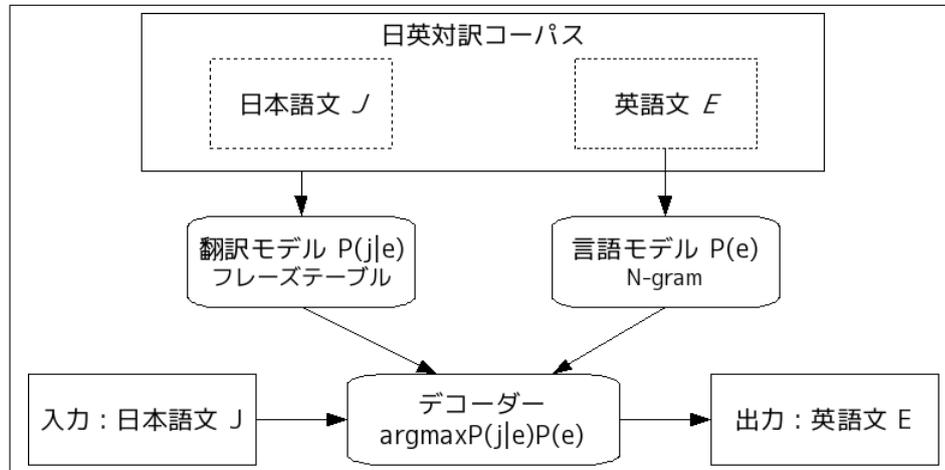


図 1: 日英統計翻訳の枠組

日英統計翻訳は、日本語文 J が与えられた場合に、翻訳モデルと言語モデルの組み合わせの中から確率が最大となる英語文 E を探索することで翻訳を行なう。以下にその基本モデルを示す。

$$E = \operatorname{argmax}_e P(e|j) \tag{1}$$

$$\simeq \operatorname{argmax}_e P(j|e)P(e) \tag{2}$$

ここで $P(j|e)$ は翻訳モデル、 $P(e)$ は言語モデルを示す。

2.2 翻訳モデル

翻訳モデルは，日本語から英語へ確率的に翻訳を行なうためのモデルである．翻訳モデルでは，フレーズテーブルと呼ばれる表によって管理される．以下にその例を示す．

フレーズテーブルの例

```
あの 工場 ||| That plant ||| 1 0.103812 0.333333 0.0502323 2.718
あの 工場 ||| that factory ||| 1 0.0763503 0.333333 0.0986164 2.718
あの 工場 ||| that plant ||| 1 0.0237359 0.333333 0.0297904 2.718
あの 工場 で ||| in that plant ||| 0.5 0.0052738 1 0.00561083 2.718
あの 工場 では ||| in that plant ||| 0.5 0.000598488 1 0.00561083 2.718
あの 工場 は ||| The factory smokes ||| 1 0.00716914 0.5 4.34963e-05 2.718
あの 工場 は ||| in that factory ||| 1 0.00199214 0.5 0.000719742 2.718
あの 工場 は 煙突 から ||| The factory smokes heavily from its chimney ||| 0.333333
0.000915843 1 5.45195e-09 2.718
```

左から日本語フレーズ，英語フレーズ，フレーズの英日方向の翻訳確率 $P(j|e)$ ，英日方向の単語の翻訳確率の積，フレーズの日英方向の翻訳確率 $P(e|j)$ ，日英方向の単語の翻訳確率の積，フレーズペナルティ(値は常に自然対数の底 $e=2.718$) である．

2.3 IBM 翻訳モデル

統計翻訳の代表的なモデルとして，IBM の Brown らによる仏英翻訳モデル [2] がある．この翻訳モデルは，model1 から model5 までの 5 つのモデルから構成され，順に複雑になっていく．IBM 翻訳モデルでは，フランス語から英語への翻訳を想定しているため，以下の説明では仏英翻訳を前提とする．本章では，原言語であるフランス語文を f ，目的言語である英語文を e として定義する．

IBM モデルでは，フランス語文 e ，英語文 f の翻訳モデル $P(F|E)$ を計算するために，アライメント a を用いる．以下に IBM モデルの基本式を示す．

$$P(f|e) = \sum_a P(f, a|e) \quad (3)$$

アライメント a は，フランス単語 F と英単語 E の対応関係を示す．IBM モデルにおいて，英単語は 1:n の対応をもち，フランス単語は 1:1 の対応をもつと仮定する．またフランス単語に適切な対応関係をもつ英単語が存在しなかった場合には，英語文の先頭に e_0 という特殊文字があると仮定し，そのフランス単語と特殊文字を対応させる．

2.3.1 Model1

式 (3) は以下の式に置き換えることができる． m はフランス語文における単語数， a_1^{j-1} はフランス語文において，1 単語目から $j-1$ 単語目までのアライメント， f_1^{j-1} はフランス語文において，1 単語目から $j-1$ 単語目までを示す．

$$P(f, a|e) = P(m|e) \prod_{j=1}^m P(a_j|a_1^{j-1}, f_1^{j-1}, m, e) P(f_j|a_1^j, f_1^{j-1}, m, e) \quad (4)$$

式 (4) では，式が複雑であるため，計算が困難である．よって，Model1 では以下の仮定によって，式を簡略化する．

- フランス語文の長さの確率は， m および e に依存しない

$$\epsilon \equiv P(m|e)$$

- アライメントの確率は英語文 e の長さ l のみに依存する

$$P(a_j|a_1^{j-1}, f_1^{j-1}, m, e) \equiv (l+1)^{-1}$$

- フランス単語は，対応する英単語のみに依存する

$$P(f_j|a_1^j, f_1^{j-1}, m, e) \equiv t(f_j|e_{a_j})$$

以上の仮定を用いることで，式 (4) は式 (5) に置き換えることができる．

$$P(f, a|e) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m t(f_j|e_{a_j}) \quad (5)$$

$$P(f|e) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \prod_{j=1}^m t(f_j|e_{a_j}) \quad (6)$$

$$= \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=0}^l t(f_j|e_i) \quad (7)$$

ラグランジェの未定係数法を用いて，制約条件 $\sum_f t(f|e)$ のもとで $P(f|e)$ の最大化を行なう問題を解くと，学習に用いたフランス語と英語の対訳文 ($f^{(s)}$ と $e^{(s)}$) において， f と e が対応付けられる回数の期待値を求めることができる．以下に期待値を求める式を示す．

$$c(f|e; f^{(s)}, e^{(s)}) = \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \sum_{j=1}^m \delta(f, f_j) \sum_{i=0}^l \delta(e, e_i) \quad (8)$$

$\delta(f, f_j)$ はフランス語文 $f^{(s)}$ に，フランス単語 f が出現する回数， $\delta(e, e_i)$ は英語文 $e^{(s)}$ に，英単語 e が出現する回数を示す．また，以下に示す EM アルゴリズムを用いて最適解を推定する．

手順1 $t(f|e)$ に適当な初期値を設定する

手順2 フランス語と英語の対訳文 ($f^{(s)}$ と $e^{(s)}$), $1 \leq s \leq S$ において, フランス単語 f と英単語 e が対応付けられる回数の期待値を式 (8) によって求める

手順3 英語文 $e^{(s)}$ に1回以上出現した英単語 e に対して

- 定数 λ_e を以下の式で計算する

$$\lambda_e = \sum_f \sum_{s=1}^S c(f|e; f^{(s)}, e^{(s)}) \quad (9)$$

- $t(f|e)$ を以下の式で再計算する

$$t(f|e) = \lambda_e^{-1} \sum_{s=1}^S c(f|e; f^{(s)}, e^{(s)}) \quad (10)$$

$$= \frac{\sum_{s=1}^S c(f|e; f^{(s)}, e^{(s)})}{\sum_f \sum_{s=1}^S c(f|e; f^{(s)}, e^{(s)})} \quad (11)$$

手順4 $t(f|e)$ が収束するまで, 手順2 と手順3 を繰り返す

2.3.2 Model2

Model1 では, アライメントの確率は英語文の長さ l のみに依存すると仮定した. しかしながら, Model2 では, 英語文の長さ l に加え, j 単語目のアライメント a_j , フランス語文の長さ m に依存すると仮定し, 以下の式で表すことができる.

$$a(a_j|j, m, l) \equiv P(a_j|a_1^{j-1}, f_1^{f-1}, m, l) \quad (12)$$

これによって, 式 (6) は以下の式に置き換えることができる.

$$P(f|e) = \epsilon \sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \prod_{j=1}^m t(f_j|e_{a_j}) a(a_j|j, m, l) \quad (13)$$

$$= \epsilon \prod_{j=1}^m \sum_{i=0}^l t(f_j|e_i) a(i|j, m, l) \quad (14)$$

ラグランジェの未定係数法を用いて, 制約条件 $\sum_f t(f|e), \sum_{i=0}^l a(i|j, m, l) = 1$ のもとで $P(f|e)$ の最大化を行なう問題を解くと, 期待値を求める式が2つ得られる. 以下に

期待値を求める2つの式を示す．

$$c(f|e; f^{(s)}, e^{(s)}) = \sum_{j=1}^m \sum_{i=0}^l \frac{t(f|e)a(i|j, m, l)\delta(f, f_j)\delta(e, e_i)}{t(f|e_0)a(0|j, m, l) + \dots + t(f|e_l)a(l|j, m, l)} \quad (15)$$

$$c(i|j, m, l; f^{(s)}, e^{(s)}) = \frac{t(f_j|e_i)a(i|j, m, l)}{t(f_j|e_0)a(0|j, m, l) + \dots + t(f_j|e_l)a(l|j, m, l)} \quad (16)$$

$c(f|e; f^{(s)}, e^{(s)})$ は，フランス語と英語の対訳文のフランス単語 f と英単語 e が対応付けられる回数の期待値を示し， $c(i|j, m, l; f^{(s)}, e^{(s)})$ は，フランス単語の位置 j と英単語の位置 i が対応付けられる回数の期待値を示す．

また，以下に示す EM アルゴリズムを用いて最適解を推定する．

手順 1 $t(f|e)$ に適当な初期値を設定する

手順 2 フランス語と英語の対訳文 ($f^{(s)}$ と $e^{(s)}$)， $1 \leq s \leq S$ において，期待値を式 (15) と式 (16) によって求める

手順 3 手順 2 からそれぞれの総和を求め，正規化を行なうことで再計算する

手順 4 $t(f|e)$ が収束するまで，手順 2 と手順 3 を繰り返す

Model2 は，上記のような EM アルゴリズムで計算した場合，複数の極大値を持つため，必ず最適解が得られる保証はない．しかしながら，Model2 の特別な場合として， $a(i|j, m, l) = (l + 1)^{-1}$ になることが考えられる．これは Model1 として考えることができる．また，Model1 は最適解が保証されているため，初期値として Model1 によって求められた値を用いる．

2.3.3 Model3

Model1，Model2 では，英単語とフランス単語の対応は 1 対 1 の場合のみと想定していた．しかしながら，実際の翻訳では，複数の単語が 1 つの単語に翻訳されることが考えられる．そこで，Model3 では，ある英単語がフランス単語に何個対応するのかを考慮し，さらに，対応する英単語とフランス単語が近い位置にあるとは限らないので，英単語とフランス単語の距離も考慮する．これらの場合を想定し，以下の 3 つのパラメータを用いる．

- 翻訳確率 $t(f|e)$ フランス単語 f が英単語 e に翻訳される確率

- 繁殖確率 $n(\phi|e)$ 英単語 e が ϕ 個のフランス語に対応する確率
- 歪み確率 $d(j|i, m, l)$ フランス語文の長さが m , 英語文の長さが l のとき , 単語位置 i の英単語が単語位置 j のフランス単語に翻訳される確率

これに加えて , 英単語に翻訳されないフランス単語の数を ϕ_0 とし , その確率 p_0 を以下の式によって求める .

$$P(\phi_0|\phi_1^l, e) = \binom{\phi_1 + \dots + \phi_l}{\phi_0} p_0^{\phi_1 + \dots + \phi_l - \phi_0} p_1^{\phi_0} \quad (17)$$

上式より , Model3 は以下の式で表すことができる .

$$\begin{aligned} P(f|e) &= \sum_{a_1=0}^l \dots \sum_{a_m=0}^l P(f, a|e) \quad (18) \\ &= \sum_{a_1=0}^l \dots \sum_{a_m=0}^l \binom{m - \phi_0}{\phi_0} p_0^{m - 2\phi_0} p_1^{\phi_0} \prod_{i=1}^l \phi_i! n(\phi_i|e_i) \times \\ &\quad \prod_{j=1}^m t(f_j|e_{a_j}) d(j|a_j, m, l) \quad (19) \end{aligned}$$

Model3 は , Model1 や Model2 と異なり , 効率的に計算できず , 全ての単語対応を考慮すると計算量が莫大であるため , 期待値は近似によって求める . また , 英単語とフランス単語の位置は絶対位置として考える .

2.3.4 Model4

Model3 では , 単語の位置を絶対位置で考えていたが , Model4 では , 英単語とフランス単語の位置を相対位置で考える . また , Model4 では各単語ごとの位置も考慮する . 例えば , 形容詞と名詞の関係のように , 英語では形容詞は名詞の前に出現し , フランス語では形容詞は名詞の後に出現する傾向がある . なお , 歪み確率 $d(j, i, m, l)$ は Model4 では2通りを考える .

2.3.5 Model5

Model4 では , 単語位置に関して直前の単語についてのみ考慮している . そのため , 複数の単語を同じ位置に配置したり , 単語が存在しない位置が生成されるという欠点がある . そこで , Model5 では , このような問題を解消するために , 単語を空白部分に配置するように制約を加える .

2.4 GIZA++

GIZA++[3] は、IBM 翻訳モデルを用いて、原言語と目的言語における単語の対応関係の確率値を計算する。

2.5 言語モデル

2.5.1 概要

言語モデルは、単語列に対して、それらの単語列が起こる確率を付与するモデルである。日英統計翻訳では、言語モデルを使用して、英語として自然な文を選出する。言語モデルの例を以下に示す。

言語モデルの例

```
-0.9121773 factory . -0.772665  
-1.571392 factory has -0.05683998  
-1.120353 factory in -0.05121826  
-1.821027 factory will -0.0660101  
-1.56243 facts do -0.2219447  
-1.232086 facts of -0.227057  
-0.1661982 faculties . -0.4716349  
-0.847985 faculty . -0.3466964  
-0.4806468 fad . -0.2219448  
-0.7325071 fade . -0.2219448
```

一番上の行に関して、左から、“factory” のあとに “.” がくる確率を常用対数で表した値 $\log_{10}(P(a | factory)) = -0.9121773$ 、2-gram で表された単語列である “factory .”、バックオフスムージングにより得られる、“factory” のあとに “.” がくる確率を常用対数で表した値 $\log_{10}(P(a | factory)) = -0.772665$ である。

また、バックオフスムージングとは、高次の N -gram が存在しない場合、低次の N -gram を用いる手法である。この低次の確率を、改良したスムージングの手法が Kneser-Ney スムージングである。言語モデルにおける N -gram 作成には、性能の観点から一般的に Kneser-Ney スムージングが用いられている。

2.5.2 N-gram モデル

N-gram モデルは，“単語列 w_1, w_2, \dots, w_n の i 番目の単語 w_i の生起確率 $P(w_i)$ は，直前の単語 ($N-1$) に依存する” という仮定に基づいているモデルである．計算式を以下に示す

$$P(w_1 w_2 \dots w_n) = \prod_{i=1}^n P(w_i | w_{i-1}) \quad (20)$$

例えば，「He is a pilot .」という文字列に対する 2-gram モデルを以下に示す．

$$P(e = \text{“He is a pilot .”}) \approx P(He) \times P(is | He) \times P(a | is) \times P(pilot | a) \times P(. | pilot)$$

また，3-gram モデルのときは $P(a | He is)$ になり，4-gram モデルのときは $P(pilot | He is a)$ になる．このように，(N-1) 単語の次にくる単語が “a” や “pilot” である確率を考える．

2.6 デコーダー

デコーダーは，翻訳モデルと言語モデルを用いて，確率が最大となる翻訳候補を探索し，出力を行なう．入力文として「彼はリンゴが好きだ .」が与えられたときの翻訳例を以下の図 2 に示す．

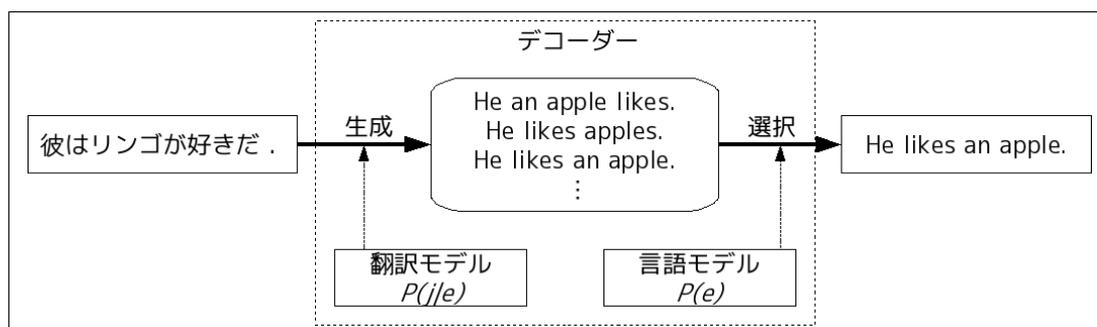


図 2: デコーダーの動作例

日英統計翻訳において， $\operatorname{argmax}_e P(j|e)P(e)$ の確率が最大となる英語文を出力するためには，適切な順序で日本語と英語の単語対応を選択する必要がある．しかしながら，適切な英語文を決定するためには，莫大な計算量が必要となり，莫大な時間が必要となる．そこで莫大な計算量を削減するために，ビームサーチ法を用いる．

2.7 パラメータチューニング

パラメータチューニングは、Minimum Error Rate Training(MERT)[4] という手法を用いて、デコーダーで用いるパラメータを最適化することである。MERT は評価関数 (一般的にはBLEU) を最大にする翻訳結果が選ばれるように、パラメータ調整を行なう。この際、ディベロップメントデータと呼ばれる、試し翻訳を行なうデータを与え、各文に対して上位 100 個程度の翻訳候補を出力し、よりよい翻訳候補が上位にくるようにパラメータを調整する。

3 ルールベース翻訳と統計翻訳の組合せ手法

3.1 過去の研究

ルールベース翻訳と統計翻訳を組み合わせる手法としては、L.Dugastら [5] や M.Simardら [6] の研究がある。彼らは、ルールベース翻訳として SYSTRAN を用いて翻訳を行ない、統計翻訳を後処理として用いた。その結果、SYSTRAN のみや統計翻訳のみのときと比較して、BLEU スコアが向上しており、彼らの手法の有効性が確認できた。

彼らの手法では、統計翻訳をルールベース翻訳の後処理として考えている。しかしながら、本研究では、ルールベース翻訳を統計翻訳の前処理として考える。

3.2 提案手法

3.2.1 概要

提案手法では、前処理として文法構造を目的言語に類似させるために、はじめに、ルールベース翻訳を使用して、日本語文を英語文に翻訳する。次に、翻訳された英語文に対して、英英統計翻訳を行なう。本研究において、この英英統計翻訳を英' 英統計翻訳と定義する。枠組を図 3 に示す。

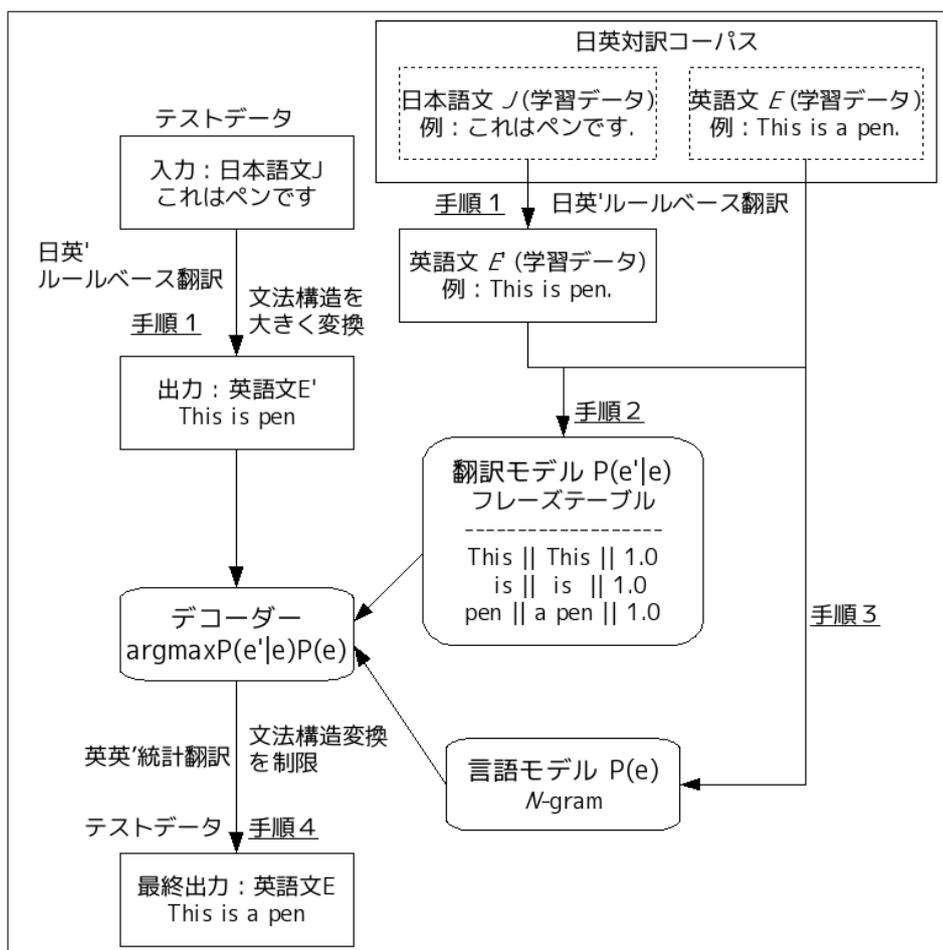


図 3: 提案手法の枠組

3.2.2 提案手法の手順

手順 1 前処理としての日英' ルールベース翻訳

前処理として、日英対訳コーパス (学習データ, テストデータ, ディベロップメントデータ) の日本語文に対して、ルールベース翻訳でそれぞれ日英' 翻訳を行なう。入力例、出力例、正解例を以下に示す。

入力例 (日本語文)

免許をとった。
 その許可が当局から下りた。
 天文学は高尚な科学である。
 彼は大声で泣いた。
 この箱はどうしても持ち上がらない。

出力例 (ルールベース翻訳)

It got the license.
The permission descended from authorities .
Astronomy is a noble science .
He cried loudly.
This box doesn't lift.

正解例 (英語文)

I obtained a license.
That permission has come from the authorities.
Astronomy is a noble science.
He cried loudly.
This box won't lift.

手順2 翻訳モデルの作成

前処理を行なった学習データの英語文と日英対訳コーパスの英語文を用いて，英' 英翻訳モデルを作成する．英' 英フレーズテーブルの例を以下に示す．

英' 英フレーズテーブルの例

It got the license ||| I obtained a license ||| (0) (1) (2) (3) ||| (0) (1) (2) (3) |||
1.000 0.000 1.000 0.000
The permission descended ||| That permission has come ||| (0) (1) (1,2,3) ||| (0)
(1,2) (2) (2) ||| 1.00 0.002 1.00 0.00
Astronomy ||| Astronomy ||| (0) ||| (0) ||| 1.00 1.00 1.00 1.00
loudly ||| loudly ||| (0) ||| (0) ||| 0.688 0.289 0.733 0.344
This box ||| This box ||| (0) (1) ||| (0) (1) ||| 0.625 0.512 0.833 0.421

手順3 言語モデルの作成

日英対訳コーパスの英語文から言語モデルを作成する．

手順4 英' 英統計翻訳

手順1で作成した言語モデルと手順3で作成した翻訳モデルを用いて，統計翻訳で英' 英翻訳を行なう．

4 実験環境

4.1 言語モデルの学習

言語モデルの学習には，“SRILM[7]”の“ngram-count”を用いる．本研究では， N -gramモデルは5-gramとする．またスムージングには，“Kneser-Ney discount”を用いる．

4.2 翻訳モデルの学習

翻訳モデルの学習には，“GIZA++[3]”を利用する“train-factored-phrase-model.perl[8]”を用いる．

4.3 デコーダのパラメータ

デコーダには，“moses[8]”を用いる．またmosesのパラメータはパラメータチューニング[4]によって最適化する．ただし，パラメータチューニングで変更されない値である，“ttable-limit”と“distortion-limit”に関しては手作業で変更する．本研究では，“ttable-limit”の値は80とする．また“distortion-limit”の値は，ベースラインでは過去の研究により-1(無限大)とし，提案手法では文法構造の変換を制限するためにデフォルト値の6とする．

4.4 実験データ

4.4.1 単文コーパス

実験には，辞書の例文より抽出した単文コーパス182,899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，ディベロップメントデータとして1,000文を用いる．

統計翻訳の前処理として，各コーパスの日本語文に対して，“chasen[10]”を用いて形態素解析を行なう．また英語文に対して，“tokenizer.perl[8]”を用いて，わかち書きを行なう．

4.4.2 重文複文コーパス

実験には，辞書の例文より抽出した重文複文コーパス 122,719 文 [11] 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，ディベロップメントデータとして 1,000 文を用いる．

統計翻訳の前処理として，各コーパスの日本語文に対して，“chasen[10]”を用いて形態素解析を行なう．また英語文に対して，“tokenizer.perl[8]”を用いて，わかち書きを行なう．

4.5 ルールベース翻訳

ルールベース翻訳とは，文構造を解析し，言語の専門家などによって厳密に定義された文法のルールに準じて行なう翻訳である．長所として，厳密に定義するので，ルールが存在する文の翻訳では精度が高いことがあげられる．短所としては，ルールが存在しない場合には翻訳精度が低いこと，人手によってルールを生成するので，コストや時間がかかることがあげられる．

また本研究では，ルールベース翻訳として，市販されている翻訳ソフトである ATLASV12(富士通)と翻訳の王様 Version5(IBM)を使用する．なお，市販翻訳では独自の形態素解析を行なうため，市販されている翻訳ソフトでは形態素解析前の日本語コーパスを使用する．

4.6 評価方法

機械翻訳における翻訳精度を自動的に評価する方法では，前もって用意した正解文と翻訳出力文を比較する方法が一般的である．本研究では，自動評価法として BLEU(Bilingual Evaluation Understudy)[12]と METEOR(Metric for Evaluation of Translation with Explicit ORdering)[13]を使用する．BLEU では 4-gram を用いて単語列を評価し，語順が正解文に近いときに高いスコアを出す．METEOR では意味などの単語属性が正しいときに高いスコアを出す．どちらのスコアも 0 から 1 の範囲で評価され，1 が最も良いスコアである．また人手評価として，対比較評価を行なう．

5 翻訳実験

本研究において，統計翻訳のみを用いて翻訳する手法をベースラインと定義する．翻訳実験では，ベースライン，ルールベース翻訳 (ATLAS) のみ，ルールベース翻訳 (翻訳の王様) のみ，提案手法の4つの手法をそれぞれ用いて，翻訳を行なう．また出力文に対して，自動評価と対比較評価を行なう．なお，翻訳対象文として，単文と重文複文を使用する．

5.1 自動評価

ベースライン，ルールベース翻訳 (ATLAS) のみ，提案手法1(ATLAS+統計翻訳)，ルールベース翻訳 (翻訳の王様) のみ，提案手法2(翻訳の王様+統計翻訳)の自動評価法による評価結果を表1と表2に示す．

表 1: 単文における評価

	BLEU	METEOR
ベースライン	0.1397	0.3844
ATLASのみ	0.1552	0.4216
提案手法1	0.1877	0.4418
翻訳の王様のみ	0.1206	0.3888
提案手法2	0.1721	0.4227

表 2: 重文複文における評価

	BLEU	METEOR
ベースライン	0.1175	0.3531
ATLASのみ	0.0942	0.3623
提案手法1	0.1364	0.3806
翻訳の王様のみ	0.0842	0.3522
提案手法2	0.1346	0.3737

結果より，ベースラインと提案手法，ルールベース翻訳 (ATLAS) のみと提案手法1，ルールベース翻訳 (翻訳の王様) のみと提案手法2を比較すると，すべて翻訳精度が向上していることが確認できる．

5.2 対比較評価

単文と重文複文において，提案手法 1 の出力文とベースラインの出力文からランダムに抽出した 100 文を用いて，対比較評価を行なった．評価基準は以下に示す．

提案手法 : 提案手法の方が良い

提案手法× : 提案手法の方が悪い

提案手法 : 提案手法と比較対象の文の単語対応や伝わる意味に差がない

提案手法= : 提案手法と比較対象の文がまったく同じ

以下の表 3 から表 10 にそれぞれの判断基準の例を示し，評価結果を表 11 に示す．

表 3: 提案手法 の例 (単文)

例 1	
日本語文	お問合せの会社は、19xx年に設立されました。
正解文	The company you inquired about was established in 19xx .
ベースライン	The company was established in the 問合せ you , 19xx .
提案手法 1	The company that inquired was established in 19xx age .
例 2	
日本語文	キャンプファイヤーの火が赤々と燃えている。
正解文	The camp fire flames are burning brightly .
ベースライン	The fire is burning brightly in the camp fire .
提案手法 1	The fire of the campfire is burning brightly .
例 3	
日本語文	旗が垂れ下がっている。
正解文	The flag is drooping .
ベースライン	are drooping flag .
提案手法 1	The flag is hanging down .

表 4: 提案手法 の例 (重文複文)

例 1	
日本語文	前田さんは受験に失敗して予備校に行っている。
正解文	Maeda flunked the entrance exams and is now attending yobiko .
ベースライン	He has gone to Mr . 前田 予備校 failed the exam .
提案手法 1	Maeda failed the exam and went to the prepschool .
例 2	
日本語文	雨がやみましたから、傘を畳みましょう。
正解文	It stopped raining . Let's close an umbrella .
ベースライン	The rain , so let's have A 畳み an umbrella .
提案手法 1	The rain has stopped , let's folding umbrella .
例 3	
日本語文	あれだけ懲らしめれば、懲りるだろう。
正解文	If you are punished so , you will probably learn your lesson .
ベースライン	He 懲りる , you will 懲らしめれ .
提案手法 1	If that is punished , it will learn .

表 5: 提案手法 × の例 (単文)

例 1	
日本語文	そんなにいらいらしないで。
正解文	Don't get so agitated .
ベースライン	That's not irritated .
提案手法 1	I don't get irritated so much .
例 2	
日本語文	そのことで君と賭けをしよう。
正解文	I'll bet you on it .
ベースライン	I'll bet with you about it .
提案手法 1	Let's go by that with you .

表 6: 提案手法 × の例 (重文複文)

例 1	
日本語文	それはとても甘い処置だ。
正解文	That is a very lenient way of dealing with the situation .
ベースライン	It was a very sweet step .
提案手法 1	It is very sweet .
例 2	
日本語文	わざわざ見に行く価値は無い。
正解文	It is not worth going out of your way to see it .
ベースライン	Do not go to see value .
提案手法 1	There is no value of him .

表 7: 提案手法 の例 (単文)

例 1	
日本語文	私にはこの文の意味が分からない。
正解文	I cannot make out the meaning of this sentence .
ベースライン	I don't understand this sentence .
提案手法 1	I don't know the meaning of this sentence .
例 2	
日本語文	肉が鍋の中でジュージューパーチパーチと音をたてた。
正解文	The meat hissed and spluttered in the skillet .
ベースライン	The meat パチパーチ and sound ジュー ジュー in the pot .
提案手法 1	Meat made ジュージューパーチパーチ and noise in the pan .

表 8: 提案手法 の例 (重文複文)

例 1	
日本語文	支払い条件について詳細に話し合いたいと思います。
正解文	We would like to discuss the terms of payment in more detail .
ベースライン	I would like to talk to details about payment terms .
提案手法 1	I would like to discuss payment terms in detail .
例 2	
日本語文	なにをさしてもそつが無い。
正解文	He acquits himself well in everything .
ベースライン	そつ even meant anything .
提案手法 1	It is not なにをさしてもそつ .

表 9: 提案手法=の例 (単文)

日本語文	彼女は我々に会いに来た。
正解文	She came to see us .
ベースライン	She came to see us .
提案手法 1	She came to see us .

表 10: 提案手法=の例 (重文複文)

日本語文	彼は歯を食いしばって痛みをこらえた。
正解文	He gritted his teeth in pain .
ベースライン	He clenched his teeth against the pain .
提案手法 1	He clenched his teeth against the pain .

表 11: 対比較評価

	提案手法	提案手法×	提案手法	提案手法=
単文	79 /100	6 /100	12 /100	3 /100
重文複文	49 /100	14 /100	33 /100	4 /100

結果より、ベースラインと提案手法を比較したときでは、単文と重文複文の両方において、“提案手法の方が良い”が選択された数が最も多く、提案手法の効果が確認できた。

6 考察

6.1 自動評価と対比較評価の不整合

単文と重文複文において，提案手法 1 の出力文とルールベース翻訳 (ATLAS) のみの出力文からランダムに抽出した 100 文を用いて，対比較評価を行なった．評価基準は 6.2 章において示したものと同様とする．判断基準の例を表 12 から 15 に示し，評価結果を表 16 に示す．

表 12: 提案手法 の例 (単文)

例 1	
日本語文	電気 コンロ の コイル が 焼き 切れた 。
正解文	The heater coil is burnt out .
ATLAS のみ	The coil of an electric stove was able to be burnt off .
提案手法 1	The coil of the electric heater has burned out .
例 2	
日本語文	酒 に おぼれ ている 。
正解文	He has an alcohol problem .
ATLAS のみ	It .. sake .. is drowned .
提案手法 1	It is drowned in drink .

表 13: 提案手法 の例 (重文複文)

例 1	
日本語文	字 が 下手 な ので 、 恥 ず かしい 。
正解文	I am embarrassed because my handwriting is so poor .
ATLAS のみ	Because the character is unskilled , it is shameful .
提案手法 1	The writing is poor , so it is embarrassing .
例 2	
日本語文	なん だ も 二 時 過 ぎ で あ っ た よ う に ぼ ん や り し た 覚 え が 有 る 。
正解文	I have a vague impression that it was past two .
ATLAS のみ	It recalls it at anything two o'clock .. dopy ...
提案手法 1	It recalls two o'clock in the dim .

表 14: 提案手法×の例 (単文)

例 1	
日本語文	クリントン 大統領 は、明らかにいら立っていた。
正解文	Clinton was apparently irritated by the situation .
ATLAS のみ	President Clinton had obviously gotten irritated .
提案手法 1	Clinton has gotten obvious .
例 2	
日本語文	彼女は野良猫を外へ放り出した。
正解文	She threw out the wild cat .
ATLAS のみ	She threw out the homeless cat to the outside .
提案手法 1	She threw the homeless in cat .

表 15: 提案手法×の例 (重文複文)

例 1	
日本語文	一つは林檎であり、もう一つはバナナであった。
正解文	One was an apple and the other was a banana .
ATLAS のみ	One was an apple , and another was a banana .
提案手法 1	There was an apple and another was a banana .
例 2	
日本語文	おどけて見せるのが好きだ。
正解文	He likes to play the joker .
ATLAS のみ	I like to joke and to show it .
提案手法 1	I like to joke .
例 3	
日本語文	この道は分かれてグローブ = シティに通じている。
正解文	This road branches off to Grove City .
ATLAS のみ	This road divides and runs to glove = city .
提案手法 1	This road is = and to gloves .

表 16: 対比較評価

	提案手法	提案手法 ×	提案手法	提案手法=
単文 (ATLAS のみ)	17 /100	9 /100	59 /100	15 /100
重文複文 (ATLAS のみ)	8 /100	42 /100	48 /100	2 /100
重文複文 (翻訳の王様のみ)	4 /100	44 /100	51 /100	1 /100

結果より，単文における対比較評価では，“提案手法の方が良い”と選択された数が，“提案手法の方が悪い”と選択された数よりも，わずかに多い．しかしながら，“差がない”とされた数が最も多く，提案手法の効果はあまり確認できなかった．

重文複文におけるルールベース翻訳のみとの対比較評価では，“提案手法の方が悪い”と選択された数が，“提案手法の方が良い”と選択された数よりも多い．この原因として，重文複文では文法構造が複雑なため，ルールベース翻訳での前処理において，英語の文法構造に類似させることができなかつたと考えている．このため，統計翻訳において，低精度に翻訳され，提案手法の翻訳精度が低下したと考えている．

しかしながら，提案手法における BLEU スコアでは，ATLAS のみと比較して 4.22% ，翻訳の王様のみと比較して 5.04% の向上が確認されている．この結果は人手による対比較評価の結果と整合してしない．この原因として，学習データとテストデータが同じ分野から抽出した文のため，統計翻訳によって，分野に適応し，BLEU スコアが向上したと考えている．なお，BLEU スコアでは 4-gram を用いて評価を行なっているため，分野に適応した文では，分野に適応していない文と比較して，BLEU スコアが高くなる傾向がある．

したがって，自動評価と対比較評価の不整合は，統計翻訳によって分野に適応した文が，BLEU スコアにおいて有利に評価されるといった問題から生じたと考えている．

6.2 未知語の影響

本研究では，市販されている翻訳ソフトを用いるため，翻訳ソフト固有の辞書データを利用している．したがって，統計翻訳のみのおときと比較して，未知語の大幅な減少が考えられる．よって，未知語に関して調査を行なった．未知語が減少した例を表 17 に示し，それぞれの未知語数を表 18 に示す．

表 17: 未知語の減少例

例 1(単文)	
日本語文	その 食事は オードブル で 始まった。
正解文	The meal opened with hors d'oeuvres .
ベースライン	The meal began with the オードブル .
提案手法 1	The meal began with the hors-d 'oeuvre .
例 2	
日本語文	組み合わせ は 籤 で きめられた。
正解文	The pairing were drawn by ballot .
ベースライン	The combination きめ in 籤 .
提案手法 1	The combination was determined by lot .
例 1(重文複文)	
日本語文	ちゃんと 縛って おけば ちゃんと 見つかる。
正解文	Safe bind , safe find .
ベースライン	If you can 縛っ properly properly .
提案手法 1	I have found it binds properly .
例 2	
日本語文	人が ごく 近く に 来る まで ウズラ は 隠れ 場 に ひそん で いる
	。
正解文	A quail will remain in its cover until one draws very near .
ベースライン	ウズラ is hiding in the very near to some people .
提案手法 1	The quail lurks in the hiding place till I call so that are very close .

表 18: 未知語数の比較

	単文		重文複文	
	総数	ユニーク数	総数	ユニーク数
ベースライン	3914	3480	4020	3576
ATLAS のみ	379	346	518	443
提案手法 1	310	297	370	348
翻訳の王様のみ	665	588	770	642
提案手法 2	446	429	492	461

結果より、ベースラインと比較して、提案手法 1 と提案手法 2 の両方において、未知語の大幅な減少が確認できた。また、ルールベース翻訳のみと比較しても、提案手法 1 と提案手法 2 の両方において、未知語の減少が確認できた。よって、未知語の減少が翻訳精度の向上に影響していると考えられる。

しかしながら、ルールベース翻訳のみと提案手法の未知語数では大きな差はない。それに関わらず、提案手法では、ルールベース翻訳のみと比較して、BLEU スコアが大きく向上していることから、BLEU スコアに対する未知語の影響は小さいものであると考えられる。

6.3 distortion-limit の影響

英' 英統計翻訳での，文法構造変換の必要性を調査するために，提案手法 1(ATLAS+統計翻訳) において，単語移動距離制限である “distortion-limit” の値を変化させて実験を行なった．例を表 19 と表 20 に示し，結果を表 21 に示す．

表 19: distortion-limit の違いによる出力例 (単文)

例 1	
日本語文	フランス語 は 父 の 知り合い の フランス 人 から 教わっ て い ま す 。
正解文	I'm learning French from a French gentleman my father knows .
ATLAS のみ	French is learnt from father 's acquaintance 's French .
distortion=1	French is learned from his father's his acquaintances French .
distortion=3	French is learned from his father's his acquaintances French .
distortion=6	French is learned from his father's his acquaintances French .
distortion=12	French is learned from his father's his acquaintances French .
distortion=-1	I am learning French from my father's his acquaintances French .
例 2	
日本語文	クリスマス は 年 に 1 度 し か こ ない 。
正解文	Christmas comes but once a year .
ATLAS のみ	Christmas comes only once a year .
distortion=1	Christmas comes once a year .
distortion=3	Christmas comes once a year .
distortion=6	Christmas comes once a year .
distortion=12	Christmas comes but once a year .
distortion=-1	Christmas comes but once a year .

表 20: distortion-limit の違いによる出力例 (重文複文)

例 1	
日本語文	この事件は裁判に訴えなければ、解決できないだろう。
正解文	If you do not take this case to court , it we not be settled .
ATLAS のみ	This event cannot be solved if it doesn 't take legal action .
distortion=1	This problem can be solved if you do not take legal action .
distortion=3	This incident can be solved if you do not take legal action .
distortion=6	This problem can be solved if you do not take legal action .
distortion=12	This problem can be solved if you do not take legal action .
distortion=-1	This incident can not be solved if you do not take legal action .
例 2	
日本語文	僕が最初米国へ行って、次に弟が行った。
正解文	I was the first to go to America , and my brother was the next to go .
ATLAS のみ	I went to the United States first , and , next , younger brother went .
distortion=1	I went to America first , and my brother went .
distortion=3	I went to America first , and then younger brother went .
distortion=6	I went to the United States , and my brother .
distortion=12	I went to the U .S . , and my brother .
distortion=-1	I went to the United States first , and my brother went .

表 21: ベースラインと提案手法の評価の比較

	単文		重文複文	
	BLEU	METEOR	BLEU	METEOR
ベースライン	0.1397	0.3845	0.1175	0.3531
ATLAS のみ	0.1552	0.4216	0.0942	0.3623
提案手法				
distortion=1	0.1878	0.4417	0.1389	0.3854
distortion=3	0.1870	0.4378	0.1356	0.3846
distortion=6	0.1877	0.4418	0.1364	0.3806
distortion=12	0.1871	0.4404	0.1398	0.3793
distortion=-1	0.1870	0.4404	0.1418	0.3865

結果より、単文では“distortion-limit”の値を変化させても、翻訳精度にあまり影響しないことが確認された。しかしながら、重文複文では“distortion-limit”の値が大きいときに、翻訳精度が高くなった。これは、単文ではルールベース翻訳による文法構造変換が高精度で行なっていたため、統計翻訳において、文法構造変換の必要性が低く、“distortion-limit”の影響が小さかったことが原因であると考えている。また、重文複文ではルールベース翻訳による文法構造変換が低精度で行なわれていたため、統計翻訳において、文法構造変換の必要性が高かったことが原因であると考えている。

このことから、前処理による文法構造変換を高精度に行なう必要があると考えている。また、対象文に応じて“distortion-limit”を変化させる必要もあると考えている。

7 おわりに

本研究では、文法構造が類似しているときには翻訳精度が高くなる統計翻訳の長所を用いるために、統計翻訳に対して、前処理としてルールベース翻訳を組み合わせた。結果として、単文において、ベースラインと比較して、BLEU スコアが 13.97%から 18.77%に 4.80%向上し、METEOR スコアが 38.44%から 44.18%に 5.74%向上した。また重文複文において、ベースラインと比較して、BLEU スコアが 11.75%から 13.64%に 1.89%向上し、METEOR スコアが 35.31%から 38.06%に 2.75%向上し、本研究の有効性が確認できた。

しかしながら、重文複文において、ルールベース翻訳のみの出力文と提案手法の出力文に対して、対比較評価したときに、提案手法の方が悪い結果となった。これは前処理において、英語の文法構造に類似させることができなかったことが原因であると考えている。よって、今後の研究では、重文複文に対しても、英語の文法構造に類似させることができる前処理の方法を調査することを考えている。

謝辞

最後に、一年間に渡り、本研究の御指導をいただきました鳥取大学工学部知能情報工学科計算機講座 C 研究室の池原悟教授，村上仁一准教授，徳久雅人講師に深く感謝するとともに厚くお礼を申し上げます。また、御指導をいただきました鳥取大学工学部知能情報工学科計算機講座 AB 研究室の菅原一孔教授，対比較評価を行なっていたいただきました山本理絵氏に厚くお礼を申し上げます。

参考文献

- [1] H.Schwenk, Marta R.Costa-jussa, and Jose A.R.Fonollosa, “Continuous space language models for the IWSLT 2006 Task”, in IWSLT, 2006, pages.166-173
- [2] Peter F.Brown, Stephen A.Della Pietra, Vincent J.Della Pietra, Robert L.Mercer, “The mathematics of statistical machine translation: Parameter Estimation”, Computational Linguistics, 1993
- [3] GIZA++
<http://www.fjoch.com/GIZA++>
- [4] Franz Josef Och, “Minimum error rate training for statistical machine translation”, Proceedings of the ACL, 2003.
- [5] L.Dugast, J.Senellart, and P.Koehn, “Statistical postediting on SYSTRAN’s rule-based translation system”, in Second Workshop on SMT, 2007, pages.179-182
- [6] M.Simard, N.Ueffing, P.Isabelle, and R.Kuhn, “Rule-based translation with statistical phrase-based post-editing”, in Second Workshop on SMT, 2007, pages.203-206
- [7] SRILM(The SRI Language Modeling Toolkit) : srilm.tgz
<http://www.speech.sri.com/projects/srilm/>
- [8] Moses : moses.2007-05-29.tgz
<http://www.statmt.org/moses/>
- [9] 西山七絵, 村上仁一, 徳久雅人, 池原悟, “単文句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
- [10] ChaSen, <http://chasen-legacy.sourceforge.jp/>
- [11] 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
- [12] NIST Open Machine Translation (OpenMT) Evaluation
<http://www.itl.nist.gov/iad/mig/tests/mt/>

[13] The METEOR Automatic Machine Translation Evaluation System

<http://www.cs.cmu.edu/~alavie/METEOR/>

[14] 福田智大, 村上仁一, 徳久雅人, 池原悟, “ルールベース翻訳を前処理に用いた統計翻訳”, 言語処理学会第 16 回年次大会 (発表予定)