英日 SMT への Head-Final 制約の導入

西村 拓哉[†]　山本 博史[††,†††]　大熊 英男[†††]　村上 仁一[†]

[†]鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
{s062044, murakami}@ike.tottori-u.ac.jp
[††]近畿大学 理工学部 情報学科
yama@info.kindai.ac.jp
[†††] National Institute of Information and Communications Technology
{hirofumi.yamamoto, hideo.okuma}@nict.go.jp

1　はじめに

現在，機械翻訳において，統計的機械翻訳 (SMT) が注目され，研究が盛んに行われている [1]．SMT では，英語-スペイン語などの構文構造が類似している言語間における翻訳精度が高い傾向がある．しかしながら，英語-日本語などの構文構造が異なる言語間においては，翻訳精度が低い傾向にある．

SMT では，一般に N-gram モデルとリオーダリングモデルを文法情報として用いる．しかし，N-gram モデルは局所的な文法情報を表現することが出来るが，大局的な文法情報を表現することができない．一方，リオーダリングモデルとしては距離依存モデルやレキシカルリオーダリングがよく用いられるが，長距離の語順入れ替えに対する制約は非常に弱く，N-gram モデルと同様に，大局的な表現ができるとは言い難い．また，語順が大きく異なる言語間の翻訳の場合，語順入れ替えが複雑になり，誤りを含む可能性が高くなる．したがって，構文構造が大きく異なる言語間では翻訳精度が低下すると考えられる．

そこで，本研究では，リオーダリングモデルに構文解析を利用した，山本らが提案する IST-ITG 制約 [2] を用いる．さらに，磯崎らの提案する Head-Final 制約 [3] を利用した手法を組み合わせて，より強力なリオーダリングモデルを作成し，このリオーダリングモデルを用いたシステムでの翻訳精度の調査を行う．

2　SMT システム

SMT では，ソース言語の入力文 $\mathbf{f}$ が与えられたとき，全ての組合せの中から確率が最大となるターゲット言語文 $\hat{e}$ を探索し，翻訳を行う [4].

$$\begin{aligned}\hat{e} &= \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) \\ &\approx \arg\max_{e} P(f|e)P(e) \qquad (1)\end{aligned}$$

$$\begin{aligned}P(f|e) &= P(\bar{f}_1^{\,I}|\bar{e_1}^I) \\ &= \prod_{i=1}^{I} (\bar{f}_i|\bar{e_i}) d(start_i, end_{i-1}) \qquad (2)\end{aligned}$$

(1) 式の $P(f|e)$ は翻訳モデル，$P(e)$ は言語モデルを表す．ここで $P(f|e)$ を分解すると，(2) 式となる．ソース言語の入力文 $\mathbf{f}$ は I 個のフレーズ $\bar{f}_1^{\,I}$ に分割され，$\bar{f}_1^{\,I}$ の各フレーズ $\bar{f}_i^{\,I}$ はターゲット言語のフレーズ $\bar{e_i}$ に翻訳される．フレーズ翻訳には確率分布 $(\bar{f}_i|\bar{e_i})$ を用いる．また，フレーズの入れ替えについてはリオーダリングモデル $d(start_i, end_{i-1})$ を用いる．

3　リオーダリングのための制約

3.1　距離依存モデル

リオーダリングモデルでは，フレーズを並べ替える際に，歪み確率分布 $d(start_i, end_{i-1})$ を用いる．$start_i$ はソース言語のフレーズが i 番目のターゲット言語のフレーズに翻訳された最初の位置であり，end_{i-1} はソース言語のフレーズが i　1 番目のターゲット言語のフレーズに翻訳された最後の位置を表している．距離依存モデル [5] では，(3) 式における，パラメータ α に適切な値を用いて計算される．

$$d(start_i, end_{i-1}) = \alpha^{|start_i - end_{i-1} - 1|} \qquad (3)$$

3.2　レキシカルリオーダリングモデル

レキシカルリオーダリングモデル [6] では，リオーダリング確率として $P_s(o|f_i, e_i)$ を用いて，各単語対 $\{f_i, e_i\}$ の右側 (r) と左側 (l) に割り当てられる．リオーダリング位置は “monotone(m)”，“swap(m)”，“discontinuous(d)” の 3 つに分類され，ソース言語の単語列 f_{i-1}, f_i に対するターゲット言語の単語列の確率は，以下の式でそれぞれ計算される．

$$\begin{aligned}\text{monotone} &: P_r(m|f_{i-1}, e_{i-1}) P_l(m|f_i, e_i) \\ \text{swap} &: P_r(s|f_{i-1}, e_{i-1}) P_l(s|f_i, e_i) \\ \text{discontinuous:} &\ P_r(d|f_{i-1}, e_{i-1}) P_l(d|f_i, e_i)\end{aligned}$$

3.3　ITG 制約

ITG 制約 [7, 8] とは，Wu 氏が提案したリオーダリングのための制約である．ITG 制約では，ソース言語側の二分木のノードを回転して得られるリオーダリングを用いる．

ソース言語側の 4 フレーズ $(e_1, ..., e_4)$ について考えると，通常のリオーダリングではリオーダリング数は $4! = 24$ 通りとなる．ITG 制約では 4 フレーズに対して考えられる全ての二分木，つまり図 1 の 3 つの二分木を考慮する．ここで，$[e_3, e_1, e_4, e_2]$ と $[e_4, e_2, e_3, e_1]$ の 2 つの組合せが削除される．したがって，リオーダリング数

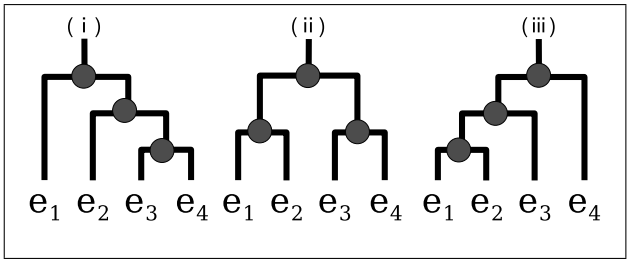

図 1　4 フレーズにおける二分木

は 24　2 = 22 通りとなる．4 フレーズの場合には 2 通りしか削減されないが，10 フレーズについて考えると，通常のリオーダリングでは 3,628,800 通りとなるが，ITG 制約では 206,098 通りとリオーダリング数は大幅に削減される．

3.4　IST-ITG 制約

IST-ITG 制約は構文解析によって得られる構文木を利用した，リオーダリングのための制約である．IST-ITG 制約は ITG 制約が元となっており，ITG 制約における二分木を，構文解析によって得られる構文木のみに限定することで ITG 制約よりも強力な制約となっている．

3.3 節と同様に，ソース言語側の 4 フレーズについて考えると，通常のリオーダリングでは 24 通りであり，ITG 制約では 22 通りである．IST-ITG 制約では構文解析によって得られた構文木の各ノードを回転して得られるリオーダリングを用いる．したがって，図 1 における 3 つの二分木の中から，構文解析で得られる構文木として，いずれかの二分木を用いた IST-ITG 制約の場合，リオーダリング数は $2^{4-1} = 8$ 通りとなる．

4　Head-Final 制約の導入

4.1　Head-Final 制約

Head-Final 制約は，日本語のように構文構造が “Subject-Object-Verb” の SOV 型である言語と，英語のような SVO 型である言語の語順を近似するための制約である．SOV 型である言語を “Head-Final” 型言語と考えると，SVO 型言語における “Head” を “Final” へ移動することで，SVO 型言語が “Head-Final” 型言語となる．したがって，SOV 型言語と SVO 型言語の双方が Head-Final 型言語となり構文構造が近似する．

4.2　磯崎法

磯崎らが提案する Head-Final 制約の例を図 2 に示す．尚，図中の “*” は “Head” を表す．

英語文 “I saw a beautiful girl yesterday” に対して構文解析を行うと，図 2 の構文木が得られる．次に，その構文木を括弧つきの文で表現すると，((I)(*(*(saw)((a)((beautiful)(girl))))(yesterday))) と表すことが出来る．そして，括弧つき文において “Head” であるノードともう一方のノードを回転することで，((I)(*(yesterday)(*((a)((beautiful)(girl)))(saw)))) となる．最終的に得られた括弧つき文に対して逐語訳を行うと，“私は昨日美しい少女を見た” となり，英語の語順が日本語の語順と近似する．

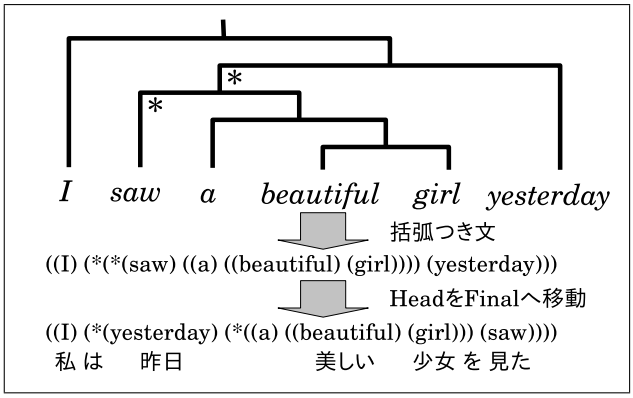

図 2　磯崎法

4.3　提案手法

提案手法では 3.4 節，4.1 節で述べた 2 つの制約を用いる．提案手法の実験手順を以下に示す．

手順 1　英語テスト文に対して構文解析を行う
手順 2　構文木から括弧つき英語テスト文を生成する
手順 3　デコーダに括弧つき英語テスト文を入力する

リオーダリングを得る際，まず IST-ITG 制約を適用してリオーダリングを得る．IST-ITG 制約では，構文木の各ノードに対する制約が存在しない．ここで提案手法では，各ノードに対して Head-Final 制約を適用する．Head-Final 制約の適用により，各ノードの自由度に制限を与えることで，IST-ITG 制約よりも強力な制約となる．

ここで注意すべきことは，本研究で用いる Head-Final 制約は，磯崎法とは適用方法が異なる点である．磯崎らの提案する Head-Final 制約はデコーディングの前処理として制約を用いる．一方，本研究では，Head-Final 制約をデコーディング時のリオーダリングのペナルティとして使用し，IST-ITG 制約での各ノードにおけるリオーダリングの重みづけに用いる．

今回使用するペナルティの計算式を以下に示す．尚，p の値は任意に設定可能である．

$$penalty = \log(p) \quad \log(1 \quad p) \tag{4}$$

5　実験環境

5.1　実験データ

実験データには，JST(新聞記事) コーパスの日英対訳文を用いる．学習データに 994,000 文，ディベロップメントデータに 2,000 文，テストデータに 1,997 文を用いる．日本語文のセグメントには，MeCab[9] を用いて形態素解析を行い，形態素と句読点の間にスペースを入れる．英語文の構文解析器には，東京大学の辻井研究室で公開されている Enju パーザ [10] を用いる．英語文のセグメントに関しては，Moses[11] に付属している tokenizer.perl と Enju パーザのセグメントの整合性をとっている．JST コーパスの例を表 1 に示す．

表 1 JST コーパスの例

日本語文	これは，適切な効率で中性子捕獲反応の後，放出する数本の高エネルギー γ 線を吸収するように可能な限り近接した照射試料を取り囲む BGO 遮蔽体を備えた 2 台の EUROBALL クラスターと 4 台のクローバー検出器から成っている。
英語文	It consists of two EUROBALL clusters and four clover detectors with BGO shields surrounding a sample as close as possible so as to absorb a few γ - rays of high energy emitted after the neutron capture reaction with reasonable e ciency .
日本語文	神経原性および低容量性ショックから意識消失発作をきたした腹直筋鞘血腫（rectus sheath hematoma，以下 RSH）の 1 例を経験した。
英語文	We reported a case of rectus sheath hematoma (RSH) which caused neurogenic and hypovolemic shock .

5.2 デコーダ

本研究では, Moses の上位互換デコーダである CleopATRa を用いる．CleopATRa では構文解析の出力である括弧つき文をデコーディングすることが出来る．尚，パラメータはディベロップメントデータを用いて最適化する [12].

6 実験結果

6.1 自動評価

自動評価には BLEU[13] を用い，その評価結果を表 2 に示す．また，提案手法における BLEU の値の変化のグラフを図 3 に示す．表中のベースラインは IST-ITG 制約のみを用いたシステムであり，提案手法の横の () はペナルティの計算で用いた p の値を表している．尚，(4) 式から $p = 0.5$ のとき，ベースラインと提案手法は同じ結果となる．

表 2 自動評価結果

	BLEU
ベースライン	30.27
提案手法 ($p = 0.49$)	**30.37**
提案手法 ($p = 0.001$)	25.93

表 2 の結果から，少しのペナルティを与えた場合 ($p = 0.49$) ではわずかに BLEU の値が向上している．しかし，p の値を小さくし ($p = 0.001$)，与えるペナルティを大きくすると BLEU の値が低下している．

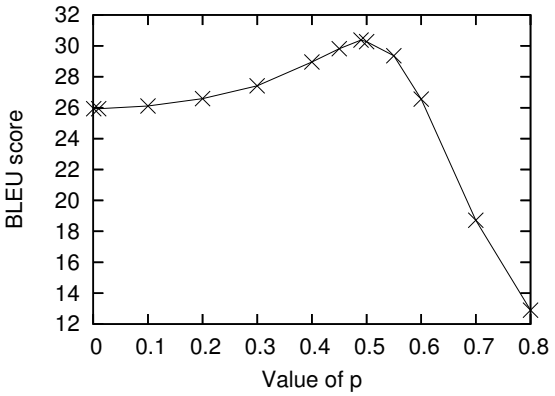

図 3 評価結果のグラフ

6.2 対比較評価

ベースラインの出力結果と $p = 0.49$ の場合での提案手法の出力結果に対して人手による対比較評価を行う．評価対象には出力が異なる結果からランダムに抽出した 100 文を用いる．尚，出力の異なる文数は 1,997 文中 518 文であった.

6.2.1 判断基準

人手による 4 つの判断基準に基づいて評価を行う．評価基準と評価例を以下に示す.

評価 1(>) 提案手法の翻訳結果がベースラインよりも優れている

入力文: On the possibility of the optical device using the sub - wavelength structure , the research from various directions succeeded as a whole .
正解文: サブ波長構造を使った光素子の可能性について，種々の方向から行った研究は全体としては成功したと言える。
ベースライン: サブ波長構造を用いた光デバイスの可能性について，全体として成功した様々な方向から研究した。
提案手法: サブ波長構造を用いた光デバイスの可能性について，全体として成功した様々な方向から研究した。

評価 2(<) 提案手法の翻訳結果がベースラインよりも劣っている

入力文: One scheme is based on the discard of TCP retransmissionpackets , and the other is based on the suppression of transmission delay fluctuation .
正解文: 一つは TCP 再送パケットを捨てることを基礎とし，他は伝送遅延変動を抑圧することを基礎とした。
ベースライン: 一つの方式をベースにした TCP の再送パケットを廃棄し，他は伝送遅延変動の抑制に基づく。
提案手法: 一つの方式は TCP の再送パケット廃棄に基づき，他は伝送遅延変動の抑制に基づく。

評価 3(≈) どちらも似たような文である，またはどちらも入力文で伝えたい情報が理解できない

入力文: In 15 eyes , the fixation point did not agree with the oral reading position .
正解文: 15 眼では固視点と音読部位が一致しなかった。
ベースライン: 15 眼では，固定点音読部位が一致しなかった。
提案手法: 15 眼では固定点音読部位が一致しなかった。

6.2.2 評価結果

対比較評価結果を表 3 に示す.

表 3 対比較実験の結果

評価 1(>)	評価 2(<)	評価 3(≈)
13	12	75

表 3 の結果から，BLEU での評価結果と同様に，対比較評価の結果においても，提案手法とベースラインの間に差がない結果となっている.

7 考察

今回の実験において，提案手法によって性能は改善したが，わずかであった．この原因として以下の 2 点が挙げられる.

1. 探索空間

   IST-ITG 制約では語順が入れ替わった場合の位置を決定する．例えば 15 単語の文の場合は探索空間を 1/14 にできる．これに対し，提案手法では，構文木の各ノードで入れ替えが生じるかどうかの判定であり，探索空間は 1/2 にしかならない．したがって，提案手法では探索空間が大きく変化しないために，効果が少なかったと考えられる.

2. 機能語と内容語

   日本語において Head-Final 制約は文節単位でないと成立しない．従って，提案した制約についても文節に対しての制約にする必要がある．この際，日本語の内容語が英語の内容語に対応し，機能語についても同様に対応しているのであれば問題はない．しかしながら，実際に対応をとることは困難である．特に，英単語の “have” や “do” などは日本語の動詞に対応する場合と助動詞に対応する場合があり，両者の区別をすることは難しい．したがって，今回提案した手法において，Head-Final 制約を成立するために，確率の形で表現せざるをえなかったため，性能がでなかったと考えられる.

8 おわりに

本研究では，リオーダリングの制約である IST-ITG 制約に Head-Final 制約を組み合わせた手法を提案した．実験の結果として，提案手法により翻訳精度は改善したが，その差はわずかであった.

この原因としては，IST-ITG 制約の各ノードに対する制限として用いる，Head-Final 制約のリオーダリングに対する影響が小さいことが挙げられる．一方で，IST-ITG 制約単体と提案手法の翻訳精度に差がほとんどないことから，IST-ITG 制約を用いて，より強力な制約を作成する必要性はないと考えられる.

謝辞

今回の研究を行うにあたり，御指導，御支援していただきました，隅田室長をはじめ，NICT 言語翻訳グループの皆様に深く御礼を申し上げます.

参考文献

[1] Richard Zens, Franz Josef Och, Hermann Ney, “Phrase-based Statistical Machine Translation”, KI, pp35-56, 2002.

[2] Hirofumi Yamamoto, Hideo Okuma, Eiichiro Sumita, “Imposing Constraints from the Source Tree on ITG Constraints for SMT”, Association for Computational Linguistics, pp.1-9, 2008.

[3] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, Kevin Duh, “Head Finalization : A simple Reordering Rule for SOV Languages”, Association for Computational Linguistics, pp.244-251, 2010.

[4] Philipp Koehn, Franz Josef Och, Daniel Marcu, “Statistical Phrase-Based Translation”, Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting, pp.48-54, 2003.

[5] Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Andrew S. kehler, Robert L. Mercer, “Language translation apparatus and method using context-based translation models”, United States Patent, Patent Number 5510981, 1996.

[6] Christoph Tillmann, “A Unigram Orientation Model for Statistical Machine Translation”, Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting, pp.101-104 , 2004.

[7] Dekai Wu, “Stochastic Inversion Transduction grammars, with Application to Segmentation, Bracketing, and Alignment of Parallel Corpora”, Proceeding of the 14th International Joint Conference on Artificial Inteligence, pp.1328-1334, 1995.

[8] Dekai Wu, “Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora”, Association for Computational Linguistics, pp.377-403, 1997.

[9] MeCab, MeCab : Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/

[10] Enju, Tsujii Laboratory : Enju - A Practical HPSG parser, http://www-tsujii.is.s.u-tokyo.ac.jp/enju/

[11] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation”, Association for Computational Linguistics, pp177-180, 2007.

[12] Franz Josef Och, “Minimum Error Rate Training in Statistical Machine Translation”, Association for Computational Linguistics, pp160-167, 2003.

[13] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, “BLEU : a Method for Automatic Evaluation of Machine Translation”, Association for Computational Linguistics, pp311-318, 2002.