目的言語の構文解析器を用いた機械翻訳のプレオーダリング

後藤 功雄 †* 内山 将夫 †† 隅田 英一郎 †† 黒橋 禎夫 †††
† 日本放送協会 †† 情報通信研究機構 ††† 京都大学

1 はじめに

日英など語順が異なる言語間の機械翻訳では、目的言語の語順の推定は重要な課題である。語順を推定するために統計的機械翻訳（SMT）では、語彙化語順推定モデル [17]、階層フレーズベース SMT[1]、構文ベース SMT[7]、プレオーダリング [19] などの手法が提案されてきた。プレオーダリングは、長距離の語順並べ替えに有用な原言語の構文構造をシンプルな方法で活用できるという特徴がある。これまでにプレオーダリングにより英日は SMT の翻訳品質が大幅に向上した [9, 4] が、これに比べて日英・中英では英日ほどの改善は実現できていない。その理由は、日・中の構文解析結果の構造やその活用方法が英語への翻訳に最適ではないためと考えられる。また、原言語の構文解析器を必要とする手法は適用できる原言語が限られる。この状況を改善する方法として、英語の句構造を日・中に射影することにより、日英・中英の翻訳に適した日・中の構文解析および語順推定のモデルを自動構築し、それを利用してプレオーダリングする。これにより、日英・中英の翻訳品質が従来手法と比べて大幅に改善した。*1

2 プレオーダリング手法

機械翻訳は原言語文 F を目的言語文 E へ変換する。プレオーダリングによる翻訳は次の 2 段階の処理で翻訳する。はじめに、F を、ほぼ目的言語の語順である原言語の単語列 F' に並べ替える（プレオーダリング）。次に、F' を E に翻訳する。

プレオーダリング手法には多くの研究がある。ほとんどの手法は、原言語の構文解析器を用いて得られる構文構造と並べ替えルールを用いる [19, 2, 9]。原言語の構文解析器の出力や既存手法の並べ替えルールは翻訳に最適であるとは限らない。また、原言語の構文解析器が利用できない場合は適用できない。この場合、構文解析器を必要としない手法 [13] が有用である。この手法は対訳コーパスと単語アラインメントを用いてシンタックスに基づかない構造（非構文の構造）の解析器を構築する。そして、原言語文の構造を解析して BTG[18] に基づいて並べ替える。

非構文の構造に比べて構文構造は、語順の推定において次の 2 点で優れていると考えられる。

- 構文構造は意味のまとまりと部分構造が一致していると考えられる。例えば、節は意味のまとまりになっておりかつ構文構造の部分構造になっている。それに対して、非構文の構造は必ずしも意味のまとまりと部分構造が一致するとは限らない。
- 構文構造は非構文の構造より情報量が多い。構文構造は多くのフレーズラベルを用いるが、非構文の構造は 1 種類のフレーズラベルしか用いない。

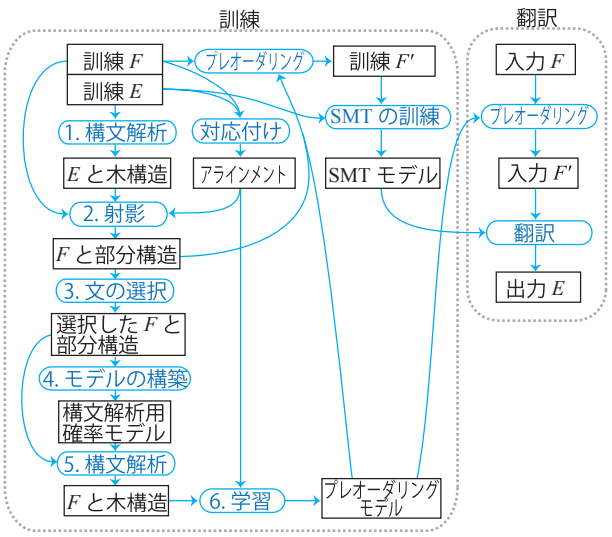

図 1 提案手法の概要

3 提案手法の概要

提案手法は、原言語の構文解析器が利用できない場合でも、目的言語の構文解析器を用いることで、構文構造を用いてプレオーダリングすることができる。対訳文では原言語の構文構造と目的言語の構文構造は類似していることが期待される [8]。我々はこの期待を利用して原言語の構文構造を構築し、ITG[18] に基づくプレオーダリングモデルを学習する。

ITG を効果的に学習するためには、同期率が高い対訳構造を用いることが重要である。なぜなら、ITG は同期している部分からのみ学習可能なためである。そこで、言語間の射影により同期率の高い構文構造を構築することで、ITG の効果的な学習を促進する。

提案手法の概要を図 1 に示す。プレオーダリングモデルは次のステップで構築する。

1. 対訳コーパスの目的言語文の 2 分木構文構造を構文解析器を用いて獲得する。
2. 目的言語文の部分的な構文構造を、単語アラインメントを用いて原言語文に射影する。（4.1 節）
3. 射影された部分構造を用いて同期率の高い対訳文を選択する。（4.2 節）

* 本研究は情報通信研究機構および日本放送協会にて実施した。
*1 より詳細については [5] を参照。

4. 射影された部分構造を用いて確率的 CFG と教師無しの確率的品詞推定モデルを構築する。(4.3 節)
5. 構築した確率モデルを用いて射影された部分構造の制約のもとで訓練データの原言語文を構文解析して同期率の高い構文構造を構築する。(4.4 節)
6. 構築した原言語の構文構造と単語アラインメントを用いて ITG に基づくプレオーダリングモデルを学習する。(4.5 節)

本稿のメインの貢献は、目的言語の構文解析器を用いたプレオーダリングの枠組みである。これに加えて、原言語の構文・品詞解析器が不要な、射影による句構造構築手法を提案する。提案手法は既存の射影による句構造構築手法 [10] と比べて次の 2 つの違いがある。(1)CFG の確率推定において、既存手法では射影から得られる曖昧性のある候補の確率に一様分布を仮定しているが、この仮定は正しくない。それに対して提案手法は各候補の確率を計算する。(2) 既存手法は原言語の品詞タグを必要とするが、提案手法は必要としない。(原言語文の単語分割は必要)

4 モデルの訓練

本節では、前記のステップ 2 以降を説明する。

4.1 部分構造の射影

単語アラインメントを用いて目的言語文の部分的な構文構造を原言語文に射影する。これによって原言語文の部分的な構文構造が得られる。例を図 2 の上部に示す。

射影は、E の部分木のスパンに対応する F のスパンを単語アラインメントを用いて同定し、E の部分木の根のフレーズラベルを F のスパンに追加することで行う。F のスパンは E の部分木中の語にアラインメントされた語のうち、左端から右端までとする。この F のスパンを最小射影スパンと呼ぶ。アラインメントされていない語が最小射影スパンに隣接している場合は、これらの語はこのスパンに含まれる可能性がある。すなわち、この場合はスパンに曖昧性がある。図 2 では、最小射影スパンを水平の実線、アラインメントされていない語の部分を水平の破線で表している。

射影された構造から木構造が構成可能で、それらがなるべく高品質なものであるように、最小射影スパンが互いに部分的に重複する場合（不整合）は、それらの部分構造を破棄する。

4.2 同期率の高い対訳文の選択

射影した部分構造を用いて構造の同期率が高い対訳文を選択する。選択した文は 4.3〜4.5 節で用いる。各対訳文対での同期率は、射影されたスパンの数を (原言語文中の語数 − 1) で割った値で計算する。この値が高ければ、不整合を起こさずに射影できた部分構造が多いことを意味する。

4.3 構文解析用の確率モデルの構築

射影された部分構造からプレオーダリングモデルの学習に用いる 2 分木構造を獲得するために、構文解析用の

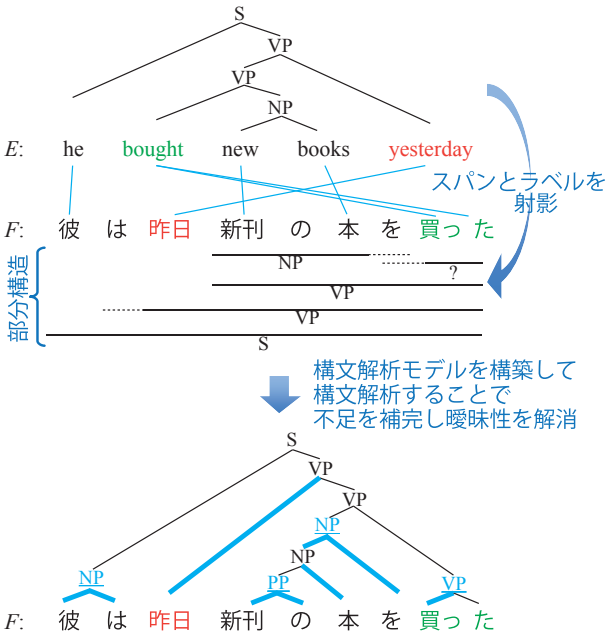

図 2　構文構造の射影と 2 分木構造の構築例

確率モデルを構築する。

入力は、F と射影された部分構造である。出力は、原言語に対する確率的文脈自由文法（PCFG）および教師無しの確率的品詞推定モデルである。我々は、Pitman-Yor 過程 (PY) [15] を用いてモデルを構築する。なぜなら、その “rich-get-richer” の特性が部分的にアノテーションされたデータからモデルを学習するのに適しているためである。

ここで用いる CFG ルール $x \to \alpha$ は、非終端記号 $x \in V$ と 2 つの非終端記号の並び α からなる。非終端記号の集合 V は $V = \mathcal{L} \cup \mathcal{T}$ で、$\mathcal{L}$ はフレーズラベルの集合である。$\mathcal{T} = \{1, 2, \cdots, |\mathcal{T}|\}$ は原言語の教師無しの品詞タグを表す数字の集合で、$|\mathcal{T}|$ は品詞タグの種類数である。$\mathcal{F}$ を訓練データ中の原言語の単語集合とし、$f \in \mathcal{F}$ を用いて、$F = f_1 f_2 \cdots f_m$ とする。木構造 D の確率は、その構成要素である CFG ルールと単語の確率の積で次のように計算する。

$$P(D) = \prod_{x \to \alpha \in \mathcal{R}} P(\alpha|x)^{c(x \to \alpha, D)} \prod_{i=1}^{m} P(f_i|t_i) \quad (1)$$

ここで、$\mathcal{R}$ は CFG ルールの集合を表し、$c(x \to \alpha, D)$ は D での $x \to \alpha$ の頻度を表し、$t \in \mathcal{T}$ は品詞タグを表し、t の添え字 i は F 中の単語位置を表す。木構造の根のフレーズラベルには、特定のフレーズラベル S を用いる。

PY モデルは CFG ルールまたは原言語の単語の確率分布として次のように表される。

$$P(\alpha|x) \sim \mathrm{PY}_x(d_{\mathrm{cfg}}, \theta_{\mathrm{cfg}}, P_{\mathrm{base}}(\alpha|x))$$
$$P(f|t) \sim \mathrm{PY}_t(d_{\mathrm{tag}}, \theta_{\mathrm{tag}}, P_{\mathrm{base}}(f|t))$$

ここで、d_{cfg}, θ_{cfg}, d_{tag}, θ_{tag} は、PY モデルのハイパーパラメータである。これらは、文献 [16] の手法を用いて最適化する。バックオフの確率分布には一様分布である

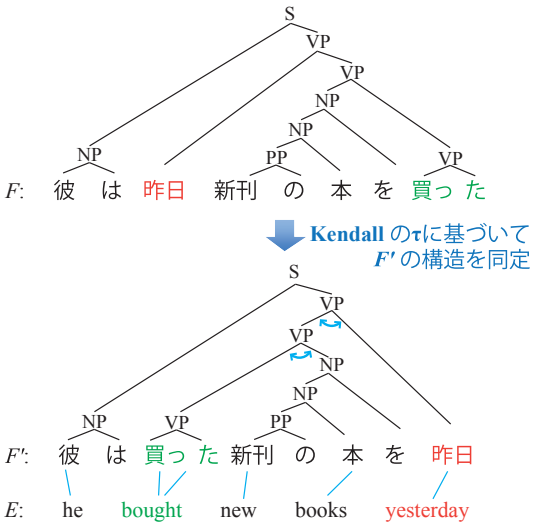

図 3　F' の構造の計算例

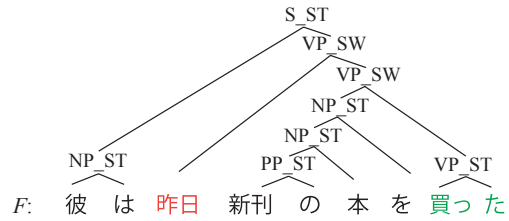

図 4　_ST と _SW が付与された F の 2 分木構造の例

$P_{\text{base}}(\alpha|x) = 1/|V|^2$ および $P_{\text{base}}(f|t) = 1/|\mathcal{F}|$ を用いる。ここで、$|V|$ は非終端記号の種類数、$|\mathcal{F}|$ は訓練データ中の原言語の単語の種類数を表す。

モデル構築のためのサンプリングは、式 (1) および次の制約に基づいて行う。最小射影スパンが存在する場合は、アラインメントされていない語を除いてそのスパンと非整合にならない構造をサンプリングする。そして、フレーズラベルが射影されているスパンが部分木としてサンプリングされる場合には、そのフレーズラベルには、射影されているフレーズラベルがサンプリングされる。

サンプリングには、動的計画法に基づいた文レベルのギブスサンプラーを用いる [11]。このサンプラーは、各文において、CYK アルゴリズムを用いてボトムアップに内側確率を計算し、次に各 CFG ルールの内側確率を用いてトップダウンで木構造をサンプリングする。計算コストを削減するために、内側確率を計算する際には文中の各語に対して確率が上位の品詞タグのみを用いた。後の実験では、上位 5 の品詞タグを利用した。

4.4　構文解析による同期率の高い構造の獲得

構築した PY モデルを用いて訓練データの原言語文を構文解析して、射影されたスパンやラベルの不足を補完し、スパンの曖昧性を解消することで、目的言語文の構造と同期率が高い 2 分木構造を獲得する。これは射影されたスパンとフレーズラベルの制約の下で、CYK アルゴリズムを用いて式 (1) が最尤の 2 分木構造を計算することで行う。例を図 2 の下部に示す。

なお、最小射影スパンは 2 分木構造に必ずしも含まれるとは限らない。例えば、括弧がスパンを表すとする

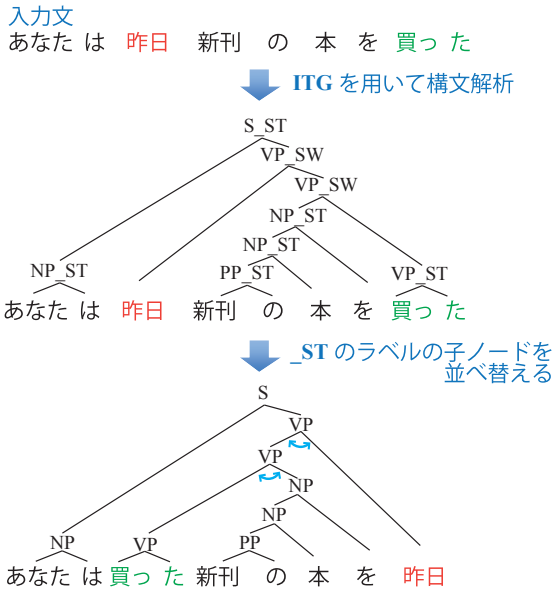

図 5　入力文のプレオーダリング例

と、最小射影スパンが $(f_1 f_2) f_3$ で f_3 がアラインメントされていない場合に、2 分木構造は $(f_1 (f_2 f_3))$ である可能性がある。

4.5　プレオーダリングモデルの学習

構築した原言語の 2 分木構文構造と単語アラインメントを用いてプレオーダリングモデルを学習する。提案手法のプレオーダリングモデルは、PCFG を用いた構文解析の枠組みと ITG の枠組みを組み合わせたモデル (ITG 構文解析モデル) として構築する。

まず、プレオーダリングモデルの訓練データを構築する。獲得した F の 2 分木構造の子ノードの順番を替えることによって得られる、E の語順に最も近い F' の構造を Kendall の τ を用いて同定する。F の構造と単語アラインメントから得られる F' の構造の例を図 3 に示す。F の構造に対して、F と F' とで子ノードの順番が替わったノードには “_SW” を付与し、順番が同じノードには “_ST” を付与する。図 4 に _SW と _ST が付与された F の 2 分木構造の例を示す。この構造は ITG からの導出と見なすことができる。次に、PCFG 学習アルゴリズムを用いてこの 2 分木構造から ITG 構文解析モデルを学習する。学習したモデルが提案手法のプレオーダリングモデルとなる。ここでは、隠れクラスを用いる PCFG 学習アルゴリズム [14] を用いる。

5　プレオーダリング

入力文をプレオーダリングするプロセスの例を図 5 に示す。まず、ITG 構文解析モデルで入力文を構文解析する。この際に、並べ替えも同時に特定される。さらに得られた 2 分木で “_SW” が付与されたノードの子ノードの順番を変えることでプレオーダリングする。

訓練データのプレオーダリングは、4.1 節で射影されたスパンの制約のもとで ITG 構文解析モデルで原言語文を構文解析してプレオーダリングする。

6 実験

NTCIR-9 と NTCIR-10 の特許機械翻訳タスク [4, 3] のデータを用いて日英翻訳と中英翻訳の実験を行った。

6.1 設定

NTCIR-9 と NTCIR-10 では、訓練データおよび開発データは同じで、テストデータは異なる。訓練データは日英が約 318 万文対で中英が 100 万文対、開発データは 2,000 文対である。テストデータは NTCIR-9 は 2,000 文、NTCIR-10 は 2,300 文である。英語の構文解析器には Enju を用い、日本語の単語分割に MeCab、中国語の単語分割に Stanford segmenter を用いた。日本語の英数字の単語分割は英単語の単位に合わせた。翻訳モデルの学習は、40 単語以下の文で英語側の文が構文解析できたもの（日英：約 206 万文対、中英：約 40 万文対）を訓練データとして用いた。単語アラインメントは GIZA++ と grow-diag-final-and ヒューリスティックおよび誤り低減の前後処理 [6] により獲得した。5-gram の言語モデルを訓練データの目的言語文で学習した。

提案手法（Propose）の学習は次のように行った。4.2 節の対訳文の選択では、上位 10 万文を選択した。4.3 節の確率モデルの学習では、$|\mathcal{T}| = 50$ を用い、ギブスサンプリングをデータ全体に対して 100 回行った。Berkeley parser [14] をプレオーダリングモデルの学習と構文解析に用いた。翻訳にはフレーズベース SMT の Moses[12] を用い、distortion limit を 6 に設定した。

比較手法として、次の 6 つの手法を用いた。

- フレーズベース SMT ＋語彙化語順推定モデル（Pbmt_L）[12]
- 階層フレーズベース SMT（Hpbmt）[1]
- String-to-tree 構文ベース SMT（Sbmt）[7]
- フレーズベース SMT ＋単語列ラベリングに基づく語順推定モデル（Pbmt_D）[6]
- 原言語の依存構造解析器を用いたプレオーダリング（Srcdep）[2][*2]
- 構文解析器不要のプレオーダリング（Lader）[13]

Pbmt_D は Moses 互換のデコーダー、他は Moses を用いて翻訳した。Pbmt_L の語順推定モデルの学習には翻訳モデルの訓練データを全て用いた。Pbmt_D の語順推定モデルの学習には 20 万文を用いた。Srcdep で利用する依存構造解析には CaboCha（日本語）と Stanford parser & tagger（中国語）を用いた。CaboCha の出力は単語の依存構造に変換して利用した [5]。Srcdep の並べ替えルールの学習には翻訳モデルの訓練データを全て用いた。Lader のプレオーダリングモデルの学習には Propose と同じ 10 万文の訓練データを用いて 100 回の繰り返し計算を行った。Hpbmt と Sbmt の max-chart-span は無制限とし、他の手法の distortion limit はシステム名の添え字で示す。

[*2] 3 種類のルール選択基準の内、crossing score 最適化を用いた。

表 1　日英翻訳評価結果

	NTCIR-9		NTCIR-10	
	RIBES	BLEU	RIBES	BLEU
$\text{Pbmt}_{L\text{-}4}$	65.48	26.73	65.53	27.44
$\text{Pbmt}_{L\text{-}20}$	68.79	30.92	68.30	31.07
Hpbmt	70.11	30.29	69.69	30.77
Sbmt	72.54	31.94	71.32	32.40
$\text{Pbmt}_{D\text{-}20}$	73.54	33.14	72.23	33.87
Srcdep$_{-6}$	71.88	29.23	71.20	29.40
Lader$_{-6}$	74.31	32.98	73.98	33.90
Propose$_{-6}$	**76.35**	**33.83**	**75.81**	**34.90**

表 2　中英翻訳評価結果

	NTCIR-9		NTCIR-10	
	RIBES	BLEU	RIBES	BLEU
$\text{Pbmt}_{L\text{-}4}$	75.02	29.22	74.24	30.65
$\text{Pbmt}_{L\text{-}10}$	76.11	31.20	75.41	32.34
Hpbmt	77.68	32.39	77.45	33.61
Sbmt	78.44	32.47	77.68	33.90
$\text{Pbmt}_{D\text{-}10}$	77.98	33.03	77.48	34.28
Srcdep$_{-6}$	76.88	28.85	76.14	29.36
Lader$_{-6}$	78.18	30.80	77.06	31.12
Propose$_{-6}$	**81.61**	**35.16**	**81.05**	**36.22**

6.2 結果

評価は、自動評価の BLEU-4 と RIBES v1.01 で行った。評価結果を表 1 と表 2 に示す。提案手法は、プレオーダリング以外の手法（Pbmt，Hpbmt，Sbmt）より高いスコアが得られた。また、提案手法は、プレオーダリング手法である、原言語の構文解析器と木構造の一部を考慮する並べ替えルールを用いる Srcdep と構文を利用しない Lader よりも高いスコアが得られた。

7 まとめ

目的言語の構文解析器を用いたプレオーダリング手法を提案した。提案手法は同期率の高い構文構造を言語間の射影により獲得し、原言語の構文解析器を必要とせずに構文構造を利用したプレオーダリングを実現した。

参考文献

[1] David Chiang. Hierarchical phrase-based translation. *CL*, 33(2):201–228, 2007.
[2] Dmitriy Genzel. Automatically learning source-side reordering rules for large scale machine translation. *Coling* 2010.
[3] Isao Goto et al. Overview of the patent machine translation task at the NTCIR-10 workshop. *NTCIR-10*, 2013.
[4] Isao Goto et al. Overview of the patent machine translation task at the NTCIR-9 workshop. *NTCIR-9*, 2011.
[5] Isao Goto et al. Pre-ordering using a target language parser via cross-language syntactic projection for statistical machine translation. *ACM TALLIP*. Accepted for publication.
[6] Isao Goto et al. Distortion model based on word sequence labeling for statistical machine translation. *ACM TALIP*, 13(1):2, 2014.
[7] Hieu Hoang et al. A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based Statistical Machine Translation. *IWSLT* 2009.
[8] Rebecca Hwa et al. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3):311–325, 2005.
[9] Hideki Isozaki et al. HPSG-based preprocessing for English-to-Japanese translation. *ACM TALIP*, 11(3):8, 2012.
[10] Wenbin Jiang et al. Relaxed cross-lingual projection of constituent syntax. *EMNLP* 2011.
[11] Mark Johnson et al. Bayesian inference for PCFGs via Markov chain Monte Carlo. *NAACL* 2007.
[12] Philipp Koehn et al. Moses: Open source toolkit for statistical machine translation. *ACL* 2007.
[13] Graham Neubig et al. Inducing a discriminative parser to optimize machine translation reordering. *EMNLP* 2012.
[14] Slav Petrov et al. Learning accurate, compact, and interpretable tree annotation. *ACL* 2006.
[15] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Ann. Prob.*, 25(2), 1997.
[16] Yee Whye Teh. A bayesian interpretation of interpolated kneser-ney. NUS School of Computing Technical Report TRA2/06, 2006.
[17] Christoph Tillman. A unigram orientation model for statistical machine translation. *HLT-NAACL* 2004.
[18] Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *CL*, 23(3):377–403, 1997.
[19] Fei Xia and Michael McCord. Improving a statistical MT system with automatically learned rewrite patterns. *Coling* 2004.