統語的タグを用いた統計的階層句機械翻訳

藤原勇†　渡辺太郎‡　村上仁一†
†鳥取大学大学院　工学研究科
{s072046,murakami}@ike.tottori-u.ac.jp
‡情報通信研究機構
taro.watanabe@nict.go.jp

1　はじめに

近年，機械翻訳の分野において，統計翻訳が主流となっている．従来の統計翻訳では，翻訳過程において統語的情報を用いていない．そのため，日本語と英語のような言語構造が大きく異なる言語間の翻訳において高い翻訳精度が得られない場合が多い．この問題に対して，様々な研究が行われている．大西らは，文書レベルの文脈情報を用いてフレーズの並び替えを制限する手法を提案し，有意な成果を得た [1]．また，Zollmann らは統語的情報を用いた機械翻訳システムを提案し，公開している [2]．しかし，統語的情報を用いた機械翻訳の問題点として，統語ラベルの爆発的増加による，解析および翻訳時間の増加が挙げられる．そこで，本研究では，日英統計的階層句機械翻訳において，'浅い' 統語的情報を用いる手法を提案する．浅い統語的情報として，日本語文の名詞句と動詞句にタグを付与する．タグを付与した日本語文を用いて学習することで，統語的情報を含む文法規則が生成され，翻訳精度が向上すると考えられる．

2　統計的階層句機械翻訳

統計的階層句機械翻訳（以下，階層型翻訳）とは，階層構造を持った句を用いて翻訳を行う機械翻訳の手法である．図 1 に階層型翻訳の枠組みを示す．句に基づく統計翻訳と比較して，大局的な語順の並びを同期文脈自由文法によって表現できるという特徴が挙げられる．

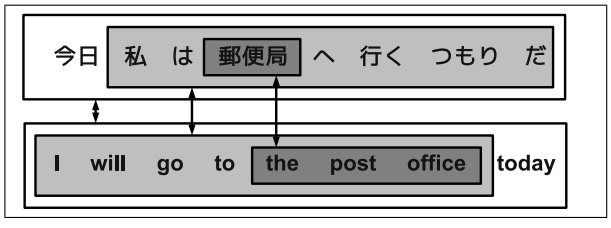

図 1　階層型翻訳の枠組

3　提案手法

階層型翻訳は，大局的な語順の並びを考慮することができる．しかし，翻訳過程において，統語的情報は用いられていない．そこで提案手法では，翻訳対象言語である日本語文の名詞句と動詞句にタグを付与することで，統語的情報を表現する．

以下に提案手法の概略を示す．

3.1　提案手法の概略

手順 1　GIZA++[3] を用いて，日本語-英語間の単語対応を得る

手順 2　日本語文を構文解析し，解析データを元に名詞句と動詞句にタグ (<NP>, </NP>, <VP>, </VP>) を付与する（表 1）

手順 3　タグが付与された部分の単語対応を修正する

手順 4　タグ付与後の日本語コーパスと，英語コーパスから文法規則を学習する

手順 5　タグが付与された日本語テスト文を翻訳し，翻訳精度を調査する

表 1　タグ付与の例

原文	そんな やり方 を すれ ば 必ず 苦情 が 起こ る 。
タグ付与後	<NP> そんな やり方 </NP> を すれ ば 必ず <VP> 苦情 が 起こ る </VP> 。

3.1.1　タグ付与の手順

名詞句のタグ付与手順を以下に示す．また，タグの付与例を表 1 に示す．

準備　日本語文を構文解析

手順 1　係り受け関係の連続しない後置詞句において，名詞を含み，動詞・助動詞を含まない部分を抽出

手順 2　係り受け関係の連続する後置詞句において，最後部の句が名詞を含み，動詞・助動詞を含まない部分を抽出

手順 3　手順 1，手順 2 で抽出したものから，末尾の助詞・記号を削除

手順 4　手順 3 で得られた名詞句の中から単語数が最大かつ単語数が 2 以上の句にタグを付与

4　実験環境

日本語文の構文解析，および分かち書きには Cabocha[4] を用いる．また，デコーダとして Joshua[5] を用いる．なお，Joshua の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を行う．また，自動評価法として BLEU[6] を用いる．

4.1　実験データ

本研究において，3 種類の異なるコーパスを用いた実験を行う．

4.1.1　単文

単文における実験データとして，単文コーパス [7] から学習文 100,000 文，ディベロップメント文 1,000 文，テスト文 10,000 文を用いる．単文の例を表 2 に示す．

表 2　単文の例

日本語文	政治は国民の生活に深い関係を持っている。
タグ付与後	政治は <NP> 国民の生活 < /NP> に <VP> 深い関係を持っている < /VP> 。
英語文	Politics have a deep bearing on the lives of the citizens.

4.1.2　重文・複文

重文・複文における実験データとして，重文・複文コーパス [8] から学習文 100,000 文，ディベロップメント文 1,000 文，テスト文 10,000 文を用いる．重文の例を表 3 に，複文の例を表 4 に示す．

表 3　重文の例

日本語文	校長先生が突然訪ねてきたので、母は即席で食事を用意した。
タグ付与後	<NP> 校長先生 < /NP> が <VP> 突然訪ねてきた < /VP> ので、母は即席で食事を用意した。
英語文	When the principal of our school paid a surprise visit to our house, my mother quickly set out some food for him.

表 4　複文の例

日本語文	家に帰る途中で食事を済ませた。
タグ付与後	<NP> 家に帰る途中 < /NP> で <VP> 食事を済ませた < /VP>。
英語文	I had a meal on my way home.

4.1.3　特許文

特許文における実験データとして，NTCIR-8 特許タスクのデータから，学習文 300,000 文，ディベロップメント文 2,000 文，テスト文 1,251 文を用いる．特許文の例を表 5 に示す．

表 5　特許文の例

日本語文	この挿入穴１７に上記プレート支持部材２３の他端部が挿入嵌合されている。
タグ付与後	この挿入穴１７に <NP> 上記プレート支持部材２３の他端部 </NP> が <VP> 挿入嵌合されている < /VP>。
英語文	Said other end part of the plate support member 23 is fitted into the insertion hole 17.

5　実験結果

5.1　自動評価

各実験データにおける自動評価結果を表 6 に示す．表中における baseline は，タグ付与を行わないコーパスを用いた翻訳結果である．

表 6　自動評価結果（BLEU）

	単文	重文・複文	特許文
baseline	10.19	7.41	26.05
提案手法	9.46	5.88	24.75

表 6 より，全ての実験において提案手法の有効性は認められなかった．

5.2　対比較調査

単文と重文・複文の実験において，baseline と提案手法の出力結果に対して人手による対比較調査を行った．評価は，出力結果からランダムに抽出した 20 文を用いる．単文における対比較調査結果を表 7 に示す．また，重文・複文における対比較調査結果を表 8 に示す．

表 7 単文における対比較調査結果

baseline ○	提案手法○	差なし	同一出力
0	0	14	6

表 8 重文・複文における対比較調査結果

baseline ○	提案手法○	差なし	同一出力
0	0	19	1

表 7，表 3 より，baseline と提案手法には，差が認められなかった．表 9 に単文における出力文例を示す．また，表 10 に重文・複文における出力文例を示す．

表 9 単文における出力文例

baseline 入力文	こちら へ おいで の 節 は ぜひ お 立ち寄り ください 。
提案手法入力文	こちら へ <NP> おいで の 節 < /NP> は <VP> ぜひ お 立ち寄り ください < /VP> 。
参照文	Please drop in on us if you happen to come this way .
baseline 出力文	When to the , I drop .
提案手法出力文	Please of of the to the , I drop .

表 10 重文・複文における出力文例

baseline 入力文	電車 に かさ を 忘れ た の を 思い出し 、 駅 に 取って返し た 。
提案手法入力文	電車 に <VP> かさ を 忘れ た の を 思い出し < /VP> 、 駅 に 取って返し た 。
参照文	I remembered I left my umbrella on the train , so I hurried back to the station .
baseline 出力文	I remembered my umbrella in the train , I to the station .
提案手法出力文	I forgot the umbrella in the train , to the station .

6 考察

実験結果より，提案手法において有効性は認められなかった．本章では以下の 5 つの視点から原因の考察を行う．

6.1 生成されたルールテーブルの考察

表 11 に，生成されたルールテーブルの例を示す．なお，表中の [X,N] は非終端記号を表している．

表 11 において，1 行目と 2 行目の例は正しく統語的情報が付与されているといえる．しかし，3 行目と 4 行目の例は明らかな誤りである．したがって，明らかに誤りのあるルールをフィルタリングすることによって翻訳精度の向上に繋がる可能性がある．

表 11 生成されたルールの例

<NP> 日本 の [X,1] < /NP> ||| The [X,1] in Japan
<VP> 新聞 に 載っ た < /VP> 。 ||| was printed in the newspaper .
< /NP> の <VP> ||| finding fault with
[X,1] < /NP> で <VP> ||| quietly by [X,1]

6.2 デコーダが考慮するフレーズ長の問題

Joshua のパラメータの 1 つである MaxPhraseSpan は，各終端記号における最大終端記号数を表している．しかし，タグを付与することにより，単語数が増加し，MaxPhraseSpan を上回る句が増加した．そのため，デコーダが考慮するフレーズが減少し，翻訳精度が減少したと考えられる．表 12 に MaxPhraseSpan を 10 から 14 まで増加させた場合における baseline と提案手法の自動評価結果を示す．

表 12 MaxPhraseSpan の増加による翻訳精度の変化

	10	11	12	13	14
baseline	26.05	26.33	26.60	26.63	26.71
提案手法	24.75	25.52	25.76	26.00	26.20

表 12 より，MaxPhraseSpan の増加に伴い，baseline と提案手法における差の減少が認められる．

6.3 動詞句タグの有効性

動詞句タグの有効性を調査するため，名詞句のみにタグを付与した場合と提案手法との比較を行った．自動評価における比較結果を表 13 に示す．なお，実験には単文を用いている．

表 13 動詞句タグの効果

名詞句タグのみ	9.21
提案手法（名詞句＋動詞句）	9.46

表 13 より，動詞句へのタグ付与は，名詞句のみへのタグ付与よりも有効であることが示された．

6.4 階層型翻訳におけるタグ付与の効果

先行研究と提案手法では，タグの付与方法が異なる．よって，提案手法のタグの付与方法の有効性を調査する

ため，句に基づく統計翻訳においてタグ付与実験を行った．句に基づく統計翻訳におけるデコーダは Moses デコーダ [9] を用いる．また，タグの付与は名詞句のみに行った．表 14 に実験結果を示す．なお，表中の baseline は通常のコーパスを用いた，句に基づく統計翻訳である．

表 14　句に基づく統計翻訳におけるタグ付与の有効性

baseline	11.01
タグ付与	12.21

表 14 より，句に基づく統計翻訳において，タグ付与手法が有効であることが示された．よって今回のタグ付与手法では，階層型翻訳においては有効ではないと考えられる．

6.5　統語的情報の不足

本研究では，統語的情報として，日本語文の名詞句と動詞句にタグを付与した．しかし，日本語文において，それぞれ最長の名詞句・動詞句のみに付与を行った．よって，複数の名詞句・動詞にタグ付与を行うことで統語的情報が増加することが考えられる．また，階層型翻訳において元言語である日本語文にタグを付与するだけでは，統語的情報としては不十分であったと考えられる．対象言語である英語にも何らかの統語的情報を付与することで，より構文的な情報を伴った文法規則が得られる可能性がある．

7　おわりに

本研究では，日英統計的階層句機械翻訳において，統語的情報として日本語文の名詞句・動詞句にタグを付与する手法を提案した．しかし，提案手法において有意な効果は認められなかった．原因として，日本語文へのタグ付与のみでは，統語的情報として不十分であることなどが挙げられる．今後の展開として，生成されたルールのフィルタリングを行う手法や，目的言語である英語側においても，統語的情報を付与する手法が考えられる．

参考文献

[1] Takashi Onishi, Masao Utiyama, Eiichiro Sumita: “Reordering Constraint Based on Document-Level Context” , The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011), pp.434-438, June 2011.

[2] Andreas Zollmann, Ashish Venugopal: “Syntax Augmented Machine Translation via Chart Parsing”, Proceedings of the Workshop on Statistical Machine Translation, pp.138-141, June 2006.

[3] Franz Josef Och, Hermann Ney: ”Improved Statistical Alignment Models”, the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000), pp.440-447, October 2000.

[4] Taku Kudo, Yuji Matsumoto: “Japanese Dependency Analysis using Cascaded Chunking”, CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops, pp.63-69, 2002.

[5] Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese and Omar Zaidan: “Joshua: An Open Source Toolkit for Parsing-based Machine Translation”, In Proceedings of the Workshop on Statistical Machine Translation (WMT09), pp.135-139, March 2009.

[6] Papineni Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu : “BLEU: a method for automatic evaluation of machine translation”, 40th Annual meeting of the Association for Computational Linguistics, pp. 311-318, 2002.

[7] 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, A2-8, pp.372-375, 2005.

[8] 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.

[9] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation”, Proceedings of the ACL 2007 Demo and Poster Sessions, pp.177-180, June 2007.