Tree-to-String翻訳における構文解析器の自己学習の効果

波多腰 優斗　Graham Neubig　Sakriani Sakti　戸田 智基　中村 哲
奈良先端奈良先端科学技術大学院大学 情報科学研究科
{hatakoshi.yuto.hq8,neubig,ssakti,tomoki,s-nakamura}@is.naist.jp

1　はじめに

統計的機械翻訳では，複数の単語からなるフレーズ間の翻訳確率を計算し，目的言語として適切な語順となるように並べ替えモデルによる局所的なフレーズの移動を行う，フレーズベース機械翻訳 (phrase-based machine translation, PBMT)[8] が広く用いられている．PBMT は翻訳モデルの学習が容易であり，多くの言語対で高い精度での翻訳が可能である一方，形態的・統語的な情報の扱いに乏しいという問題点もある．また，並べ替えモデルによる長距離の並べ替え確率の推定が困難であり，英語と日本語のように語順が大きく異なる言語対では翻訳精度が低下することが知られている．

これに対して，翻訳パターンに構文木の部分木の構造を用いる統語ベース翻訳 [24] は，文法構造が大きく異なる言語対においてPBMT よりも正確な翻訳を実現することが多い．統語ベース翻訳の中でも原言語側のみの構文情報を用いた，同期木置換文法 (STSG)[3] に基づく T2S(Tree-to-String) 翻訳 [9] は，語彙選択と並べ替えを同時に行う翻訳パターンを利用することで，語順が大きく異なる言語対においても高精度かつ高速な訳出が可能である．一方で，構文解析の結果を用いて翻訳を行うため，訳出の精度は構文解析器の精度に大きく依存する．英日・日英翻訳の実験において，より精度の高い解析モデルを用いることで，T2S の翻訳精度が向上することが報告されており [16]，構文解析器の精度は重要な要素の一つとなっている．

構文解析の精度を高めるための単純かつ有用な手法として，自己学習 (self-training) が挙げられる [11]．自己学習では，はじめに既存のモデルにより構文解析を行い，その結果を用いてモデルの再学習を行う．再学習したモデルにより構文解析を行うことで，解析対象のデータへの適応を補助する効果が得られ，解析精度が向上する．McClosky らは，自己学習により既知語の組み合わせによる未知の bi-gram の解析精度が向上したことから，既知語が通常と異なる文脈で現れるときに有効であると報告している [12]．このような構文解析器の自己学習は，構文情報を用いて翻訳を行う統語ベース翻訳の精度向上にも寄与すると考えられるが，その詳細な効果については検証されていない．

そこで本研究では，T2S 翻訳における構文解析器の自己学習の効果について調査を行う．実験では英日・日英翻訳を対象とし，2 つの専門分野 (特許・Wikipedia 記事) への適応を行う．実験の結果，構文解析器の自己学習により T2S の精度が大きく向上することが確認された．

2　Tree-to-String翻訳

T2S 翻訳は原言語文の構文解析結果を利用することで，2 言語間の関係を統語的な構造により捉えることができ，正確な翻訳が可能となる．翻訳パターンは PBMT のように単語列ではなく，変数 x を含むルールも利用して原言語文の部分木の構造として表現される．以下の例は，置き換え可能な 2 つの NP を翻訳パターンに直接含んでおり，x0，x1 に当てはまる候補の確率と翻訳パターン自体の確率を考慮して訳文を生成する．

S (x0:NP) (VP (VBD hit) (x1:NP)))
→ x0 は x1 を打つ

原言語文の部分木を用いることで訳出候補が少なくなるメリットもあり，探索空間が小さくなることで翻訳に必要な時間も短縮される．一方で，構文解析の結果を基に翻訳パターンを生成するため，翻訳精度が構文解析器の精度に依存してしまうというデメリットもある．

このような問題点を緩和する手法として，複数の構文木の解釈を一つの超グラフ構造で保持した構文森を入力として用いる F2S(Forest-to-String) 翻訳 [13] がある．これにより，いくつかの構文木の解釈の中から精度の高い訳出を行うための構文解析結果を選択することが可能となる [25]．

3 構文解析における自己学習

構文解析における自己学習に関する実験は，Charniak により初めて報告されており，WSJ コーパス [10] により学習された確率文脈自由文法 (probabilistic context-free grammar, PCFG) に対して自己学習を行ったところ精度の改善は見られなかったとしている [1]．また，語彙化木接合文法 (lexicalized tree adjoining grammar, LTAG) および語彙化 PCFG においては，構文解析モデルの学習データが小さいときに限り精度が改善されるものの，十分な量のデータにより学習されたモデルに対しては自己学習の効果が得られない [21]．これに対して，McClosky らはラベルなしのデータを効果的に利用することで，語彙化 PCFG の精度を自己学習により改善している [11]．この手法では，構文解析結果の n-best のリランキングを行い，その結果を自己学習に用いている．

一方で，PCFG-LA (PCFG with Latent Annotations) は自己学習によって解析精度が大きく向上するモデルとして知られている [4]．PCFG-LA の解析精度は，語彙化 PCFG と比較して自己学習による改善の幅が大きく，大量のデータによって学習されたモデルに対する効果も示されている．この理由として，PCFG-LA が高精度なモデルであるため自己学習に用いる構文木の精度が高いことや，EM アルゴリズムにより，正解木と自動で生成した構文木の両方から複雑な文法規則を獲得できることを挙げている．本研究では，これを踏まえて PCFG-LA に対して自己学習を行い，統語ベース翻訳の精度に及ぼす影響を調査する．

4 統計的機械翻訳における自己学習

前節で述べた自己学習による構文解析精度の向上と，先行研究 [16] により報告されている，構文解析精度が統語ベース翻訳に及ぼす影響を考慮すると，構文解析器の自己学習により翻訳精度が向上する可能性は十分にあると考えられる．本研究では，この効果を定量化することを主な目的とする．

統語ベース翻訳における前例は筆者の知る限り存在しないが，関連深い先行研究として，事前並べ替え [23] に用いる構文解析器に対して自己学習を行い，PBMT の翻訳精度を向上させた研究が挙げられる [6]．この研究では，通常の自己学習に加えて標的自己学習 (targeted self-training) の枠組みを取り入れており，複数の構文解析候補の中から事前並べ替えの精度が最も高くなるものを選択し，構文解析器を再学習を行っている．報告によると，自己学習は一定の効果を示し，標的自己学習により更なる向上が実現可能である．本研究では通常の自己学習に焦点を合わせ，同様の効果が統語ベース翻訳に対しても得られるか調査を行う．

表 1: KFTT・NTCIR7 のデータ内訳

KFTT	Words (En)	Words (Ja)	Sentences
train	10.6M	10.9M	430k
dev	24.3k	26.8k	1.17k
test	26.7k	28.3k	1.16k
NTCIR	Words (En)	Words (Ja)	Sentences
train	99.0M	117M	3.08M
dev	28.6k	33.5k	0.82k
test	44.3k	52.4k	1.38k

5 実験的評価

5.1 実験条件

実験は英日・日英翻訳を対象とし，京都フリー翻訳タスク (KFTT)[1]，および NTCIR-7 特許機械翻訳テストコレクション [2] の英日翻訳データを用いた．実験データに関して，学習データ (train)，開発データ (dev)，テストデータ (test) の詳細を表 1 に示す．

T2S は Travatar [15] [2] に実装されているものをデフォルトの設定で用いた．単語アライメントを取るツールとして Nile [20] [3] を用い，目的言語である日本語の言語モデルは SRILM[22] を用いて 5-gram で学習した．実験における翻訳精度は，BLEU[18] と RIBES[5] の 2 つの自動評価尺度を用いて測った．各素性の重みは BLEU が最大となるように MERT[17] を用いて最適化した．日本語の単語分割には KyTea [4] を用いた．

構文解析は PCFG-LA モデル [19] を用いた Ckylark [26] [5] により行った．英語および日本語の構文解析には，WSJ コーパス [10]，JDC コーパス [14] によりそれぞれ学習されたモデルを用いた．JDC は係り受けコーパスであるため，Travatar のツールキットに含まれる主辞ルールを利用して PCFG-LA の学習に必要な形式に変換した [6]．構文解析器の自己学習は，WSJ/JDC のモデルにより解析された KFTT と NTCIR の train データのみを用いて行った．自己学習に用いるデータは train データの上位 5k∼50k まで変化させ，各条件における翻訳精度を評価した．

5.2 実験結果

実験結果を表 2 に示す．表中の太字は，ブートストラップ・リサンプリング法 [7] を用いてベースライン

[1] http://www.phontron.com/kftt/index-ja.html
[2] http://www.phontron.com/travatar/
[3] https://code.google.com/p/nile/
[4] http://www.phontron.com/kytea/index-ja.html
[5] http://odaemon.com/?page=tools_ckylark
[6] ja-adjust-dep.pl, ja-dep2cfg.pl

表 2: 自己学習のデータ量と英日・日英翻訳の精度

	KFTT(En-Ja)		NTCIR(En-Ja)		KFTT(Ja-En)	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
WSJ/JDC (Baseline)	21.91	71.63	39.73	79.16	19.91	67.74
self-training sentences (k)						
5	21.53	70.94	32.36	76.39	20.37	68.04
15	22.05	71.97	**41.20**	**79.80**	20.30	68.16
20	**22.53**	**72.24**	**41.14**	**80.50**	**20.49**	68.14
25	22.14	71.55	36.01	77.78	**20.40**	**68.82**
30	22.23	71.57	**41.19**	**80.32**	**20.48**	67.90
35	**22.34**	72.04	**41.12**	**80.10**	**20.51**	**68.59**
40	22.17	71.89	**40.99**	**80.19**	**20.63**	**68.58**
45	**22.70**	**72.28**	**41.51**	**80.23**	**20.86**	68.23
50	22.37	**72.26**	**40.96**	**80.43**	**20.59**	**68.67**

表 4: 構文解析器の自己学習による英日翻訳の訳出の改善例

source	··· , comprises a plurality of male contacts 98 aligned in a row in an electrically insulative housing 97 as shown in the figure .
Reference	···、図示のように、絶縁ハウジング９７内に雄コンタクト９８を整列保持して構成される。
Baseline	··· 、複数の雄型コンタクト列に整列されて同図に示すように、電気絶縁性のハウジング９７９８から構成されている。
self-training 45k	···、図に示すように、絶縁性のハウジング９７に一列に配列された複数の雄型コンタクト９８から構成されている。

表 3: NTCIR(En-Ja):45k のテスト文に対する各条件の構文木の精度

self-training sentences (k)	20	25	30
F-measure	85.64	74.83	88.57

と比較した結果，統計的有意性が認められた数値を示している ($p < 0.05$)．

実験の結果，各翻訳タスクにおいて構文解析器の自己学習により BLEU と RIBES が向上した．それぞれの実験条件における BLEU は最大で，KFTT(En-Ja):0.79, NTCIR(En-Ja):1.68，KFTT(Ja-En):0.95 ポイント上昇した (Baseline vs self-training 45k)．

一方で，翻訳精度が改善されない条件も多く見られた (ex.NTCIR:25k, KFTT:25k-30k-40k-50k)．表 3 は NTCIR(En-Ja) において最も BLEU 値が高かった self-training 45k のテスト文を正解木としたときの各条件における構文木の精度を示している (Evalb[7] により算出)．翻訳精度が低かった self-training 25k の条件において構文木の F 値も低くなっており，自己学習による効果が得られていないことが分かる．本研究では，train データにおける上位の行をそのまま構文解析器の再学習に用いており，解析精度の向上に寄与するデータを効率的に抽出できていない．先行研究 [4] のように構文解析器の再学習に用いるデータを選択し，効果的な自己学習を行うことで，学習に用いるデータ量に対して安定した結果が得られるようになると考えられる．

5.3 自己学習の効果

表 4 に英日翻訳の訳出の改善例を示す．構文解析器の自己学習によって表中の下線部のような名詞句の訳出の改善が多く確認された．図 1 は表 4 の原言語文の構文解析結果である．Baseline における構文木は名詞句 “male contacts 98” が 2 つの名詞句に分解されてしまっているが (図 1(a))，自己学習により 1 つのスパンの名詞句として正しく解析された (図 1(b))．これは先行研究 [12] で報告されていたように，WSJ と NTCIR において異なる文脈で現れていた既知語の解析精度が向上した結果であると考えられる．このような構文解析精度の向上によって翻訳モデルの構築時により良い翻訳規則が学習され，翻訳精度が改善されたと言える．

6 おわりに

本研究では，T2S 翻訳における構文解析器の自己学習の効果について検証を行った．自己学習の効果が特に大きいとされる PCFG-LA モデルによる構文解析器を用い，2 つの翻訳タスクにおいて英日・日英翻訳を行った．構文解析器の自己学習の効果は T2S 翻訳の結果にも反映され，訳出の精度が向上することが確認

[7] http://nlp.cs.nyu.edu/evalb/

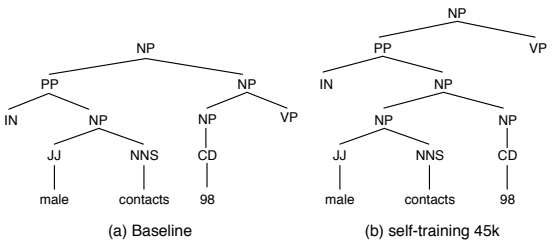

図 1: 自己学習による構文解析の改善

された．今後の課題として，他の言語対や F2S 翻訳に対する効果の検証を予定している．また，先行研究では標的自己学習の効果が示されており，統語ベース翻訳に対してこのような枠組みを用いることも考えていきたい．

謝辞

本研究の一部は，JSPS 科研費 25730136 の助成を受け実施した．

参考文献

[1] E. Charniak. Statistical parsing with a context-free grammar and word statistics. *AAAI/IAAI*, 2005:598–603, 1997.

[2] A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, T. Ehara, H. Echizen-ya, and S. Shimohata. Overview of the patent translation task at the NTCIR-7 workshop. In *Proceedings of the 7th NTCIR Workshop Meeting*, pp. 389–400, 2008.

[3] J. Graehl, K. Knight, and J. May. Training tree transducers. *Computational Linguistics*, pp. 391–427, 2008.

[4] Z. Huang and M. Harper. Self-training PCFG grammars with latent annotations across languages. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 832–841, 2009.

[5] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 944–952, 2010.

[6] J. Katz-Brown, S. Petrov, R. McDonald, F. Och, D. Talbot, H. Ichikawa, M. Seno, and H. Kazawa. Training a parser for machine translation reordering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–192, 2011.

[7] P. Koehn. Statistical significance tests for machine translation evaluation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 388–395, 2004.

[8] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *North American Chapter of the Association for Computational Linguistics*, pp. 48–54, 2003.

[9] Y. Liu, Q. Liu, and S. Lin. Tree-to-string alignment template for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 609–616, 2006.

[10] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of English: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[11] D. McClosky, E. Charniak, and M. Johnson. Effective self-training for parsing. In *North American Chapter of the Association for Computational Linguistics*, pp. 152–159, 2006.

[12] D. McClosky, E. Charniak, and M. Johnson. When is self-training effective for parsing? In *International Conference on Computational Linguistics (COLING)*, pp. 561–568, 2008.

[13] H. Mi, L. Huang, and Q. Liu. Forest-based translation. In *ACL*, pp. 192–199, 2008.

[14] S. Mori, H. Ogura, and T. Sasada. A Japanese word dependency corpus. In *Proceedings of the Nineth International Conference on Language Resources and Evaluation*, 2014.

[15] G. Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. *Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 91, 2013.

[16] G. Neubig and K. Duh. On the elements of an accurate tree-to-string machine translation system. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 143–149, 2014.

[17] F. J. Och. Minimum error rate training in statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 160–167, 2003.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

[19] S. Petrov and D. Klein. Improved inference for unlexicalized parsing. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 404–411, 2007.

[20] J. Riesa and D. Marcu. Hierarchical search for word alignment. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 157–166, 2010.

[21] M. Steedman, M. Osborne, A. Sarkar, S. Clark, R. Hwa, J. Hockenmaier, P. Ruhlen, S. Baker, and J. Crim. Bootstrapping statistical parsers from small datasets. In *European Chapter of the Association for Computational Linguistics (EACL)*, pp. 331–338, 2003.

[22] A. Stolcke, J. Zheng, W. Wang, and V. Abrash. SRILM at sixteen: Update and outlook. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 5, 2011.

[23] F. Xia and M. McCord. Improving a statistical mt system with automatically learned rewrite patterns. In *International Conference on Computational Linguistics (COLING)*, p. 508, 2004.

[24] K. Yamada and K. Knight. A syntax-based statistical translation model. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 523–530, 2001.

[25] H. Zhang and D. Chiang. An exploration of forest-to-string translation: does translation help or hurt parsing? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 317–321, 2012.

[26] 小田, G. Neubig, S. Sakti, 戸田, 中村. 解析失敗の発生しにくい PCFG-LA 句構造構文解析. 言語処理学会第 21 回年次大会 (NLP2015), 2015.