構文情報を利用した対訳データ選択手法

丹生 伊左夫　Graham Neubig　Sakriani Sakti　戸田 智基　中村 哲
奈良先端科学技術大学院大学 情報科学研究科
{isao-ni, neubig, ssakti, tomoki, s-nakamura}@is.naist.jp

1　はじめに

統計的機械翻訳 (Statistical machine translation; SMT) の主な翻訳方式は，対象の言語間で得られる対応関係から翻訳モデルを学習し，目的言語側の言語モデルなど様々な統計モデルを用いることで，入力される原言語文を自動的に翻訳先の目的言語文に変換するというものである．SMT において，翻訳の精度は学習に用いる対訳データの量に大きく依存することが報告されている [1]．しかし，大規模な対訳データを用いて翻訳モデルを学習する際，大量の時間を要するという問題がある．これは，SMT を研究または開発する上で，効率を下げる要因となる．また，大規模なデータで学習した統計モデルはサイズも大規模になるため，翻訳システムを搭載するデバイスにサイズの制限がある場合は，統計モデルのサイズを小規模化させる必要もある．

学習に用いる対訳データを選択的に小規模化させることで，学習時間の短縮とモデルサイズの縮小に取り組む研究がなされている [2, 3, 4]．Bilingual Sentence Selection (BSS) タスクと呼ばれ，SMT システムを学習させるために利用できる全文対から，最適な文対を選択してくる問題とされている．

Gascó らは n-gram に着目して対訳データを選択する手法 [4] を提案している．学習データ中で低頻度な n-gram を多く含む文を，対訳データから選択して学習データに追加していく方法である．そして，SMT の代表的な翻訳方式であるフレーズベース機械翻訳 (Phrase-based machine translation; PBMT)[5] を用いた英語から仏語への翻訳で，その効果を報告している [4]．しかし，n-gram という局所的な単語間の関係性だけを考慮している手法なため，構文情報を利用した翻訳方式に対しては構文解析誤りや不当な翻訳規則の学習を招く可能性もある．さらに，英語と日本語のように語順が大きく異なる言語間では，n-gram という局所的な情報だけでは語同士の関係性を捉えることはさらに困難である．

そこで本稿では，英語と日本語のように語順の大きく異なる言語間に対して，翻訳精度の維持と翻訳モデルの学習時間の短縮を図って構文構造を利用した対訳データ選択手法を提案する．特に，原言語側の構文情報を利用して対訳データを選択する方式である．結果として，方式の大きく異なる 2 つの翻訳方式で実験し，特に構文情報を用いた翻訳方式に対して提案法は Gascó らの選択方法やランダムな選択よりも，翻訳精度の面で効果的に機能することを確認した．さらに，それに伴う翻訳モデルの学習時間やサイズの評価と，選択したデータを分析したので報告する．

2　対訳データ選択

対訳データの選択に関する先行研究として，異なるサブデータそれぞれに対して重みを付与することで学習に用いるデータを選択する手法 [3] や，翻訳される文に従って重み付けされるよう異なるサブモデルを生成し，このモデルを利用した対訳データ選択手法 [6] などが提案されている．ただし，特定のドメインにおける実験だけで提案法の効果を示している．つまり，実環境としてどのような場面で用いられるか特定できない場合や，様々なドメインで構成される学習データの場合，性能が向上するとは一概に言い切れない．

2.1　n-gram に基づく対訳データ選択手法

Gascó らはオープンドメインなデータに対する対訳データ選択手法 [4] を提案し，その効果を示している．具体的には，Gascó らの低頻度 n-gram 補完 (Infrequent n-gram recovery) は，学習データにあまり含まれていない n-gram を多く含む文を選択する手法である．学習データの情報量を向上させる文を選択することを目的としている．また，PBMT は語の系列を利用して翻訳モデルを学習するため，様々な n-gram を含む学習データを得ることは，翻訳規則の多様性を向上させることに繋がる．

対象の対訳データの各文対にスコアを付与し，値の高い文対から順に選択していくことで学習データを得る．各文対へのスコア $\mathrm{i}(f)$ は式 (1) で計算される．

$$\mathrm{i}(f) = \sum_{w \in \mathrm{X}_d(f)} \min(1, \mathrm{N}(w)) \max(0, t - \mathrm{C}(w)) \quad (1)$$

ここで，$\mathrm{X}_d(f)$ はスコア付与対象の原言語文 f に含まれる長さ d 以下の n-gram 集合で，w は $\mathrm{X}_d(f)$ の一要素である．そして，$\mathrm{N}(w)$ は f における w の出現回数で，$\mathrm{C}(w)$ は学習データにおける w の出現回数である．t は学習データ中に含まれる n-gram の出現回数の閾値を表す．式 (1) の計算方法によって対訳データの各文対にスコアを付与し，スコアが最も高い文対を学習データに追加する．そして，学習データ中に含まれる n-gram の情報を更新し，式 (1) によって残りの対訳データのスコアが再計算される．この処理を選択する文対の数に応じて繰り返すことで，対訳データから学習データを選択的に得られる．

2.2　低頻度 n-gram 補完の問題点

Gascó らの低頻度 n-gram 補完は，学習データに含まれない n-gram が多く含まれる文に高いスコアを付与するため，含まれる単語数が多い文ほどスコアが高くなる確率も上がり，結果的に長い文が選ばれやすい

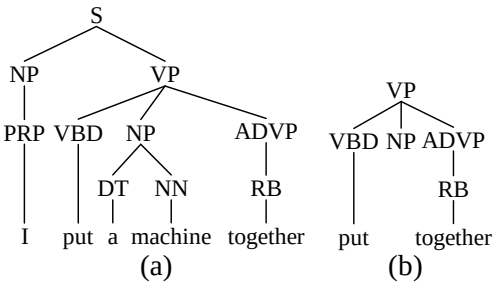

図 1: 構文木 (a) と部分木 (b) の例

傾向にある．長い文ばかり選択されてしまうと，単語アライメントなどの後段の処理に悪影響を及ぼす可能性もある．我々は予備実験でこのような問題を確認したため，文 f に対して計算されるスコアは文に含まれる単語数に依存することから，式 (1) を文長 $\mathrm{X}_1(f)$ で正規化した式 (2) を新たに定義する．

$$\mathrm{i}(f) = \frac{\sum_{w \in \mathrm{X}_d(f)} \min(1, \mathrm{N}(w)) \max(0, t - \mathrm{C}(w))}{\mathrm{X}_1(f)} \quad (2)$$

式 (2) でスコアを計算することで，長い文に不当に高いスコアが付与されることを防ぐことが可能になる．

しかし，このように補正しても依然として課題が残る．Gascó らは翻訳方式として PBMT に対する効果のみを示しており，統語ベース機械翻訳に対しても効果的であるとは言い切れない．統語翻訳に有用でありながら n-gram で捉えられない情報の一例を図 1(a) に示す．この場合，Gascó らが提案するように n-gram を 3-gram まで拡張しても，“put together” という高頻度な句動詞の関係性が考慮されていない．

3 構文構造に基づく対訳データ選択

構文情報を利用した Tree-to-string (T2S) 翻訳や Forest-to-string (F2S) 翻訳があり，文法が大きく異なる言語間では高い精度を誇っている [7]．これらの手法は学習に用いる対訳データの原言語文に含まれない構造を翻訳規則として学習できない．つまり，様々な構造を含んだ学習データを用いることで，翻訳モデルも多様な規則を学習できることになる．そこで，多様な構造を利用するために対象の原言語文の構文情報に着目する．特に，構文構造内の局所的な関係性から大域的な関係性まで捉えるために，構文木内の内部ノードの異なる部分木を n-gram の代わりに利用する．

図 1 に示す構文木ついて説明する．Gascó らの手法は離れた場所に位置する単語間の関係を捉えることが困難であったが，図 1 の (a) から内部ノード数 4 の部分木に着目すると，(b) に示す構造を抽出できる．つまり，対象の文から内部ノード数の異なる集合を順に取り出していくことで，構文木内の多様な構文構造を捉えることが可能になる．内部ノード数を順に拡張していくと，最終的に構文木全体を表現する構造を抽出できることになる．そして，T2S 翻訳などの翻訳手法では “put [NP] together” という翻訳規則を学習できることに繋がる．

そして，これらの部分木の集合を利用して，対訳データ中の各文対に付与するスコアを計算する．式 (1) に基づいて，対象の文 f に付与するスコア $\mathrm{s}(f)$ の計算は式 (3) で求める．

$$\mathrm{s}(f) = \frac{\sum_{x \in \mathrm{T}_d(f)} \min(1, \mathrm{N}(x)) \max(0, t - \mathrm{C}(x))}{|\mathrm{X}_1(f)| + |\mathrm{T}_1(f)|} \quad (3)$$

ここで，$\mathrm{T}_d(f)$ は文 f に含まれる内部ノード数 d 以下の全ての部分木の集合を表し，x は $\mathrm{T}_d(f)$ の一要素である．また，$|\mathrm{X}_1(f)|$ は原言語文 f の単語数を表し，$|\mathrm{T}_1(f)|$ は文 f 中の内部ノード数 1 の要素数を表す．対象の原言語文の構文解析結果に，同じ構造が多く含まれることによって不当に高いスコアが付与されることを防ぐために，各部分木は対象文中で一回のみ出現を許す．以上の提案法を適用した対訳データ選択処理の流れを図 2 に示す．

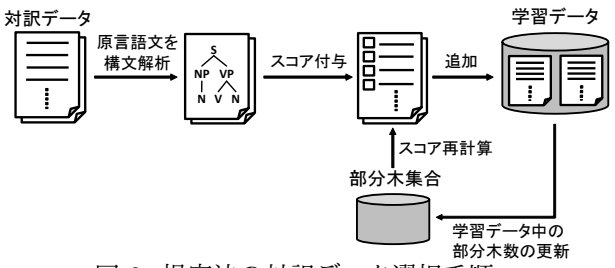

図 2: 提案法の対訳データ選択手順

4 実験

対訳データ選択手法を適用した学習データを用いて，方式の異なる翻訳方式それぞれに対して実験した．

4.1 実験条件

実験に用いた対訳データは特許コーパス NTCIR-7,8 の PatentMT[8, 9] で約 3M 文対，チューニングデータとテストデータは NTCIR-7 の PatentMT の約 1k 文対を利用して，英語から日本語への翻訳を実験した．翻訳方式として，Moses[10] に実装されているフレーズベース機械翻訳 (PBMT) と，Travatar[11] に実装されている複数の構文木から単語列へと変換する F2S 翻訳を実験に用いた．上記それぞれの翻訳方式に対して，従来法である Gascó らの低頻度 n-gram 補完 (Baseline) と，ランダムな対訳データ選択 (Random)，そして提案法 (Proposed) による選択手法をそれぞれ適用したデータで学習した．従来法のスコア計算方法に式 (1) を適用した場合，翻訳精度が低下し，学習時間を大幅に要したため，式 (2) を用いて閾値 t は 1 で実験した．そして，提案法のスコア計算式 (3) に関して，閾値 t を 1 に固定し，内部ノード数 d は実験的に 5 に設定した．ただし，学習データについて，100% を全対訳データとして，50%，25%，12.5%，6.3%，3.1%，1.6% の順に小規模化させて実験した．

英語側の単語分割には Stanford Parser[12] を，そして日本語側の単語分割には KyTea[13] を利用した．また，原言語である英語の文に対して用いる構文解析器は Egret[14] を利用した．そして，各手法において対訳データ間のアライメントを取るツールとして，GIZA++[15] を利用し，目的言語である日本語の言語モデルは IRSTLM を用いて 5-gram で学習し，各素性

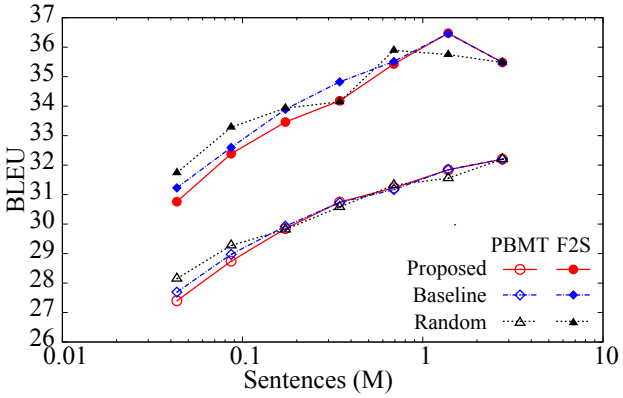

図 3: 各選択手法を適用した場合の BLEU

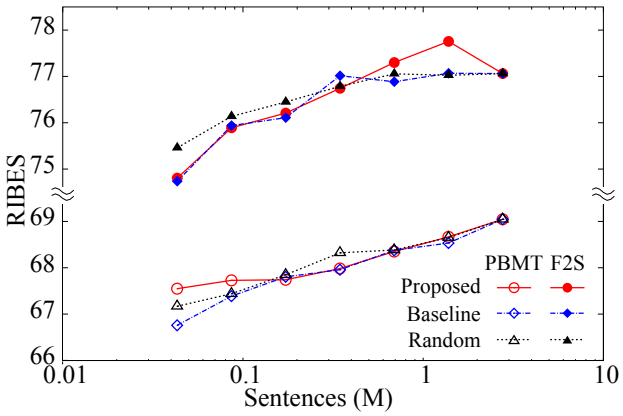

図 4: 各選択手法を適用した場合の RIBES

の重みは BLEU が最大となるように MERT[16] を用いて調整した．評価した項目は，翻訳方式と対訳データ選択手法それぞれを組み合わせた場合の翻訳精度と，翻訳モデルの学習時間，そして，学習した翻訳モデルのサイズである．翻訳精度の評価には，BLEU[17] と RIBES[18] の 2 つの自動翻訳評価尺度を用いた．ただし，MERT を用いたチューニングにはランダム性があるため，[19] に従って 3 度のチューニングを行い，BLEU と RIBES はその平均値を算出した．

4.2 翻訳性能と学習時間

翻訳方式に対する対訳データ選択手法の関係性として，翻訳性能と学習時間の実験結果について述べる．まず翻訳性能に関して，PBMT と F2S に対して各選択手法を適用した学習データ量別の BLEU を図 3 に示す．BLEU で測った場合，両翻訳方式において対訳データ選択手法別の翻訳精度に有意な差はほとんどなかった．しかし，ランダム選択とは異なり，提案法と従来法を F2S 翻訳に利用した場合，全対訳データを用いるより 50% のデータで精度の向上を確認した．さらに，RIBES で評価した結果を図 4 に示す．各学習データ量毎に近い結果となっている部分も存在するが，50% に小規模化したデータで学習した F2S の翻訳精度は，提案法が有意な差で高い結果となった．つまり，提案法で選択した 50% の学習データを F2S に利用した場合，両翻訳評価尺度で精度の向上を確認した．しかし，小規模なデータで学習した場合，提案法と従来法はランダム選択より若干低い精度となったが，どの

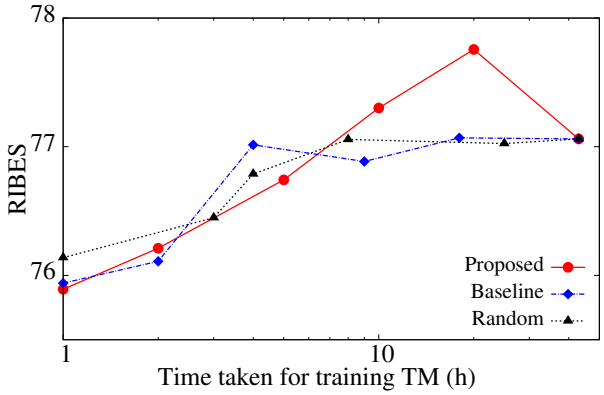

図 5: F2S の翻訳モデル学習時間と RIBES の関係

表 1: 学習データのサイズ別平均文長

文対数 (%)	言語	平均文長 提案法	従来法	ランダム
2.8M (100%)	En	29.3		
	Ja	34.7		
1.4M (50%)	En	32.5	29.6	29.3
	Ja	37.8	35.9	34.6
691k (25%)	En	33.6	28.5	29.3
	Ja	38.9	35.6	34.6

手法も翻訳精度が大幅に低下していることから現実的でない設定であると言える．

次に，各対訳データ選択手法を適用した学習データで実験した場合の，翻訳モデルの学習時間と翻訳精度の関係について記す．F2S の翻訳モデルの学習に要した時間と，翻訳精度として RIBES を用いた場合の関係を図 5 に示す．縦軸が翻訳精度 RIBES を表し，横軸は翻訳モデルの学習時間を対数目盛で表している．図 5 から，提案法を用いて選択したデータで学習した場合，他の選択手法と同程度の学習時間でも，50% の比較的大きいデータで学習した F2S は高い翻訳精度であるとわかる．また，PBMT の場合，従来法で選択したデータは全体的に他の選択手法よりわずかに短い学習時間であったため，従来法は PBMT と相性が良いと言える．

4.3 対訳データ選択の結果と分析

対訳データ選択手法によって 50% と 25% に小規模化された学習データの詳細を表 1 に示す．選択されたデータに関して，提案法は従来法やランダム選択よりも数単語多く含む傾向にあった．また，式 (2) を用いた従来法は，ランダム選択と同じか数単語少ない平均文長であった．

また，選択された学習データの n-gram と部分木に着目し，テストデータに対するカバー率を調べた．50% と 25% に小規模化された学習データのカバー率を表 2 に示す．学習データのサイズ毎に計算したが，n-gram に関しては基本的に従来法が高いカバー率となった．1,2,3-gram のカバー率を評価し，従来法は提案法やランダム選択よりも数% の差をつけて高い値となった．そして，部分木のカバー率に関しては，提案法が従来法やランダム選択より高いカバー率となった．これらの結果から，n-gram に関してはテストデータの

表 2: テストデータに対する学習データのカバー率

データ	評価項目	提案法	従来法	ランダム
100%	n-gram	83.4%		
	部分木	83.1%		
50%	n-gram	80.2%	81.4%	79.8%
	部分木	81.0%	80.8%	79.8%
25%	n-gram	76.1%	77.5%	76.4%
	部分木	77.8%	77.3%	76.9%

表 3: 対訳データ選択手法別の翻訳モデルサイズ

文対数 (%)	提案法	従来法	ランダム
2.8M (100.0%)	1600MB		
1.4M (50.0%)	906MB	985MB	818MB
691k (25.0%)	528MB	568MB	490MB
346k (12.5%)	301MB	317MB	267MB
173k (6.3%)	170MB	173MB	151MB
86k (3.1%)	96MB	94MB	82MB
43k (1.6%)	53MB	50MB	44MB

カバー率が高いデータで PBMT を学習してもあまり翻訳精度に効果は期待できないが，部分木のカバー率が高い提案法を用いた選択は F2S で学習すると翻訳規則がテストデータを多くカバーすることに繋がり，RIBES で測る翻訳精度を向上させることに寄与していることがわかる．

4.4 翻訳モデルサイズ

それぞれ異なる対訳データ選択手法を適用したデータで，F2S によって学習された翻訳モデルのサイズを表 3 に示す．結果として，ランダム選択，提案法，従来法の順でモデルが大きくなる傾向にあったが，提案法は従来法に比べて選択されたデータの平均文長が長く，含まれる単語数も多かったにも関わらず，学習データが大きくなるにつれて従来法より小さい翻訳モデルを得られる傾向にあった．PBMT で学習される翻訳モデルの場合は，従来法と提案法は同じ程度のサイズで，わずかにランダム選択が小さい結果となった．

5 おわりに

大量の対訳データを利用して SMT システムを構築する場合，膨大な学習時間を要すると共にモデルサイズも大規模になるという問題を受けて，翻訳モデルの学習に用いる対訳データの選択手法を提案した．そこで，日本語と英語のような語順の大きく異なる言語間で大域的な単語間の関係性を捉えるために，構文情報を利用した．特に，対象の対訳データのうち原言語側の構文情報を用いて対訳データを選択するために，内部ノード数の異なる部分木に着目した．結果として，PBMT と F2S の両翻訳手法において，提案法を適用し 50% に小規模化させた対訳データで学習しても，全対訳データを利用するより翻訳精度が向上することを確認した．特に，F2S で提案法は従来の方法やランダムに文対を選択するより高い RIBES の翻訳精度を得られた．さらに，学習時間やモデルサイズも他の選択手法と同程度の結果を得られた．結論として，構文構造を利用することは，翻訳性能に大きく寄与するだけでなく，対訳データ選択でも有益な情報を捉えることが可能になると言える．

今後の課題として，提案法は構文解析の結果を利用しているため，構文解析誤りに頑健な選択手法が必要である．また，対訳データを小規模化させることで，PBMT と F2S の両翻訳手法で翻訳精度の向上を確認できたため，どの程度まで小規模化させると翻訳精度を最大化できるのかについての検討と，精度が向上した原因についてさらに調査したい．

謝辞

本研究の一部は JSPS 科研費 25730136 の助成を受け実施したものである．

参考文献

[1] Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large language models in machine translation. In *Proc. EMNLP*, pages 858–867, 2007.
[2] WenHan Chao and ZhouJun Li. Improved graph-based bilingual corpus selection with sentence pair ranking for statistical machine translation. *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, pages 446–451, 2011.
[3] Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. Discriminative corpus weight estimation for machine translation. In *Proc. EMNLP*, pages 708–717, 2009.
[4] Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. Does more data always yield better translations? In *Proc. EACL*, pages 152–161, 2012.
[5] Phillip Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proc. HLT*, pages 48–54, Edmonton, Canada, 2003.
[6] Yajuan Lu, Jin Huang, and Qun Liu. Improving statistical machine translation performance by training data selection and optimization. In *Proc. EMNLP*, pages 343–350, 2007.
[7] Haitao Mi and Liang Huang. Forest-based translation rule extraction. In *Proc. EMNLP*, pages 206–214, 2008.
[8] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the patent translation task at the NTCIR-7 workshop. In *Proc. NTCIR*, 2008.
[9] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the patent translation task at the ntcir-8 workshop. In *Proc. NTCIR*, pages 293–302, 2010.
[10] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL*, pages 177–180, Prague, Czech Republic, 2007.
[11] Graham Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proc. ACL Demo Track*, Sofia, Bulgaria, August 2013.
[12] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proc. ACL*, pages 423–430, 2003.
[13] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proc. ACL*, pages 529–533, Portland, USA, June 2011.
[14] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proc. ACL*, pages 433–440, 2006.
[15] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.
[16] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proc. ACL*, 2003.
[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318, Philadelphia, USA, 2002.
[18] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proc. EMNLP*, pages 944–952, 2010.
[19] Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. ACL*, pages 176–181, 2011.