Vol.2015-NL-224 No.19
2015/12/4

構文木と句の極大性に基づく機械翻訳のための能動学習

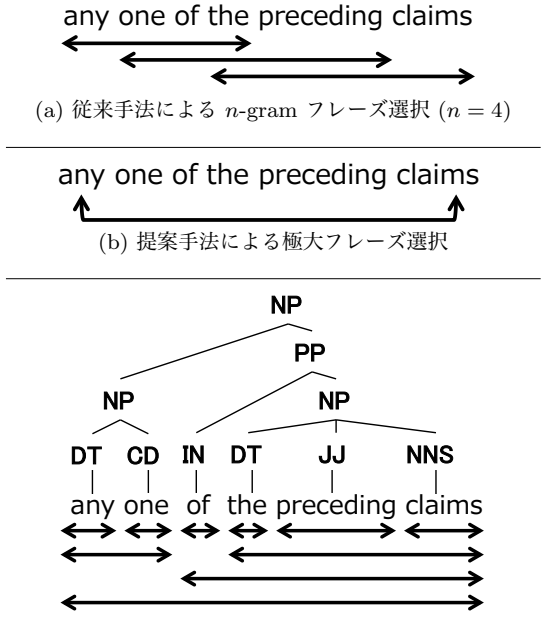
三浦 明波[1,a)]　Graham Neubig[1,b)]　Michael Paul[2,c)]　中村 哲[1,d)]

概要：能動学習は機械学習において，逐次的にデータを選択し，専門知識を有する者が情報を付与してモデルの更新を繰り返すことで，少量の人手作業で効率的に学習を行う枠組みである．この枠組みを機械翻訳に適用することで，人手翻訳のコストを抑えつつ高精度な翻訳モデルを学習可能である．機械翻訳の能動学習には，原言語コーパスの n-gram 頻度に基づきカバレッジを最大化する手法の有効性が知られている．一方で，フレーズの最大長が制限されることにより，慣用表現の断片のみが提示されて，人手翻訳が困難になる場合がある．また，フレーズの重複が多いため，単語数あたりの精度向上率を損なう問題も考えられる．本研究では原言語コーパスの構文解析結果を用いて慣用表現を保存しつつ，包含関係にある極大長のフレーズを抽出する学習データ選択手法を提案する．また，翻訳モデルを逐次更新するシミュレーション実験により，本提案手法の有効性が示された．

1. はじめに

統計的機械翻訳 (Statistical Machine Translation: SMT [1]) で高い翻訳精度を達成するには，学習に用いる対訳コーパスの質と量が不可欠である．特に，質の高い対訳データを得るためには，専門家による人手翻訳が必要となるが，時間と予算の面で高いコストを要するため，翻訳対象は厳選しなければならない．このように，正解データを得るための人手作業を抑えつつ高い精度を達成する手法として，能動学習 (Active Learning) が知られている．統計的機械翻訳においても，能動学習を用いることで人手翻訳のコストを抑えつつ高精度な翻訳モデルを学習可能である [2][3][4][5][6][7].

翻訳候補を含む原言語コーパスの中から次の翻訳対象を選択する基準として，翻訳済みデータにカバーされていない表現をなるべく多く含む文を選択する手法が先ず考えられる [2]．高頻度の未知語や未知のフレーズが優先的にカバーされることで，効率的に翻訳モデルのカバレッジを高められるため，翻訳精度の向上が期待できる．しかし，毎回新しく文全体を選択するため，翻訳済みデータに含まれるフレーズを多く含みやすく，カバー済みのフレーズ長だけ余分な翻訳コストがかかる欠点がある．



(a) 従来手法による n-gram フレーズ選択 ($n = 4$)

(b) 提案手法による極大フレーズ選択

(c) 提案手法による構文部分木選択

図 1　データ選択手法の例，および従来手法と提案手法の比較

文の選択手法では翻訳済みフレーズを冗長に含んでしまう問題に対処するため，原言語コーパスの n-gram 頻度に基づき，最高頻度の未カバーフレーズを順次選択する手法が提案されている [5]．この手法では，選択されたフレーズ全体が必ず翻訳モデルのカバレッジ向上に寄与し，余分な単語を選択しないため，文選択手法よりも少ない単語数の人手翻訳で精度向上が得られやすく，費用対効果に優れている．しかし，この手法には 2 つの問題点が挙げられる．先ず，図 1 (a) に示すように，フレーズの選択手法では互

1　奈良先端科学技術大学院大学 情報科学研究科
Nara Institute of Science and Technology
2　株式会社 ATR-Trek
a)　miura.akiba.lr9@is.naist.jp
b)　neubig@is.naist.jp
c)　michael.paul@atr-trek.co.jp
d)　s-nakamura@is.naist.jp

いに重複する部分が多いため冗長な翻訳作業が発生し，単語あたりの精度向上率を損なう問題がある (フレーズ間の重複問題)．また，最大フレーズ長が $n = 4$ などに制限されるため，“any one of the” ように複合句の一部が不完全な形で作業者に提示されて人手翻訳が困難になる問題もある (句構造の断片化問題)．

フレーズ間の重複問題に対しては，図 1 (b) に示すように包含関係を持つフレーズを統合して，より少ないフレーズでカバレッジを保つことで学習効率の向上が可能と考えられる (フレーズの極大性に基づく選択手法)．重複を取り除き，なるべく長いフレーズを抽出する基準として，本研究では極大部分文字列 [8] の定義を単語列に適用し，極大長となるフレーズの頻度を素性に用いる．また，句構造の断片化問題に対しては，図 1 (c) に示すように，原言語コーパスの句構造解析を行い，構文解析結果の部分木をなすようなフレーズのみを選択することで，人手翻訳が容易になると考えられる (構文木に基づくフレーズ選択手法)．

本研究で提案するデータ選択手法による能動学習効率への影響を調査するため，英日翻訳および英仏翻訳において逐次的にデータ追加・モデル更新・評価を行うシミュレーション実験を実施し，その結果，より少ない追加単語数でカバレッジの向上や翻訳精度の向上を達成することができた．

2. 機械翻訳における能動学習

本節では，機械翻訳における能動学習手法について述べる．ここで，フレーズとは任意の長さの単語列を表すものとし，文そのものや単語もひとまとめにフレーズと呼ぶこととする．翻訳対象の候補となるデータを含む原言語コーパスから，逐次的に新しい原言語フレーズを選択し翻訳，正解データとして対訳コーパスに加える手順をまとめると下表のように一般化できる．

Algorithm 1 能動学習手法

1: **Init:**
2: $SrcPool \leftarrow$ 翻訳候補の原言語コーパス
3: $Translated \leftarrow$ 翻訳済みの対訳コーパス
4: $Oracle \leftarrow$ 入力フレーズの正解訳を与えるオラクル
5: **Loop Until** 停止条件**:**
6: $TM \leftarrow TrainTranslationModel(Translated)$
7: $NewSrc \leftarrow SelectNextPhrase(SrcPool, Translated, TM)$
8: $NewTrg \leftarrow GetTranslation(Oracle, NewSrc)$
9: $Translated \leftarrow Translated \bigcup \{\langle NewSrc, NewTrg\rangle\}$

1 行目から 4 行目でデータの定義，初期化を行う．$SrcPool$ は原言語コーパスの各行を要素とする集合である．$Translated$ は翻訳済みの原言語フレーズと目的言語フレーズの対を要素とする集合であり，初期状態は空でもよいが，既に対訳データが与えられている場合には，$Translated$ を設定することで効率的に追加データの選択を行うことができる．$Oracle$ は任意の入力フレーズに対して正解訳を与えることができるオラクルであり，人手翻訳を模したモデルである．本研究では全対訳データを用いて学習した翻訳モデルをオラクルとして用いており，詳細は 5 節で述べる．

5 行目から 9 行目で逐次的なデータ学習を行う．5 行目の停止条件には，任意の終了タイミングを設定できるが，実際の利用場面では一定の翻訳精度に達成した時点や，予算の許容する単語数を翻訳し終えた時点などで能動学習を打ち切ることになるだろう．6 行目では，その時点で保持している対訳コーパス $Translated$ を用いて翻訳モデルの学習を行う．また，実験的評価においては，翻訳モデルの学習直後に翻訳精度の評価を行う．7 行目では $SrcPool$, $Translated$, TM を判断材料として，次に翻訳対象となる原言語フレーズを選択する．ここでデータ選択時に基準となる要素として，学習済みモデルにおける各データの信頼度，コーパス中に出現する各フレーズの代表性，翻訳候補のフレーズから正解訳を得るためのコストなどが考えられる．

3. *n*-gram 頻度に基づくデータ選択手法

本節では，従来手法である n-gram 頻度に基づく文選択手法とフレーズ選択手法について紹介する．

3.1 *n*-gram 頻度に基づく文選択手法

n-gram 頻度に基づく文選択手法では，原言語コーパスに含まれる単語数が n 以下の全フレーズのうち，翻訳済みの原言語データに出現せず，かつ頻度が最大となるようなものを含む文を選択する．逐次的に文を追加していき，翻訳済みのデータが原言語コーパスの全 n-gram フレーズをカバーした時点で能動学習を停止する．この手法によって最頻出の n-gram フレーズを効率的にカバー可能であり，翻訳コストを抑えつつ高い精度を達成できる．Bloodgood らは，$n = 4$ の n-gram 頻度に基づく文選択手法を用いた能動学習のシミュレーション実験によって，原言語データ全てを翻訳する場合に比べて，80% 未満の文数で同等の BLEU スコア [9] を達成できたと報告している [5]．

しかし，文全体を選択するため，翻訳済みのデータに既にカバーされているフレーズも多く含んでおり，重複部分の単語数だけ余分な翻訳コストがかかると考えられる．そのため，文全体ではなく高頻出のフレーズのみを選択する手法を 3.2 節から紹介する．

3.2 *n*-gram 頻度に基づくフレーズ選択手法

n-gram 頻度に基づくフレーズ選択手法では，3.1 節の文選択手法とは異なり，原言語コーパス中で翻訳済みデータにカバーされていない単語数 n 以下のフレーズそのものを頻度順に選択する．この手法では，文全体の選択を行うよ

りも少ない単語数の追加でカバレッジを高めることができるため，翻訳コストの低減による精度向上効率が期待できる．Bloodgood らは，ベースとなる対訳データを元に，追加の原言語データ中の高頻度の未カバー n-gram フレーズを順次選択し，アウトソーシングサイトを用いた人手翻訳実験により，少ない追加単語数と短い翻訳時間でベースシステムよりも大幅に BLEU スコアの向上を確認できたと報告している [5]．

ただし，このフレーズ選択手法では，1 節で述べたようにフレーズ長が $n = 4$ などに制限されるため，選択されるフレーズどうしの重複が多い問題や，複合句の断片が選択される問題があり，また長いフレーズ対応を学習できないことも機械翻訳を行う上で不利である．$n = 5$ などの，より長いフレーズ長を設定することは根本的な解決にならないばかりか，さらに多くのフレーズの重複が発生して逆効果となり得る．

4. 構文木とフレーズの極大性に基づく手法

本節では，提案手法であるフレーズの極大性に基づく選択手法と，構文木に基づくフレーズ選択手法について説明する．

4.1 フレーズの極大性に基づく選択手法

本節では，n-gram 頻度に基づくフレーズ選択ではフレーズ長の制限により発生する，フレーズの重複問題を解消するために，極大部分文字列の定義を利用したフレーズ選択手法を提案する．極大部分文字列は Okanohara らによって効率的に文書分類器を学習するために提案された素性であり [8]，以下のような半順序関係の定義を用いて示すことができる．

$$s_1 \preceq s_2 \Leftrightarrow \exists \alpha, \beta :\ s_1 = \alpha s_2 \beta \wedge occ(s_1) = occ(s_2) \quad (1)$$

ここで s_1, s_2, α, β は長さ 0 以上の要素列であり，$occ(\cdot)$ は文書中の要素列の出現回数である．例えば，

p_1 = “one of the preceding”, $occ(p_1) = 200,000$
p_2 = “one of the preceding claims”, $occ(p_2) = 200,000$
p_3 = “any one of the preceding claims”, $occ(p_3) = 190,000$

のようなフレーズが原言語コーパス中に出現している場合，$p_1 = \alpha p_2 \beta$，α = “”，β = “claims” が成り立ち，すなわち p_1 は p_2 の部分単語列であり，同様に p_2 は p_3 の部分単語列である．p_1 は p_2 の部分単語列であり，コーパス中の出現頻度について $occ(p_1) = occ(p_2) = 200,000$ が成り立つため，式 1 により $p_1 \preceq p_2$ が成り立つ．一方，p_2 は p_3 の部分単語列であるが，$occ(p_2) = 200,000 \neq 190,000 = occ(p_3)$ であるため，$p_2 \preceq p_3$ とはならない．式 1 で定義される半順序 $\preceq$ を用いて，単語列 s_1 について $s_1 \preceq s_2$ となるような s_2 が s_1 自体を除いて存在しない場合に，s_1 は極大性を有し，本稿では極大フレーズと呼ぶこととする．先述の例では，$p_1 \preceq p_2$ であるため p_1 は極大フレーズではなく，$p_2 \preceq p$ となるような p は p_2 自体を除いて存在しないため p_2 は極大フレーズである．

このような極大フレーズとその頻度は，拡張接尾辞配列 [10] を用いて線形時間で列挙可能である．ただし，本提案手法では，極大フレーズが改行文字を含む場合は分割し，また，出現回数が 2 以上のものを列挙するようにしている．

フレーズの極大性に基づく選択手法では，原言語コーパス中の極大フレーズを列挙し，翻訳済みデータに出現していないフレーズを頻度順に選択する．これによって，先の例では，$p_1 \preceq p_2$ であるため，p_1 は翻訳候補に選ばれず，フレーズの重複を削減できたことになる．ところが，p_2 と p_3 では出現頻度が互いに近いが一致はしないため，p_2，p_3 ともに極大フレーズとして選択されてしまう．

このように，極大フレーズでは，包含関係にあるフレーズの出現頻度が一致しない場合には重複を削除することができない．しかし，前述の p_2 と p_3 の例のように，完全一致ではないがほとんど一致している際にもフレーズを統合したい場合が考えられる．そこで，式 1 の制約を緩めた半順序関係を下記のように提案する．

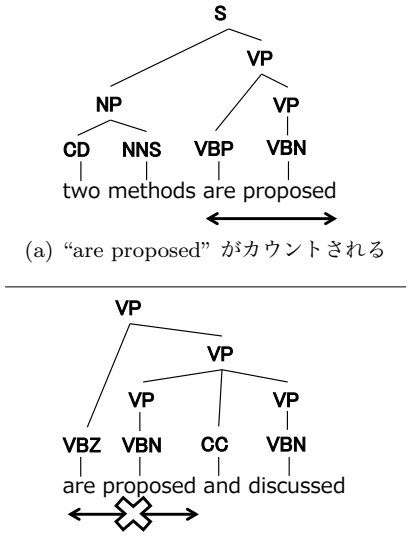
$$s_1 \overset{*}{\preceq} s_2 \Leftrightarrow \exists \alpha, \beta :\ s_1 = \alpha s_2 \beta \wedge \frac{occ(s_1)}{2} < occ(s_2) \quad (2)$$

式 2 により，半順序 $\overset{*}{\preceq}$ を用いて極大性を定義可能であり，通常の極大フレーズと区別するため半極大フレーズと呼び，このような特徴を持つフレーズを列挙し，未カバーフレーズを頻度順に追加する手法を提案する．

半極大フレーズに基づくデータ選択手法では，通常の極大フレーズに基づく手法に比べ，包含関係を持つフレーズが大幅に削除されるが，長いフレーズを残すことにより頻度は低下し，データ選択の優先順位が他の手法とは大きく変化するため，能動学習への影響は明らかではない．

4.2 構文木に基づくフレーズ選択手法

本節では 4.1 節で述べた提案手法とは別に，原言語コーパスの構文解析結果に基いてフレーズを選択する手法を提案する．本手法では，図 2 に示すように，翻訳候補となる原言語コーパスの全文を句構造解析器で処理し，得られた構文木の全部分木をたどりながらフレーズを数え上げ，その後にフレーズを頻度順に選択する．部分木の構造が異なっても，単語列が一致していれば同一のものとしてカウントする．これにより，木をまたがるようなフレーズ選択は行われないため，複合句が分断されるような問題は発生せず，選択されるデータは構文的にまとまった意味を持つと考えられる．本手法で選択された翻訳候補のフレーズは，統語情報を用いない他の手法と比べて，人手翻訳を行う際に有用で，同じ追加単語数でも短い時間で正解データが得られるものと期待できる．n-gram 頻度や極大性に基



(a) “are proposed” がカウントされる

(b) “are proposed” はカウントされない

図 2　構文木に基づく手法のフレーズカウント条件

いてフレーズを選択する手法では，表層的な単語列を数え上げるため，“two methods are proposed” というフレーズがあると，その一部である “are proposed” も頻度に加えるが，構文木に基づく場合，図 2 (b) に示すように “are proposed” は部分木をまたがるために頻度に加えない．このため，構文木に基づくフレーズ選択手法では，フレーズの頻度が他の手法による表層的な数え上げよりも小さくなる傾向があり，結果として 2 単語以上からなるフレーズを選択する優先順位が低くなりやすい．

この手法では，全部分木のフレーズを数え上げるため，n-gram 頻度に基づくフレーズ選択手法と同様に，フレーズの重複により追加単語数あたりの精度向上率に悪影響が出る可能性がある．従って，4.1 節で提案した半極大フレーズと併用することで，重複を取り除き，選択するフレーズを絞り込む手法も同時に提案する．

5. 実験的評価

5.1 実験設定

4 節で提案したデータ選択手法が，機械学習のための能動学習にどのような影響を与えるかを調査するため，逐次的にデータを追加して翻訳モデルを更新するシミュレーション実験を実施し，各ステップにおける翻訳精度の比較評価を行った．本実験では，高精度な構文解析器を利用可能な英語を原言語とし，目的言語には日本語とフランス語を選択した．対訳コーパスが全く存在しない状態から能動学習を用いることも可能であるが，より実際的な利用方法を考慮し，一般分野の対訳コーパスが存在している状態に，専門分野の追加コーパスからフレーズを選択し，翻訳モデルの高精度化を目指す．日英翻訳には，日常的な英会話表現を広くカバーする英辞郎例文データをベースの対訳コーパスとし，科学論文の概要を元に抽出された

言語対	分野	データセット	文数/単語数
En-Ja	一般 (ベース)	Train	414k 文 En: 6.72M 単語 Ja: 9.69M 単語
	科学論文 (追加)	Train	1.87M 文 En: 46.4M 単語 Ja: 57.6M 単語
		Test	1790 文
		Dev	1790 文
En-Fr	一般 (ベース)	Train	1.89M 文 En: 47.6M 単語 Fr: 49.4M 単語
	医療 (追加)	Train	15.5M 文 En: 393M 単語 Fr: 418M 単語
		Test	1000 文
		Dev	500 文

表 1　対訳コーパスのデータ内訳 (有効数字 3 桁)

ASPEC[*1] を追加の対訳コーパスとして用いた．日仏翻訳には，WMT2014[*2] の翻訳タスクで用いられた欧州議会議事録の Europarl コーパスをベースとし，医療翻訳タスクで用いられたデータのうち EMEA，PatTR，Wikipedia タイトルを合わせて追加コーパスとした．前処理として，日本語コーパスの単語分割には KyTea を用いており，学習に用いるトレーニングデータのうち，単語数が 60 を超える行は取り除いた．前処理後の対訳データの内訳を表 1 にまとめる．

翻訳の枠組みには，フレーズベース翻訳 [11] を用いた．ただし，少量の対訳を追加して単語アラインメントの再学習およびフレーズテーブルの再構築を行うには計算コストが非常に大きい．そのため，単語アラインメントには GIZA++ を逐次学習に対応させた inc-giza-pp[*3] を用いており，翻訳モデルの学習には Moses の MMSAPT(Memory-mapped Dynamic Suffix Array Phrase Tables) 機能を利用して，フレーズ抽出を行わずに接尾辞配列による動的なフレーズテーブルの構築を行った．言語モデルの学習には KenLM [12] を用いて，ベースコーパスと追加コーパスの全トレーニングデータから $n = 5$ の n-gram 言語モデルを学習した．デコード時のパラメータ調整には MERT [13] を用いたが，データ追加の度に最適化を行うのは時間的に現実的でないため，ベースコーパス全文で学習した翻訳モデルに対して，追加コーパス用の開発データセットで自動評価尺度の BLEU スコア [9] が最大となるよう学習を行い，その後はパラメータを固定し能動学習を行った．能動学習に用いるデータ選択手法には従来手法と提案手法を含め，以下のように 8 つのタスクを設定した．

[*1] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[*2] http://statmt.org/wmt14/
[*3] https://code.google.com/p/inc-giza-pp/

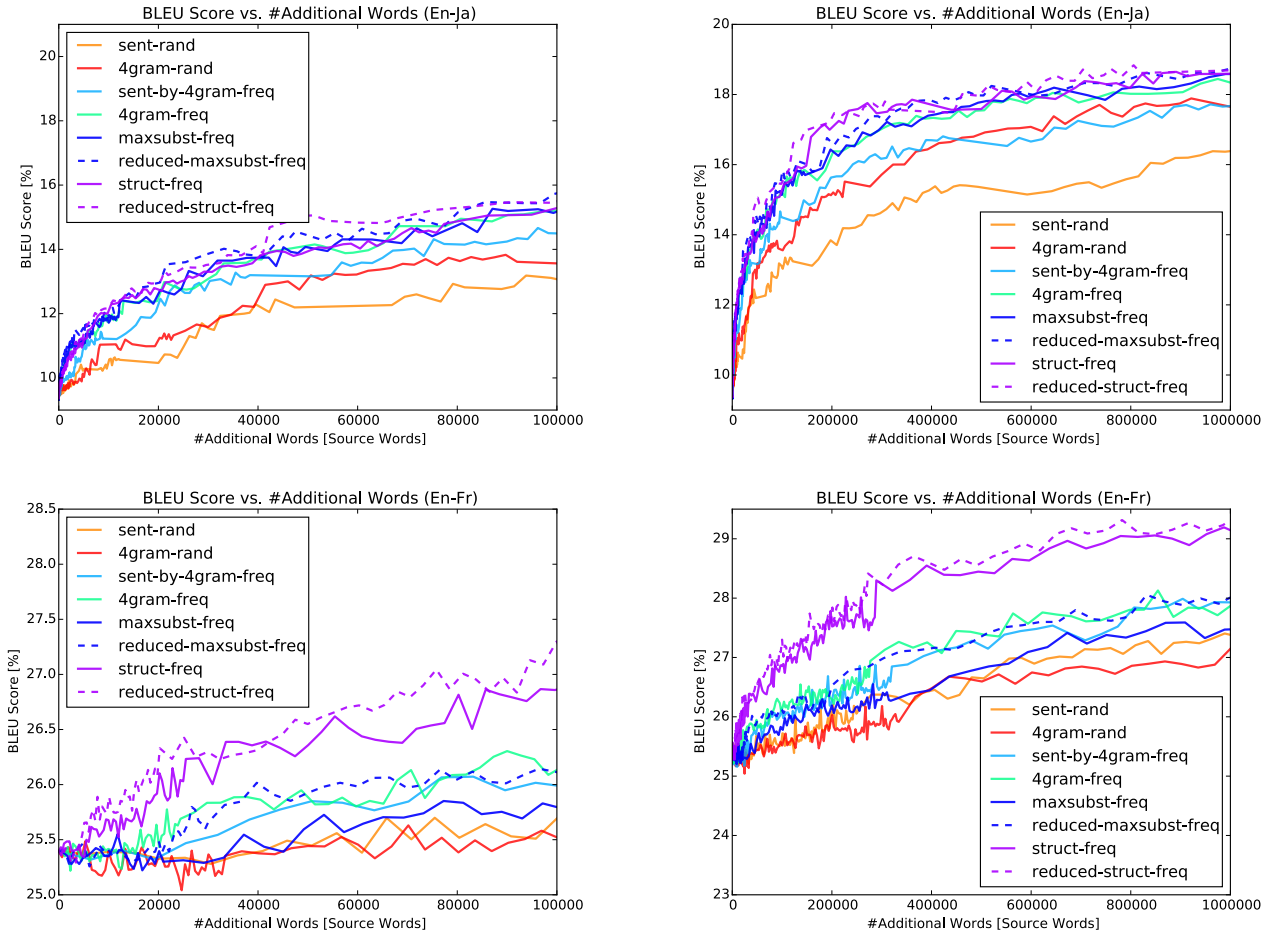


図 3　各手法における追加単語数あたりの BLEU スコア (左上: 10 万単語まで追加の英日翻訳，右上: 100 万単語まで追加の英日翻訳，左下: 10 万単語まで追加の英仏翻訳，右下: 100 万単語まで追加の英仏翻訳)

文の乱択 (sent-rand):
追加コーパスの順序をシャッフルし，順次選択

フレーズの乱択 (4gram-rand):
ベースコーパス中に含まれない追加コーパス中の単語数 4 以下のフレーズを列挙後にシャッフルし，順次選択

4-gram 頻度に基づく文選択 (sent-by-4gram-freq):
翻訳済みデータに含まれず，単語数 4 以下で最高頻度のフレーズを含む文を順次選択 (3.1 節)

4-gram 頻度に基づくフレーズ選択 (4gram-freq):
翻訳済みデータに含まれず，単語数 4 以下で最高頻度のフレーズを順次選択 (3.2 節)

極大性に基づく手法 (maxsubst-freq):
翻訳済みデータに含まれず，追加コーパス中で最高頻度の極大フレーズを順次選択 (4.1 節)

構文木に基づく手法 (struct-freq):
追加コーパスの句構造解析結果を元に，部分木を成すようなフレーズの中から翻訳済みデータに含まれず最高頻度のものを順次追加 (4.2 節)

半極大性に基づく手法 (reduced-maxsubst-freq):
翻訳済みデータに含まれず，追加コーパス中で最高頻度の半極大フレーズを順次選択 (4.1 節)

構文木と半極大性に基づく手法 (reduced-struct-freq):
追加コーパスの句構造解析結果を元に，部分木を成すような半極大フレーズの中から翻訳済みデータに含まれない最高頻度のものを順次追加 (4.1 節, 4.2 節)

それぞれの手法で選択されたデータの正解訳を得るために，文の選択に対しては対応する対訳文をそのまま選択，フレーズの選択に対してはベースコーパスと追加コーパスの全文を用いて学習した翻訳モデルをオラクルとして，翻訳結果を対訳フレーズとした．フレーズの出現頻度を扱う全ての手法において，計算資源を節約するため頻度 1 のものは取り除いた．構文木に基づく手法では，句構造解析を行うために Ckylark [14] を使用した．

5.2　実験結果

シミュレーション実験により得られた結果から，それぞれの言語対で 10 万単語まで追加した場合と 100 万単語まで追加した場合の単語数と翻訳精度の変化を図 3 に示す．英日翻訳シミュレーション結果からは，4gram-freq と

言語対	データ選択手法	全フレーズ追加			1 万単語追加	
		フレーズ数	単語数	平均フレーズ長	フレーズ数	平均フレーズ長
En-Ja	sent-by-4gram-freq	1.28M	33.6M	26.3	560	17.8
	4gram-freq	8.48M	26.0M	3.07	4.70k	2.13
	maxsubst-freq	7.29M	25.8M	3.54	4.51k	2.22
	reduced-maxsubst-freq	6.06M	21.7M	3.58	4.76k	2.10
	struct-freq	1.45M	4.85M	3.34	6.64k	1.51
	reduced-struct-freq	1.10M	3.33M	3.03	6.73k	1.49
En-Fr	sent-by-4gram-freq	10.6M	269M	25.4	310	32.1
	4gram-freq	40.1M	134M	3.34	3.62k	2.76
	maxsubst-freq	62.4M	331M	5.30	2.39k	4.17
	reduced-maxsubst-freq	45.9M	246M	5.36	2.95k	3.39
	struct-freq	14.1M	94.2M	6.68	4.01k	2.49
	reduced-struct-freq	7.33M	41.3M	5.63	4.55k	2.20

表 2 手法ごとに選択されるデータ内訳 (有効数字 3 桁)

言語対	データ選択手法	1-gram / 4-gram カバレッジ [%]			
		追加なし	1 万単語	10 万単語	100 万単語
En-Ja	sent-rand	94.36 / 5.38	94.81 / 5.63	95.99 / 6.59	97.54 / 10.06
	4gram-rand		94.80 / 5.38	96.10 / 5.46	97.67 / 5.98
	sent-by-4gram-freq		95.10 / 5.84	96.28 / **7.23**	97.64 / **11.39**
	4gram-freq		95.64 / 5.97	96.87 / 7.14	97.97 / 10.43
	maxsubst-freq		95.59 / 5.96	96.83 / 7.07	97.91 / 10.20
	reduced-maxsubst-freq		95.73 / **6.00**	96.97 / 7.19	98.00/10.57
	struct-freq		96.60 / 5.44	97.80 / 5.79	98.58 / 7.02
	reduced-struct-freq		**96.64** / 5.44	**97.84** / 5.80	**98.61** / 7.14
En-Fr	sent-rand	92.72 / 10.60	92.93 / 10.60	93.73 / 10.71	95.94 / 11.30
	4gram-rand		92.95 / 10.60	93.99 / 10.60	96.42 / 10.64
	sent-by-4gram-freq		92.95 / 10.60	93.96 / **10.72**	96.25 / **11.55**
	4gram-freq		92.92 / 10.60	94.46 / 10.66	96.60 / 11.16
	maxsubst-freq		92.79 / 10.60	93.61 / 10.62	95.99 / 10.92
	reduced-maxsubst-freq		92.92 / 10.60	94.38 / 10.66	96.55 / 11.13
	struct-freq		93.63 / 10.60	96.15 / 10.65	97.84 / 11.28
	reduced-struct-freq		**94.02** / 10.60	**96.38** / 10.69	**98.00** / 11.38

表 3 各データ選択手法がカバレッジに与える影響（小数点第三位を四捨五入），太字は一定の単語数追加時点でのカバレッジ最大値を示す

maxsubst-freq でほとんど精度差が出ず，変動がほぼ一致している．struct-freq や reduced-struct-freq も追加単語数が少ないうちは 4gram-freq とあまり大きな差は見られなかったが，約 4 万単語追加時点から他の手法よりも精度が高くなっており，約 50 万単語追加時点からは頻度に基づく手法の精度がほぼ横這いとなった．一方，英仏翻訳シミュレーション結果では，最初から構文木に基づく手法での精度の伸びが良く，他の手法よりも精度が大きく上回り，100 万単語追加時点でも差はほとんど縮まらなかった．また，全てのグラフに共通して，maxsubst-freq よりも reduced-maxsubst-freq が，struct-freq よりも reduced-struct-freq が多くの場合に精度向上が確認でき，長いフレーズを優先することで，結果的に少ない単語数でカバレッジを向上できたと考えられる．

手法毎に翻訳対象のフレーズ選択基準が異なるため，フレーズ長制限の有無や重複の削減方法の違いによって，翻訳対象を選び尽くした場合のフレーズ数等に大きな差が出ることになる．フレーズ頻度に基づくそれぞれの手法によって選択されるデータの傾向を調べるため，翻訳候補を全て追加し終えた時点および約 1 万単語のみ追加した時点でのフレーズ数，単語数，平均フレーズ長を表 2 にまとめる．どの手法においても，カバレッジに寄与するフレーズがこれ以上存在しないと判断した際には能動学習を停止する．言い換えれば，全翻訳対象を翻訳し終えた時点でカバ

レッジが収束するため，翻訳対象の単語数が少ないほどカバレッジの収束が速く，翻訳精度の向上しやすいと考えられる．一方，一度に追加するフレーズが長いほど，同時に複数の n-gram をカバーできるため，平均フレーズ長が大きいほど 4-gram カバレッジ等を向上させる上で有利と考えられる．提案手法による選択データの平均フレーズ長が英日翻訳で 3.03〜3.58 単語，英仏翻訳で 5.30〜6.68 単語と大きく差が開いているが，これは原言語側のベースコーパスと追加コーパスの組み合わせのみに依存しており，目的言語には当然依存しない．また，表 2 のフレーズ頻度に基づく手法において，全データ追加時の平均フレーズ長に比べ，1 万単語追加時の平均フレーズ長が短いことを確認できる．短いフレーズほど高頻度となりやすく優先的に選択されるため当然であるが，構文木に基づく手法では 1 万単語追加時の平均フレーズ長が極端に小さくなっており，長いフレーズの頻度が大幅に下がりやすい傾向が見られる．

また，各手法によって，翻訳済みのデータが実際に評価データをどの程度カバーしているかを調査する．各手法でフレーズを 1 つずつ選択していき，追加単語数が 1 万，10 万，100 万にそれぞれ達する時点での評価データの 1-gram カバレッジおよび 4-gram カバレッジを表 3 にまとめる．この結果から，reduced-struct-freq ではどの場合でも最も 1-gram カバレッジが向上していることが分かり，効率的に未知語がカバーされることになる．一方で，4-gram カバレッジに関しては，3 単語以下のフレーズを追加しても全く影響が出ないため，長いフレーズを追加する方が有利であることは明らかであり，sent-by-4gram-freq で最も効率的に向上が見られる．英仏翻訳では，1 万単語追加時点で 4-gram カバレッジの上 4 桁に変化が見られなかった．このように，フレーズ選択時に長いフレーズを選ぶか，短いフレーズを選ぶかは，カバレッジの影響を考える際にトレードオフの関係が生じるが，半極大性に基いて重複を取り除くことによって，1-gram カバレッジと 4-gram カバレッジを両立して向上させられることが確認できた．

6. おわりに

本稿では，機械翻訳のための能動学習において，フレーズの極大性と半極大性を導入し，構文解析結果と組み合わせてフレーズ抽出を行い頻度順に学習データの選択を行う手法を提案した．英仏翻訳シミュレーションでは提案手法により，従来手法よりも大幅に少ない翻訳コストで翻訳精度が向上したが，英日翻訳では提案法の優位性はあまり現れず，本手法の有効性がデータ依存であることも示唆された．また，提案手法によって評価データのカバレッジが効率的に向上していることも確認できた．本手法は構文解析結果に基づいたフレーズ選択を行っているため，従来手法よりも人手翻訳が容易で短時間に高品質な対訳データを得られることが期待できるが，本稿のために行ったシミュレーション実験では確認できなかったため，今後の課題としたい．また，提案手法では文選択手法に比べてあまり 4-gram カバレッジが向上しなかったが，より効率的なフレーズ選択手法を提案することで，翻訳精度の向上を試みたいと考えている．

謝辞： 本研究は，(株)ATR-Trek の助成を受け実施したものである．

参考文献

[1] Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 19, pp. 263–312, 1993.

[2] Matthias Eck, Stephan Vogel, and Alex Waibel. Low Cost Portability for Statistical Machine Translation based in N-gram Frequency and TF-IDF. In *Proc. IWSLT*, pp. 61–67, 2005.

[3] Gholamreza Haffari and Anoop Sarkar. Active Learning for Multilingual Statistical Machine Translation. In *Proc. ACL*, pp. 181–189, August 2009.

[4] Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. A Semi-Supervised Batch-Mode Active Learning Strategy for Improved Statistical Machine Translation. In *Proc. CoNLL*, pp. 126–134, July 2010.

[5] Michael Bloodgood and Chris Callison-Burch. Bucking the Trend: Large-Scale Cost-Focused Active Learning for Statistical Machine Translation. In *Proc. ACL*, pp. 854–864, July 2010.

[6] Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. Active learning for interactive machine translation. In *Proc. EACL*, pp. 245–254, April 2012.

[7] Spence Green, Sida I. Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D. Manning. Human Effort and Machine Learnability in Computer Aided Translation. In *Proc. EMNLP*, pp. 1225–1236, October 2014.

[8] Daisuke Okanohara and Jun'ichi Tsujii. Text Categorization with All Substring Features. In *Proc. SDM*, pp. 838–846, 2009.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*, pp. 311–318, July 2002.

[10] Toru Kasai, Gunho Lee, Hiroki Arimura, Setsuo Arikawa, and Kunsoo Park. Linear-Time Longest-Common-Prefix Computation in Suffix Arrays and Its Applications. In *Proc. CPM*, pp. 181–192, 2001.

[11] Phillip Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-Based Translation. In *Proc. HLT*, pp. 48–54, 2003.

[12] Kenneth Heafield. KenLM: Faster and Smaller Language Model Queries. In *Proc, WMT*, July 2011.

[13] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. ACL*, pp. 160–167, 2003.

[14] Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Ckylark: A More Robust PCFG-LA Parser. In *Proc. NAACL*, pp. 41–45, June 2015.