統語的一貫性と非冗長性を重視した機械翻訳のための能動学習手法

三浦 明波[†]・Graham Neubig[†,††]・Michael Paul[†††]・中村 哲[†]

能動学習は機械学習において，逐次的に選択されたデータに対してのみ正解ラベルを付与してモデルの更新を繰り返すことで，少量のコスト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．機械翻訳のための能動学習では，人手翻訳の対象となる文またはフレーズをどのように選択するかが学習効率に大きな影響を与える要因となる．既存研究による代表的な手法として，原言語コーパスの単語 n-gram 頻度に基づき n-gram カバレッジを向上させる手法の有効性が知られている．この手法は一方で，フレーズの最大長が制限されることにより，句範疇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，能動学習の過程で選択されるフレーズには，共通の部分単語列が繰り返し出現するため，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句構造解析結果を用いて句範疇を保存しつつ，包含関係にある極大長のフレーズのみを人手翻訳の候補とするフレーズ選択手法を提案する．本研究の提案手法の有効性を調査するため，機械翻訳による擬似対訳を用いたシミュレーション実験および専門の翻訳者による人手翻訳と主観評価を用いた実験を実施した．その結果，提案手法によって従来よりも少ない単語数の翻訳で高い翻訳精度を達成できることや，人手翻訳時の対訳の品質向上に有効であることが示された．

キーワード：統計的機械翻訳, 能動学習, 人手翻訳, 対訳コーパス, 構文解析, 句構造解析

Selecting Syntactic, Non-redundant Segments in Active Learning for Machine Translation

Akiva Miura[†], Graham Neubig[†,††], Michael Paul[†††] and Satoshi Nakamura[†]

Active learning is a framework that makes it possible to efficiently train statistical models by selecting informative examples from a pool of unlabeled data. Previous work has found this framework effective for machine translation (MT), making it possible to train better translation models with less effort, particularly when annotators translate short phrases instead of full sentences. However, previous methods for phrase-based active learning in MT fail to consider whether the selected units are coherent and easy for human translators to translate, and also have problems with

[†] 奈良先端科学技術大学院大学 情報科学研究科,
Graduate School of Information Science, Nara Institute of Science and Technology
[††] カーネギーメロン大学 言語技術研究所, Language Technologies Institute, Carnegie Mellon University
[†††] 株式会社 ATR-Trek, ATR-Trek Co. Ltd.

selecting redundant phrases with similar content. In this paper, we tackle these problems by proposing two new methods for selecting more syntactically coherent and less redundant segments in active learning for MT. Experiments using both simulation and extensive manual translation by professional translators find the proposed method effective, achieving both greater gain of BLEU score for the same number of translated words, and allowing translators to be more confident in their translations.

Key Words: *Statistical Machine Translation, Active Learning, Manual Translation, Parallel Corpus, Syntactic Parsing, Phrase Strucuture Analysis*

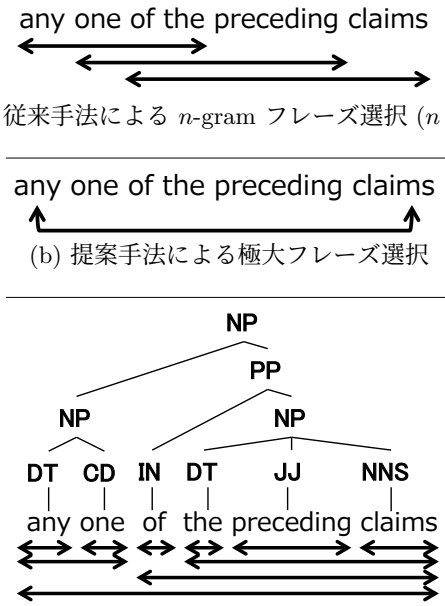
1 はじめに

統計的機械翻訳 (Statistical Machine Translation: SMT (Brown, Pietra, Pietra, and Mercer 1993)) で高い翻訳精度 [1] 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．SMT 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 (Eck, Vogel, and Waibel 2005; Turchi, De Bie, and Cristianini 2008; Haffari, Roy, and Sarkar 2009; Haffari and Sarkar 2009; Ananthakrishnan, Prasad, Stallard, and Natarajan 2010a; Bloodgood and Callison-Burch 2010; González-Rubio, Ortiz-Martínez, and Casacuberta 2012; Green, Wang, Chuang, Heer, Schuster, and Manning 2014)．

SMT や，その他の自然言語処理タスクにおける多くの能動学習手法は，膨大な文書データの中からどの**文**をアノテータに示すか，という点に注目している．これらの手法は一般的に，幾つかの基準に照らし合わせて，SMT システムに有益な情報を多く含んでいると考えられる文に優先順位を割り当てる．単言語データに高頻度で出現し，既存の対訳データには出現しないような**フレーズ** [2] を多く含む文を選択する手法 (Eck et al. 2005)，現在の SMT システムにおいて信頼度の低いフレーズを多く含む文を選択する手法 (Haffari et al. 2009)，あるいは翻訳結果から推定される SMT システムの品質が低くなるような文を選択する手法 (Ananthakrishnan et al. 2010a) などが代表的である．これらの手法で選択される文は，機械学習を行う上で有益な情報を含んでいると考えられるが，その反面，既存システムに既にカバーされているフレーズも多く含んでいる可能性が高く，余分な翻訳コストを要する欠点がある．

[1] SMT システムの性能を評価する場合，評価用原言語コーパスの翻訳結果が目標となる正解訳にどの程度近いかを示す自動評価尺度を翻訳精度の指標とすることが多く，本稿では最も代表的な自動評価尺度と考えられる BLEU スコア (Papineni, Roukos, Ward, and Zhu 2002) を用いて評価する．

[2] 本稿では，フレーズとは特定の文中に出現する任意の長さの部分単語列を表すものとし，文全体や単語もフレーズの一種として扱う．また，後述する句構造文法における句とは区別して扱うこととする．



(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) 提案手法による部分構文木選択

図 1 フレーズ選択手法の例，および従来手法と提案手法の比較

このように文全体を選択することで過剰なコストを要する問題に対処するため，自然言語処理タスクにおいては短いフレーズからなる**文の部分的アノテーション**を行うための手法も提案されている (Settles and Craven 2008; Tomanek and Hahn 2009; Bloodgood and Callison-Burch 2010; Sperber, Simantzik, Neubig, Nakamura, and Waibel 2014)．特に SMT においては，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，原言語コーパスの単語 n-gram 頻度に基づき，対訳コーパスにカバーされていない原言語コーパス中で最高頻度の n-gram 自体を翻訳対象のフレーズとして選択する手法が提案されている (Bloodgood and Callison-Burch 2010)．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 **n-gram カバレッジ**[3] 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翻訳精度を向上させやすく，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，n-gram 頻度に基づくフレーズの選択手法では複数のフレーズ間で共有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，“one of the preceding” のように句範疇の一部がたびたび不完全な形で翻訳者に提示

[3] 入力されるデータに対して，その構成要素がどの程度モデルに含まれているかという指標をカバレッジ (被覆率) と呼ぶ．本稿では，原言語コーパス中の n-gram が翻訳モデル中に含まれる割合に着目する．

されて人手翻訳が困難になる問題もある (**句範疇の断片化問題**).

本研究では，前述の 2 つの問題に対処するために 2 種類の手法を提案し，部分アノテーション型の**能動学習効率** [4] と翻訳結果に対する**自信度**の向上を目指す (4 節)．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 (**極大フレーズ選択手法**, 4.1 節)．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** (Yamamoto and Church 2001; Okanohara and Tsujii 2009) の定義を単語列に適用し，極大長 [6] となるフレーズの頻度を素性に用いる．句範疇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，部分木をなすようなフレーズを**統語的に整ったフレーズ**とみなして選択することで，人手翻訳が容易になると考えられる (**部分構文木選択手法**, 4.2 節)．また，これら 2 つの手法を組み合わせ，フレーズの極大性と構文木を同時に考慮する手法についても提案する (4.2 節).

本研究で提案するフレーズ選択手法による能動学習効率への影響を調査するため，先ず英仏翻訳および英日翻訳において逐次的にフレーズ対の追加・モデル更新・評価を行うシミュレーション実験 (5 節) を実施し，その結果，2 つの提案手法を組み合わせることで従来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．次に，部分構文木選択手法が人手翻訳に与える影響を調査するため，専門の翻訳者に翻訳作業と主観評価を依頼し，述べ 120 時間におよぶ作業時間で収集された対訳データを用いて実験と分析を行った結果 (6 節)，同様に高い能動学習効率が示された．また，翻訳者は構文木に基づくフレーズ選択手法において，より長い翻訳時間を要するが，より高い自信度の翻訳結果が得られるという傾向も得られた [5].

2 機械翻訳のための能動学習

本節では，機械翻訳のための能動学習手法について述べる．翻訳対象の候補となるフレーズ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，学習用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 Algorithm 1 のように一般化できる．

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，$Translated$

[4] 人手翻訳に要した一定のコストに対する翻訳精度の上昇値を本稿における学習効率とし，作業時間あたりの精度向上と必要予算あたりの精度向上に注目する．
[5] 本稿の内容は (三浦，Neubig，Paul，中村 2015, 2016) および (Miura, Neubig, Paul, and Nakamura 2016) で報告されている．

Algorithm 1 機械翻訳のための能動学習手法

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: $Translated \leftarrow Translated \bigcup \{\langle NewSrc, NewTrg \rangle\}$

を設定することで効率的に追加フレーズ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．

5 行目から 9 行目で翻訳モデルの逐次的な学習を行う．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．ここでフレーズ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フレーズ対の信頼度，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．

次節からは，先述のアルゴリズム 7 行目で述べたフレーズ選択基準に用いられる具体的な手法として，既存のフレーズ選択手法 (3 節) および本研究の提案手法 (4 節) について述べる．

3　単語 n-gram 頻度に基づく文・フレーズ選択手法

本節では，従来手法である単語 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　単語 n-gram 頻度に基づく文選択手法

単語 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 (Papineni et al. 2002) を達成できたと報告している (Bloodgood and Callison-Burch 2010)．

しかし，1 節で述べたように，この手法は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 単語 n-gram 頻度に基づくフレーズ選択手法

単語 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 n-gram 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って高い能動学習効率が期待できる．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 (Bloodgood and Callison-Burch 2010)．

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，句範疇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間の重複が発生して逆効果となり得る．

4 極大フレーズ選択手法と部分構文木選択手法

本節では，提案手法である極大フレーズ選択手法と，部分構文木選択手法，また，それらの組み合わせ手法について説明する．

4.1 極大フレーズ選択手法

本節では，単語 n-gram 頻度に基づくフレーズ選択手法でフレーズ長の制限によって発生する，フレーズ間の重複問題を解消するために，極大部分文字列 (Yamamoto and Church 2001; Okanohara and Tsujii 2009) 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，形式的には**「その部分文字列**

を常に包含するような，より長い部分文字列が存在しない」という性質を持った部分文字列として定義される．この極大部分文字列の定義は，文字列を任意の要素列に読み替えて，極大部分要素列とすることができる．

極大部分要素列は下記のような半順序関係の定義を用いて示すことができる．

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

ここで $s_1,\ s_2,\ \alpha,\ \beta$ 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，

$$\begin{aligned} p_1 &= \text{“one of the preceding”}, & occ(p_1) &= 200,000 \\ p_2 &= \text{“one of the preceding claims”}, & occ(p_2) &= 200,000 \\ p_3 &= \text{“any one of the preceding claims”}, & occ(p_3) &= 190,000 \end{aligned}$$

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，$p_2 = \alpha p_1 \beta$，$\alpha =$ “”，$\beta =$ “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 p について $p \preceq q$ となるような q が p 自体を除いて存在しない場合に，p は**極大性** [6] 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，$p_2 \preceq q$ となるような q 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．

原言語コーパス中のすべての極大フレーズは拡張接尾辞配列 (Kasai, Lee, Arimura, Arikawa, and Park 2001) を用いて原言語コーパスの単語数 N に対して線形時間 $O(N)$ で効率的に列挙可能であるが, 出現頻度を同時に得るためには二分探索のためにそれぞれ $O(logN)$ 回の文字列比較が必要なため，合計 $(NlogN)$ 回の文字列比較が必要となる (Okanohara and Tsujii 2009)．列挙されうる極大フレーズの数は高々 $N-1$ 個であるが, 頻度で降順に列挙するためには $O(NlogN)$ のソートアルゴリズムを用いることができる．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．これは，原言語コーパス中のほとんどの文を含めた膨大な部分単語列が出現頻度 1 の極大フレーズとして選択されることを防止するためである．

極大フレーズのみを人手翻訳の対象とし，翻訳済みデータに出現していない最高頻度の極大フレーズを順次選択する手法を**極大フレーズ選択手法**として提案する．本提案手法には 2 つの

[6] 極大性 (maximality) とは，代数学の用語であり，半順序関係 $\preceq$ と集合 S とその元 $x \in S$ について，$x \preceq y$ となるような $y \in S, y \neq x$ が存在しない場合に，x は S の極大元であると言う．

利点があると考えられる．1 つ目の利点は，互いに重複するような複数のフレーズを 1 つの極大フレーズにまとめ上げて翻訳対象とすることで，1 度の人手翻訳で複数の高頻度フレーズを同時にカバーすることが可能となり，翻訳コスト減少による能動学習効率の向上が見込めることである．2 つ目の利点は，既存手法でフレーズ長が 4 単語などの固定長に制限される問題を解消できることである．

ただし，先述の例で述べたが，p_2 は p_3 の一部であり，二者の出現頻度も近いが一致はしておらず，そのため二者とも極大フレーズとなる．実際の用途を考慮すると，このように出現頻度が完全に一致していなくてもほとんどの場合に重複して出現するフレーズは統合することが望ましいが，すべての極大フレーズをそのまま翻訳候補とする実装では重複を取り除けない場合がある．そこで，式 (1) の制約をパラメータ λ で緩和して，より一般化した半順序関係を下記のように定義する．

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

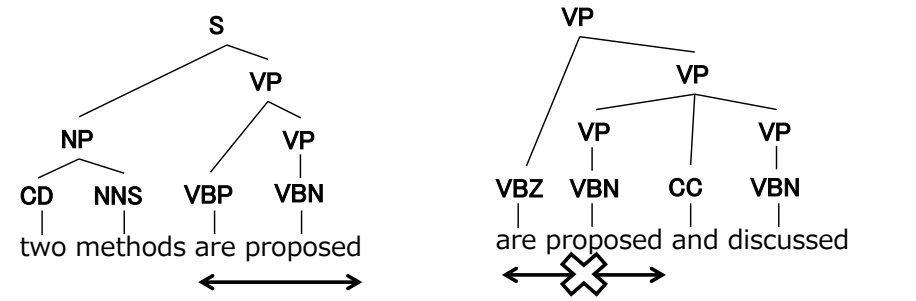
ここで，λ は 0 から 1 の間の実数値を取る．この半順序 $\overset{*}{\preceq}$ を用いた場合にも極大性を定義可能であり，通常の極大フレーズ (以下，**標準極大フレーズ**とする) と区別するため **λ-極大フレーズ**と呼ぶことにし，このような特徴を持つフレーズを列挙し，未カバーフレーズを頻度順に追加する手法を **λ-極大フレーズ選択手法**として併せて提案する．

λ-極大フレーズ選択手法のパラメータ λ を 1 より小さく設定することで，2 つの重複するフレーズの一致条件を取り除き，近似する出現頻度を許容するようになる．特殊な場合として $\lambda = 1 - \epsilon$ のときには標準極大フレーズ選択手法と同一であり (ϵ は正の極小値)，$\lambda = 0$ のときには部分的アノテーションを行わない，文の乱択手法となる．両者の利点を両立できる可能性を考慮して，本研究では特に中間の値となる $\lambda = 0.5$ を用いた際の影響を他の手法と比較に用いている．

$\lambda < 1$ における λ-極大フレーズは，常に標準極大フレーズの条件を満たすため，原言語コーパス中の λ-極大フレーズの候補は，すべての標準極大フレーズの中から探せばよい．標準極大フレーズは $O(N)$ 時間で列挙可能であることは先に述べた通りであるが，これは接尾辞配列に対応する接尾辞木の内部ノードをたどりながら列挙する．この時に，極大フレーズ p に対応するノードから祖先ノードをたどっていき，対応する祖先ノードのフレーズ p_1 が $\lambda \cdot occ(p_1) < occ(p)$ である場合，p_1 は λ-極大フレーズの条件を満たさないため除外できる．接尾辞木のすべての内部ノードについて，このような処理を行うことで，除外されなかったフレーズは λ-極大であり，高々 $N-1$ 個の内部ノードに対し，$O(logN)$ 回の文字列比較で出現頻度比較を行い，根ノードから帰りがけ順で処理を行えば $O(NlogN)$ 回の文字列比較で λ-極大フレーズを列挙できる．また，標準極大フレーズと同様に，λ-極大フレーズとその出現頻度は，$O(NlogN)$ 時間で頻度順

に列挙可能である．

4.2 部分構文木選択手法



(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

本節では 4.1 節で述べた提案手法とは別に，原言語コーパスの句構造解析結果に基くフレーズ選択手法を提案する．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．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，句範疇が分断されるような問題は発生せず，選択されるフレーズは構文的にまとまった意味を持つと考えられる．本研究では選択されるフレーズが能動学習効率に与える影響の調査を目的とするため，他の手法と比較しやすいように構文木をフレーズ抽出のみに用いており，そのため異なる構造の部分木であっても単語列が一致している場合には同一のものとしてカウントする．

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質の高い正解データが得られるものと期待できる．

n-gram 頻度や極大フレーズの選択手法では，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 and discussed” の一部である “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，単語 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能動学習効率に悪影響を及ぼす可能性がある．従って，4.1 節で提案した λ-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 (λ-**極大部分構文木選択手法**).

5 シミュレーション実験

5.1 実験設定

4 節で提案したフレーズ選択手法が，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，本研究では先ず，逐次的にフレーズの対訳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，高精度な句構造解析器を利用可能な英語を原言語とし，目的言語にはフランス語と日本語を選択した．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，より現実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．英仏翻訳には，WMT2014[7] 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 [8] (Koehn 2005) 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[9] (Tiedemann 2009)，PatTR[10] (Wäschle and Riezler 2012)，Wikipedia タイトルを合わせて追加コーパスとした．英日翻訳には，日常的な英語表現を広くカバーする英辞郎例文データ [11] 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 ASPEC[12] (Nakazawa, Yaguchi, Uchimoto, Utiyama, Sumita, Kurohashi, and Isahara 2016) を追加の対訳コーパスとして用いた．前処理として，日本語コーパスの単語分割には KyTea(Neubig, Nakata, and Mori 2011) を用いており，句構造解析と単語アラインメント推定の精度を確保するため，学習用対訳コーパスのうち，単語数が 60 を超える文の対訳は取り除いた．前処理後の対訳データの内訳を表 1 にまとめる．

本実験では，逐次的なデータの追加とモデルの再学習を行うものの，各ステップで 1 フレーズずつ追加するのでは数十万フレーズ以上ある翻訳候補すべての影響を現実的な時間で評価できないと判断したため，ステップ毎の追加フレーズ数は次式に従い可変とした [13]．

$$\#\text{additional\_phrases} = \left\lfloor \frac{\#\text{accmulated\_additional\_phrases}}{10} \right\rfloor + 1 \tag{3}$$

翻訳の枠組みには，フレーズベース機械翻訳 (Koehn, Och, and Marcu 2003) を用い，Moses ツールキット [14](Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, and Herbst 2007) を利用して翻訳モデルの学習やデコードを行った．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

[7] http://statmt.org/wmt14/
[8] http://www.statmt.org/europarl/
[9] http://opus.lingfil.uu.se/EMEA.php
[10] http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/
[11] http://eijiro.jp
[12] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[13] 例として，追加された累積フレーズ数は能動学習開始から 0, 1, 2, …, 9, 10, 12, 14, …, 20, 23, … と変化する．
[14] http://www.statmt.org/moses

言語対	分野	データセット	文数/単語数
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	1,000 文
		Dev	500 文
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	1,790 文
		Dev	1,790 文

表 1 対訳コーパスのデータ内訳 (有効数字 3 桁)

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++(Och and Ney 2003) を逐次学習に対応させた inc-giza-pp[15] 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 (Germann 2014)) 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 (Heafield 2011) を用いて，ベースコーパスと追加コーパスの全学習用データから $n = 5$ 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 (Och 2003) を用いたが，フレーズ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 (Papineni et al. 2002) 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．能動学習に用いるフレーズ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．

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

[15] https://github.com/akivajp/inc-giza-pp/

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 (ベースライン，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 (ベースライン，3.2 節)

標準極大フレーズ選択手法 (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標準極大フレーズを順次選択 (提案手法，4.1 節)

λ-極大フレーズ選択手法 (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 λ-極大フレーズ ($\lambda = 0.5$) を順次選択 (提案手法，4.1 節)

部分構文木選択手法 (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 (提案手法，4.2 節)

λ-極大部分構文木選択手法 (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 λ-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 (提案手法，4.1 節, 4.2 節)

それぞれの手法で選択されたフレーズ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．構文木に基づく手法では，句構造解析を行うために Ckylark[16] (Oda, Neubig, Sakti, Toda, and Nakamura 2015) を使用した [17]．maxsubst-freq や reduced-maxsubst-freq では出現頻度 1 のものを取り除くことを 4.1 節で述べたが，文を含まないフレーズの出現頻度を扱う他の全ての手法においても条件を揃えるために，出現頻度 1 のフレーズは除外した [18]．

言語対	フレーズ選択手法	各時点における BLEU スコア [%]				
		追加なし	1 万単語追加	10 万単語追加	100 万単語追加	全フレーズ追加
En-Fr	sent-rand	25.39	25.57	25.72	27.35	30.02
	4gram-rand		25.53	25.52	27.16	28.32
	sent-by-4gram-freq		25.55	26.12	27.93	30.69
	4gram-freq		25.61	26.16	27.89	28.75
	maxsubst-freq		25.55	25.84	27.49	29.60
	reduced-maxsubst-freq		**25.63**	26.10	27.91	29.81
	struct-freq		**25.85**	**26.86**	**29.06**	30.03
	reduced-struct-freq		† **26.08**	† **27.18**	† **29.40**	30.20
En-Ja	sent-rand	9.37	10.44	13.03	15.58	21.22
	4gram-rand		10.57	13.37	17.61	19.71
	sent-by-4gram-freq		11.14	14.49	17.66	21.06
	4gram-freq		11.49	15.07	18.27	19.74
	maxsubst-freq		**11.72**	**15.13**	**18.58**	19.88
	reduced-maxsubst-freq		**11.87**	† **15.72**	**18.71**	19.59
	struct-freq		**12.02**	**15.44**	**18.61**	19.97
	reduced-struct-freq		† **12.27**	**15.66**	† **18.91**	19.83

表 2　各手法における BLEU スコアの推移 (100 万単語追加直後までの各時点において下線は乱択手法とベースライン手法の中でスコア最大であることを示し，このスコアを上回る提案手法のスコアをボールド体で示す．また，全手法でスコア最大のものには短剣符 † を付記した．)

5.2　実験結果

能動学習効率の比較: シミュレーション実験により得られた結果から，追加単語なし，1 万単語追加直後，10 万単語追加直後，100 万単語追加直後，全フレーズ追加時点における各手法の BLEU スコアの推移を表 2 に示す [19]．全フレーズ追加時点でのスコアは，各手法で選択されうるフレーズの全対訳を用いて学習した翻訳精度であるため，各手法の性能限界と考えられるが，追加される単語数が大きく異なるため能動学習効率という観点では単純比較ができない．

この表から，2 つの乱択手法と 2 つのベースライン手法と比較した場合，10 万単語追加直後までは安定して 4gram-freq での精度の伸びが良く，文全体ではなく高頻度のフレーズのみを選択することによる利点を確認できる．ただし，英仏翻訳における 100 万単語追加時点では，

[16] https://github.com/odashi/Ckylark
[17] 本実験では専門分野のコーパスに対して句構造解析を行ったため，科学用語などの多くは句構造解析モデル中に含まれていないが，Ckylark は単語の並びから未知語に対しても何らかの品詞を推定できるため，本研究では未知語による解析の失敗などは特に考慮していない．
[18] 予備実験により，頻度 1 のフレーズを含めた場合と取り除いた場合のフレーズ数を比較したが，頻度 1 の異なるフレーズは頻度 2 以上のフレーズの等倍以上，maxsubst-freq などでは 10 倍以上列挙された．これら頻度 1 のフレーズは計算資源を圧迫し，カバレッジへの影響も非常に小さいため除外して実験を行うこととした．
[19] 英日翻訳におけるベースシステムの BLEU スコアが 10 を下回る低い値から開始しているが，専門分野の対訳が極端に不足しているためである．先行研究 (Haffari et al. 2009; Bloodgood and Callison-Burch 2010; Ananthakrishnan et al. 2010a; Ananthakrishnan, Prasad, Stallard, and Natarajan 2010b) においても，対訳不足の状態から能動学習によって分野適応する際に共通して BLEU スコアを用いているため，本稿でも同様の評価を行った．

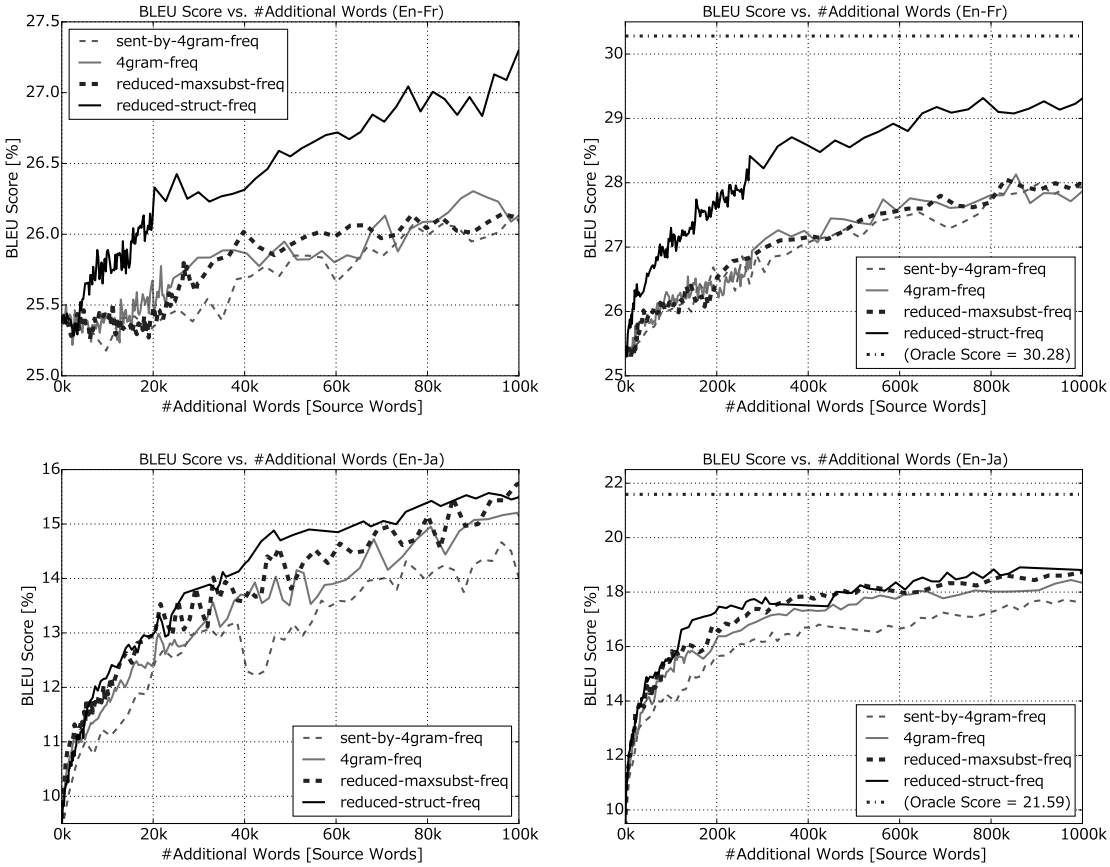


図 3 追加単語数あたりの BLEU スコア (左上: 10 万単語まで追加の英仏翻訳，右上: 100 万単語まで追加の英仏翻訳，左下: 10 万単語まで追加の英日翻訳，右下: 100 万単語まで追加の英日翻訳)

4gram-freq のスコアが sent-by-4gram-freq を下回っており，また，両言語対における全フレーズ追加時点では 4gram-freq のスコアが sent-by-4gram-freq や sent-rand を下回っていることから，一定量以上の単語数を追加する場合には 4gram-freq では文選択手法より能動学習効率が低下し，性能限界も高くないことが分かる．これは，3.2 節でも述べたように，4gram-freq では選択されるフレーズの最大長が 4 単語までに制限されており，より長いフレーズの対応を学習できないことが翻訳する上で不利であるためと考えられる．

次に，提案手法とベースライン手法との比較を行う．提案手法の中では，reduced-maxsubst-freq は maxsubst-freq よりほぼ常に高スコアであり，reduced-struct-freq は struct-freq よりほぼ常に高スコアであったため，λ-極大性を利用して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が向上したと考えられる．このため，λ-極大性を利用する 2 つの提

案手法と 2 つのベースライン手法との，より詳細な比較を行いたい．

図 3 には，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追加単語数と翻訳精度の変化を示す．また，ベースコーパスと追加コーパスの全対訳データを用いて学習・評価したスコアをオラクルスコアとして，右側の 100 万単語追加までのグラフに併せて示す．

reduced-maxsubst-freq は，英日翻訳ではベースライン手法よりも安定して高いスコアであったが，英仏翻訳では 100 万単語追加時点まで 4gram-freq とほぼ同程度のスコアとなった．ただし，両言語対において全フレーズ追加時点でのスコアは 4gram-freq や 4gram-rand を大きく上回っているため性能限界は高く，提案手法では先述のような最大フレーズ長制限の問題が発生しないことが大きな原因と考えられる．また，英仏翻訳において reduced-maxsubst-freq と 4gram-freq で大きな差が見られなかった原因として，両手法で選択された高頻度フレーズを見たところ，最高頻度順に “according to claim” (1,502,455 回)，“claim 1” (1,133,243 回)，“characterized in that” (858,404 回) などと共通のフレーズが選択されており，1 節で述べたような，フレーズ間の重複はあまり発生しておらず，句範疇の断片化の方が目立っていた．複数の高頻度な 4-gram フレーズが共通の部分単語列を多く有するという状態は，特に高頻度の $n > 4$ 単語からなる長い未カバーフレーズが出現する場合であり，本実験で用いた医療文書コーパス中の高頻度の長いフレーズは，専門的な表現をあまり含んでいないために一般分野の大規模コーパス中に既に含まれていたことが原因と考えられる．一方，英日翻訳では，4gram-freq で選択された高頻度フレーズは “results suggest that” (6,352 回)，“these results suggest” (5,115 回)，“these results suggest that” (4,791 回) などのように多くの重複が見られ，reduced-maxsubst-freq では，こういったフレーズを 1 つにまとめられたことが大きいと考えられる．特に，英日翻訳で用いた一般分野のコーパスは訳 40 万文と比較的小規模であり，日常表現をまとめたものであるため長いフレーズはあまりカバーされていなかったことの影響もあるだろう．

reduced-struct-freq は，両言語対においてほぼ安定して最高スコアのフレーズ選択手法であった．英日翻訳における 10 万単語追加時点のみ，reduced-maxsubst-freq が最大スコアとなっているが僅差であり，学習曲線の振れ幅も大きいため誤差の範囲であろう [20]．特に英仏翻訳シミュレーション結果では，最初から reduced-struct-freq や struct-freq 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．一方で，英日翻訳の追加単語数が少ないうちは reduced-maxsubst-freq や 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フレーズ選択手法の精度がほぼ横這いとなった．

[20] ブートストラップ・リサンプリング法 (Koehn 2004) で統計的有意差を検定したところ，$p < 0.1$ の有意さも見られなかった

言語対	フレーズ選択手法	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-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
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

表 3 手法ごとに選択されるフレーズ内訳 (有効数字 3 桁)

選択されたフレーズ長の傾向: 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フレーズ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 3 にまとめる．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，翻訳精度が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．提案手法によって選択されたフレーズ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 3 のフレーズ頻度に基づく手法において，全フレーズ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 万単語追加時点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．

カバレッジの影響: 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 4 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

言語対	フレーズ選択手法	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-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
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

表 4　各フレーズ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），ボールド体は一定の単語数追加時点でのカバレッジ最大値を示す

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．struct-freq や，reduced-struct-freq において，他の手法よりも高い 1-gram カバレッジを得られた理由としては，4.2 節で述べたように木構造に基づくフレーズ選択手法では，表層的な単語列の出現回数ではなく特定の部分木の句範疇をなすフレーズとして出現頻度を数え上げるため，2 単語以上からなるフレーズの頻度は大きく低下し，優先的に高頻度の未カバー 1-gram を選択したことの影響が大きいと考えられる．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，λ-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．

削減された単語数: 3.2 節では，4gram-freq の問題として，選択されるフレーズ間で重複して出現する共通の部分単語列が多いことを挙げ，この問題に対処するために 4.1 節で λ-極大フレーズ選択手法を提案した．表 5 に，4gram-freq で 1 万単語追加直後，10 万単語追加直後，100 万単語追加直後の各時点で選択されたフレーズが，maxsubst-freq や reduced-maxsubst-freq でよ

言語対	フレーズ選択手法	削減された単語数 (削減割合)		
		1 万単語追加	10 万単語追加	100 万単語追加
En-Fr	maxsubst-freq	92 (0.92%)	2,077 (2.11%)	34,917 (3.49%)
	reduced-maxsubst-freq	5,079 (50.79%)	42,622 (42.62%)	378,938 (37.89%)
En-Ja	maxsubst-freq	138 (1.38%)	686 (1.61%)	41,046 (4.10%)
	reduced-maxsubst-freq	2,560 (25.6%)	24,697 (24.70%)	24,697 (24.70%)

表 5 4gram-freq で重複して選択されるフレーズの提案手法による削減量

り長いフレーズに統合されて削減された単語数と割合をまとめる．表から，英仏翻訳においても英日翻訳においても，maxsubst-freq では 1%から 4%程の少量の単語数しか削減できていないが，これは 4.1 節で述べたように，標準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完全一致するという厳しい制約があるため，多くのフレーズが極大フレーズとなったことに起因する．一方，両言語対において，reduced-maxsubst-freq では 24.70%以上，最大で 50.79%の単語が削減された．この結果からも，フレーズの出現頻度の一致条件を緩めることで，包含されたフレーズを効果的により多く削減することができると言えるだろう．

6 人手翻訳実験

6.1 実験設定

前節のシミュレーション実験で得られた結果が，現実の人手翻訳による能動学習を行う際にも有効と言えるかどうかを調査するため，外部委託機関を通じて翻訳作業を依頼し，それによって得られた結果を用いて従来手法との比較評価を行った．特に，翻訳に要した実作業時間や，得られる対訳の自信度評価も能動学習の効果を比較する上で重要である．

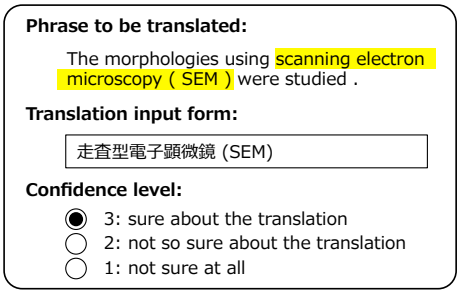


図 4 人手翻訳ユーザーインターフェイスのイメージ

作業手順:

1.「Phrase to be translated」の項目に表示されている英文のうち、**黄色くハイライト表示された部分に対応する日本語訳**を「Translation」の記入欄に入力。文全体がハイライトされている場合は文全体を翻訳する。辞書や Web サイトの情報などを用いてもよい。

2. 入力した日本語訳に対する自信度を「Confidence」に表示されている 3 段階から選択。選択の目安としては本ガイドラインの作業例の項目を参考にすること。

3.「Submit」ボタンを押して翻訳内容を送信。正しく送信された場合には「Workspace」の領域に次の翻訳候補が表示される。日本語訳が未記入であったり、自信度を選択していない場合にはエラーが表示されるので、記入漏れなどを確認すること。

4. 3. で日本語訳が正常に送信されると、フォームの下部にこれまで翻訳を行った（英語側の）単語数やフレーズ数が表示されるので確認する。翻訳済みの単語数が全部合わせて **3 万単語＋α（99 単語以下）**になると翻訳タスクの終了である。（終了条件を満たすと自動的に Workspace の入力フォームが消える。ちょうど 3 万単語にはならないと考えられるため、αは端数の調整用である。）

注意事項:

* 今回、翻訳作業を依頼する翻訳データは、科学論文の概要を元に選択されているため、科学技術関係の専門用語が多く出現します。必要に応じて辞書やオンライン検索などを用いて頂いて問題ありません。

* 上記と同等の理由から、英語の直訳が自然でない可能性があります。たとえば、英語で受動態で書かれている文は日本語で能動態の方が自然だったり、主語の「we」や「this paper」が日本語では訳されない方が自然であったりします。厳格な翻訳規定はありませんが、作業例を参考に日本語で自然になるように柔軟に対応してください。

* 基本的に日本語訳を未記入のまま送信することはできませんが、自信度で「1: not sure at all」を選択した場合のみ、翻訳欄を未記入のまま送信して次の翻訳を行うことができます。どのような場合に 1 を送信すべきかは作業例の項目を参考にして下さい。

作業例:

翻訳対象: In addition, the critical current was estimated from magnetization measurements .

翻訳例: また，磁化測定から臨界電流を推定した。

自信度の選択例:「3: sure about the translation」を選択（専門用語は多いが対訳として問題ないと考えたため）

翻訳対象: The wind velocity can be measured with the accuracy of 1 m/s , and the wind direction with 10Åwithout influence of the running direction of the car.

翻訳例: で測定される

自信度の選択例:「2: not so sure about the translation」を選択 (意味は難しくないが、半端な位置でハイライトされているため、翻訳しにくい)

翻訳対象: Sucrose , glucose and fructose were identified by high performance liquid chromatography (HPLC) .

翻訳例: クロマトグラフィー (HPLC

自信度の選択例:「2: not so sure about the translation」を選択（一応翻訳したが、丸括弧の断片のみが含まれており、また複合語の一部のみ選択されているため、ハイライト部分のみの翻訳は無理がある。）

翻訳対象: Crude-peroxidase was prepared from madake (Phyllastachys bambusoides Sieb . et Zucc .) bamboo shoots .

翻訳例: . et Zucc

自信度の選択例:「1: not sure at all」を選択（「Phyllostachys bambusoides Sieb. et Zucc.」で「マダケ」の学名であるが、一部だけを和訳するのは非常に難しいため、無理に翻訳をしない。）

表 6　翻訳作業ガイドライン

人手翻訳の依頼を行うため，図 4 に示すような作業用ユーザーインターフェイスを持つ Web ページを作成した．翻訳対象のフレーズのみ提示されても翻訳が困難であったり多くの時間が必要となる可能性があるため，Bloodgood らの実験手法 (Bloodgood and Callison-Burch 2010) に従い，翻訳候補のフレーズを含むような文を表示して文脈を明らかにした上で，ハイライトされたフレーズのみを翻訳するよう依頼した．フレーズを含む文は複数存在し得るが，本実験では単純に最も短い文を選択して表示した．各フレーズの翻訳後には，翻訳者がその翻訳結果

にどの程度確証を持てるかという主観的な自信度を 3 段階で評価するよう併せて依頼した．翻訳時の自信度評価は表 6 に掲載した翻訳作業ガイドラインを基準とし，翻訳が困難で自信度評価が 1 の場合のみ，翻訳のスキップを許容した．また，翻訳候補が表示されてから対訳が送信されるまでの時間の記録も行った．

比較評価に用いたフレーズ選択手法には，ベースラインとして従来手法の $n = 4$ における単語 n-gram 頻度に基づく文選択手法 (sent-by-4gram-freq) および単語 n-gram 頻度に基づくフレーズ選択手法 (4gram-freq) の 2 つを，提案手法として，前節のシミュレーション実験で最も高い能動学習効率を示した λ-極大部分構文木選択手法 (reduced-struct-freq) を用いて比較評価を行った．翻訳作業を行ったのは専門の翻訳者 3 名であり，それぞれの手法で 1 万単語以上のフレーズに対する翻訳が得られるよう発注を行った．翻訳者毎の能力や評価の偏りによる影響を小さくするため，毎回異なる手法からフレーズを選択して新しい翻訳対象の表示を行った．

実験に用いたデータやツールは，英日翻訳のシミュレーション実験で用いたものと同一である (5 節)．しかし，人手翻訳によって収集した対訳データは，ベースシステムの学習に用いた対訳データと比較して非常に小規模であり，追加されたフレーズ対が与える影響が非常に小さくなってしまう可能性があるため，フレーズ対を追加する度に，ベースシステムの対訳データと収集した追加データで個別に 5-gram 言語モデルを学習し，開発用データセットにおけるパープレキシティが最小となるよう SRILM (Stolcke 2002) を用いて二者を線形補間で合成して用いた [21]．

6.2 実験結果

能動学習効率: 図 5 のグラフは，本実験で収集した対訳データを用いて翻訳モデルを学習した際の翻訳精度の推移を表している．翻訳者が翻訳をスキップしたフレーズに関しては，追加単語数には含めず，累計作業時間には含めているが，実際にスキップされたフレーズ数は極少数であったため，この影響は小さい．左のグラフから，reduced-struct-freq で，従来手法よりも急激に翻訳精度が向上している様子が分かる．クラウドソーシングのような形で翻訳作業を委託する際には，文章量，特に単語数に応じた予算が必要となることから (Bloodgood and Callison-Burch 2010)，こういった状況では提案手法で高い費用対効果を発揮できることになる．一方，右のグラフから，作業時間あたりの能動学習効率は，4gram-freq を上回ることはなかった．これは，前節でテーブル 4 をもとに議論したように，部分構文木選択手法では，未カバーの 1-gram，即ち未知語を優先的にカバーする傾向があるため，本実験タスクでは科学分野の専門用語が多く，翻訳に多くの時間を要したものと考えられるため，後の議論で詳細な分析を行う．

[21] 言語モデル L における文 e の尤度を $P_L(e)$ とすると，2 つの言語モデル L_1, L_2 を線形補間で合成したモデル L_{1+2} の尤度は $P_{L_{1+2}}(e) = \alpha P_{L_1}(e) + (1-\alpha)P_{L_2}(e)$ となる．α は 0 から 1 の間を取る補完係数であり，開発用データ E_{dev} に対して，$PPL(E_{dev}) = exp\left(\frac{\sum_{e \in E_{dev}} -logP_{L_{1+2}}(e)}{|E_{dev}|}\right)$ が最小となるように調整される．

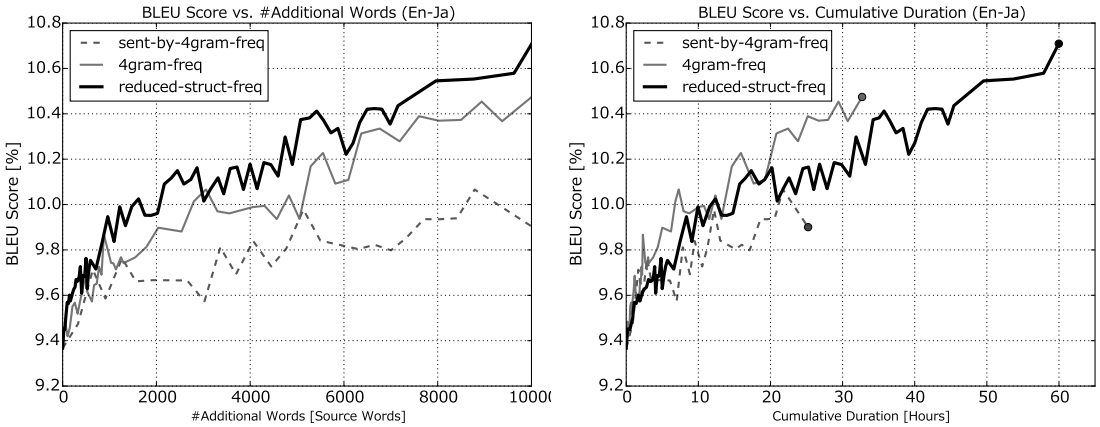


図 5　各手法における追加単語数あたりの BLEU スコア推移（左）と累計作業時間あたりの BLEU スコア推移（右）

手法	合計作業時間 [時間]	平均自信度 (3 段階)	自信度評価の割合			
			スキップ	自信度 1	自信度 2	自信度 3
sent-by-4gram-freq	25.22	2.689	1.77%	0.00%	30.74%	69.08%
4gram-freq	32.70	2.601	0.53%	1.69%	35.48%	62.29%
reduced-struct-freq	59.97	**2.771**	0.52%	1.51%	18.82%	**79.15%**

表 7　合計実作業時間と自信度評価の統計

手法	フレーズ数					
	1 単語	2 単語	3 単語	4 単語	5 単語以上	合計
sent-by-4gram-freq	-	-	-	-	565	566
4gram-freq	1,185	2,061	1,045	390	0	4,681
reduced-struct-freq	4,688	1,038	884	96	38	6,744

表 8　各手法で選択されたフレーズ数の内訳

作業時間と自信度評価: 表 7 に，各手法で 1 万単語をすべて翻訳し終えるのに要した時間と，3 段階で主観評価を行った自信度評価の統計値をまとめる．表 8 は，各手法で選択されたフレーズ数の内訳である．提案手法では，合計作業時間が他の手法の倍近い値になっているが，提案手法では先述のように専門用語を重点的に選択する傾向が確認されており，表 8 からは，4gram-freq と比較していて 4 倍近い数の未知語が選択され，全体的なフレーズ数も多いことが大きな要因

手法	平均作業時間 [秒]				
	1 単語	2 単語	3 単語	4 単語	5 単語以上
sent-by-4gram-freq	-	-	-	-	160.64
4gram-freq	30.14	24.76	21.77	21.12	-
reduced-struct-freq	35.61	25.23	21.72	28.13	22.82

表 9 各手法におけるフレーズの翻訳に要した平均時間

手法	平均自信度評価 (3 段階)				
	1 単語	2 単語	3 単語	4 単語	5 単語以上
sent-by-4gram-freq	-	-	-	-	2.689
4gram-freq	2.885	2.585	2.422	2.300	-
reduced-struct-freq	2.802	2.796	2.778	2.708	2.737

表 10 各手法におけるフレーズ長ごとの平均自信度評価

であろう．一方で，選択されたフレーズの翻訳作業に対する自信度評価は提案手法が最大で，全体の約 79%のフレーズ翻訳作業で最大評価の 3 が選択されており，質の高い対訳を得られたと考えられる．この結果は，句構造を保つようなフレーズが選択されることで，構文的に対応の取れた翻訳を行えた点が大きく影響していると考えられる．

表 9 には，各手法で選択されたフレーズの翻訳に要した平均時間の傾向を示す．手法の内外で翻訳候補のフレーズ長に大きく差があり，単純な比較を行うことができないため，フレーズ長に応じて個別に平均作業時間を求めて比較を行うことにした．この表から，1 単語の翻訳作業に要した平均時間は，2 〜 4 単語からなるフレーズの翻訳よりも長くなるという現象が見られるが，未カバーの単語はほとんどが専門用語であるため，辞書やオンライン検索で慎重に意味を調べる必要性を考えれば納得できる．また，これらは 1 フレーズの翻訳に要した平均時間であるため，1 単語の翻訳に要する時間コストとして換算した場合，1 単語フレーズの翻訳時間は 2 単語フレーズの単語翻訳時間の倍以上であり，専門用語の翻訳に要するコストがいかに大きいかが分かる．

各手法におけるフレーズ長ごとの平均自信度評価を表 10 に示す．この表から，提案手法では 1 単語の翻訳時の平均自信度はベースライン手法よりも低くなっているのが分かるが，ベースライン手法では 1 単語のみ選択されることが少なく，提案手法では多くの専門用語が選択されたことが原因と考えられる．一方で，従来手法ではフレーズ長が長くなるほど劇的に自信度が下がる傾向が見られるが，提案手法においては長いフレーズに対しても安定して高い自信度が得られており，構文的に整ったフレーズを選択する手法の有効性が如実に現れている．専門用

手法	BLEU スコア [%]		
	自信度 1 以上 (全フレーズ)	自信度 2 以上	自信度 3
sent-by-4gram-freq	9.88	9.92	9.85
4gram-freq	10.48	10.54	10.36
reduced-struct-freq	10.70	**10.72**	10.67

表 11　保証値以上の自信度を持つフレーズ対のみを学習に用いた場合の翻訳精度

語の対訳を得るには時間がかかるが，調べれば対応する訳語を得られる可能性も高いため，対訳の自信度が高くなる傾向も見られた．

自信度帯による翻訳精度: 各手法によって得られたフレーズ対をすべて学習に用いた場合の翻訳精度を表 11 に示す．また，それぞれのフレーズ対に自信度評価が記録されているため，最低保証値を定めて全フレーズ対のうち自信度が 2 以上や 3 のフレーズ対のみを学習に用いた翻訳モデルの評価も行った．その結果，どの手法においても自信度 1 の対訳を除去して 2 以上のフレーズ対のみを用いた場合の方が，全フレーズ対を用いる場合よりも翻訳精度の向上が見られた．一方，自信度 3 の対訳のみを用いる場合は精度がかえって減少したが，これは大幅に対訳データを削ってしまうことによる悪影響であろう．追加データ無しのベースシステムでは BLEU スコアが約 9.37% であったが，提案手法によって収集した 1 万単語分の追加データのうち自信度 2 以上のものを用いて翻訳モデルを学習することで，BLEU スコアは約 10.72% となり，約 1.35 ポイントの翻訳精度向上を達成することができた．

7　まとめ

本研究では，機械翻訳のための能動学習における，新しいフレーズ選択手法として，フレーズの極大性を導入し，それをパラメータ λ で一般化して頻度順に追加する λ-極大フレーズ選択手法と，句構造解析結果から部分木のみを頻度順に選択する部分構文木選択手法，およびそれらの組み合わせである λ-極大部分構文木選択手法を提案した．提案手法の有効性を調査するため，先ず，人手翻訳によるアノテーション作業を擬似的に SMT で行うシミュレーション実験を実施したところ，冗長に選択されるフレーズを λ-極大性に基づいて削減することで，従来より少ない追加単語数で精度向上を達成することができた．また，λ-極大部分構文木選択手法が実際の人手翻訳に与える影響を調査するため，人手翻訳実験を実施したところ，翻訳精度と翻訳者の自信度評価のどちらにおいても従来手法より高い結果が得られた．

しかし，今回用いた手法では専門用語が重点的に選択されるため，従来のフレーズ選択手法よりも長い翻訳時間を要することも示された．そのため，翻訳時間を短縮しつつ，有効にモデ

ルを高度化させられるような能動学習手法を考案することが今後の課題である．具体的には，未知語の獲得手法 (Daumé III and Jagarlamudi 2011) や，時間効率を最適化するフレーズ選択手法 (Sperber et al. 2014) が改善の手がかりとなり得る．また，フレーズ中の部分フレーズを “one of the preceding X” のように変数でテンプレート化する手法 (Chiang 2007) を用いて能動学習に利用する手法の検討も興味深いと考えている．

謝　辞

本研究は，(株)ATR-Trek の助成を受け実施されたものです．また，(株) バオバブには人手翻訳実験のための翻訳作業を支援して頂きました．

参考文献

Ananthakrishnan, S., Prasad, R., Stallard, D., and Natarajan, P. (2010a). “A Semi-Supervised Batch-Mode Active Learning Strategy for Improved Statistical Machine Translation.” In *Proc. CoNLL*, pp. 126–134.

Ananthakrishnan, S., Prasad, R., Stallard, D., and Natarajan, P. (2010b). “Discriminative Sample Selection for Statistical Machine Translation.” In *Proc. EMNLP*, pp. 626–635.

Bloodgood, M. and Callison-Burch, C. (2010). “Bucking the Trend: Large-Scale Cost-Focused Active Learning for Statistical Machine Translation.” In *Proc. ACL*, pp. 854–864.

Brown, P. F., Pietra, V. J., Pietra, S. A. D., and Mercer, R. L. (1993). “The Mathematics of Statistical Machine Translation: Parameter Estimation.” *Computational Linguistics*, **19**, pp. 263–312.

Chiang, D. (2007). “Hierarchical Phrase-Based Translation.” *Computational Linguistics*, **33** (2), pp. 201–228.

Daumé III, H. and Jagarlamudi, J. (2011). “Domain adaptation for machine translation by mining unseen words.” In *Proc. ACL*, pp. 407–412.

Eck, M., Vogel, S., and Waibel, A. (2005). “Low Cost Portability for Statistical Machine Translation based in N-gram Frequency and TF-IDF.” In *Proc. IWSLT*, pp. 61–67.

Germann, U. (2014). “Dynamic phrase tables for machine translation in an interactive post-editing scenario.” In *Proc. AMTA 2014 Workshop on Interactive and Adaptive Machine Translation*, pp. 20–31.

González-Rubio, J., Ortiz-Martínez, D., and Casacuberta, F. (2012). “Active learning for interactive machine translation.” In *Proc. EACL*, pp. 245–254.

Green, S., Wang, S. I., Chuang, J., Heer, J., Schuster, S., and Manning, C. D. (2014). "Human Effort and Machine Learnability in Computer Aided Translation." In *Proc. EMNLP*, pp. 1225–1236.

Haffari, G., Roy, M., and Sarkar, A. (2009). "Active Learning for Statistical Phrase-based Machine Translation." In *Proc. NAACL*, pp. 415–423.

Haffari, G. and Sarkar, A. (2009). "Active Learning for Multilingual Statistical Machine Translation." In *Proc. ACL*, pp. 181–189.

Heafield, K. (2011). "KenLM: Faster and Smaller Language Model Queries." In *Proc, WMT*, pp. 187–197.

Kasai, T., Lee, G., Arimura, H., Arikawa, S., and Park, K. (2001). "Linear-Time Longest-Common-Prefix Computation in Suffix Arrays and Its Applications." In *Proc. CPM*, pp. 181–192.

Koehn, P. (2004). "Statistical Significance Tests for Machine Translation Evaluation." In Lin, D. and Wu, D. (Eds.), *Proc. EMNLP*, pp. 388–395.

Koehn, P. (2005). "Europarl: A parallel corpus for statistical machine translation." In *MT summit*, Vol. 5, pp. 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proc. ACL*, pp. 177–180.

Koehn, P., Och, F. J., and Marcu, D. (2003). "Statistical Phrase-Based Translation." In *Proc. NAACL*, pp. 48–54.

Miura, A., Neubig, G., Paul, M., and Nakamura, S. (2016). "Selecting Syntactic, Non-redundant Segments in Active Learning for Machine Translation." In *Proc. NAACL*, pp. 20–29.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). "ASPEC: Asian Scientific Paper Excerpt Corpus." In *Proc. LREC*, pp. 2204–2208.

Neubig, G., Nakata, Y., and Mori, S. (2011). "Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis." In *Proc. ACL*, pp. 529–533.

Och, F. J. (2003). "Minimum Error Rate Training in Statistical Machine Translation." In *Proc. ACL*, pp. 160–167.

Och, F. J. and Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics*, **29** (1), pp. 19–51.

Oda, Y., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2015). "Ckylark: A More Robust PCFG-LA Parser." In *Proc. NAACL*, pp. 41–45.

Okanohara, D. and Tsujii, J. (2009). "Text Categorization with All Substring Features." In *Proc.*

SDM, pp. 838–846.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation." In *Proc. ACL*, pp. 311–318.

Settles, B. and Craven, M. (2008). "An Analysis of Active Learning Strategies for Sequence Labeling Tasks." In *Proc. EMNLP*, pp. 1070–1079.

Sperber, M., Simantzik, M., Neubig, G., Nakamura, S., and Waibel, A. (2014). "Segmentation for Efficient Supervised Language Annotation with an Explicit Cost-Utility Tradeoff." *TACL*, **2**, pp. 169–180.

Stolcke, A. (2002). "SRILM - an extensible language modeling toolkit." In *Proc. ICSLP*, pp. 901–904.

Tiedemann, J. (2009). "News from OPUS-A collection of multilingual parallel corpora with tools and interfaces." In *Proc. RANLP*, Vol. 5, pp. 237–248.

Tomanek, K. and Hahn, U. (2009). "Semi-Supervised Active Learning for Sequence Labeling." In *Proc. ACL*, pp. 1039–1047.

Turchi, M., De Bie, T., and Cristianini, N. (2008). "Learning Performance of a Machine Translation System: a Statistical and Computational Analysis." In *Proc. WMT*, pp. 35–43.

Wäschle, K. and Riezler, S. (2012). "Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus." *Multidisciplinary Information Retrieval*, pp. 12–27.

Yamamoto, M. and Church, K. W. (2001). "Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus." *Computational Linguistics*, **27** (1), pp. 1–30.

三浦明波，Neubig Graham，Paul Michael，中村哲 (2015). 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. 情報処理学会 第 224 回自然言語処理研究会 (SIG-NL), 19 号, pp. 1–7.

三浦明波，Neubig Graham，Paul Michael，中村哲 (2016). 構文情報に基づく機械翻訳のための能動学習手法と人手翻訳による評価. 言語処理学会 第 22 回年次大会 (NLP2016), pp. 605–608.

略歴

三浦 明波：2013 年イスラエル国テクニオン・イスラエル工科大学コンピュータ・サイエンス専攻卒業．2016 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，同大学院博士後期課程在学．機械翻訳，自然言語処理に関する研究に従事．情報処理学会，言語処理学会，ACL 各会員．

Graham Neubig: 2005 年米国イリノイ大学アーバナ・シャンペーン校工学部コンピュータ・サイエンス専攻卒業．2010 年京都大学大学院情報学研究科修士課程修了．2012 年同大学院博士後期課程修了．2012〜2016 年奈良先端科学

技術大学院大学助教．現在，カーネギーメロン大学言語技術研究所助教，奈良先端科学技術大学院大学客員准教授．機械翻訳，自然言語処理に関する研究に従事．

Michael Paul: 1988 年 ドイツサーランド大学コンピュータ・サイエンス専攻卒業．2006 年 神戸大学工学博士．1995 年 ATR 音声翻訳通信研究所研究員，2000 年 ATR 音声言語コミュニケーション研究所主任研究員，2006 年 (独) 情報通信研究機構研究センター主任研究員．2013 年 株式会社 ATR-Trek．音声翻訳，自然言語処理に関する研究・ビジネスソリューション開発に従事．第 58 回前島密賞受賞，アジア太平洋機械翻訳協会 (AAMT) 長尾賞受賞．

中村 哲：1981 年京都工芸繊維大学工芸学部電子工学科卒業．京都大学工学博士．シャープ株式会社．奈良先端科学技術大学院大学助教授，2000 年 ATR 音声言語コミュニケーション研究所室長，所長，2006 年 (独) 情報通信研究機構研究センター長，けいはんな研究所長などを経て，現在，奈良先端科学技術大学院大学教授．ATR フェロー．カールスルーエ大学客員教授．音声翻訳，音声対話，自然言語処理の研究に従事．情報処理学会喜安記念業績賞，総務大臣表彰， 文部科学大臣表彰，Antonio Zampoli 賞受賞．ISCA 理事、IEEE SLTC 委員，IEEE フェロー．