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

三浦 明波 †,a)　Graham Neubig†,b)　Michael Paul‡,c)　中村 哲 †,d)
† 奈良先端科学技術大学院大学 情報科学研究科
‡ 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

1　はじめに

統計的機械翻訳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 [1]．特に，質の高い対訳データを得るためには人手翻訳の作業が必要となるが，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．アノテーション付きデータを作成する際，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が学習可能である [2, 3, 4]．

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]．高頻度の未カバー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む傾向があり，カバー済みのフレーズ長だけ余分な翻訳コストを要する欠点がある．

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 [4]．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る傾向があり，費用対効果に優れている．しかし，この手法では最大フレーズ長が $n = 4$ などに制限されるため，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．

この問題を解決するため，我々は n-gram 頻度のような表層的な単語列の数え上げの代わりに，構文解析結果を用いて各部分木からなるフレーズを出現頻度順に選択することで，句構造の断片化を防ぐ手法を提案している [5]．選択されたフレーズの機械翻訳結果として得られた擬似対訳を追加の学習データに用いるシミュレーション実験により，構文情報に基づくフレーズ選択手法で，n-gram 頻度に基づく文選択手法やフレーズ選択手法よりも少ない追加単語数で高い翻訳精度が得られた．しかし，現実の人手翻訳を用いた評価が行われていないため，構文情報を用いることが人手翻訳においてどのような影響を与えるかは明らかにされていない．

本研究では，機械翻訳のための能動学習手法において，構文情報に基づくフレーズ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，収集したデータを用いて実験と分析を行った．これにより，従来のフレーズ選択手法と比較して，構文情報を用いることによって翻訳者はより自信を持って翻訳作業を行うことができ，翻訳モデルの再学習では，従来手法によって得られた対訳を用いるよりも高い翻訳精度を達成することができた．

2　n-gram 頻度に基づく選択手法

本節では，従来の 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．

2.1　n-gram 頻度に基づく文選択

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 フレーズが効率的にカバーされるため，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翻訳精度を達成できたと報告している [4].

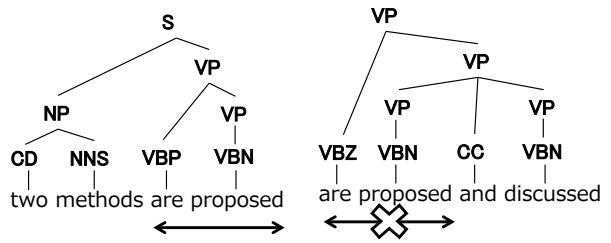
しかし，文全体を選択するため，翻訳済みのデータに既にカバーされているフレーズも多く含んでおり，重複部分の単語数だけ余分な翻訳コストを要すると考えられる．そのため，文全体ではなく高頻出のフレーズのみを選択する手法を次に紹介する．

2.2 n-gram 頻度に基づくフレーズ選択

n-gram 頻度に基づくフレーズ選択手法では，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，翻訳コストの削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，クラウド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 [4].

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の重複が多い問題や，複合句の断片が選択される問題が発生する．

3 構文木に基づくフレーズ選択手法



(a) “are proposed” がカウントされる　(b) “are proposed” はカウントされない

図 1: 構文木に基づく手法のフレーズカウント条件

本節では，構文木に基づくフレーズ選択手法について紹介する．本提案手法では図 1 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，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，より質の高い対訳データが得られるものと期待できる．

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．そこで，下記に定義するような半順序関係を用いることで，重複して選択されるフレーズの削減を行っている．

$$s_1 \overset{*}{\preceq} s_2 \Leftrightarrow \exists \alpha, \beta : s_1 = \alpha s_2 \beta \wedge \lambda \cdot occ(s_1) < occ(s_2). \quad (1)$$

ここで s_1, s_2, α, β は長さ 0 以上の単語列であり，$\exists \alpha, \beta : s_1 = \alpha s_2 \beta$ は s_1 が s_2 の部分単語列であることを表している．$occ(\cdot)$ は文書中の単語列の出現回数である．λ は 0 から 1 の間の実数値を取るパラメータであり，0 に近いほど重複するフレーズを積極的に除外することになる．この半順序関係を用いて，$s_1 \overset{*}{\preceq} s_2$ となるような 2 つのフレーズ s_1, s_2 が存在する場合には s_1 は翻訳候補から除外する．シミュレーション実験では $\lambda = 1$ の場合よりも $\lambda = 0.5$ の場合の方が高い精度向上を得られており [5]，本研究でも $\lambda = 0.5$ で固定して用いる．

4 実験的評価

4.1 実験設定

構文情報を用いたフレーズ選択手法が人手翻訳と能動学習の効率に与える影響を調査するため，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，実作業時間や，得られる対訳の信頼度評価も能動学習の効果を比較する上で重要である．

本実験では，能動学習によって英日翻訳モデルの高精度化を目指す．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 [1] をベースの対訳コーパスとしてベースシステムを構築し，科学論文の概要を元に抽出された ASPEC[2] を追加の対訳コーパスとして能動学習を行う．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

[1] http://eijiro.jp
[2] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

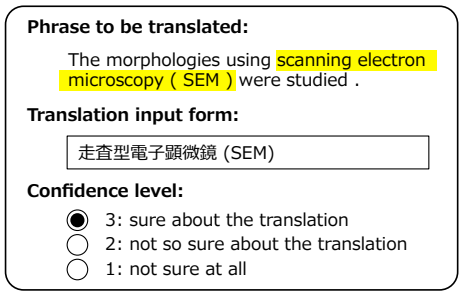


図 2: 人手翻訳ユーザーインターフェイスのイメージ

ングデータのうち，単語数が 60 を超える行は取り除いた．

人手翻訳の依頼を行うため，図 2 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 UI 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 [4] 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．各フレーズの翻訳後には，作業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．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．

比較評価に用いたデータ選択手法は以下の 3 つである．

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 (2.1 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 (2.2 節)

構文木に基づくフレーズ選択 (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フレーズを $\lambda = 0.5$ で重複削減し，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 (3 節)

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データを選択して新しい翻訳対象の表示を行った．

翻訳の枠組みには，フレーズベース翻訳 [6] を用いた．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 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，メモリ上で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，目的言語側のベースデータと追加対訳データを個別に用いて学習した 2 つの n-gram 言語モデル ($n = 5$) を，SRILM を用いて線形補間で合成した．構文木に基づく手法では，句構造解析を行うために Ckylark を使用した．

手法	合計作業時間 [時間]	平均信頼度 [3 段階]
sent-by-4gram-freq	25.22	2.689
4gram-freq	32.70	2.601
reduced-struct-freq	59.97	**2.771**

表 1: 合計実作業時間と平均信頼度

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 2: 各手法におけるフレーズの翻訳に要した平均時間

4.2 実験結果

学習効率: 図 3 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．図 3 左のグラフから，構文情報に基づく提案手法で，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．一方，右のグラフから，作業時間あたりの翻訳精度の向上効率は，4-gram 頻度に基づくフレーズ選択手法を上回ることはなかった．これは構文木に基づくフレーズ選択手法では，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 [5]，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．

作業時間と信頼度評価: 表 1 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信頼度の平均値および標準偏差をまとめる．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一方で，選択されたフレーズの翻訳作業に対する信頼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．これはやはり，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 2 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．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

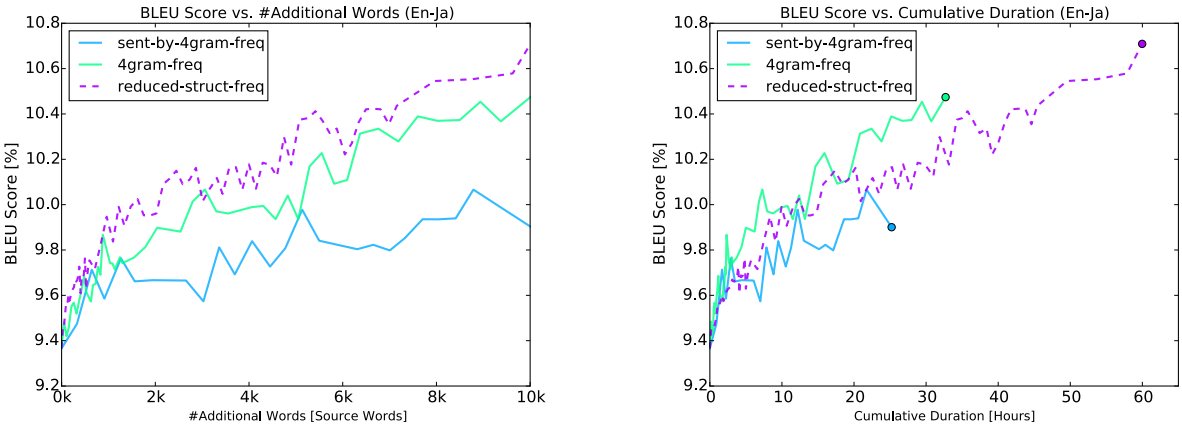


図 3: 各手法における追加単語数あたりの BLEU スコア推移（左）と累計作業時間あたりの BLEU スコア推移（右）

手法	BLEU スコア [%]		
	信頼度 1+ (All)	信頼度 2+	信頼度 3
sent-by-4gram-freq	9.88	9.92	9.37
4gram-freq	10.48	10.54	10.36
reduced-struct-freq	10.70	**10.72**	10.67

表 3: 保証値以上の信頼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

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．

信頼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訳をすべて学習に用いた場合の翻訳精度を表 3 に示す．また，それぞれのフレーズ対に信頼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信頼度が 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信頼度 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，信頼度 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，提案法によって収集した 1 万単語分の追加データのうち信頼度 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% の翻訳精度向上を達成することができた．

5 おわりに

本研究では機械翻訳のための能動学習において，構文情報に基づくフレーズ選択手法が人手翻訳に与える影響を調査するための実験を行った．その結果，構文情報を用いることによって，他の手法よりも質の高い対訳を得ることができ，少ない単語数で翻訳精度の向上を実現できた．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の考案を今後の課題として考えている．

謝辞

本研究は，(株)ATR-Trek の助成を受け実施したものである．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いた．

参考文献

[1] Christopher Dyer, Aaron Cordova, Alex Mont, and Jimmy Lin. Fast, easy, and cheap: construction of statistical machine translation models with MapReduce. In *Proc. WMT*, pp. 199–207, 2008.

[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, Maxim Roy, and Anoop Sarkar. Active Learning for Statistical Phrase-based Machine Translation. In *Proc. ACL*, pp. 415–423, June 2009.

[4] 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.

[5] 三浦明波, Graham Neubig, Michael Paul, 中村哲. 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), 2015.

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