

木構造を中間表現とするピボット翻訳手法

三浦 明波^{1,a)} Graham Neubig^{1,b)} 中村 哲^{1,c)}

概要：統計的機械翻訳において、特定の言語対で十分な量の対訳コーパスが得られない場合、中間言語を用いたピボット翻訳が有効な手法の一つである。ピボット翻訳手法の中でも、特に中間言語を介する2つの翻訳モデルを合成するテーブル合成手法で、高い翻訳精度を得られることが知られている。しかし、従来のテーブル合成手法では、フレーズペアの中間言語側の単語列一致をもとに新しい翻訳規則を推定するが、多品詞語などの影響で曖昧性が高く、誤った規則に高い翻訳確率が推定されてしまう問題がある。そこで、本研究ではフレーズの構文構造が一致する場合のみフレーズペアを結び付けるという条件を導入し、曖昧性解消に取り組む。このようにして得られる翻訳モデルは高信頼度だがカバレッジを損なうことが考えられるため、従来モデルも併用することで信頼度とカバレッジを両立する手法も提案する。本手法の有効性を評価するため、国連文書コーパスを用いた多言語翻訳の実験で精度の比較を行った。

1. はじめに

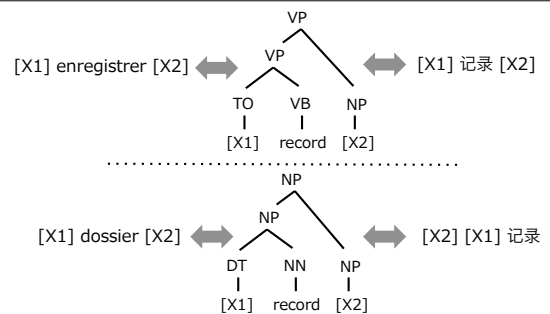
統計的機械翻訳 (Statistical Machine Translation: SMT[1]) では、学習に用いる対訳コーパスが大規模になるほど、高精度な訳出結果を得られることが知られている [2]。一方、英語を含まない言語対などを考慮すれば、多くの言語対において、大規模な対訳コーパスを取得することは困難と言える。特定の言語対で十分な量の対訳コーパスが得られない場合、中間言語を用いたピボット翻訳が有効な手段の一つである [3]。

中間言語を用いる手法も様々なものが考案されている [3], [4], [5] が、特に原言語・中間言語、中間言語・目的言語の2つの翻訳モデルを合成し、新しく得られた原言語・目的言語の翻訳モデルによって翻訳を行うテーブル合成手法で、高い翻訳精度を得られることが示されている [6]。ところが、語義曖昧性や言語間の用語法の差異により、原言語・目的言語間のフレーズ対応を正確に推定することは困難である。

テーブル合成手法では、原言語・中間言語および中間言語・目的言語それぞれの言語対における記号列の対応から、中間言語側で記号列の一致する対応を結び付ける。図1 (a) に、表層的な記号列の一致によってフレーズ対応を結び付ける従来のテーブル合成手法の例を示す。英単語 “record” は文脈に応じて複数の品詞を取りうるため、この語を含む



(a) 従来の単語列を軸とした合成



(b) 提案手法による部分木を軸とした合成

図1 部分構文木の一致による曖昧性解消の例 (仏-英-中), [X1] や [X2] は任意のフレーズを導出可能な非終端記号

単語列も動詞句や名詞句など複数の候補があり、文脈の情報が与えられなければ一意に解釈することはできない。例えば、フランス語の動詞 “enregistrer” は英語の動詞 “record” に結びつくが、フランス語の名詞 “dossier” は名詞としての “record” に結び付く。一方で、中国語は品詞に応じた活用をしない言語であるため、“record” が異なる品詞として用いられる文脈でも、中国語の単語 “记录” が対応するが、語順によって構文的な役割が変化するため、本来それぞれ “[X1] enregistrer [X2]” は “[X1] 记录 [X2]” と、“[X1] dossier [X2]” は “[X2] [X1] 记录” と結び付くべきフレーズであっても、対応する英語フレーズ “[X1] record [X2]” が共通しているため、“[X1] enregistrer [X2]” と “[X2] [X1]

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

a) miura.akiba.lr9@is.naist.jp

b) neubig@is.naist.jp

c) s-nakamura@is.naist.jp

記録”のように対応しないはずのフレーズの組み合わせも学習されてしまう。このように、表層的な記号列の一致条件では誤った対応を結び付けてしまうことも多く、高い翻訳スコアが推定された場合には誤訳の原因となりやすい。

前述の問題に対しては、事前に句構造タグのような統語的情報が与えられている場合には曖昧性解消が容易になると考えられる。図 1 (b) に、提案手法によって中間言語フレーズの構文木の一致によってフレーズ対応を結び付ける例を示す。これにより、英語側の記号列 “[X1] record [X2]” が動詞句として用いられる文脈ではフランス語でも同様の役割を持つ記号列が対応し、名詞句の場合とは区別されて適切な語彙選択を行いやすいと考えられる。また、中間言語側でも文法的役割を一致させることによって誤った対応が結び付かず、結果的に単語並び替えの問題も改善されることが期待できる。

このような構文構造の一致条件を用いるピボット翻訳手法を実現するには、フレーズの対応のみを翻訳規則に用いる翻訳手法ではなく、言語対の片側で構文解析結果を用いて部分構文木とフレーズの対応を翻訳規則に用いる翻訳手法の 2 つの翻訳モデルから、部分構文木を軸にして 1 つの翻訳モデルに合成する手法が考えられる。本稿では 2 節で先ず、木構造ベースの翻訳方式で翻訳規則を規定するために用いられる同期文脈自由文法 (2.1 節) を、そしてこの文法に従った翻訳規則を獲得するために用いられる階層的ルール抽出 (2.2 節) と Tree-to-String ルール抽出 (2.3 節) について紹介する。3 節では従来のテーブル合成手法の手続きと問題点について触れる。4 節では、提案手法である木構造を中間表現とするテーブル合成手法 (4.1 節) について述べる。また、4.1 節で紹介する手法は従来よりも厳しい条件で翻訳モデルを合成し、従来手法よりもカバレッジを損なうおそれがあるため、線形補間によって従来モデルのカバレッジを適宜補う手法 (4.2 節) についても提案する。

本提案手法の有効性を調査するため、国連多言語コーパス [7] から英語を中間言語とし、異なる 5 つの言語を用いて実験を行ったところ、部分構文木の一致を条件に合成された翻訳モデルを従来モデルと組み合わせることによって、多くの言語対において有意な精度向上が見られた。5 節では実験手順と結果の分析を述べ、6 節に本稿の結論をまとめる。

2. 機械翻訳の枠組

2.1 同期文脈自由文法

本節では先ず、木構造ベースの様々な翻訳方式で用いられる同期文脈自由文法 (SCFG) について紹介する。SCFG は、以下のような導出ルールによって構成される。

$$X \rightarrow \langle \bar{s}, \bar{t} \rangle \quad (1)$$

ここで、 X は導出ルールの親記号であり、 \bar{s} と \bar{t} はそれ

ぞれ原言語と目的言語における終端記号と非終端記号からなる記号列である。 \bar{s} と \bar{t} にはそれぞれ同じ数の非終端記号が含まれ、対応する記号に対して同じインデックスが付与される。以下に導出ルールの例を示す。

$$X \rightarrow \langle X_0 \text{ of } X_1, X_1 \text{ 的 } X_0 \rangle \quad (2)$$

このような導出ルールは、対訳文と単語アラインメント [1] を基に自動獲得可能であり、階層的ルールとして抽出する方法 [8]、Tree-to-String ルールとして抽出する方法 [9] が代表的である。それぞれ 2.2 節、2.3 節において紹介する。抽出された各々のルールには、双方向のフレーズ翻訳確率 $\phi(\bar{s}|\bar{t})$, $\phi(\bar{t}|\bar{s})$, 双方向の語彙翻訳確率 $\phi_{lex}(\bar{s}|\bar{t})$, $\phi_{lex}(\bar{t}|\bar{s})$, ワードペナルティ (\bar{t} の終端記号数), フレーズペナルティ (定数 1) の計 6 つのスコアが付与される。

翻訳時には、導出に用いられるルールのスコアと、生成される目的言語文の言語モデルスコアの和を導出確率として最大化するよう探索を行う。言語モデルを考慮しない場合、CKY+法 [10] によって効率的な探索を行ってスコア最大の導出を得ることが可能である。言語モデルを考慮する場合には、キューブ枝刈り [8] などの近似法により探索空間を抑えつつ、目的言語モデルを考慮した探索が可能である。

2.2 階層的ルール

階層的フレーズベース翻訳 (Hierarchical Phrase-Based Translation: Hiero [8]) で用いられる階層的ルールの特徴と例について述べる。階層的ルールは、開始記号 S および単語と単一の親記号 X からなる導出ルール、そして接着ルール $S \rightarrow \langle X_0, X_0 \rangle$, $S \rightarrow \langle S_0 X_1, S_0 X_1 \rangle$, $S \rightarrow \langle S_0 X_1, X_1 S_0 \rangle$ によって構成される。導出ルールは、原言語と目的言語で行ごとに意味の対応の取れた対訳コーパスと、対訳コーパスを用いて学習・推定された単語アラインメントをもとにフレーズ抽出 [11]、再帰的な階層的フレーズ抽出 [8] を経て自動的に獲得可能である。

以下に導出ルールの例を示す。

$$X \rightarrow \langle \text{Officers, 主席 成員} \rangle \quad (3)$$

$$X \rightarrow \langle \text{the Committee, 委会} \rangle \quad (4)$$

$$X \rightarrow \langle X_0 \text{ of } X_1, X_1 \text{ 的 } X_0 \rangle \quad (5)$$

ここから下記のような導出が可能であり、入力文に対する翻訳候補が得られる。

$$\begin{aligned} S &\rightarrow \langle X_0, X_0 \rangle \\ &\Rightarrow \langle X_1 \text{ of } X_2, X_2 \text{ 的 } X_1 \rangle \\ &\Rightarrow \langle \text{Officers of } X_2, X_2 \text{ 主席 成員} \rangle \\ &\Rightarrow \langle \text{Officers of the Committee, 委会 的 主席 成員} \rangle \end{aligned}$$

このようなヒューリスティックを用いたルール抽出によって、言語対に非依存な方法で高度な並び替えに対応できる翻訳規則を得られることが Hiero の利点であるが、一方で統語的情報を用いない総当りのな手法でもあるため、計算量が膨大になりやすいことや不適切な導出も行われやすい点が課題として残されている。

2.3 Tree-to-String 翻訳ルール

Tree-to-String (T2S) 翻訳 [9] で用いられる T2S ルールの特徴と例について述べる。T2S ルールは原言語側で構文解析結果を用いる点が階層的ルールと異なっており、導出ルールの親記号は X に限定されず、構文木の句構造タグに応じた非終端記号が付与される。導出ルールは対訳コーパスと単語アラインメントの他に、対訳コーパスの原言語側の構文解析結果を用いて抽出され、導出を行う際にも入力文の構文解析結果を用いる。

以下に導出ルールの例を示す。

$$X_{NP} \rightarrow \langle (NP (NNS Officers)), \text{主席 成員} \rangle \quad (6)$$

$$X_{NP} \rightarrow \langle (NP (DT the) (NNP Committee)), \text{委会} \rangle \quad (7)$$

$$X_{PP} \rightarrow \langle (PP (IN of) X_{NP,0}), X_0 \text{ 的} \rangle \quad (8)$$

$$X_{NP} \rightarrow \langle (NP X_{NP,0} X_{PP,1}), X_1 X_0 \rangle \quad (9)$$

ここで、原言語側の木構造は S 式によって表現されている。ここから下記のような導出が可能であり、入力文の構文解析結果に対する翻訳候補が得られる。

$$\begin{aligned} X_{\text{ROOT}} &\rightarrow \langle X_{NP,0}, X_0 \rangle \\ &\Rightarrow \langle (NP X_{NP,1} X_{PP,2}), X_2 X_1 \rangle \\ &\Rightarrow \langle (NP (NP (NNS Officers) X_{PP,2})), X_2 \text{ 主席 成員} \rangle \\ &\Rightarrow \left\langle \begin{array}{l} (NP \\ (NP (NNS Officers)) \\ (PP (IN of) \\ (NP (DT the) \\ (NNP Committee)))) \end{array}, \text{委会 的 主席 成員} \right\rangle \end{aligned}$$

このように、T2S 翻訳では原言語文の構文解析結果を元に句構造に応じた導出を行うことで、人手による翻訳規則の記述を行わずに文法に則した翻訳に近い結果が得られる。このため Hiero の場合のような不適切な導出ルールが抽出される可能性も低くなり、よりコンパクトで高速な翻訳が可能となる。翻訳精度は構文解析器の性能に左右されるが、英語のように高精度な句構造解析を行える言語を原言語とした翻訳において、特に優れた単語並び替え精度や翻訳精度を得られることが知られている [12], [13]。

3. 従来のテーブル合成手法

本節では、ピボット翻訳手法の中でも特に高い翻訳精度を得られることで知られているテーブル合成手法 (Triangulation [4]) の標準的な手順を示す。

従来のテーブル合成手法では、まず原言語・中間言語および中間言語・目的言語の翻訳モデルを対訳データによ

て学習し、それぞれをルールテーブル T_{SP} , T_{PT} として格納する。そして、テーブル T_{SP} , T_{PT} のそれぞれに共通する中間言語フレーズを含む導出ルールを結び付けることで新たに原言語・目的言語の導出ルールを合成し、ルールテーブル T_{ST} を作成する。

$$\begin{aligned} X &\rightarrow \langle \bar{s}, \bar{t} \rangle \in T_{ST} \\ \text{s.t. } X &\rightarrow \langle \bar{s}, \bar{p} \rangle \in T_{SP} \wedge X \rightarrow \langle \bar{p}, \bar{t} \rangle \in T_{PT} \end{aligned} \quad (10)$$

合成されたすべての規則について、フレーズ翻訳確率 $\phi(\cdot)$ と語彙翻訳確率 $\phi_{lex}(\cdot)$ を以下の計算によって推定する。

$$\phi(\bar{t}|\bar{s}) = \sum_{\bar{p} \in T_{SP} \cap T_{PT}} \phi(\bar{t}|\bar{p}) \phi(\bar{p}|\bar{s}) \quad (11)$$

$$\phi(\bar{s}|\bar{t}) = \sum_{\bar{p} \in T_{SP} \cap T_{PT}} \phi(\bar{s}|\bar{p}) \phi(\bar{p}|\bar{t}) \quad (12)$$

$$\phi_{lex}(\bar{t}|\bar{s}) = \sum_{\bar{p} \in T_{SP} \cap T_{PT}} \phi_{lex}(\bar{t}|\bar{p}) \phi_{lex}(\bar{p}|\bar{s}) \quad (13)$$

$$\phi_{lex}(\bar{s}|\bar{t}) = \sum_{\bar{p} \in T_{SP} \cap T_{PT}} \phi_{lex}(\bar{s}|\bar{p}) \phi_{lex}(\bar{p}|\bar{t}) \quad (14)$$

式 (11)-(14) は、以下のような条件を満たす無記憶通信路モデルに基いている。

$$\phi(\bar{t}|\bar{p}, \bar{s}) = \phi(\bar{t}|\bar{p}) \quad (15)$$

$$\phi(\bar{s}|\bar{p}, \bar{t}) = \phi(\bar{s}|\bar{p}) \quad (16)$$

例えば式 (15) では、共通する中間言語フレーズに対応する原言語フレーズと目的言語フレーズが存在する場合、中間言語フレーズから目的言語フレーズへ翻訳される確率は対応する原言語フレーズの影響を受けないという仮定を置いていることになる。しかし実際の自然言語処理ではこの仮定に反する現象があり、図 1 の例のように、中間言語フレーズに複数の解釈がある場合には、中間言語の単語列にそれぞれ対応する原言語フレーズと目的言語フレーズが本来意味的に結び付かない場合でも、翻訳確率の掛け合わせによって高い翻訳確率が推定されてしまい、誤訳の原因となりやすい。

4. 構文木を中間表現とするピボット翻訳手法

4.1 構文木の一一致条件を用いたテーブル合成手法

3 節では、ピボット翻訳における従来のテーブル合成手法について説明し、中間言語フレーズの曖昧性が原因で翻訳確率が不適切に推定され、翻訳精度が低下しう問題について述べた。前述の問題に対処するため、中間言語フレーズの意味や文法的役割が複数存在する曖昧な状況においては、記号列が一致していても異なる状態であるよう区別されるべきだと考えられる。特に文法的役割は構文解析によって容易に明示的な区別が可能であり、本節では構文木を中間表現とするテーブル合成手法を提案する。

本提案手法では、先ず原言語・中間言語および中間言語・目的言語それぞれの言語対の対訳コーパスの中間言語側に対して句構造解析を行い、中間言語を入力側とするような T2S 翻訳モデルを学習し、それぞれをルールテーブル T_{PS} , T_{PT} として格納する。ここで T_{PS} , T_{PT} の導出ルールはそれぞれ $X \rightarrow \langle \hat{p}, \bar{s} \rangle$, $X \rightarrow \langle \hat{p}, \bar{t} \rangle$ のような形を取り、 \hat{p} は中間言語の部分構文木を表す記号列 (S 式)、 \bar{s} , \bar{t} はそれぞれ原言語、目的言語における記号列を意味する。導出ルールの合成の手順は式 (10)-(14) と同様に行うが、 T_{SP} の代わりに T_{PS} (逆方向の翻訳モデル) を用いる点や、中間言語フレーズ \bar{p} の代わりに中間言語の部分構文木 \hat{p} を用いる点で異なる。原言語側や目的言語側には構文解析を用いていないため、合成される導出ルールは階層的ルールとなる。

本手法は、従来の表層的な単語列の一致よりも厳しい、木構造の一致を条件に翻訳規則を結び付けるため、誤ったフレーズ対応を結び付けることが減り、結果的に信頼度の高い翻訳モデルが合成されると考えられる。しかし、厳しい一致条件で合成することで、結び付けるフレーズ対応の数自体も減少するため、合成されたモデルのカバレッジを損なう問題も考えられる。SMT システムは翻訳モデルでカバーされた表現の中から訳出候補を選び出すため、カバーされていない表現の翻訳は不可能であり、翻訳確率推定の信頼度とカバレッジを両立することが重要である。そのため、基本的には本提案手法で得られる高信頼度の翻訳モデルを用いつつも、低信頼度ながらカバレッジの高い従来手法で得られる翻訳モデルも考慮する手法について 4.2 節で提案する。

4.2 翻訳モデルの線形補間

翻訳モデルの線形補間 (Linear Interpolation [4]) はカバレッジの増加や分野適応を効率的に行うために利用され、4.1 節で提案したテーブル合成手法のカバレッジの欠如を従来モデルで補間することで、より高い性能が得られると考えられる。

共通の言語対を対象とする 2 つの翻訳モデル T_1 , T_2 が存在する場合、それぞれの翻訳確率 $\phi_1(\cdot)$, $\phi_2(\cdot)$ を以下の数式に従って $\phi_{1+2}(\cdot)$ へと線形補間して新しい翻訳モデル T_{1+2} に併合することができる。

$$\phi_{1+2}(\bar{t}|\bar{s}) = \alpha \cdot \phi_1(\bar{t}|\bar{s}) + (1 - \alpha) \phi_2(\bar{t}|\bar{s}) \quad (17)$$

ここで、 α は補間係数であり 0 から 1 の間の実数値が入る。翻訳確率の線形補間は順・逆方向のフレーズ翻訳確率・語彙翻訳確率それぞれにおいて式 (17) と同様に行うことができる。 α は片方の翻訳モデルを重視したい度合いに応じて定められ、例えば T_1 を重要視している場合には $\alpha = 0.9$ を設定、 T_1 , T_2 を均等に扱いたい場合は $\alpha = 0.5$ を設定するなど、事前にパラメータを決めて併合する場合

もあるが、最適なパラメータを事前に求めることは容易ではない。

近年の SMT は厳密な確率モデルとしてそのまま翻訳確率を扱うのではなく、近似的に対数線形モデルとして対数確率スコア等に対する重みの最適化問題として考えるのが一般的である [14]。対数線形モデルを用いることで、開発データセットに対して翻訳精度が最大となるようなパラメータ調整を効率的に行えるため [14]、式 (17) の代わりに以下のような式で対数線形補間を行うことを考える。

$$\phi_{1+2}(\bar{t}|\bar{s}) = (\phi_1(\bar{t}|\bar{s}))^\alpha \cdot (\phi_2(\bar{t}|\bar{s}))^{(1-\alpha)} \quad (18)$$

$$\begin{aligned} w \cdot \log \phi_{1+2}(\bar{t}|\bar{s}) \\ = w \cdot \alpha \cdot \log \phi_1(\bar{t}|\bar{s}) + w \cdot (1 - \alpha) \cdot \log \phi_2(\bar{t}|\bar{s}) \\ = w_1 \cdot \log \phi_1(\bar{t}|\bar{s}) + w_2 \cdot \log \phi_2(\bar{t}|\bar{s}) \end{aligned} \quad (19)$$

これにより、最初から対数確率スコア $\log \phi_1$ と $\log \phi_2$ を SMT システムの新しい素性として組み込み、パラメータ最適化時に w_1 , w_2 を求めることで、結果的に最適な対数線形補間パラメータと対数確率への重みを求めることと等価の問題として扱える。また、このような重み w_1 , w_2 は、2 つのモデルの順方向・逆方向のフレーズ翻訳確率・語彙翻訳確率それぞれに対して個別に求めることができる。

5. 実験的評価

5.1 実験設定

本実験の提案手法の有効性を評価するため、国連多言語コーパス [7] を用いて様々な言語の組み合わせで各手法の翻訳精度を比較した。国連多言語コーパスは、英語 (En)、アラビア語 (Ar)、スペイン語 (Es)、フランス語 (Fr)、ロシア語 (Ru)、中国語 (Zh) の 6 カ国語に対応により複数の語族をカバーしており、それぞれの言語対で 1,100 万文以上の対訳文を含む大規模コーパスであるため、多言語翻訳の比較評価に相応しいと考えられる。

本実験では、英語が現実的なピボット翻訳で多用される点や高精度な句構造解析ツールが利用可能な点を考慮して、英語を中間言語として固定し、その他の 5 言語の全ての組み合わせでピボット翻訳を行い、手法毎の翻訳精度を比較した。利用可能な約 1,100 万文を全て翻訳モデルの学習に利用するのが理想的ではあるが、Hiero の学習・評価には膨大な計算資源が必要であるため、本実験では学習用データの最初の 100 万文を取り出し、各翻訳モデルおよび言語モデルの学習に利用した。翻訳の評価およびパラメータの調整には、あらかじめ用意されたそれぞれ 4,000 文のデータセットから 2,000 文ずつを用いた。SCFG のデコーダには Travatar [12] を使い、付属の階層的ルール抽出コードと T2S ルール抽出コードを用いて、それぞれ Hiero 翻訳モデルと T2S 翻訳モデルの学習を行った。英語コーパスの句構造解析には Ckylark [15] を用いた。翻訳結果の評価

Source	Target	BLEU Score [%]			
		Direct	Tri. Phrase (baseline)	Tri. Subtree (proposed 1)	LogInt. Subtree+Phrase (proposed 2)
Ar	Es	35.37	31.46	‡ 32.01	‡ 32.74
	Fr	28.83	26.05	26.06	† 26.38
	Ru	22.28	19.99	19.91	‡ 20.42
	Zh	19.14	17.55	17.43	17.55
Es	Ar	18.24	15.10	‡ 15.74	‡ 15.80
	Fr	39.03	35.17	35.44	‡ 35.70
	Ru	26.58	23.79	† 24.12	‡ 24.33
	Zh	22.30	20.07	‡ 20.86	‡ 21.08
Fr	Ar	16.04	13.36	12.72	‡ 13.72
	Es	42.53	38.82	38.93	† 39.13
	Ru	25.26	22.07	21.96	22.04
	Zh	20.39	18.43	† 18.81	† 18.80
Ru	Ar	15.48	13.10	† 13.41	‡ 13.77
	Es	35.77	32.91	32.47	33.15
	Fr	30.61	27.78	27.60	27.86
	Zh	21.39	19.50	19.52	19.52
Zh	Ar	13.50	11.86	11.60	‡ 12.31
	Es	29.84	26.82	26.59	26.87
	Fr	24.15	21.62	† 22.02	‡ 22.34
	Ru	20.04	18.28	18.03	18.45

表 1 ピボット翻訳手法毎の翻訳精度比較. 太字はそれぞれの言語対において提案手法がベースラインよりも BLEU スコアが高いことを示し, 短剣符は提案手法の翻訳精度が従来手法よりも統計的に有意であることを示す (†: $p < 0.05$, ‡: $p < 0.01$).

には, 自動評価尺度 BLEU [16] を用い, 各翻訳モデルは MERT [14] により, パラメータ調整用データセットに対して BLEU スコアが最大となるように各パラメータの調整を行った.

次節では以下の 4 つの翻訳手法を比較評価する.

Direct:

理想的な状況における翻訳精度の上限値として比較を行うため, 中間言語を用いず原言語・目的言語の直接対訳コーパスで学習した Hiero 翻訳モデルによって翻訳

Tri. Phrase:

原言語・中間言語および中間言語・目的言語の Hiero 翻訳モデルを, 標準的なテーブル合成手法によって合成し, 原言語・目的言語の Hiero 翻訳モデルによって翻訳 (3 節, ベースライン)

Tri. Subtree:

中間言語・原言語および中間言語・目的言語の T2S 翻訳モデルを, 部分構文木の一致条件を用いたテーブル合成手法で合成し, 原言語・目的言語の Hiero 翻訳モデルによって翻訳 (4.1 節, 提案手法 1).

LogInt. Subtree+Phrase

上記の Tri. Phrase と Tri. Subtree で得られた Hiero 翻訳モデルの翻訳確率を個別の素性として併合して最適化を行った対数線形補間モデルによって翻訳 (4.2 節, 提案手法 2)

5.2 実験結果

表 1 に各ピボット翻訳手法で得られた翻訳精度を示す. 表の結果から, 1 つ目の提案手法である構文木の一致条件を用いるテーブル合成手法は, 従来のテーブル合成手法と比較した場合, スペイン語から中国語への翻訳で BLEU スコアが 0.79 向上するなど大幅な改善が見られる場合もあるが, 一方でフランス語からアラビア語への翻訳で BLEU スコアが 0.62 下がってしまうなど, 安定した結果が得られなかった. これは 4.1 節で述べたように, テーブル合成時に構文木の一致という厳しい制約が課されることで, 翻訳モデルの信頼度向上とカバレッジ低下が同時に発生していることが原因と考えられる.

一方で構文木の一致条件で合成した翻訳モデルと従来手法で合成した翻訳モデルを線形補間して用いる 2 つ目の提

案手法では、ほとんどの言語対においてベースラインよりも BLEU スコアの向上と、多くの場合には統計的有意差も確認できた。アラビア語から中国語、フランス語からロシア語への翻訳の場合のみ、BLEU スコアが向上しなかったが大きく減少したわけでもなく、ほぼ同等のスコアが得られた点からも、線形補間による翻訳モデルの併合によって概ね安定した精度向上が得られたと考えられる。補間係数 α は開発用データセットのスコアに最適化され、過学習をしない限り、従来モデルに重み 1 を与えることもできるので従来モデルよりもスコアが下がることはない。

本実験で得られた結果としては、木構造の一致条件を用いてテーブル合成を行う提案手法では安定した精度向上が得られなかったため、木構造の完全一致だけで曖昧性を解消するには不十分であると考えられる。翻訳モデルを線形補間して用いた手法では、結果として木構造の一致で合成した導出ルールの翻訳スコアを従来モデルに対する追加の素性として用いているのと同様であり、このように統計的情報を考慮して翻訳スコアを補正することでカバレッジを保ちつつ曖昧性の解消に役立てられるものと期待できる。

6. おわりに

本稿では、ピボット翻訳において統語的情報を用いた曖昧性解消の手法として、中間言語の構文木を一致条件として 2 つの T2S 翻訳モデルを Hiero 翻訳モデルに合成するテーブル合成手法、および、前述の手法によって合成された翻訳モデルと従来手法で合成した翻訳モデルを線形補間によって同時に考慮して翻訳を行う手法を提案した。実験結果から、木構造を一致条件とするテーブル合成手法のみでは安定した精度向上は得られなかったものの、従来モデルとの線形補間によって多くの言語対で有意に翻訳精度が向上した。これにより、ピボット翻訳において中間言語の統語的情報を有効利用することで曖昧性解消に役立てられることが示唆された。今後の計画として、木構造の完全一致という厳しい条件ではなく、一致条件を緩めた場合の評価や、補間係数パラメータの最適化時に起こる過学習を防ぐための方策、異なる対訳データを用いた際の評価などに取り組んでいきたい。

謝辞：本研究の一部は、JSPS 科研費 16H05873 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] Chris 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.
- [3] Adrià de Gispert and José B. Mariño. Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish. In *Proc. of LREC 5th Workshop on Strategies for developing machine translation for minority languages*, pp. 65–68, 2006.
- [4] Trevor Cohn and Mirella Lapata. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proc. ACL*, pp. 728–735, 2007.
- [5] Akiva Miura, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Improving Pivot Translation by Remembering the Pivot. In *Proc. ACL*, pp. 573–577, 2015.
- [6] Masao Utiyama and Hitoshi Isahara. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Proc. NAACL*, pp. 484–491, 2007.
- [7] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations Parallel Corpus v1.0. In *Proc. LREC*, pp. 3530–3534, 2016.
- [8] David Chiang. Hierarchical Phrase-Based Translation. *Computational Linguistics*, Vol. 33, No. 2, pp. 201–228, 2007.
- [9] Yang Liu, Qun Liu, and Shouxun Lin. Tree-to-String Alignment Template for Statistical Machine Translation. In *Proc. ACL*, pp. 609–616, 2006.
- [10] Jean-Cédric Chappelier, Martin Rajman, et al. A Generalized CYK Algorithm for Parsing Stochastic CFG. In *Proc. TAPD*, Vol. 98, p. 5. Citeseer, 1998.
- [11] Phillip Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-Based Translation. In *Proc. NAACL*, pp. 48–54, 2003.
- [12] Graham Neubig. Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers. In *Proc. ACL Demo Track*, pp. 91–96, 2013.
- [13] Graham Neubig and Kevin Duh. On the Elements of an Accurate Tree-to-String Machine Translation System. In *Proc. ACL*, pp. 143–149, 2014.
- [14] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. ACL*, pp. 160–167, 2003.
- [15] Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Ckylark: A More Robust PCFG-LA Parser. In *Proc. NAACL*, pp. 41–45, 2015.
- [16] 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, 2002.