NAIST-IS-MT1451101

修士論文

中間言語モデルを用いた多言語機械翻訳の精度向上

三浦 明波

2016 年 3 月 8 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士（工学）授与の要件として提出した修士論文である.

三浦 明波

審査委員：

中村 哲 教授	（主指導教員）
松本 裕治 教授	（副指導教官）
Graham Neubig 助教	（副指導教官）

中間言語モデルを用いた多言語機械翻訳の精度向上*

三浦 明波

内容梗概

統計的機械翻訳において，特定の言語対で十分な文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．複数のピボット翻訳手法が考案されている中でも，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．ところが，従来のテーブル合成手法では，フレーズ対応推定時に用いた中間言語の情報は消失し，翻訳時には利用できない問題が発生する．本研究では，合成時に用いた中間言語の情報も記憶し，ピボットの言語モデルを追加の情報源として翻訳に利用する新たなテーブル合成手法を提案する．また，欧州議会議事録による多言語コーパスを用いた実験により，本手法で評価を行った全ての言語の組み合わせで従来手法よりも有意に高い翻訳精度が得られた．

キーワード

統計的機械翻訳, 多言語翻訳, ピボット翻訳, 同期文脈自由文法, 言語モデル, 対訳コーパス

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文, NAIST-IS-MT1451101, 2016 年 3 月 8 日.

Improving Multilingual Machine Translation using Pivot Language Models*

Akiva Miura

Abstract

In statistical machine translation, the pivot translation approach allows for translation of language pairs with little or no parallel data by introducing a third language for which data exists. In particular, the triangulation method, which translates by combining source-pivot and pivot-target translation models into a source-target model is known for its high translation accuracy. However, after the conventional triangulation method, information of pivot phrases is forgotten, and not used in the translation process. In this research, we propose a novel approach to remember the pivot phrases in the triangulation stage, and use a pivot language model as an additional information source at translation phase. Experimental results on the Europarl corpus showed significant improvements in all tested combinations of languages.

Keywords:

statistical machine translation, multilinguality, pivot translation, synchronous context-free grammars, language models, parallel corpora

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1451101, March 8, 2016.

目次

図目次

表目次

第 1 章　緒言

1.1　背景

言語は，人間にとって主要なコミュニケーションの道具であると同時に，話者集団にとっては社会的背景に根付いたアイデンティティーでもある．母国語の異なる相手と意思疎通を取るためには，翻訳は必要不可欠な技術であるが，専門の知識が必要となるため，ソフトウェア的に代行できる機械翻訳の技術に期待が高まっている．英語と任意の言語 (母国語など) 間での翻訳で機械翻訳の実用化を目指す例が多いが，利用者は必ずしも英語を熟知しているわけではないため，文章の原意を汲むためには原文から母国語に直接翻訳を行えることが好ましい．

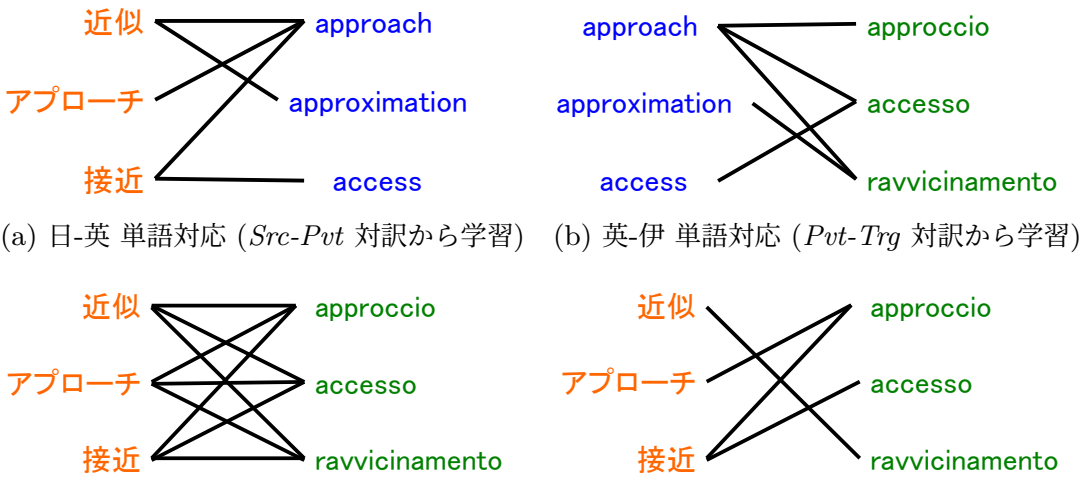
人手で翻訳規則を記述するルールベース機械翻訳 (Rule-Based Machine Translation; RBMT [1]) では，対象の二言語に精通した専門家の知識が必要であり，多彩な表現を広くカバーすることも困難である．そのため，近年主流の機械翻訳方式と考えられる，機械学習技術を用いて対訳コーパスから自動的に翻訳規則を獲得する統計的機械翻訳 (Statistical Machine Translation; SMT [2]) について本論文では議論を行う．対訳コーパスとは，行単位で意味の対応する二言語間の文章を集めたデータのことを指すが，SMT では学習に使用する対訳コーパスが大規模になるほど，高精度な訳出結果を得られることが知られている [3]．しかし，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直ちに取得することは困難と言える．このような，容易に対訳コーパスを取得できないような言語対においても，既存の言語資源を有効に用いて高精度な機械翻訳を実現することができれば，機械翻訳の実用の幅が大きく広がることになる．

特定の言語対で十分な文量の対訳コーパスが得られない場合，中間言語 (Pvt) を用いたピボット翻訳が有効な解法の一つである [4, 5, 6, 7]．中間言語を用いる方法も様々であるが，一方の出力言語と他方の入力言語が一致するような 2 つの機械翻訳システムを利用できる場合，それらをパイプライン処理する逐次的ピボット翻訳 (Cascade Translation [4]) 手法が先ず考えられる．より高度なピボット翻訳の手法としては，原言語・中間言語 (Src-Pvt) と中間言語・目的言語 (Pvt-Trg) の 2 組の言語対のためにそれぞれ学習された SMT システムのモデルを合成し，新しく得られた原言語・目的言語 (Src-Trg) の SMT システムを用いて翻訳を行うテーブル合

成手法 (Triangulation [5]) も提案されており，この手法で特に高い翻訳精度を得られることが知られている．

しかし，PBMT において議論されてきたピボット翻訳手法は，異なる SMT の枠組みでも有効であるかどうかは明らかにされていない．

例えば英語と日本語，英語と中国語といった語順の大きく異なる言語間の翻訳では，同期文脈自由文法 (Synchronous Context-Free Grammar; SCFG [8]) のような木構造ベースの SMT によって高度な単語並び替えに対応可能であり，PBMT よりも高い翻訳精度を達成できることが知られている．PBMT において有効性の知られているピボット翻訳手法が，SCFG による翻訳でも有効であるなら，並び替えの問題に高度に対応しつつ直接 *Src-Trg* の対訳コーパスを得られない状況にも対処できることになる．



(a) 日-英 単語対応 (*Src-Pvt* 対訳から学習)　(b) 英-伊 単語対応 (*Pvt-Trg* 対訳から学習)

(c) 日-伊 単語対応推定時の全候補　(d) 正しく推定された 日-伊 単語対応

図 1.1　2 組の単語対応から新しい単語対応を推定

また，テーブル合成手法では，*Src-Pvt* フレーズ対応と *Pvt-Trg* フレーズ対応から，正しい *Src-Trg* フレーズ対応と確率スコアを推定する必要がある．図 1.1 に示す例では，個別に学習された (a) の日英翻訳および (b) の英伊翻訳における単語対応から，日伊翻訳における単語対応を推定したい場合，(c) のように単語対応を推定する候補は非常に多く，(d) のように正しい推定結果を得ることは困難である．その上，図 1.1(c) のように推定された *Src-Trg* の単語対応からは，原言語と目的言

語の橋渡しをしていた中間言語の単語情報が分からないため，翻訳を行う上で重要な手がかりとなり得る情報を失ってしまうことになる．このように語義曖昧性や言語間の用語法の差異により，ピボット翻訳は通常の翻訳よりも本質的に多くの曖昧性の問題を抱えており，さらなる翻訳精度の向上には課題がある．

1.2 研究目的

本研究では，多言語機械翻訳，とりわけ対訳コーパスの取得が困難である低資源言語対における機械翻訳の高精度化を目指し，従来のピボット翻訳手法を調査，問題点を改善して翻訳精度を向上させることを目的とする．ピボット翻訳の精度向上に向けて，本論文では 2 段階の議論を行う．

第 1 段階目では，従来の PBMT で有効性の知られているピボット翻訳手法が異なる枠組みの SMT でも有効であるかどうかを調査する．前節で述べたように，テーブル合成手法が PBMT におけるピボット翻訳手法において特に高い翻訳精度を得られていることが知られているため，木構造ベースの SMT である SCFG による翻訳で同等の処理を行うための応用手法を提案する．SCFG とテーブル合成手法によるピボット翻訳が，逐次的ピボット翻訳や，PBMT におけるピボット翻訳手法よりも高い精度を得られるどうかを比較評価することで，次の段階への予備実験とする．

第 2 段階目では，テーブル合成手法において発生する曖昧性の問題を解消し，翻訳精度を向上させるための新たな手法を提案する．従来のテーブル合成手法では，図 4.1(c) に示したように，フレーズ対応の推定後には中間言語フレーズの情報が失われてしまうことを前節で述べた．この問題を克服するため，本論文では原言語と目的言語を結び付けていた中間言語フレーズの情報も翻訳モデル中に保存し，原言語から目的言語と中間言語へ同時に翻訳を行うための確率スコアを推定することによって翻訳を行う新しいテーブル合成手法を提案する．通常の SMT システムでは，入力された原言語文から，目的言語における訳出候補を選出する際に，文の自然性を評価し，適切な語彙選択を促すために目的言語の言語モデルを利用する．一方，本手法で得られる翻訳モデルを用いた SMT システムでは，原言語文に対して目的言語文と中間言語文を同時に翻訳を行うため，目的言語モデルのみではなく，

中間言語モデルも同時に考慮して訳出候補の探索を行う．本手法の利点は，英語のように中間言語として選ばれる言語は豊富な単言語資源を得られる傾向が強いため，このような追加の言語情報を翻訳システムに組み込み，翻訳の質を向上させられることにある．

1.3 論文構成

本論文では，2 章で様々な SMT の要素技術について説明し，PBMT と木構造に基づく翻訳の違いを示す．3 章では，PBMT におけるピボット翻訳手法を紹介し，後半ではピボット翻訳手法の 1 つであるテーブル合成手法を SCFG で応用するための方法を説明し，各手法の翻訳精度を実験により比較評価し，考察を行う．4 章では，中間言語情報を記憶し，目的言語と中間言語への同時翻訳を行うピボット翻訳手法を提案する．5 章で提案手法と従来手法の翻訳精度を実験により比較評価し，その結果を元に考察を行う．最後に 6 章で本論文のまとめと今後の課題について述べる．

第 2 章　統計的機械翻訳

機械翻訳とは，計算機を用いて，ある言語の文を異なる言語の文に変換する技術である．機械翻訳を実現するため，歴史の中で様々な枠組が提案されてきており，代表的なものとしてルールベース機械翻訳 (Rule-Based Machine Translation; RBMT [1])[1]，用例ベース機械翻訳 (Example-Based Machine Translation; EBMT [9])[2]，統計的機械翻訳 (Statistical Machine Translation; SMT [2]) およびニューラルネット機械翻訳 (Neural Machine Translation; NMT [10, 11]) などが存在する．

RBMT では，人手で翻訳規則を記述するため，文法規則を意識しながら翻訳結果を制御できることは利点であるが，言語対毎に原言語と目的言語の二ヶ国語に精通した専門家の知識が必要であり，膨大で複雑多岐なルールの記述が必要であり，多彩な表現までカバーすることは困難なことが欠点である．それ以外の手法では，対訳コーパスを元に翻訳規則を自動獲得する点などで共通点が多い．EBMT では，対訳の用例データベースを元に，入力文に対して類似度が高いと判断される用例を組み合わせて目的言語文を結合する．この枠組は特に，用例に用いる対訳コーパスと，翻訳を行いたい文の分野が適合している場合に高い性能を発揮することが知られているが，複雑な組み合わせ問題を考慮する必要があり，用例を選択する際のスコア付等にモデルが仮定されておらず，複雑な要素の扱いがシステム開発者にゆだねられることになる．本論文で中心的に議論する SMT は，対訳データから単語やフレーズの対応関係を自動的に翻訳規則として獲得し，それぞれに確率的スコアを付与し，入力文に対して翻訳確率スコアが最大となるような目的言語文を探索して出力する．用いる翻訳規則に応じてさらに細かい SMT の枠組があるため次節以降で紹介するが，総じて確率モデルに基づくこの枠組の汎用性は高く，線形対数モデルで一般化されたモデルは効率的な探索や最適化を実現可能である．近年は大規模な計算資源の運用と充実した言語資源の整備が可能になったため，SMT の研究・開発に拍車がかかっており，実用の機械翻訳システムのほとんどが SMT を採用している．ニューラルネットワークを用いて機械翻訳を実現する NMT の研究者も近年増加の傾向にあり，様々な手法が考案されている．SMT のモデル学習を強化す

[1] 知識に基づく機械翻訳 (Knowledge-Based Machine Translation; KBMT) という呼称も有名
[2] アナロジーに基づく機械翻訳 (Mechanical Translation by Principal Analogy) とも

るためにニューラルネットワークを使う場合もあるが，対訳コーパスから直接翻訳を行う手法もあり，通常の SMT では困難なハイパーパラメータを含めた最適化や，長期の依存関係を学習できるため注目度は高い．しかし，高度な並列計算環境が必須である点や，学習されたモデルの意味付けや制御が難しい点もあり，実用にはまだ課題が残されている．

それぞれの枠組に長所と短所があるが，本論文では SMT に的を絞って議論を行う．本研究で SMT を用いる理由として，対訳コーパスから翻訳規則を自動獲得できる点や，汎用性が高く現行の実用システムが SMT を採用している利点もあるが，特に各要素が確率モデルに基いてスコア付けされているため，モデルを合成する際に意味付けを行いやすく，ピボット翻訳のような問題と相性が良いと考えられる点が大きい．

次節以降では，SMT の要素技術について紹介し (2.1 節)，翻訳モデルと，翻訳規則を自動獲得する上で主軸となる単語アラインメント (2.1.1 節)，翻訳候補の自然性・流暢性の評価に用いられる言語モデル (2.1.2 節)，SMT の中でも特に代表的な翻訳方式であるフレーズベース機械翻訳 (PBMT, 2.2 節) と木構造に基づく翻訳方式である同期文脈自由文法 (SCFG, 2.3 節)，SCFG を 3 言語以上に対応できるよう一般化して拡張された複数同期文脈自由文法 (Multi-Synchronous Context-Free Grammar; MSCFG, 2.4 節) について説明する．

2.1 対数線形モデル

SMT の基本的なアイディアは，雑音のある通信路モデル [12] に基いている．ある原言語の文 $\boldsymbol{f}$ に対して，可能なさまざまな目的言語の翻訳文の集合を $\mathcal{E}(\boldsymbol{f})$ とする．$\boldsymbol{f}$ が目的言語の文 $\boldsymbol{e} \in \mathcal{E}(\boldsymbol{f})$ へと翻訳される確率 $Pr(\boldsymbol{e}|\boldsymbol{f})$ をすべての $\boldsymbol{e}$ について計算可能とする．SMT では，$Pr(\boldsymbol{e}|\boldsymbol{f})$ を最大化する $\hat{\boldsymbol{e}} \in \mathcal{E}(\boldsymbol{f})$ を求めることにより翻訳誤りが最小な目的言語の文を生成する．

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e} \in \mathcal{E}(\boldsymbol{f})} Pr(\boldsymbol{e}|\boldsymbol{f}) \tag{2.1}$$

$$= \arg\max_{\boldsymbol{e} \in \mathcal{E}(\boldsymbol{f})} \frac{Pr(\boldsymbol{f}|\boldsymbol{e})Pr(\boldsymbol{e})}{Pr(\boldsymbol{f})} \tag{2.2}$$

$$= \arg\max_{\boldsymbol{e} \in \mathcal{E}(\boldsymbol{f})} Pr(\boldsymbol{f}|\boldsymbol{e}) P(\boldsymbol{e}) \tag{2.3}$$

しかし，このままではモデル化が困難であるため，近年では以下のような対数線形モデルに基づく重みの最適化問題として考えることが一般的である [13].

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e} \in \mathcal{E}(f)} Pr(\boldsymbol{e}|\boldsymbol{f}) \tag{2.4}$$

$$\approx \arg\max_{\boldsymbol{e} \in \mathcal{E}(\boldsymbol{f})} \frac{\exp\left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}\right)}{\sum_{e'} \exp\left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e'})\right)} \tag{2.5}$$

$$= \arg\max_{\boldsymbol{e} \in \mathcal{E}(\boldsymbol{f})} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}) \tag{2.6}$$

ここで，$\boldsymbol{h}$ は素性ベクトルと呼ばれ，翻訳の枠組毎に決められた次元数を持ち，翻訳モデル (2.1.1) や言語モデル (2.1.2) 等の対数確率スコア，導出に伴う単語並び替え，各種ペナルティなどを与える．$\boldsymbol{w}$ は $\boldsymbol{h}$ と同じ次元を持っており，素性ベクトルの各要素に対する重み付けを行う．$\boldsymbol{w}$ の各要素を最適な値に調整するためには，対訳コーパスを学習用データや評価用データとは別に切り分けた，開発用データを利用し，原言語文の訳出と参照訳 (目的言語側の正解訳) との類似度を評価するための自動評価尺度 BLEU[14] などが最大となるようパラメータを求める [13]．2.2 節以降で説明する各種翻訳枠組も，この対数線形モデルに基いているが，用いる素性はそれぞれで異なる．

2.1.1 翻訳モデル

翻訳モデル $Pr(\boldsymbol{f}|\boldsymbol{e})$ は，訳語の尤もらしさを規定するための統計モデルであり，対訳コーパスから学習を行う．翻訳モデルは，訳語の尤もらしさを規定するための統計モデルであり，対訳コーパスから学習を行う．翻訳モデルは直接 $\boldsymbol{e}$ と $\boldsymbol{f}$ を関連づけず，なんらかのステップを経て $\boldsymbol{e}$ から $\boldsymbol{f}$ が生成されると仮定し，そのステップを導出 $\boldsymbol{d}$ と呼ぶ．統計モデルでは，導出 $\boldsymbol{d}$ は潜在変数として扱われ，式 (2.3) は

$$\begin{aligned}\hat{\boldsymbol{e}} &= \arg\max_{\boldsymbol{e}\in\mathcal{E}(\boldsymbol{f})} Pr(\boldsymbol{f}|\boldsymbol{e})Pr(\boldsymbol{e}) \\ &= \arg\max_{\boldsymbol{e}\in\mathcal{E}(\boldsymbol{f})} \sum_{\boldsymbol{d}\in\mathcal{D}(\boldsymbol{f},\boldsymbol{e})} Pr(\boldsymbol{f},\boldsymbol{d}|\boldsymbol{e})Pr(\boldsymbol{e}) \end{aligned} \tag{2.7}$$

となる．$\mathcal{D}(\boldsymbol{f},\boldsymbol{e})$ を，$\boldsymbol{f}$ および $\boldsymbol{e}$ が与えられたときの導出の集合とする．

導出の具体例として，二言語間の単語対応によって $\boldsymbol{f}$ と $\boldsymbol{e}$ を関連付け，翻訳確率を最大化する IBM モデル [2] が知られている．IBM モデルでは，$Pr(\boldsymbol{f}|\boldsymbol{e})$ は，$\boldsymbol{e}$ から $\boldsymbol{f}$ を生成する確率モデルであり，単語アラインメント $\boldsymbol{a}$ を介して単語単位に生成するモデルとして定義される．

$$Pr(\boldsymbol{f}|\boldsymbol{e}) = \sum_{\boldsymbol{a}} Pr(\boldsymbol{f},\boldsymbol{a}|\boldsymbol{e}) \tag{2.8}$$

統計モデルとして表現されることで，単語アラインメントの決定は，翻訳モデルの条件付き確率を最大化する問題として捉えられる．

$$\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}} Pr(\boldsymbol{f},\boldsymbol{a}|\boldsymbol{e}) \tag{2.9}$$

単語アラインメント $\boldsymbol{a}$ は，対訳文 $\langle\boldsymbol{f},\boldsymbol{e}\rangle$ の単語単位の対応 $\langle f_j, e_i\rangle$ を表した集合であり

$$\boldsymbol{a} = \{\cdots,(j,i),\cdots\} \tag{2.10}$$

のような，点の集合として表現される．

このような単語単位での翻訳は，単語の置き換えだけで意味が通じるほど似通った言語対に対しては有効であるが，現実には単語が一対一対応しない場合が多く，語順の考慮も困難であるため，このままでは不十分である．そのため，IBM モデルで学習を行った単語アラインメントを元に，連続する単語列の対応を最小の翻訳単

位とするフレーズベース機械翻訳 (PBMT) によって単語ベースの SMT よりも翻訳精度は劇的に向上した．しかし，語順の大きく異なる言語対に関しては PBMT でも複雑な並び替え問題が発生するため翻訳は難しく，このような高度な並び替えの問題に対応すべく木構造ベースの SMT も提案された．これらの翻訳モデルについては 2.2-2.3 節で詳細を述べる．

2.1.2 言語モデル

言語モデル $Pr(\boldsymbol{e})$ は，与えられた文の単語の並びが目的言語においてどの程度自然で流暢であるかを評価するために用いられる．優れた言語モデルは正確に自然な文に高い確率を与え，不自然な文に低い確率を与える．この情報を翻訳の過程で参照すると，翻訳候補の中からより自然な文を選択することができ，より流暢な翻訳結果の生成につながる．本節では機械翻訳で広く用いられている n-gram モデルについて説明する．

先ず，確率連鎖によって目的言語文 $\boldsymbol{e}$ の自然性を以下のように計算することを考える．

$$P(e_1^I) = \prod_{i=1}^{I+1} P_{ML}(e_i|e_0^{i-1}) \tag{2.11}$$

ここで，I は $\boldsymbol{e}$ の長さであり，$\boldsymbol{e} = e_1^I = e_1 \cdots e_I$ とする．また，$e_0 = \langle s \rangle$ は文頭記号，$e_{I+1} = \langle /s \rangle$ は文末記号を表す．各条件付き確率は最尤推定を用いて以下のように求めることができる．

$$P_{ML}(e_i|e_0^{i-1}) = \frac{c_{train}(e_0^i)}{c_{train}(e_0^{i-1})} \tag{2.12}$$

ここで c_{train} は学習データ中における単語列の出現頻度を表す．

しかし，このままでは学習データ中に出現しない文に対しては確率 0 を与えてしまうため，多くの文に対して確率 0 が推定されてしまい，訳出候補の自然性評価に優劣が付けられず問題が生じてしまう．そこで，e_i の条件付き確率を算出する時

に，e_i より前に現れるすべての単語列 e_0^{i-1} ではなく，直前の $n-1$ 単語 e_{i-n+1}^{i-1} のみを考慮した条件付き確率を利用する．e_i と直前の $n-1$ 単語を含む n 単語の列 e_{i-n+1}^{i} を n-gram という．これに従って，式 (2.11) を以下のように変形する．

$$P(e_i^I) \approx \prod_{i=1}^{I+1} P_{ML}(e_i|e_{i-n+1}^{i-1}) \tag{2.13}$$

このように学習された n-gram 言語モデルでは，学習データ中に存在しない文に対してでも確率スコアを推定することができる．しかし，式 (2.13) でも，学習データ中に出現しない n-gram を含む文に対しては依然として確率 0 を推定してしまう問題が残る．この問題を解決するために，n-gram の条件付き確率 $P(e_i|e_{i-n+1}^{i-1})$ と $(n-1)$-gram の条件付き確率 $P(e_i|e_{i-n+2}^{i-1})$ を組み合わせて確率スコアを推定する平滑化という手法が存在する．平滑化にも様々な手法が提案されているが，代表的な平滑化手法として，線形補間や Kneser-Ney 法などがよく用いられる [15].

2.2 フレーズベース翻訳

Koehn らによる フレーズベース機械翻訳 (PBMT [16]) は SMT で最も代表的な翻訳枠組である．PBMT 翻訳モデルを学習する際には，先ず対訳コーパスから単語アラインメントを学習し，アラインメント結果をもとに複数の単語からなるフレーズを抽出し，各フレーズ対応にスコア付けを行う．

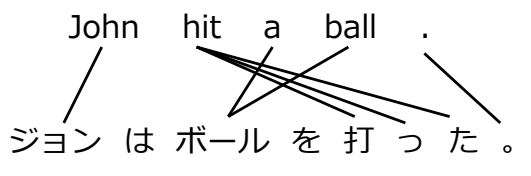


図 2.1　英-日 単語アラインメント

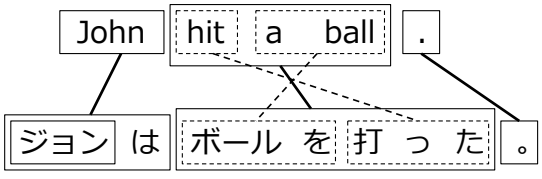


図 2.2　英-日 フレーズ抽出

例えば，2.1.1 節で説明した手法によって，学習用対訳データから図 2.1 のような単語対応が得られたとする[1]．得られた単語対応からフレーズの対応を見つけ出し

[1] 日本語や中国語，タイ語のように，通常の文では単語をスペースで区切らないような言語では，先ず単語分割を行うツールを用いて分かち書きを行う必要がある

て抽出を行う例を図 2.2 に示す．図のように，与えられた単語対応から抽出されるフレーズ対応の長さは一意に定まらず，複数の長さのフレーズ対応が抽出される．ただし，抽出されるフレーズ対応には，フレーズの内外を横断するような単語対応が存在しないという制約が課され，フレーズの最大長等も制限される．このようにして抽出されたフレーズ対の一覧を元に，フレーズ対や各フレーズの頻度を計算し，PBMT の翻訳モデルが学習される．

単語ベースの翻訳モデルと違い，フレーズベースの翻訳モデルは抽出されたフレーズを翻訳の基本単位とし，これによって効率的に慣用句のような連続する単語列の翻訳規則を学習し，質の高い翻訳が行えるようになる．フレーズの区切り方によって，与えられた原言語文から，ある目的言語文へ翻訳されるための導出も複数の候補があり，それぞれの導出で用いられるフレーズ対の確率スコアや並び替えも考慮して最終的な翻訳確率を推定する．式 (2.6) の対数線形モデルによって確率スコア最大の翻訳候補を探索するが，素性関数として用いられるものには，双方向のフレーズ翻訳確率，双方向の語彙翻訳確率，単語ペナルティ，フレーズペナルティ，言語モデル，並び替えモデルなどがある．

PBMT は，翻訳対象である 2 言語間の対訳コーパスさえ用意すれば，容易に学習し，高速な翻訳を行うことが可能であるため，多くの研究や実用システムで利用されている．しかし，文の構造を考慮しない手法であるため，単語の並び替えが効果的に行えない傾向にある．高度な並び替えモデルを導入することは可能であるが [17, 18]，長距離の並び替えは未だ困難であり，ピボット翻訳で用いることは容易ではない．

2.3 同期文脈自由文法

本節では，木構造に基づく SMT の枠組である同期文脈自由文法 (SCFG [8]) について説明する．SCFG は，階層的フレーズベース翻訳 (Hierarchical Phrase-Based Translation; Hiero [8]) を代表とする様々な翻訳方式で用いられている．SCFG は，以下のような同期導出規則によって構成される．

$$X \longrightarrow \langle \bar{s},\ \bar{t} \rangle \tag{2.14}$$

ここで，X は同期導出規則の親記号であり，$\bar{s}$ 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．以下に置換規則の例を示す．

$$X \longrightarrow \langle X_0 \text{ of } X_1,\ X_1 \text{ の } X_0 \rangle \tag{2.15}$$

Hiero 翻訳モデルのための SCFG の学習手法では，先ず PBMT と同等のアイディアで，対訳コーパスから学習された単語アラインメントを元にフレーズを抽出する．そしてフレーズ対応中の部分フレーズ対応に対しては，非終端記号 X_i で置き換えてよいというヒューリスティックを用いて，多くの SCFG ルールが自動抽出される．例えば，図 2.1 の単語アラインメントを用いて，以下のような同期導出規則を得ることができる．

$$X \longrightarrow \langle X_0 \text{ hit } X_1 \text{ .},\ X_0 \text{ は } X_1 \text{ を 打 っ た 。} \rangle \tag{2.16}$$
$$X \longrightarrow \langle \text{John},\ \text{ジョン} \rangle \tag{2.17}$$
$$X \longrightarrow \langle \text{a ball},\ \text{ボール} \rangle \tag{2.18}$$

また，初期非終端記号 S と初期導出規則 $S \longrightarrow \langle X_0,\ X_0 \rangle$，抽出された上記の導出規則を用いて，以下のような導出が可能である．

$$\begin{aligned} S &\Longrightarrow \langle X_0,\ X_0 \rangle && (2.19)\\ &\Longrightarrow \langle X_0 \text{ hit } X_1 \text{ .},\ X_0 \text{ は } X_1 \text{ を 打 っ た 。} \rangle && (2.20)\\ &\Longrightarrow \langle \text{John hit } X_1 \text{ .},\ \text{ジョン は } X_1 \text{ を 打 っ た 。} \rangle && (2.21)\\ &\Longrightarrow \langle \text{John hit a ball .},\ \text{ジョン は ボール を 打 っ た。} \rangle && (2.22) \end{aligned}$$

対訳文と単語アラインメントを元に自動的に SCFG ルールが抽出される．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 [19] 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，キューブ枝狩り [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．

SCFG によって翻訳を行う SMT では，並び替えと語彙選択を同時に行うルールを利用することで，PBMT よりも高い並び替え精度を得ることが可能となる．一方で，Hiero によって自動抽出される SCFG の翻訳規則は，一種類の非終端記号 X のみが用いられ，統語的な情報を考慮していないため，総当り的な抽出手法によってモデルサイズが肥大化し，多くの計算量やメモリ使用量が必要になる欠点もある．そのため，原言語側か目的言語側，あるいはその両側で構文解析を行い，句構造単位で翻訳規則の抽出を行うことでコンパクトで高品質な翻訳モデルを学習し，より文法的規則に近い導出を行って翻訳精度を高める Tree-to-String, String-to-Tree, Tree-to-Tree 翻訳なども知られている [20].

2.4 複数同期文脈自由文法

複数同期文脈自由文法 (MSCFG [21]) は，SCFG を複数の目的言語文の同時生成に対応できるよう拡張された手法である．SCFG では生成規則中の目的言語記号列 $\overline{t}$ が単一であったが，MSCFG では以下のように N 個の目的言語記号列を有する．

$$X \longrightarrow \left\langle \overline{s},\ \overline{t_1}, \cdots, \overline{t_N} \right\rangle \tag{2.23}$$

通常の MSCFG 学習手法では，SCFG ルール抽出手法を一般化し，行アラインメントの取れた多言語コーパスから多言語置換規則が抽出され，複数の目的言語を考慮したスコアが付与される．

MSCFG で複数の目的言語モデルを考慮した探索を行う場合，言語毎に導出中の単語列を記憶し，組み合せ毎に状態を区別する必要があるため，探索手順にも複数の手法が考えられる．SCFG の探索における単一の目的言語文を単純に複数の目的言語文に拡張して同時に展開する同時探索では，探索幅の制限により主要な目的言

語文の多様性が失われてしまう可能性がある．そこで先ず，第一の目的言語文のみを考慮した組み合わせで探索して多様性を確保し，続いてその他の目的言語文との組み合わせに展開する逐次探索により，主要な目的言語を重視した効率的な探索が行える．

2.5 多言語応用における問題

2.5.1 対訳コーパスの取得困難性

ここまで，SMT は対訳コーパスから自動的に翻訳規則を行い，統計に基づいたモデルによって翻訳確率スコアが最大となるような翻訳を行うことを述べてきた．統計モデルであるため，言語モデルの学習に用いる目的言語コーパスと翻訳モデルの学習に用いられる対訳コーパスが大規模になるほど確率推定の信頼性が向上し，翻訳スコアが高くなる．言語モデルについては，目的言語の話者数やインターネット利用者数の影響はあるものの，比較的取得が容易であるため問題になることは少ない．一方で対訳コーパスは SMT の要であり，学習データにカバーされていない単語や表現の翻訳は不可能なため，多くの対訳データ取得が望ましく，実用的な SMT システム構築には数百万文の対訳が必要と言われている．例えば，日本語と英語，フランス語と英語のような言語対は，企業のマニュアル，特許，科学論文，医療文書などから抽出された質の高い対訳データを合計千万文以上インターネットで取得することが可能である[2]．

ところが英語を含まない言語対，例えば日本語とフランス語のような言語対を考えると，それぞれの言語では単言語コーパスが豊富に取得可能であるにも関わらず，100 万文を超えるような大規模な対訳データを短時間で獲得することは困難である．また，スペインの地方公用語であるカタルーニャ語のように，英語との対訳コーパスの取得も困難である言語も少なくはない．このように，SMT の大前提である対訳コーパスは多くの言語対で十分な文量を取得できないことが多く，このため，任意の言語対で翻訳を行うには課題がある．

[2]多くの対訳コーパスは無償で取得し実験に用いることができるが，高価で取引されるものや，商用利用は有償である対訳データも多い

2.5.2 ピボット翻訳

2.5.1 節で述べた，特定の言語対で十分な対訳を得られないような場面では，中間言語 (ピボット言語) を介したピボット翻訳が有効であることが知られている．例えば，日本語とフランス語の例の場合は，日本語と英語，英語とフランス語でそれぞれ大規模な対訳コーパスが得られるため，英語を介した日本語とフランス語の翻訳が可能となる．また，英語とカタルーニャ語の場合，英語とスペイン語であれば欧州議会議事録などから得られた大規模な対訳コーパスが取得可能であり，スペイン語とカタルーニャ語であれば，スペイン国内の新聞などから大規模な対訳コーパスが取得可能であるため，スペイン語を介した英語とカタルーニャ語の翻訳が可能である．しかし，一般的には英語を含む言語対の方が，英語を含まない言語対よりも対訳コーパスを得やすい傾向が強いため，本研究では英語を含まない言語対で英語を介してピボット翻訳を行うことに焦点をあてて実験を実施する．

中間言語を用いて翻訳を行う方法として，目的言語から中間言語へ翻訳し，その後に中間言語から目的言語へと翻訳を行う逐次的ピボット翻訳 [4] が考えられるが，この方法では中間言語に一度翻訳した際に発生する情報落ちが目的言語への翻訳時にも伝播されるため，質の高い翻訳はあまり期待できない．そのため，ピボット翻訳では目的言語の情報をいかに損なわずに目的言語へ伝達するかが課題となり，中間言語の扱い方も様々に考案されてきている．代表的なピボット翻訳手法について 3 章で説明する．

2.5.3 本研究の位置づけ

日本語とフランス語のような言語対は，対訳コーパスの取得が困難であり，英語を介してピボット翻訳を行えることを説明してきたが，日本語と英語，日本語とフランス語のような語族の異なる言語対では一般的に語順も大きく異なり，長距離の並び替え問題を考慮した翻訳が必要になる．本章で述べてきた，言語対毎の対訳コーパスの取得性や言語構造の類似度の問題を表 2.1 にまとめる．

本研究は多言語翻訳，特に対訳コーパスの取得が困難である低資源言語対における翻訳精度向上を目指している．また，語順の大きく異なる言語対でピボット翻訳を行う際，PBMT では十分な精度が得られないことが予想されるため，SCFG の

言語対 (代表例)	対訳コーパスの取得性	言語構造の類似度	手法 (代表例)
英語 ↔ フランス語	○	○	句に基づく手法 (PBMT)
英語 ↔ 日本語	○	×	木に基づく手法 (SCFG など)
英語 ↔ カタルーニャ語 (via スペイン語)	×	○	**ピボット翻訳 w/ PBMT**
日本語 ↔ フランス語 (via 英語)	×	×	**ピボット翻訳 w/ SCFG?**

表 2.1　多言語翻訳における課題

ように高度な並び替え問題に対処できる翻訳枠組でピボット翻訳が有効であるかを調査し，予備実験を行う．その後に，従来のピボット翻訳手法の問題点に対処し，翻訳精度を向上させるための手法を提案し，実験と考察を行う．

第 3 章　ピボット翻訳手法

PBMT におけるピボット翻訳手法が数多く考案されており，本章の 3.1 節では代表的なピボット翻訳手法について紹介する．また，3.2 節では，PBMT で有効性の知られているピボット翻訳手法であるテーブル合成手法を SCFG で応用するための手法を提案し，実験による比較評価と考察を述べる．本章では原言語を *Src*，目的言語を *Trg*，中間言語を *Pvt* と表記し，これらの言語対を *Src-Pvt, Src-Trg, Pvt-Trg* のように表記して説明を行うこととする．

3.1　フレーズベース翻訳におけるピボット翻訳手法

3.1.1　逐次的ピボット翻訳手法

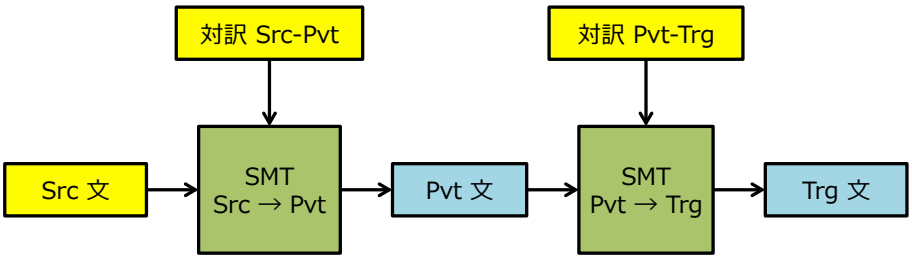


図 3.1　逐次的ピボット翻訳

逐次的ピボット翻訳手法 (Cascade)[4] によって *Src* から *Trg* へと翻訳を行う様子を図 3.1 に示す．この方式では先ず，*Src-Pvt, Pvt-Trg* それぞれの言語対で，対訳コーパスを用いて翻訳システムを構築する．そして *Src* の入力文を *Pvt* へ翻訳し，*Pvt* の訳文を *Trg* に翻訳することで，結果的に *Src* から *Trg* への翻訳が可能となる．この手法は機械翻訳の入力と出力のみを利用するため，PBMT である必然性はなく，任意の機械翻訳システムを組み合わせることができる．優れた 2 つの機械翻訳システムがあれば，そのまま高精度なピボット翻訳が期待できることや，既存のシステムを使い回せること，実現が非常に容易であることが利点と言える．逆に，最初の翻訳システムの翻訳誤りが次のシステムに伝播し，加法性誤差によって精度が落ちることは欠点となる．

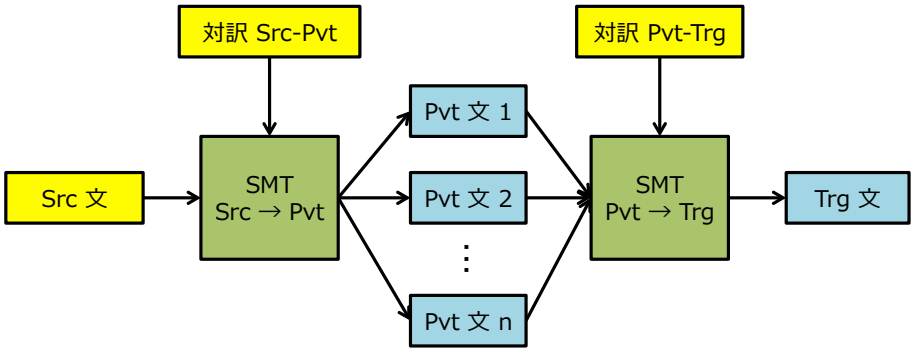


図 3.2　マルチセンテンス方式

本手法では通常，*Src-Pvt* 翻訳時点で訳出候補のうち最も確率スコアが高くなる *Pvt* の 1 文のみが選択されるが，図 3.2 のように，*Src-Pvt* 翻訳システムで確率スコアの高い上位 n 文の訳出候補を出力し，*Pvt-Trg* 翻訳における探索の幅を拡げるマルチセンテンス方式も提案されている [6]．しかし，通常より n 倍の探索時間が必要であり，大きな精度向上も報告されていない．

3.1.2　擬似対訳コーパス手法

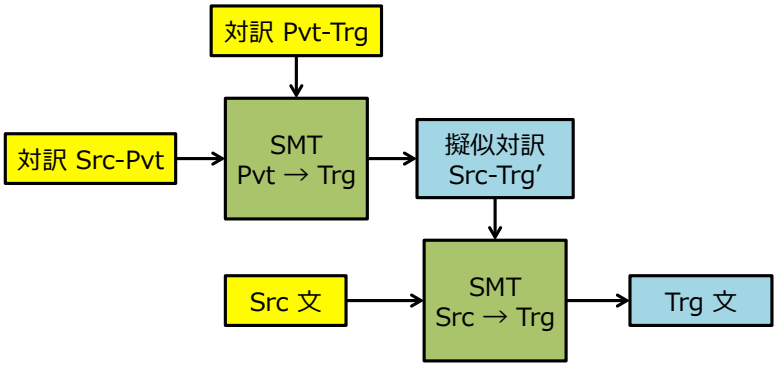


図 3.3　擬似対訳コーパス手法

擬似的に *Src-Trg* 対訳コーパスを作成することで SMT システムを構築する**擬似対訳コーパス手法 (Synthetic)** によって，*Src-Trg* 翻訳を行う様子を図 3.3 に

示す．この手法では先ず，*Src-Pvt, Pvt-Trg* のうちの片側，図の例では *Pvt-Trg* の対訳コーパスを用いて SMT システムを構築する．そして *Src-Pvt* 対訳コーパスの *Pvt* 側の全文を *Pvt-Trg* 翻訳にかけることで，*Src-Trg* 擬似対訳コーパスが得られる．これによって得れられた *Src-Trg* 擬似対訳コーパスを用いて，SMT の翻訳モデルを学習することが可能となる．対訳コーパスの翻訳時に少しの翻訳誤りが含まれていても，統計モデルの学習に大きく影響しなければ，高精度な訳出が期待できる．既存のシステムから新しい学習データやシステムを作り直すことになるため，一度擬似対訳コーパスを作ってしまえば，それ以降は通常の SMT と同じ学習手法を用いられることは利点となる．

De Gispert らは，スペイン語を中間言語としたカタルーニャ語と英語のピボット翻訳で，逐次的ピボット翻訳手法と擬似対訳コーパス手法によるピボット翻訳手法の比較実験 [4] を行った．その結果，これらの手法間で優位な差は示されなかった．

3.1.3　テーブル合成手法

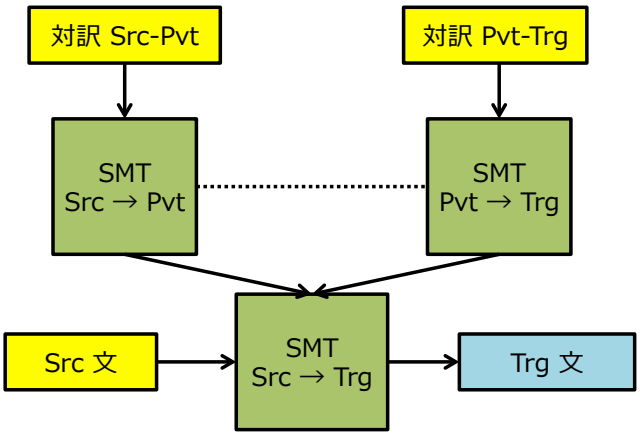


図 3.4　テーブル合成手法

PBMT，SCFG では，対訳コーパスによってフレーズ対応を学習してスコア付けした翻訳モデルを，それぞれフレーズテーブル，ルールテーブルと呼ばれる形式で格納する．フレーズテーブルを合成することで *Src-Trg* のピボット翻訳を行う様子を図 3.4 に示す．Cohn らによるフレーズテーブル合成手法 [5] では，先ず *Src-Pvt*

および *Pvt-Trg* の翻訳モデルを対訳コーパスによって学習し，それぞれをフレーズテーブル T_{SP}, T_{PT} として格納する．得られた T_{SP}, T_{PT} から，*Src-Trg* の翻訳確率を推定してフレーズテーブル T_{ST} を合成する．T_{ST} を作成するには，フレーズ翻訳確率 $\phi(\cdot)$ と語彙翻訳確率 $\phi_{lex}(\cdot)$ を用いて，次式のように翻訳確率の推定を行う．

$$\phi\left(\overline{t}|\overline{s}\right) = \sum_{\overline{p}\in T_{SP}\cap T_{PT}} \phi\left(\overline{t}|\overline{p}\right)\phi\left(\overline{p}|\overline{s}\right) \tag{3.1}$$

$$\phi\left(\overline{s}|\overline{t}\right) = \sum_{\overline{p}\in T_{SP}\cap T_{PT}} \phi\left(\overline{s}|\overline{p}\right)\phi\left(\overline{p}|\overline{t}\right) \tag{3.2}$$

$$\phi_{lex}\left(\overline{t}|\overline{s}\right) = \sum_{\overline{p}\in T_{SP}\cap T_{PT}} \phi_{lex}\left(\overline{t}|\overline{p}\right)\phi_{lex}\left(\overline{p}|\overline{s}\right) \tag{3.3}$$

$$\phi_{lex}\left(\overline{s}|\overline{t}\right) = \sum_{\overline{p}\in T_{SP}\cap T_{PT}} \phi_{lex}\left(\overline{s}|\overline{p}\right)\phi_{lex}\left(\overline{p}|\overline{t}\right) \tag{3.4}$$

ここで，$\overline{s}$, $\overline{p}$, $\overline{t}$ はそれぞれ *Src, Pvt, Trg* のフレーズであり，$\overline{p} \in T_{SP} \cap T_{PT}$ はフレーズ $\overline{p}$ が T_{SP}, T_{PT} の双方に含まれていることを示す．式 (3.1)-(3.4) は，以下のような条件を満たす無記憶通信路モデルに基づいている．

$$\phi\left(\overline{t}|\overline{p},\overline{s}\right) = \phi\left(\overline{t}|\overline{p}\right) \tag{3.5}$$

$$\phi\left(\overline{s}|\overline{p},\overline{t}\right) = \phi\left(\overline{s}|\overline{p}\right) \tag{3.6}$$

この手法では，翻訳確率の推定を行うために全フレーズ対応の組み合わせを求めて算出する必要があるため，大規模なテーブルの合成には長い時間を要するが，既存のモデルデータから精度の高い翻訳を期待できる．

Utiyama らは，英語を中間言語とした複数の言語対で，逐次的ピボット翻訳手法とテーブル合成手法によるピボット翻訳で比較実験を行った [6]．その結果，テーブル合成手法では，$n = 1$ の単純な逐次的ピボット翻訳や，$n = 15$ のマルチセンテンス方式よりも高い BLEU スコアが得られたと報告している．

3.2 同期文脈自由文法におけるテーブル合成手法の応用

3.1 節で説明したピボット翻訳手法のうち，逐次的ピボット翻訳および擬似対訳コーパス手法は SMT の枠組にとらわれない手法であるため，SCFG を用いる SMT でもそのまま適用可能であるが，テーブル合成手法は本来，PBMT のフレーズテーブルを合成するために提案されたものである．SCFG を用いる翻訳方式では，式 (2.14) のように表現される同期導出規則をルールテーブルという形式で格納する．次節以降では，SCFG ルールテーブルを合成することで，PBMT におけるテーブル合成手法と同等のピボット翻訳を行うための手法について説明し，その後に PBMT および SCFG における複数のピボット翻訳手法による翻訳精度の差を実験によって比較評価し，考察を行う．

3.2.1 同期規則の合成

SCFG ルールテーブル合成手法では，先ず *Src-Pvt, Pvt-Trg* それぞれの言語対について，対訳コーパスを用いて同期導出規則を抽出し (2.3 節)，各規則の確率スコア等の素性を算出してルールテーブルに格納する．その後，*Src-Pvt, Pvt-Trg* ルールテーブル双方に共通の *Pvt* 記号列を有する導出規則 $X \rightarrow \langle \overline{s},\ \overline{p} \rangle$, $X \rightarrow \langle \overline{p},\ \overline{t} \rangle$ をすべて見つけ出し，新しい導出規則 $X \rightarrow \langle \overline{s},\ \overline{t} \rangle$ の翻訳確率を，式 (3.1)-(3.4) に従って推定する．PBMT においては $\overline{s}, \overline{p}, \overline{t}$ が各言語のフレーズ (単語列) を表しており，SCFG においては非終端記号を含む各言語の記号列を表す点で異なるが，計算式については同様である．また， $X \rightarrow \langle \overline{s},\ \overline{t} \rangle$ のワードペナルティおよびフレーズペナルティは $X \rightarrow \langle \overline{p},\ \overline{t} \rangle$ と同じ値に設定する．

3.2.2 比較評価実験

3.2.1 節で説明したルールテーブル合成手法によるピボット翻訳が，他の手法や他の翻訳枠組と比較して有効であるかどうかを調査するため，後述する手順によって比較実験を行った．

実験設定: 3.1 節および 3.2.1 節で紹介したピボット翻訳手法のうち，実現が非

常に容易で比較しやすい逐次的ピボット翻訳手法と，PBMT で高い実用性が示されたテーブル合成手法によるピボット翻訳を，PBMT および SCFG において実施し，翻訳精度の比較評価を行った．PBMT モデルの構築には Moses [22]，SCFG 翻訳モデルの構築には Travatar [23] の Hiero 学習ツールを利用した．複数の言語の組み合わせで PBMT，SCFG のそれぞれについて以下のように SMT の学習と評価を行い，ピボット翻訳手法の違いによる翻訳精度を比較した．

Direct (**直接翻訳**):
: 翻訳精度の上限値を得て比較を行うため，*Pvt* を用いず *Src-Trg* の直接対訳コーパスを用いて翻訳モデルを学習し評価

Cascade (**逐次的ピボット翻訳**):
: *Src-Pvt, Pvt-Trg* それぞれの対訳で学習された翻訳モデルでパイプライン処理を行い，*Src-Trg* 翻訳を評価

Triangulation (**テーブル合成方式**):
: *Src-Pvt, Pvt-Trg* それぞれの対訳で学習された翻訳モデルから，翻訳確率の推定により *Src-Trg* 翻訳モデルを合成し評価

本実験では，対訳コーパスとして，欧州諸言語を広くカバーし，ピボット翻訳のような多言語翻訳タスクで広く用いられる Europarl コーパス [24] を用いた．一般的なピボット翻訳タスクを想定し，英語 (En) を中間言語として固定し，欧州でも特に話者数の多いドイツ語 (De)，スペイン語 (Es)，フランス語 (Fr)，イタリア語 (It) の 4 言語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．5 言語間の対訳コーパスを得るためには，先ず Gale-Church アラインメント法 [25] によって行アラインメントの取れた多言語コーパスを取得し，それぞれの翻訳モデルと目的言語モデルの学習のために 10 万文ずつ，開発用と評価用にそれぞれ 1,500 文ずつを取り出した．また，言語構造の異なる言語対におけるピボット翻訳の評価も行うため，比較的小規模であるが聖書の翻訳を元に作成された多言語コーパス[1]を元に，日本語 (Ja)，英語 (En)，ヘブライ語 (He)，ギリシャ語 (El) の 4 言語でピボット翻訳の実験を行い，英語以外の言語を中間言語とした場合の評価も行った．これは，旧約聖書の原典はヘブライ語，新約聖書の原典はギリシャ語である点を考

[1] http://homepages.inf.ed.ac.uk/s0787820/bible/

Source	Target	MT Method	BLEU Score [%]		
			Direct	Triangulation	Cascade
De	Es	PBMT	29.44	**25.58**	24.23
		Hiero	27.10	**25.31**	25.05
	Fr	PBMT	27.25	**23.43**	22.38
		Hiero	25.65	**24.12**	23.86
	It	PBMT	22.47	**20.84**	19.16
		Hiero	23.04	**21.27**	20.76
Es	De	PBMT	24.40	**22.08**	20.06
		Hiero	20.11	**18.77**	18.52
	Fr	PBMT	36.36	**31.26**	26.81
		Hiero	33.48	**29.54**	27.00
	It	PBMT	30.07	**26.26**	24.40
		Hiero	27.82	**25.11**	22.57
Fr	De	PBMT	19.78	**18.52**	17.98
		Hiero	19.69	**18.73**	18.01
	Es	PBMT	38.04	**36.41**	33.11
		Hiero	34.36	**30.31**	27.26
	It	PBMT	28.21	**26.30**	22.70
		Hiero	28.48	**25.31**	22.73
It	De	PBMT	20.06	**17.52**	14.81
		Hiero	19.09	**17.35**	14.03
	Es	PBMT	36.51	**30.75**	27.33
		Hiero	31.99	**28.85**	25.64
	Fr	PBMT	33.34	**30.17**	28.89
		Hiero	31.39	**28.48**	25.87

表 3.1　ピボット翻訳手法毎の翻訳精度比較 (欧州議会議事録 100k 文)

慮している．聖書コーパスを用いた翻訳では，学習用データに 2 万 8 千文，開発用と評価用に千文ずつを用いた．翻訳結果の評価には，自動評価尺度 BLEU [14] を用い，各 SMT システムについて MERT [13] により，開発用データセットに対して BLEU スコアが最大となるようにパラメータ調整を行った．

実験結果: 様々な言語と機械翻訳方式の組み合わせについて Direct, Triangulation, Cascade の各ピボット翻訳手法で翻訳を行い評価した結果を表 3.1 と表 3.2 に示す．表 3.1 は Europarl，表 3.2 は聖書コーパスを用いて学習と翻訳を行った

Source	Target	Pivot	MT Method	BLEU Score [%]		
				Direct	Triangulation	Cascade
En	Ja	El	PBMT	33.39	**32.02**	29.52
			Hiero	38.27	**34.53**	30.08
		He	PBMT	33.39	**23.89**	17.12
			Hiero	38.27	**20.99**	18.04
El	Ja	En	PBMT	32.70	**30.97**	30.59
			Hiero	35.72	**32.16**	31.89
		He	PBMT	32.70	**21.16**	17.49
			Hiero	35.72	**21.25**	17.32
He	Ja	En	PBMT	23.45	**23.41**	22.45
			Hiero	24.36	**24.12**	23.81
		El	PBMT	23.35	**20.60**	19.38
			Hiero	24.36	18.10	**18.68**
Ja	En	El	PBMT	34.25	**30.63**	27.39
			Hiero	44.24	**35.78**	33.73
		He	PBMT	34.25	**21.67**	15.24
			Hiero	44.24	**24.77**	18.96
	El	En	PBMT	28.23	**27.52**	26.82
			Hiero	34.01	**36.46**	35.82
		He	PBMT	28.23	**27.52**	12.48
			Hiero	34.01	**22.17**	16.37
	He	En	PBMT	25.23	**15.40**	13.55
			Hiero	25.71	**16.11**	15.96
		El	PBMT	25.23	**10.40**	9.68
			Hiero	25.71	**16.11**	11.81

表 3.2　ピボット翻訳手法毎の翻訳精度比較 (聖書コーパス 280k 文)

結果であり，太字は言語と翻訳枠組の各組み合わせで精度の高いピボット翻訳手法を示す．先行研究では，PBMT のピボット翻訳手法において Triangulation が Cascade よりも高い精度の翻訳を行えることが示されており，このことは実験結果の表からも，すべての言語の組み合わせで再現されている．また，SCFG を用いた場合も，ギリシャ語を介したヘブライ語・日本語翻訳など，僅かに Cascade より低くなる場合もあるが，ほとんどのケースで Triangulation の精度が高くなっている．欧州の言語間の翻訳では，大きく語順が変わることは少ないため，PBMT で

も SCFG でも翻訳精度に大きな差は見られなかった．一方，表 3.2 のように，日本語と，言語構造の異なるような言語では SCFG の翻訳精度が高くなることが見られた．

また，同じ言語対の翻訳でも，中間言語によってピボット翻訳の精度が大きく変化し得ることも確認できる．例えば，ギリシャ語から日本語や逆方向のピボット翻訳では，英語を中間言語とする場合よりもヘブライ語を中間言語とした場合の方が大きく精度が下がっている．これは，英語とギリシャ語の場合には比較的語順や語彙が近いとされるが，ヘブライ語は日本語ともギリシャ語とも言語構造が大きく異なるため，それぞれの翻訳モデルの質も低く，それらを合成して得られる翻訳モデルでも多くの翻訳誤りが発生してしまうことが原因として考えられる．

3.2.3 考察および関連研究

本章前半では，PBMT で提案されてきた代表的なピボット翻訳手法について説明し，後半ではテーブル合成手法を SCFG のルールテーブルに適用するための手法について述べ，また言語対・機械翻訳方式・ピボット翻訳手法の組み合わせによって翻訳精度の影響を比較評価した．その結果，SCFG においてもテーブル合成手法によって高い翻訳精度を得られることが示され，また言語対や用いるデータによっては PBMT の場合よりも高い精度が得られることも分かった．ピボット翻訳におけるその他の関連研究は，PBMT のテーブル合成手法をベースに，さらに精度を上げるための議論が中心である [7, 26, 27]．テーブル合成手法では，翻訳精度をいかに正しく推定するかが問題となる．

Zhu らは，*Src-Pvt, Pvt-Trg* それぞれのフレーズ対の翻訳確率から直接 *Src-Trg* の翻訳確率を推定するのではなく，先ず *Src-Trg* フレーズ対の共起頻度を推定し，それを元に翻訳確率を推定する手法を提案しており，これにより不均衡な 2 つのフレーズテーブルの合成でも安定した翻訳精度が得られるとしている [7]．

Levinboim らは，テーブル合成時のフレーズ対応推定の中でも，特に単語対応の翻訳確率推定が困難であることに対処すべく，単語の分散表現 [28] の技術を用いて，直接対訳の取れない単語の対応に対しても翻訳確率を推定して，テーブル合成の質を高める手法を提案している [26]．

本章の実験結果からも中間言語の選び方によって精度が変化することを確認したが，中間言語がピボット翻訳に与える影響については Paul らの研究で詳しく議論されている [29]．現実には，中間言語を幾つも選べるような状況は多くはないが，複数の中間言語を介して同じ規模の対訳が得られるような場合には，構造の似通った言語を優先して選択すべきである．

また，中間言語は必ずしも 1 つに限定する必要はなく，複数の中間言語を同時に考慮するような手法も提案されている．その場合，それぞれの中間言語でテーブル合成を行って得られた，複数の *Src-Trg* テーブルに対して，線形補間で 1 つのテーブルに集約させたり，複数の翻訳モデルを同時に考慮して探索を行うような手法などがある [27]．

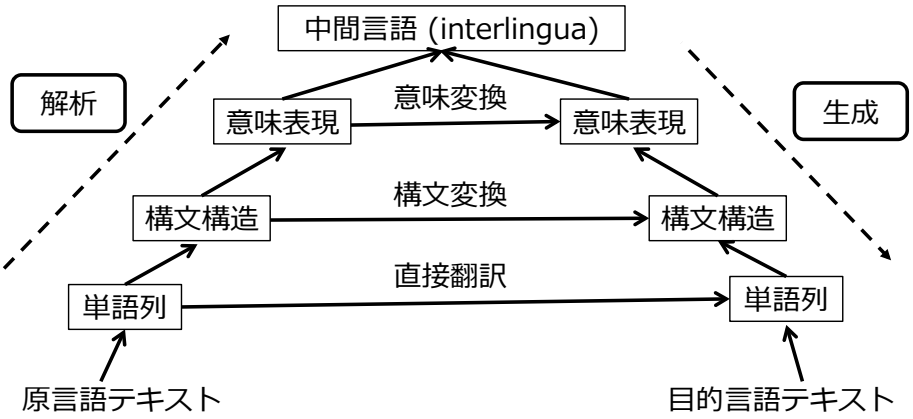


図 3.5　翻訳のレベル (Vauquois の三角形)

ここで, ピボット翻訳における中間言語 (Pivot Language) とは異なるが, RBMT における中間言語 (Interlingua) との関係性について触れておきたい．中間言語を用いる翻訳の試みというのは歴史が深く，SMT が発明されるよりも以前，RBMT が考案された当初から翻訳のレベルという議論がなされており [1]，図 3.5 に示すような，Vauquois の三角形と呼ばれる解析と生成の段階を表す図が有名である．より高い段階まで解析して言語変換を行うことで，原意を汲んだ翻訳が可能と言われており，三角形の頂点である中間言語 (Interlingua) とは，あらゆる言語の表現を包括できるような理想的な言語であるが，そのような自然言語は実在しない．過去には，人手で翻訳規則を記述する RBMT で，人工的な中間言語に変換するような

試みもなされてきた歴史があるが，ルールベースの機械翻訳では多彩な分野，多彩な表現を専門家の手でカバーしきることは不可能であり，また，あらゆる言語に精通している人間も存在しないのが実際である．

SMT の発明により，対訳コーパスを用いることで専門家によるルール記述よりも遥かに効率的に高精度な機械翻訳を行えるようになったが，今日の SMT は Vauquois の三角形の底辺を彷徨っているとも言われる．任意の自然言語を Interlingua のような中間言語として用いた場合，言語の表現力に依存して情報が失われてしまうため，生成の過程で原言語の情報を再現できなくなることが考えられる．そのため，中間言語には単純な単語列よりも高い表現力を持った形式を用いられるべきであり，一例として，オントロジーを中間表現に用いる機械翻訳手法 [30] も提案されている．また，ニューラルネット機械翻訳の手法として，複数の言語対の翻訳タスクを共通のエンコーダとして学習させることで翻訳精度を高める方法もあり [31]，これは翻訳規則の分散表現をうまく中間言語のように扱っていると考えることもできる．

第 4 章　中間言語情報を記憶するピボット翻訳手法の提案

第 3 章では，SMT で用いられているピボット翻訳手法について紹介し，従来手法の中で特に高い翻訳精度が得られることで知られるテーブル合成手法を SCFG で応用するための手順について説明した．また，比較評価実験により，SCFG においても PBMT と同様，テーブル合成手法によって逐次的ピボット翻訳手法よりも高い精度を得られることができた．しかし，直接の対訳を用いて学習した場合と比較すると，翻訳精度の差は未だ大きいため，精度が損なわれてしまう原因を特定し，解消することができれば，さらなる翻訳精度の向上が期待できる．テーブル合成手法で翻訳精度が損なわれる原因の一つとして，翻訳時に重要な手がかりとなるはずの中間言語の情報はテーブル合成後には失われてしまい，不正確に推定された *Src-Trg* のフレーズ対応と翻訳確率のみが残る点が挙げられる．本章では，従来では消失してしまう中間言語情報を記憶し，この追加の情報を翻訳時に用いることで精度向上に役立てる，新しいテーブル合成手法を提案する．

4.1　従来のテーブル合成手法の問題点

従来のテーブル合成手法の問題点について，第 1 章緒言の中でも紹介したが，本説で改めて説明を行う．テーブル合成手法では，*Src-Pvt, Pvt-Trg* それぞれの言語対におけるフレーズの対応と翻訳確率のスコアが与えられており，この情報を元に，*Src-Trg* 言語対におけるフレーズ対応と翻訳確率の推定を行う．ところが，語義曖昧性や言語間の用語法の差異により，*Src-Trg* のフレーズ対応を正確に推定することは困難である．

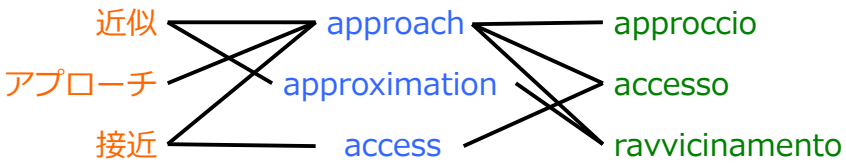


図 4.1　モデル学習に用いるフレーズ対応 (日-英-伊)

図 4.1 はテーブル合成手法によって対応を推定するフレーズの例を示しており，

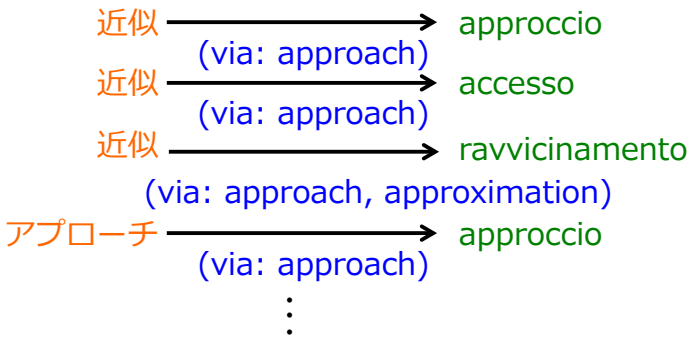


図 4.2　従来手法によって得られるフレーズ対応

図中では日本語とイタリア語それぞれにおける 3 つの単語が，語義曖昧性を持つ英単語「approach」に結び付いている．このような場合，*Src*-*Trg* のフレーズ対応を求め，適切な翻訳確率推定を行うのは複雑な問題となってくる．その上，図 4.2 に示すように，従来のテーブル合成手法では，合成時に *Src* と *Trg* の橋渡しをしていた *Pvt* フレーズの情報が，合成後には保存されず失われてしまう．現実の人手翻訳の場合を考えても，現在着目しているフレーズに関する追加の言語情報が与えられているなら，その言語を知る者にとって重要な手がかりとなって曖昧性解消などに用いることができる．そのため，*Src*-*Trg* を結び付ける *Pvt* フレーズは重要な言語情報であると考えられ，本研究では，この情報を保存することで機械翻訳にも役立てるための手法を提案する．

4.2　中間言語情報を記憶するテーブル合成手法

前節で述べた問題を克服するため，本研究では *Src* と *Trg* を結び付けていた *Pvt* フレーズの情報も翻訳モデル中に保存し，*Src* から *Trg* と *Pvt* への同時翻訳確率を推定することによって翻訳を行う新しいテーブル合成手法を提案する．図 4.3 に，本提案手法によって得られるフレーズ対応の例を示す．本手法の利点は，英語のように中間言語として選ばれる言語は豊富な単言語資源も得られる傾向が強いため，このような追加の言語情報を翻訳システムに組み込み，翻訳の質を向上させられることにある．

中間言語フレーズの情報を翻訳時に役立てるため，SCFG(2.3 節) を複数の目的

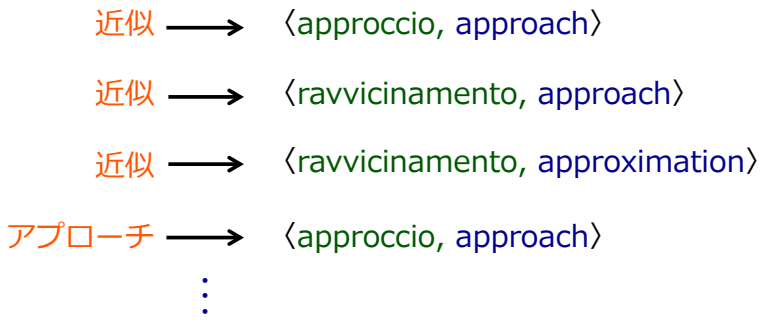


図 4.3　提案手法によって得られるフレーズ対応

言語文の同時生成に対応できるよう拡張した MSCFG(2.4 節) を用いて翻訳モデルの合成を行う．MSCFG による翻訳モデルを構築するためには，*Src-Pvt, Pvt-Trg* の SCFG 翻訳規則が格納されたルールテーブルを元に，SCFG ルールテーブルとしてではなく，*Src-Trg-Pvt* の MSCFG ルールテーブルとして合成し，これによって *Pvt* フレーズを記憶する．訳出候補の探索時には，生成文の自然性を評価し，適切な語彙選択を促すために言語モデル (2.1.2 節) を用いるが，目的言語モデルのみでなく，中間言語モデルも同時に用いた探索を行う．次節から，SCFG 翻訳モデルの同期規則から MSCFG 翻訳モデルの複数同期規則を合成するための手順について説明する．

4.3　同期規則を用いた複数同期規則の合成

3.2.1 節では，SCFG の同期規則を合成するために，*Src-Pvt, Pvt-Trg* ルールテーブル双方に共通の *Pvt* 記号列を有する導出規則 $X \to \langle \bar{s},\ \bar{p} \rangle$, $X \to \langle \bar{p},\ \bar{t} \rangle$ を見つけ出し，新しい導出規則 $X \to \langle \bar{s},\ \bar{t} \rangle$ の翻訳確率を，式 (3.1)-(3.4) に従って確率周辺化を行い推定することを述べた．一方，中間言語情報を記憶するテーブル合成手法では $X \to \langle \bar{s},\ \bar{p} \rangle$, $X \to \langle \bar{p},\ \bar{t} \rangle$ を元に，以下のように複数同期規則を合成する．

$$X \longrightarrow \langle \bar{s},\ \bar{t},\ \bar{p} \rangle \tag{4.1}$$

このような規則を用いて翻訳を行うことによって，同時生成される中間言語文を

通じて中間言語モデルなどのような追加の素性を取り入れることが可能となる．式 (3.1)-(3.4) に加えて，*Trg* と *Pvt* を同時に考慮した翻訳確率 $\phi\left(\overline{t},\overline{p}|\overline{s}\right)$，$\phi\left(\overline{s}|\overline{p},\overline{t}\right)$ を以下のように推定する．

$$\phi\left(\overline{t},\overline{p}|\overline{s}\right) = \phi\left(\overline{t}|\overline{p}\right)\phi\left(\overline{p}|\overline{s}\right) \tag{4.2}$$

$$\phi\left(\overline{s}|\overline{p},\overline{t}\right) = \phi\left(\overline{s}|\overline{p}\right) \tag{4.3}$$

Src-Pvt の翻訳確率 $\phi\left(\overline{p}|\overline{s}\right)$，$\phi\left(\overline{s}|\overline{p}\right)$，$\phi_{lex}\left(\overline{p}|\overline{s}\right)$，$\phi_{lex}\left(\overline{s}|\overline{p}\right)$ はルールテーブル T_{SP} のスコアをそのまま用いることが可能である．これら 10 個の翻訳確率 $\phi\left(\overline{t}|\overline{s}\right)$，$\phi\left(\overline{s}|\overline{t}\right)$，$\phi\left(\overline{p}|\overline{s}\right)$，$\phi\left(\overline{s}|\overline{p}\right)$，$\phi_{lex}\left(\overline{t}|\overline{s}\right)$，$\phi_{lex}\left(\overline{s}|\overline{t}\right)$，$\phi_{lex}\left(\overline{p}|\overline{s}\right)$，$\phi_{lex}\left(\overline{s}|\overline{p}\right)$，$\phi\left(\overline{t},\overline{p}|\overline{s}\right)$，$\phi\left(\overline{s}|\overline{p},\overline{t}\right)$ に加えて，$\overline{t}$ と $\overline{p}$ に含まれる非終端記号数を 2 つのワードペナルティとし，定数 1 のフレーズペナルティの，合わせて 13 個のスコアが MSCFG ルールにおける素性となる．

4.4 同期規則のフィルタリング

前節で説明した，中間言語情報を記憶するテーブル合成手法は，このままでは $\left\langle\overline{s},\overline{t}\right\rangle$ ではなく，$\left\langle\overline{s},\overline{t},\overline{p}\right\rangle$ の全組み合わせを記録するため，従来より大きなルールテーブルが合成されてしまう．計算資源を節約するためには，幾つかのフィルタリング手法が考えられる．Neubig らによると，主要な目的言語 T_1 と補助的な目的言語 T_2 で翻訳を行う際には，T_1-フィルタリング手法 [21] が効果的である．このフィルタリング手法を，提案するテーブル合成手法に当てはめると，$T_1 = Trg$，$T_2 = Pvt$ であり，原言語フレーズ $\overline{s}$ に対して，先ず Trg において $\phi\left(\overline{t}|\overline{s}\right)$ が上位 L 個までの $\overline{t}$ を残し，それぞれの $\overline{t}$ に対して $\phi\left(\overline{t},\overline{p}|\overline{s}\right)$ が最大となるような $\overline{p}$ を残す．

第 5 章　実験的評価

5.1　実験設定

第 4 章で提案した中間言語情報を記憶するテーブル合成手法の有効性を検証するため，多言語コーパスを用いたピボット翻訳の比較評価実験を実施した．用いたデータやツールは，3.2.2 節の実験と大部分が共通しているが，本節で改めて説明する．本実験では先ず，欧州議会議事録を元にした Europarl コーパス [24] を用いて，英語 (En) を中間言語とするドイツ語 (De)，スペイン語 (Es)，フランス語 (Fr)，イタリア語 (It) の 4 言語の組み合わでピボット翻訳の翻訳精度を比較した．翻訳モデルと目的言語の学習用に 10 万文，最適化用および評価用に 1,500 文ずつを互いに重複しないように取り出した．また，多くの場合，英語においては大規模な単言語資源が取得可能であるため，最大 200 万文までのデータを用いて段階的に学習を行った複数の中間言語モデルを用意した．SCFG および MSCFG を用いるデコーダとして Travatar [23] を用い，付属の Hiero ルール抽出プログラムを用いて SCFG 翻訳モデルの学習を行った．翻訳結果の比較には，自動評価尺度 BLEU [14] を用い，各翻訳モデルは MERT [13] により，開発用データに対して BLEU スコアが最大となるようにパラメータ調整を行った．提案手法のテーブル合成手法によって得られた MSCFG ルールテーブルは，$L = 20$ の T_1-フィルタリング手法によって枝刈りを行った．本実験では以下の 6 つの翻訳手法を比較評価する．

Direct
: 翻訳精度の上限値を得て比較を行うため，中間言語を用いず *Src-Trg* の直接対訳コーパスで学習した SCFG で翻訳

Cascade:
: *Src-Pvt* および *Pvt-Trg* の SCFG モデルで逐次的ピボット翻訳 (3.1.1 節)

Tri. SCFG:
: *Src-Pvt* および *Pvt-Trg* の SCFG モデルを合成し，*Src-Trg* の SCFG モデルによって合成 (3.1.3 節)

Tri. MSCFG:
: *Src-Pvt* および *Pvt-Trg* の SCFG モデルを合成し，*Src-Trg-Pvt* の MSCFG

Src	Trg	BLEU Score [%]					
		Direct	Cascade	Tri. SCFG (baseline)	Tri. MSCFG -PivotLM	Tri. MSCFG +PivotLM 100k	Tri. MSCFG +PivotLM 2M
De	Es	27.10	25.05	25.31	25.38	25.52	† **25.75**
	Fr	25.65	23.86	24.12	24.16	24.25	† **24.58**
	It	23.04	20.76	21.27	21.42	† 21.65	‡ **22.29**
Es	De	20.11	18.52	18.77	18.97	19.08	† **19.40**
	Fr	33.48	27.00	29.54	† 29.87	† 29.91	† **29.95**
	It	27.82	22.57	25.11	25.01	25.18	‡ **25.64**
Fr	De	19.69	18.01	18.73	18.77	18.87	† **19.19**
	Es	34.36	27.26	30.31	30.53	† 30.73	‡ **31.00**
	It	28.48	22.73	25.31	25.50	† 25.72	‡ **26.22**
It	De	19.09	14.03	17.35	† 17.99	‡ 18.17	‡ **18.52**
	Es	31.99	25.64	28.85	28.83	29.01	† **29.31**
	Fr	31.39	25.87	28.48	28.40	28.63	† **29.02**

表 5.1　各手法による翻訳精度

モデルによって翻訳 (4 章).「-PivotLM」は中間言語モデルを用いないことを示し,「+PivotLM 100k/2M」はそれぞれ 10 万文，200 万文で学習した中間言語を用いることを示す.

5.2　翻訳精度の比較

表 5.1 に，英語を介したすべての言語対におけるピボット翻訳の結果を示す．太字はそれぞれの言語対において最も 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 ($\dagger : p < 0.05, \ddagger : p < 0.01$)．評価値から，提案したテーブル合成手法で中間言語モデルを考慮した翻訳を行った場合，すべての言語対において従来のテーブル合成手法よりも BLEU スコアが上昇していることが確認できる．すべての組み合わせにおいて，テーブル合成手法で中間言語情報を記憶し，200 万文の言語モデルを考慮して翻訳を行った場合に最も高いスコアを達成しており，従来法に比べ 0.4 から 1.2 ほどの BLEU 値の向上が見られる。このことから，中間言語情報を記憶し，これを翻訳に利用す

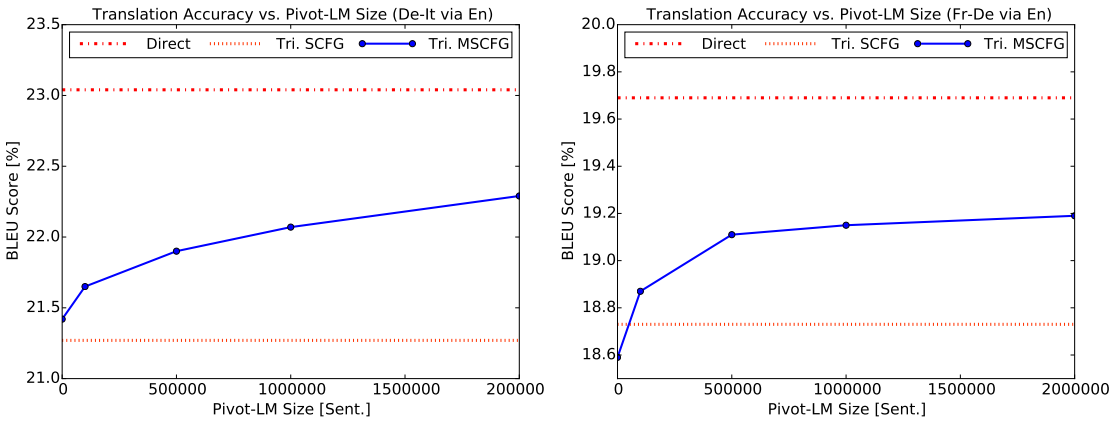


図 5.1　中間言語モデル規模がピボット翻訳精度に与える影響

ることが曖昧性の解消に繋がり，安定して翻訳精度を改善できたと考えられる．

また，異なる要因による影響を切り分けて調査するため，MSCFG へ合成するが，中間言語モデルを用いずに翻訳を行った場合の比較も行った．この場合，保存された中間言語情報が語彙選択に活用されないため，本手法の優位性は特に現れないものと予想されたが，実際には，SCFG に合成する場合よりも多くの言語対で僅かに高い翻訳精度が得られた．これは，追加の翻訳確率などのスコアが有効な素性として働き，パラメータ調整を行った上で，適切な語彙選択に繋がったことなどが原因として考えられる．

5.3　中間言語モデルの規模が翻訳精度に与える影響

中間言語モデルの規模がピボット翻訳精度に与える影響の大きさは言語対によって異なってはいるが，中間言語モデルの学習データサイズが大きくなるほど精度が向上することも確認できる．図 5.1 は，ドイツ語・イタリア語 (左) およびフランス語・ドイツ語 (右) のピボット翻訳において異なるデータサイズで学習した中間言語モデルが翻訳精度に与えるを示す．図からも中間言語モデルが曖昧性を解消して翻訳精度向上に寄与している様子が見られるが，一方で 100 万文および 200 万文で学習した中間言語モデルを用いた場合の精度の差はあまり大きくないため，さらに大規模な中間言語モデルを用いても，さらなる精度向上の見込みは薄く，計算コスト

が大きくなる欠点の方が相対的に大きくなると考えられる.

5.4 曖昧性が解消された例と未解決の問題

本提案手法によって中間言語側で曖昧性が解消されて翻訳精度向上に繋がったと考えられる訳出の例を示す.

入力文 (ドイツ語):

Ich bedaure , daß es keine **gemeinsame Annäherung** gegeben hat .

参照文 (イタリア語):

Sono spiacente del mancato **approccio comune** .

対応する英文:

I regret that there was no **common approach** .

Tri. SCFG:

Mi rammarico per il fatto che non si **ravvicinamento comune** . (BLEU+1: 13.84)

Tri. MSCFG+PivotLM 2M:

Mi dispiace che non esiste un **approccio comune** . (BLEU+1: 25.10)

I regret that there is no **common approach** . (同時生成された英文)

上記の Tri. MSCFG+PivotLM 2M で導出に用いられた MSCFG ルールでは, イタリア語「approccio」と英語「approach」が結び付いており, 生成される英文中の単語の前後関係から適切な語彙選択を促し, 精度向上に繋がったものと考えられる.

逆に, 提案手法では語彙選択がうまくいかず, 直接対訳で学習した場合よりも精度が落ちた訳出の例を示す.

入力文 (フランス語):

Vous avez tout à fait raison et **je vous remercie** d'avoir attiré l'attention sur ce point.

参照文 (スペイン語):

Tiene usted toda la razón y **le agradezco** que nos llame la atención sobre este punto.

対応する英文:

You are quite right and **I thank you** for drawing our attention to this matter .

Direct:

Tiene usted razón y **le agradezco** que haya llamado la atención sobre este punto . (BLEU+1: 56.00)

Tri. MSCFG+PivotLM 2M:

Tiene usted mucha razón y **gracias** por haber conseguido la atención sobre este punto . (BLEU+1: 38.91)

You have quite right and **I thank you** for having courageously brought the attention on this point . (同時生成された英文)

この例では，英単語「thank」に対応するスペイン語の動詞活用形「agradezco (私は感謝する)」と名詞の間投詞的用法「gracias (感謝，ありがとう)」に訳が分かれてしまっている．この例では品詞が異なっても意味としては大きく変化しておらず，単純に 1 つしか存在しない正解訳と異なるためにスコアが下がっている．しかし，このような多品詞語によって生じる曖昧性の問題もピボット翻訳では多く直面し，品詞を誤ることで文法を誤る可能性だけでなく，異なる意味の文になる場合もある．

本節で示した例のように，本研究の提案手法でテーブル合成手法に中間言語情報と中間言語モデルを用いることで，導出に関係する中間言語フレーズの単語の並びを評価して曖昧性を解消できた部分と，これだけでは不十分な部分があることが分かる．

品詞	出現頻度	F-Measure [%]		
		Direct	Tri. MSCFG +PivotLM 2M	Tri.SCFG
NC (一般名詞)	8,360	78.94	78.19 (+0.01)	78.18
P (前置詞)	6,553	92.58	92.46 (+1.02)	91.44
DET (定冠詞)	5,363	91.04	90.63 (+1.50)	89.13
PUNC (区切り記号)	4,258	93.89	93.77 (+0.44)	93.33
ADJ (形容詞)	3,725	71.36	69.85 (+0.67)	69.18
V (動詞)	3,324	69.52	63.86 (+1.06)	62.80
ADV (副詞)	2,238	81.98	77.81 (+0.34)	77.47

表 5.2　独仏翻訳における品詞ごとの翻訳精度

5.5　品詞ごとの翻訳精度

前節では，英語の多品詞語の問題について触れた．また，英語は他の欧州諸言語と比較して，性・数・格に応じた活用などが簡略された言語として有名であり，語形から統語情報が失われることで発生する曖昧性の問題もある．本節では，英語を介したドイツ語・フランス語の両方向の翻訳において，誤りの発生しやすい品詞について調査する．先ず，独仏・仏独翻訳における評価データの参照訳および各ピボット翻訳手法の翻訳結果に対して Stanford POS Tagger [32, 33] を用いて品詞付与を行い，参照訳と翻訳結果を比較して，語順は考慮せずに適合率と再現率の調和平均である F 値を算出した．

表 5.2 および表 5.3 は，各翻訳における高頻出品詞の正解率を表している．出現頻度は，参照訳中の各品詞の出現回数を意味し，丸括弧内に示された数値は，提案手法とベースライン手法の F 値の差分である．結果は言語対依存であるが，特に目的言語に強く依存していることが明らかである．

表 5.2 の独仏翻訳の例では，提案手法によって，ベースライン手法よりも特に前置詞，定冠詞，動詞で F 値が大きく向上している．一般名詞，形容詞，動詞，副詞は重要な内容語であり，語彙選択の幅も広いため，どの手法でも全体的に F 値が低

品詞	出現頻度	F-Measure [%]		
		Direct	Tri. MSCFG +PivotLM 2M	Tri.SCFG
NN (一般名詞)	10,813	78.04	75.42 (+0.53)	74.89
ART (冠詞)	4,502	90.38	87.21 (+1.24)	85.97
CARD (数詞)	3,849	94.56	94.38 (−0.44)	94.82
APPR (前置詞)	2,577	89.66	84.73 (+1.50)	83.23
ADJA (形容詞)	2,311	65.96	65.29 (+0.47)	64.82
NE (固有名詞)	2,005	77.52	75.34 (+0.47)	74.87
ADV (副詞)	1,848	74.32	70.13 (+0.18)	69.95
PPEF (再帰代名詞)	1,665	92.64	87.46 (+1.67)	85.79

表 5.3　仏独翻訳における品詞ごとの翻訳精度

くなっている．一般名詞に関しては，通常のテーブル合成手法でも Direct と大きな差は出ておらず，そのため提案手法でもほとんど改善されなかった．一方で，動詞の F 値は提案手法で大きく向上しており，中間言語モデルによって語の並びを考慮して語彙選択を行うことで翻訳精度向上に繋がったと考えられる．しかし，それでも Direct には大きく及ばず，頻度のあまり高くない内容語の語彙選択を適切に行うことは非常に難しいことが分かる．一方で，機能語においては提案手法において Direct と近い F 値が得られた．

表 5.3 の仏独翻訳の例では，表 5.2 の場合と少し変わっており，冠詞，前置詞，再帰代名詞のような機能語で，提案手法によって F 値が大きく改善されているものの，Direct と比べると大きな差があり，ピボット翻訳で精度が大きく落ちる原因と考えられる．冠詞や前置詞は機能語であるものの，ドイツ語では男性系・女性系に加えて，英語にもフランス語にも無い中世系の活用を持っているため，フランス語よりも活用の種類が多く，また冠詞や前置詞も格に応じた活用をすることが知られている．原言語を英語に翻訳した際には統語的情報が失われることが多いため，機能語にも活用幅があるような目的言語に対してはピボット翻訳が特に困難になるこ

とが多い.

5.6 考察

本章では，第 4 章で提案した，中間言語情報を記憶したテーブル合成手法と，中間言語モデルを考慮したピボット翻訳手法について，複数の言語対について従来のテーブル合成手法との比較実験を実施した．その結果，すべての言語対において，従来のテーブル合成手法よりも高い翻訳精度が得られ，特に大規模な中間言語モデルを用いることでピボット翻訳の質を高められることが示された．しかし，提案手法においても，直接の対訳を用いて学習した翻訳モデルと比較すると，未だ大きな開きがあり，さらに大規模な中間言語モデルを用いても差を縮めることは困難であると考えられる．これは，中間言語の語順を考慮することで解消できた曖昧性の問題がある一方，それだけでは不十分な曖昧性の問題もあり，Direct モデルとの翻訳精度の差だけ未解決の問題があるということである．

前節では，品詞ごとの単語正解率の分析を行ったが，言語対ごとに特定品詞の翻訳精度が落ちる減少が見られた．英語には同じ語形で多品詞の語が多いため，単語の対応だけで統語的な役割を判断するのは不可能な場合もある．統語的な情報を汲み取るには複数の単語からなるフレーズを考慮する必要があるが，文頭と文末のような離れた位置での依存関係も存在するため，単語列としてではなく構文構造を考慮することも重要と考えられる．そこで，本研究の提案手法の今後の課題として，構文構造を中間言語の表現として用いるピボット翻訳手法を検討している．

第 6 章　結言

6.1　本論文のまとめ

本研究の目的は，多言語機械翻訳における翻訳精度の向上を目指し，従来のピボット翻訳手法を調査，問題点を改善して翻訳精度を図ることであった．そのために，PBMT で既に有効性が示されている，テーブル合成手法によるピボット翻訳を SCFG に適用し，どのような処理が有効であるかに着目した．さらに，従来のテーブル合成手法では中間言語情報が失われ，曖昧性により翻訳精度が減少する問題に対処するため，中間言語情報を記憶し，中間言語モデルを利用して自然な語順の語彙選択を促すことで精度を向上させるテーブル合成手法についても提案した．

第 2 章では, SMT の要素技術について説明し, 単語列に基いて翻訳を行う PBMT と，木構造に基いて翻訳を行う SCFG の 2 つの翻訳方式について比較を行った．また，SCFG を複数の目的言語生成に拡張させた MSCFG について説明した．そして，任意の言語対で高精度な機械翻訳を実現するための課題と，ピボット翻訳の必要性について述べた．

第 3 章では，PBMT で提案されている代表的なピボット翻訳手法について説明し，その中でも特に高い翻訳精度を達成できることで知られるテーブル合成手法を SCFG で応用するための手順について述べた．また，複数の言語と翻訳枠組とピボット翻訳手法の組み合わせによる比較実験を実施し，SCFG におけるテーブル合成手法の有効性を検証した．

第 4 章では，従来のテーブル合成手法の問題点を指摘した上で，その問題点を解消すべく，中間言語情報を記憶するテーブル合成手法の提案を行った．本手法は個別に学習した 2 つの SCFG ルールテーブルを，1 つの MSCFG ルールテーブルに合成し，原言語から目的言語と中間言語に同時に翻訳を行うものである．

第 5 章では，実験により提案法の有効性を検証した．欧州議会議事録を元にした多言語コーパスで 5 言語のデータを用いてピボット翻訳手法の比較評価を行い，その結果，すべての言語の組み合わせで従来のテーブル合成手法よりも高い翻訳精度が得られた．また，特に大規模な中間言語モデルを用いることで，より適切な語彙選択が促されて翻訳の質を高められることが分かった．しかし，提案手法でも，直接の対訳コーパスを用いて学習を行った理想的な状況と比較すると，精度の開きが

大きく，中間言語モデルをさらに大きくするだけでは解決できないであろう点も示唆された．本提案手法で解決できなかった曖昧性の問題を特定すべく，特定の言語対で品詞ごとの単語正解率を求めたところ，語順を考慮するだけでは解消されない曖昧性の問題もあり，構文情報を用いてこの点を改善することを今後の課題として述べた．

6.2 今後の課題

今後の課題を以下に挙げる．

ピボット翻訳の曖昧性の問題は，主として中間言語の表現力に起因しており，中間言語の単語列だけでは原言語の情報が失われてしまい，目的言語側で確率的に正しく再現することは困難である．そのため，中間言語を特定の言語の単語列としてではなく，より高い表現力を持った構文構造を中間表現とすることで，中間言語側の多品詞語の問題に対処したり，原言語側の情報を保存して，より正確に目的言語側で情報を再現するための手法を検討する．

1 つ目は，中間表現に統語情報を用いた翻訳規則テーブルの軽量化・高精度化である．本研究で提案した手法では，SCFG の翻訳モデルを学習するために，階層的フレーズベース翻訳という枠組みの翻訳手法で翻訳規則を獲得している．これは，翻訳において重要な，単語並び替え問題を高精度に対処できる点で優れているが，統語情報を用いず，総当り的な手段で非終端記号を含んだ翻訳規則を学習するため，テーブルサイズが肥大化する傾向がある．その上，テーブル合成時には，中間表現の一致する組み合わせによってテーブルサイズはさらに増加する．これは，曖昧性により多くの誤ったフレーズ対応も保存されることを意味するため，不要な規則は除去してサイズを削減するべきである．このような，中間表現が一致する組み合わせの中には，文法上の役割は異なるが表記上同じようなものも含まれるため，意味の対応しない組み合わせに対して高い翻訳確率が推定されてしまう場合もある．こういった問題は，中間言語の表現に統語情報を組み込むことで，品詞や句構造が異なればフレーズの対応も結び付かないという制約が働き，誤った句対応を容易に除外できるため，テーブルサイズは減少し，曖昧性が解消されて翻訳精度の向上が期待できる．

2 つ目は，中間表現に原言語の統語情報を保存するピボット翻訳手法の提案である．ピボット翻訳においては，中間言語の表現力が悪影響を及ぼして，原言語の情報が失われてしまうことを述べてきたが，これは機械翻訳に限らず，人手による翻訳でも度々起こる問題である．例えば，英語には人称接尾辞のような活用体系が無いため，英語に訳した際に性・数・格などの統語的情報が失われ，結果的に英語を元にした翻訳では原意と大きく異なってしまう現象などがある．本枠組みでは，前述の中間言語側の統語情報と組み合わせることで，より原意を汲んだ翻訳の実現を目指す．

謝辞

本学情報科学研究科の中村哲教授には主指導教官として，研究全般に渡り大変貴重な御助言を頂き，本論文を執筆することができました．心より感謝いたします．

本学情報科学研究科の松本裕治教授には副指導教官として，研究全般に渡り貴重な御助言を頂きました．心より感謝いたします．

名古屋大学情報科学研究科の戸田智基教授には，本学勤務時において，研究全般に渡り貴重な御助言を頂きました．心より感謝いたします．

本学情報科学研究科の Graham Neubig 助教には研究における御指導や，大変貴重な御助言を頂きました．また対外発表論文執筆や語学面においても的確なアドバイスを頂きました．心より感謝いたします．

本学情報科学研究科の Sakriani Sakti 助教，鈴木優特任准教授，吉野幸一郎特任助教には研究における貴重な御助言を頂きました．心より感謝いたします．

知能コミュニケーション研究室の秘書である松田真奈美様には，諸手続き等で大変お世話になりました．心より感謝いたします．

知能コミュニケーション研究室の先輩諸氏には，公私に渡り大変お世話になりました．そして同期諸君には，研究面においてのアドバイスなどお世話になりました．心より感謝いたします．

そして母と，修士論文準備中に急逝した父には，返しきれない恩を感じつつ，心より感謝いたします．

参考文献

[1] Sergei Nirenburg. Knowledge-Based Machine Translation, *Machine Translation*, Vol. 4, No. 1, pp. 5–24, 1989.

[2] 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.

[3] 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.

[4] 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*, 2006.

[5] Trevor Cohn and Mirella Lapata. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora In *Proc. ACL*, pp. 728–735, 2007.

[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] Xiaoning Zhu, Zhongjun He, Hua Wu, Conghui Zhu, Haifeng Wang, and Tiejun Zhao. Improving Pivot-Based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs In *Proc. EMNLP*, 2014.

[8] David Chiang. Hierarchical phrase-based translation, *Computational Linguistics*, Vol. 33, No. 2, pp. 201–228, 2007.

[9] Makoto Nagao. A framework of a Mechanical Translation between Japanese and English by Analogy Principle In *Proc. International NATO Symposium on Artificial and Human Intelligence*, pp. 173–180, 1984.

[10] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to Sequence

Learning with Neural Networks In *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate, *arXiv preprint arXiv:1409.0473*, 2014.

[12] Claude E. Shannon. A Mathematical Theory of Communication, *Bell System Technical Journal*, Vol. 27, No. 3, pp. 379–423, 1948.

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

[14] 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.

[15] Stanley F. Chen and Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling In *Proc. ACL*, pp. 310–318, 1996.

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

[17] Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. What's in a Translation Rule? In *Proc. NAACL*, pp. 273–280, 2004.

[18] Isao Goto, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. Distortion Model Considering Rich Context for Statistical Machine Translation In *Proc. ACL*, pp. 155–165, 2013.

[19] Jean-Cédric Chappelier, Martin Rajman et al. A Generalized CYK Algorithm for Parsing Stochastic CFG., *Proc. TAPD*, Vol. 98, No. 133-137, p. 5, 1998.

[20] Haitao Mi, Liang Huang, and Qun Liu. Forest-Based Translation In *Proc. ACL*, pp. 192–199, 2008.

[21] Graham Neubig, Philip Arthur, and Kevin Duh. Multi-Target Machine Translation with Multi-Synchronous Context-free Grammars In *Proc. NAACL*, 2015.

[22] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Mar-

cello 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*, pp. 177–180, 2007.

[23] Graham Neubig. Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers In *Proc. ACL Demo Track*, pp. 91–96, 2013.

[24] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation In *MT summit*, Vol. 5, pp. 79–86, 2005.

[25] William A Gale and Kenneth W Church. A program for aligning sentences in bilingual corpora, *Computational linguistics*, Vol. 19, No. 1, pp. 75–102, 1993.

[26] Tomer Levinboim and David Chiang. Supervised Phrase Table Triangulation with Neural Word Embeddings for Low-Resource Languages In *Proc. EMNLP*, pp. 1079–1083, 2015.

[27] Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages In *Proc. NAACL*, pp. 1192–1202, 2015.

[28] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations In *Proc. NAACL*, pp. 746–751, 2013.

[29] Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. On the Importance of Pivot Language Selection for Statistical Machine Translation In *Proc. NAACL*, pp. 221–224, 2009.

[30] Eduard Hovy. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses In *Proc. LREC*, pp. 535–542, 1998.

[31] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-Task Learning for Multiple Language Translation In *Proc. ACL*, pp. 1723–1732, 2015.

[32] Kristina Toutanova and Christopher D. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger In *Proc. EMNLP*, pp. 63–70, 2000.

[33] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network In *Proc. NAACL*, pp. 173–180, 2003.

発表リスト

国際会議

[1] Akiva Miura, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura. Improving Pivot Translation by Remembering the Pivot, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL2015), 2015 年 7 月, 北京.
[2] Akiva Miura, Graham Neubig, Michael Paul, Satoshi Nakamura. Selecting Syntactic, Non-redundant Segments in Active Learning for Machine Translation, Proceedings of the 15th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2016), 2016 年 6 月 (発表予定), San Diego, USA.

大会講演

[1] 三浦 明波, Graham Neubig, Michael Paul, 中村 哲. 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), 2016 年 3 月, 仙台.

研究会

[1] 三浦 明波, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲. 階層的フレーズベース翻訳におけるピボット翻訳手法の応用, 情報処理学会第 219 回自然言語処理研究会 (SIG-NL), 2014 年 12 月, 神奈川.
[2] 三浦 明波, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲. 中間言語モデルを用いたピボット翻訳の精度向上, 情報処理学会第 222 回自然言語処理研究会 (SIG-NL), 2015 年 7 月, 東京.
[3] 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 年 12 月, 名古屋.