CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Tsukuba Repository

統計的機械翻訳における語義に基づく大局的な情報を用いた語彙モデルに関する研究

著者	大山 鉄郎
内容記述	筑波大学修士（情報学）学位論文・平成26年3月25日授与（32640号）
発行年	2014
URL	http://hdl.handle.net/2241/00123829

統計的機械翻訳における
語義に基づく大局的な情報を用いた
語彙モデルに関する研究

筑波大学
図書館情報メディア研究科
２０１４年３月
大山 鉄郎

目 次

表 目 次

図 目 次

第1章　はじめに

1.1　本研究の目的と概要

本研究は，統計的機械翻訳において，語義に基づく大局的な情報を用いた語彙モデルを提案する．語義に基づく大局的な情報を用いることで，単語やフレーズだけでは翻訳できない曖昧な語や，文の時制など，文脈を考慮した翻訳を実現する．

統計的機械翻訳では，入力文に対する仮説（翻訳候補）を複数のモデルで評価し，最も評価値が高くなる仮説を出力文とする．一般的なモデルとして，単語の連接を評価する言語モデルや，単語がどの単語に翻訳されやすいかを評価する翻訳モデルがある．言語モデルは，単語の連接に確率を与えることで，尤もらしい文を生成するするように働く．翻訳モデルは，文をフレーズ単位で扱うことで単語より大きな単位で翻訳するように働く．そのため，これらのモデルもある程度の文脈を考慮しているといえる．しかし，この場合の文脈とは，翻訳する単語周辺の数単語だけの局所的なものであり，翻訳に有用である文脈を全て考慮しているとはいえない．

本研究では，大局的な情報 [2, 15] を語義から構築することで，広範囲の文脈を捉えることができる語彙モデルを提案する．語義を利用することで，表層形は異なるが意味は同じ単語を 1 つの素性と見なす．同じ語義は同じ文脈に出現すると仮定して目的言語を推定することで，より柔軟に文脈を考慮した目的言語の推定が行える．提案手法の有用な例として，離れた場所に生起する副詞を考慮した時制の選択，トピック的に共起しやすい単語の選択などが挙げられる．

提案手法では，語義に基づく大局的な情報として，翻訳する単語と同じ文に含まれている，原言語の全ての単語の表層形と語義を用いる．目的言語の単語は，表層形と語義の集合から推定される．提案手法では，目的言語の単語それぞれに対して，原言語に含まれる全ての単語が素性として用いられるため，広い範囲の文脈を考慮した翻訳が行える．一方，単語に対する素性は，同じ文に出現する全ての単語になるため，その組み合わせは膨大な数になり，有効な素性として働く単語は一部に限られる．

提案手法では，表層形に加えて語義を用いることで，この問題を軽減している．原言

語の単語を語義に変換することで，意味が同じ単語を同一の素性として扱う．意味が同じ単語は出現する文脈にもある程度の類似があると考えられ，語義によってまとめあげることで，素性の種類を抑えつつ，柔軟な文脈を考慮した目的言語の推定ができる．単語を語義に変換する際には，誤った語義に変換することも考えられる．提案手法では，目的言語の語義を用いることで，単語が不適切な語義に変換されることを防いでいる．

本論文では，語義に基づく大局的な情報を用いた語彙モデルの提案を行い，提案手法を組み込んだ翻訳システムが，既存の翻訳システムと比較して，翻訳精度が向上することを示す．

1.2 研究背景

情報技術の発展により，情報の入手は飛躍的に容易になった．現在では，国や地域を超えて世界中の情報にアクセスすることができる．しかし，それらの情報は，基本的に発信者の母語や，利用者の多い言語（i.e. 英語）で記述されている．そのため，アクセスすることはできるが，利用することはできないという，言語の壁による情報格差は未だ大きな問題として残っている．

この問題を軽減するための方法として，機械的に言語を別の言語に翻訳する,「機械翻訳」と呼ばれる技術がある．初期の機械翻訳は，翻訳の規則を人手で定義して，その規則通りに変換する方法が主流だった．この手法はルールベース翻訳と呼ばれ，現在でも多くの製品で用いられている．ルールベースによる翻訳は，定義された規則の範囲内であれば，適切な訳文を返すことができる．しかし，各言語対ごとに規則を定義する必要があり，同じ単語でも文脈によって翻訳が異なる場合もある．そのため，必要な規則は非常に膨大な量になり，すべての規則を定義することは事実上不可能であるといえる．

ルールベース翻訳の問題を踏まえて，1990 年代に，二言語間で同じ意味を持つ文の集合である対訳コーパスから，単語対応に基づいてモデルを構築して翻訳を行う，統計的な手法 [3] が提案された．この手法では，対訳コーパスを準備すれば後は自動で学習，翻訳を行い，候補から最も確率の高い文を翻訳結果として出力する．このため，ルールベース翻訳と比較して非常に小さなコストでシステムを構築することができる．この手法は，統計的機械翻訳（Statistical Machine Translation: SMT）と呼ばれている．

統計的機械翻訳の 原言語$\boldsymbol{f}$から目的言語$\boldsymbol{e}$ への翻訳を以下に示す.

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} P(\boldsymbol{e}|\boldsymbol{f}) \tag{1.1}$$

$$= \arg\max_{\boldsymbol{e}} \frac{P(\boldsymbol{e})P(\boldsymbol{f}|\boldsymbol{e})}{P(\boldsymbol{f})} \tag{1.2}$$

$$= \arg\max_{\boldsymbol{e}} P(\boldsymbol{e})P(\boldsymbol{f}|\boldsymbol{e}) \tag{1.3}$$

式（1.3）の $P(\boldsymbol{e})$ を言語モデル, $P(\boldsymbol{f}|\boldsymbol{e})$ を翻訳モデルと呼ぶ．言語モデルは 目的言語$\boldsymbol{e}$ の言語としての尤もらしさを，翻訳モデルは 原言語$\boldsymbol{f}$から目的言語$\boldsymbol{e}$ への翻訳確率を表している．$\arg\max_{\boldsymbol{e}} P(\boldsymbol{e})P(\boldsymbol{f}|\boldsymbol{e})$ が最大になるような仮説を求め，それを翻訳結果とする．統計的機械翻訳では，単語の翻訳を繰り返して文の翻訳を行う．そのため，慣用句など，複数の単語で意味を持つ部分の翻訳に対して不適切に働くことがあった．その後，2000年代に，統計的機械翻訳の発展として，翻訳単位を単語からフレーズに拡張させた翻訳手法 [19] が提案された．フレーズは，対訳コーパスの文中における単語対応から一定のルールで抽出される．また，同時期に対数線形モデルを利用することで，複数の素性を統合するモデル [17] が提案されており，フレーズモデルを用いる場合にはこのモデルを用いることが多い．これらの手法は統計的機械翻訳手法の主流となっているが，先の統計的機械翻訳手法と区別して，フレーズベース統計的機械翻訳（Phrase Based Statistical Machine Translation: PBSMT）と呼ばれることもある．

フレーズベース統計的機械翻訳の翻訳は以下で表される．

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} \sum_{i=0} \lambda_i h_i(\boldsymbol{e}, \boldsymbol{f}) \tag{1.4}$$

$\boldsymbol{e}$ は出力である目的言語文の仮説，$\boldsymbol{f}$ は入力である原言語文を表す．$h(\boldsymbol{e}, \boldsymbol{f})$ は素性関数であり，λ は各素性関数に対する重みである．統計的機械翻訳の式（1.3）は式（1.4）の対数線形モデルにおいて，$h_1(\boldsymbol{e}, \boldsymbol{f}) = P(\boldsymbol{e})$，$h_2(\boldsymbol{e}, \boldsymbol{f}) = P(\boldsymbol{f}|\boldsymbol{e})$ とすることで表現できる．本研究では，フレーズベース統計的機械翻訳の枠組みで研究を行うが，表記の簡略化のため，以下では統計的機械翻訳と示す.

本研究では，提案する語彙モデルを式（1.4）に組み込める素性関数の形で実装する．式（1.4）によって，提案手法と，既存の翻訳モデル，言語モデルは調整された重みによって組み合わされる.

提案する語彙モデルは，大局的な文脈を考慮して目的言語を推定する．これは，局所的な文脈を考慮する言語モデルや翻訳モデルに対して，相補的な働きをする．提案手法を導入することによって，文中での時制のずれや，訳語の組み合わせにおける翻訳誤りを軽減することができる.

1.3 論文の構成

2章以降の本論文の構成は以下のようになっている．2章では，本研究の関連研究について述べる．関連研究として，統計的機械翻訳の代表的なモデル，語彙モデル，語義曖昧性解消手法に関する研究，機械翻訳の評価尺度について紹介する．3章では，語義に基づく大局的な情報の働きを整理し，提案手法である語義に基づく大局的な情報を利用した語彙モデルについて述べる．4章では，実験として提案手法を組み込んだ統計的機械翻訳システムの翻訳精度について述べる．5章では，実験結果を受けて，提案手法による改善点や問題点についての考察を述べる．最後に，6章で本研究の結論を述べる．

第2章　関連研究

本章では，本研究に関連する研究を紹介する．2.1 節では，統計的機械翻訳システムで用いられる代表的なモデルについて説明する．2.2，2.3 節では，本研究に関連する研究として，語彙モデル，語義曖昧性解消の研究について紹介する．2.4 節では機械翻訳の評価方法について説明する．2.5 節では関連研究を踏まえて，本研究の位置付けを説明する．

2.1　統計的機械翻訳における代表的なモデル

本節では，統計的機械翻訳の代表的なモデルとして，単語がどの単語に翻訳されやすいかを評価する翻訳モデルと，文が自然かどうかを評価する N-gram 言語モデル [5] について説明する．

2.1.1　翻訳モデル

翻訳モデルは，単語がどの単語に翻訳されるのかを表すモデルである．最も一般的に用いられる IBM モデル [3] を説明する．IBM モデルは，式（2.1）のように原言語と目的言語の対応関係を表す潜在変数を導入することで $p(f|e)$ をモデル化する．IBM モデルは 1 から 5 までの 5 つのモデルがあり，数値が大きくなるほど複雑になる．低い数値のモデルで推定したパラメータを高い数値のモデルの初期値にすることで，最適なパラメータを推定する．

$$p(\boldsymbol{f}|\boldsymbol{e}) \;=\; \sum_{a} p(\boldsymbol{f}, a|\boldsymbol{e}) \tag{2.1}$$

式（2.1）は，以下のように展開できる．m は原言語文の長さ，a_j は，f_j と対応する e とのアライメントを表している．

$$p(\boldsymbol{f}, a|\boldsymbol{e}) \;=\; p(m|\boldsymbol{e}) \prod_{j=1}^{m} p(a_j|a_1^{j-1}, f_1^{j-1}, m, \boldsymbol{e}) p(f_j|a_1^{j}, f_1^{j-1}, m, \boldsymbol{e}) \tag{2.2}$$

IBM モデル 1 では，式（2.2）に対して，いくつかの仮定を加えて式を変形する．まず，原言語文の長さは目的言語文には依存しないと仮定して，定数に置き換える．

$$p(m|\boldsymbol{e}) \;=\; \epsilon \tag{2.3}$$

次に，アライメントは，目的言語文の長さ l のみに依存すると仮定する．

$$p(a_j|a_1^{j-1}, f_1^{j-1}, m, \boldsymbol{e}) \;=\; (l+1)^{-1} \tag{2.4}$$

最後に，翻訳確率は，原言語の単語にアライメントされた目的言語の単語のみに依存すると仮定する．

$$p(f_j|a_1^{j}, f_1^{j-1}, m, \boldsymbol{e}) \;=\; t(f_j|e_{a_j}) \tag{2.5}$$

式（2.3）から式（2.5）より，式（2.2）は，以下の式で表される．

$$p(\boldsymbol{f}, a|\boldsymbol{e}) \;=\; \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.6}$$

$$p(\boldsymbol{f}|\boldsymbol{e}) \;=\; \frac{\epsilon}{(l+1)^m}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.7}$$

$$\;=\; \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{l} t(f_j|e_{a_j}) \tag{2.8}$$

なお，目的言語の単語 e と対応する f の確率の和は 1 になるため，$t(f|e)$ は以下を満たす．

$$\sum_{f} t(f|e) = 1 \tag{2.9}$$

式（2.9）の制約を満たし，式（2.8）を最大にする $t(f|e)$ を，ラグランジュ未定乗数法を用いて求め，整理すると以下の式を得る．

$$t(f|e) \;=\; \lambda_e^{-1}\sum_{s=1}^{S} c(f|e; \boldsymbol{f^{(s)}}, \boldsymbol{e^{(s)}}) \tag{2.10}$$

$$\lambda_e \;=\; \sum_{f}\sum_{s=1}^{S} c(f|e; \boldsymbol{f^{(s)}}, \boldsymbol{e^{(s)}}) \tag{2.11}$$

$$c(f|e; \boldsymbol{f}, \boldsymbol{e}) \;=\; \frac{t(f|e)}{t(f|e_0)+\cdot+t(f|e_l)}\sum_{j=1}^{m}\delta(\boldsymbol{f}, f_j)\sum_{i=0}^{l}\delta(\boldsymbol{e}, e_i) \tag{2.12}$$

$$\delta(x,y) \;=\; \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases} \tag{2.13}$$

$\boldsymbol{f}^{(s)}$,$\boldsymbol{e}^{(s)}$ は，s 番目の対訳文，δ は，単語が文に出現するか否かを判定する関数である．$t(f|e)$ に適当な初期値を与え，$t(f|e)$ が収束するまで，式（2.11）と式（2.12）を繰り返すことで最適解を推定する．

IBM モデル 2 から 5 は，IBM モデル 1 に加えて以下のことを考慮したモデルである．

- IBM モデル 2：単語の絶対位置を考慮
- IBM モデル 3：単語が複数の単語と対応する場合を考慮
- IBM モデル 4：単語の相対位置を考慮
- IBM モデル 5：単語位置の重複・欠損を考慮

IBM モデルによって得られる単語対応は方向を持つため，原言語から目的言語，目的言語から原言語の 2 つの対応関係は細部が異なる．この 2 つの異なる対応関係を用いて，より正確な単語対応を求める．また，得られた単語対応からフレーズの抽出を行う．

IBM モデルによって，2 つの単語対応が得られたとする．これを単語対応候補として，単語対応は以下のステップで行われる．

1. 2 つの単語対応候補の積集合を取り，それを単語対応とする（intersection）
2. 得られた単語対応の縦横，または対角方向に単語対応候補の和集合のにおける対応点があればそれを追加する（grow-diag）
3. 得られた単語対応で，両方向に単語対応がない部分に，和集合の対応点があればそれを追加する（final-and）

この単語対応の取り方は，grow-diag-final-and と呼ばれる．

最後に，獲得した単語対応からフレーズの抽出を行う．フレーズ抽出では，まず，複数の単語対応を包含する矩形を定義する．この矩形の縦横方向に単語対応がなかったとき，その矩形に含まれる単語対応をフレーズとする．フレーズの抽出例を図 2.1 に示す．

図 2.1 では，“I” と “私は” のような短いフレーズから，“live in a studio apartment” と “ワンルームマンションに住んでいる” といった長いフレーズまで様々なフレーズが抽出されていることがわかる．

フレーズの翻訳確率は，相対頻度を用いて，フレーズ e, f の共起頻度を $C(e, f)$ とすると以下のような式で与えられる．これらの処理により，単語とフレーズを 1 章で示し

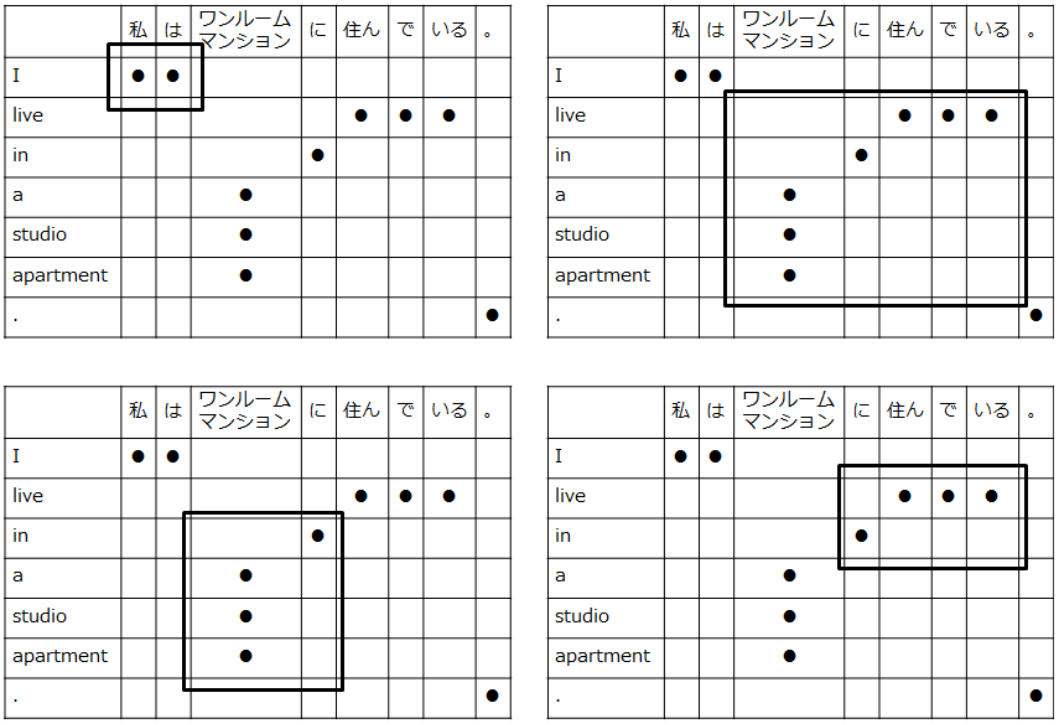

図 2.1: 抽出されるフレーズ

た式（1.4）において同様に扱うことができる．このモデルをフレーズ翻訳モデル，またはフレーズモデルという．

$$p(f|e) = \frac{C(e,f)}{\sum_f C(e,f)} \tag{2.14}$$

翻訳モデルによって得られる単語・フレーズ対は，提案手法の入力として使用される．提案手法は翻訳候補を生成することはせず，翻訳モデルが生成する翻訳候補に，語義に基づいた大局的な情報による確率を与えるモデルである．翻訳候補の生成を翻訳モデルに任せることは，推定の計算量と他のモデルとの組み合わせにおいて有利に働く．

2.1.2 N-gram 言語モデル

言語モデルは，文が自然かどうかを評価するモデルである．本節では最も一般的に用いられる N-gram 言語モデル [5] について述べる．

N-gram 言語モデルは，単語列の生起確率に，$N-1$ 次のマルコフ性を仮定したモデルである．長さ L の単語列 $w_1^L = w_1, w_2, ..., w_L$が与えられたとき，その生起確率は，式

(2.15) となる．N-gram 言語モデルは，式 (2.16) で与えられる．

$$p(\boldsymbol{w}) = \prod_{i=1}^{L} p(w_i|w_{i-1}w_{i-2}\cdots w_1) \tag{2.15}$$

$$\simeq \prod_{i=1}^{L} p(w_i|w_{i-N+1}^{i-1}) \tag{2.16}$$

$p(w_i|w_{i-N+1}^{i-1})$ は，単語列 w_i^L の訓練データにおける出現頻度を $C(w_i^L)$ とすると，最尤推定により式 (2.17) を得る．

$$p(w_i|w_{i-N+1}^{i-1}) = \frac{C(w_{i-N+1}^{i})}{C(w_{i-N+1}^{i-1})} \tag{2.17}$$

しかし，式 (2.17) では，N に対してパラメータが指数的に増加する．また，訓練データに出現しない単語列の確率が 0 になってしまう．適切な確率を与えるためには，スムージングと呼ばれる手法が用いられる．本研究では，最も一般的に用いられる Interpolated Kneser-Ney [11] を使用する．

Interpolated Kneser-Ney [11] は，式 (2.18) で与えられる．

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i) - D}{C(w_{i-1})} + \beta(w_i)\frac{|\{w_{i-1} : C(w_{i-1}w_i) > 0\}|}{\sum_{w_i}|\{w_{i-1} : C(w_{i-1}w_i) > 0\}|} \tag{2.18}$$

式 (2.18) は bigram の例を示している．定数 D は，出現頻度 $C(w)$ が不当に高い確率を与えることを防ぐために導入されている．2 項目の $\sum_{w_i}|\{w_{i-1} : C(w_{i-1}w_i) > 0\}|$ は，単語 w_{i-1} に続く単語の異なり語数を表す．異なり語数を考慮することで，高頻度だが，特定の N-gram でしか出現しないような unigram に対する確率を抑制する．係数 β は，単語 w_i に対する重みであり，$\sum_i \beta(w_i) = 1$ となる．

言語モデルは，文中の N-gram を評価しており，文脈を考慮して自然な文に高いスコアを与える優れたモデルである．しかし，考慮しているのは対象単語の直前の N-gram のみであり，局所的な文脈だといえる．一方，提案手法で用いる語義に基づいた大局的な情報は，翻訳する単語の文に含まれる全ての単語を含む．提案手法は，広い範囲の文脈を考慮するモデルだといえる．1 章で示した式 (1.4) によって，2 つのモデルは組み合わされ，翻訳の際にはお互いが相補的な働きを示す．

2.2 語彙モデルに関する研究

語彙モデルは，機械翻訳において，単語単位の翻訳を与えるモデルである．統計的機械翻訳において，語彙選択の問題は非常に重要である．IBM モデル [3] も語彙モデルの

一種である．IBM モデルは，単語のアライメントとアライメントされた単語の翻訳確率をモデル化している．

近年では，アライメントされた単語だけではなく，様々な情報を用いた語彙選択モデルが提案されている [10, 2, 15]．本節では，語彙選択に用いられる情報に着目して関連研究を紹介する．

IBM モデルに近い手法としては，Hasan ら [10] の Triplet Lexicon Model が挙げられる．Triplet Lexicon Model は，$p(f|e)$ の条件部に他の単語をトリガーとして考慮し，$p(f|e,e')$ によって翻訳確率をモデル化する．トリガーとなる e' は，文の位置に関係なく，全ての単語がトリガーとなることができる．トリガーは，単語対応だけでは拾いきれない文脈を捉える働きをする．Triplet Lexicon Model は，その構造から IBM モデル 1 の拡張だと見なせる．モデルの訓練も IBM モデルと同様の手順で行うことができる．

より大局的な情報を用いた語彙モデルとして，Bangalore ら [2] や Mauser ら [15] がある．Bangalore ら [2] は，原言語の N-gram を素性として，その素性が目的言語に含まれる確率をモデル化した．

$$BOW_{\boldsymbol{e}} \quad = \quad \{e|p(e|BOW(\boldsymbol{f})) > \theta\} \tag{2.19}$$

$BOW(\boldsymbol{f})$ は，原言語から生成される N-gram，θ は，閾値である．$p(e|BOW(\boldsymbol{f}))$ が閾値以上であれば，その e を出力文に含まれる候補とする．$BOW_{\boldsymbol{e}}$ から生成される候補文は，言語モデルによって評価される．

Mauser ら [15] は，目的言語の単語に対して，対応する原言語に含まれる全ての単語の表層形を用いることで，局所的でない文脈を利用する方法を提案した．この手法では，原言語の単語セット $\boldsymbol{f}$ が与えられた際の目的言語文 $\boldsymbol{e}$ の確率を，素性関数 $\phi(f,\boldsymbol{f})$ と，重み λ を用いて，個々の要素である最大エントロピーモデル $p(e|\boldsymbol{f})$ からモデル化する．$\boldsymbol{V_E}$ は目的言語側の語彙のセットを表している．また，目的言語側の語彙 e が目的言語文 $\boldsymbol{e}$ に含まれるとき e^+，含まれないとき e^- を使用する．

$$p \quad (\boldsymbol{e}|\boldsymbol{f}) = \prod_{e \in \boldsymbol{e}} p(e^+|\boldsymbol{f}) \cdot \prod_{e \in \boldsymbol{V_E} \backslash \boldsymbol{e}} p(e^-|\boldsymbol{f}) \tag{2.20}$$

$$\phi \quad (f,\boldsymbol{f}) = \begin{cases} 1 & \text{if } f \in \boldsymbol{f} \\ 0 & \text{else} \end{cases} \tag{2.21}$$

$$p \quad (e^+|\boldsymbol{f}) = \frac{\exp\left(\sum_{f \in \boldsymbol{f}} \lambda_{f,e^+} \phi(f,\boldsymbol{f})\right)}{\sum_{e \in \{e^+,e^-\}} \exp\left(\sum_{f \in \boldsymbol{f}} \lambda_{f,e^+} \phi(f,\boldsymbol{f})\right)} \tag{2.22}$$

Mauser らの手法は，仮説をフレーズモデルが生成するかどうかに依存している点で，Bangalore らの手法と異なる．Bangalore らの手法ではフレーズモデルを必要としない代わりに，出力候補の単語の選択に閾値を必要とする．2 つの手法とも，原言語に含まれる全ての単語を用いて目的単語を推定している．これらの手法で用いる情報は，文脈を捉えた情報と見なすことができる．

本研究では，語彙モデルの構築において，Mauser らのモデルを参考にしている．提案手法では，語義を利用して目的言語を推定する．語義で単語をまとめあげることで，意味が同じ単語は同じ素性として扱われる．そのため，Mauser らの表層形のみによる推定と比較すると，推定に用いられる素性は密なデータになり，推定はより柔軟に文脈を考慮するといえる．

2.3 語義曖昧性解消に関する研究

大局的な情報を用いて文脈を考慮するモデルは，文脈の素性から単語の意味を推定する語義曖昧性解消（Word Sense Disambiguation: WSD）と関係がある．語義曖昧性解消の技術は多義語の翻訳において適切な訳語を選択し，統計的機械翻訳の精度を向上させるための要素技術として認識されている [4, 24].

統計的機械翻訳における語義曖昧性解消では，曖昧性を解消する単語とその語義を対訳コーパスから取得することが一般的である [1, 4, 24]．アライメントによって対訳文それぞれの単語対応が取得され，原言語側の単語が語義曖昧性解消の対象語に，目的言語側の単語がその語義として扱われる．語義の決定には教師あり学習がよく用いられ，対象単語と対応する翻訳候補から適切な訳語を選択する単語翻訳 [24] や，n-best の翻訳仮説を WSD を用いてリランキングする手法 [1] など，さまざまな研究がされている．また，単語だけではなく，フレーズモデルへの対応も行われており，Carpuat ら [4] は，曖昧性解消の対象をフレーズモデルの生成した仮説とすることで，フレーズ意味曖昧性解消（Phrase Sense Disambiguation: PSD）を提案している．曖昧性解消の単位をフレーズモデルに任せることで，翻訳精度が向上することが示されている．

また，語義曖昧性解消は，語義を推定するための言語資源として WordNet を使用することがある．WordNet [7] は，英語の概念辞書である．英単語が synset と呼ばれる同義語集合にまとめられている．それぞれの synset には，簡単な定義や他の synset との関係などが記述されている．これを元に日本語を付与したものが，日本語 WordNet [14] として公開されている．WordNet を用いることで，単語の文字列だけでなく，意味的な類似

を用いることができる．

WordNet を用いた研究には，語義曖昧性解消の際の言語資源として用いた Pedersen ら [21] の研究や，機械翻訳に対して WordNet を用いた Vintar ら [25] の研究がある．

まず，WordNet について説明する．WordNet は synset と呼ばれる同義語集合が階層構造で表現されている．図 2.2 に WordNet の概念の構造を表す．図 2.2 では，“13134947-n” という synset に同義語として，“fruit” や “果実” といった単語が登録されている．“13134947-n” の下位には “果物” や “edible fruit” が登録された “07705931-n” や “さや” 等が登録された “13139055-n” の synset があり，さらにその下に “apple” や “peach” といった同義語集合を持つ synset が存在している．synset に含まれる同義語集合や，各 synset のパスの長さなどを利用することで，synset の関係や，意味的な類似度を取得することができる．

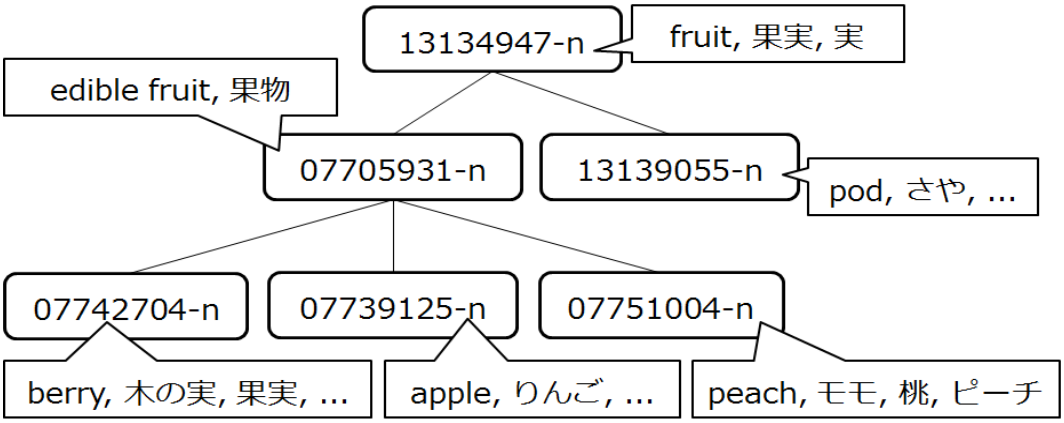

図 2.2: WordNet [14] における概念の階層構造

機械翻訳における WordNet の有用性については，Vintar ら [25] が既存の翻訳システムとの比較を行っている．Vintar らは，WordNet を用いた語義曖昧性解消による翻訳と数種類の既存の機械翻訳システムとの比較において，一定の単語の翻訳において，WordNet を用いた語義曖昧性解消が有効であることを示した．ただし，この研究では，WordNet を用いた翻訳が，既存の翻訳システムの誤りの一部を正しく翻訳することを示しただけであり，全体の精度の向上には至っていない．

WordNet を用いた汎用的な語義曖昧性解消の手法としては，Pedersen ら [21] の研究がある．Pedersen らは，語義曖昧性解消の対象語を含む文脈に対して，WordNet を用いて語義間に定義される任意の関連性尺度を定義することで語義を推定する手法を提案して

いる．Pedersen らの語義推定は，以下の式で行われる．

$$\arg\max_{i=1}^{m_i} \sum_{j=1,j\neq t}^{n} max_{k=1}^{m_j} rel(s_{ti}, s_{jk}) \tag{2.23}$$

$w_1, w_2, ..., w_n$ は，文脈内の語を，w_t は語義を付与する対象語を表す．$s_{i1}, s_{i2}, ..., s_{im_i}$ は，w_i が持つ，m_i 個の語義である．rel は語義のペアに対してその関連性スコアを与える．関連性スコアは，synset 間のパスの長さなどから決定される．Pedersen らの手法は，任意の語義ペアに評価値を与える汎用性の高い手法であるが，前処理として全ての synset 間の関連性スコアを保持することは難しいため，処理の都度，WordNet へ問い合わせて計算を行う必要がある．統計的機械翻訳システムに組み込むことを考えると，入力文を処理するたびに WordNet への問い合わせが必要なため，動作速度的に実用的ではない．

本研究では，単語から語義への変換に WordNet を使用する．提案手法は，翻訳する単語の文に出現する全ての単語を素性として，目的言語の単語を推定する．目的言語の単語ごとに，共起した原言語の単語全てがパラメータを持つことになるため，推定に用いる素性は非常に疎なデータになる．提案手法では，語義が同じ単語は，それが出現する文脈にも類似性があると仮定して，WordNet を用いて単語を語義に変換してから推定を行う．単語は複数の語義を持つことがあるため，単純に単語を WordNet に問い合わせて該当する語義を用いたのでは，不適切な語義を多く含んでしまう．不適切な語義を含まないようにするため，原言語だけではなく，WordNet に登録されている目的言語の単語も用いて語義への変換を行う．

語義への変換は，モデルの訓練の前処理として行われる．訓練されたモデルは，語義から目的言語の単語へのパラメータを持つ．このモデルをそのまま翻訳に用いると，翻訳時に WordNet への問い合わせが必要になる．WordNet への問い合わせは，翻訳速度の大きな遅延になるため，訓練されたモデルは，原言語の単語から目的言語の単語に対するパラメータに前もって変換しておく．

2.4 機械翻訳の評価手法

機械翻訳の評価には，人手評価と自動評価がある．人手評価はシステムの出力文を評価者が判定していく方法であり，評価者の技能による部分もあるが，自動評価より高い精度かつ，柔軟な評価が行えるというメリットがある．

自動評価はあらかじめ入力文に対する正解訳を用意しておき，システムの出力文と正解訳を比較することで評価する．正解訳を用意する手間はあるが，評価，改良という流

れが速やかに行えるため，一般的に自動評価の方がよく用いられる．自動評価は，正解訳と出力文の比較で評価値を出すため，正解訳と文字列的に異なる出力文には，たとえそれが適切な場合にも低い評価値を与えてしまうことがある．

本節では，自動評価尺度として $BLEU$（Bilingual Evaluation Understudy） [20] と $RIBES$（Rank-based Intuitive Bilingual Evaluation Score）[26] を紹介する．どちらの評価尺度も，出力文と，入力文の正解訳に基づいて評価値を決定する．

$BLEU$ は機械翻訳の自動評価の尺度として最も頻繁に用いられている評価尺度である．$BLEU$ は以下の式で与えられる．

$$BLEU = BP \times \exp(\frac{1}{N}\sum_{n=1}^{N}\log p_n) \tag{2.24}$$

$$p_n = \frac{\sum_i \text{正解文}\, i \,\text{と一致した}\, Ngram \,\text{数}}{\sum_i \text{出力文}\, i \,\text{中の全}\, Ngram \,\text{数}} \tag{2.25}$$

$$BP = \begin{cases} 1 & (|c| \geq |r|) \\ \exp(1 - \frac{|r|}{|c|}) & (otherwise) \end{cases} \tag{2.26}$$

式 (2.25) は出力文と正解文の N-gram の一致率を表している．一致率は 1-gram から N-gram まで計算され，その幾何平均が p_n の値となる．一般に N の値は 4 が用いられ，本研究でも $N = 4$ で扱っている．

式 (2.26) は $|c|$ が翻訳文の単語数, $|r|$ が正解文の単語数を 表している．翻訳文が正解文より短い場合，不当に BLUE 値が高くなるため，このようなペナルティをかける．$BLEU$ は [0,1] で与えられ，1 に近いほど翻訳精度が高い．

$RIBES$ は，正解訳と出力文で共に出現した単語の順序を Kendall の順位相関係数で評価する．$RIBES$ は以下の式で与えられる．

$$RIBES = NKT \times P^{\alpha} \tag{2.27}$$

$$NKT = \frac{\tau + 1}{2} \tag{2.28}$$

$$\tau = \frac{\sum_{i=1}^{n-1} K_1 - \sum_{i=1}^{n-1} L_1}{\frac{n(n-1)}{2}} \tag{2.29}$$

$$P = \frac{n}{h} \tag{2.30}$$

$RIBES$ は正規化された Kendall の順位相関係数 NKT と，ペナルティ P とその重み $\alpha(0 \leq \alpha \leq 1)$ の積で計算される．

$RIBES$ の計算には，まず，正解訳と出力文で共通する単語のみを抽出する．その単語集合を正解訳に出現した順にソートして，リスト $\boldsymbol{r}$ を得る．次に，$\boldsymbol{r}$ の各要素に対応

する単語が出力文の何番目に出てきたかを表すリスト $\boldsymbol{h}$ を得る．$\boldsymbol{r}$ の 1 番目の要素が出力文の 5 番目の要素に対応する場合，$\boldsymbol{h}$ の 1 番目の要素は 5 となる．得た 2 つのリストから式（2.29）の Kendall の順位相関係数を計算する．K_i は，h_i について，$h_i < h_j$ となる場合の数，L_i は $h_i > h_j$ となる場合の数を表す．ただし，$j = i+1, ..., n$ である．式（2.29）は，[-1,1] の値を取るため，[0,1] の値をよるように，式（2.28）で正規化する．また，極端に共通する単語が少ない場合，不当に高いスコアを与えるため，式（2.30）によって，出力文の単語が正解訳に含まれる割合をペナルティとして与える．n は，出力文と正解訳で共通な単語数，h は出力文の単語数である．$RIBES$ も $BLEU$ と同様に，[0,1] で与えられ，1 に近いほど翻訳精度が高い．

$BLEU$ は，N-gram の一致数から計算される，シンプルかつ直感的な評価尺度であるが，問題点として，N-gram までの局所的な単語列の一致しか見ないため，局所的には正しいが，大きく語順が誤っている文に対して，不当に高い評価値を与えることがある．$RIBES$ は，共通して出現する単語の出現順序に着目した評価尺度であり，参照訳に対して語順が大きく異なる文に対して適切な評価値を与えることができる．本研究では，$BLEU$ と $RIBES$ の 2 つを評価尺度として用いる．

2.5 本研究の位置づけ

本研究では，語義に基づく大局的な情報を用いた語彙モデルを提案する．本研究では，同じ意味を持つ単語は，出現する文脈に類似性があるという仮定のもと，大局的な情報として語義を用いることで，表層形が異なるが同じ意味を持つ単語を同じ素性と見なす．この処理は，語義によって単語をまとめあげることに等しい．単語の表層形による大局的な情報 [2, 15] を用いた推定手法と比較して，文脈をより柔軟に考慮することができる．

提案手法は，1 章で示した式（1.4）によって，他のモデルと組み合わせて用いられる．フレーズモデルは，対訳文におけるフレーズ対の共起から翻訳確率を与える．提案手法は，フレーズモデルによる翻訳候補に対して，語義に基づいた大局的な情報による確率を与えるモデルである．文に対して，尤もらしさから評価値を与える言語モデルと提案手法は，相補的な関係にある．言語モデルが局所的な文脈を，提案手法が大局的な文脈を考慮している．

単語から語義への変換には，WordNet を使用する．単純に単語を WordNet に問い合わせて該当する語義を用いると，不適切な語義を多く含んでしまう．そのため，原言語だけではなく，目的言語も用いて語義への変換を行う．また，構築したモデルは，翻訳

時に WordNet への問い合わせが発生しないように，モデルの訓練後，語義を単語へ戻す処理が行われる．

本研究では，語義に基づく大局的な情報を用いることで，表層形が異なるだけの異表記語や，同じ概念を持つ語を同じ素性として目的言語の単語の推定を行う．これにより，単語やフレーズだけでは翻訳できない曖昧な語や，文の時制など，文脈を考慮した翻訳を実現する．

第3章　語義に基づく大局的な情報を用いた語彙モデル

本章では，本研究で提案する語義に基づく大局的な情報を用いた語彙モデルについて説明する．3.1 節では，本研究で用いる語義に基づく大局的な情報について述べる．3.2 節では，語義に基づく大局的な情報を用いた語彙モデルについて示す．

3.1　訳語選択における語義に基づく大局的な情報の役割

本節では，訳語選択の際に語義に基づく大局的な情報がどのように働くのかについて述べる．まず，関連研究でも用いられている，大局的な情報 [2, 15] が翻訳においてどのような働きするのかを整理する．次に，語義に基づく大局的な情報の概要を示し，翻訳における働きについて述べる．

3.1.1　大局的な情報

大局的な情報とは，ある単語の翻訳を考えた際に，その単語が生起した文における他の全ての単語の表層形 [2, 15] である．大局的な情報は，文脈を表す素性であり，これを考慮した語彙モデルを用いることで，訳語選択において文脈を考慮することができる．

大局的な情報と語義に基づく大局的な情報を整理するために，まず，大局的な情報を考慮しないモデルについて例を示す．以下のような入力文が与えられたときの，太線部 “いる” の翻訳を考える．

入力文: 誰 か が 私 を 訪ね て来 たら 、 私 は コーヒー ショップ に
いる と 伝え て下さい .

翻訳例: If anyone asks for me , tell them I will be in the coffee shop .

フレーズモデルが生成した候補として，be, are, will be, am at の 4 つが与えられたとする．フレーズモデルでは，対応が取られた目的言語の単語，フレーズから翻訳確率が

与えられる．そのため，“いる” の翻訳では，“いる” と対応が取られた be, are, will be, am at の候補から最も翻訳確率が高い単語，フレーズが選択される[1]．しかし，実際の人手による翻訳では，ある単語を翻訳するとき，その単語のみで訳語を決定するわけではなく，文脈を考慮することが一般的である．言語モデルは文脈を考慮しているが，対象単語の周辺のみであり，十分に文脈を考慮しているとはいえない．

次に，大局的な情報を用いた翻訳を考える．大局的な情報に用いられる原言語文の単語は，翻訳の対象となる語から距離的な制約を持たず，また，順序も関係なく扱われる．距離的な制約を設けないのは，言語モデルでは捉えきれない大局的な文脈を捉えるためである．“いる” の翻訳においては，以下のように下線部の情報が大局的な情報として扱われる．

入力文: 誰 か が 私 を 訪ね て来 たら 、 私 は コーヒー ショップ に **いる** と 伝え て下さい .

例文では，“いる” の翻訳として “will be” が適切であるが，“will be” を選択する最も大きい手がかりは，文の最初に生起する “誰かが私を訪ねて来たら...” である．このような文脈における “いる” は，まだその場所にはいないが，近い将来にいるということを表すことが多い[2]．

大局的な情報は，距離的な制約を持たずに，文脈における全ての単語の表層形を用いて，目的言語の単語を推定する．そのため，時制に関連する翻訳や，文脈を考慮することで訳語が決定されるような単語やフレーズの翻訳において非常に有用であるといえる．しかし，1 つの単語に対してその単語と共起する全ての単語を考慮することになるため，推定に必要なデータ量は増加する傾向がある．

3.1.2 語義に基づく大局的な情報

大局的な情報を用いた研究 [2, 15] では，目的言語の単語それぞれに対して，生起する原言語の単語やフレーズがパラメータを持っていた．この場合の単語は表層形を指し，文字列で区別されている．例えば，パソコンと鉛筆，コーヒーと珈琲，ショップとストア，商店，店，販売店といった語は全て区別される．これらは，パソコンと鉛筆のように

[1] この例は単語に着目してフレーズモデルのみを考えた場合であり，実際には複数の仮説から他のモデルの評価値も考慮するため，翻訳確率が高い候補が選択されるとは限らない．

[2] 現在地がコーヒーショップで，電話で伝言を頼んでいるという状況も考えられるが，ここでは，より一般的な対面での会話を想定している．

文字列も意味も訳語も異なるものや，コーヒーと珈琲のように表層形が異なるだけで意味と訳語は同じもの，ショップとストアのようにある文脈において意味は同じだが，文字列と訳語は異なるものがある．

大局的な情報は，目的言語の単語を推定するために有用である文脈を考慮するために，文中に生起する単語の表層形という素性を利用している．そのため，生起する文脈が同じ単語は同一のものとして扱うことで，より適切に文脈を考慮できると考える．具体的には，コーヒーと珈琲のような表層形だけが異なる単語や，ショップとストアのように意味が同じ単語を同一のものとして扱う．

本研究では，先ほどの大局的な情報に加えて，単語を WordNet を用いて語義に変換して用いることで，より柔軟に文脈を考慮する．単語を語義に置き換えることで，単語のままでは区別されていたものが，同じ素性として利用できる．これは，目的言語の単語に対する素性を，語義によってまとめあげることに等しい．語義に変換することで表層形が異なる単語だけでなく，概念の近い語も同じ文脈内に生起する語として扱える．“いる” の翻訳においては，大局的な情報で用いられていた下線部の単語のうち，非活用語が WordNet を通して語義に変換される．例文で該当する語を二重下線で示した．二重下線部の単語を変換して得た語義に登録されている単語には次のようなものがある．語義に基づく大局的な情報では，これらの単語も同じ文脈に生起する単語として用いる．

入力文: 誰 か が 私 を 訪ね て来 たら 、 私 は コーヒー ショップ に **いる** と 伝え て下さい .

コーヒー: カフェー，カフェ，珈琲

ショップ: 店鋪，ストア，売店，販売店，商店，店

本研究では，WordNet を用いて単語から語義への変換を行うが，この際のデメリットして，過剰な汎化が起きることがある．WordNet では，1 つの単語に複数の語義が登録されていることがある．単語から語義への変換の際には，どの語義が文における単語の意味を表しているかはわからないため，一旦全ての語義へ変換する．しかし，全ての語義を使用すると，関係のない語義を含んでしまうため，関係のない単語を同じ文脈に生起する単語として扱ってしまう．

提案手法では，WordNet の語義に複数言語の単語が登録されていることを利用して，不適切な語義の追加される問題を防ぐ．例として，“流れ” という名詞を挙げて説明する．WordNet において，“流れ” は，以下のような語義に登録されている．

(a) 流れるか，流れ込む行為; 流れる行為
登録されている語: ストリーム，流れ，フロー，stream，flow

(b) 一般的に変化する傾向の（意見などで）
登録されている語: 動向，風潮，トレンド，流れ，trend，movement

(c) 出来事が途絶えることなく続くこと
登録されている語: 連続，流れ，streak，run

行為としての “流れ” や，流行や意見の傾向を表す “流れ” など，“流れ” は様々な意味を持つ．これらは生起する文脈も異なっており，別の単語として扱うことが望ましい．しかし，原言語文から取得できるのは，“流れ” という文字列のみであり，このまま変換すると，不適切な語義を含んでしまう．そこで，目的言語の語義も併用して用いる．“流れ” の語義に登録されている目的言語の単語には，行為としての “流れ” には，flow や stream，流行や意見の傾向を表す “流れ” には，movement や trend がある．もし，対訳文において，原言語で “流れ” が，目的言語で “movement” が生起していたとする．“movement” は，以下のような語義に登録されている．

(a) ある場所から別の場所へ位置を変える行為
登録されている語: 移転，変位，motion，movement

(b) 一般的に変化する傾向の（意見などで）
登録されている語: 動向，風潮，トレンド，流れ，trend，movement

“流れ” の語義と，“movement” の語義の積集合を取ることによって，正しい語義である，流行や意見の傾向を表す語義とそれに登録されている単語を取得することができる．

提案手法では，原言語と目的言語の文から別々に語義を構築し，積集合を取ったものを素性として用いる．この処理によって，不適切な語義を素性に含んでしまうことを防止している．具体的なモデルについては次節で説明する．

3.2 提案する語彙モデル

提案手法は，原言語と目的言語から生成する語義と表層形の集合を用いて目的言語の単語を推定する．語義の生成には，名詞と副詞を用いる．推定に用いられる集合は，翻訳する単語が出現する文に含まれる全ての単語から生成され，目的言語の単語との距離や単語対応に影響を受けない．

提案手法では，原言語と目的言語から生成した集合 $\boldsymbol{s}$ が与えられた際の目的言語文 $\boldsymbol{e}$ の確率を，素性関数 $\phi(s, \boldsymbol{s})$ と，重み λ を用いて，個々の要素である最大エントロピーモデル $p(e|\boldsymbol{s})$ からモデル化する．本モデルは Mauser ら [15] のものを参考にしている．Mauser らのモデルでは，目的言語の単語に対して，対応する原言語文で生起する全ての単語がパラメータを持つ．それらのパラメータは，原言語で生起する単語が目的言語のある単語に翻訳される程度を示している．パラメータは非常に多くなるため，全ての単語が有効な情報とはならず，一部の単語が翻訳する単語と同じ文中に生起すると，ある単語に翻訳されやすくなるという働きをする．提案手法は，語義に基づく大局的な情報を用いることで，表層形による大局的な情報を用いるモデルよりも柔軟に文脈を考慮する．

$\boldsymbol{V_E}$ は目的言語側の語彙の集合を表している．また，目的言語側の語彙 e が目的言語文 $\boldsymbol{e}$ に含まれるとき e^+，含まれないとき e^- を使用する．

$$p(\boldsymbol{e}|\boldsymbol{s}) = \prod_{e \in \boldsymbol{e}} P(e^+|\boldsymbol{s}) \cdot \prod_{e \in \boldsymbol{V_E} \backslash \boldsymbol{e}} P(e^-|\boldsymbol{s}) \tag{3.1}$$

$$\phi(s, \boldsymbol{s}) = \begin{cases} 1 & \text{if } s \in \boldsymbol{s} \\ 0 & \text{else} \end{cases} \tag{3.2}$$

$$p(e^+|\boldsymbol{s}) = \frac{\exp\left(\sum_{s \in \boldsymbol{s}} \lambda_{s,e^+} \phi(s, \boldsymbol{s})\right)}{\sum_{e \in \{e^+, e^-\}} \exp\left(\sum_{s \in \boldsymbol{s}} \lambda_{s,e^+} \phi(s, \boldsymbol{s})\right)} \tag{3.3}$$

$$\boldsymbol{s} = \{s | s \in \boldsymbol{f}^{syn} \cap \boldsymbol{e}^{syn} \vee s \in \boldsymbol{f}\} \tag{3.4}$$

$$\boldsymbol{f}^{syn} = \{syn(f) | f \in \boldsymbol{f} \wedge f \text{ は名詞か副詞 }\} \tag{3.5}$$

$$\boldsymbol{e}^{syn} = \{syn(e) | e \in \boldsymbol{e} \wedge e \text{ は名詞か副詞 }\} \tag{3.6}$$

$$syn(word) : WordNet \text{ を用いて単語を語義に変換} \tag{3.7}$$

提案手法は，語義と表層形の集合 $\boldsymbol{s}$ から目的言語を推定している．まず，次の手順で語義と表層形の集合 $\boldsymbol{s}$ を作成する．

1. 式（3.5）より，原言語文に含まれる非活用語と品詞のペアを WordNet に問い合わせ，語義に変換する．WordNet が取り扱っている品詞は動詞，名詞，形容詞，副詞の 4 種類である．そのため，日本語かつ非活用な品詞としては，名詞，副詞が該当する．この問い合わせの際，1 つの語に対して複数の語義が提示されることもあるが，全ての語義を利用する．

2. 式（3.6）より，目的言語も原言語と同様に WordNet へ問い合わせて，語義を取得する．

3. 式（3.4）より，先の手順で得た原言語の語義集合と目的言語の語義集合の積集合を取る．積集合を取ることで得た語義集合と，原言語の表層形から $\boldsymbol{s}$ を作成する．$\boldsymbol{s}$ は語義と原言語の混同された集合である．

語義への変換の際に非活用語のみ使用する理由は，活用語の語形変化は文脈全体の時制の変化など，訳語が変化する場合が多いためである．WordNet へ問い合わせて単語を語義を変換する式（3.7）では，単語の生起した文脈での意味とは異なった語義を含む，単語に対する語義が存在しないことがある，という 2 つの問題が発生する．単語の生起した文脈での意味とは異なった語義を含む問題については，目的言語側の語義を取得し，その積集合を取ることで，過剰な汎化を防いでいる．単語に対する語義が存在しない問題については，式（3.4）のように，語義に加えて，原言語文の全ての表層形も素性に含むことで影響を軽減している．

作成された語義と表層形の集合 $\boldsymbol{s}$ を用いて，式（3.1）から式（3.3）によって目的言語文 $\boldsymbol{e}$ を推定する．式（3.1）の目的言語文 $\boldsymbol{e}$ は，語義と表層形の集合 $\boldsymbol{s}$が与えられた際の単語 e の確率 $p(e|\boldsymbol{s})$ の積で表される．

目的言語のある単語の確率 $p(e|\boldsymbol{s})$ は最大エントロピーモデルで表現される．式（3.3）は，ある単語に対して，与えられた語義と表層形の集合 $\boldsymbol{s}$ に生起する要素に割り振られた確率を足し合わせたものを，全ての単語に対して与えられた語義と表層形の集合 $\boldsymbol{s}$ に生起する要素に割り振られた確率を足し合わせたもので割った確率で表される．これは，ある単語と共起するかつ，他の単語と共起しない素性ほど大きい確率を持つ．

$\boldsymbol{s}$ は，原言語の表層形と語義が混同された集合である．そのため，式（3.1）の形で翻訳システムに組み込むと，入力の度に名詞と副詞を WordNet へ問い合わせて語義に変換する必要が生じる．WordNet への問い合わせの遅延を失くすため，実装では，語彙モデルの訓練後に語義を表層形に変換する処理を行っている．そのため，前節で説明した入力文中の “コーヒー” や “ショップ” の語義への変換は，実際の処理としては行われない．代わりに，語義で作られたモデルが表層形へ変換されることで同等の処理を行っている．語義から単語への変換の際，表層形で登録されていた単語と，語義から変換される単語で重複が起きることがある．重複した場合は，より有効な値が優先される．この処理によって，語彙モデルの実体は，目的言語の単語に対して，それが生起する文脈の原言語の単語とパラメータという三つ組のデータの集合で表される．

提案手法自体は，単語対応を求めることはせず，フレーズモデルが生成した仮説に対して，語義に基づく大局的な情報を用いて確率を与えるモデルである．

第4章　翻訳精度の評価実験

本章では，提案手法の評価実験について説明する．翻訳精度の評価実験として，統計的機械翻訳システム Moses[12] を用いて日英翻訳を行い，提案モデルの有無での翻訳精度を比較した．4.1 節では，実験に用いたデータについて説明する．4.2 節は，実験方法について，使用したモデルや使用したツール等を説明する．最後に 4.3 節で実験結果とその概要について述べる．

4.1　実験に用いるデータ

本実験で使用する，日英翻訳エンジン学習・評価用対訳コーパスと NTCIR-9 特許機械翻訳テストコレクションについて説明する．

日英翻訳エンジン学習・評価用対訳コーパスは，高度言語情報融合フォーラム ALAGIN[1]で公開されている対訳コーパスである．本コーパスは，翻訳に関する評価型ワークショップである IWSLT(International Workshop on Spoken Language Translation)[2]の日英翻訳タスクで使用された，基本旅行会話データセットから作成されている．本コーパスのサンプルを表 4.1 に示す．

NTCIR-9 特許機械翻訳テストコレクション [9] は，情報アクセス技術の評価型ワークショップである NTCIR-9 で提供された対訳コーパスである．1993 年から 2007 年までの日本公開特許公報全文と米国特許庁特許全文を元に作成されている．本コーパスのサンプルを表 4.2 に示す．

本実験で用いたコーパスのデータサイズは表 4.3 のようになっている．なお，NTCIR-9 特許機械翻訳テストコレクションについては，約 320 万文対のデータから一部を抽出して使用している．また，特許機械翻訳テストコレクションのテストセットとして NTCIR-8[8] で提供されたものも用いた．テストセットのサイズに続く括弧内の文字列は，各テストセットの名前である．本実験では，各コーパスで 2 つのテストセットを用いた．

[1] http://alaginrc.nict.go.jp/
[2] http://www.is.cs.cmu.edu/iwslt2005/

表 4.1: 日英翻訳エンジン学習・評価用対訳コーパスのサンプル

原言語文 (日本語文)	目的言語文 (英語文)
荷物をトランクに入れて下さい.	Put the baggage in the trunk, please.
信号は赤でした.	The light was red.
窓際の席を御願いします.	We want to have a table near the window.
暗証番号を押して下さい.	Please input your pin number.
海に潜るのは初めてです.	This is my first time diving.
えーと、何泊なさいますか.	Let me see. How many nights will you be staying?
ジョンと一緒に働いています.	I work with John.
彼女が私達のパーティーに来ると思いますか.	Do you think she'll come to our party?
私がそこ迄行けるように地図を描いて下さい.	Can you draw a map which can get me there?
ツアーガイドが朝の九時頃御迎えに参ります.	The tour guide will be here to pick you up around nine o'clock in the morning.

日英翻訳エンジン学習・評価用対訳コーパスと NTCIR-9 特許機械翻訳テストコレクションを比較すると，NTCIR-9 特許機械翻訳テストコレクションの方が文長が長いことがわかる．それに伴って文の構造も複雑になる傾向が見られる．また，内容についても，特許文から構築された NTCIR-9 特許機械翻訳テストコレクションの方が専門用語多く出てくるため，語彙も多様な傾向がある．また，訳語が複数単語になるような単語も多い．日英翻訳エンジン学習・評価用対訳コーパスは比較的平易な単語が多い．しかし，原言語文の主語が省略されていたりと，単語単位の対応で見ると合致しないような文も多くある．また，疑問文も多く含まれている．

表 4.4 にコーパスに含まれる対訳文の平均単語数，平均語彙数，総語彙数を示した．日英翻訳エンジン学習・評価用対訳コーパスと NTCIR-9 特許機械翻訳テストコレクションを比較すると，NTCIR-9 特許機械翻訳テストコレクションは一文あたりの単語数，語彙数において，日英翻訳エンジン学習・評価用対訳コーパスの倍以上の値となった．

表 4.2: NTCIR-9 特許機械翻訳テストコレクションのサンプル

原言語文 (日本語文)	目的言語文 (英語文)
オペレーティングシステム３０２は、選択されたアプリケーション３０１を実行する.	An operating system 302 executes a selected application 301.
例えば、厚さ約３００ｎｍのシリコン基板表面層をエッチングする.	For example, a silicon substrate surface layer of about 300 nm thick is etched.
コンデンサ２２６は、ＰＭＯＳトランジスタ２３４のゲート電圧を、漸減して変化させる.	The capacitor 226 changes the gate voltage of the PMOS transistor 234 to gradually decrease.
この後、マスク材が除去される.	Subsequently, the mask material is removed.
ここで, 灰色で示すものが, 回路パターンである.	The circuit patterns are shown in gray.
次に、エンジン１の始動及び停止について説明する.	Next, starting and stopping of the engine 1 will be described.
そして所定周期Ｔ３毎にＴ３期間内の判定結果を調べ、全て正常である場合には故障検出カウンタをクリアし、そうでない場合には故障検出カウンタを１加算する.	Then, after the determination results, during a predetermined period T3, each interval T3 are checked, if all of them are normal, a malfunction detection counter is cleared, while if not, one is added to the malfunction detection counter.
このことにより、シール部３が相手部材１３に引っ張られてゴム部材２から剥離するのを防止できる.	The sealing portion 3 is prevented from being pulled by the corresponding member 13 and peeled off the rubber member 2 thereby .

表 4.3: コーパスのサイズ

	日英翻訳エンジン 学習・評価用対訳コーパス	NTCIR-9 特許機械翻訳 テストコレクション
訓練セット	19,972	300,000
開発セット	506	500
テストセット	500（IWSLT-4）	1,251（NTCIR-8）
テストセット 2	506（IWSLT-5）	2,000（NTCIR-9）

表 4.4: コーパス（訓練セット）の基礎統計量

	日英翻訳エンジン 学習・評価用対訳コーパス	NTCIR-9 特許機械翻訳 テストコレクション
平均単語数（日本語）	10.4	25.5
平均単語数（英語）	9.4	23.5
平均語彙数（日本語）	9.8	22.1
平均語彙数（英語）	8.9	19.8
総語彙数（日本語）	8,463	40,136
総語彙数（英語）	7,067	50,423

4.2 実験方法

統計的機械翻訳システム Moses による日英の翻訳実験において，使用したツール，手順について示す．

データの整形に用いる形態素解析器には，日本語には MeCab[13]，英語には Moses 付属のスクリプトを使用した．また，提案手法の語彙モデルの構築には，WordNet[14] に加えて，英語の品詞情報と基本形が必要になる．それらの取得には GENIA Tagger[23] を用いた．モデルの構築では，単語対応の取得に GIZA++[18] を，フレーズモデルの構築には Moses を用いた．フレーズの抽出は，日英で最も良いとされている grow-diag-final-and を用いた．言語モデルは，SRILM[22] を用いて構築した．提案手法の学習には，最大エントロピーモデルのツールキットである MegaM[6] を使用した．各素性関数の重み付けには，MERT（Minimum Error Rate Training）[16] を用いた．評価尺度には $BLEU$[20] と $RIBES$[26] を使用した．

実験では，次の 3 つの手法を用いた．それぞれで MERT による素性関数の重み付けを行った．その他使用したモデル等の条件は全て同じである．

ベースライン: 統計的機械翻訳システム Moses[12]

Mauser: ベースラインに，語彙モデルとして Mauser の手法 [15] を追加

提案手法: ベースラインに，語彙モデルとして提案手法を追加

4.3 実験結果

4 種類のテストセットに対して 2 種類の評価尺度で翻訳精度を評価した結果，提案手法は 7 つの評価で最も良い結果を示した．1 つの評価では，ベースラインが最も良い結果を示した．

最も改善したテストセット（IWSLT-4）では，ベースラインから $BLEU$ で 1.51 ポイント，$RIBES$ で 0.51 ポイントの向上が見られた．

日英翻訳エンジン学習・評価用対訳コーパスと NTCIR-9 特許機械翻訳テストコレクションでは，日英翻訳エンジン学習・評価用対訳コーパスの $BLEU$ の平均が 47.08，$RIBES$ の平均が 86.94，NTCIR-9 特許機械翻訳テストコレクションの $BLEU$ の平均が 27.01，$RIBES$ の平均が 62.09 と，全体的に日英翻訳エンジン学習・評価用対訳コーパスの方が高い値を示した．

Mauser の手法は，IWSLT-4 の $RIBES$ 以外は，ベースラインを下回る結果となった．ベースラインと比較した全体の平均では，$BLEU$ が 0.61 ポイント，$RIBES$ が 0.77 ポイントの低下となった．

ベースラインから提案手法の改善は，日英翻訳エンジン学習・評価用対訳コーパスで $BLEU$ の平均が 1.30 ポイント，$RIBES$ の平均が 0.34 ポイントの向上，NTCIR-9 特許機械翻訳テストコレクションでは $BLEU$ の平均が 0.09 ポイント，$RIBES$ の平均が 0.37 ポイントの向上と日英翻訳エンジン学習・評価用対訳コーパスの方が高い値を示した．ベースラインから提案手法の精度向上の平均は，$BLEU$ が 0.70 ポイントの向上，$RIBES$ が 0.36 ポイントの向上となった．$BLEU$ と $RIBES$ では，$BLEU$ でより改善が見られたが，$RIBES$ では，4 種類のテストセットに対して，その全てで提案手法がベースラインを上回ったのに対して，$BLEU$ では，NTCIR-8 のテストセットでベースラインを下回るなど，テストセットによって大きく異なる結果となった．

表 4.5: WordNet を用いた単語から語義への変換

	日英翻訳エンジン 学習・評価用対訳コーパス		NTCIR-9 特許機械翻訳 テストコレクション	
	IWSLT-4	IWSLT-5	NTCIR-8	NTCIR-9
名詞	51.51	42.10	**27.63**	26.81
副詞	51.26	41.38	26.94	26.03

表 4.6: 翻訳精度（$BLEU$）

	日英翻訳エンジン 学習・評価用対訳コーパス		NTCIR-9 特許機械翻訳 テストコレクション	
	IWSLT-4	IWSLT-5	NTCIR-8	NTCIR-9
ベースライン	51.51	42.10	**27.63**	26.81
Mauser	51.26	41.38	26.94	26.03
提案手法	**53.02**	**43.19**	27.35	**27.27**

表 4.7: 翻訳精度（$RIBES$）

	日英翻訳エンジン 学習・評価用対訳コーパス		NTCIR-9 特許機械翻訳 テストコレクション	
	IWSLT-4	IWSLT-5	NTCIR-8	NTCIR-9
ベースライン	89.06	84.86	62.08	62.62
Mauser	89.52	83.61	61.05	61.35
提案手法	**89.57**	**85.03**	**62.58**	**62.86**

第5章　考察

本章では，実験結果の分析，考察を行う．5.1 節では，提案手法で用いた語義に基づく大局的な情報が，翻訳の改善にどのように寄与したかを述べる．5.2 節は，旅行会話文と特許文という 2 つのドメインのコーパスにおいて，提案手法の働きの差異について述べる．5.3 節は，大局的な情報と提案手法を適用した際の問題点についてそれぞれ述べる．

5.1　目的言語の推定における語義に基づく大局的な情報の有効性

表 4.6，表 4.7 より，提案手法は若干の精度向上が見られた．本節では，提案手法によって翻訳精度が改善された要因について，語義に基づく大局的な情報が単語推定においてどのような働きをしたかを，前節の実験による出力文から考察する．

提案手法では，主に次の 2 つの改善点が見られた．以下でそれぞれについて詳細に述べる．

1. 単語，フレーズ単位では適切だが，文脈を考慮すると不適切な訳語選択
2. 訳語は適切だが，文全体での整合性を考慮すると，より適切な候補が存在する場合の訳語選択

5.1.1　原言語文の文脈に応じた適切な訳語の選択

提案手法では，ベースラインで翻訳が誤っていた単語に対して，文脈を考慮することで適切な訳語を選択する例が見られた．以下の例は，提案手法で訳語が改善された例である．

入力文: 図 9 中 、 矢印 は 熱 の 流れ を 示し て いる .

参照訳: In FIG . 9 , the arrows indicate the flow of heat .

ベースライン: FIG . 9 shows a flow of the arrows in the thermal .
Mauser: FIG . 9 shows a flow of the arrows in the heat .
提案手法: In FIG . 9 , the arrow shows a flow of heat .

ベースラインと Mauser には文法的な誤りもあるが，本節では訳語選択について考察する．

ベースラインと，Mauser の手法，提案手法とを比較すると，“熱” の訳語が異なっている．Mauser と提案手法では，“熱” を “heat” と訳しているが，ベースラインでは，“thermal” と訳している．“thermal” は “熱” 単体の訳語としては誤りではないが，入力文の文脈では，適切な翻訳とは言えない．

入力文で登場する “熱” は，図示される流れとしての “熱” であり，この場合は，“熱の流れ” をまとめて，“flow of heat” と翻訳することが適切である．しかし，本実験で用いた訓練データからは，“熱の流れ” というフレーズは取得されなかった．そのため，“熱” と “流れ” は分割して訳す必要がある．ただし，全ての手法において，“流れ” は “flow” と適切に訳されていたため，ここでは，“熱” の訳語選択のみが問題になる．

“熱” を “thermal” と訳した際のベースラインの翻訳モデルの働きについて考える．ある単語の翻訳は，その単語とアライメントされた単語から選択される．この場合のアライメントされた単語とは，“thermal” や “heat” が該当する．アライメントされた単語の中で翻訳確率が高い単語が出力として選択されやすい．されやすいとは，まず，仮説として出力文の候補をいくつか生成し，それに対してその他のモデルの評価値の合算によって評価されるため，翻訳確率が評価の全てではないことを表す．この場合では，“熱” の翻訳候補として，“thermal” が “heat” より高い翻訳確率を持っていたため，“thermal” が有利に働き訳語として採用された．その結果，単語単位での翻訳は誤っていないが，文の翻訳として一部分が不適切な出力文が生成された．

提案手法では，“熱” と “流れ” を ‘flow of heat” という適切なフレーズに翻訳している．提案手法では，語彙モデルのパラメータとして，語義に基づく大局的な情報が用いられている．先ほどと同じく，“熱” を翻訳することを考える．この際，同文中に出てきた “流れ” や “矢印” といった語が語義に基づく大局的な情報として用いられる．活用語は表層形を，非活用語である名詞と副詞は表層形に加えて語義が素性として使用される．例では，図や熱，流れが語義として扱われ，語義に含まれる単語，図表や，図式，熱エネルギー，フローなどが同じ文脈内に出現する単語として扱われる．訓練データ上で，任意の目的言語に訳される場合，それらが生起していれば，翻訳において有利に働くように考慮される．このケースでは，“熱の流れ” は，慣例的に “flow of heat” と訳されるこ

とが多く，語彙モデルは，“thermal” よりも “heat” に高い評価値を与えた．

この例のように提案手法は，語義に基づいた大局的な情報を用いることで，単語，フレーズ単位では適切だが，文脈を考慮すると不適切な訳語選択を改善する働きがある．これは，語義によって大局的かつ柔軟な文脈を捉えているためである．

5.1.2 文単位での整合性を保つ訳語の選択

提案手法は，各フレーズ，単語の語彙選択において，文単位で整合性を保つような訳語を選択する傾向が見られた．以下に翻訳例を示す．

入力文: ビーフ が とても 美味しかっ た です .
参照訳: The beef was great .
ベースライン: I really enjoyed it beef .
Mauser: Beef is really enjoyed it .
提案手法: The beef was excellent .

翻訳例では，“とても 美味しかっ た です .” に対する訳語が変化している．ベースラインと Mauser は，“really”，“enjoyed it” を，提案手法は “was excellent” を選択している．“美味しかっ た です .” と “enjoyed it .” はフレーズ対として登録されており，翻訳確率も他の表現と比較して高い値を持つ．これらがフレーズ対として対応関係が取られていること自体は，“it” に当たる原言語側の単語が省略されやすいことから考えても適切である．“enjoyed” の表現を使用すること自体は適切であるが，問題は，その場合は目的語を “beef” にする必要がある入力文に対して “enjoyed it” の訳語を選択したことである．このため，ベースラインでは，“really enjoyed” と “beef” の間に不要な “it” が挿入された出力文を生成してしまった．また，Mauser では，主語が “I” になるような文において，誤って “Beef” を選択している．

提案手法では，“とても 美味しかっ た です .” を翻訳する際に，それぞれの候補（“enjoyed it”，“was excellent”，etc.）で原言語の語義を考慮する．例として，訓練データ中には，“enjoyed it” と “beef” が同時に出現する文は存在しないが，“was excellent” と “beef” が同時に出現する文は存在する．そのため，単語，フレーズの翻訳確率では有利であった “enjoyed it” の確率より，“was excellent” が選択されやすくなる．

このように，語義に基づいた大局的な情報を用いることで，訓練データ中に類似する文がある場合，提案手法はそれを再現するよう訳語に対して有利に働く傾向がある．こ

れは，単語，フレーズの訳語選択が適切になるだけではなく，文単位で適切な組み合わせになるような訳語を選択する．

Mauser も同じように原言語を考慮するが，原言語の表層形だけではデータが足りず，素性がばらついて適切な推定が行えなかったため，適切な翻訳にならなかったと考えられる．

例のような誤りは，アライメントされた単語，フレーズの翻訳確率が，文脈と関係なく，単語対，フレーズ対だけを考慮して計算されていることが原因である．各単語やフレーズ対は，原言語側と目的言語側で n 対 n 対応をしているわけではなく，“美味しかった です ．” と “enjoyed it .” のように，どちらかの単語が省略されていることも多くある．そういった単語やフレーズを組み合わせて文を生成する場合は，文に含まれる単語やフレーズの訳語同士が組み合わせられるかどうかを考慮する必要がある．この問題については，局所的な文脈を捉える言語モデルがある程度の補正をかけるが，ベースラインの例のように完全ではない．

提案手法は，語義に基づく大局的な情報を考慮して訳語を推定することで，単語，フレーズ単位だけでなく，文全体の整合性を考慮することができる．

このとき，提案手法は，語義に基づく大局的な情報という文単位で統一の素性を用いる．そのため，文の整合性が保たれやすいという利点を持つ．このような訳語の変化は，誤りではなくても，コーパス上で一般的な表現で翻訳するように働く．以下の文では，“以下に示す” という翻訳表現の変化が見られる．

入力文: 以下 に 示す コラム 選択 回路 １ １ ０ は 、 ライト アンプ の 機能 を 併せ もつ ．

参照訳: The following column selection circuit 110 also functions as a write amplifier .

ベースライン: Also , the write amplifier has a function of column select circuit 110 to be described below .

Mauser: The column selecting circuit 110 has a function of the write amplifiers are shown below .

提案手法: Following column selecting circuit 110 is also functions of the write amplifier .

ベースラインでは，“described below”，Mauser では，“shown below”，提案手法では，“following ” と，全ての手法で誤りではないが，異なる表現が取られている．ベースラ

インは文全体で見ると誤りが多いが，Mauser の手法と提案手法は，文全体での意味もある程度適切である．しかし，Mauser の手法で訳された “shown below” は，“〜〜を以下に示す．” という訳され方が一般的であり，入力文と比較すると微妙に言い回しが変化している．そのため，Mauser の手法は，正確な翻訳とは言えない．入力文に対しては，提案手法が選択した “following ” が一番適切な訳語である．

5.2 コーパスによる差異

実験の翻訳には，英翻訳エンジン学習・評価用対訳コーパスと NTCIR-9 特許機械翻訳テストコレクションを使用した．英翻訳エンジン学習・評価用対訳コーパスの特徴としては，旅行英会話文，文長は短め，疑問文が多く含まれていることなどが挙げられる．NTCIR-9 特許機械翻訳テストコレクションの特徴としては，特許文書，文長は長め，専門用語が多く含まれていることなどが挙げられる．

提案手法の翻訳精度は，ベースラインと比較して，日英翻訳エンジン学習・評価用対訳コーパスで $BLEU$ の平均が 1.30 ポイント，$RIBES$ の平均が 0.34 ポイントの向上，NTCIR-9 特許機械翻訳テストコレクションでは $BLEU$ の平均が 0.09 ポイント，$RIBES$ の平均が 0.37 ポイントの向上と日英翻訳エンジン学習・評価用対訳コーパスの方が高い値を示した．

提案手法は，英翻訳エンジン学習・評価用対訳コーパスにおいて，より有効だった．この原因としては，動詞や助動詞などの時制の変化や疑問文などを文の素性として上手く利用できたことが挙げられる．NTCIR-9 特許機械翻訳テストコレクションは，過去形はほとんど出現せず，疑問文は一文も出てこない．例として，“あり ます か” や “ですか” といったフレーズは，疑問文かつ現在形であると推定できるため，文全体の単語に対して有用な素性であるといえる．その一方で，表層形を素性として用いた Mauser の手法は，ベースラインより精度が落ちる結果となった．主な原因として，訓練データの不足による，文脈によるものではない誤った素性の偏りがあったことが考えられる．提案手法では，単語を語義として扱うため，データ不足による偏りは軽減される．そのため，適切に文脈を考慮できたと考えらえれる．

NTCIR-9 特許機械翻訳テストコレクションにおける提案手法の翻訳精度は，英翻訳エンジン学習・評価用対訳コーパスにおける翻訳精度の改善と比較すると小さく，提案手法がベースラインを僅かに上回る結果であった．しかし，提案手法が改善することのできる誤り自体は，英翻訳エンジン学習・評価用対訳コーパスよりも多く存在した．これ

は，提案手法の原言語の表層形と語義によって目的言語の単語を推定するという処理が，適切に動作しなかったことを表している．しかし，いずれの結果も Mauser の手法よりは精度が向上している．また，Mauser の手法はベースラインを下回っている．以上のことから，特許文の翻訳においては表層形の素性が，精度向上よりもノイズになる割合の方が多いと考えらえる．これは，特許文において，分野には依存しないが，出現は偏っている “FIG” など，同じ表層形だが，分野ごとに意味が異なる単語が多かったためと考えられる．また，文長の長さと文構造の複雑さの影響も考えられる．

また，提案手法とベースラインを比較して，英翻訳エンジン学習・評価用対訳コーパスほど精度が向上しなかった理由の 1 つには，名詞と副詞を WordNet を用いて汎化させる処理が，専門用語の多い特許文ではうまく働かなかったことが挙げられる．

以上のことから，提案手法が有効に働くコーパスの条件として次の 2 点が挙げられる．

1. 疑問文やいくつかの時制を含んでおり，それに伴う語彙の変化がある
2. コーパスに生起する単語の多くが WordNet に登録されている

提案手法は語義に基づいた大局的な情報として，原言語の表層形と，名詞，副詞の語義を用いて目的言語の単語を推定する．語義は，目的言語も利用して構築され，原言語から構築された語義との積集合が用いられる．1 つ目の条件を満たしていることで，表層形と語義の素性は，文脈の特徴を捉えることができる．2 つ目の条件は，語義を用いて，意味が同じだが表層形が異なる単語をまとめて扱うために必要である．これにより，柔軟な文脈を考慮できる．

本研究における実験では，日英翻訳エンジン学習・評価用対訳コーパスは，2 つの条件を満たしていた．NTCIR-9 特許機械翻訳テストコレクションは，両方の条件を十分には満たしていなかったが，2 つ目の条件はある程度満たしていたため，翻訳精度におけるある程度の向上が見られた．その他，条件には挙げていないが，文長と文構造も大きく影響すると考えられる．文長は，短すぎると十分な素性を取得できず，長すぎても不要な素性がノイズになると考えられる．文構造については，単純なものほど有利に働きやすい傾向は見られたが，具体的にどのような構造になると精度が落ちるのかは，今回の実験では明らかにならなかった．

5.3 大局的な情報と提案手法における課題

大局的な情報は，全ての単語の翻訳において有効なわけではなく，ノイズとなる場合もある．大局的な情報がノイズとなった例として，以下に，ベースラインと提案手法で

は適切な訳語を選択したが，Mauser では誤った翻訳を示す．

入力文: 本 発明 は この うち 、 ガラス 製 の ロッド レンズ の 耐候 性 を 改善 する 効果 が ある .

参照訳: Of those , the present invention is effective for improving the weather resistance of the glass rod lenses .

ベースライン: The present invention has the effect of improving the glass rod lens of the weather resistance .

Mauser: The glass rod lens of the present invention is to improve the effect of weather resistance .

提案手法: The present invention has the effect of improving the weather resistance of the glass rod lens .

“ 改善 する” の翻訳に対して，ベースラインと提案手法では，“improving”，Mauser の手法では，“to improve” を選択している．

Mauser の手法では，単語の表層形によって目的言語の単語が推定される．しかし，単語の表層形には，“ 改善 する” と “improving” を強く結びつけるような素性はなく，僅かながら “improving” よりも “to improve” に高い評価値を与えたため翻訳を誤った．この例のように，ある単語の翻訳に関して，その単語が生起した文の単語が訳語選択に有効でない場合，大局的な情報は，不適切であるといえる．この例では，語義に基づく大局的な情報を用いた提案手法では，“improving” と適切な訳語を選択しているが，これは語義に基づく大局的な情報が有効に働いたわけではなく，どの訳語候補にも，同等の評価値を与えたためである．このように，大局的な情報，語義に基づく大局的な情報による推定は不適切な単語も多くある．現在は，その場合，どの訳語候補にも同等の評価値を与えるという動きが望ましいが，素性の少しの偏りによって誤った訳語選択をすることも多くある．

提案手法は，フレーズモデルが生成した仮説に対して評価を行う．そのため，アライメントの誤りが原因で，不適切なフレーズや単語とその仮説が入力されることがある．誤ったアライメントのフレーズ，単語との仮説に対する翻訳確率は，一般的に翻訳確率も低くなることが多い．しかし，提案手法では，それらを考慮せずに評価を行うため，高い評価値を与えてしまうことがある．その誤りは，ある程度までは翻訳確率が補正するが，提案手法の評価値がそれを上回る評価値を与えることがある．

語義に基づく大局的な情報は，翻訳する全ての単語に対して評価値を与える．しかし，翻訳する単語全てが文脈に依存するわけではなく，文脈を用いずに翻訳したほうが適切な単語，フレーズも存在する．提案手法は全ての単語，フレーズに対して評価値を与えるため，不要な単語，フレーズに評価値を与えてしまうことがある．

第6章　おわりに

6.1　まとめ

本研究では，統計的機械翻訳において，語義に基づく大局的な情報を用いた語彙モデルを提案した．本研究で提案した語彙モデルを，統計的機械翻訳システム Moses による日英翻訳実験によって評価した．提案手法は，コーパスに日英翻訳エンジン学習・評価用対訳コーパスと NTCIR-9 特許機械翻訳テストコレクションを用いた場合，提案手法なしのベースラインと比較して，自動評価尺度 $BLEU$ の平均が 1.30 ポイントと 0.09 ポイントの向上，$RIBES$ の平均が 0.34 ポイントと 0.37 ポイントの向上を示した．

提案手法は，単語を翻訳する際に，その同一文中に出現する単語の表層形とその語義を素性として用いて，文脈を考慮した目的言語の推定を行う．単語を語義に置き換えることで，単語のままでは区別されていたものが，同じ素性として利用できる．これは，目的言語の単語に対する素性を，語義によってまとめあげる働きをする．語義に変換することで表層形が異なる単語だけでなく，概念の近い語も同じ文脈内に出現する語として扱える．語義を用いることにより，表層形のみの使用と比較して，柔軟に文脈を考慮することができる．

語義に基づく大局的な情報を用いて目的言語を推定することにより，単語，フレーズ単位では適切だが，文脈を考慮すると不適切な訳語選択を改善することを確認した．また，文単位で見た場合の訳語も適切だが，文全体での整合性を考慮すると，単語，フレーズが適切に組み合わすことができない訳語選択についても，文の整合性を保つように訳語が選ばれることを示した．これは，提案手法は同文中の単語の翻訳には同じ素性を用いて推定を行うため，訓練中に同時に出現していた単語，フレーズの組み合わせが再現されやすくなるためである．

6.2 今後の課題

本研究では，語義に基づいた大局的な情報が翻訳に有用であることを示した．本研究では，全ての単語，フレーズを区別することなく提案した語彙モデルを適用した．現在は，同文中に出現する単語やフレーズに対しては同じ素性から同じ評価値が与えられている．今後の課題として，単語やフレーズの粒度や条件に合致した，語義に基づいた大局的な情報の分析，適用を行う事で，より高性能な語彙モデルが構築できると考える．

謝辞

本研究を進めるにあたり，ご指導を賜った筑波大学図書館情報メディア系の 関洋平助教，および，佐藤哲司教授に心から感謝致します。また，審査を引き受けて頂いた同研究系の上保秀夫准教授に感謝致します。

本研究で使用した，NTCIR 特許機械翻訳テストコレクション及び，日英翻訳エンジン学習・評価用対訳コーパスについて，構築，公開に当たってご尽力された方々に対して感謝致します。

最後に，研究を進めるにあたって，様々な面でお世話になりました，コミュニケーション理解研究室，コンテンツ工学研究室の皆様に感謝致します。

参考文献

[1] Marianna Apidianaki, Guillaume Wisniewski, Artem Sokolov, Aurelien Max, and Francois Yvon. WSD for N-best Reranking and Local Language Modeling in SMT. In *Proceedings of Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, pp. 1–9, Jeju, Korea, 2012.

[2] Srinivas Bangalore, Patrick Haffner, and Stephan Kanthak. Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL2007)*, pp. 152–159, Prague, Czech Republic, 2007.

[3] Peter E Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[4] Marine Carpuat and Dekai Wu. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL2007)*, pp. 61–72, Prague, Czech Republic, 2007.

[5] Stanley F Chen and Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[6] Hal Daum. Notes on CG and LM-BFGS Optimization of Logistic Regression. 2004.

[7] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[8] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata. Overview of the Patent Translation task at the NTCIR-8 workshop. In *Proceedings of the 8th NTCIR Workshop*

Meeting on Evaluation of Information Access Technologies (NTCIR-8), pp. 371–376, Tokyo, Japan, 2010.

[9] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K Tsou. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9)*, pp. 559–578, Tokyo, Japan, 2011.

[10] Saša Hasan, Juri Ganitkevitch, Hermann Ney, and Human Language Technology. Triplet Lexicon Models for Statistical Machine Translation. In *Processings of the 2008 Conference on Empirical Methods in Natural Language (EMNLP2008)*, pp. 372–381, Hawaii, USA, 2008.

[11] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 2nd edition, 2009.

[12] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello 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 *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL2007) demonstation session*, pp. 177–180, Prague, Czech Republic, 2007.

[13] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Processings of the 2004 Conference on Empirical Methods in Natural Language (EMNLP2004)*, pp. 230–237, Barcelona, Spain, 2004.

[14] Kow Kuroda, Francis Bond, and Kentaro Torisawa. Why Wikipedia Needs to Make Friends with WordNet. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*, pp. 9–16, Mumbai, India, 2010.

[15] Arne Mauser, Saša Hasan, and Hermann Ney. Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models. In *Processings of the 2009 Conference on Empirical Methods in Natural Language (EMNLP2009)*, pp. 210–218, Singapore, 2009.

[16] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL2003)*, pp. 160–167, Sapporo, Japan, 2003.

[17] Franz Josef Och and Hermann Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002)*, pp. 295–302, Philadelphia, USA, 2002.

[18] Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[19] Franz Josef Och and Hermann Ney. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417–449, 2004.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU:a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002)*, pp. 311–318, Philadelphia, USA, 2002.

[21] Ted Pedersen, Satanjeev Banerjee, and Siddharth Patwardhan. Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. In *Research Report UMSI*, 2005.

[22] Andreas Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002)*, pp. 901–904, Denver, USA, 2002.

[23] Yoshimasa Tsuruoka and Tsujii Junichi. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP2005)*, pp. 467–474, Vancouver, Canada, 2005.

[24] David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. Word-Sense Disambiguation for Machine Translation. In *Proceedings of Human Language Tech-*

nology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP2005), pp. 771–778, Vancouver, Canada, 2005.

[25] Spela Vintar, Darja Fiser, and Aljosa Vrscaj. Were the clocks striking or surprising? Using WSD to improve MT performance. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL2012)*, pp. 87–92, Avignon, France, 2012.

[26] 平尾努, 磯崎秀樹, Kevin Duh, 須藤克仁, 塚田元. RIBES：順位相関に基づく翻訳の自動評価法. 言語処理学会第 17 回年次大会 (NLP2011), pp. 1115–1118, 2011.