修　　士　　論　　文

単語の意味表現の翻訳
およびその統計的機械翻訳への応用

Translation of Word Semantic Representations
and its Application to Statistical Machine Translation

指導教員　　喜連川 優教授

東京大学 大学院情報理工学系研究科
電子情報学専攻

氏　名　　48-146404 石渡 祥之佑

提 出 日　　平成28年2月4日

概 要

近年，マイクロブログや SNS 等，複数の言語で使われる共通のコミュニケーション基盤が普及し，個人の情報発信が世界レベルで盛んになっている．しかし，母国語と異なる言語で書かれた情報を（恐らく有用な情報があると分かっていても）誰もが自由に参照できているとは言い難く，言語の壁は依然深刻である．このように多言語で発信される情報を利活用する上では，言語横断的にことばの意味を取り扱う技術が求められる．現在，ことばの意味を計算処理可能なかたちで扱う方法としては,「ある単語の意味は，その単語と共に出現している単語群によって特徴づけられる」という分布仮説 [1, 2] に基づくアプローチが一般的となっている．しかし，こうした分布仮説に基づく意味表現には，異なる言語の単語間の類似度を測ることができないため，多言語処理技術の枠組みで活用するのが難しいという問題点が存在する．そこで，本研究では「対訳辞書」と「表層の近さ」を意味表現ベクトルの次元間の対応付けに活用することで，高精度で言語依存の意味表現を翻訳する新しい手法を提案した．さらにこの手法を統計的機械翻訳に応用し，未知語を含む文の翻訳品質の向上を実現した．

謝辞

本研究を進めるにあたり，非常に多くの方々のお世話になりました．まずはじめに，指導教官の喜連川優教授に感謝致します．喜連川先生に「君にとって本当に大事なのは，些末な成果を捻り出して論文を書くことではなく，どうすれば君を育ててくれたこの社会が良くなるような貢献をできるのかを考えて行動することです.」と言われてから，私は折にふれて「自分の研究が社会にどんな影響を与えうるか」を考えるようになりました．

次に，本研究の方針からスライドの細部に至るまで，的確な助言を下さった豊田正史准教授に感謝致します．学会の運営や本郷キャンパスでの授業で大変お忙しい時でも，いつも夕方になると研究室に帰ってきてくださり,「なにか相談はありますか」と常に気にかけてくださり，ありがとうございました．

吉永直樹特任准教授と鍛治伸裕特任准教授 (現・Yahoo! JAPAN 研究所 上席研究員) にも感謝致します．お二人にはプログラミングのコツや論文の書き方，発表の仕方，学会での振る舞い方など，研究に関するあらゆることを丁寧に，辛抱強く教えていただきました．吉永先生には修士課程の 2 年間を通して，一番長い時間ご指導頂きました．論文の締切が迫る度に何度も繰り返し文章を添削していただいたり，深夜遅くまで議論に付き合っていただき，ありがとうございました．鍛治先生とは，私がはじめて参加した国際会議の合間に万里の長城へ観光に行ったことが非常に楽しい思い出として印象に残っています．吉永先生と鍛治先生のご指導を受け，自然言語処理分野の研究の面白さに触れたことが，私が博士課程への進学を決めた直接的なきっかけになりました．

続いて，伊藤正彦特任准教授と横山大作特任助教に感謝致します．伊藤先生には毎週の修士ミーティングで研究に関するコメントをいただいたり，輪講前に何度も

繰り返し発表練習に付き合っていただきました．また，横山先生に喜連川研究室の強みである計算機クラスタの使い方を詳細に教えていただいたおかげで，通常は難しいであろう大規模な実験を気軽に行うことができました．

また，卒論の時に研究の楽しさを教えていただき，修士課程の期間中も貴重なデータやアドバイスを下さった Liverpool 大学の Danushka Bollegala 准教授にも感謝致します．Danushka 先生の勧めで修士課程から喜連川研究室に進学したおかげで，充実した 2 年間の研究生活を送ることができました．

研究生活の苦楽を共にした同期の加藤千裕さん，川本貴史君，金洪善さん，小矢島諒君，谷川祐一君に感謝致します．同期の皆と将来の科学技術や自分たちの未来について話しあったり，研究の合間に卓球やテニスでリフレッシュしたり，時には昼間からサボって渋谷で遊んだりしたことは，私にとってかけがえのない楽しい時間でした．皆と共に研究し，共に遊んだことは私の人生の誇りです．

先輩の鈴木有さん，槇佑馬さん，伊東直弘さん，清水翔太さん，劉さんに感謝致します．研究室に入ったばかりの私たちに研究室での生活について教えて下さったり，就職活動について相談に乗って下さり，ありがとうございました．鈴木さんが昨年作られた「修論文字数可視化システム」は，今年も大活躍しました．

後輩の岩成達哉君，小泉実加さん，佐藤翔悦君に感謝致します．3 人はそれぞれの研究テーマについて非常に詳しいので，研究について議論するのがとても刺激的で楽しいです．研究以外についても，コーヒーメーカーを導入してくれたり，修論締め切り直前にアイスをくれたり，スライドの図を書いてくれたりと，私は先輩なのに助けてもらってばかりでした．ありがとうございます．

研究に集中できる最高の研究室にしてくださっている井崎さん，池田さん，越水さんにも深く感謝致します．

最後に,「学費のことは何も気にせず，好きなだけ好きな勉強をしろ」と言ってくれた父と,「実験上手くいった？論文書くの大変？」と笑いながら夕食を用意してくれる母，大騒ぎして家の空気を明るくしてくれる妹と，辛い時に私を励まして前向きな気分にしてくれる丹下恵里さんに，感謝の意を捧げます．

2016 年 2 月 4 日

目 次

図 目 次

表 目 次

第1章　はじめに

近年，マイクロブログやSNS等，複数の言語で使われる共通のコミュニケーション基盤が普及している．これにより，個人の情報発信が世界レベルで盛んになっているが，言語の壁は依然深刻である．たとえばテロや感染病などの社会問題や五輪など世界的イベントについて国際的な世論を知りたい場面や，海外情勢の調査，海外で当地の情報収集を行う場面において，母国語と異なる言語で書かれた情報を（恐らく有用な情報があると分かっていても）誰もが自由に参照できているとは言い難い．こうした状況から，言語をまたぐ情報推薦や言語横断検索，機械翻訳といった様々な多言語処理技術の重要性が増している．

多言語処理の問題をさらに難しくしているのは，解析対象であるテキストに含まれることばの意味や表現が変化し続けている点である．たとえばソーシャルメディア上では，日々新たな話題について世界中の利用者による議論がなされるのに伴い，新しい語や新しい言い回しが絶え間なく生まれている．そのため，機械翻訳や言語横断情報検索といった技術を用いて解析を行う場面では，システムが解釈できない未知語が頻繁に出現する．このような状況下において，我々が様々な言語で書かれた情報を即座に利活用するためには，変化し続ける言葉に柔軟に適応することができる多言語処理技術の実現が必要不可欠である．

ことばが変化するということは，すなわち常に解析対象の文書に未知語や未知の表現が含まれる状況を想定しなければならない，ということである．人間であれば，文書内に出現する未知語の意味をある程度文脈から類推することが可能である．たとえば，“ググる”という語を知らない人であっても，“わからなければそのiPhoneとか使ってググってくれよ”という文を見れば，“ググる”は動詞であり，情報端末で行える動作を表し，なにかを調査する意味を持つ語であることは容易に推測でき

る．これと同様の類推を計算機で実現するための基礎技術として，単語の意味表現を教師なし学習に基づき獲得する手法 [4, 5, 6] が存在する．これらの手法は，いずれも「ある単語の意味は，その単語と共に出現している単語群によって特徴づけられる」という分布仮説 [1, 2] の考え方に基づき，単語の意味をベクトルで表現する．代表的な意味表現としては，単語の意味を共起語の頻度に基づくベクトルとして表現する手法 [4] が挙げられる．単語の意味をベクトルで表現することにより，単語間の意味的な類似性の計算を，ベクトル間の類似度計算 (コサイン類似度など) に帰着させることが可能となる．この手法を用いて未知語のベクトルを作成し，同様に作成した既知語のベクトルと比較することによって，計算機が未知語の意味を文脈から類推することが可能となった．

しかし，これらの分布仮説に基づく意味表現には，異なる言語の単語間の類似度を測ることができないため，多言語処理技術の枠組みで活用するのが難しいという問題点が存在する．これは言語によって単語の分布が全く異なっているからであり，ロシア語を読めない日本人が“м е д у з а”という語の意味を文脈から類推できないのと同様の問題である．この問題に対し，単語のベクトルを他の言語に翻訳することにより，言語横断的に意味表現を比較することを目指した研究が行われている．代表的なものとしては，1990 年代後半に Fung らによって提案された対訳辞書に基づくアプローチ [7] と，2010 年代前半に Mikolov らによって提案された教師あり学習に基づくアプローチ [3] が挙げられる．これらの手法についての詳細は 3.1 節にて述べるが，この 2 種類のアプローチには，いずれも翻訳において重要な情報を，翻訳の過程で活用できていないという問題点が存在する．

上記の問題をふまえ，本研究ではまず (1) 意味表現のより正確な翻訳手法を提案し，(2) それが多言語処理技術の性能向上に寄与することを検証した．(1) に関して，より正確な意味表現の翻訳を実現するためには，Fung らの用いた辞書と Mikolov らの用いた学習という個別の手がかりを双方とも利用できることが理想である．しかし，辞書で定義された単語間の離散的な対応関係を考慮しつつ，連続的な数値最適化を行うことは容易ではない．そこで本研究では，辞書を，翻訳行列を最適化する際の目的関数の報酬項として表現し，この課題を解決する．報酬項により，辞書に

含まれる単語対に対応する翻訳行列の要素の重みをより大きくするよう学習が行われるので，結果として，辞書で定義された対応関係を重視しつつ，同時にそれ以外の要素も適切な重みを持つような翻訳行列を得ることが可能になる．さらに，報酬項を用いて，教師なしで獲得できる単語の表層的な知識 (例: 英語の “importance” とスペイン語の “importante” は表層的に類似している) も学習に組み込む．これにより，さらに高精度な意味表現の翻訳を実現した．実験では提案手法を用いて日本語，中国語，英語，スペイン語の 4 カ国語間の意味表現の翻訳を行い，未知語の訳語選択タスクによる評価を行った．その結果，実験したほぼ全ての言語対において先行研究の手法を大きく上回る翻訳精度が得られた．

(2) に関して，意味表現の翻訳を通して未知語に適切な訳語を与えることができれば，変化し続ける言葉に柔軟に適応することができる多言語処理技術の実現に大きく寄与するはずである．しかし，現時点では上記で述べた意味表現のベクトル翻訳手法を用いたとしても，100%正しい未知語の翻訳は実現できない．そのため，ベクトル翻訳を多言語処理技術に応用するためには，ノイズを含む未知語の翻訳を適切に各アプリケーションに取り込む必要がある．そこで本研究では，未知語の問題が最も顕著となる多言語処理技術である機械翻訳に焦点を当て，この課題の解決に取り組んだ．具体的には，まず (i) ベクトル翻訳によって得られる未知語の訳語候補にベクトルの距離関数に基づく確信度 (翻訳確率) を与え，(ii) それをフレーズベース統計的機械翻訳における翻訳モデル [1] のバックオフモデル [8] として扱う手法を提案する．これにより，文脈からの意味類推を行うベクトル翻訳と,「言語としての自然さ」を担保する機械翻訳システムの言語モデルが補完しあうことにより，文書内に出現した未知語に適切な訳語を与える枠組みが実現した．実験では，機械翻訳において未知語が多く出現する設定で日英翻訳，英日翻訳を行い，提案手法を導入することによりシステムの翻訳品質が大きく改善することを示した．

以下に本研究の貢献をまとめる．

- 単語の意味表現の翻訳における 2 種の先行研究の欠点を補いあうモデルを提案し，当該分野における 2 つの流れを統合した．

[1] 翻訳元言語の句と，翻訳先言語の句の間の翻訳確率を表す．

- 単語の意味表現の翻訳に「単語の表層」という教師無しで獲得できる知識を導入し，先行研究の手法を大きく上回る翻訳精度を達成した．
- 単語の意味表現の翻訳手法を機械翻訳に応用し，未知語を含む文の翻訳品質の向上を実現した．

以降，本論文は以下のように構成されている．

第 2 章　ことばの数学的表現と統計的機械翻訳に関して，本論文の前提となる基本的な事項を説明する．

第 3 章　単語の意味表現の翻訳手法に関する研究と，機械翻訳における未知語の翻訳に関する研究についてそれぞれまとめ，本研究との関連について述べる．

第 4 章　「対訳辞書」や「言語間に存在する表層の類似性」といったあらゆる手がかりを活用し，正確に単語の意味表現を翻訳する，本研究の提案手法について述べる．また，提案手法の有用性を確認するために行った訳語選択タスクでの評価についても述べる．

第 5 章　前章までに述べた意味表現の翻訳手法を，機械翻訳システムに組み込む提案手法について述べる．また，実際に機械翻訳タスクで行った評価について述べ，提案手法の有用性を示す．

第 6 章　全体のまとめと今後の課題について述べる．

第2章　基礎知識

2.1　ことばの意味の数学的表現

ことばの意味を計算処理可能な形式で表現することは，情報検索やウェブマイニングなど，自然言語テキストを対象としたアプリケーションの高度化に必要となる要素技術であり，これまでに多くの研究が行われてきた [9, 10, 11]．本節では以降，コーパスからこうした意味表現を獲得する手法を説明する．

2.1.1　文書内の単語分布に基づく意味表現

多くの言語において，なんらかの意味を表すことばの最小単位は単語である [1] が，はじめて計算処理可能な形で表現されたのは単語ではなく，文書であった [12]．Salton らによって開発された情報検索システム SMART[13] に，世界ではじめて文書をベクトルの形で表現する手法 [14] が導入された．各文書を，含まれる各単語の頻度を次元とするベクトルで表現することによって，文書間の類似度を定量的に測ることが可能となる．これにより，検索クエリに含まれるキーワードにマッチする文書だけでなく，キーワードはマッチしないが関連度の高い文書も検索結果として出力する検索システムが実現された．Salton らによる文書のベクトル表現からヒントを得て，1990 年代初頭に Deerwester ら [15] は単語をベクトルのかたちで表現する手法を提案した．Salton らの文書ベクトルが単語の出現頻度を各次元の値として持つ一方，Deerwester らの単語ベクトルは出現する文書を各次元の値として持つ．この単

[1] 言語によっては文字や接辞も意味を表しうるが，それらが単独で表現できる意味は少ないので，ここでは言及しない．

語ベクトルの類似度を計算することにより，単語の意味の類似度を測る事が可能となった．

2.1.2 分布仮説に基づく意味表現

前節で紹介した意味の表現手法は，いずれも「似た文書は似た単語の分布を持つ」という前提に基づくものであった．しかし，ウェブの普及により個人の情報発信が世界レベルで盛んになっている現代においては，マイクロブログや SNS など，形式や規模が異なる様々な文書が混在している．そのため，文書よりも細かいことばの単位に基づく意味表現が求められている．

現在広く用いられている意味表現は,「ある単語の意味は，その単語と共に出現している単語群によって特徴づけられる」という分布仮説 [1] に基づいている．代表的な手法としては，ある単語の意味を，コーパスにおける共起語の頻度に基づくベクトルとして表現するという方法 [4] が挙げられる．たとえば，以下のような文で構成されるコーパスが与えられたとき，これを用いて *valley* という単語の意味をベクトルで表現することを考える．

I have a dream that one day every valley shall be exalted.

単語 *valley* のベクトル表現は，前後に出現している単語の出現頻度に基づいて構成される．このとき，*valley* の前後何語まで考慮するかを定める必要があるが，ここでは前後 3 語を考えるものとする．そうすると，*valley* の周囲には *one*，*day*，*every*，*shall*，*be*，*exalted* が 1 度ずつ出現していることになる．そのため，単純な共起頻度を使って得られる *valley* のベクトルは各次元をそれぞれ共起語 *I*，*have*，*a*，*dream*，*that*，*one*，*day*，*every*，*valley*，*shall*，*be*，*exalted* に対応させることで，以下のようになる．

$$(0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1)$$

たとえば，上記のベクトルの 6 番目の要素の値が 1 であるが，これは *valley* が *one* と 1 回共起したことを表している．

この例では，簡単のため共起頻度の値をそのまま各次元の値としているが，一般的には，式(2.1)に示す正の自己相互情報量(Pointwise mutual information, PPMI)[16, 17]など共起度を表す尺度に変換したものが用いられる．

$$\mathrm{PPMI}(w, f) = \max\left(\log_2 \frac{P(w, f)}{P(w)P(f)}, 0\right) \tag{2.1}$$

ただし，ここで $P(w)$ は見出し語 w の出現確率，$P(f)$ はベクトルのある次元に対応する語 f の出現確率，$P(w, f)$ は w と f がコーパス内で同時に出現する確率をそれぞれ表す．上記の例ではある単語の前後3語をベクトルの要素として利用したが，このように対象の前後何語までを共起しているとみなすかを文脈窓と呼ぶ．文脈窓の長さは，短くするほど結果のベクトルが単語の統語的な性質を捉え，逆に長くするほど，単語のトピックの性質を捉えることが知られている [18]．

上記で説明した意味表現手法は，単語が他の語と共起する回数を数え上げて得られるため，カウントベースのベクトル [19] とも呼ばれる．本研究の提案手法も，カウントベースのベクトルを対象としたものである．

2.2 統計的機械翻訳

機械翻訳のアイデアは1947年にWarren WeaverとNobert Wienerの手紙のやりとり [20, pp.13–17] で初めて議論された．このときWeaverによって記された「読めない言語を機械によって解読する」というアイデアが，現在の機械翻訳技術の基本となっている．

1950年代に研究が始められた初期の機械翻訳技術は，翻訳対象となる2つの言語に精通した専門家が書き下した言語規則に従って翻訳を行うルールベースの手法が主流であった．一方，1980年代後半に研究が始まった統計的機械翻訳 [21] は大規模な対訳文の集合から自動的に規則を抽出する，統計ベースの手法である．統計的機械翻訳では人間が各言語に固有の文法規則を記述する必要は無く，あらゆる言語対において利用できる点で汎用性に優れているため，この統計ベースの手法が現在主

流となっている．以降，本節では統計的機械翻訳の仕組みと，翻訳結果の評価手法を概観する．

2.2.1 雑音のある通信路モデルとしての機械翻訳

上述した Weaver の「読めない言語を機械によって解読する」というアイデアは，Shannon の雑音のある通信路モデル (Noisy channel model)[22] によって表現できる．原言語 (翻訳前の言語，たとえばフランス語) の文 $\boldsymbol{f}$ に対して，目的言語 (翻訳後の言語，たとえば英語) の全ての文の集合 $\boldsymbol{E}$ の中から，適切な文 $\boldsymbol{e}$ を選んで与える統計的機械翻訳翻訳システムは次式で表される．

$$\boldsymbol{e}^{\star} = \underset{\boldsymbol{e}\in\boldsymbol{E}}{\operatorname{argmax}} P(\boldsymbol{e}|\boldsymbol{f}) \tag{2.2}$$

この式はベイズの定理を用いて以下のように書き換えられる．

$$\begin{aligned} \boldsymbol{e}^{\star} &= \underset{\boldsymbol{e}\in\boldsymbol{E}}{\operatorname{argmax}} \frac{P(\boldsymbol{f}|\boldsymbol{e})P(\boldsymbol{e})}{P(\boldsymbol{f})} && (2.3) \\ &= \underset{\boldsymbol{e}\in\boldsymbol{E}}{\operatorname{argmax}} P(\boldsymbol{f}|\boldsymbol{e})P(\boldsymbol{e}) && (2.4) \end{aligned}$$

式 (2.3) の分母は定数であるため，翻訳の問題は式 (2.4) の最大化問題に帰着させることができる．機械翻訳において，式 (2.4) の $P(\boldsymbol{f}|\boldsymbol{e})$ は翻訳モデル，$P(\boldsymbol{e})$ は言語モデルと呼ばれる．翻訳モデルは文 $\boldsymbol{e}$ から文 $\boldsymbol{f}$ への翻訳の正しさを表現しており，言語モデルは文 $\boldsymbol{e}$ の言語としての流暢さを表現している．機械翻訳システムはこの 2 つのモデルの積が最大となるような目的言語の文を選ぶことにより，意味が正しく，かつ自然な翻訳を出力することを可能とする．

2.2.2 言語モデル

前節で述べたように，言語モデルは統計的機械翻訳システムが言語として自然な文 $\boldsymbol{e}$ を出力することを保証する．自然な文とは，すなわち (1) 文法が正しく，かつ (2) 語彙の選択が適切な文のことである．たとえば，以下の訳文候補が生成されているとする．

- $\boldsymbol{e_1}$: I drink coffee.
- $\boldsymbol{e_2}$: I coffee drink.
- $\boldsymbol{e_3}$: I am coffee.

この 3 つの文を比較すると，$\boldsymbol{e_1}$ は文法も語彙の選択も正しいが，$\boldsymbol{e_2}$ は文法に誤りがあり，$\boldsymbol{e_3}$ は語彙の選択が不適切である．言語モデルの役割とは，このうち $\boldsymbol{e_1}$ のような訳文候補に高い確率を与え，システムが正しい文が出力されやすくすることである．

正しい文の生成確率が高くなるような言語モデルは，大規模な英語の文書集合 (以降，コーパス) を学習データとして用いることで獲得することが可能である．たとえば，100 万文で構成されるコーパスに “I drink coffee.” という文が 200 回出現していれば，$P(\boldsymbol{e} = \text{I drink coffee.}) = 0.0002$ と求められる．しかし，この単純なモデルには，長い文の確率が 0 になりやすいという問題点も存在している．これは，長い文であるほどコーパス内に含まれない可能性が高くなることに起因している．この問題に対処するため，統計的機械翻訳の分野では n-gram 言語モデル [2] と呼ばれる手法が広く利用されている．n-gram 言語モデルでは文全体の出現確率を求めるのではなく，文を構成する単語の系列の出現確率を求める．たとえば上記の確率は，文頭記号 $\langle s \rangle$ と文末記号 $\langle /s \rangle$ を導入し，次式のように変換される．

$$\begin{aligned} & P(\boldsymbol{e} = \text{I drink coffee.}) \\ & = P(\langle s \rangle, \text{I}, \text{drink}, \text{coffee}, \langle /s \rangle) \end{aligned} \tag{2.5}$$

[2] n-gram とは，隣接して出現する n 単語のことである．$n = 1$ であれば unigram，$n = 2$ であれば bigram，$n = 3$ であれば trigram と呼ぶ．$n > 3$ のときは 4-gram，5-gram などと呼ぶ．

式 (2.5) に確率の連鎖規則を適用すると，次式のように条件付き確率の積で書き換えられる．

$$
\begin{aligned}
&P(\langle s\rangle, \mathrm{I}, \mathrm{drink}, \mathrm{coffee}, \langle /s\rangle) \\
&= P(e_1 = \mathrm{I} | e_0 = \langle s\rangle) \\
&\times P(e_2 = \mathrm{drink} | e_0 = \langle s\rangle, e_1 = \mathrm{I}) \\
&\times P(e_3 = \mathrm{coffee} | e_0 = \langle s\rangle, e_1 = \mathrm{I}, e_2 = \mathrm{drink}) \\
&\times P(e_4 = \langle /s\rangle | e_0 = \langle s\rangle, e_1 = \mathrm{I}, e_2 = \mathrm{drink}, e_3 = \mathrm{coffee})
\end{aligned} \tag{2.6}
$$

ここで，n-gram 言語モデルは上式に近似を適用し，e_i よりも前に現れる全ての単語ではなく，直前の $n-1$ 単語のみを考慮した条件付き確率とする．たとえば $n = 2$ のとき，(2.6) の条件付き確率は次式のように近似される．

$$
\begin{aligned}
&P(\langle s\rangle, \mathrm{I}, \mathrm{drink}, \mathrm{coffee}, \langle /s\rangle) \\
&= P(e_1 = \mathrm{I} | e_0 = \langle s\rangle) \\
&\times P(e_2 = \mathrm{drink} | e_1 = \mathrm{I}) \\
&\times P(e_3 = \mathrm{coffee} | e_2 = \mathrm{drink}) \\
&\times P(e_4 = \langle /s\rangle | e_3 = \mathrm{coffee})
\end{aligned} \tag{2.7}
$$

この近似により，コーパス内に存在しない長い文に対しても確率を与えることが可能となる．

2.2.3 フレーズベース統計的機械翻訳

現在主流の統計的機械翻訳は，2000 年代前半に提案されたフレーズベース統計的機械翻訳 [23] と呼ばれる手法に基づくものである．フレーズベース機械翻訳における翻訳モデルは，文を構成する句単位に翻訳する句翻訳モデルと，句翻訳モデルが生成した句翻訳を適切な順番に並び替える役割を担う並び替えモデルの 2 つで構成

される．具体的には，式 (2.4) で示した機械翻訳システムを次式のかたちで表現する．

$$\begin{aligned} \boldsymbol{e}^{\star} &= \underset{\boldsymbol{e}\in\boldsymbol{E}}{\operatorname{argmax}}\, P(\boldsymbol{f}|\boldsymbol{e})P(\boldsymbol{e}) \\ &= \underset{\boldsymbol{e}\in\boldsymbol{E}}{\operatorname{argmax}} \sum_{\boldsymbol{\phi},\boldsymbol{\alpha}} P(\boldsymbol{f},\boldsymbol{\phi},\boldsymbol{\alpha}|\boldsymbol{e})P(\boldsymbol{e}) \\ &\approx \underset{\boldsymbol{e}\in\boldsymbol{E}}{\operatorname{argmax}} \sum_{\boldsymbol{\phi},\boldsymbol{\alpha}} P(\boldsymbol{f},\boldsymbol{\alpha}|\boldsymbol{\phi})P(\boldsymbol{\phi}|\boldsymbol{e})P(\boldsymbol{e}) \end{aligned} \tag{2.8}$$

ここで，新たに導入した隠れ変数 $\boldsymbol{\alpha}$ は句の並び順を表すベクトルであり，$\boldsymbol{\phi}$ は句の翻訳ペアを要素として持つベクトルである．式 (2.8) の右辺は，目的言語の文 $\boldsymbol{e}$ が (1) 言語モデル $P(\boldsymbol{e})$ により生成され，(2) 句翻訳モデル $P(\boldsymbol{\phi}|\boldsymbol{e})$ により句ごとに翻訳され，(3) 並び替えモデル $P(\boldsymbol{f},\boldsymbol{\alpha}|\boldsymbol{\phi})$ により並び替えられた結果，原言語の文 $\boldsymbol{f}$ が得られることを示している．

ここで用いる句翻訳モデルと並び替えモデルの学習は，$P(\boldsymbol{f}|\boldsymbol{e})$ を人手によって作成された対訳文 $\langle \boldsymbol{e},\boldsymbol{f}\rangle$ へ適用した際の尤度を最大化する問題に帰着する．また，フレーズベース機械翻訳による翻訳は，未知の入力文 $\boldsymbol{f}$ に対して目的言語の全ての文 $\boldsymbol{E}$ を生成し，式 (2.8) の最適化問題を解くことにより，最適な翻訳 $\boldsymbol{e}^{\star}$ を選択することに帰着する．ただし，実際のシステムにおいては目的言語の全ての文 $\boldsymbol{E}$ を生成することは不可能であるため，並び替えモデルの並び替えパターンに制約を設けることにより，探索空間を削減する．

現在までに，フレーズベース統計的機械翻訳は言語モデル，翻訳モデルを問わず多くの研究がなされている．本研究もこれに倣い，オープンソースの統計的機械翻訳ツールキットである Moses[3] を用いて実験を行った．

2.2.4 機械翻訳の評価尺度

翻訳品質の評価手段には，人間による主観評価と機械による自動評価が存在する．機械翻訳の目的は人間に理解できる訳文を出力することであるので，主観評価は妥当な評価手法であると考えられる．しかし，人間による主観評価には (1) 異なる評

[3] http://www.statmt.org/moses/

価者で一定した評価を保つのが困難であり，(2) 再現性がなく，(3) 非常に時間とコストがかかるといった問題点が存在する [24, pp.47–57]．また，機械翻訳システムは調整が必要な統計モデルのパラメタを多く含むため，自動的に，簡単な統計量で評価できる手法が求められていた．

2000 年代前半に Papineni らが提案した BLEU(bilingual evaluation understudy) は，システムの出力文を構成する n-gram が人手により作成された正解の訳文 (以降，参照訳と呼ぶ) と符合する割合をもとに計算される．具体的には，次式のように unigram から N-gram までの適合率 p_n を計算し，それらの幾何平均をとることにより計算される．

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.9}$$

上式の適合率 p_n には，システムが確信度の高い単語のみを翻訳し，極端に短い文を出力した場合でも高くなってしまうという問題点が存在する．この問題に対処するため，Papineni らは次式の罰則項 BP(brevity penalty) を導入し，出力文の単語数 c が参照訳の単語数 r よりも短い出力文の BLEU スコアが低くなるようにした．

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{2.10}$$

上記で紹介した BLEU は機械翻訳の自動評価尺度として多くの研究に採用されているため，本研究でもこの評価尺度により実験を行った．

第3章　関連研究

以降，本章では単語の意味表現の翻訳に関する研究（3.1 節）と，機械翻訳における未知語の翻訳に関する研究（3.2 節）についてそれぞれまとめ，本研究との関連について述べる．

3.1　単語の意味表現の翻訳

1 章で述べたように，意味表現は「似た意味の語は似たベクトルで表現される」という性質を持ち，類義語検出 [19] や言い換え検出 [25]，対話分析 [26] など，多くの自然言語処理技術で応用されている．しかし，これらの意味表現は一般的に単言語コーパスから学習されるため，得られる意味の表現は単一の言語内に閉じている．すなわち，言語が異なる単語（例：日本語の “猫” と英語の “cat”）の意味類似性を計算することができない．この問題を解決するアイデアとして，片方のベクトル（例：“猫” のベクトル）をもう一方のベクトルの言語（例：英語）へ翻訳してから，ベクトル間の比較を行うというものがある．ここでいうベクトルの翻訳には，大きく分けて 2 種のアプローチがある．

1 つめは，辞書を用いて原言語 (翻訳元の言語) から目的言語 (翻訳先の言語) へベクトルを翻訳する手法 [7] である．この研究は機械翻訳における対訳辞書の自動抽出を目的とし，1998 年に Fung らによって始められた．Fung らの研究が対象としていた意味表現はカウントベースの単語ベクトル [1][4] である．カウントベースの単語ベクトルは，2.1.2 節で紹介した各次元が他の単語との共起頻度に対応しているベクトルのことであり，他の意味表現手法と同様に，似た意味を持つ 2 つの語のベクト

[1] Distributional representation（分布表現）とも呼ばれる．

ル間の類似度が高くなることが知られている [12, 27]．この研究で提案されている単語ベクトルの翻訳手法では，既存の対訳辞書を用いて対応する次元間の値を直接移行することでベクトルを（部分的に）変換する．この手法には，翻訳に用いる対訳辞書のカバレッジが低いと，翻訳元言語の単語ベクトルの多くの次元が翻訳先言語のベクトルの次元に対応させることができないという弱点が存在する．

2つめのアプローチは，原言語から目的言語への翻訳を教師あり学習によって学習する手法 [3, 28] である．2013 年に Mikolov らによって提案されたこの手法は，ベクトル間の対応をベクトル間の線形変換（翻訳行列）を求める問題として定式化し，教師あり学習によってベクトル間の対応関係を間接的に求めることで，前述した Fung らの手法の弱点に対処している．Lazaridou ら [28] は Mikolov らの用いた最小二乗法ではなく，マージン最大化による学習を用いる手法を提案し，翻訳精度さらに向上させた．Mikolov らと Lazaridou らの研究ではカウントベースの単語ベクトルは用いられておらず，ニューラルネットワーク言語モデルにより学習するベクトル [2][5, 6] が意味表現として利用されている．この意味表現手法はカウントベースのベクトルと異なり，各次元が明示的に他の単語との共起頻度を表さないため，人間によるモデルの解釈が困難であるという欠点がある．また，同一言語内の意味の類似性判定タスクに用いる場合，カウントベースの意味表現手法よりもニューラルネットワークベースの意味表現手法を用いたほうが優れた精度が得られたという報告 [19] もあるが，Levy ら [29] による最新の報告では，カウントベースのベクトルに特異値分解などを施すことで，その質をニューラルネットワークベースの単語ベクトルと同程度に向上させることが可能であるとされている．

上記で述べた意味表現の翻訳とは異なるアプローチであるが，異言語に共通のベクトル表現を学習する手法も研究されている．Hermann ら [30] は「複数の言語から単語の表現を学習することで，言語非依存なもの (すなわち，意味) を精度よく捉えられる」と主張し，翻訳の必要ない言語非依存なベクトル表現を提案している．しかし，こうした単語の表現を学習するためには，莫大な対訳コーパスが必要となるため，対訳コーパスの存在しない言語対には適用できない．また，対訳が学習デー

[2] Distributed representation（分散表現）とも呼ばれる．

タに存在しない新語や，語の新用法にも適用できないという短所も存在する．

本研究では，以上で述べた異言語に共通のベクトル表現を学習する手法の欠点をふまえ，Fung ら [7] や Mikolov ら [3] と同じく，単語ベクトルを翻訳するアプローチを採用する．このアプローチは，(1) 小規模な単語の対訳辞書のみを学習データとして利用する点と，(2) 学習データに存在しない語に対しても訳語を与えられる点で，上述した異言語に共通のベクトル表現を学習するアプローチよりも優れている．本研究は，Mikolov らの教師あり学習を用いたベクトル翻訳という枠組みに，Fung らが用いた辞書の情報を学習の手がかりとして取りこむことで，これらの短所を補い，長所を組み合わせた手法となっている．

3.2 機械翻訳における未知語の翻訳

統計的機械翻訳の大きな課題として，学習データのドメインとテストデータのドメインが異なる場合に翻訳精度が著しく落ちる [31] 点が挙げられる．これは，翻訳モデルが対訳コーパスに現れた表現しか適切に訳出できないという性質に起因する．たとえば，特許の対訳文から学習されたモデルを用いて，旅行会話の文を正確，かつ流暢に翻訳することは非常に困難となる．こうした問題は旅行会話ドメインの対訳コーパスが存在すれば容易に回避可能であるが，プロの翻訳者による対訳コーパスの作成は多大なコストを要する．そのため，任意のドメインに対して対訳コーパスを作成することは現実的な解決方法とは言えない．

以上で述べたこうした問題意識から，既存の対訳コーパスで学習されたモデルをテストデータのドメインに適応させることで，異なるドメインの文に対しても正確な翻訳を与える研究が数多くなされている．これらドメイン適応の研究の多くは，対象ドメインの対訳コーパスが十分に得られない状況を想定して行われている．しかし，本研究のように対象ドメインの対訳コーパスが全く存在しない状況を想定している研究は少ない．ドメイン適応に対象ドメインの対訳コーパスを全く必要としない点で，Yamamoto ら [32] の研究は本研究と共通している．彼らは学習データのドメインも，テストデータのドメインも未知である状況を想定し，教師なし学習に

基づくドメイン適応の手法を提案した．Yamamoto らの提案手法では，学習データである対訳コーパスをクラスタリングにより分割し，各クラスタをひとつのドメインとみなす．分割されたコーパスを用いて複数の翻訳モデルを学習し，テスト時にはいずれかの翻訳モデルを選択して利用することで，テストデータと関連の強い学習データの影響を受けやすくしている．

また，Mathur ら [33] は個別に作成された複数の対訳コーパスを活用して，翻訳モデルのドメイン適応を行う手法を提案している．彼らは TED talk，ニュース，ソフトウェアのマニュアル等 11 種類の対訳コーパスを元に翻訳モデルを学習し，それらのモデルを線形補間により組み合わせることで近似的にドメイン適応を行う手法を提案した．Mathur らは実験により，対象ドメインの対訳コーパスが全く存在しない場合でも，任意のテストデータに対してある程度ドメイン適応が可能であることを示した．

これらと異なるアプローチとして，Wu ら [34] が提案した対象ドメインの対訳辞書を用いる手法も存在する．Wu らは対象ドメインの対訳辞書を用いて人工的に翻訳モデルを作成し，対訳コーパスから自動的に得られた翻訳モデルと組み合わせることで，学習データと異なるドメインの文の翻訳精度を向上させた．

Yamamoto らと Mathur らの手法は，いずれも対訳コーパスから得られる翻訳モデルを調整することで，一部の学習データが翻訳に及ぼす悪影響を減らすアプローチと考えることができる．これに対し，Wu らの手法は未知語に訳を与えるものであり，学習データに存在しない情報を追加するアプローチと考えることができる．この点において，Wu らの研究は本研究と共通している．しかし，Wu らが人手で作られた対象ドメインの対訳辞書を用いている一方，本研究ではドメイン特化しない汎用的な対訳辞書を用いる点では異なっており，本研究の提案手法の方がより汎用性が高い．

第4章　単語ベクトルの翻訳

1 章で述べたように，単語の意味を表現するベクトル (以降，単語ベクトルと呼ぶ) は一般的に同じ言語のテキストコーパスから作られるため，異なる言語に属する 2 単語の意味表現をそのまま比較することはできない．以降，4.1 節ではこの問題に対処するために我々が提案する，ある言語における単語ベクトルを他の言語のベクトルに翻訳する手法について述べる．続いて，4.2 節で提案手法の翻訳翻訳精度を評価するために行った，訳語選択タスクでの評価実験について述べる．

4.1　提案手法：正確な単語ベクトルの翻訳

4.1.1　線形変換による単語ベクトルの翻訳

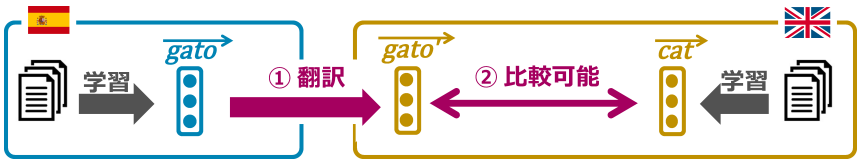

図 4.1: 単語ベクトルの翻訳 [3] による言語の異なる単語の類似性判定

一般的な単語ベクトルには，異なる言語の単語間の類似性判定に用いることが難しいため，言語を横断するアプリケーションに利用しにくいという問題点が存在する．この問題に対し，Mikolov ら [3] は単語ベクトルを他の言語のベクトルに翻訳する手法を提案し，言語をまたいで単語の意味的な類似性を比較することを可能とした．たとえば，図 4.1 に示すように，この手法によりスペイン語の “*gato*” のベクトルを英語の空間へ翻訳することにより，本来比較できなかった “*cat*” のベクトルと

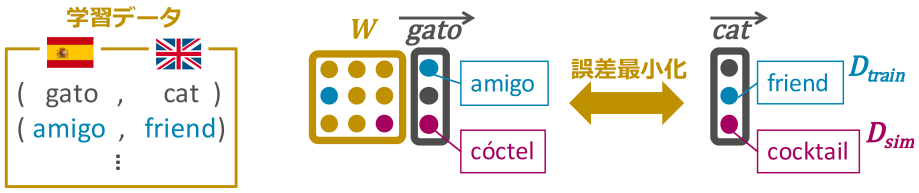

図 4.2: 提案手法：翻訳行列を学習する過程で，文脈語ペアと対応する行列の成分 (図中 $\boldsymbol{W}$ の青い要素と赤い要素) に重み付けを行う．

の類似性判定が可能となる．本節では，提案手法の基礎となるこの手法について説明する．

Mikolov らの提案する単語ベクトルの翻訳手法は，行列変換に基づく．これは，ある単語 x の単語ベクトル $\boldsymbol{x}$ に翻訳行列 $\boldsymbol{W}$ を乗じることによって，その訳語のベクトル表現の近似を得るというものである．翻訳行列 $\boldsymbol{W}$ は，訳語ペアの集合 $\{\boldsymbol{x_i}, \boldsymbol{z_i}\}_{i=1}^{n}$ を用いて，以下の最適化問題を解くことによって得られる．

$$\boldsymbol{W}^{\star} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \sum_{i=1}^{n} \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{z}_i\|^2 + \frac{\lambda}{2}\|\boldsymbol{W}\|^2 \tag{4.1}$$

ここで 2 つめの項は L_2 正則化項である．この正則化項は，Mikolov らの元論文では用いられていない．しかし，正則化項には過学習を防ぐ効果があり，一般的にモデル学習の精度を高めるために有効である．予備実験においてもその効果が確認されたため，本論文では上記の学習方法をベースラインとして議論を進める．

4.1.2 翻訳可能な文脈語ペア

本節では，提案手法の核となるアイデアについて述べる．以降，本研究で扱う意味表現は，3.1 節で述べた Fung らと同様に，カウントベースの単語ベクトル [19] に限定する．カウントベースの単語ベクトルは，各次元が他の単語との共起頻度に対応しているため，前節で述べた翻訳行列の学習に用いる原言語 (翻訳前の言語) のベクトルと目的言語 (翻訳後の言語) のベクトルの次元は，互いに異なるコーパスにおける共起単語，すなわち文脈語にそれぞれ独立に対応している．しかしながら，こ

の 2 言語のベクトルの次元間には，対訳や翻字，発音の類似など，何らかの意味的な対応関係が存在する可能性がある．

たとえば，スペイン語のベクトルから英語のベクトルへの翻訳行列 $\boldsymbol{W}$ を学習することを考える．図 4.2 に例示したように，スペイン語のベクトルの各次元は “*amigo*” や “*comer*”，“*cóctel*” といった，スペイン語の共起語と対応している．一方で，英語のベクトルの各次元は “*eat*” や “*friend*”，“*cocktail*” といった，英語の共起語と対応している．ここで，2 つの共起語 “*amigo*” と “*friend*” は互いの対訳になっている．こうした言語をまたいだ文脈語間，すなわち次元間の対応関係は，式 (4.1) の翻訳行列 $\boldsymbol{W}$ を学習する際にも活用しうるものである．

本研究は，このような 2ヶ国語のベクトルの次元間の対応関係を自動的に見つけるための 2 つの方法を提案する．一つめは，翻訳行列を学習するための学習データを転用する方法である．この学習データは単語の対訳辞書であるため，次元間の対訳関係を見つけるのにそのまま用いることができる．たとえば，図 4.2 では対訳ペア (“*amigo*”, “*friend*”) が学習データに含まれていることから，2 つのベクトルの青い次元間の対応関係を見つける．ただし，ある言語における単語 (例: スペイン語の “*amigo*”) が，他の言語においては複数の訳語を持つ (例: 英語の “*friend*”, “*fan*”, “*supporter*”) 場合は少なくないため，$\mathcal{D}_{train}$ 内の対訳ペアは 1:1 に対応するとは限らない点に留意する．以降，上記の方法で見つけられた次元間の対訳ペアの集合は $\mathcal{D}_{train}$ と記す．

次元間の対応関係を自動的に見つける二つめの方法は，共起語間の表層の近さを用いる方法である．言語はその進化の過程で，他の言語から語や概念を借用することがある．こうした現象によって他言語から取り入れられた言葉は，もとの言語における言葉と似通った (あるいは完全に同じ) 綴りを持つことが多い．たとえば，図 4.2 ではスペイン語の “*cóctel*” と，英語の “*cocktail*” の表層の類似度が高いことから，2 つのベクトルの赤い次元間の対応関係を見つける．本研究では，下記の距離関数を用いて単語の表層の近さを測った．

$$\mathrm{DIST}(r, s) \quad = \quad \frac{Levenshtein(r, s)}{\min(len(r), len(s))}$$

ただし，ここで $Levenshtein(r, s)$ は 2 単語の編集距離を表し，$len(r)$ は単語 r の文字数を表す．また，式 (4.2) の分母は編集距離の正規化のために導入している．本研究では，適当な閾値[1] を設定し，閾値よりも表層の距離が小さい単語のペア (r, s) を全て翻訳可能な文脈語ペアとした．以降，上記の方法で見つけられた次元間の対訳ペアの集合は $\mathcal{D}_{sim}$ と記す．

4.1.3 文脈語ペアの目的関数への導入

前節では，翻訳可能な文脈語ペアの集合 $\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ を見つける方法についてそれぞれ述べた．本節では得られた $\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ を，翻訳行列を学習する際の手がかりとして用いる方法について述べる．

前節と同様に，スペイン語から英語へベクトルの翻訳を行うことを考える．このとき，直観的にはスペイン語のベクトルの “*amigo*” の次元は英語のベクトルの “*friend*” の次元に，スペイン語のベクトルの “*cóctel*” の次元は英語のベクトルの “*cocktail*” の次元と強い相関を持つはずである．この直観に従い，仮にスペイン語のベクトルの j 次元目が “*amigo*” との共起頻度を表し，英語のベクトルの k 次元目が “*friend*” という語との共起頻度を表すとすると，翻訳行列 $\boldsymbol{W}$ の k 行 j 列目 (図 4.2 中 $\boldsymbol{W}$ の青い成分) の値が他の成分と比較して相対的に大きくなるよう補正することで，より正確なベクトルの翻訳が得られるはずである．同様に，スペイン語のベクトルの l 次元目が “*cóctel*”，英語のベクトルの m 次元目が “*cocktail*” とそれぞれ対応している場合，$\boldsymbol{W}$ の m 行 l 列目 (図 4.2 中 $\boldsymbol{W}$ の赤い成分) の値が相対的に大きくなるようにするべきである．

このように対訳関係にある次元のペアにかかる重みを強くするため，式 (4.1) に報酬項を加えた下記の最適化問題を解くことによって，翻訳行列を学習することを提

[1] 今回は 0.5 に固定した．

案する．

$$\begin{aligned} \boldsymbol{W}^{\star} &= \operatorname*{argmin}_{\boldsymbol{W}} \sum_{i=1}^{n} \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{z}_i\|^2 + \frac{\lambda}{2}\|\boldsymbol{W}\|^2 \\ &\quad -\beta_{train} \sum_{(j,k)\in\mathcal{D}_{train}} w_{kj} \\ &\quad -\beta_{sim} \sum_{(l,m)\in\mathcal{D}_{sim}} w_{ml} \end{aligned} \tag{4.2}$$

ここで，上記の式における 3 つめと 4 つめの項は，$\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ に含まれている文脈語ペアに関して，$\boldsymbol{W}$ の対応する成分の値を大きくするように働く報酬項である．この 2 項の係数 β_{train} と β_{sim} はハイパーパラメータであり，$\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ の最適な重みを開発データにより求める．本稿で行った実験では，予備実験の結果から $\beta_{train} = \beta_{sim} = 5.0$ とした．

4.1.4 確率的勾配降下法による学習

前節で提案した最適化問題を，確率的勾配降下法 [35] の一種である Pegasos [36] で解く．τ 番目の学習例 $(\boldsymbol{x}_\tau, \boldsymbol{z}_\tau)$ が与えられた時，翻訳行列 $\boldsymbol{W}$ を次式に基づいて更新する．

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta_\tau \nabla E_\tau(\boldsymbol{W})$$

ここで η_τ は学習率であり，Pegasos では λ をハイパーパラメータとして $\eta_\tau = \frac{1}{\lambda\tau}$ で表現される．また，$\nabla E_\tau(\boldsymbol{W})$ は τ 番目の学習例 $(\boldsymbol{x}_\tau, \boldsymbol{z}_\tau)$ に基いて求められた勾配であり，次式で表すことができる．

$$2(\boldsymbol{W}\boldsymbol{x}_\tau - \boldsymbol{z}_\tau)\boldsymbol{x}_\tau^{\mathrm{T}} - \beta_{train}\boldsymbol{A} - \beta_{sim}\boldsymbol{B} + \lambda\boldsymbol{W}$$

ここで $\boldsymbol{A}$ と $\boldsymbol{B}$ は 4.1.3 節で導入した 2 つの報酬項にそれぞれ対応している．$\boldsymbol{A}$ は $(j,k) \in \mathcal{D}_{train}$ のとき $a_{kj} = 1$ となり，それ以外の場合は $a_{kj} = 0$ となる行列である．$\boldsymbol{B}$ も同様に，$(l,m) \in \mathcal{D}_{sim}$ のとき $b_{ml} = 1$ となり，それ以外の場合は $b_{ml} = 0$ となる行列である．

4.2 訳語選択タスクにおける評価実験

本節では，前節までに述べた提案手法の有用性を評価するため行った，訳語選択タスクでの評価実験について述べる．実験では日本語 (Ja)，中国語 (Zh)，英語 (En)，スペイン語 (Es) の 4 カ国語の単語ベクトルを用いた．この 4 カ国語間の翻訳を通して，教師あり学習の目的関数に組み込まれた 2 種類の文脈語ペアがそれぞれ翻訳精度に与える影響を評価した．

4.2.1 データセット

日本語，中国語，英語，スペイン語の 4ヶ国語の単語ベクトルを作成するため，下記の手順で各言語の Wikipedia ダンプデータ [2] を加工した．まず，wp2txt[3] を用いて XML タグの除去を行う．得られた 4ヶ国語のテキストデータのうち，文中の単語が空白等で句切られない言語である日本語と中国語は形態素解析ソフトウェア MeCab[4] と Stanford Word Segmenter[5] をそれぞれ用いて単語単位に分割した．続いて，語形変化が存在する英語，スペイン語，日本語のデータについてはそれぞれ Stanford POS tagger[6]，Pattern[7]，MeCab を用いて見出し語化し，最終的なテキストコーパスとした．

次に，上記のテキストコーパスを用いてカウントベースの単語ベクトルを作成する．この際，前後 5 単語以内に出現している語を共起語とみなすが，接続詞や助詞といった機能語は共起語から除外する．こうして作られた単語ベクトルは非常に高次元でスパースであるため，全言語について頻出の共起語上位 10,000 語のみを文脈語として残し，10,000 次元のベクトルを得た．また，単語の出現頻度の差異を吸収するため，全てのベクトルの各次元の値を PPMI [16] に変換した後，正規化を行ってスケールを統一した．

[2] http://dumps.wikimedia.org/から入手．バージョンはそれぞれ Ja: 2014/11/04，Zh: 2014/12/04，En: 2014/10/08，Es: 2015/02/07 である．
[3] https://github.com/yohasebe/wp2txt/
[4] http://taku910.github.io/mecab/
[5] http://nlp.stanford.edu/software/segmenter.shtml
[6] http://nlp.stanford.edu/software/tagger.shtml
[7] http://www.clips.ua.ac.be/pages/pattern

表 4.1: 語彙，学習データ，$\mathcal{D}_{train}$ 及び $\mathcal{D}_{sim}$ の件数

	語彙 (原言語)	語彙 (目的言語)	学習データ	$\mathcal{D}_{train}$	$\mathcal{D}_{sim}$
(Ja → Zh)	10000	10641	42037	9552	3189
(Zh → Ja)	10000	20356	69619	9552	3189
(Ja → En)	10000	15060	50300	18296	2234
(En → Ja)	10000	28275	84451	18296	2234
(Zh → En)	10000	15784	41144	9292	3551
(En → Zh)	10000	14770	38854	9292	3551
(En → Es)	10000	10247	34034	15567	12764
(Es → En)	10000	19917	48125	15567	12764

最後に，Open Multilingual Wordnet[8] を用いて，学習データと評価データとして用いる対訳辞書を作成した．本研究では，提案手法を応用して対訳辞書の拡張を行う状況を想定し，頻出語であるほど既に辞書に含まれている可能性は高いことをふまえて，得られた対訳辞書を学習データと評価データに分割した．具体的には，原言語のコーパス内での出現頻度が高い順に全ての単語を並べ，上位 10,000 位までの単語を学習データとし，続く 10,001 位から 11,000 位までの単語を開発データ，11,001 位から 12,000 位までの単語を評価データとした．このとき，多義語が存在した場合はそれぞれを独立した対訳ペアとして学習データに含める．そのため，表 4.1 に示すように，学習データ内の原言語の語彙数は全て 10,000 語で固定されているのに対し，目的言語の語彙数と学習データの数は 10,000 件よりも大きくなっている．また，原言語と目的言語の単語が 1:1 に対応していないため，$\mathcal{D}_{train}$ の件数も 10,000 件を超える場合が存在する．表 4.1 には言語対 (Ja → Es)，(Es → Ja)，(Zh → Es)，(Zh → Es) が含まれていないが，これは Open Multilingual Wordnet に含まれる対訳データが少なく，12,000 位までの単語について対訳辞書を作成できなかったためである．そのため，この 4 種の言語対を除いた計 8 つの言語対について，上記の対訳辞書を作成した．

[8] http://compling.hss.ntu.edu.sg/omw/

4.2.2 比較手法

提案手法の有用性を評価するため，以下に述べる 3 種類の比較手法を実装した．

ベースライン 式 (4.1) に基づき，翻訳行列を学習する．このとき，提案手法と同様にカウントベースの単語ベクトルを翻訳に用いる．ベースラインと提案手法を比較することで，提案手法で用いる文脈語ペアの集合 $\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ が単語ベクトルの翻訳精度に与える影響を明らかにする．

CBOW 式 (4.1) に基づき，翻訳行列を学習する．このとき，提案手法とは異なる，ニューラルネットワークによって学習された Continuous bag-of-words モデル (CBOW) [3] の単語ベクトルを翻訳に用いる．この手法と前項のベースラインを比較することで，ベクトル表現手法の差異が翻訳精度に与える影響を明らかにする．ただし，4.2.1 節で述べたとおり，提案手法及び前項のベースラインで用いたカウントベースのベクトルは出現頻度上位 10,000 位までの単語のみ共起語として考慮している．一方で，CBOW モデルのベクトルはコーパス内に出現した全ての単語を共起語として考慮している．本研究では，word2vec[9] を用いて 4 カ国語の単語ベクトルを得た [10]．ベクトルの最適な次元数については，原言語のベクトルの次元数が目的言語のベクトルの次元数の 2-4 倍になるように設定した際の翻訳精度が最も良いという報告 [3] がある．本研究ではこれをふまえ，原言語のベクトルを m 次元 ($m = 100, 200, 300$) に，目的言語のベクトルを n 次元 ($n = 2m, 3m, 4m$) に設定し，開発データを用いて最適な m と n の組み合わせを探索した．

Direct Mapping 学習された翻訳行列による翻訳が有用であることを確認するため，翻訳行列を用いずに単語ベクトルを翻訳する手法を実装した．この手法では，提案手法と同様にカウントベースの単語ベクトルを翻訳に用いる．4.1.2 節で述べた学習

[9] https://code.google.com/p/word2vec/

[10] "*a*" や "*the*" といった頻出語の影響を抑制するため，サブサンプリングの閾値パラメタを 1e-3 に設定した．

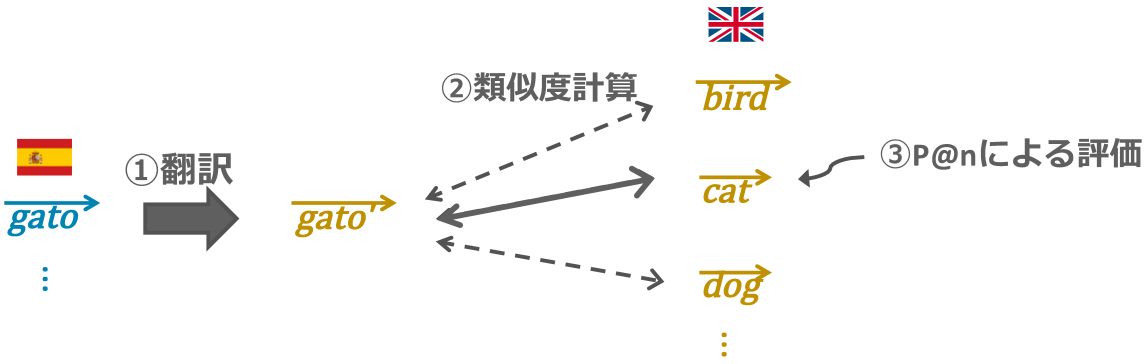

図 4.3: 単語ベクトルの翻訳の評価手順

データから得られる $\mathcal{D}_{train}$ を用いて，原言語のベクトルの次元を，目的言語のベクトルの対応する次元に直接変換する [7]．この際，原言語の次元と目的言語の次元が 1:1 に対応しない場合も存在する．その場合，各訳語に対して目的言語内での出現頻度に基づき，Reciprocal Rank(順位の逆数) による重み付け和 [37] を用いた．

4.2.3 評価手法

4 節で述べたように，学習した翻訳行列がどれほど正確に言語間の翻訳を行うことができるかを，ある単語のベクトル $\boldsymbol{x}$ を翻訳した結果 $\boldsymbol{W}\boldsymbol{x}$ が，どれほど対訳語のベクトル $\boldsymbol{z}$ に近くなるかで評価する [3]．具体的には，図 4.3 に示す以下 3 つの手順で評価を行った．(1) まず原言語から目的言語へベクトルの翻訳を行い，(2) 次に得られた翻訳後のベクトルを目的言語の全ての単語ベクトルと比較し，コサイン類似度が高いもの上位 n 語 ($n = 1, 5$) を選択する．(3) この n 語の中に，正解の訳語が含まれているかどうかを精度 (P@n) で評価する．

表 4.2: 実験結果: 単語ベクトルの翻訳精度

Testset	ベースライン		CBOW		Direct Mapping		提案手法 sim 無し		提案手法	
	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5
Ja → Zh	0.6%	1.6%	5.4%	13.8%	9.3%	22.2%	11.1%	26.2%	**15.5%**	**34.0%**
Zh → Ja	0.3%	1.2%	2.9%	11.3%	11.6%	26.9%	7.8%	21.6%	**13.1%**	**27.9%**
Ja → En	0.2%	1.0%	6.5%	19.1%	22.3%	37.4%	32.3%	51.0%	**32.5%**	**51.9%**
En → Ja	0.3%	1.1%	4.9%	13.3%	5.4%	13.9%	18.5%	36.4%	**19.3%**	**37.1%**
Zh → En	0.2%	0.8%	3.4%	11.8%	**23.3%**	40.6%	22.3%	40.4%	23.1%	**42.0%**
En → Zh	0.2%	1.1%	5.1%	13.7%	4.5%	11.8%	9.1%	22.1%	**9.5%**	**23.0%**
En → Es	0.2%	1.0%	7.1%	18.9%	11.9%	26.1%	28.7%	45.7%	**31.3%**	**49.6%**
Es → En	0.0%	0.6%	7.5%	22.0%	45.7%	61.1%	46.6%	62.4%	**54.7%**	**67.6%**

4.2.4 実験結果

4.2.4.1 訳語選択タスクにおける翻訳精度

表 5.2 に各言語対における単語ベクトルの翻訳精度を示す．**提案手法**は**ベースライン**と比較して，ほぼ全ての言語対において翻訳精度が大きく向上していることがわかる．**CBOW** はカウントベースの単語ベクトルのかわりに CBOW モデルのベクトルを用いることで，**ベースライン**よりも良い精度を達成しているが，**提案手法**と比較すると僅かな改善に留まっている．

提案手法 sim 無し は文脈語ペアを学習データのみから探し，文脈語間の表層に関する情報は利用しない．すなわち，$\beta_{sim} = 0$ と設定することで，文脈語ペアの集合 $\mathcal{D}_{sim}$ の影響力を除外する．この手法は特に (Ja, En)，(En, Es) の各組み合わせについて **Direct Mapping** よりも大きく優れており，翻訳行列を学習する手法の有用性を示している．ただし，(Zh → Ja) と (Zh → En) については **Direct Mapping** の方が**提案手法 sim 無し** よりも高精度となっている．また，逆方向の (Ja → Zh) と (En → Zh) についても，**Direct Mapping** から**提案手法 sim 無し** への改善は僅かである．この現象に関して，表 4.1 を参照すると，(Ja, Zh) や (Zh, En) の各組み合わせと比較して，(Ja, En) や (En, Es) の各組み合わせは $\mathcal{D}_{train}$ の 1.6 倍から 2 倍程度多いことが読み取れる．すなわち，特に $\mathcal{D}_{train}$ として活用できる対訳辞書が十分存在している状況であれば，辞書から得られる情報 $\mathcal{D}_{train}$ が翻訳精度の向上に大きく貢献するといえる．

また，全ての言語対について，**提案手法**は**提案手法 sim 無し** の結果を上回っており，

表 4.3: $\mathcal{D}_{sim}$ の有無による正解例の数の変化．全ての言語対について，評価データは 1,000 例ずつ存在する

	不正解 → 正解	**正解 → 不正解**
(Ja → Zh)	53	9
(Zh → Ja)	60	7
(Ja → En)	9	7
(En → Ja)	11	3
(Zh → En)	14	6
(En → Zh)	8	4
(En → Es)	52	26
(Es → En)	108	27

その差は特に (Ja, Zh)，(En, Es) といった言語対において顕著である．日本語と中国語，英語とスペイン語はそれぞれ文字体系が近く，語彙の交換が頻繁に存在する．こうした表層的に類似している言語対においては，表層の類似度から得られる情報 $\mathcal{D}_{sim}$ が翻訳精度の向上に大きく貢献すると考えられる．

文脈語の表層の類似度が翻訳行列に与える影響を分析するため，表 4.3 に $\mathcal{D}_{sim}$ の有無による正解例の数の変化を示す．(Ja, Zh)，(En, Es) といった言語対においては $\mathcal{D}_{sim}$ が翻訳精度の向上に大きく貢献することは前述したが，この 2 種類の言語対における $\mathcal{D}_{sim}$ の役割は少し異なっていることが表 4.3 から読み取れる．すなわち，(Ja, Zh) 間の翻訳では，不正解→正解となる例が増える一方で正解→不正解の例はあまり増加しておらず，$\mathcal{D}_{sim}$ は確実に正しい方向に，慎重にバイアスをかけているといえる．しかし，(En, Es) 間の翻訳では不正解→正解の増加とともに正解→不正解も大きく増加していることから，(En, Es) 間の翻訳行列を学習する場合，$\mathcal{D}_{sim}$ は小さい間違いは無視して，よりアグレッシブにバイアスをかける働きをしていると考えられる．

4.2.4.2 学習データの規模による影響

図 4.4 - 図 4.7 は，各言語対について，学習データの大きさを変化させた時の翻訳精度 (Precision@1) の変化をプロットしたものである．ただし，学習データは翻訳行列の学習だけではなく，**Direct Mapping** における次元の変換と，提案手法の報酬項にも用いられている．図から，**Direct Mapping** と**提案手法** $_{\textbf{sim 無し}}$ を比較すると，後者は学習データが少ない時には有効でないことがわかる．これは後者が学習する翻訳行列のパラメータ数と比較して，学習データが少ないことから起こる過学習の影響であると推測される．しかし，**Direct Mapping** では学習データを増加させても精度向上に寄与しにくいことから，十分学習データが存在している場面では翻訳行列の学習を行った方が良いと言える．

また，表層が似ている言語対 (Ja, Zh) と (En, Es) の各組み合わせについては，特に学習データが多い時に**提案手法**の精度が**提案手法** $_{\textbf{sim 無し}}$ を大きく上回っていることがわかる．このことから，文脈語の類似度による手がかり $\mathcal{D}_{sim}$ が翻訳行列の学習を正しい方向に誘導する役割を果たしているといえる．

4.2.4.3 各翻訳手法の出力例

表 5.4，表 4.5，表 4.6 に (Zh → Ja)，(En → Ja)，(Es → En) の出力結果を一部示す．(太字は正解の翻訳語を表す)

3 つの言語対全てにおいて，**ベースライン**は設問に依存せず，どれも似通った翻訳候補を出力している．この現象は Hubness 問題 [28] として報告されており，線形回帰を写像関数として用いることによって，説明変数が目的変数よりも原点に近い位置に写像される傾向があることが原因である [38] ことが知られている．**CBOW** の出力を**ベースライン**と比較すると，正解の関連語が多く存在していることが分かる．たとえば，(En → Ja) における sorceress (魔法使い/魔女) の訳語候補には「化け物」や「生霊」が，また xenon (キセノン) の訳語候補には「微粒子」が存在している．こうした例から，CBOW モデルのベクトルを用いることにより，正解の関連語が出力されやすくはなるものの，正解は出力できていないことがわかる．

表 4.4: (Zh → Ja) の翻訳例

ベースライン	CBOW	Direct Mapping	提案手法 sim 無し	提案手法
校験位 → パリティ /パリティビット/パリティー				
1 違う	考慮	プリミティブ	縫い目	**パリティビット**
2 動く	相	クライアント	言葉	**パリティ**
3 持つ	把握	結び目	**パリティ**	縫い目
4 周囲	規準	用語	正しい	クライアント
5 十分	正しい	ディレクトリ	一見	言葉
焼瓶 → フラスコ				
1 周囲	卵殻	空気	**フラスコ**	**フラスコ**
2 軽い	微粒子	**フラスコ**	空気	空気
3 動く	**フラスコ**	溶解	磨る	活栓
4 小さい	小片	滴	寒天	磨る
5 持つ	薄片	寒天	天井	寒天
小農 → 小作農				
1 周囲	変質	**小作農**	困窮	困窮
2 見る	溜め込む	困窮	把握	保護
3 現れる	無駄	配慮	好ましい	把握
4 かなり	不潔	把握	配慮	**小作農**
5 動く	腐敗	作物	保護	配分

表 4.5: (En → Ja) の翻訳例

ベースライン	CBOW	Direct Mapping	提案手法 sim 無し	提案手法
sorceress → 魔法使い/魔女				
1 思う	化け物	魔物	魔物	**魔法使い**
2 恐ろしい	生霊	恐ろしい	**魔法使い**	魔物
3 捨てる	狂気	**魔法使い**	呪い	**魔女**
4 怒る	暴君	思う	**魔女**	呪い
5 邪魔	人殺し	呪い	怪物	エルフ
xenon → キセノン				
1 逆	微粒子	気体	放射	**キセノン**
2 実際	蒸散	放射	**キセノン**	放射
3 ある程度	変化	粒子	気体	気体
4 弱い	吸い込む	特性	粒子	粒子
5 小さい	縮む	小さい	重力	重力
abduct → 連れ去る				
1 思う	追い払う	逃げる	**連れ去る**	襲う
2 捨てる	騙す	逃げ出す	襲う	**連れ去る**
3 恐ろしい	庇う	襲う	殺害	殺害
4 逃げる	責める	思う	殺し	殺し
5 怒る	見捨てる	恐ろしい	逃げ出す	逃げ出す

表 4.6: (Es → En) の翻訳例

ベースライン CBOW		Direct Mapping	提案手法 sim 無し	提案手法
clericalismo → clericalism				
1 call	attitude	struggle	attitude	**clericalism**
2 describe	banality	attitude	negativity	attitude
3 intend	self-consciousness	turn	struggle	struggle
4 make	fatalistic	**clericalism**	**clericalism**	negativity
5 ignore	egoism	espouse	fatalistic	chauvinism
papio → baboon				
1 call	crab	elephant	cow	**baboon**
2 turn	dwarf	antelope	ichthyosaur	crocodile
3 make	elephant	cow	parcel	dwarf
4 describe	crocodile	parcel	elephant	elephant
5 intend	hairy	bovid	crocodile	obscure
yambo → iamb				
1 call	fairy	call	**iamb**	interrogative
2 turn	pluck	turn	caesura	stanza
3 describe	stick	stanza	interrogative	**iamb**
4 make	dark	set	gesture	caesura
5 intend	croak	**iamb**	stanza	gesture

また，**Direct Mapping** と**提案手法** sim 無し の出力は多くが共通している．この 2 手法は，いずれも学習データに含まれる対訳例をベクトルの翻訳に用いている．前者はベクトルの次元の変換に直接対訳例を用いるのに対し，後者は $\mathcal{D}_{train}$ として学習の目的関数の報酬項として取り込まれている．両手法で翻訳に利用できる手がかりは完全に等しいため，出力結果や翻訳精度が類似していると考えられる．

提案手法では，**提案手法** sim 無し で用いる対訳例の情報に加えて，さらに文脈語の表層の類似度を手がかりとして活用している．この手がかりを $\mathcal{D}_{sim}$ として目的関数に加える事で，**提案手法**はさらに正確な翻訳を出力できている．たとえば，(Zh → Ja) における (パリティビット) の訳語候補を比較すると，**提案手法** sim 無し では翻訳候補の 3 位に正解「パリティ」が出現しているが，**提案手法**では 1 位，2 位がともに正解 (「パリティビット」「パリティ」) となっている．(Zn → Ja) における小農（小作農）や (Es → En) の yambo (iamb) の翻訳のように，提案手法で正解を出力できなくなった例も存在するが，多くの正解は翻訳候補の上位として出力されていることが分かる．

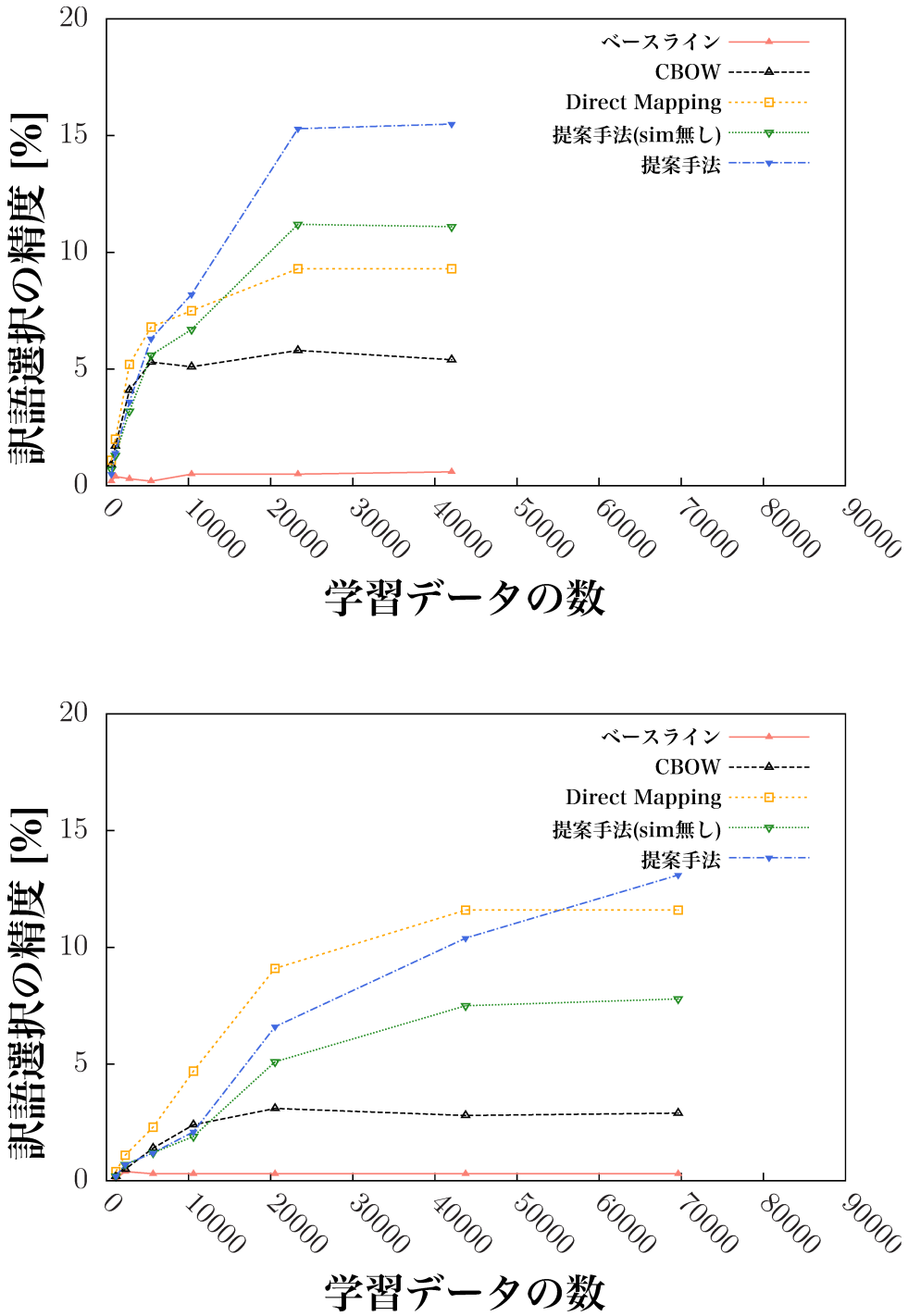

図 4.4: 学習データの規模が翻訳精度に与える影響 (上：Ja → Zh, 下：Zh → Ja)

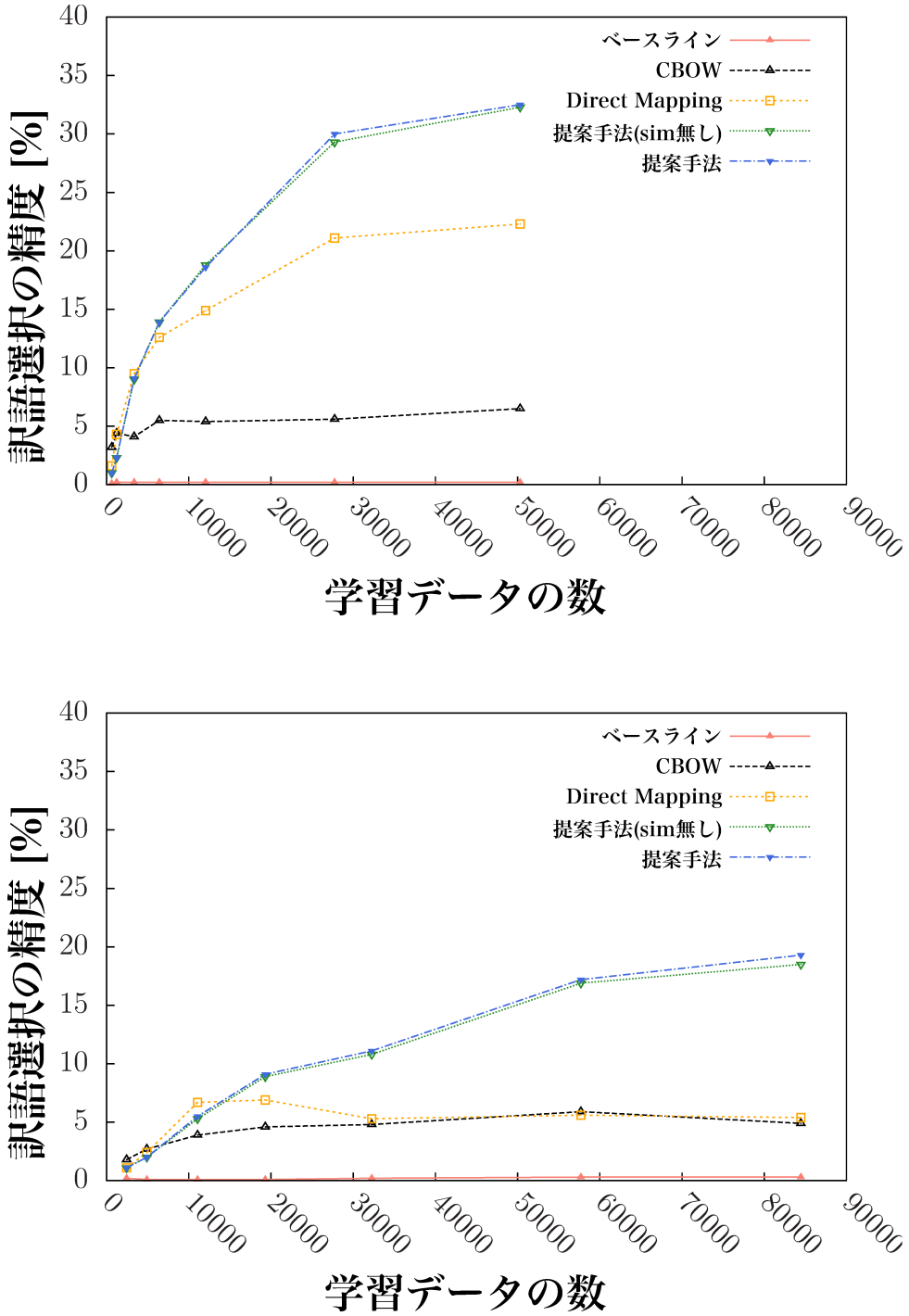

図 4.5: 学習データの規模が翻訳精度に与える影響 (上：Ja → En, 下：En → Ja)

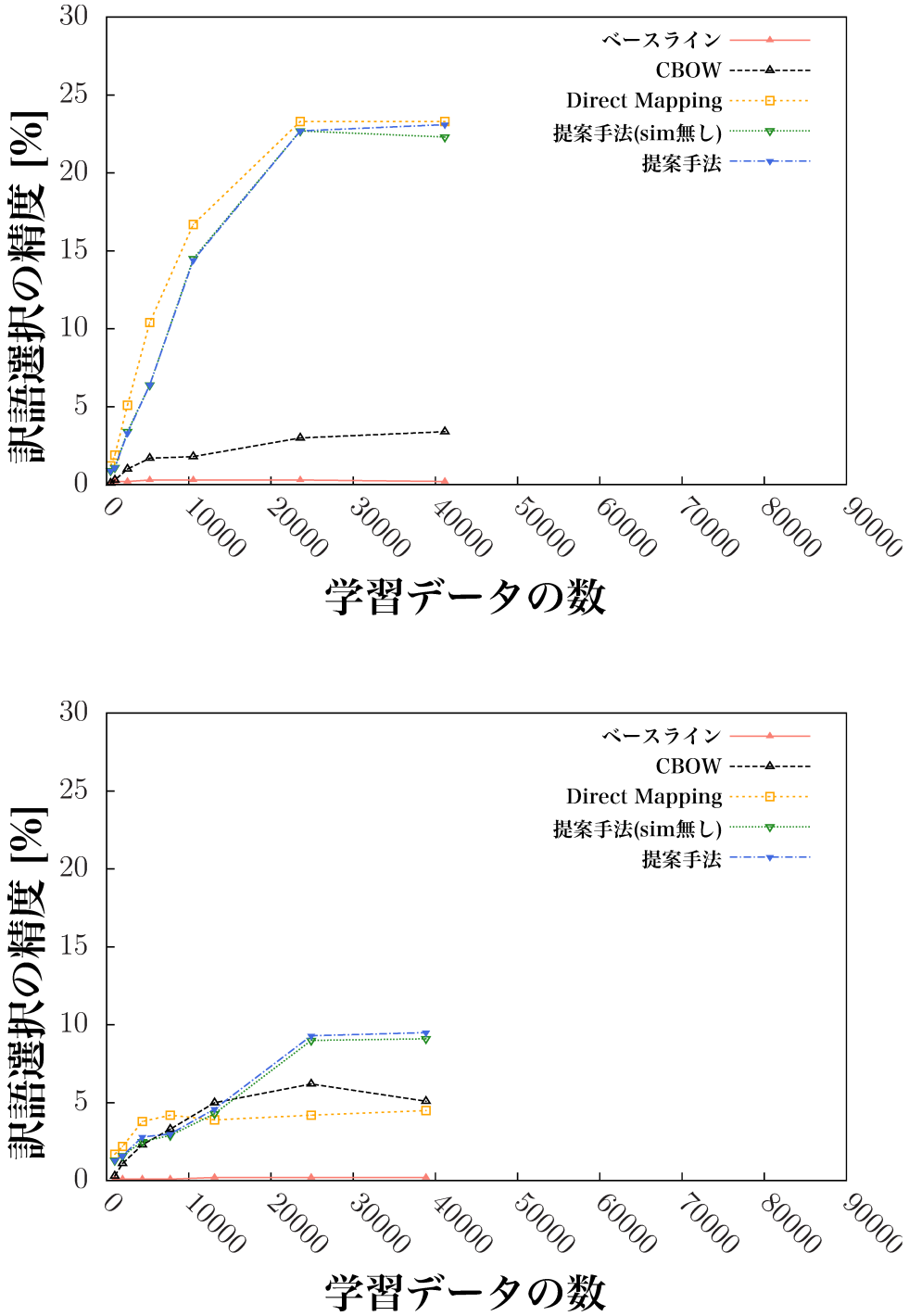

図 4.6: 学習データの規模が翻訳精度に与える影響 (上：Zh → En, 下：En → Zh)

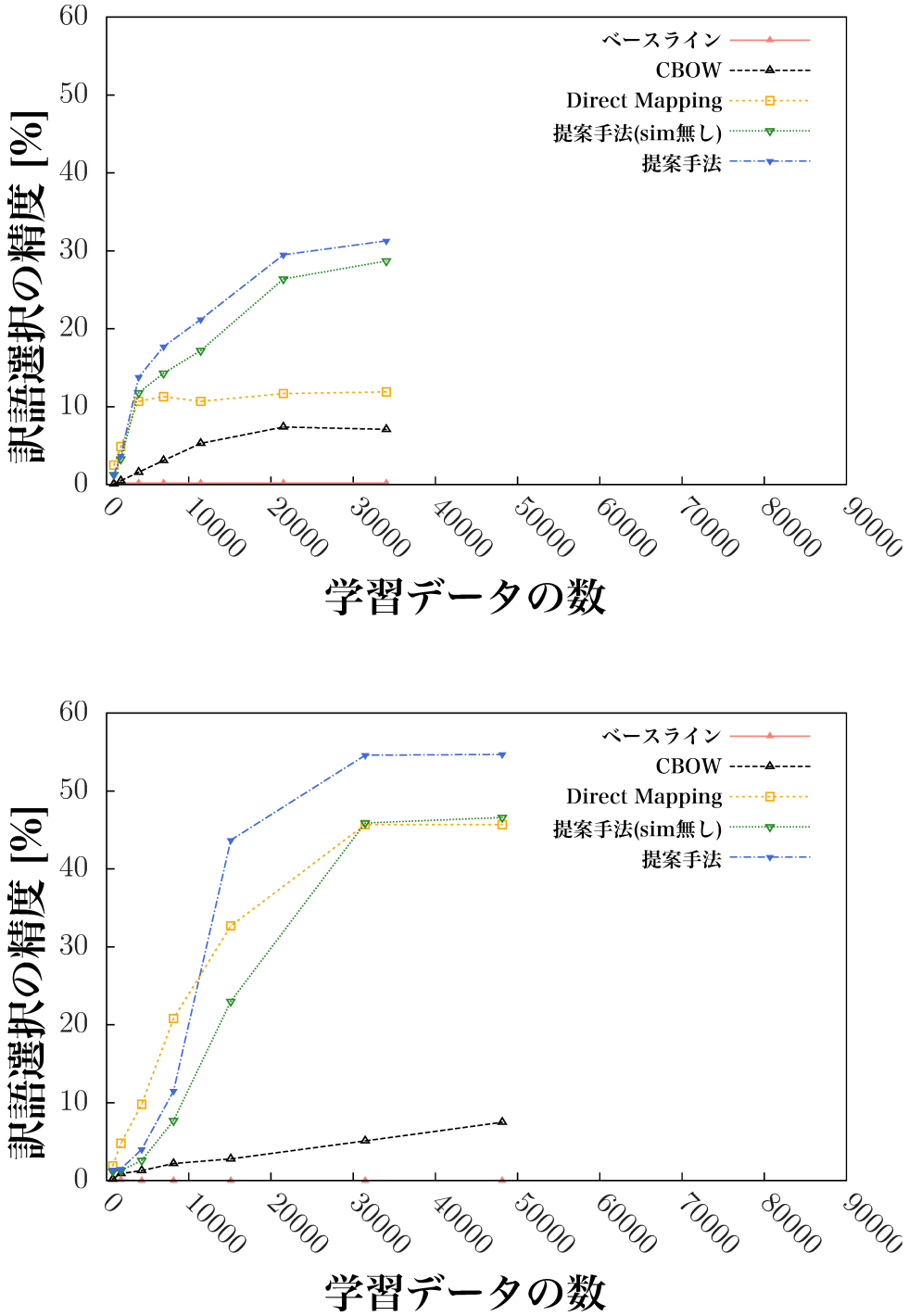

図 4.7: 学習データの規模が翻訳精度に与える影響 (上：En → Es, 下：Es → En)

第5章　統計的機械翻訳における未知語の翻訳

前章で提案した単語ベクトルの翻訳手法を用いることで，言葉の意味を言語横断的に比較することが可能となった．本章では，言語をまたいだ単語間の意味の類似性判定を活用することで，機械翻訳の高精度化を実現する方法を述べる．以降，5.1節ではベクトル翻訳を活用することで未知語の翻訳を行う手法を述べる．続く5.2節では提案手法の有用性を評価するために行った，機械翻訳における未知語翻訳タスクにおける評価実験について述べる．

5.1　提案手法：単語ベクトルを用いた未知語の翻訳

翻訳対象の文とは異なるドメインのコーパスから学習されたモデルを使う際に最も難しい問題となるのは，学習データに存在しない未知の語や表現，統語構造の翻訳をいかに行うかという点である [31]．本節では以降，これらの中でも特に顕著な問題となる未知語の翻訳に取り組む．未知語，すなわち翻訳モデルに存在しない語に訳語を与えることを考えた場合，最も単純なアプローチは，ドメインごとに対訳辞書を作成する方法である．しかし，ドメインごとに辞書を整備するのは容易ではない上，文書のドメインは翻訳前に未知であることも多い [32] ため，より簡便な手法が望まれる．これに対し，ドメインに特化しない汎用的な対訳辞書と，対象ドメインの単言語コーパスであれば，比較的容易に利用できる．前者については，たとえ

ば日英であればEDICT[1]や英辞郎[2]，その他の言語対でもOpen Multilingual Wordnet[3]といった容易に入手できるデータセットが存在する．後者についても対象ドメインの文書を機械翻訳システム利用者から入手するか，ウェブからクロールすることにより容易に獲得可能である．こうした状況をふまえ，本節で提案する手法は小規模な汎用対訳辞書と，対象ドメインの単言語コーパスのみを用いて未知語の翻訳を行う．

提案手法は図5.1に示すように，(1) まず単言語コーパスから原言語と目的言語の単語ベクトルをそれぞれ学習し，(2) 対訳辞書を用いて言語間のベクトル翻訳を学習した後，(3) それらを機械翻訳システムに導入することで，未知語に対して適切な訳語を与えることを可能とする．(1)，(2) については3.1節で述べた手続きと同様に行うため，以降は (3) の手順について詳細に述べる．

ステップ1　まず，機械翻訳システムに未知語 x を含む文が入力された場合，事前に原言語の単言語コーパスから学習した x の単語ベクトル $\boldsymbol{x}$ に翻訳行列 $\boldsymbol{W}$ を乗じることで，目的言語のベクトル $\boldsymbol{Wx}$ へと翻訳する．

ステップ2　次に，$\boldsymbol{Wx}$ と目的言語の全ての単語ベクトルとの比較を行い，コサイン類似度が高いもの上位10語を x の訳語候補とする．各訳語候補に対して，$\boldsymbol{x}$ と上位10語のベクトルとのコサイン類似度の和が1になるように正規化することにより，翻訳確率を与える．これらの確率は，以降翻訳モデルにおけるバックオフモデル [8] として扱う．すなわち，上記の手続きで得られた (未知語 x，x の訳語候補，翻訳確率) の組の集合は，対訳コーパスから得られた翻訳モデルに訳語が存在しない語を翻訳する際にのみ参照される．

ステップ3　最後に，事前に学習した翻訳モデルと言語モデル，及び上記で作成したバックオフモデルを統計的機械翻訳システムのデコーダに入力し，翻訳文候補の生成，選択を行う．ステップ2で生成されたバックオフモデルは誤りを含む可能性

[1] http://www.edrdg.org/
[2] http://www.eijiro.jp/
[3] http://compling.hss.ntu.edu.sg/omw/

表 5.1: 実験に使用したコーパスの規模

コーパス	日本語	英語
京都関連記事 (対訳)	30MB (440k文)	31MB (440k文)
料理レシピ (対訳)	13MB (150k文)	11MB (150k文)
Wikipedia(単言語)	4.4GB	16GB

が考えられるが，その誤りはデコードの段階で言語モデルが「目的言語の文としてもっともらしい」出力文を選択することにより，ある程度補正されることが期待される．

5.2 未知語翻訳タスクにおける評価実験

3.2 節で述べたように，現在の統計的機械翻訳技術には，学習データのドメインとテストデータのドメインが異なる場合に翻訳精度が著しく落ちる [31] という問題が存在する．また，ドメインが異なることにより特に顕著な問題となるのは，未知語の出現頻度が大きく増加することである．

これらの点をふまえ，本節で述べる評価実験においても，学習に用いた対訳コーパスと異なるドメインの文に対して翻訳を行う．この際，未知語 (すなわち，対訳コーパスから学習された翻訳モデルでは翻訳できなかった語) に対しては前節で導入した提案手法を用いることにより，翻訳品質の向上を目指す．以降，本節では実際に行った評価実験について詳細に述べる．

5.2.1 データセット

本実験では，まず機械翻訳システムの学習／評価に用いるデータとして，京都翻訳フリータスク [39] で公開されている Wikipedia の京都関連記事の日英対訳コーパスと，ユーザ投稿型レシピサイトである Cookpad[4] のレシピの日英対訳コーパスを用意した．前者のドメインには日本の歴史上の人物や寺社仏閣に関する語彙が多く，

[4]http://cookpad.com/

だ／である調で記述されている．一方，後者のドメインには調理器具や食材に関する語彙が多く，です／ます調や砕けた表現が頻出する．このうち，料理レシピの対訳コーパスは学習データ (144,178 文対) とテストデータ (10,000 文対) に分割した．

次に，ベクトル表現の学習に用いる単言語コーパスとして，翻訳対象ドメインと同一のものと，ドメイン特化していないものの 2 種類を用意した．このうち，前者は Cookpad の対訳コーパスの学習データ (144,178 文対) を日本語と英語に分割し，それぞれを単言語コーパスとして用いた．後者は 4.2 節と同様に，Wikipedia のダンプデータ [5] を単言語コーパスとして用いた．以上の対訳コーパス，および単言語コーパスのデータサイズは表 5.1 に示す．

最後に，ベクトル表現間の翻訳行列を学習するための学習データとして，4.2 節と同様に Open Multilingual Wordnet[6] を用意した．

5.2.2 評価手法

実験では 4 種類の手法を実装し，下記の手順に従って評価を行った．

ベースライン　オープンソースの機械翻訳ツールキットである Moses[7] を用いてベースラインシステムを構築した．ベースラインシステムにおいて，目的言語の言語モデルは京都関連記事のデータと料理レシピのデータを用いて行い，2 カ国語間の翻訳モデルは京都関連記事の対訳コーパスのみを用いて行った．これらの言語モデルと翻訳モデルを用いて，テストデータの翻訳品質を機械翻訳の自動評価指標である BLEU[40] により評価した．

提案手法　まず，5.1 節で述べた提案手法を用いて，前項**ベースライン**の評価において出現した未知語に対してバックオフモデルを構築した．ただし，バックオフモデルはドメイン特化していない Wikipedia コーパスから作成したもの **(general)** と，翻

[5] http://dumps.wikimedia.org/から入手．日本語は 2014/11/04 版，英語は 2014/10/08 版．
[6] http://compling.hss.ntu.edu.sg/omw/
[7] http://www.statmt.org/moses/

表 5.2: 各手法の翻訳品質 (BLEU)

手法	全データ 日英	全データ 英日	未知語あり 日英	未知語あり 英日
ベースライン	5.58	3.37	5.36	3.16
提案手法 (general)	6.09	3.50	5.91	3.46
提案手法 (in-domain)	**7.29**	**3.67**	**7.23**	**3.89**
対訳コーパス	20.88	16.69	20.72	17.01

表 5.3: テストデータ (10,000 件) が含む未知語の統計量

	日英	英日
未知語 (出現回数)	21,218	4,639
未知語 (語彙数)	3,464	1,613
未知語を含む文	8,742	3,636

訳対象ドメインである料理レシピコーパスから作成したもの **(in-domain)** の 2 種類を用意した．次に，作成した 2 種類のバックオフモデルを用いて，再度**ベースライン**と同様のテストデータで評価を行った．

対訳コーパス　前項で述べた 2 種類の提案手法の比較手法として，料理レシピの対訳コーパスを用いて翻訳モデルの学習を行った．ただし，本研究では対象ドメインの対訳コーパスは利用できないと仮定しているため，この手法により得られる翻訳品質は本論文における upper bound となる．

5.2.3 実験結果

実験結果を表 5.2 に示す．表 5.2 の**全データ**は料理レシピコーパスのテストデータ 10,000 件を用いて評価を行ったものであり，**未知語あり**は 10,000 件のうち，未知語を含む文のみを用いて評価を行ったものである．テストデータに含まれる未知語の統計量は表 5.3 に示すとおりである．

表 5.2 から，提案手法を用いて未知語の翻訳を行うことにより，翻訳品質が向上することがわかる．さらに，**提案手法 (in-domain)** の方が**提案手法 (general)** よりも

表 5.4: レシピドメインにおける英日翻訳の出力例．太字は未知語とその訳語．

入力文	when the chocolate has melted completely , add the brandy or rum to mix in .
ベースライン	チョコレートが完全に溶けて、**rum** の **brandy** を混ぜたものである。
提案手法 (general)	チョコレートが完全に溶けて、**砂糖** を入れて、**砂糖** を混ぜたものである。
提案手法 (in-domain)	チョコレートが完全に溶けて、**ラム 酒** を加え、混ぜたものである。
対訳コーパス	チョコが溶けたら、**洋酒** を入れて混ぜる。
参照訳	チョコが全てとけたら **洋酒** を加えて混ぜます。
入力文	you can also use broccoli .
ベースライン	**broccoli** も使われることができる。
提案手法 (general)	**野菜** も使われることができる。
提案手法 (in-domain)	**ブロッコリー** も使われることができる。
対訳コーパス	**ブロッコリー** を使ってもＯＫです。
参照訳	**ブロッコリー** バージョンも美味。
入力文	preheat the oven to 200 ° c .
ベースライン	オーブンは 200 度に **preheat** 。
提案手法 (general)	オーブンは 200 度に **加熱** 。
提案手法 (in-domain)	オーブンは 200 度に **予熱** 。
対訳コーパス	オーブンを 200 ℃に **予熱** しておきます。
参照訳	オーブンを 200 度に **予熱** しておく。

良い翻訳品質を達成していることから，ベクトル翻訳に用いる単語ベクトルの作成には，一般的なドメインのコーパスよりもテストデータと同じドメインのコーパスを利用した方が良いという知見が得られた．これは，ベクトルを構成する単語の共起頻度に関する情報が，対象ドメインと強い相関を持つためであると考えられる．また，**対訳コーパス**と比較すると，他の手法の翻訳品質が非常に低くなっているが，これは京都関連記事ドメインと料理レシピドメインにおける語彙や言葉遣い等が著しく異なり，細かい表現の訳し分けが非常に困難となるからである．

さらに考察を深めるため，英日翻訳における出力例を表 5.4 に示す．表から，**ベースライン**では翻訳できなかった “rum” や “brandy”，“broccoli”，“preheat” といった未知語が提案手法では翻訳できていることが読み取れる．2 つの提案手法の差異に着目すると，**提案手法 (general)** では多くの翻訳が関連語や上位語となっている（“rum” → “砂糖” や，“broccoli” → “野菜” など）一方で，**提案手法 (in-domain)** ではより正解に近い訳語（“rum” → “ラム酒”，“broccoli” → “ブロッコリー” など）を出力できていることがわかる．このことからも，単語ベクトルを学習する際のコーパスはテストデータと同じドメインのものを利用する方が有効であると言える．

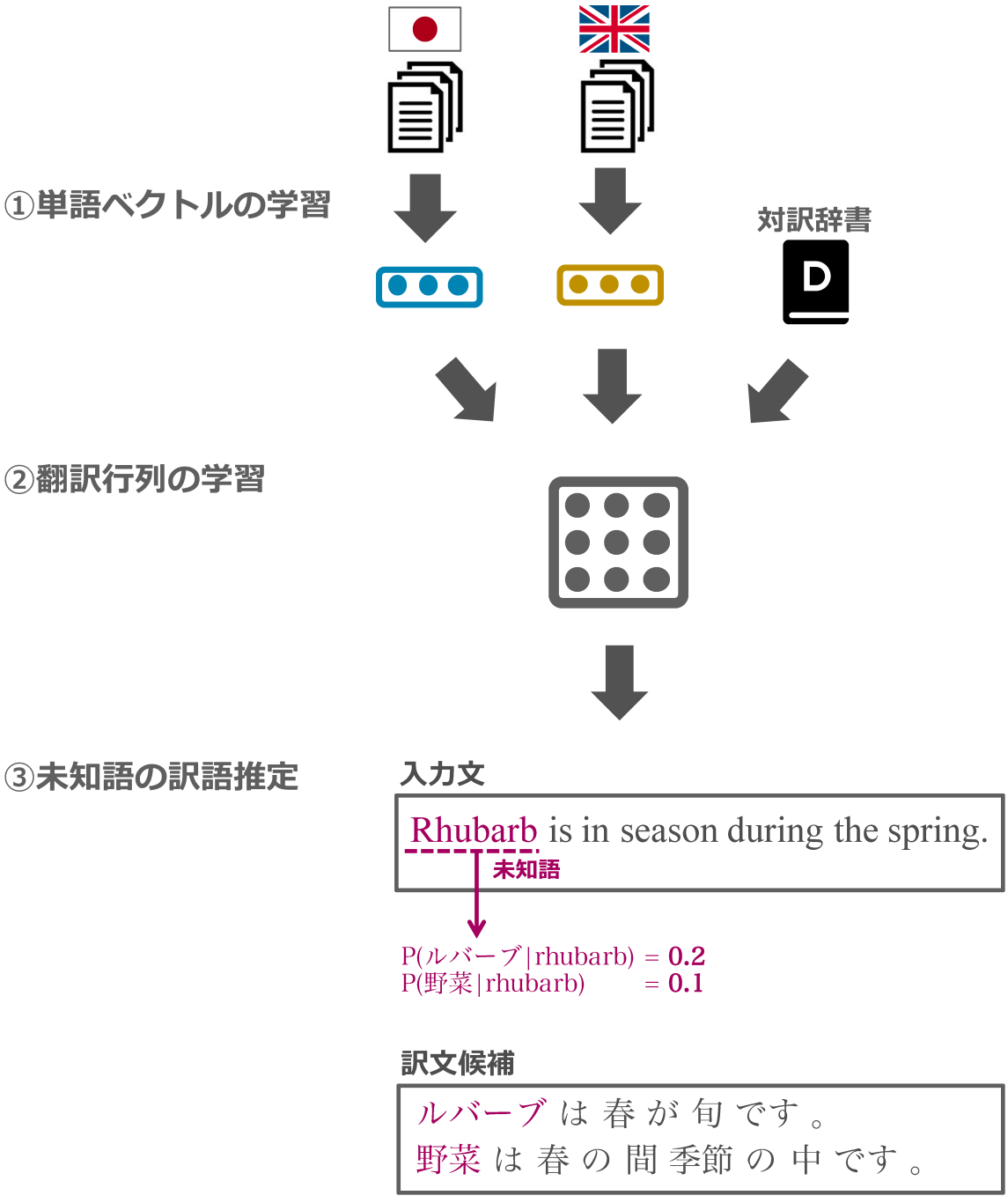

図 5.1: 提案手法：ベクトル翻訳を用いた未知語の翻訳

第6章　おわりに

本研究は機械翻訳や言語横断情報検索といった多言語処理技術の高度化を目的とし，(1) ことばの意味表現を言語横断的に扱えるようにすることと，(2) それを機械翻訳技術の枠組み内で扱えるようにすることの 2 点に取り組んだ．

(1) については 4 章で述べた．まず，我々は (i) 意味表現の翻訳を得るための学習データとして用いられる対訳辞書と，(ii) 言語対に存在する表層の類似性という 2 種類の手がかりから得られる情報を考慮可能な目的関数を新たに設計した．続いて，我々は 4 カ国語間の訳語選択タスクでの実験を行い，この提案手法が日本語，中国語，英語，スペイン語における意味表現の翻訳精度を大きく向上させることを示した．

(2) については 5 章で述べた．まず，我々は上記で提案した意味表現の翻訳手法を活用し，機械翻訳システムにおける未知語に対しても訳語を与えうるバックオフモデルを構築する手法を提案した．次に，我々は現在の機械翻訳技術が特に苦手とする状況，すなわち学習データのドメインとテストデータのドメインが異なる設定における翻訳実験を行った．この実験により，我々の手法で構築したバックオフモデルを用いることで，日英翻訳，英日翻訳のいずれにおいても翻訳品質が向上することを示した．

今後の課題としては，(1) 意味表現の翻訳の高度化と，(2) 機械翻訳技術の高精度化，(3) 機械翻訳以外の多言語処理技術への応用が挙げられる．以下，この 3 点について詳細に述べる．

意味表現の翻訳の高度化　本研究で提案した意味表現の翻訳手法は，適用先が単語から単語への翻訳に限定されている．しかし，本来異なる言語における意味の対応関係は単語:単語に限定されない．たとえば，日本語の単語 “クワガタムシ” は，英

語では 2 単語からなる “stag beetle” という句に対応している．提案手法の枠組みで正しく “クワガタムシ” → “stag beetle” の翻訳を実現するためには，“stag beetle” の意味表現を適切に与える必要がある．

ここで考えうる最も簡単なアプローチは，単言語コーパス内に出現する “stag beetle” を 1 語とみなして，他の語と同様に意味表現を学習する方法である．しかし，このアプローチには句の語数が増加するに従って，データの過疎性により有意な意味表現が獲得できないという根本的な欠陥が存在する [11] ことがわかっている．

より現実的なアプローチとしては，意味表現の翻訳の前処理として，構成的意味論 [9] に基づく意味表現の合成を行う方法が考えられる．意味表現の合成としては，たとえば “stag” のベクトルと “beetle” のベクトルの要素ごとの和や積を求める手法 [9] や，「“stag” は “beetle” に係る」といった係り受け構造を考慮した手法 [41, 42] 等を検討すべきである．

機械翻訳技術の高精度化　本研究で行った機械翻訳に関する実験では，意味表現の翻訳手法の適用先を未知語に限定した．この実験設定は，「対訳コーパスから得られる翻訳例 (とその翻訳確率) が最も正確な情報源であり，それが使えない場合にのみベクトル翻訳に頼ればよい」という観点から設計されたものである．しかし，現実のシステムにおいては対訳コーパスから得られる翻訳例が常に正しいとは限らない．たとえば，英日翻訳を行う際 “book” という語は “本” や “歌詞”，“‘予約する” などさまざまな語に翻訳されうるが，システムがどの訳語を選択しやすいかは学習データのドメインに依存する．書籍ドメインの対訳コーパスを学習データとして用いていれば “本” が，音楽ドメインの対訳コーパスであれば “歌詞” が，旅行会話ドメインの対訳コーパスであれば “予約する” がそれぞれ選ばれやすくなる．

学習データとテストデータのドメインが異なる状況においては，本研究が取り組んだ未知語の問題以外にも上記のような問題が存在する．こうした問題も，意味の表現手法を適用することにより解決しうるので，今後取り組む必要があると考えている．具体的には，対訳辞書の自動拡張を行う際に概念辞書 WordNet[43] を用いて多義語の語義曖昧性解消 (Word sense disambiguation, WSD) を行う手法 [44] や，新

語義検出 (Word sense induction, WSI) に基づいて翻訳モデルを学習する手法 [45] 等を検討すべきである．

機械翻訳以外の多言語処理技術への応用　本研究では，統計的機械翻訳システムに意味表現の翻訳手法を応用し，未知語を含む文の翻訳精度を向上させた．しかし，意味表現の翻訳手法は機械翻訳以外の多言語処理技術にも適用しうるものである．具体的には，言語横断情報検索システム [46] におけるクエリ翻訳やクエリ拡張の処理へ適用することが考えられる．

以上で述べた 3 つの課題については，著者が博士課程進学後に引き続き取り組む予定である．

参考文献

[1] Zellig S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[2] John R. Firth. A synopsis of linguistic theory. *Studies in Linguistic Analysis*, pages 1–32, 1957.

[3] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint*, 2013.

[4] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, June 1996.

[5] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, February 2003.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations (ICLR)*, 2013.

[7] Pascale Fung. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 1–17, 1998.

[8] Philipp Koehn and Barry Haddow. Interpolated backoff for factored translation models. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012.

[9] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 236–244, 2008.

[10] Reinhard Rapp. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the ninth Machine Translation Summit (MT SUMMIT IX)*, pages 315–322, 2003.

[11] Peter D. Turney. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research (JAIR)*, 44:533–585, July 2012.

[12] Peter D. Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)*, 37(1):141–188, February 2010.

[13] Gerard Salton. *The SMART retrieval system—experiments in automatic document processing*. Prentice-Hall, 1971.

[14] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[15] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[16] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[17] John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526, 2007.

[18] Dekang Lin and Xiaoyun Wu. Phrase clustering for discriminative learning. In *Proceedings of Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, pages 1030–1038, 2009.

[19] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–247, 2014.

[20] Sergei Nirenburg and Harold L. Somers. *Readings in machine translation*. MIT Press, 2003.

[21] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

[22] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[23] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, pages 48–54, 2003.

[24] 渡辺 太郎, 今村 賢治, 賀沢 秀人, Neubig Graham, 中澤 敏明, and 奥村 学. **機械翻訳**. コロナ社, 2014.

[25] Katrin Erk and Sebastian Padó. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 897–906, 2008.

[26] Nal Kalchbrenner and Phil Blunsom. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of ACL Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126, 2013.

[27] Katrin Erk. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653, 2012.

[28] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 270–280, 2015.

[29] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[30] Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 58–68, 2014.

[31] Marta R Costa-Jussà. Domain adaptation strategies in statistical machine translation: a brief overview. *The Knowledge Engineering Review*, 30(05):514–520, 2015.

[32] Hirofumi Yamamoto and Eiichiro Sumita. Bilingual cluster based models for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 514–523, 2007.

[33] Prashant Mathur, Fondazione Bruno Keseler, Sriram Venkatapathy, and Nicola Cancedda. Fast domain adaptation of SMT models without in-domain parallel data. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1114–1123, 2014.

[34] Hua Wu, Haifeng Wang, and Chengqing Zong. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 993–1000, 2008.

[35] Léon Bottou. Stochastic learning. In *Advanced lectures on machine learning*, pages 146–168. Springer, 2004.

[36] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical programming*, 127(1):3–30, 2011.

[37] Emmanuel Prochasson, Emmanuel Morin, and Kyo Kageura. Anchor points for bilingual lexicon extraction from small comparable corpora. In *Proceedings of the 12th Machine Translation Summit (MT SUMMIT XII)*, pages 284–291, 2009.

[38] 重藤 優太郎, 鈴木 郁美, 原 一夫, 新保 仁, and 松本 裕治. Zero-shot learning における線形回帰の影響. 2015-NL-222(4):1–8, 2015.

[39] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[40] 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 (ACL)*, pages 311–318, 2002.

[41] Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1183–1193, 2010.

[42] Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. Multi-step regression learning for compositional distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*, pages 131–142, 2013.

[43] George A. Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[44] Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL) Short Papers*, pages 759–764, 2013.

[45] Deyi Xiong and Min Zhang. A sense-based translation model for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1459–1469, 2014.

[46] David A. Hull and Gregory Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pages 49–57, 1996.

発表文献

論文誌

1. 石渡祥之佑, 鍛治伸裕, 吉永直樹, 豊田正史, 喜連川優, 文脈語間の対訳関係を用いた単語の意味ベクトルの翻訳. 人工知能学会論文誌 (JSAI), 2016. (投稿中)

国際会議

1. Shonosuke Ishiwatari, Nobuhiro Kaji, Naoki Yoshinaga, Masashi Toyoda, Masaru Kitsuregawa, Accurate Cross-lingual Projection between Count-based Word Vectors by Exploiting Translatable Context Pairs. Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL 2015). pages 300–304. Beijing, China, 2015.

査読付き国内会議

1. 石渡祥之佑, 鍛治伸裕, 吉永直樹, 豊田正史, 喜連川優, ウェブ上の言語資源を用いた単語のベクトル表現の翻訳. 第 7 回 Web とデータベースに関するフォーラム (WebDB Forum 2014), 東京, 2014.

査読なし国内会議

1. 石渡祥之佑, 吉永直樹, 豊田正史, 喜連川優, 未知語の分布表現の翻訳に基づく機械翻訳のドメイン適応. 言語処理学会第 22 回年次大会 (NLP2016), 宮城, 2016. (発表予定)

ポスター発表

1. 石渡祥之佑, 鍛治伸裕, 吉永直樹, 豊田正史, 喜連川優, 文脈語間の対訳関係を用いた単語ベクトルの翻訳. 言語処理若手の会 第10回シンポジウム (YANS2015), 石川, 2015.