機械翻訳における
大規模フレーズ翻訳知識の獲得と利用
に関する研究

２０１８年　３月

龍　梓

機械翻訳における
大規模フレーズ翻訳知識の獲得と利用
に関する研究

龍 梓

システム情報工学研究科
筑波大学

2018年3月

概 要

機械翻訳とは，自然言語の文や文章を別の自然言語の等価な文や文章へ機械的に変換することである．現在までに，いくつかの機械翻訳方法が提案されてきたが，その中でも主要なものとして，「人手によって作成された規則に基づく機械翻訳」(Rule-based Machine Translation; RBMT),「統計的機械翻訳」(Statistical Machine Translation; SMT)，および,「ニューラルネットワーク機械翻訳」(Neural Machine Translation; NMT) がある．このうち，RBMT の長所は，文法理論・構文構造に従った文構造の解析と生成を行うことができる点であるが，その膨大な規模の文法規則・翻訳辞書・翻訳規則を人手で開発・管理する必要がある点が最大の欠点であった．その人的コストを軽減することを目的として，機械翻訳システムの開発者が手作業で辞書や文法規則，翻訳規則を作成するのではなく，大量に蓄積した翻訳例を計算機処理することにより，機械翻訳システムを自動学習する方式に基づくパラダイムとして，SMT および NMT が考案された．SMT においては，大規模な対訳テキストに対して統計的な数学モデルを適用してそのパラメータを推定し，目的言語文を生成する．その際，句を連結して文を生成し，数単語程度の単語列の出現確率の大小によって句の連結の良否を判断する．したがって，文中において複数個連結された句の列の間の連結の不自然さの問題が頻発し，その結果，出力文には，自然言語としての自然さや流暢さが欠ける．一方，NMT は，原言語文の単語や文を，意味的に似たものが近くに配置されるベクトルに変換し，その原言語文の意味ベクトルから直接目的言語文を生成するため，流暢な目的言語文を生成できる．しかし，NMT の弱点の一つとして，扱う語彙のサイズの増加に伴い，NMT モデルの訓練および翻訳に要する時間が増すため，扱える語彙に限りがある点が知られている．

ここで，本論文では，機械翻訳において翻訳対象となる文書ジャンルの中でも，最もその重要性が大きいジャンルの一つである特許文書の翻訳をとりあげる．大規模対訳特許文を情報源として，機械翻訳における大規模フレーズ翻訳知識の獲得と利用の観点から,「対訳専門用語の同定」,「同義対訳専門用語の同定」,「統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラル機械翻訳」の課題について取り組む.「対訳専門用語の同定」，および,「同義対訳専門用語の同定」の課題においては，機械翻訳，特に，RBMT のパラダイムにおいて利用するための大規模フレーズ翻訳知識の獲得手法に取り組む．一方,「統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラル機械翻訳」の課題においては，NMT のパラダイムにおいて大規模フレーズ翻訳知識を利用する方式について取り組む．この方式においては，SMT の長所である大語彙フレーズ翻訳，および，NMT の長所である自然かつ流暢な訳文生成の両方の長所を併せ持った NMT 方式を実現した．

第 1 章では，全体の序論として，研究の背景について述べる．第 2 章では，本論文でとりあげる機械翻訳のパラダイムについて述べる．第 3 章では，本論文で用いた言語資源であるコーパス，SMT によってコーパスから自動生成されたフレーズ翻訳テーブル，および，フレーズ翻訳テーブルの利用法について述べる

第 4 章では，大規模フレーズ翻訳知識の獲得における研究課題の一つとして，「対訳専門用語の同定手法」について述べる．この研究では，SMT が持つ特異な機能として，大規模フレーズ翻訳テーブルの自動生成機能に着目し，これを利用して，RBMT のための大規模フレーズ翻訳辞書作成支援方式を考案した．この研究においては，まず，対訳文およびフレーズ翻訳テーブルを用いた訳語推定を行い，フレーズ翻訳テーブルに含まれる大部分の誤訳・ノイズを排除して，対訳専門用語の候補を獲得する．そして，複数の対訳文から得られた素性を用いた分類器学習方式を適用することによって，対訳専門用語の同定を行った．評価実験の結果として，提案手法により，適合率を最大化する調整を行った場合の適合率は 90%以上で，F 値を最大化する調整を行った場合の F 値は 80%以上となった．

次に，第 5 章において，獲得された大規模フレーズ翻訳知識の高機能化の研究課題として，獲得された対訳専門用語に対して「同義対訳専門用語を同定」する手法について述べる．この研究では，第 4 章の研究課題において各対訳文を情報源として専門用語対訳対を同定する際に，それぞれの対訳文から同定された専門用語対訳対の間の関係性を考慮していない点に着目し，この問題点の解消に取り組む．この研究では，原言語・目的言語方向，および，目的言語・原言語方向，の両方向において専門用語の訳語推定を繰り返すことによって，人手で選定した中心的対訳対に対して同義の可能性のある対訳対を網羅する同義対訳専門用語候補集合を生成する．そして，分類器学習方式を適用することによって同義集合および異義集合を同定する．評価実験において，日中パテントファミリーから抽出した 360 万対の日中対訳文に対して提案手法を適用し，同義関係にある日中対訳専門用語の同定において，再現率が 25%以上という条件のもとで，約 90%の適合率を達成した．

さらに，第 6 章では，SMT によって獲得された大規模フレーズ翻訳知識の利用における研究課題として，SMT による大語彙フレーズ翻訳との併用による NMT 方式について述べる．標準的な NMT モデルにおいては，扱える語彙のサイズに限界があるため，モデル中には低頻度語を含むことができない．したがって，いかに流暢な訳文を生成できたとしても，特許文書の正確な翻訳において不可欠である新語・固有表現が未知語となって訳出されない．そこで，訓練用対訳文からフレーズ間の二言語対応の情報を収集し，二言語間で対応済みのフレーズ対訳対を同一のトークンに置き換えた後，NMT モデルの訓練を行う．次に，翻訳時においては，NMT モデルの語彙集合によって対応可能な部分については NMT モデルによる訳文生成がなされ，一方，その他のフレーズまたは単語語彙部分については SMT モデルによって翻訳がなされる．日中，中日，日英，英日の各方向の翻訳

において評価を行い，提案手法の有効性を検証した．結果として，ベースラインである標準的 SMT モデル，および，提案手法が適用されていない NMT モデルを上回る性能を達成した．

最後に，第 7 章において，今後の展望について論じる．

目 次

図 目 次

表 目 次

第1章　序論

機械翻訳とは，自然言語の文や文章を別の自然言語の文や文章へ機械的に変換することである．現在までに，いくつかの機械翻訳方法が提案されてきたが，その中でも主要なものとして,「人手によって作成された規則に基づく機械翻訳」(Rule-based Machine Translation; RBMT),「統計的機械翻訳」(Statistical Machine Translation; SMT)，および,「ニューラルネットワーク機械翻訳」(Neural Machine Translation; NMT) がある．RBMT の長所は，文法理論・構文構造に従った文構造の解析と生成を行うことができる点である．しかし，RBMT においては，膨大な規模の文法規則・翻訳辞書・翻訳規則を人手で開発・管理する必要がある点が最大の欠点であった．それらの開発・管理における人的コストを軽減することを目的として，機械翻訳システムの開発者が手作業で辞書や文法規則，翻訳規則を作成するのではなく，大量に蓄積した翻訳例を計算機処理することにより，機械翻訳システムを自動学習する方式に基づくパラダイムとして，SMT および NMT が考案された．標準的な SMT の原理は，原言語文 f から目的言語文 e への翻訳において，確率 $P(e \mid f)$ を最大化する目的言語文 e を求める確率最大化問題 $\arg\max_e P(e \mid f)$ を，別の最大化問題 $\arg\max_e P(e)P(f \mid e)$ に置き換えて解くことである．ここで，$P(e)$ と $P(f \mid e)$ は，それぞれ,「目的言語における文 e の流暢さ」，および,「文 e の，文 f の翻訳としての忠実さ」を表しており，それらの値は，いずれも，統計的な数学モデルを大規模な対訳テキストに適用することによって推定する．しかし，SMT にも解決すべき問題が多く存在する．SMT では，句を連結して文を生成し，数単語程度の単語列の出現確率の大小によって句の連結の良否を判断する．したがって，文中において複数個連結された句の列の間の連結の不自然さの問題が頻発し，その結果，出力文には，自然言語としての自然さや流暢さが欠ける．一方，NMT は，SMT をのように句単位での翻訳を行うのではなく，深層学習によって，原言語文の単語や文を，意味的に似たものが近くに配置されるベクトルに変換し，その原言語文の意味ベクトルから直接目的言語文を生成する方式である．SMT と比べると，NMT は，表現の幅が大きく，かつ，流暢な目的言語文を生成できる点において大きな利点を有する．しかしながら，NMT の弱点の一つとして，扱う語彙のサイズの増加に伴い，NMT モデルの訓練および翻訳に要する時間が増すため，扱える語彙に限りがある点が知られている．

ここで，本論文では，機械翻訳において翻訳対象となる文書ジャンルの中でも，最もその重要性が大きいジャンルの一つである特許文書の翻訳をとりあげる．近

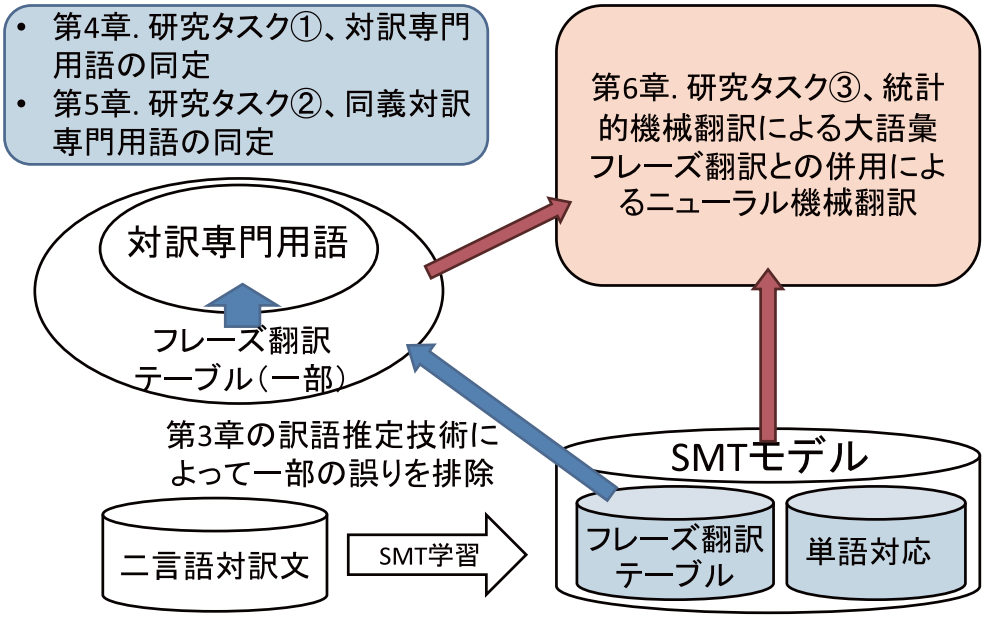


図 1.1: 本論文の全体像

年，特許出願数が急増し，知的財産権の保護意識が急速に高まり，国際特許情報サービス業が発展するとともに，多言語特許翻訳のニーズも高まりを見せている．しかし，現在，特許翻訳，特に，特許出願時の特許文書の翻訳業務においては，以前として人手の介在が不可欠であり，大量の特許出願文書の高速かつ高精度翻訳には至っていないのが実情である．このことから，特許翻訳の分野においては，特許文書の高精度自動翻訳が喫緊な課題となっている．他のジャンルの文書の翻訳と比べて，特許文書の翻訳において特に問題となることとして，(1) 新出の専門用語が多数出現すること，および，(2) 一つの文が長く，かつ，回りくどくて難解な構文で書かれること，が挙げられる．ここで，本論文では，特許文書の翻訳特有の二つの問題の中でも，特に (1) の専門用語の翻訳の問題に焦点を当てる．

本論文において取り組む三つの項目の全体像を図 1.1 に示す．まず，一つ目の研究課題として，機械翻訳，特に，RBMT のパラダイムにおいて利用するための大規模フレーズ翻訳知識の獲得手法について研究を行った．この研究では，SMT が持つ特異な機能として，大規模フレーズ翻訳テーブルの自動生成機能に着目し，これを利用して，RBMT のための大規模フレーズ翻訳辞書作成支援方式を考案した．ここで特に，SMT において自動生成されるフレーズ翻訳テーブルにおいては，貴重なフレーズ翻訳知識が大規模に含まれる反面，翻訳精度に悪い影響を与える誤訳・ノイズの類も多く含まれる．そこで，この研究においては，複数の対訳文から得られた素性を用いた分類器学習方式を適用することによって，対訳専門用語の同定を行った．次に，二つ目の研究課題として，一つ目の研究課題において獲得された大規模フレーズ翻訳知識を高機能化するための研究課題に取り組んだ．この課題では，一つ目の研究課題において各対訳文を情報源として専門用語対訳対を

同定する際に，それぞれの対訳文から同定された専門用語対訳対の間の関係性を考慮していない点に着目し，この問題点の解消に取り組んだ．具体的には，それらの複数の専門用語対訳対の間の同義・異義関係を同定することにより，獲得された対訳辞書のカバレージを改善した．最後に，三つ目の研究課題として，NMT のパラダイムにおいて大規模フレーズ翻訳知識を利用する方式について研究を行った．標準的な NMT モデルにおいては，扱える語彙のサイズに限界があるため，モデル中には低頻度語を含むことができない．したがって，いかに流暢な訳文を生成できたとしても，特許文書の正確な翻訳に不可欠である新語・固有表現が未知語となって訳出されない．そこで，NMT のパラダイムにおいて大規模フレーズ翻訳知識を利用する方式について研究を行った．具体的には，SMT において生成される大規模フレーズ翻訳知識を用いることによって，SMT による大語彙フレーズ翻訳と NMT による文翻訳とを併用する方式を考案した．これによって，SMT の長所である大語彙フレーズ翻訳，および，NMT の長所である自然かつ流暢な訳文生成の両方の長所を併せ持った NMT 方式を実現した．

対訳専門用語の同定

対訳特許文から自動的に対訳専門用語の辞書を構築する技術について研究を行う．機械翻訳，特に，翻訳規則および翻訳辞書を必須とするパラダイムである RBMT や，人手によって翻訳を行う場合，高い翻訳品質を保つためには，大規模で正確な対訳辞書が必要である．ここで，各国では，年々新しい技術開発が行われ，新しい専門用語が作られ，特許が申請されている．しかし，人手によって対訳辞書を作成するためには膨大な時間と労力を要するため，年々新しく作られる専門用語を迅速に専門用語対訳辞書に追加していくためには，自動もしくは半自動的に専門用語対訳辞書を構築する手法が必要である．

これまでに行われてきた研究を大別すると，初期の研究としては，二言語間で文間の対応がつけられた文対応対訳コーパスを情報源として，対訳文中の共起頻度を用いる手法がよく研究された．また，文対応対訳コーパスよりも利用可能性の高いコンパラブルコーパスを情報源とする手法も，初期の時期から近年に至るまでよく研究されている．さらに，近年では，多言語文書を収集する情報源として，ウェブ上の多様な言語・分野・ジャンルの文書を利用する研究も数多く進められている．例えば，複合語である専門用語の構成要素の訳語を連結して訳語の候補を生成する要素合成法，および，ウェブから収集した目的言語の専門分野コーパスを用いて，生成された訳語候補を検証する手法，検索エンジン等を利用して訳語が併記された文書を収集し，訳語対を獲得する手法訳語対を獲得する手法，多言語文書である Wikipedia を情報源とする手法等がある．ここで，これらの研究においては，対訳辞書を作成するためにの情報源の特性に応じて適切な手法が選択され，対訳辞書作成におけるそれぞれの手法の有用性が評価された．

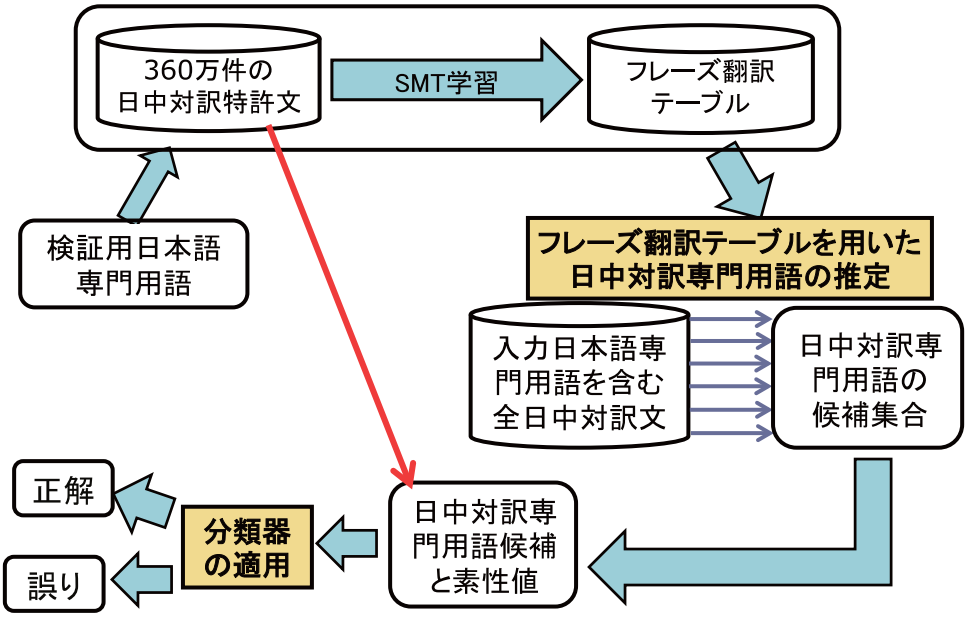


図 1.2: 対訳専門用語の同定

本項目では，日本語・中国語間の RBMT によって特許文書を翻訳するための対訳辞書を構築する課題に取り組む．具体的には，図 1.2 に示すように，日中パテントファミリーから抽出した日中対訳特許文を言語資源として，句に基づく SMT モデルにより学習されたフレーズ翻訳テーブルを用いて，対訳専門用語を獲得する手法を提案する．まず，専門用語対訳辞書獲得の情報源として用いる日中対訳文対に対して，句に基づく SMT モデルを適用することにより，フレーズ翻訳テーブルを学習する．ここでは，各日本語専門用語に対して，全対訳データから日本語専門用語を含む日中対訳文を収集し，句に基づく SMT モデルのフレーズ翻訳テーブル用いた訳語推定を行う．実験では，評価対象である 360 万対の日中対訳特許文から 78 例の日本語専門用語を選定した．そして，SMT モデルのフレーズ翻訳テーブルを用いて訳語推定を行い,「中国側が形態素単位のフレーズ翻訳テーブル」を用いた場合では，2,533 例の日中対訳専門用語を生成し，全 2,533 例中，専門用語の対訳対である正例を 1,531 例，負例を 1,002 例とした.「中国側が文字単位のフレーズ翻訳テーブル」を用いた場合では，2,092 例の日中対訳専門用語を生成し，全 2,092 例中，専門用語の対訳対である正例を 1,255 例，負例を 1,002 例とした．次に，このフレーズ翻訳テーブルを用いて日本語専門用語の中国語訳語推定を行う．最後に，獲得した日中対訳専門用語に対して，複数の対訳文から得られる情報を素性として，Support Vector Machines (SVMs) を適用する．素性として，単言語素性と二言語素性の二種類の素性を用いた．評価実験において適合率を最大化する調整を行い,「中国側が形態素単位のフレーズ翻訳テーブル」を用いた場合において 94.2% の適合率を，また,「中国側が文字単位のフレーズ翻訳テーブル」を用いた場合において 93.9%の適合率を，それぞれ達成した．なお，本論文では，

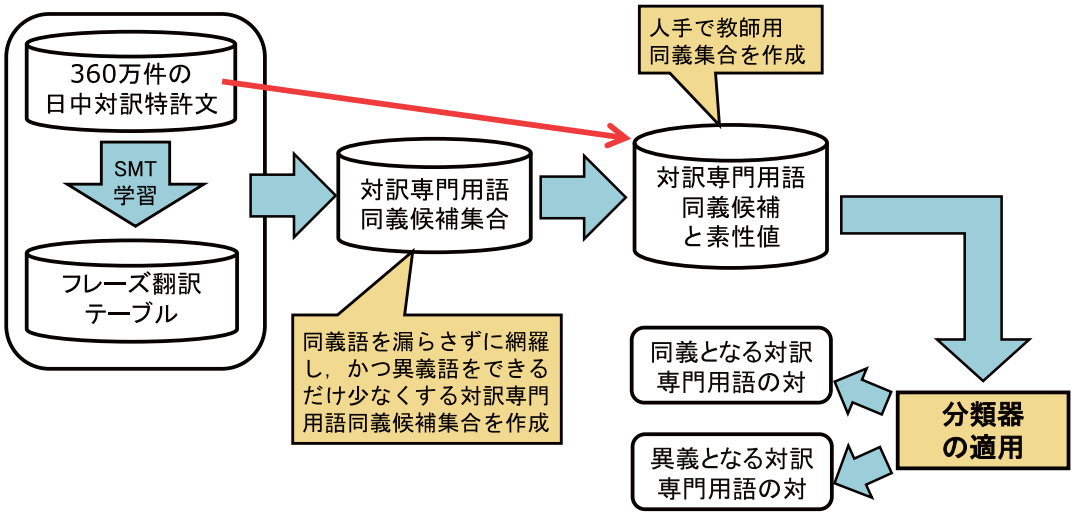


図 1.3: 同義対訳専門用語の同定

日中対訳特許文からの専門用語対訳対獲得について述べるが，本項目の手法は特定の言語対には依存しない．

同義対訳専門用語の同定

対訳特許文から獲得した対訳専門用語の同義・異義関係を同定する手法について研究を行う．本項目では，前項目の研究課題において各対訳文を情報源として専門用語対訳対を同定する際に，それぞれの対訳文から同定された専門用語対訳対の間の関係性を考慮していない点に着目し，この問題点の解消に取り組む．具体的には，それらの複数の専門用語対訳対の間の同義・異義関係を同定することにより，獲得された対訳辞書のカバレージを改善する．

前項目において，対訳特許文を情報源として専門用語対訳対獲得を行う手法では，ある日本語専門用語の訳語推定の際に，その日本語専門用語が出現する一つの対訳文に出現する訳語のみを推定対象としていた．したがって，他の対訳文に出現している同義の専門用語対訳対とは全く独立に訳語推定が行われており，本来同義関係にある複数の専門用語対訳対の間の関係を同定できない，という問題点があった．そこで，ある日本語専門用語およびその同義語候補が出現する複数の対訳文を入力として，同義の専門用語対訳対を同定する手法を提案する．

図 1.3 に示すように，実際の手順においては，まず，ある日本語専門用語を種として，同義関係にある専門用語対訳対の候補を生成・収集する．生成・収集した候補集合の中から同義判定を行うための中心的対訳対を選び，中心的対訳対のうちの日本語専門用語に対して，専門用語対訳対同義候補集合を再生成する．次に，再生成した候補集合に対して SVM 分類器を適用することにより，同義集合・異義

集合を同定する．SVM 分類器の素性としては，入力対訳対の特性を規定するもの，および，入力対訳対と中心的対訳対の間の関係を規定するものの二種類を利用した．評価実験において，日中パテントファミリーから抽出した 360 万対の日中対訳文に対して提案手法を適用し，同義関係にある日中対訳専門用語の同定において，再現率が 25%以上という条件のもとで，約 90%の適合率を達成した．さらに，比較対象として，日英同義対訳専門用語の同定を対象とした先行研究における素性と同等の素性のもとで日中同義対訳専門用語の同定を行った評価結果との比較を行い，本論文で提案した素性の組合わせによって大幅に性能が改善されることを示した．

統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラル機械翻訳

大語彙フレーズの翻訳に対応した NMT モデルについて研究を行う．近年，従来の SMT に代わって NMT モデルが盛んに研究されている．NMT は，原言語文を固定長ベクトルへ写像し，その固定長ベクトルから目的言語文を生成するため，意味的要素の翻訳に非常に優れており，SMT を上回る翻訳精度を達成している．しかしながら，NMT の弱点の一つとして，扱える語彙に限りがある点が知られている．具体的には，扱う語彙のサイズの増加に伴い，NMT モデルの訓練および翻訳に要する時間が増す点が課題となっている．

NMT においては，語彙辞書に含まれていない単語は未知語トークンとして出力されるため，これが誤訳となる．そこで，これまでにも，NMT が扱える語彙の規模を拡大する方式について研究が行われてきた．従来の手法として，訓練用対訳文における単語対応の情報に基づいて，語彙辞書に含まれていない未知語単語を，単語間の対応関係を特定できるトークンに置き換えた後，NMT モデルの訓練を行う方式が提案された．この方式では，出力文に含まれた未知語トークンが対応する原言語の単語を推定しその訳語に置き換えることによって，NMT の出力文において出力可能となる語彙の規模を拡大した．しかし，この方式は，単語単位での語彙規模の拡大にとどまる点が弱点であった．この弱点のため，複合語によって構成される専門用語が多数含まれる特許文の翻訳精度の改善においては限界があった．

以上の背景のもとで，本項目においては，特許文を対象としたニューラルネットワーク翻訳において，大規模フレーズ語彙に対応する方式について研究を行った．図 1.4 に示すように，本方式においては，訓練用対訳文においてフレーズ間の二言語対応の情報を収集し，二言語間で対応済みのフレーズ対訳対を同一のトークンに置き換えた後，NMT モデルの訓練を行う．本方式による特許文の翻訳時には，NMT モデルの語彙集合中の語彙部分に対しては，NMT モデルによる訳文生成がなされ，一方，その他のフレーズまたは単語語彙部分に対しては，SMT モデルによる翻訳がなされる．日中，中日，日英，英日の各方向の翻訳において評価を行い，提案手法の有効性を検証した．本方式を用いない従来型の NMT モデル

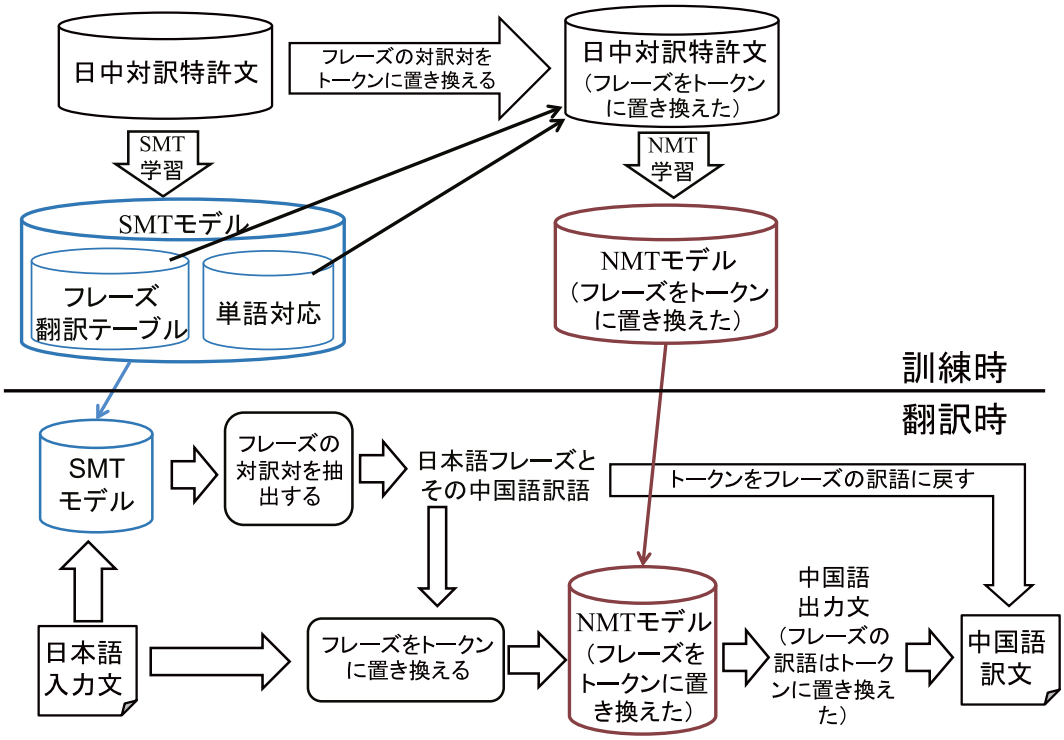


図 1.4: 統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラル機械翻訳

と，本方式との間で翻訳精度の比較を行った結果，0.7 ポイント以上の BLEU の向上を達成できた．さらに，NMT の弱点である訳抜けの改善においては，提案手法が適用されていない NMT モデルによる訳抜けを約 30%減らすことができた．

本論文の構成

本論文は，以下の各章から構成される．第 1 章では，全体の序論として，研究の背景について述べる．第 2 章では，本論文でとりあげる機械翻訳のパラダイムについて述べる．第 3 章では，本論文で用いた言語資源であるコーパス，SMT によってコーパスから自動生成された単語対応とフレーズ翻訳テーブル，および，単語対応とフレーズ翻訳テーブルの利用法について述べる．第 4 章では，大規模フレーズ翻訳知識の獲得における研究課題の一つとして，対訳専門用語の同定手法について述べる．次に，第 5 章において，獲得された大規模フレーズ翻訳知識の高機能化の研究課題として，獲得された対訳専門用語間の同義・異義関係の同定手法について述べる．さらに，第 6 章では，SMT によって獲得された大規模フレーズ翻訳知識の利用における研究課題として，SMT による大語彙フレーズ翻訳との併用による NMT 方式について述べる．最後に，第 7 章において，本論文の今後の展望について論じる．

第2章　機械翻訳のパラダイム

機械翻訳 (Machine Translation; MT) とは，自然言語の文や文章を別の自然言語の等価な文や文章へ機械的に変換することである．機械翻訳に関する歴史の詳細は [40] に記述されている．本章では，その中でも主要なもので代表的な機械翻訳の手法である,「人手によって作成された規則に基づく機械翻訳 (Rule-based Machine Translation; RBMT)」(2.1 節) ,「統計的機械翻訳 (Statistical Machine Translation; SMT)」(2.2 節)，および,「ニューラルネットワーク機械翻訳 (Neural Machine Translation; NMT)」(2.3 節) について述べる．

2.1　人手によって作成された規則に基づく機械翻訳

規則に基づく機械翻訳 (Rule-based Machine Translation; RBMT) では，翻訳のために必要な知識を，専門家が人手によって記述し，記述されたルールに基づいて翻訳が生成される [38].

SMT や NMT といった自動学習に基づく機械翻訳方式が台頭するまでの時期において中心的位置を占めていた機械翻訳パラダイムの一つが RBMT であるが，その中でも，多くの商用機械翻訳システムが採用していた翻訳方式は，構文トランスファーと呼ばれる方式である．構文トランスファー方式に基づく機械翻訳の一例を図 2.1 に示す．構文トランスファー方式においては，語彙と文法に関する言語的な知識を利用して原言語の文の構文を解析して，文の構文構造 (図 2.1 の入力文の依存構造) を生成し，これを目的言語の文の構文構造 (図 2.1 の出力文の依存構造) に変換し，最後に，変換された構文構造から目的言語の表層表現を生成する方式である．

まず，図 2.1 の「1. 解析処理」の部分では，入力文の単語・形態素列を作成する．単語分割においては，同時に，品詞推定，名詞の単複，動詞の原形の推定を行う．そして，構文解析を行い，文の動詞を根として，単語と単語の間の修飾関係を表す依存構造を生成し出力する．さらに，構文構造をふまえて，動詞から見た名詞句の意味役割 (格と呼ぶ) を認定して付与する．次に，図 2.1 の「2. 構造変換」の部分では，原言語文の依存構造を目的言語文の依存構造へ変換する．変換過程においては，予め人手によって作成された格パターン対応規則を適用するとともに，対訳辞書を参照して単語の翻訳を行い，当該単語部分の部分木を変換す

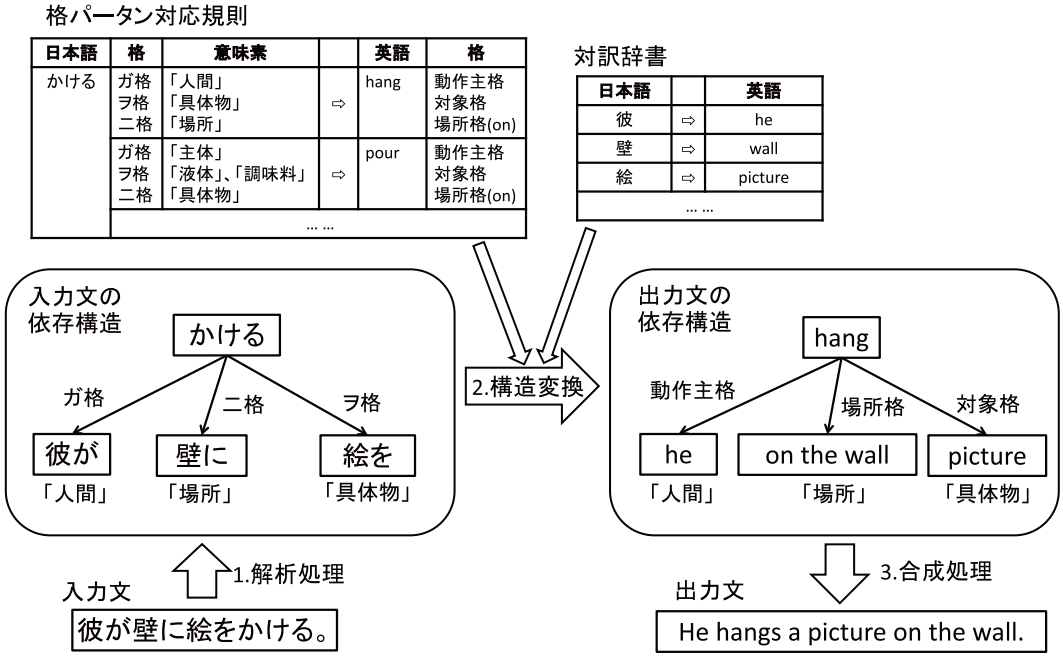


図 2.1: 構文トランスファー方式に基づく機械翻訳

る．単語の訳し分けはこの時点で行われ，訳すべき単語のみではなく，文中の他の単語も参照して訳語を決定する．例えば，格パターン対応規則を参照して,「かける」のガ格やヲ格などの意味素に基づいて英語訳語を “hang” を選択する．また，格要素の名詞については，対訳辞書を参照して訳語を埋め込む．最後に，図 2.1 の「3. 生成処理」の部分では，目的言語の依存構造を単語列に変換し，文を生成する．図 2.1 の例では，英語文を生成する際に，動詞 “hang” を三人称単数 “hangs” に変換し，“wall” と “picture” に適切な冠詞を付与することによって，最終的に英語文 “He hangs a picture on the wall.” を出力する．

RBMT では，文の解析・変換・生成の知識を人手で用意することによって翻訳を行うが，言語現象は多様であり，すべての言語現象に対応する規則を人手で用意することはほぼ不可能である．また，新たな翻訳規則の追加によって，他の規則との間で矛盾を引き起こすことがないように細心の調整が不可欠となるが，これらの作業は，専門家がすべて手作業によって解決することが必須である．このように，RBMT においては，システム構築の労力が極めて大きいため，一個人で新たなシステムを設計し実装することは容易ではない．

そこで，機械翻訳システムの開発者が手作業で辞書や文法規則，翻訳規則を作成するのではなく，大量に蓄積した翻訳例を計算機処理することにより，文法規則・翻訳辞書・翻訳規則などの開発・管理における人的コストを軽減することを目的

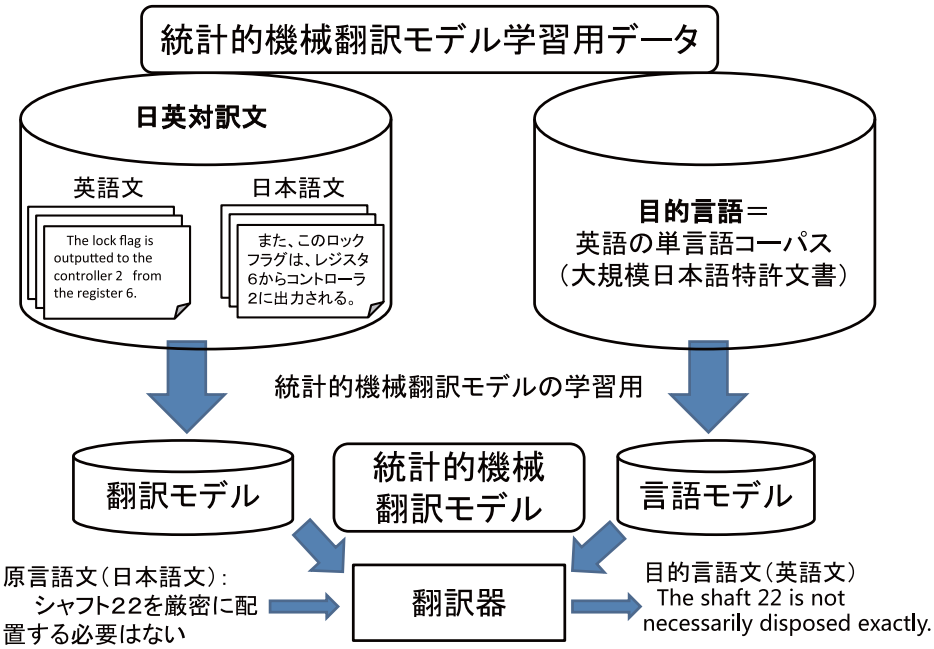


図 2.2: 統計的機械翻訳の枠組 ([53] から抜粋)

として，機械翻訳システムを自動学習する方式に基づくパラダイムとして，SMT および NMT が考案された.

2.2 統計的機械翻訳

統計的機械翻訳 (Statistical Machine Translation; SMT) は，NMT と並んで，大量に蓄積した翻訳例を計算機処理することにより，機械翻訳システムを自動学習する方式に基づく機械翻訳パラダイムの一つである [1].

一般的な統計的機械翻訳の枠組を図 2.2 に示す．SMT においては，原言語文を $\boldsymbol{f} = f_1, \ldots, f_N$(図 2.2 の場合，日本語文)，目的言語文を $\boldsymbol{e} = e_1, \ldots, e_M$(図 2.2 の場合，英語文) として，文 $\boldsymbol{f}$ の翻訳結果として，確率 $P(\boldsymbol{e} \mid \boldsymbol{f})$ を最大にする翻訳文 $\hat{\boldsymbol{e}}$ を求める.

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} P(\boldsymbol{e} \mid \boldsymbol{f})$$

ここで，ベイズの定理を適用して以下の式変形を行う．翻訳文の選択の際には定数扱いとなる分母の項 $P(\boldsymbol{f})$ を省略すると，二項の確率の積 $P(\boldsymbol{f} \mid \boldsymbol{e})\ P(\boldsymbol{e})$ を最大

[1] 本節における SMT の定式化は，[39] における定式化に従う.

にする翻訳文を求める問題に帰着する．

$$\begin{aligned}\hat{\boldsymbol{e}} &= \arg\max_{\boldsymbol{e}} \frac{P(\boldsymbol{f} \mid \boldsymbol{e})\ P(\boldsymbol{e})}{P(\boldsymbol{f})} \\ &= \arg\max_{\boldsymbol{e}} P(\boldsymbol{f} \mid \boldsymbol{e}) P(\boldsymbol{e}) \end{aligned} \tag{2.1}$$

式 (2.1) の二項のうち，第二項の $P(\boldsymbol{e})$ は，翻訳結果である $\boldsymbol{e}$ が目的言語文として自然な文である度合いを表す確率モデルであり，これを言語モデルと呼ぶ．一般に，言語モデルは，次式で示す単語の n 個の連接の確率の積 (n-gram) で表現されており，次式中の各パラメータは，目的言語の大規模な単言語コーパスを用いて推定される．

$$\begin{aligned} P(\boldsymbol{e}) &= P(e_1, \ldots, e_M) \\ &= P(e_1)P(e_2 \mid e_1)P(e_3 \mid e_1^2) \ldots P(e_M \mid e_1^{M-1}) \\ &\approx \prod_{i=1}^{M} P(e_i \mid e_{i-n+1}^{i-1}) \end{aligned} \tag{2.2}$$

一方，式 (2.1) の二項のうち，第一項の $P(\boldsymbol{f} \mid \boldsymbol{e})$ は，目的言語文 $\boldsymbol{e}$ から原言語文 $\boldsymbol{f}$ を生成する際の確からしさを表す条件付き確率によって表現された確率モデルであり，これを翻訳モデルと呼ぶ．

SMT においては，これまでに何種類かのモデル化が提案されているが，それらを大別すると，翻訳モデルにおいて用いられる情報によって，句に基づく統計的機械翻訳モデル，あるいは，構文木に基づく統計的機械翻訳等に分類される．それらのうち，特に，句に基づく統計的機械翻訳モデルは，原言語・目的言語の文法構造を考慮したモデル化を必要とせず，句の二言語間対応，および，二言語間並べ替えに基づいて翻訳結果を生成する手法であり，これまでにも，様々な言語対に対して SMT による翻訳システムが開発されている．句に基づく統計的機械翻訳により，原言語文 (図 2.3 の場合，日本語文) を目的言語文 (図 2.3 の場合，英語文) に翻訳する処理の流れを図 2.3 に示す．まず,「(1) フレーズ分割」においては，フレーズ翻訳テーブルに登録されている原言語フレーズを参照して，入力文をフレーズに分割する．そして，可能な全てのフレーズ分割の結果に対して，次の「(2) フレーズ翻訳」を行い，可能な翻訳の組み合わせをすべて出力する．ここでのフレーズ翻訳テーブルとは，句に基づく統計的機械翻訳モデルにおける翻訳モデルの一部で，その中には，訓練文から抽出されたすべての対訳フレーズ組とその翻訳確率が含まれる．次に，原言語と目的言語の語順の違いを吸収するために，句に基づく統計的機械翻訳モデルにおける翻訳モデルの一部である「語順並べ替え確率モデル」を参照しながら,「(3) 語順並べ替え確率モデル適用」を行い，目的言語側のフレーズの語順として可能な組み合わせをすべて出力する．次に，目的言語の n-gram 言語モデルを適用し，出力された目的言語の単語列が目的言語文としてどの程度自然かをを表す確率値を算出する．最後に，フレーズ翻訳確率，語

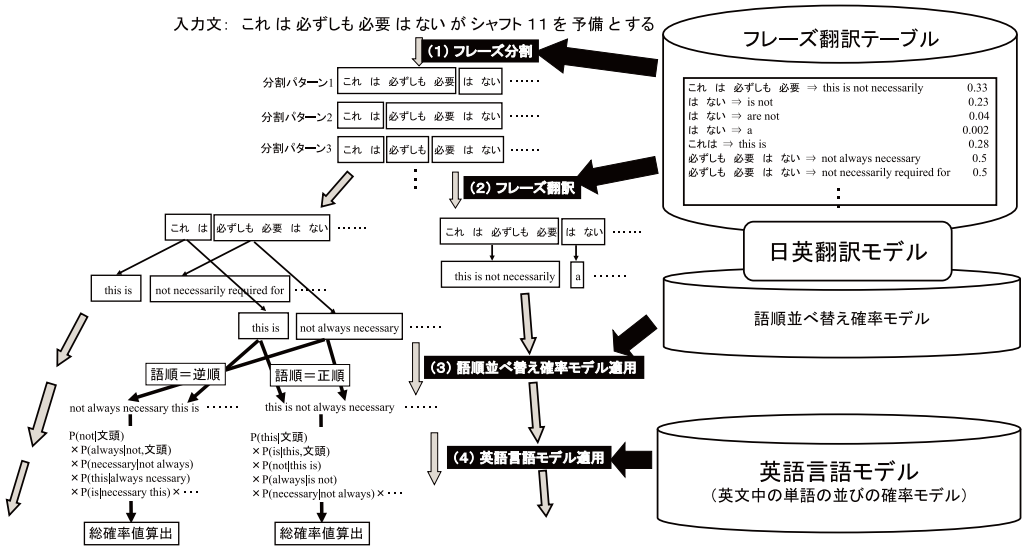


図 2.3: 句に基づく統計的機械翻訳: 翻訳の流れ ([53] から抜粋)

順並べ替え確率，および，言語モデル確率値を掛け合わせることによって，総確率値を算出し，総確率値最大となる目的言語文 $\hat{\boldsymbol{e}}$ を出力する．

以下では，統計的機械翻訳モデルの中でも，本論文の研究において中心的に利用するフレーズ翻訳テーブルに対して，その作成過程の詳細を説明する．まず，フレーズ翻訳テーブルの作成過程における対訳フレーズ組の抽出部分の流れを図 2.4 に示す．まず，二言語対訳文を訓練事例として，EM アルゴリズムを適用することによって，対訳単語対応の同定を行う．次に，同定済みの単語対応情報をふまえ，「文単位の句対応制約」(3.2.1 節) を考慮しながら，訓練用の二言語対訳文の全体から対訳フレーズ組を抽出し，式 (2.3) の条件付き確率を計算して付与し，フレーズ翻訳テーブルに格納する [23]．ここで，式 (2.3) の条件付き確率を求める際には，フレーズペア $\langle p_f, p_e \rangle$ に対して，訓練用の二言語対訳文の全体から抽出された全ての対訳フレーズ組 Φ 中におけるフレーズペア $\langle p_f, p_e \rangle$ の頻度を $count(p_f, p_e)$ とし，最尤推定により次式 $P(p_f \mid p_e)$ を求める．

$$P(p_f \mid p_e) = \frac{count(p_f, p_e)}{\sum_{p'_f \in \Phi} count(p'_f, p_e)} \tag{2.3}$$

以上において説明したように，SMT においては，句を連結して文を生成し，数単語程度の単語列の出現確率の大小によって句の連結の良否を判断する．したがって，文中において複数個連結された句の列の間の連結の不自然さの問題が頻発し，結果的に，出力文において，自然言語としての自然さや流暢さが欠けるという問題が発生する．一方，SMT と比べると，NMT においては，原言語文の意味ベクトルから直接目的言語文を生成するため，生成される文における表現のバリエー

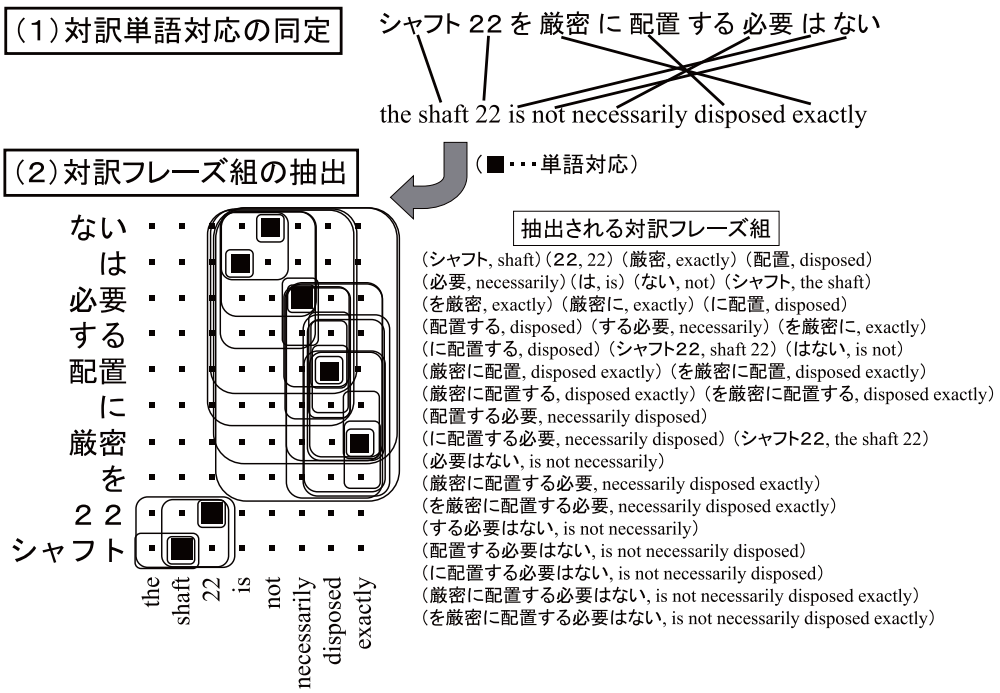


図 2.4: 句に基づく統計的機械翻訳: 対訳フレーズ組の抽出 ([53] から抜粋)

ションが大きく，かつ，流暢な目的言語文を生成できる点において大きな利点を有する．

2.3 ニューラルネットワーク機械翻訳

近年，自然言語分野においては，ニューラル言語モデルや分散表現など，さまざまな深層学習のモデルを利用する方式が提案されてきた [49]．それらの深層学習モデルによる自然言語処理方式の中で，ニューラルネットワーク機械翻訳 (Neural Machine Translation; NMT) とは，1 つの多層ニューラルネットワークによって原言語文を目的言語文に翻訳する過程を定式化する機械翻訳方式の総称ととらえることができる．その中でも，現在，標準的な NMT モデルとして最も広く利用されている方式は，原言語文の系列を目的言語文の系列へ変換する確率をモデル化する系列変換モデルに基づく NMT である [48, 47].

本節では，本論文において中心的に利用する注意機構付き多層双方向系列変換モデル方式の NMT [1, 56] の定式化の詳細を述べる．注意機構付き双方向系列変換モデル方式の NMT を用いて入力文である原言語文 $\boldsymbol{x} = (x_1, \ldots, x_N)$ を出力文である目的言語文 $\boldsymbol{y} = (\hat{y}_1, \ldots, \hat{y}_M)$ に翻訳する流れを図 2.5 に示す．符号化部 (encoder) においては，$i(i = 1, \ldots, N)$ 番目の入力 x_i に対して，前向きと後向きの両方向の多層隠れ層のベクトル $\overrightarrow{h}_i$ と $\overleftarrow{h}_i$ が計算される．ベクトル h_i は，入力文における i

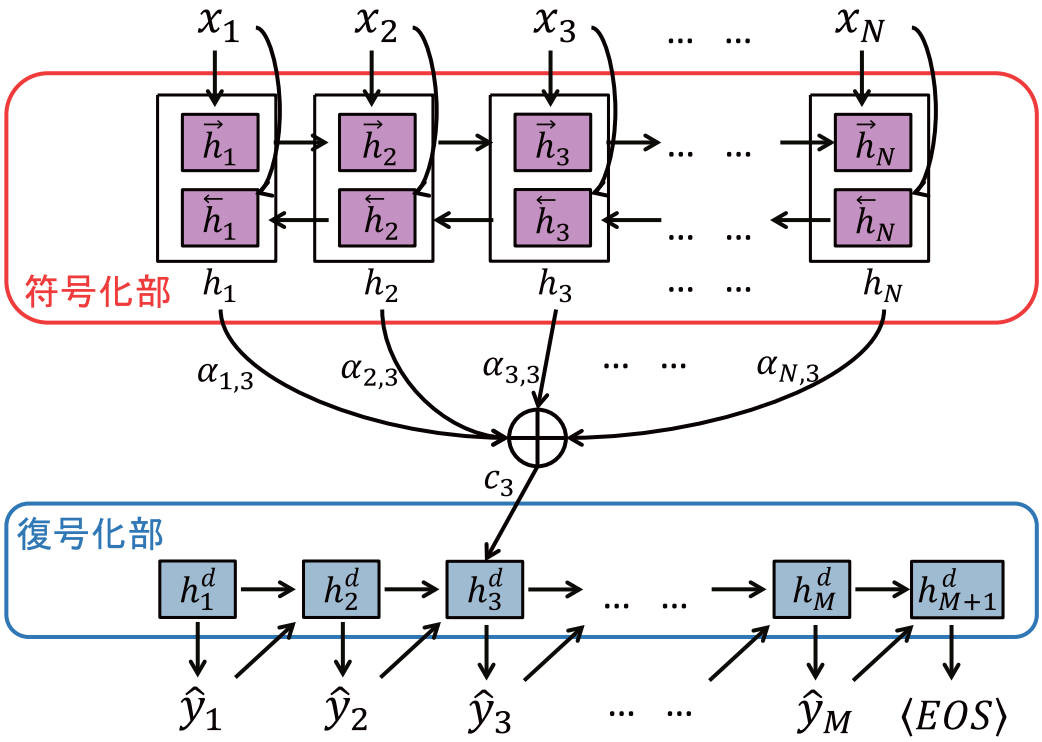


図 2.5: 注意機構付き双方向系列変換モデル方式の NMT [1]

番目時点での意味表現を表すベクトルであり，前向きと後向きの両方向の多層隠れ層のベクトル $\overrightarrow{h}_i$ と $\overleftarrow{h}_i$ を連結して得られる．同様に，復号化部 (decoder) においても，$j(j = 1, \ldots, M)$ 番目の出力を決める時点での多層隠れ層のベクトル h_j^d の計算を行い，このベクトルに基づいて出力 $\hat{y}_j$ が求められる．ここで，復号化部の多層隠れ層のベクトル h_j^d の計算を行う際には，入力文における $i(i = 1, \ldots, N)$ 番目時点での意味表現を表すベクトル h_i に対して重み $\alpha_{i,j}$ を付けて足し合わせることによってコンテキストベクトル c_j を求め，これと，$j-1$ 番目の隠れ層のベクトル h_{j-1}^d および出力 $\hat{y}_{j-1}$ とから h_j^d を求める．以上の流れによって，出力単語 $\hat{y}_j(j = 1, \ldots, M)$ が逐次的に出力される．以下では，これらの定式化の詳細について述べる．

まず，符号化部において原言語文 $\boldsymbol{x} = (x_1, \ldots, x_N)$ の各時点での意味表現を計算する．図 2.5 に示すように，符号化部においては，原言語文 $\boldsymbol{x}$ に対して，回帰型ニューラルネットワークである長短期記憶 (long short-term memory; LSTM) [15] により隠れ層の値を計算して記憶する．ここでは，多層 LSTM を二つ用いて，原言語文 $\boldsymbol{x}$ を，文頭 (x_1) から文末 (x_N) の前向き方向，および，文末 (x_N) から文頭 (x_1) の後向き方向の両方向に符号化することによって頑健性を増す．原言語文における i 番目時点での前向きと後向きの両方向の多層隠れ層のベクトル $\overrightarrow{h}_i$ と $\overleftarrow{h}_i$

は，それぞれ，次式で表現される．

$$\overrightarrow{h}_i = f(\overrightarrow{W}^e[\overrightarrow{h}_{i-1}; \bar{x}_i] + \overrightarrow{b}^e)$$
$$\overleftarrow{h}_i = f(\overleftarrow{W}^e[\overleftarrow{h}_{i+1}; \bar{x}_i] + \overleftarrow{b}^e)$$
$$\bar{x}_i = E^e x_i$$

ただし，E^e は入力語彙の単語の分散表現を並べた埋め込み行列である．入力語彙の埋め込み行列に，入力文中の i 番目の単語に対応する one-hot ベクトル x_i をかけることによって，i 番目の単語の分散表現である埋め込みベクトル $\bar{x}_i$ が得られる．$\overrightarrow{W}_e$ と $\overrightarrow{b}_e$，および，$\overleftarrow{W}_e$ と $\overleftarrow{b}_e$ は，それぞれ，符号化部の前向きベクトル，および，後向きベクトルを計算する際のパラメーターである．$f(\cdot)$ は隠れ層のベクトルを計算するための活性化関数で，tanh 関数等の非線形的な関数を用いる．“;” は二つのベクトルの次元をそのまま結合して，結合後のベクトルの次元数が，元の二つのベクトルの次元数の和となる演算を表し，ベクトル a と b の次元をそのまま結合したベクトルを $[a; b]$ と表記する．そして，原言語文における $i(i = 1, \ldots, N)$ 番目時点での意味表現を，$\overrightarrow{h}_i$ と $\overleftarrow{h}_i$ の結合

$$h_i = [\overrightarrow{h}_i; \overleftarrow{h}_i]$$

とする．

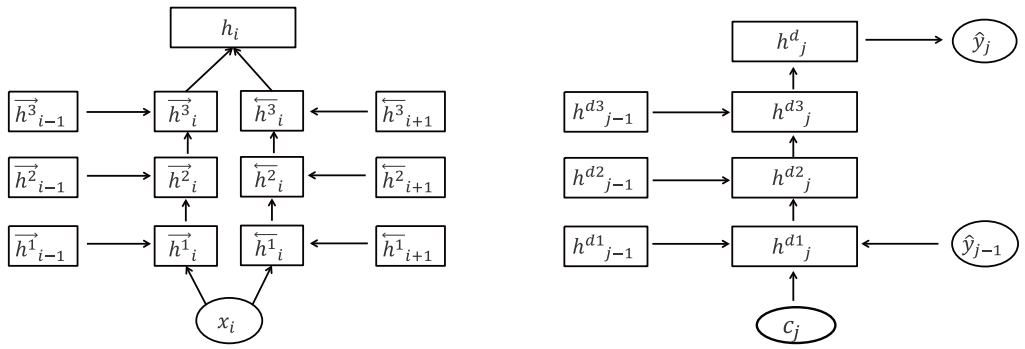
次に，復号化部においては，次式のように，$j(j = 1, \ldots, M)$ 番目時点での多層隠れ層のベクトル h_j^d を用いて，j 番目の出力を決めるための条件付き確率 $P(y \mid \hat{y}_{<j}, \boldsymbol{x})$ を計算し，この確率を最大化する単語 $\hat{y}_j$ を目的言語の語彙集合 $V_{\mathcal{Y}}$ から選択する．以上の系列を繰り返すことによって，出力単語 $\hat{y}_j(j = 1, \ldots, M)$ の系列が逐次的に出力される．

$$P(y \mid \hat{y}_{<j}, \boldsymbol{x}) = g(W^o h_j^d + b^o) \tag{2.4}$$
$$\hat{y}_j = \arg\max_{y \in V_{\mathcal{Y}}} P(y \mid \hat{y}_{<j}, \boldsymbol{x}) \tag{2.5}$$

ただし，W^o と b^o は復号化部の出力層のパラメーターである．$g(\cdot)$ は出力ベクトルを計算するための活性化関数で，一般的には softmax 関数が用いられる．

復号化部の多層隠れ層のベクトル h_j^d の計算を行う際には，多層 LSTM を用いて，次式によって，原言語文における $i(i = 1, \ldots, N)$ 番目時点での意味表現を表すベクトル h_i に対して重み $\alpha_{i,j}$ をかけて足し合わせることによってコンテキストベクトル c_j を求め，これと，$j - 1$ 番目時点での隠れ層のベクトル h_{j-1}^d および出力 $\hat{y}_{j-1}$ とから h_j^d を求める．

$$c_j = \sum_{i=1}^{N} \alpha_{i,j} h_i$$
$$h_j^d = f(W^d[h_{j-1}^d; \bar{y}_{j-1}; c_j] + b^d)$$
$$\bar{y}_{j-1} = E^d \hat{y}_{j-1}$$



(a) 符号化部　　(b) 復号化部

図 2.6: 3 層 NMT モデル

ここで，E^d は出力語彙に対する埋め込み行列で，この埋め込み行列を $j-1$ 番目の出力単語に対応する one-hot ベクトル $\hat{y}_{j-1}$ にかけることによって，$j-1$ 番目の出力単語の埋め込みベクトル $\bar{y}_{j-1}$ を求めた後，h_j^d の計算で用いている．また，W^d と b^d は復号化部の隠れ層ベクトルを計算する際のパラメーターである．$f(\cdot)$ は符号化部の場合と同様に，隠れ層のベクトルを計算するための活性化関数で，tanh 関数等の非線形的な関数を用いる．重み $\alpha_{i,j}$ は次式で定義される．

$$\alpha_{i,j} = \frac{\exp(e_{ij})}{\sum_{i'=1}^{N} \exp(e_{i'j})}$$

$$e_{ij} = u \cdot \tanh(W^{ed} h_i + W^{dd} h_{j-1}^d)$$

ここで，u は重みを計算するためのパラメータの一つである実数ベクトルである．W^{ed}，および，W^{dd} は，それぞれ，重み $\alpha_{i,j}$ の計算において，符号化部の隠れ層ベクトル h_i，および，複合化部の隠れ層ベクトル h_{j-1}^d にかけられるパラメータである．重み $\alpha_{i,j}$ は，j 番目の出力 y_j に対して i 番目の入力 x_i が持つ影響の強さの度合いを表しており，この仕組みのことを注意機構 (attention machanism) と呼ぶ．

また，符号化部および復号化部の隠れ層は，[55] に基づき，図 2.6 の三層の隠れ層とする．図 2.6(a) の符号化部に示すように，三層ある隠れ層の第三層のベクトル $\overrightarrow{h}_i^3$ と $\overleftarrow{h}_i^3$ を符号化部の隠れ層のベクトルとして用い，$h_i = [\overrightarrow{h}_i^3; \overleftarrow{h}_i^3]$ を $i(i = 1, \ldots, N)$ 番目時点の意味表現として用いる．同様に，図 2.6(b) の復号化部に示すように，三層ある隠れ層の第三層のベクトル h_j^{d3} を $j(j = 1, \ldots, M)$ 番目時点における復号化部の隠れ層のベクトル h_j^d として用いる．

以上のように，NMT は，ニューラルネットワークによって原言語文の分散表現から，直接，翻訳文を生成する仕組みを実現している．この仕組みによって，NMT は意味的要素の翻訳に非常に優れており，表現の幅が大きく，かつ，流暢な目的言語文を生成できる点が大きな長所であり，その翻訳精度においても SMT を大きく上回っている．ただし，短所として，目的言語文を生成時において，目的言語の

語彙集合 V_y 中のすべての単語を列挙した後，確率を最大化する単語を探索する必要がある．この際，式 (2.4) において，活性化関数 $g(\cdot)$ として softmax 関数を用いるためには，語彙集合 V_y 中のすべての単語について総和をとる必要があり，この計算コストが障壁となって，大規模な語彙へと対応することが困難となっている．そのため，訓練文中の高頻度語のみによってい語彙集合 V_y を構成し，それ以外の単語をすべて未知語トークン $\langle unk \rangle$ に置き換えた後，NMT モデルの訓練を行っており，このことが誤訳の大きな原因の一つとなっている．

第3章　統計的機械翻訳における大規模フレーズ翻訳知識

3.1　コーパス

本論文では，訓練・評価用データとして，二種類の日中対訳特許文および一種類の日英対訳特許文を用いた．第 4 章および第 5 章においては，日本特許情報機構 (JAPIO) から提供された日中対訳特許文の一部である 360 万対訳文を用いた．一方，第 6 章においては，関連研究における翻訳手法との性能比較を行うことを目的として，第 4 回 Asian Translation ワークショップ (WAT2017) の特許翻訳タスクにおいて提供された日中対訳特許文および日英対訳特許文を用いた．

3.1.1　日中対訳特許文

360 万対の日中対訳特許文

この日中対訳特許文は，2004-2012 年発行の日本公開特許広報全文と 2005-2010 年中国特許全文を対象として，以下の手順で得られたものである．

1. 2004-2012 年発行の日本公開特許広報全文と 2005-2010 年中国特許全文を対象として，総計 312,492 件の日中パテントファミリー全文を得る．
2. 日中パテントファミリーに対して，[52] の手法によって日中間で文対応を付ける [1]．
3. 抽出された約 2,400 万件の日中対訳文のうち，スコア降順上位の 360 万文対を抽出する．

100 万対の日中対訳特許文

この日中対訳特許文は，WAT2017 の特許翻訳タスク [2] において配布された日中 100 万件の対訳特許文であり，100 万件の訓練用対訳特許文とは別に，2,000 件の

[1] 文対応付けにおいては，約 170,000 見出し語の中日対訳辞書を用いた．
[2] http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/index.html

開発用対訳特許文，および，2,000 件の評価用対訳特許文も併せて配布されており，本論文でもこれらを用いる．

3.1.2 日英対訳特許文

この日英対訳特許文は，WAT2017 の特許翻訳タスクにおいて配布された日英 100 万件の対訳特許文であり，100 万件の訓練用対訳特許文とは別に，2,000 件の開発用対訳特許文，および，2,000 件の評価用対訳特許文も併せて配布されており，本論文でもこれらを用いる．

3.2 句に基づく統計的機械翻訳モデルを用いた訳語推定

3.2.1 単語対応およびフレーズ翻訳テーブルの作成

単語対応およびフレーズ翻訳テーブルの作成においては，3.1 節で述べた各対訳特許文に対して，句に基づく統計的機械翻訳モデルのツールキットである Moses [22] (バージョン 2.1) を適用することにより，原言語・目的言語の句の組，及び，原言語・目的言語の句が対応する確率を推定し記録したものを作成する．Moses によって単語対応およびフレーズ翻訳テーブルを作成する過程を以下に示す．

1. 文対応データに対する前処理として，単語の数値化，単語のクラスタリング，共起単語表の作成を行う．

2. IBM モデルにより文対応データから単語対応を生成するツール GIZA++ [37] を用いて，原言語から目的言語，目的言語から原言語の両方向に対して最尤な単語対応を得る [3]．

3. 原言語から目的言語，目的言語から原言語両方向の単語対応から，(パラメータ `alignment` を `grow-diag-final-and` とする) ヒューリスティックスを用いて対称な単語対応を得る．

4. 対称な単語対応を用いて，可能な全ての原言語・目的言語の句の組を作成し，各組に対して，「文単位の句対応制約」[4] の条件に対する違反の有無をチェックする (違反しない句の組を有効な対応とみなす)．

[3] 各 IBM モデルの繰り返し回数は，デフォルトのものをそのまま用いた．

[4] 原言語文の単語・形態素列中の単語・形態素を文頭から順に $w_s^1, w_s^2, \ldots, w_s^n$，目的言語語の単語・形態素列中の単語・形態素を文頭から順に $w_t^1, w_t^2, \ldots, w_t^m$ として，原言語句を $t_s (= w_s^p \cdots w_s^{p'})$ とし，目的言語句を $t_t (= w_t^q \cdots w_t^{q'})$ とする．ここで，原言語・目的言語句の組 $\langle t_s.t_t \rangle$ が含まれるある一つの対訳文対 $\langle S_s, S_t \rangle$ 中において得られているあらゆる単語対応 $\langle w_s^i, w_t^j \rangle$ について，「$p \leq i \leq p' \Leftrightarrow q \leq j \leq q'$」が成り立つ場合に，$t_s$ と t_t は対訳文対 $\langle S_s, S_t \rangle$ において「文単位の句対応制約」に違反しない，と定義する．

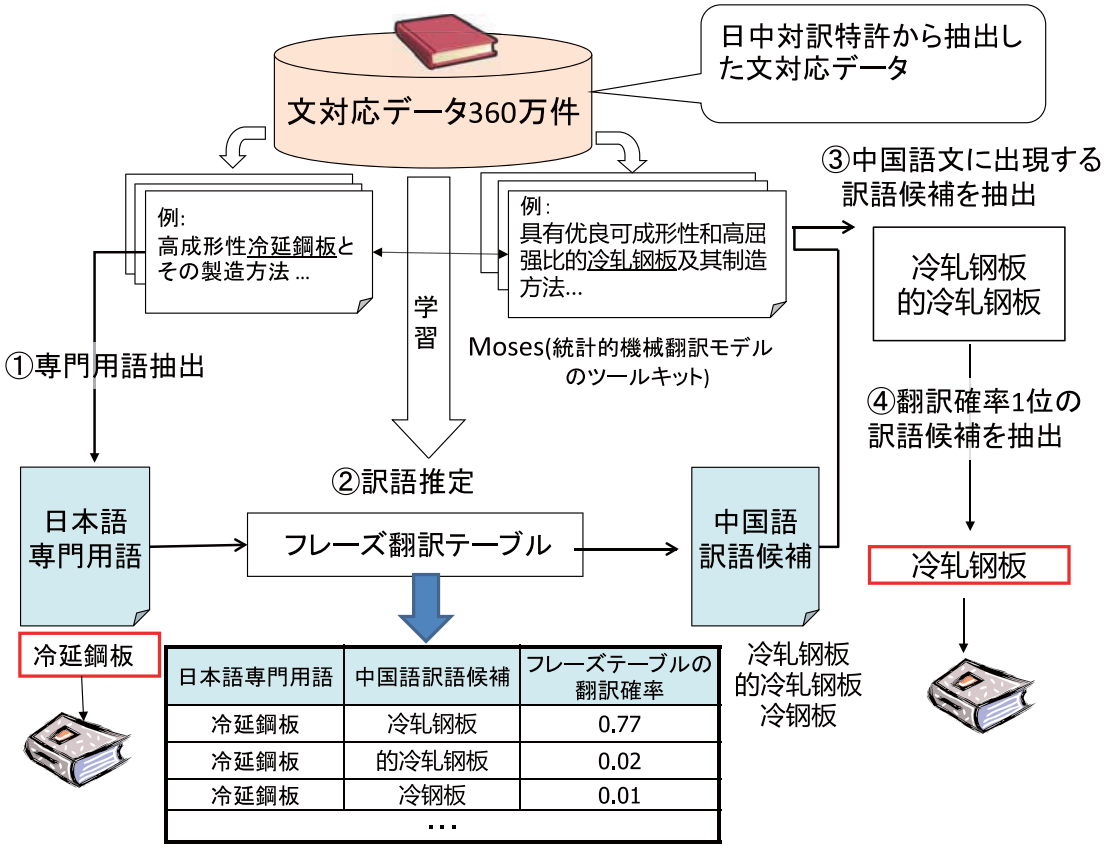


図 3.1: 一組の対訳文およびフレーズ翻訳テーブルを用いた対訳専門用語獲得の流れ

5. 文対応データにおける原言語・目的言語の句の対応数に基づいて，各句の対応に翻訳確率等のパラメータを付与する．

本節の手順の出力としては，上述した手順 3 において得られる対称な単語対応，および，手順 5 において得られる翻訳確率等のパラメータが付与された句の対応を記録したフレーズ翻訳テーブルが得られる．本論文の訳語推定の過程においてはこれらを用いる．また，フレーズ翻訳テーブルを用いて得られる目的言語訳語候補のスコアとしては，句対応の原言語・目的言語翻訳確率 P(目的言語句 t_t | 原言語句 t_s) を用いる．なお，原言語句の見出し語ごとに，目的言語句を翻訳確率の降順に順位付けする．

3.2.2 一組の対訳文およびフレーズ翻訳テーブルを用いた訳語推定

本節では，図 3.1 に沿って，フレーズ翻訳テーブルおよび対訳文を用いて，用語の訳語推定を行う．訳語推定手法において，一組の対訳文を対象として，その対訳文に出現する用語の訳語対 (用語対訳対) を推定する．

原言語用語 t_s に対して，t_s が出現する対訳文のうちの一組 $\langle S_s, S_t \rangle$ およびフレーズ翻訳テーブルを用い，以下の条件をすべて満たす訳語候補 t_t を推定する [5].

(a) t_s の訳語として，フレーズ翻訳テーブルに存在する．

(b) (a) を満たす訳語であり，かつ，与えられた対訳文 $\langle S_s, S_t \rangle$ の目的言語文 S_t に出現する．

(c) (a)，(b) を満たす訳語の中で，フレーズ翻訳テーブルにおける翻訳確率 $P(t_t \mid t_s)$ が最大である．

この手順を経ることにより，各対訳文からは，フレーズ翻訳テーブルに含まれる対訳対のうち，誤りに相当する部分文字列等が削除され，対訳対として適切である可能性の高いものだけが得られる可能性が高くなる．

本節で述べた訳語推定手法は，第 4 章，第 5 章，および，第 6 章の各章において用いる．

3.2.3 一組の対訳文および単語対応を用いた訳語推定

本節では，単語対応および対訳文を用いて，用語の訳語推定を行う．訳語推定手法において，一組の対訳文を対象として，その対訳文に出現する用語の訳語対 (用語対訳対) を推定する．

原言語用語 $t_s = v_s^1 \cdots v_s^n$ (n は原言語用語に含まれた形態素・単語の数) に対して，t_s が出現する対訳文のうちの一組 $\langle S_s, S_t \rangle$ ($S_s = w_s^1 \cdots w_s^N$，$S_t = w_t^1 \cdots w_t^M$，N と M はそれぞれ原言語文と目的言語文に含まれた形態素・単語の数)，および，対訳文対 $\langle S_s, S_t \rangle$ 中の単語対応

$$align(\langle S_s, S_t \rangle) \;=\; \left\{ \langle w_s^i, w_t^j \rangle \middle| 1 \leq i \leq N, 1 \leq j \leq M \right\}$$

を用い，以下の条件をすべて満たす訳語候補 $t_t = v_t^1 \cdots v_t^m$ (m は推定できた訳語候補に含まれる形態素・単語の数) を推定する．

(a) 訳語候補 t_t に含まれる全ての形態素・単語 v_t^y ($y = 1, \cdots, m$) に対して，単語対応 $\langle v_s^x, v_t^y \rangle \in align(\langle S_s, S_t \rangle)$ ($1 \leq x \leq n$) を満たす v_t^x が存在する．

(b) t_t は，(a) を満たす訳語のうち，構成要素となる形態素・単語の数が最大である．

[5] ここでは，Moses を用いたデコーディングによって訳語を推定したのではなく，フレーズ翻訳テーブル中の訳語を直接参照して訳語推定を行う．

(c) t_t は，(a)，(b) を満たす訳語の中で，与えられた対訳文 $\langle S_s, S_t \rangle$ の目的言語文 S_t において，連続する形態素・単語列として出現する

$$t_t = v_t^1 \cdots v_t^m = w_t^k \cdots w_t^{k+m-1} \quad (1 \leq k \leq k+m-1 \leq M)$$

本節で述べた訳語推定手法は，第 6 章において，3.2.2 節の手法では訳語推定できないフレーズに対して，3.2.2 節で述べた訳語推定手法を補完して訳語推定を行うために用いる．

第4章　対訳専門用語の同定

4.1　はじめに

特許文書の翻訳は，他国への特許申請や特許文書の言語横断検索などといったサービスにおいて不可欠である．機械翻訳[1]においても，また，翻訳者による翻訳においても，大規模で正確な対訳辞書は，高い訳質を保証するための必要な情報源である．しかし，人手によって対訳辞書を作成し，継続的に収録語数を増やし辞書を維持・管理していく作業は膨大な時間と労力を要する．そこで，自然言語処理分野においては，多様なテキストデータを情報源として，対訳辞書を自動もしくは半自動的に作成する技術に関する研究が行われてきた．

これまでの研究を大まかに分類すると，初期の研究としては，二言語間で文間の対応がつけられた文対応対訳コーパスを情報源として，対訳文中の共起頻度を用いる [32] がよく研究された．また，文対応対訳コーパスよりも利用可能性の高いコンパラブルコーパスを情報源とする手法 (例えば，[10, 2, 34]) も，初期の時期から近年に至るまでよく研究されている．さらに，近年では，多言語文書を収集する情報源として，ウェブ上の多様な言語・分野・ジャンルの文書を利用する研究も数多く進められている．例えば，複合語である専門用語の構成要素の訳語を連結して訳語の候補を生成する要素合成法，および，ウェブから収集した目的言語の専門分野コーパスを用いて，生成された訳語候補を検証する手法 [45]，検索エンジン等を利用して訳語が併記された文書を収集し，訳語対を獲得する手法訳語対を獲得する手法 [16, 26]，多言語文書である Wikipedia を情報源とする手法 [7] 等がある．それらの研究の中で，特に，専門用語対訳辞書の作成においては，対訳特許文書を情報源として，専門用語対訳対を自動獲得する手法の研究が行われてきた．[35] では，NTCIR-7 の特許翻訳タスク [9] において配布された日英 180 万件の対訳特許文を情報源として専門用語対訳対獲得を行った．この研究では，句に基づく統計的機械翻訳モデル [22] を用いることにより日英対訳文から学習されたフレーズ翻訳テーブル，要素合成法 [45]，Support Vector Machines(SVMs) [54] による分類器学習を用いることによって，専門用語対訳対獲得における適合率 91.9%，

[1]ここでの「機械翻訳」とは，狭義には，「人手によって作成された規則を用いる古典的機械翻訳」のことを指すが，5.4 節末尾において簡単に触れるように，近年の統計的機械翻訳の研究においては，コーパスから獲得した語彙知識を有効に活用して翻訳性能およびカバレージを改善する方式も提案されている．

再現率 70%を達成した.

そこで，本論文では，[35] と同様の考え方に基づき，[52] の手法を適用することによって，3.1.1 節で述べた日中パテントファミリーから抽出した 360 万件の日中対訳特許文を言語資源としてフレーズ翻訳テーブルを学習し，対訳特許文から日中対訳専門用語を獲得する手法を提案する．具体的には，まず，専門用語対訳辞書獲得の情報源として用いる日中対訳文対に対して，句に基づく統計的機械翻訳モデルを学習することにより，フレーズ翻訳テーブルを作成する．次に，対訳特許文およびフレーズ翻訳テーブルを用いた訳語推定を行い，日本語専門用語の中国語訳語候補を獲得する．最後に，獲得した日中対訳専門用語に対して，複数の対訳文から得られる情報を素性として，SVMs [54] を適用する．評価結果として，提案手法により，適合率最大の場合，90%以上の適合率を達成し，F 値最大の場合，80%以上の適合率を達成した.

4.2 句に基づく統計的機械翻訳モデルのフレーズ翻訳テーブル

3.1.1 節で述べた 360 万文の日中対訳特許文に対して，3.2.1 節の手順を適用することにより，日中の句の組および日中の句の組が対応する確率を記録した日中フレーズ翻訳テーブルを作成する．ここで，3.2.1 節の手順 (1) に用いられた対訳文は，IPAdic[2] を用いた MeCab[3] によって形態素解析された形態素単位の日本語文一文に対して，Chinese Penn Treebank を用いた Stanford Word Segment [46] によって形態素解析された形態素単位の中国語文，および，文字単位 [4] の中国語文の 2 種類を用意し，作成されたものである．このような 2 種類の対訳文に対して，独立に Moses を適用することより,「中国語側が形態素単位のフレーズ翻訳テーブル」および「中国語側が文字単位のフレーズ翻訳テーブル」をそれぞれ作成した．ただし，Moses を適用する過程においては，日本語文の形態素数もしくは中国語文の文字数が 101 以上となる日中対訳特許文を除外する [5]．また，日本語フレーズの形態素数の上限，中国語側形態素単位フレーズ翻訳テーブルの中国語形態素数の上限，および，中国語側文字単位フレーズ翻訳テーブルの中国語文字数の上限を，いずれも 15 とした.

表 4.1: 評価対象の日本語専門用語の数

頻度レンジ	日本語名詞句の総数	無作為に抽出した日本語名詞句	日本語専門用語	一般名詞句または区切り位置誤り
1	599,026	90	56	34
2∼5	378,286	90	57	33
6∼10	102,632	90	57	33
11∼15	44,046	90	56	34
16∼20	24,812	90	55	35
21∼30	27,481	90	50	40
31∼50	24,741	90	44	46
51∼100	20,210	90	40	50
101∼200	11,263	90	46	44
201∼500	7,264	90	39	51
501∼1000	2,530	90	38	52
1001∼10,000	2,098	90	28	62
10,000 以上	91	68	12	56
合計	1,244,480	1,148	578	570

4.3 専門用語の対訳対の訓練・評価用集合の作成

4.3.1 訳語推定対象の選定

3.1.1 節で述べた 360 万件の日中対訳特許文を言語資源として，訳語推定対象とする日本語名詞句を以下の手順で選定する．

1. 全 360 万件の対訳文データの日本語文を形態素解析してから，120 万件の日本語名詞句を自動的に抽出する [6]．以下に該当する日本語名詞句は評価対象外とした．

   (i) 語頭または語尾が不適切である日本語名詞句．具体的には，「上記，下記，当該，該，各」が語頭，または「等，毎，上，下，中，内，前，後，側，時，式，的，さ」が語尾である日本語名詞句．

[2] http://sourceforge.jp/projects/ipadic/

[3] http://mecab.sourceforge.net

[4] ただし，連続する数字とアルファベットは，それぞれ一個のトークンとして扱う．

[5] この結果，4,291 件の日中対訳特許文が除外され，残った 3,595,709 件の日中対訳特許文を評価実験に用いた．

[6] 日本語文の形態素解析結果に対して，名詞・接頭辞・接尾辞・未知語のいずれかの品詞の形態素からなる最長の形態素列に加えて，アルファベットの列が連接することを許容したものを日本語名詞句として抽出する．

表 4.2: 日本語名詞句の分類および例

日本語名詞句の分類	例
専門用語	有機 化合 物 膜
	希土類 焼 結 磁石
一般名詞句	製造 方法
	化合 物
区切り位置誤り	二 次 加
	式 リフト クレーン

(ii) 記号および数字を含む日本語名詞句．例えば「メタンガス濃縮装置 M1」，「クロック 信号 /」,「3 通り」．

2. 抽出した全日本語名詞句に対して，表 4.1 に示す 13 の頻度レンジを設定した．以降の手順においては，13 の各頻度レンジごとに日本語名詞句の選定を行った．

3. 各頻度レンジに対して，無作為に 90 例の日本語名詞句 (合計 1,148 例)[7] を抽出し，評価対象の候補とする．

4. ここで，これらの日本語名詞句 1,148 例を大別すると，表 4.1 および表 4.2 に示すように，日本語専門用語，および，それ以外の一般名詞句，区切り位置誤り，あるいはそれらの混在したものに分類することができる．ただし，ここで「区切り位置誤り」と呼んでいるものは，主として，品詞判定の誤りによって抽出された専門用語の部分文字列である．例えば，表 4.2 の区切り位置誤り例のうち,「二 次 加」は「二 次 加 圧」の部分文字列である．また,「式 リフト クレーン」は「移動 式 リフト クレーン」部分文字列である．本論文では，日本語名詞句 1,148 例から，表 4.1 における 578 例の日本語専門用語のみを人手で選定して，評価対象の日本語名詞句とした．

4.3.2 訓練・評価用集合の作成手順

以下では，対訳特許文およびフレーズ翻訳テーブルを用いて，専門用語対訳対の訓練・評価用集合を作成する流れを示す．以下の手順においては，訓練・評価用の専門用語対訳対候補を生成するため，訳語推定対象として日本語用語を選定し，そこから専門用語対訳対候補集合を生成する．

[7]頻度レンジ「10,000 以上」の場合のみ，68 例のみを抽出した．

表 4.3: 訳語候補集合における正例・負例数の内訳

		正例数		負例数		総数	
中国語側が形態素単位のフレーズ翻訳テーブルを用いた場合	形態素単位の集合のみ含まれる	115	1,537	527	1,002	642	2,533
	文字単位の集合と共通	1,416		475		1,891	
中国語側が文字単位のフレーズ翻訳テーブルを用いた場合	文字単位の集合のみ含まれる	44	1,255	376	837	420	2,092
	形態素単位の集合と共通	1,211		461		1,672	

(1) まず，入力の日本語用語 t_J に対して，日中対訳特許文の中から t_J が出現する対訳文 $\langle S_J^i, S_C^i \rangle$ $(i = 1, \ldots, n_1$，ただし，n_1 は日本語文に t_J が含まれる日中対訳特許文の数) をすべて収集する．

(2) 次に，収集した各対訳文に対して，フレーズ翻訳テーブルを参照して 3.2.2 節で述べた手法[8] を適用することにより，t_J の中国語訳語を推定する．

(3) 推定された中国語訳語を t_C^j $(j = 1, \ldots, n_2$，ただし，n_2 は推定された中国語訳語の数) として，すべての対訳対 $\langle t_J, t_C^j \rangle$ を生成する．

(4) 最後に，生成された対訳専門用語に対して，専門用語の対訳対として適切か否かの判定を人手によって行い，対訳専門用語を正例と負例に分類して，訓練・評価用の専門用語対訳対候補集合を選定する．

4.3.3 訓練・評価用集合の作成結果

前節の手順に従い，専門用語対訳の訓練・評価用集合の作成を行った結果を以下に示す．まず，4.3.1 節において得られた日本語専門用語 578 例に対して，前節の手順 (2) に従い，4.2 節で作成された「中国語側が形態素単位のフレーズ翻訳テーブル」および「中国語側が文字単位のフレーズ翻訳テーブル」をそれぞれ独立に用いて，訳語推定を行った．それらの対訳専門用語候補集合における専門用語の

[8]複数の訳語候補が翻訳確率 1 位となる場合には，そのすべての訳語候補を順位 1 位とみなして，中国語訳語推定結果とする．

総数を表 4.3 に示す[9].「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合では，2,533 例の対訳専門用語が生成され,「中国語側が文字単位のフレーズ翻訳テーブル」を用いた場合では，2,092 例の対訳専門用語が生成された．次に，前節の手順 (4) に従って，専門用語の対訳対として適切か否かの判定を行い，生成された対訳専門用語を正例と負例に分類した結果，表 4.3 に示すように,「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合では全 2,533 例中，正例が 1,531 例，負例が 1,002 例となり,「中国語側が文字単位のフレーズ翻訳テーブル」を用いた場合では，全 2,092 例中，正例が 1,255 例，負例が 837 例となった．更に，獲得できた日中対訳専門用語に対して，日本語専門用語の頻度レンジ (jf) および日中間の共起頻度レンジ (jcf) ごとに正例の割合を求めた結果を表 4.4 と表 4.5 に示す．また，高・中・低の各頻度レンジにおける日中対訳専門用語の正例の例を表 4.6 に示す．

4.4 複数の日中対訳文からの情報を素性とするSVMの適用

本節では，SVM を用いて対訳専門用語を同定する手法について述べる．

4.4.1 SVM の適用

まず，4.3 節で生成した日中対訳専門用語を全事例集合として，正例と負例の数が均等になるように，互いに素な部分集合に 10 分割した．ただし，同一の日本語専門用語を共有する複数の日中対訳専門用語は，同一の部分集合に含めた．本論文では，TinySVM[10] を利用して，評価実験を行った．カーネル関数としては，一次多項式カーネルおよび二次多項式カーネルを評価し，相対的に高い性能を達成できた二次多項式カーネルを用いた．また，SVM の分離平面から評価事例までの距離を信頼度とし，正例判定における信頼度の下限を設定した．具体的には，10 個の部分集合のうち，8 個を訓練用事例集合として SVM の訓練を行い，残りのうちの 1 個を調整用事例集合とし，最後の 1 個を評価用事例集合とした．調整用事例集合を用いたパラメータの調整においては，分離平面から評価用事例までの距離の下限のパラメータの調整を行った．本論文では，日中対訳専門用語の適合率および F 値を最大化する調整を行った．ただし，適合率を最大化する場合は，再現率が 60%以上となるという条件のもとで，パラメータの調整を行った．以上の

[9]表 4.3 では，中国語側の形態素解析誤りが原因で，同一の文字列に対する形態素分割のパターンが 2 通り以上出現する場合があるため,「文字単位の集合と共通」となる対訳対数が,「形態素単位の集合と共通」となる対訳対数よりも多くなっている．

[10]http://chasen.org/~taku/software/TinySVM

表 4.4: 日本語専門用語の頻度 (jf) の各レンジおよび日中間共起頻度 (jcf) の各レンジごとの正例割合 (正例数 / (負例数+正例数)) (「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合)

	jf= 1	2≦jf ≦5	6≦jf ≦10	11≦jf ≦15	16≦jf ≦20	21≦jf ≦30	31≦jf ≦50	51≦jf ≦100	101≦jf ≦200	201≦jf ≦500	501≦jf ≦1,000	1,001≦jf ≦10,000	10,001≦jf	総計
jcf_m=1	36/58 =62.1%	27/35 =77.1%	20/26 =76.9%	27/43 =62.8%	31/55 =56.4%	21/37 =56.8%	25/45 =55.6%	26/51 =51.0%	36/102 =35.3%	32/85 =37.6%	52/149 =34.9%	21/99 =21.2%	4/57 =7.0%	358/842 =42.5%
2≦jcf_m ≦5		56/57 =98.2%	32/36 =88.9%	17/23 =73.9%	25/37 =67.6%	18/27 =66.7%	28/35 =80.0%	27/50 =54.0%	41/72 =56.9%	33/62 =53.2%	35/89 =39.3%	37/82 =45.1%	5/52 =9.6%	354/622 =56.9%
6≦jcf_m ≦10			36/37 =97.3%	25/25 =100%	21/24 =87.5%	22/24 =91.7%	21/23 =91.3%	10/20 =50.0%	19/26 =73.1%	17/27 =62.9%	31/53 =58.5%	16/27 =59.3%	8/27 =29.6%	226/313 =72.2%
11≦jcf_m ≦15				29/30 =96.7%	16/17 =94.1%	11/11 =100%	11/11 =100%	4/5 =80.0%	11/12 =91.7%	9/17 =52.9%	10/15 =66.7%	11/23 =47.8%	1/11 =9.1%	113/152 =74.3%
16≦jcf_m ≦20					20/20 =100%	17/17 =100%	4/4 =100%	2/3 =66.7%	9/10 =90.0%	3/4 =75.0%	10/17 =58.8%	12/21 =57.1%	2/9 =22.2%	79/105 =75.2%
21≦jcf_m ≦30						19/19 =100%	10/11 =90.9%	3/4 =75.0%	7/12 =58.3%	12/15 =80%	10/12 =83.3%	3/8 =37.5%	6/10 =60.0%	67/91 =73.6%
31≦jcf_m ≦50							20/21 =95.2%	18/19 =94.7%	12/13 =92.3%	11/14 =78.6%	9/14 =64.3%	3/11 =27.3%	4/11 =36.4%	78/104 =75.0%
51≦jcf_m ≦100								21/22 =95.5%	22/22 =100%	6/11 =54.5%	11/12 =91.7%	7/13 =53.8%	5/14 =35.7%	72/94 =76.6%
101≦jcf_m ≦200									21/21 =100%	18/18 =100%	7/8 =87.5%	9/15 =60.0%	7/14 =50.0%	62/76 =81.6%
201≦jcf_m ≦500										20/21 =95.2%	17/18 =94.4%	6/10 =60.0%	3/4 =75.0%	46/53 =86.8%
501≦jcf_m ≦1,000											26/26 =100%	14/16 =87.5%	1/2 =50.0%	41/44 =93.2%
1,001≦jcf_m ≦10,000												18/18 =100%	7/9 =77.8%	25/27 =92.6%
10,001≦jcf_m													10/10 =100%	10/10 =100%
総計	36/58 =62.1%	83/92 =90.2%	88/99 =88.9%	98/121 =81.0%	113/153 =73.9%	108/135 =80.0%	119/150 =79.3%	111/174 =63.8%	178/290 =61.4%	159/275 =57.8%	218/413 =52.8%	157/343 =45.8%	63/230 =27.4%	1,521/2,533 =60.4%

表 4.5: 日本語専門用語の頻度 (jf) の各レンジおよび日中間共起頻度 (jcf) の各レンジごとの正解割合 (正例数 / (負例数+正例数)) (「中国語側が文字単位のフレーズ翻訳テーブル」を用いた場合)

	jf=1	2≦jf≦5	6≦jf≦10	11≦jf≦15	16≦jf≦20	21≦jf≦30	31≦jf≦50	51≦jf≦100	101≦jf≦200	201≦jf≦500	501≦jf≦1,000	1,001≦jf≦10,000	10,001≦jf	総計
jcf_c=1	37/62 =59.7%	20/26 =77.0%	12/19 =63.2%	20/33 =60.7%	21/37 =56.8%	11/20 =55.1%	24/40 =60.0%	19/49 =38.8%	25/70 =35.8%	22/76 =29.0%	36/121 =29.8%	8/69 =11.6%	1/28 =3.6%	256/650 =39.4%
2≦jcf_c≦5		55/58 =94.9%	24/28 =85.8%	12/15 =80.0%	16/24 =66.7%	14/23 =60.9%	22/32 =68.8%	21/42 =50.0%	28/55 =51.0%	23/55 =41.9%	22/86 =25.6%	34/79 =43.1%	1/28 =3.6%	272/525 =51.9%
6≦jcf_c≦10			41/43 =95.4%	16/16 =100.0%	14/16 =87.5%	14/16 =87.5%	16/18 =88.9%	9/15 =60.0%	15/25 =60.0%	12/19 =63.2%	23/39 =59.0%	14/31 =45.2%	1/17 =5.9%	175/255 =68.7%
11≦jcf_c≦15				38/40 =95.0%	13/14 =92.9%	9/9 =100.0%	6/7 =85.8%	4/6 =66.7%	8/8 =100.0%	10/15 =66.7%	7/10 =70.0%	6/17 =35.3%	1/6 =16.7%	102/132 =77.3%
16≦jcf_c≦20					31/31 =100.0%	7/7 =100.0%	8/8 =100.0%	1/2 =50.0%	7/8 =87.5%	4/6 =66.7%	8/9 =88.9%	7/11 =63.7%	1/4 =25.1%	74/86 =86.1%
21≦jcf_c≦30						32/32 =100.0%	6/8 =75.0%	3/4 =75.0%	6/10 =60.0%	5/7 =71.5%	6/9 =66.7%	2/8 =25.1%	1/5 =20.1%	61/83 =73.5%
31≦jcf_c≦50							25/26 =96.2%	15/17 =88.3%	7/9 =77.8%	7/11 =63.7%	8/10 =80.0%	3/5 =60.0%	3/13 =23.1%	68/91 =74.8%
51≦jcf_c≦100								24/25 =96.0%	21/21 =100.0%	7/9 =77.8%	8/8 =100.0%	6/6 =100.0%	4/14 =28.6%	70/83 =84.4%
101≦jcf_c≦200									25/25 =100.0%	17/18 =94.5%	5/6 =83.4%	7/8 =87.5%	4/8 =50.0%	58/65 =89.3%
201≦jcf_c≦500										23/24 =95.9%	10/10 =100.0%	6/6 =100.0%	3/4 =75.0%	42/44 =95.5%
501≦jcf_c≦1,000											31/31 =100.0%	5/5 =100.0%	1/1 =100.0%	37/37 =100.0%
1,001≦jcf_c≦10,000												24/24 =100.0%	6/7 =85.8%	30/31 =96.8%
10,001≦jcf_c													10/10 =100.0%	10/10 =100.0%
総計	37/62 =59.7%	75/84 =89.3%	77/90 =85.6%	86/104 =82.7%	95/122 =77.9%	87/107 =81.4%	107/139 =77.0%	96/160 =60.0%	142/231 =61.5%	130/240 =54.2%	164/339 =48.4%	122/269 =45.4%	37/145 =25.6%	1255/2092 =60.0%

表 4.6: 日本語専門用語の頻度 (jf) および日中間共起頻度 (中国語側が形態素単位の場合 jcf_m，中国語側が文字単位の場合 jcf_c) の低・中・高の各レンジごとの正例の例

	低頻度レンジ ($1 \leqq jf \leqq 15$)	中頻度レンジ ($16 \leqq jf \leqq 100$)	高頻度レンジ ($101 \leqq jf$)
低頻度レンジ ($1 \leqq jcf_m, jcf_c \leqq 15$)	<水車羽根型発電装置, 水轮机叶片型发电装置> (jf=1, jcf_m=1, jcf_c=1)	<リンパ球増殖, 淋巴细胞增殖> (jf=21, jcf_m=10, jcf_c=10)	<スクリーン装置, 筛装置> (jf=104, jcf_m=11, jcf_c=14)
中頻度レンジ ($16 \leqq jcf_m, jcf_c \leqq 100$)		<高圧水素ガス, 高压氢气> (jf=37, jcf_m=37, jcf_c=37)	<窒素化合物, 氮化合物> (jf=112, jcf_m=91, jcf_c=97)
高頻度レンジ ($101 \leqq jcf_m, jcf_c$			<反応混合物, 反应混合物> (jf=30,620, jcf_m=29,803, jcf_c=29,860)

訓練，調整，評価の手順を 10 通り繰り返し，その評価結果のマイクロ平均を算出し，日中対訳専門用語判定の性能評価を行った．

4.4.2 素性

本論文の手法において用いた素性は，表 4.7 に示すように，大きく，単言語素性と二言語素性に分けられる．単言語素性としては，言語資源として用いられた全 360 万対訳文における日本語専門用語 t_J の頻度 (f_1)，および，中国語専門用語 t_C の頻度 (f_2) を用いた．

二言語素性としては，フレーズ翻訳テーブルによって各訳語候補に付与された翻訳確率の素性 (f_3)，および，フレーズ翻訳テーブル中での，同一日本語専門用語に対する翻訳確率の降順順位の素性 (f_4)[11] を用いた．その他，全 360 万対訳文における日中対訳専門用語 $\langle t_J, t_C \rangle$ の共起頻度の素性 (f_5) を用いた．また，日本語専門用語の頻度と日中対訳専門用語の共起頻度の差の素性 (f_6)，同一日本語専門用語に対する中国語訳語候補の数の素性 (f_7)，対訳文およびフレーズ翻訳テーブルを用いた訳語推定を行う際の文単位の句対応制約 [35] の違反のない対訳文の割合の素性 (f_8) を用いた．さらに，日本語専門用語 t_J と中国語専門用語 t_C との間で，要素合成法に基づき，フレーズ翻訳テーブルを用いて構成要素の間の翻訳を行った際の翻訳確率の積を求めた値を，要素合成法の翻訳確率の素性 (f_9) として用いた．ただし，フレーズ翻訳テーブルを用いて構成要素を翻訳する際の翻訳確率に対して下限値 (本論文では 0.005) を設けるとともに，日本語専門用語 t_J および中国語専門用語 t_C を構成要素に分割する際に二通り以上の分割の仕方が可能

[11] ここで素性として用いる順位としては，本論文の手順においては訳語候補としては残らないが，フレーズ翻訳テーブル中には含まれる訳語候補を含む順位を用いる．

表 4.7: 日中対訳専門用語同定のための素性

分類	素性名	定義
単言語素性	f_1: 日本語専門用語の頻度	日本語専門用語が属する頻度レンジの番号 (1~13)
	f_2: 中国語専門用語の頻度	中国語専門用語が属する頻度レンジの番号 (1~13)
二言語素性	f_3: 翻訳確率	フレーズ翻訳テーブルにおける翻訳確率
	f_4: 訳語候補の順位	同一日本語専門用語に対する訳語候補の順位 (フレーズ翻訳テーブルにおける順位)
	f_5: 日中対訳専門用語の頻度	日中対訳専門用語が属する頻度レンジの番号 (1~13)
	f_6: 日本語専門用語と対訳共起頻度の頻度差	日本語専門用語の頻度 – 日中対訳共起頻度が上限値 (本論文では 105) 以下の場合 1，上限値を超える場合 0
	f_7: 訳語数	同一の日本語専門用語に対する中国語訳語候補数
	f_8: 文単位の句対応制約の違反のない対訳文の割合	$f_8 = \frac{\text{文単位の句対応制約の違反のない対訳文対の数}}{\text{当該日中対訳専門用語の共起頻度}}$
	f_9: 要素合成法の翻訳確率	要素合成法により出力された訳語候補の翻訳確率

な場合には，それぞれの分割の仕方における翻訳確率の平均を用いた.

4.4.3 評価結果

訳語推定性能の評価結果を表 4.8 に示す．ベースラインとして「4.3 節で生成した日中対訳専門用語全事例が正しいと判定する」という規則を用いた.「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合では，ベースライン手法での適合率は 60.4%，再現率は 100%，F 値は 75.3%で,「中国語側が文字単位のフレーズ翻訳テーブル」を用いた場合では，ベースライン手法での適合率は 60.0%，再現率は 100%，F 値は 75.0%となった．一方,「中国語側が形態素単位」の場合，SVM を用いて適合率を最大化する調整を行った場合の適合率は 94.2%，F 値を最大化する調整を行った場合の F 値は 82.3%となった.「中国語側が文字単位」の場合，SVM を用いて適合率を最大化する調整を行った場合の適合率は 93.9%，F 値を最大化する調整を行った場合の F 値は 83.7%となった.

一方，表 4.9 と表 4.10 においては，SVM を用いて適合率を最大化する調整を行った場合について，日本語専門用語の頻度 (jf) の各レンジ，および，日中間共起頻度 (jcf) の各レンジごとの適合率・再現率・F 値を示す．これらの結果から，いずれの頻度レンジにおいても高い適合率が達成できており，頻度レンジを問わ

表 4.8: 日中対訳専門用語同定の適合率・再現率・F 値 (%)

(a) 中国語側が形態素単位のフレーズ翻訳テーブルを用いた場合

		適合率	再現率	F 値
ベースライン		60.4	100	75.3
SVM	適合率最大	94.2	60.0	73.3
	F 値最大	75.2	90.9	82.3

(b) 中国語側が文字単位のフレーズ翻訳テーブルを用いた場合

		適合率	再現率	F 値
ベースライン		60.0	100	75.0
SVM	適合率最大	93.9	59.0	72.5
	F 値最大	80.6	87.2	83.7

ず安定した適合率が実現できていることが分かる．その一方で，対訳専門用語同定の難易度が相対的に高い頻度レンジにおいては，高い適合率を保つため再現率が犠牲になっていることが分かる．具体的には，日本語専門用語の頻度レンジと再現率の相関に関しては，高頻度になるほど再現率が低下している．この原因は，主として，日本語専門用語の頻度が大きくなるほど，日中対訳特許文における中国語訳語の多様性が増すとともに，訳語候補生成過程において負例となる対訳対が多く生成される点にある．逆に，日中間共起頻度の頻度レンジと再現率の相関に関しては，低頻度になるほど再現率が低下している．これは，日中間共起頻度が大きくなるほど，当該対訳専門用語用語適切な対訳対である可能性が高くなり，正例の割合が高くなることが原因である．

「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合について，SVM によって選定された訳語候補の正解例および誤り例を表 4.11 に示す．

表 4.11(a)「SVM による正解例」のうち，日本語専門用語「水性/樹脂/組成/物」および中国語訳語“水性/树脂/组合物”の組においては，フレーズ翻訳テーブルにおける翻訳確率 (素性 f_3) は 0.95 であり，フレーズ翻訳テーブルにおける中国語訳語の順位 (素性 f_4) は 1 位である．また,「水性/樹脂/組成/物」の訳語数 (素性 f_7) は 1 であり，要素合成法により出力された訳語候補の翻訳確率 (素性 f_9) は 0.73 である．これらの素性の効果によって，SVM により正解の対訳専門用語であると判定できた．一方，日本語専門用語「気/液/分離/器」および中国語訳語“气液/反应器”の組においては，フレーズ翻訳テーブルにおける翻訳確率および要素合成法を用いた確率の素性 (f_3，f_9) が十分に小さい値となった．これらの素性の効果によって，当該日中対訳専門用語を誤りの対訳専門用語であると判定できた．

表 4.9: 日本語専門用語の頻度 (jf) の各レンジおよび日中間共起頻度 (jcf) の各レンジごとの適合率・再現率・F 値 (%) (「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合)

	jf= 1	2≦jf≦5	6≦jf≦10	11≦jf≦15	16≦jf≦20	21≦jf≦30	31≦jf≦50	51≦jf≦100	101≦jf≦200	201≦jf≦500	501≦jf≦1,000	1,001≦jf≦10,000	10,001≦jf	総計
jcf_m=1	78.8/72.3/75.4	100/40.8/57.9	100/45.0/62.1	100.0/37.1/54.1	92.4/38.8/54.6	80.0/19.1/30.8	100/16.1/27.6	100/19.3/32.3	85.8/16.7/28.0	100/9.4/17.2	100/5.8/11.0	100/4.8/9.1	0/0/0	89.6/26.3/40.7
2≦jcf_m≦5		98.1/91.1/94.5	96.3/81.3/88.2	92.4/70.6/80.0	87.5/56.1/68.3	84.7/61.2/71.0	93.4/50.0/65.2	68.8/40.8/51.2	93.4/34.2/50.1	100.0/9.1/16.7	100/14.3/25.1	100/13.6/23.9	0/0/0	92.3/46.9/62.2
6≦jcf_m≦10			100/97.3/98.6	100/100/100	100/95.3/97.6	100/91.0/95.3	94.2/76.2/84.3	60.0/30.1/40.0	100/52.7/69.0	83.4/29.5/43.5	90.0/29.1/44.0	100/37.5/54.6	100/25.1/40.0	96.8/66.9/79.1
11≦jcf_m≦15				100/100/100	94.2/100/97.0	100/100/100	100/91.0/95.3	100/50.0/66.7	100/91.0/95.3	77.8/77.8/77.8	66.7/40.0/50.0	60.0/27.3/37.5	0/0/0	93.0/81.5/86.8
16≦jcf_m≦20					100/100/100	100/94.2/97.0	100/75.0/85.8	50.0/50.0/50.0	83.4/55.6/66.7	100/33.4/50.0	85.8/60.0/70.6	80.0/66.7/72.8	0/0/0	92.4/76.0/83.4
21≦jcf_m≦30						100/100/100	90.0/90.0/90.0	75.0/100/85.8	62.5/71.5/66.7	100/66.7/80.0	100/100/100	100/100/100.0	100/50.0/66.7	92.1/86.6/89.3
31≦jcf_m≦50							95.3/100/97.6	94.5/94.5/94.5	100/75.0/85.8	85.8/50.0/63.2	100/88.9/94.2	100/100/100	50.0/25.1/33.4	94.2/82.1/87.7
51≦jcf_m≦100								95.5/100/97.7	100/91.0/95.3	71.5/83.4/77.0	100/100/100	75.0/42.9/54.6	80.0/80.0/80.0	92.8/88.9/90.8
101≦jcf_m≦200									100/100/100	100/94.5/97.2	100/71.5/83.4	83.4/55.6/66.7	100/85.8/92.4	98.2/87.1/92.4
201≦jcf_m≦500										100/100/100	100/100/100	100/100/100	100/33.4/50.0	100/95.7/97.8
501≦jcf_m≦1,000											100/96.2/98.1	100/92.9/96.3	100/100/100	100/95.2/97.5
1,001≦jcf_m≦10,000												100/94.5/97.2	100/71.5/83.4	100/88.0/93.7
10,001≦jcf_m													100/100/100	100/100/100
総計	78.8/72.3/75.4	98.5/74.7/85.0	98.6/79.6/88.1	98.8/77.6/86.9	95.4/72.6/82.5	96.5/75.0/84.4	95.0/63.9/76.4	85.2/56.8/68.2	94.4/56.2/70.5	92.5/46.0/61.4	96.3/47.3/63.4	92.5/46.5/61.9	91.7/52.4/66.7	94.2/60.0/73.3

表 4.10: 日本語専門用語の頻度 (jf) の各レンジおよび日中間共起頻度 (jcf) の各レンジごとの適合率・再現率・F 値 (%) (「中国語側が文字単位のフレーズ翻訳テーブル」を用いた場合)

	jf= 1	2≦jf≦5	6≦jf≦10	11≦jf≦15	16≦jf≦20	21≦jf≦30	31≦jf≦50	51≦jf≦100	101≦jf≦200	201≦jf≦500	501≦jf≦1,000	1,001≦jf≦10,000	10,001≦jf	総計
jcf_c=1	100/27.1 /42.6	100/25.1 /40.0	100/25.1 /40.0	100/35.0 /51.9	100/23.9 /38.5	50.0/9.1 /15.4	100/4.2 /8.0	33.4/5.3 /9.1	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	89.2/12.9 /22.6
2≦jcf_c≦5		97.9/83.7 /90.2	93.4/58.4 /71.8	85.8/50.0 /63.2	90.0/56.3 /69.3	77.8/50.0 /60.9	100/18.2 /30.8	85.8/28.6 /42.9	75.0/10.8 /18.8	0/0/0	0/0/0	0/0/0	0/0/0	91.4/35.0 /50.6
6≦jcf_c≦10			100/95.2 /97.5	100/93.8 /96.8	100/85.8 /92.4	92.9/92.9 /92.9	92.4/75.0 /82.8	80.0/44.5 /57.2	75.0/20.1 /31.6	100/16.7 /28.6	100/21.8 /35.8	100/35.8 /52.7	0/0/0	96.5/62.9 /76.2
11≦jcf_c≦15				100/97.4 /98.7	92.9/100 /96.3	100/100 /100	100/66.7 /80.0	50.0/50.0 /50.0	100/62.5 /77.0	85.8/60.0 /70.6	83.4/71.5 /77.0	71.5/83.4 /77.0	0/0/0	91.5/84.4 /87.8
16≦jcf_c≦20					100/96.8 /98.4	100/100 /100	100/87.5 /93.4	50.0/100 /66.7	87.5/100 /93.4	80.0/100 /88.9	100/100 /100	83.4/71.5 /77.0	0/0/0	94.6/93.3 /93.9
21≦jcf_c≦30						100/96.9 /98.5	75.0/100 /85.8	75.0/100 /85.8	66.7/100 /80.0	100/80.0 /88.9	85.8/100 /92.4	66.7/100 /80.0	100/100 /100	88.1/96.8 /92.2
31≦jcf_c≦50							100/96.0 /98.0	87.5/93.4 /90.4	100/85.8 /92.4	71.5/71.5 /71.5	80.0/100 /88.9	66.7/66.7 /66.7	50.0/33.4 /40.0	88.3/88.3 /88.3
51≦jcf_c≦100								95.9/95.9 /95.9	100/95.3 /97.6	75.0/85.8 /80.0	100/100 /100	100/100 /100	100/50.0 /66.7	95.6/92.9 /94.3
101≦jcf_c≦200									100/96.0 /98.0	100/100 /100	100/100 /100	100/85.8 /92.4	100/50.0 /66.7	100/93.2 /96.5
201≦jcf_c≦500										100/87.0 /93.1	100/90.0 /94.8	100/100 /100	100/66.7 /80.0	100/88.1 /93.7
501≦jcf_c≦1,000											100/96.8 /98.4	100/100 /100	100/100 /100	100/97.3 /98.7
1,001≦jcf_c≦10,000												100/100 /100	100/50.0 /66.7	100/90.0 /94.8
10,001≦jcf_c													100/100 /100	100/100 /100
総計	100/27.1 /42.6	98.1/68.0 /80.4	98.3/72.8 /83.6	98.5/75.6 /85.6	97.2/72.7 /83.2	94.5/78.2 /85.6	95.1/54.3 /69.1	83.1/56.3 /67.1	91.4/52.2 /66.4	91.5/49.3 /64.1	94.4/51.3 /66.5	93.0/54.1 /68.4	91.7/59.5 /72.2	93.9/59.0 /72.5

表 4.11: SVM による正解例および不正解例 (「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合)

(a) SVM による正解例

日本語専門用語	中国語専門用語	素性 f_1	素性 f_3	素性 f_4	素性 f_5	素性 f_7	素性 f_9	人手による判断	SVMによる判断
水性/樹脂/組成/物	水性/树脂/组成物	$11 \leq jf \leq 15$	0.95	1	$11 \leq jcf_m \leq 15$	1	0.73	正解	正解
気/液/分離/器	气液/反应器	$1{,}001 \leq jf \leq 10{,}000$	0.008	9	$jcf_m = 1$	13	0	誤り	誤り

(b) SVM による不正解例

日本語専門用語	中国語専門用語	素性 f_1	素性 f_3	素性 f_4	素性 f_5	素性 f_7	素性 f_9	人手による判断	SVMによる判断
生物/処理/反応/槽	生物/处理/反应/槽中	$21 \leq jf \leq 30$	0.06	3	$2 \leq jcf_m \leq 5$	5	0.05	誤り	正解
非/晶/質/シリコン	非晶质/硅	$501 \leq jf \leq 1{,}000$	0.05	7	$jcf_m = 1$	13	0	正解	誤り

表 4.11(b)「SVM による不正解例」のうち，日本語専門用語「生物/処理/反応/槽」および中国語訳語“生物/处理/反应/槽中”の組においては，素性 (f_3,f_4,f_7,f_9) の値が原因となって，誤りの対訳対を正解訳語と判定されてしまった．逆に，日本語専門用語「非/晶/質/シリコン」および中国語訳語“非晶质/硅” の組においては，要素合成法を用いた翻訳確率の素性 (f_9) が 0 となったことが原因で，SVM によって誤りの対訳専門用語であると判定されてしまった．これら二組の例においては，いずれも，中国語側の形態素解析の誤りの影響が原因となって，素性の値は正確に対訳専門用語の特徴を表すことができず，SVM による判定結果が不正解となっている．一つ目の例においては，日本語専門用語中の構成要素「槽」に対応するべき中国語側の形態素は“槽”であるが，中国語文の形態素解析結果において,「の中」を意味する“中”と“槽”が分割されなかったため，中国語訳語が“生物/处理/反应/槽中”となってしまった．一方，二つ目の例では，日本語側が「非/晶/質」と一文字ごとに一形態素へと分割されたのに対して，中国語側が“非晶质” と三文字が一形態素に連結された形態素解析結果となったため，要素合成法において構成要素に分割して訳語を推定することができず，要素合成法を用いた翻訳確率が 0 となってしまった．これらの二例における誤りを回避するためには，中国語側に一文字を一単語として学習した文字単位フレーズ翻訳テーブルを併用し，この文字単位フレーズ翻訳テーブルを扱うための素性を新たに導入する必要があると考えられる．

次に,「中国語側が文字単位のフレーズ翻訳テーブル」を用いた場合について，SVM によって選定された訳語候補の正解例および誤り例を表 4.12 に示す．

表 4.12(b)「SVM による不正解例」のうち，日本語専門用語「カバー/絶縁/層」および中国語訳語 “盖/绝/缘/层” の組においては，素性 f_3，f_4，および，f_9 の値が相対的に高く，素性 f_7 の値が相対的に小さいことが原因となって，誤りの対訳対が正解訳語と判定されてしまった．実際に，中国語用語 “盖” は日本語用語「カバー」

表 4.12: SVM による正解例および不正解例 (「中国語側が文字単位のフレーズ翻訳テーブル」を用いた場合)

(a) SVM による正解例

日本語専門用語	中国語専門用語	素性 f_1	素性 f_2	素性 f_3	素性 f_4	素性 f_7	素性 f_9	人手による判断	SVMによる判断
置換/基	取/代/基	$1{,}000 \leqq jf \leqq 10{,}000$	0.8	1	$1{,}000 \leqq jcf_c \leqq 10{,}000$	6	0.12	正解	正解
気/液/分離/器	气/液/反/应/器	$1{,}000 \leqq jf \leqq 10{,}000$	0.0008	14	$jcf_c = 1$	14	0	誤り	誤り

(b) SVM による不正解例

日本語専門用語	中国語専門用語	素性 f_1	素性 f_2	素性 f_3	素性 f_4	素性 f_7	素性 f_9	人手による判断	SVMによる判断
カバー/絶縁/層	盖/绝/缘/层	$501 \leqq jf \leqq 1{,}000$	0.05	3	$21 \leqq jcf_c \leqq 30$	10	0.57	誤り	正解
駆動/回路	驱/动/器/电/路	$10{,}001 \leqq jcf$	0.006	3	$101 \leqq jcf_c \leqq 200$	5	0	正解	誤り

の正解中国語訳語 “覆盖” の部分列で,「中国語側が文字単位のフレーズ翻訳テーブル」を用いた場合では，当該対訳対以外にも，⟨ カバー/絶縁/層，覆/盖/绝/缘/层 ⟩ が生成されて，SVM によって正解と判定された．逆に,「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合では，対訳専門用語 ⟨ カバー/絶縁/層，覆/盖/绝/缘/层 ⟩ のみが正解と判定された.「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合では，⟨ カバー/絶縁/層，盖/绝/缘/层 ⟩ の要素合成法の翻訳確率 (f_9) が小さくなったことが原因となって，SVM によって誤りと判定された．一方，日本語専門用語「駆動/回路」および中国語訳語 “驱/动/器/电/路” の組においては，要素合成法を用いた翻訳確率の素性 (f_9) が 0 となったことが原因で，SVM によって誤りの対訳専門用語であると判定されてしまった．ここでの要素合成法の過程においては，構成要素の対訳対 ⟨ 駆動，驱/动/器 ⟩ はフレーズ翻訳テーブルには存在するものの，翻訳確率が下限値 (0.0005) 以下であるため，要素合成法を用いて訳語を推定することができず，要素合成法を用いた翻訳確率が 0 となってしまった.

4.5 関連研究

訳語対の自動獲得手法において，本研究以外に，統計的機械翻訳モデルを用いたものとして，[17, 27, 35, 59, 12] がある.

[17] においては，統計的機械翻訳モデルによって生成された単語対応を用いて，原言語名詞句と目的言語名詞句から対訳対を推定する手法を提案した．生成された対訳対に対して混合ガウスモデルを用いた分類器を構築し，対訳対が適切か否かの検証を行った．[59] においては，日本語漢字から中国語簡体字への文字対応情

報と既存の統計的機械翻訳モデルを用いて，日中特許対訳コーパスから対訳辞書を段階的に自動構築した．[59] においては，句に基づくフレーズ翻訳テーブルおよび階層的句に基づくフレーズ翻訳テーブルの二種類のフレーズ翻訳テーブルから，共通する訳語候補を抽出することにより，訳語対を生成する．[12] においては，統計的機械翻訳モデルを用いて対訳対の候補を生成し，対数尤度を用いて適切な対訳対を選定し，対訳辞書の自動構築を行う手法を提案した．本研究と [17, 59, 12] の間の最も大きな相違点として，本研究においては，フレーズ翻訳テーブルから得られた訳語候補のうち，日中対訳文対の中国語文に出現する訳語候補を抽出することにより，訳語推定を行う点，および，生成された訳語候補に対して，複数の対訳文から得られる素性を用いたSVMによって高信頼度な対訳専門用語を同定する点，が挙げられる．

[27] においては，英中パテントファミリーを情報源として，英中対訳特許文を収集し，SVMを用いた分類器によって対訳専門用語を同定した．本研究と [27] の間の最も大きな相違点として，本研究において用いた素性の方が [27] において用いられた素性よりも多いという点が挙げられる．[27] においては，対訳専門用語の同定のための素性として，英語専門用語の頻度と中国語専門用語の頻度の比率，および，フレーズ翻訳テーブルの各種スコアを利用した．一方，本研究においては，単言語専門用語の頻度情報およびフレーズ翻訳テーブルの翻訳確率以外に，同一日本語専門用語に対する翻訳確率の降順順位 (f_4)，日中対訳専門用語の共起頻度の素性 (f_5)，日本語専門用語の頻度と日中対訳専門用語の共起頻度の差 (f_6)，同一日本語専門用語に対する中国語訳語候補の数 (f_7)，対訳文およびフレーズ翻訳テーブルを用いた訳語推定を行う際の文単位の句対応制約の違反のない対訳文の割合 (f_8)，および，要素合成法を用いた訳語推定の翻訳確率 (f_9) などの情報を素性として用いた．これらの素性の性能を比較した予備実験の結果においては，本研究において用いた素性の方が，[27] で用いられた素性よりも高い性能を達成した．

[35] においては，知識源として，句に基づく統計的機械翻訳モデルのフレーズ翻訳テーブルおよび既存の対訳辞書を併用して，日英間の訳語推定を行った．一方，本論文では，日中の対訳特許文のみを知識源として，句に基づく統計的機械翻訳モデルを学習することによって作成されたフレーズ翻訳テーブルを用いて，日中間の訳語推定を行った．さらに，本論文においては，複数の対訳文から得られる素性を用いたSVMによって，高信頼度な日中間対訳専門用語を同定する手法を提案しており，[35] と比べると，これらの点が本研究の新規性となる．例として挙げると，本研究では，新しく導入した素性として，全対訳文から推定できた日本語専門用語の中国語訳語候補の数の素性 (f_7) を用いた．また，[35] での句対応制約に違反するかの二値素性と比べて，本研究においては，対訳文およびフレーズ翻訳テーブルを用いた訳語推定を行う際の文単位の句対応制約の違反のない対訳文の割合 (f_8) を素性として用いた．さらに，[35] での要素合成法で訳語を推定できる否かの二値素性と比べて，本研究においては，要素合成法を用いた訳語推定の翻訳確率の素性 (f_9) を用いた．これらの素性の性能を比較した予備実験の結果

においては，本研究において用いた素性の方が，[35] で用いられた素性よりも高い性能を達成した．

4.6 本章のまとめ

本章では，日中対訳特許文に対して，句に基づく統計的機械翻訳モデルにより学習されるフレーズ翻訳テーブルを用いて，対訳専門用語を同定する手法を提案した．提案手法では，日本語名詞句を頻度別に分類し，評価対象として，頻度レンジごとに均等に抽出した日本語名詞句から日本語専門用語を選定し，対訳特許文から学習されたフレーズ翻訳テーブルを用いることによって，日中対訳専門用語の候補を生成した．そして，生成した日中対訳専門用語候補に対して，複数の対訳文から得られる素性を用いたSVMを適用した．適合率を最大化する調整を行うことにより，90%以上の適合率を達成した．今後は，中国語側の区切り単位として，形態素および文字の二種類の単位を併用した上でSVMを適用することによって，日中対訳専門用語同定の性能を改善する方式に取り組む．この方式が確立されれば，日中対訳専門用語の同義集合を同定する方式 (次章) 等，本研究の手法を要素技術として利用する関係にある他の研究課題における性能改善が期待できる．

第5章　同義対訳専門用語の同定

5.1　はじめに

第4章では，360万件の日中対訳特許文を言語資源として，句に基づく統計的機械翻訳モデルにより学習されるフレーズ翻訳テーブル，要素合成法，SVMによる分類器学習を用いることによって，日中専門用語対訳対を獲得した．しかし，この手法においては，ある日本語専門用語の訳語推定の際に，その日本語専門用語が出現する一つの対訳文に出現する訳語のみを推定対象としていた．したがって，他の対訳文に出現している同義の専門用語対訳対とは全く独立に訳語推定が行われており，本来同義関係にある複数の専門用語対訳対の間の関係を同定できない，という問題点があった．そこで，本論文では，ある日本語専門用語およびその同義語候補が出現する複数の対訳文を入力として，同義の専門用語対訳対を同定する手法を提案する．提案手法では，対訳特許文および句に基づく統計的機械翻訳モデルのフレーズ翻訳テーブルを用いて，対訳関係にある日中専門用語の対 (「日中対訳専門用語」と呼ぶ) $\langle t_J, t_C \rangle$(ただし，t_J，t_C はそれぞれ日本語専門用語，及び中国語専門用語) を多数収集する．そして，二組の日中対訳専門用語 $\langle t_J, t_C \rangle$ および $\langle t'_J, t'_C \rangle$ の間で以下の同義関係を定義し，この同義関係の判定を行うタスクに対してSVMを適用するというアプローチをとる．

$\langle t_J, t_C \rangle$ と		t_J と t'_J，および，
$\langle t'_J, t'_C \rangle$ が	$\longleftrightarrow$	t_C と t'_C の組が
同義である		それぞれ同義である．

本論文の実際の手順においては，まず，ある日本語専門用語を種として，同義関係にある専門用語対訳対の候補を生成・収集する．生成・収集した候補集合の中から同義判定を行うための**中心的対訳対**を選び，中心的対訳対のうちの日本語専門用語に対して，専門用語対訳対同義候補集合を再生成する．再生成した候補集合に対してSVM分類器を適用することにより，同義集合・異義集合を同定する．そして，日中パテントファミリーから抽出した360万対の日中対訳文に対して提案手法を適用し，同義関係にある日中対訳専門用語の同定において，再現率が25%以上という条件のもとで，約90%の適合率を達成した．さらに，比較対象として，日英同義対訳専門用語の同定を対象とした先行研究 [50] における素性と同等の素性のもとで日中同義対訳専門用語の同定を行った評価結果との比較を行い，本論文で提案する素性の組合わせによって大幅に性能が改善されることを示した．

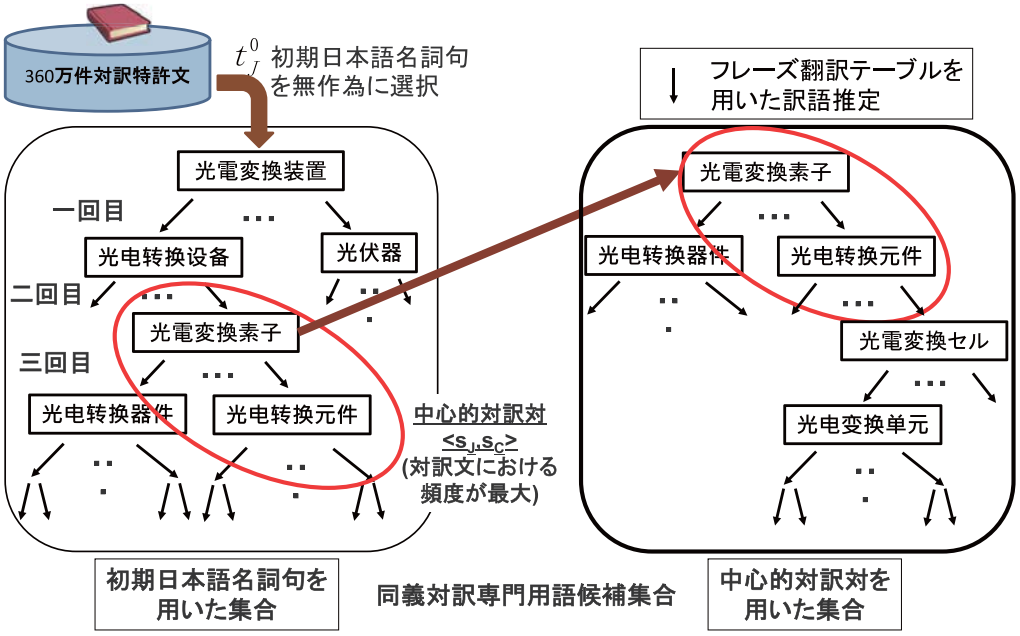


図 5.1: 専門用語対訳対の訓練・評価用同義・異義集合の作成

5.2 専門用語対訳対の訓練・評価用同義・異義集合の作成

5.2.1 作成手順

全体的手順

以下では，図 5.1 に沿って，対訳特許文およびフレーズ翻訳テーブルを用いて，専門用語対訳対の訓練・評価用同義・異義集合を作成する流れを示す．以下の手順においては，訓練・評価用の同義・異義専門用語対訳対の候補を生成するための種として，**中心的対訳対**と呼ぶ対訳対を選定し，この中心的対訳対を初期対訳対として，そこから訓練・評価用の同義・異義専門用語対訳対の候補を生成する．ただし，中心的対訳対を選定するに当たっては，無作為に抽出した初期日本語名詞句を用いることにより，まず，同義・異義専門用語対訳対の初期候補集合を生成し，この初期候補集合中の要素のうち一定の基準を満たす対訳対を中心的対訳対として選定するという方式を用いる[1]．

(1) まず，日中対訳特許文から無作為に初期日本語名詞句 t_J^0 を抽出する．

[1]本論文の主たる主張は，専門用語対訳対の同義候補集合を生成した後，分類器学習手法によって同義対訳専門用語を同定する手法を提案することである．本節においては，中心的対訳対を用いて同義候補集合を作成する手順を示しているが，これはあくまで経験的な知見を示したものに過ぎない．特に，専門用語対訳対の同義候補集合において十分な数の正例・負例を含むためには，中心的対訳対の候補集合を作成する段階において十分に大きく，かつ，一定以下の大きさに制限された集合を作成する必要があるという経験的知見を得ており，この知見に基づき，本論文で述べる手順を用いる．

(2) そして，初期日本語名詞句 t_J^0 に対して，次節に示す「反復手続き: 専門用語対訳対同義候補集合の生成」を適用することにより，図 5.1 左半分の過程を行い初期候補集合 $CBP(t_J^0)$ を生成する．この過程においては，対訳特許文，および，フレーズ翻訳テーブルを用いる[2]，初期候補集合 $CBP(t_J^0)$ を生成する．ここで，集合 $CBP(t_J^0)$ の要素数が m_0 以上である (すなわち，$|\ CBP(t_J^0)\ | \geq m_0$) ならば，以降の手続きを続ける．

(3) 次に，t_J^0 を初期日本語名詞句として作成した同義候補集合 $CBP(t_J^0)$ の要素の中から，5.2.2 節の「中心的対訳対の選定手順」に従い，**中心的対訳対** $s_{JC} = \langle s_J, s_C \rangle$ を選定する．ただし，5.2.2 節の「中心的対訳対の選定手順」においては，初期候補集合 $CBP(t_J^0)$ 中に「専門用語の対訳対」が存在しない場合は，その初期候補集合 $CBP(t_J^0)$ は，その時点で棄却される．

(4) 次に，中心的対訳対 $s_{JC} = \langle s_J, s_C \rangle$ の日本語用語 s_J に対して，次節に示す「反復手続き: 専門用語対訳対同義候補集合の生成」を適用することにより，図 5.1 の右半分の過程を行い，対訳特許文，および，フレーズ翻訳テーブルを用いて，専門用語対訳対の同義候補集合 $CBP(s_J)$ を生成する．

(5) 最後に，人手によって，同義候補集合 $CBP(s_J)$ を，中心的対訳対 s_{JC} と同義となる対訳対の集合 (訓練・評価用同義対訳専門用語集合)$SBP(s_{JC})$，および，その他の対訳対の集合 (訓練・評価用異義対訳専門用語集合) $NSBP(s_{JC})$ に分割する．

反復手続き: 専門用語対訳対の同義候補集合の生成

ステップ 1 入力の日本語用語 t'_J に対して，日中対訳特許文の中から t'_J が出現する対訳文をすべて収集する．そして，収集した各対訳文に対して，フレーズ翻訳テーブルを参照して 3.2.2 節で述べた手法[3] を適用することにより，t'_J の中国語訳語を推定する[4]．そして，推定された中国語訳語を t_C^i $(i = 1, \ldots, n_1$, ただし，n_1 は推定された中国語訳語の数) として，すべての対訳対 $\langle t'_J, t_C^i \rangle$ を収集し，専門用語対訳対の同義候補集合の初期集合 $CBP(t'_J)$ を作成する．

[2]この「反復手続き: 専門用語対訳対同義候補集合の生成」においては，訳語推定回数を 6 回としている．これは，予備実験において，6 回の訳語推定において，同義対訳専門用語として獲得することが望ましい専門用語対訳対のうちの大半が生成され，しかも，同義関係にある専門用語対訳対以外の候補の生成数が最少となったためである．なお，これらの反復手続きにおいては，同一の日本語用語，および，中国語用語を検出することにより，それらの重複生成は回避されている．

[3]ただし，ここにおいて，翻訳確率が最大となる訳語候補が複数存在する場合には，いずれの訳語候補も信頼性が低いとみなし，不採用とする．

[4]この手順を経ることにより，各対訳文からは，フレーズ翻訳テーブルに含まれる対訳対のうち，誤りに相当する部分文字列等が削除され，対訳対として適切である可能性の高いものだけが得られる可能性が高くなる．

ステップ 2 同様に，すべての中国語用語 $t_C \in CBP(t'_J)$ に対して，対訳特許文の中から t_C を含む対訳文をすべて収集し，t_C の日本語訳語を推定する．そして，推定された日本語訳語を t_J^j $(j = 1, \ldots, n_2$，ただし，n_2 は推定された日本語訳語の数) として，すべての対訳対 $\langle t_J^j, t_C \rangle$ を $CBP(t'_J)$ に追加する．

ステップ 3 同様に，すべての日本語用語 $t_J \in CBP(t'_J)$ に対して，対訳特許文の中から t_J を含む対訳文をすべて収集し，t_J の中国語訳語を推定する．そして，推定された中国語訳語を t_C^k $(k = 1, \ldots, n_3$，ただし，n_3 は推定された中国訳語の数) として，すべての対訳対 $\langle t_J, t_C^k \rangle$ を $CBP(t'_J)$ に追加する．

ステップ 4 「ステップ 2」の処理を実行する．

ステップ 5 「ステップ 3」の処理を実行する．

ステップ 6 「ステップ 2」の処理を実行する．

ただし,「ステップ 1」から「ステップ 6」までのすべてのステップにおいて収集対象とする対訳対を選定する際には，5.2.2 節の「収集対象対訳対の選定手順」に従う．

5.2.2 作成手順における詳細設定

本節では，本論文において，5.2.1 節の作成手順に沿って実際に専門用語対訳対の訓練・評価用同義・異義集合の作成を行う際にどのような具体的な設定を用いたのかの詳細を説明する．

まず，情報源として用いた対訳特許文は，3.1.1 節で述べた 360 万対訳特許文である．また，5.2.1 節の「全体手順」の手順 (2) における初期候補集合 $CBP(t_J^0)$ の要素数の下限値 m_0 は 10 とする．その他，5.2.1 節の「反復手続き: 専門用語対訳対の同義候補集合の生成」において，収集対象とする対訳対を選定する手順，および，5.2.1 節の「全体手順」の手順 (3) において中心的対訳対を選定する手順の詳細は以下の各節の通りである．

「反復手続き: 専門用語対訳対の同義候補集合の生成」における収集対象対訳対の選定手順

本節では，5.2.1 節の「反復手続き: 専門用語対訳対の同義候補集合の生成」において，収集対象とする対訳対を選定する手順を述べる．

5.2.1 節の「反復手続き: 専門用語対訳対の同義候補集合の生成」の「ステップ 1」から「ステップ 6」までのすべてのステップにおいて,「一般語の対訳対」をで

きるだけ除去し,「専門用語の対訳対」をできるだけ多く残すための予備調査を行ったところ，以下の条件をすべて満たす対訳対の7割以上が「専門用語の対訳対」であったのに対して，以下の条件のうち一つしか満たさない対訳対のうち「専門用語の対訳対」であるものは2割以下であった．また，その大半においては，対訳となる日中用語の一方のみの頻度が12,500以上となっており,「専門用語の対訳対」としても相対的な適切さの度合いが低い対訳対であった．以上の予備調査の結果をふまえて,「ステップ1」から「ステップ6」までのすべてのステップにおいて,以下の条件をすべて満たす対訳対$\langle t_J, t_C \rangle$(ただし，t_J，t_Cはそれぞれ日本語専門用語，及び中国語専門用語) のみを収集対象として残し，その他の組を枝刈りする．

1. t_J，t_Cのいずれの頻度も12,500未満.
2. t_J，t_Cのいずれの頻度も700未満，又は，長さの下限[5]を満す.
3. t_J，t_Cいずれも語頭及び語尾が機能語，数字，句読点でない (これらはいずれも，フレーズ自動抽出時に自動生成されたものであり，専門用語の語頭・語尾としては不適切なものである).
4. $\langle t_J, t_C \rangle$の頻度が3,000未満.

中心的対訳対の選定手順

本節では，5.2.1節の「全体手順」の手順(3)において，t_J^0を初期日本語名詞句として作成した同義候補集合$CBP(t_J^0)$の要素の中から中心的対訳対を選定する手順を述べる．

本論文において,「一般語の対訳対」と「専門用語の対訳対」を区別するための予備調査を行ったところ，以下で述べる条件を一つも満たさない場合に「専門用語の対訳対」となる割合は十分に低いが，以下で述べる条件を少なくとも一つ満たす場合には一定の割合で「専門用語の対訳対」となることが分かった．そこで，以下で述べる条件を少なくとも一つ満たす場合に，その対訳対は「一般語の対訳対」でないというヒューリスティクスを用いた．

(a) 日本語用語が以下のいずれかを満たす.

  (i) 漢字または平仮名を含む場合は，3文字以上.

  (ii) カタカナ語の場合は，複合語である.

(b) 中国語用語が4文字以上，または形態素数が3以上.

[5] t_Jが (i) 連続する漢字長が3以上，(ii) 漢字数が4以上，(iii) 文字数が6以上，かつ，形態素数が2以上，(iv) 一形態素の場合は10文字以上，のいずれかを満たし，かつ，t_Cが (i) 文字数が4以上，(ii) 形態素数が2以上の場合は3文字以上，のいずれかを満たす.

表 5.1: 作成された専門用語対訳対同義候補集合中の対訳対数

(a) 中国語側が形態素単位のフレーズ翻訳テーブルを用いた場合

		総要素数		114 個の集合の間の平均対数	
同義候補集合 $\bigcup_{s_J} CBP(s_J)$	形態素単位の集合のみに含まれる	12,640	24,621	110.9	216.0
	文字単位の集合と共通	11,981		105.1	
人手で同定した同義集合 $\bigcup_{s_{JC}} SBP(s_{JC})$	形態素単位の集合のみに含まれる	228	2,473	2.0	21.7
	文字単位の集合と共通	2,245		19.7	

(b) 中国語側が文字単位のフレーズ翻訳テーブルを用いた場合

		総要素数		114 個の集合の間の平均対数	
同義候補集合 $\bigcup_{s_J} CBP(s_J)$	文字単位の集合のみに含まれる	6,358	17,478	55.8	153.3
	形態素単位の集合と共通	11,120		97.5	
人手で同定した同義集合 $\bigcup_{s_{JC}} SBP(s_{JC})$	文字単位の集合のみに含まれる	287	2,318	2.5	20.3
	形態素単位の集合と共通	2,031		17.8	

そして，初期候補集合 $CBP(t_J^0)$ の要素のうち，このヒューリスティクスによって一般語の対訳対であると判定された対訳対を除去し，残った対訳対を対象として，360 万対訳文中の共起頻度が最大となる対訳対を人手で確認し，専門用語の対訳対として適切であると判定された場合は，その対訳対を「中心的対訳対」とする．それ以外の場合は，初期候補集合 $CBP(t_J^0)$ 中の全ての対訳対のうち，最も適切な対訳対を「中心的対訳対」とする．ただし，初期候補集合 $CBP(t_J^0)$ 中に「専門用語の対訳対」が存在しない場合は，その初期候補集合 $CBP(t_J^0)$ は，その時点で棄却する．

5.2.3 作成結果

5.2.1 節の「全体手順」に従い，専門用語対訳対の訓練・評価用同義・異義集合の作成を行った結果を以下に示す．まず，手順 (1) において無作為に選択した 4,000 個の初期日本語名詞句に対して，手順 (2) に従い初期候補集合 $CBP(t_J^0)$ を生成した結果，およそ 500 個の集合が要素数の下限を満たした．さらに，これらの初期候補集合に対して，手順 (3) に従って中心的対訳対の選定を行った結果，合計 114 個の中心的対訳対が選定された．この後，手順 (4) に従うことにより，4.2 節で述

表 5.2: 専門用語対訳対の同義・異義同定のための素性 (1) (提案手法)

分類	素性名	定義 (ただし，$X \in \{J,C\}$，$(Y,Z) \in \{(J,C),(C,J)\}$)
対訳対 $\langle t_J, t_C \rangle$ の特性を規定	f_1: 共起頻度	対訳特許文における $\langle t_J, t_C \rangle$ の共起頻度の二進対数.
	f_2: 中国訳語の順位	条件付き確率 $\mathrm{P}(t_C \mid t_J)$ の降順に t_C を順位付けしたときの t_C の順位の二進対数.
	f_3: 日本語訳語の順位	条件付き確率 $\mathrm{P}(t_J \mid t_C)$ の降順に t_J を順位付けしたときの t_J の順位の二進対数.
	f_4: 日本語文字数	t_J の文字数.
	f_5: 中国語文字数	t_C の文字数.
	f_6: 訳語推定における繰り返しの回数	s_J から訳語推定を開始し，訳語として t_Y を生成した直後に t_Y から t_Z を訳語推定 した場合の，s_J から t_Z までの繰り返し訳語生成回数.

べた「中国語側が形態素単位のフレーズ翻訳テーブル」及び「中国側が文字単位のフレーズ翻訳テーブル」をそれぞれ独立に用いて，選定された 114 個の中心的対訳対に対して，中国語側が形態素単位のフレーズ翻訳テーブルを用いた場合の専門用語対訳対同義候補集合，と，中国語側が形態素単位のフレーズ翻訳テーブルを用いた場合の専門用語対訳対同義候補集合を，それぞれ，生成した．それらの同義候補集合における専門用語対訳対の総数および平均対訳対数を表 5.1 に示す[6]．最後に，手順 (5) に従って，中心的対訳対と同義となる対訳対の選定を行った結果，表 5.1 に示すように，114 個の専門用語対訳対の候補集合において，中心的対訳対と同義となる対訳対の総数および平均数は,「中国語側が形態素単位のフレーズ翻訳テーブル」を用いた場合では，2,473 個および 21.7 個となり,「中国語側が文字単位のフレーズ翻訳テーブル」を用いた場合では，2,318 個および 20.3 個となった.

5.3 分類器学習を用いた同義対訳専門用語の同定

本節では，SVM を用いて同義対訳専用語を同定する手法について述べる.

5.3.1 適用手順

まず，114 個の専門用語対訳対同義候補集合 $CBP(s_J)$ の和集合を全事例集合 CBP とし，互いに素な部分集合 $CBP_i (i = 1, \ldots, 10)$ に 10 分割する[7]．本論文で

[6]表 5.1 では，中国語側の形態素解析誤りが原因で，同一の文字列に対する形態素分割のパターンが 2 通り以上出現する場合があるため，表 5.1(a) において「文字単位の集合と共通」となる対訳対数が，表 5.1(b) において「形態素単位の集合と共通」となる対訳対数よりも多くなっている.

[7]各 $CBP_i (i=1, \ldots, 10)$ における正例 (中心的対訳対と同義)・負例 (中心的対訳対と異義) の数が，各 $CBP_i (i=1, \ldots, 10)$ の間で均等になるように，中心的対訳対の集合を分割した.

表 5.3: 専門用語対訳対の同義・異義同定のための素性 (2) (提案手法)

分類	素性名	定義 (ただし，$X \in \{J,C\}$，$(Y,Z) \in \{(J,C),(C,J)\}$)
対訳対 $\langle t_J, t_C \rangle$ と中心的対訳対 $\langle s_J, s_C \rangle$ の間の関係を規定する	f_7: 日本語用語が同一	$t_J = s_J$ ならば，1 となる.
	f_8: 中国語用語が同一	$t_C = s_C$ ならば，1 となる.
	f_9: 編集距離類似度	$f_9(t_X, s_X) = 1 - \frac{ED(t_X, s_X)}{\max(\|t_X\|, \|s_X\|)}$ ：ED は t_X と s_X の間の編集距離，$\|\ t\ \|$ は t に含まれる文字数を表す.
	f_{10}: バイグラム類似度	$f_{10}(t_X, s_X) = \frac{\|bigram(t_X) \cap bigram(s_X)\|}{\max(\|t_X\|, \|s_X\|) - 1}$ ： $bigram(t)$ は，t に含まれる文字単位のバイグラムの集合.
	f_{11}: 日本語用語の同一形態素の割合	$f_{11}(t_J, s_J) = \frac{\|const(t_J) \cap const(s_J)\|}{\max(\|const(t_J)\|, \|const(s_J)\|)}$ ： $const(t)$ は日本語用語 t に含まれる形態素単語の集合.
	f_{12}: 中国語用語の同一文字数の割合	$f_{12}(t_C, s_C) = \frac{\|const(t_C) \cap const(s_C)\|}{\max(\|const(t_C)\|, \|const(s_C)\|)}$ ： $const(t)$ は中国語用語 t に含まれる文字の集合.
	f_{13}: 日本語用語の文字列の包含関係もしくは異表記	t_J と s_J は，以下のいずれかの関係を満たす. (i) 構成要素の差分は接尾辞のみ，(ii) 構成文字列の差分は，長音「ー」のみ，(iii) 構成文字列の差分は，送り仮名の違いのみ.
	f_{14}: 中国語用語の文字列の包含関係	t_C と s_C の構成要素の差分は語頭・語尾でない「的」のみ.
	f_{15}: フレーズ翻訳テーブルの共通訳の割合	$f_{15}(t_X, s_X) = \frac{\|trans(t_X) \cap trans(s_X)\|}{\max(\|trans(t_X)\|, \|trans(s_X)\|)}$ ： $trans(t)$ は，フレーズ翻訳テーブルから得られる用語 t のすべての訳語の集合.
	f_{16}: 全非共有箇所に対しフレーズ翻訳テーブルにおける共通訳の割合	t_X と s_X の間で文字列が一致しない箇所 $x_t^1, \ldots, x_t^m$，$x_s^1, \ldots, x_s^n$ に対して，$x_t^i (i = 1, \ldots, m)$ と $x_s^j (j = 1, \ldots, n)$ の 1 対 1 対応に対して，フレーズテーブルから得られる訳語の集合 $trans(x_t^i)$ および $trans(x_s^j)$ 中の共通訳の割合を求め，その共通訳の割合の積 $(i = 1, \ldots, m,\ j = 1, \ldots, n)$ が最大となる 1 対 1 対応において，共通訳の割合の積を素性値とする.
	f_{17}: フレーズ翻訳テーブルの訳語関係が存在	フレーズ翻訳テーブル中に t_Y と s_Z の訳語関係が存在する．($\langle t_J, s_C \rangle$ または $\langle s_J, t_C \rangle$ のどちらか一方のみの訳語関係が存在することを表す素性，および，$\langle t_J, s_C \rangle$ と $\langle s_J, t_C \rangle$ の両方の訳語関係が存在することを表す素性の二種類を区別して用いる).

は，TinySVM[8] を利用して，評価実験を行った．カーネル関数としては，一次多項式カーネルおよび二次多項式カーネルを評価し，大きな性能差が観測されなかったため，評価実験においては一次多項式カーネルを用いた．また，SVM の分離平面から評価事例までの距離を信頼度とし，正例 (中心的対訳対と同義) 判定において信頼度の下限を設定した．$CBP_1, \ldots, CBP_{10}$ の 10 個の部分集合のうち，8 個を訓練用事例集合として SVM の訓練を行い，残りのうちの 1 個を調整用事例集合とし，最後の 1 個を評価用事例集合とした．調整用事例集合を用いたパラメータの調整においては，分離平面から評価用事例までの距離の下限のパラメータの調整を行った[9]．以上の訓練，調整，評価の手順を 10 通り繰り返し，その評価結果のマイクロ平均を算出し，同義判定の性能評価を行った．

5.3.2 同義・異義判定のための素性

表 5.2 と表 5.3 に示すように，同義対訳専門用語の同定に用いた素性は，大きく，対訳対 $\langle t_J, t_C \rangle$ の特性を規定するもの，および，対訳対 $\langle t_J, t_C \rangle$ と中心的対訳対 $\langle s_J, s_C \rangle$ の間の関係を規定するものの 2 種類に分けられる．以下にその詳細を述べる．

対訳対の特性を規定する素性

対訳対の特性を規定する素性としては，対訳特許文における対訳対の共起頻度の素性 (f_1)，訳語の翻訳確率における順位の素性 (f_2，f_3)，用語の文字列長・単語長の素性 (f_4，f_5)，訳語推定において，中心的対訳対の日本語用語 s_J から t_C または t_J を生成するまでの繰り返し訳語生成回数の素性 (f_6) を用いる．

対訳対と中心的対訳対の間の関係を規定する素性

対訳対と中心的対訳対の間の関係を規定する素性としては，用語表記の同一性の素性 (f_7，f_8)，用語文字列の編集距離類似度の素性 (f_9)，用語文字列の 2 グラム類似度の素性 (f_{10})，日本語用語間における同一形態素数の素性 (f_{11})，中国語用語間における同一文字数の割合 (f_{12}) 日本語用語文字列包含関係・異表記の素性 (f_{13})，中国語用語文字列包含関係の素性 (f_{14})，フレーズ翻訳テーブルから得られる訳語のうち共通なものの割合の素性 (f_{15})，文字列の非共有箇所のみに対してフレーズ

[8] http://chasen.org/~taku/software/TinySVM/

[9] SVM のソフトマージンを制約するパラメータについても調整を行い性能への影響を評価したが，SVM light(http://svmlight.joachims.org/) におけるデフォルト値 (訓練事例の素性値の二乗和の平均値の逆数) と比べて大きな変化が観測されなかったため，SVM のソフトマージンを制約するパラメータはデフォルト値に固定して評価を行った．

表 5.4: 同義対訳専門用語同定の評価結果 (%)

(a) 中国語側が形態素単位のフレーズ翻訳テーブルを用いた場合

手法 (素性・分離平面からの距離下限調整の基準)		適合率	再現率	F 値
ベースライン		71.4	40.0	51.3
SVM(全素性)	適合率最大	**86.5**	26.5	40.5
	F 値最大	64.3	64.1	**64.2**
SVM (適合率 最大となる素性の組合わせ: $f_{1\sim6} + f_{9\sim16}$)	適合率最大	**89.0**	23.9	37.7
SVM ([50] の素性)	適合率最大	72.6	26.1	38.4
	F 値最大	71.0	54.7	61.5

(b) 中国語側が文字単位のフレーズ翻訳テーブルを用いた場合

手法 (素性・分離平面からの距離下限調整の基準)		適合率	再現率	F 値
ベースライン		74.0	40.1	52.0
SVM(全素性)	適合率最大	**89.0**	26.1	40.4
	F 値最大	63.5	65.3	**64.4**
SVM (適合率最大となる素性の組合わせ: $f_{2,3} + f_{6\sim9} + f_{11,12,15,16}$)	適合率最大	**90.4**	25.5	40.4
SVM ([50] の素性)	適合率最大	74.4	36.7	49.2
	F 値最大	72.7	53.7	61.8

翻訳テーブルから得られる訳語のうち共通なものの割合の素性 (f_{16})，フレーズ翻訳テーブルにおいて s_J と t_C または s_C と t_J の間に訳語関係が存在するか否かの素性 (f_{17}) を用いる [10] ．このうち，f_{15} および f_{16} は，フレーズ翻訳テーブルにおいてどの程度の割合で共通の訳語を持つかという情報と，単言語において同義関係にある度合いとの間の相関に着目した素性であり，次節の評価結果において示すように，性能に大きな影響を持つ重要な素性である．

5.3.3 評価結果

表 5.4 に，同義判定における性能の評価結果を示す．ベースラインとしては，

> 「t_J と s_J が同一，または，t_C と s_C が同一の場合に，対訳対 $\langle t_J, t_C \rangle$ は中心的対訳対 $\langle s_J, s_C \rangle$ と同義である」

という規則を用いた．まず，分離平面からの距離下限のパラメータに対して，同義判定の適合率を最大化する調整 [11] を行った．「中国語側が形態素単位」の場合，

[10] ただし，素性 f_9，f_{10}，f_{15}，f_{16} においては，それぞれ日本語側素性および中国語側素性の二種類の素性を用いる．

[11] ただし，再現率が 25%以上となるという条件のもとで，パラメータの調整を行った．

表 5.5: 「適合率最大の場合」との間で有意差 (有意水準 5%) のない適合率となる 2 種類の素性情報の組とその評価結果 (%)

(a) 中国語側が形態素単位のフレーズ翻訳テーブルを用いた場合

素性	適合率	再現率	F 値
f_{15}(日中)+ f_{16}(日中)	85.6	25.4	39.2
f_{9}(日中)+ f_{16}(日中)	86.8	24.9	38.7
f_{13}(日) + f_{14}(中) +f_{16}(日中)	86.8	24.8	38.6

(b) 中国語側が文字単位のフレーズ翻訳テーブルを用いた場合

素性	適合率	再現率	F 値
f_{9}(日中) + f_{15}(日中)	87.4	25.4	39.3

全素性を用いた場合 (表 5.4「SVM(全素性)」欄) には 86.5%，適合率最大となる素性の組合わせ ($f_{1\sim6} + f_{9\sim16}$) を用いた場合 (表 5.4「SVM(適合率最大となる素性の組合わせ)」欄) には 89.0%の適合率を達成した．一方,「中国語側が文字単位」の場合，全素性を用いた場合には 89.0%，適合率最大となる素性の組合わせ ($f_{2,3} + f_{6\sim9} + f_{11,12,15,16}$) を用いた場合には 90.4%の適合率を達成した．ただし,「中国語側が形態素単位」の場合，および,「中国側が文字単位」の場合，いずれにおいても，全素性を用いた場合と適合率最大となる素性の組み合わせを用いた場合との間で適合率の差には有意差 (有意水準 5%) はない．次に，全素性を用いて，分離平面からの距離下限のパラメータに対して，同義判定の F 値を最大化する調整を行ったところ,「中国語側が形態素単位」の場合 64.2%の F 値を,「中国語側が文字単位」の場合 64.4%の F 値を，それぞれ達成した [12] ．

性能に大きな影響を持つ素性を同定するために，適合率最大の場合との間で有意差 (有意水準 5%) のない適合率となる素性の組合わせのうち，二種類の素性 (一つの素性で日中二言語の情報を記述するもの，もしくは，同種類の情報を記述する日本語素性および中国語素性の二つの素性) からなる場合の性能を表 5.5 に示す．この結果から，f_{15} および f_{16} のように，単言語の各専門用語またはその断片の間にフレーズ翻訳テーブルにおける共通の訳語が存在するか否かを記述する素性が，重要な素性の一つであることが分かる．この f_{15} および f_{16} は，[25] における素性の組合わせを改善する形で新たに導入された「フレーズ翻訳テーブルにおける共通訳の割合」の考え方に基づく素性であるが，本節の評価結果より，この新素性が性能に大きな影響を持つ重要な素性であることが示された．

また，比較対象として，日英同義対訳専門用語の同定を対象とした先行研究 [50] における素性と同等の素性の組合わせを表 5.6 のように設計し，5.3.1 節に示した訓

[12]その他,「中国語側が形態素単位」と「中国語側が文字単位」の間で判定結果の AND 条件をとった場合の適合率の評価も行ったが,「中国語側が形態素単位」単独，および,「中国語側が文字単位」単独の場合の適合率を有意に改善することはできなかった．

表 5.6: 専門用語対訳対の同義・異義同定のための素性 ([50] の手法)

分類	素性名	定義
基本素性	h_{1J}，h_{1C}: 第一文字の一致	日中各単言語において中心的対訳対の専門用語との間で第一文字が一致するか否か.
	h_{2J}，h_{2C}: 編集距離類似度	f_9 と同じ.
	h_{3J}，h_{3C}: バイグラム類似度	f_{10} と同じ.
	h_{4J}，h_{4C}: 部分文字列の一致数	日中各単言語において中心的対訳対の専門用語との間で部分文字列が一致する回数. ([50] では，中心的対訳対の専門用語の部分文字列との間で既知の同義関係が成り立つ回数も併せて数えているが，本論文では，利用可能な既知の同義関係の情報がないため，部分文字列の一致回数のみ素性として用いる．)
	h_{5J}，h_{5C}: フレーズ翻訳テーブルの訳語関係が存在	f_{17} と同じ. ([50] では，フレーズ翻訳テーブルの代りに訓練用の対訳辞書を用いている．)
	h_6: 中国語用語の文字列の包含関係	f_{14} と同じ. ([50] では，英語における頭字語を扱うための素性として用いているが，本論文では，中国語における「的」のための素性に置き換える．)
	h_7: 日本語用語の文字列の包含関係もしくは異表記	f_{13} と同じ. ([50] では，片仮名語の異表記のみを扱うための素性として用いているが，本論文では，片仮名語の長音「ー」の他，接尾辞および送り仮名の異表記を扱うための素性に置き換える．)
複合素性	$h_{1J} \wedge h_{1C}$	—
	$\sqrt{h_{2J} \cdot h_{2C}}$	—
	$\sqrt{h_{3J} \cdot h_{3C}}$	—
	$h_{5J} \wedge h_{5C}$	—
	$h_6 \cdot h_{2J}$	—
	$h_7 \cdot h_{2C}$	—

練，調整，評価の手順をそのまま適用して性能評価を行った結果を表 5.4「SVM([50] の素性)」欄に示す．この結果から，提案手法によって，先行研究 [50] における素性と同等の素性の組合わせの性能を大幅に改善することが分かる.

次に，ベースラインによる同義判定の結果を，SVM によって改善する例を表 5.7 に示す.

表 5.7 (a)「SVM のみで同義と判定し正解」の例においては，専門用語対訳対と中心的対訳対の日本語表記および中国語表記の両方とも異なる場合 ($t_J \neq s_J$, $t_C \neq s_C$)，ベースラインでは異義であると判定されたが，提案手法では，「f_{17}: フレーズ翻訳テーブルの訳語関係が存在」(フレーズ翻訳テーブルにおいて「グラス転移温度」の訳語として "玻璃态转化温度" が存在し，「グラス転移点」の訳語として "玻璃化转变温度" が存在しており．$f_{17}(\langle t_J, t_C\rangle, \langle s_J, s_C\rangle) = 1$) となる素性の効果によって，同義と判定できた.

表 5.7: 同義判定における SVM による改善例

ベースライン:	t_J と s_J が同一，または，t_C と s_C が同一の場合に，対訳対 $\langle t_J, t_C \rangle$ は中心的対訳対 $\langle s_J, s_C \rangle$ と同義である
SVM:	中国語側が形態素単位のフレーズ翻訳テーブルを用いた場合，適合率が最大となる下限を用いたモデル

(a) SVM のみで同義と判定し正解

中心的対訳対 $\langle s_J, s_C \rangle$	専門用語対訳対 $\langle t_J, t_C \rangle$	人手による同義・異義判定	ベースラインによる判定	SVM による判定
⟨グラス転移温度，玻璃化转变温度⟩	⟨グラス転移点，玻璃态转化温度⟩	同義	異義	同義

(b) SVM のみで異義と判定し正解

中心的対訳対 $\langle s_J, s_C \rangle$	専門用語対訳対 $\langle t_J, t_C \rangle$	人手による同義・異義判定	ベースラインによる判定	SVM による判定
⟨集電装置，集电器⟩	⟨コレクト，集电器⟩	異義	同義	異義

表 5.8: 同義判定における提案手法の誤り例

(a) 提案手法により同義と判定し不正解

中心的対訳対 $\langle s_J, s_C \rangle$	専門用語対訳対 $\langle t_J, t_C \rangle$	日本語側		中国語側		素性 f_{17}（両方の訳語関係が存在）	素性 f_{17}（片方の訳語関係のみが存在）	人手による同義・異義判定	提案手法による判定
		素性 f_9	素性 f_{10}	素性 f_9	素性 f_{10}				
⟨断熱体，绝热体⟩	⟨インシュレータ，绝缘件⟩	0	0	0.33	0	1	1	異義	同義

(b) 提案手法により異義と判定し不正解

中心的対訳対 $\langle s_J, s_C \rangle$	専門用語対訳対 $\langle t_J, t_C \rangle$	日本語側		中国語側		素性 f_{17}（両方の訳語関係が存在）	素性 f_{17}（片方の訳語関係のみが存在）	人手による同義・異義判定	提案手法による判定
		素性 f_9	素性 f_{10}	素性 f_9	素性 f_{10}				
⟨成膜室，成膜室⟩	⟨成膜チャンバー，膜成形室⟩	0.29	0.17	0.5	0	0	1	同義	異義

一方，表 5.7 (b)「SVM のみで異義と判定し正解」の例においては，専門用語対訳対の中国語表記と中心的対訳対の中国語語表記が同一のため ($t_C = s_C$)，ベースラインでは同義であると判定されたが，提案手法では，日本語用語 t_J「集電装置」および s_J「コレクト」の文字列の間で，素性「f_9: 編集距離類似度」および素性「f_{10}: バイグラム類似度」のいずれも値が 0 となった ($f_9(\langle t_J, t_C \rangle, \langle s_J, s_C \rangle) = 0$, および，$f_{10}(\langle t_J, t_C \rangle, \langle s_J, s_C \rangle) = 0$)．提案手法では，これらの素性の効果によって異義と判定できた．

最後に，提案手法による誤り例を表 5.8 に示す．

表 5.8(a)「提案手法により同義と判定し不正解」の例では，素性「f_{17}: フレーズ翻訳テーブルの訳語関係が存在」において，フレーズ翻訳テーブル中に誤った対訳対 ⟨ 断熱体，绝缘件⟩ および ⟨ インシュレータ，绝热体⟩ が含まれることが原因

で,「f_{17}: フレーズ翻訳テーブル中に $\langle t_J, s_C \rangle$, $\langle s_J, t_C \rangle$ 両方の訳語関係が存在」および「f_{17}: フレーズ翻訳テーブル中に $\langle t_J, s_C \rangle$ または $\langle s_J, t_C \rangle$ の片方の訳語関係のみが存在」の両方の値が 1 となってしまい，最終的に誤って同義と判定されてしまった．この場合，フレーズ翻訳テーブル中の対訳対の正誤判定を行う分類器の訓練・適用過程を導入することによって，素性 f_{17} の判定精度を高めることにより誤りを改善できると考えられる．

一方，表 5.8(b)「提案手法により異義と判定し不正解」の例では，素性「f_{17}: フレーズ翻訳テーブルの訳語関係が存在」において，対訳対 ⟨成膜チャンバー, 成膜室⟩ のみがフレーズ翻訳テーブルに含まれることから,「f_{17}: フレーズ翻訳テーブル中に $\langle t_J, s_C \rangle$ または $\langle s_J, t_C \rangle$ の片方の訳語関係のみが存在」の値は 1 となるものの「f_{17}: フレーズ翻訳テーブル中に $\langle t_J, s_C \rangle$, $\langle s_J, t_C \rangle$ 両方の訳語関係が存在」の値が 0 となっている．また，中国語文字列 “成膜” と “膜成形” は実際は同義関係にあるにも関わらず，文字列が逆順となっていることが原因でバイグラム類似度が 0 となっている．主としてこれらが原因となって，最終的に誤って異義と判定されてしまった．この場合，文字列の順序の異なりを反映しない文字列類似度に相当する素性を導入することによって，誤りが改善できると考えられる．

5.4 関連研究

テキストから二言語対訳辞書を獲得する一連の研究の中で，[50] においては，専門用語対訳対の同義判定手法を提案しており，また，手法として分類器学習を適用している．したがって，手法の点においても，また，分類器学習で用いている素性の点においても，本論文の手法と密接に関連している．しかし，[50] においては，同義判定の対象とする専門用語対訳対の収集を手動で行っており，手法の適用範囲が限定される点が短所である．一方，本論文の手法においては，毎年公開される対訳特許文書を情報源として，同義判定の対象とする専門用語対訳対を収集しており，中心的対訳対の選定過程を除けば，その他の全過程がほぼ自動化されている．したがって，本論文の手法においては，[50] と比較した場合の重要な長所として，手法の適用範囲を大幅に拡大できている点を挙げることができる．また，前節で示したように，分類器学習において用いる素性の性能比較の点においても，提案手法の素性によって，[50] で用いられた素性の性能を大幅に改善することが実現できている．

一方，[25] においては，日英パテントファミリーから抽出した日英対訳特許文を対象として日英の同義対訳専門用語の同定を行っている．[25] の手法において同義・異義の専門用語対訳対の候補集合を生成する枠組み，および，分類器学習を適用して同義・異義の同定を行う枠組みは，本論文の枠組みとほぼ同等であるが，[25] は日英同義対訳専門用語同定タスクに対して枠組みを提案するにとどまり，分類器学習における素性の組合わせの網羅的な評価および各素性の有効性に関する

詳細な評価にまでは至っていない点が問題である．一方，本論文においては，[25] において日英を対象として提案された素性に対して，日中を対象とする場合の素性を定義しなおすとともに，[25] における素性の組合わせを改善する形で導入された新たな素性として「フレーズ翻訳テーブルにおける共通訳の割合」を提案した．さらに，分類器学習における素性の組合わせの網羅的な評価および各素性の有効性に関する詳細な評価を行い，新素性である「フレーズ翻訳テーブルにおける共通訳の割合」が性能に大きな影響を持つ重要な素性であることを示した[13]．

その他，本論文で対象とした同義関係の同定に関連して，言い換え知識や翻訳知識を獲得するとともに，それらの知識を利用することにより統計的機械翻訳の翻訳性能やカバレージが改善できることが報告されている [26, 3, 13]．具体的には，訳語が併記されたウェブ文書を情報源として獲得された訳語対を利用することにより統計的機械翻訳の翻訳性能を改善した事例 [26]，および，対訳コーパスを情報源として獲得した言い換え知識を利用することにより，統計的機械翻訳の翻訳性能およびカバレージを改善した事例 [3, 13] が報告されている．これらの成果をふまえると，本論文において同定された同義関係についても，それらを有効に利用することにより，統計的機械翻訳の翻訳性能の改善が期待できる．

5.5 本章のまとめ

本論文では，専門用語対訳対の獲得というタスクにおける同義語同定問題を解決する手法を提案した．提案手法では，対訳特許文および句に基づく統計的機械翻訳モデルのフレーズ翻訳テーブルを用いて専門用語対訳対を自動収集し，それに対して，SVM を適用することにより，専門用語対訳対間の同義・異義関係の判定を行った．日中パテントファミリーから抽出した 360 万対の日中対訳文に対して提案手法を適用し，同義関係にある日中対訳専門用語の同定において，再現率が 25%以上という条件のもとで，約 90%の適合率を達成した．今後の課題として，再現率を改善するため，[25] で提案された，人手の介在を併用する半自動的な同義対訳専門用語の同定の枠組を開発することが重要であると考えられる．

[13]ここで，新素性「フレーズ翻訳テーブルにおける共通訳の割合」は言語対に対して独立な素性であるので，本論文で対象とした日中間の同義対訳専門用語同定タスクだけでなく，[25] における対象である日英間の同義対訳専門用語同定タスクにおいても効果的である可能性は十分にあると考えられる．

第6章　統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラル機械翻訳

6.1　はじめに

近年の機械翻訳の研究分野においては，従来のSMTモデルに代わってNMTモデルによる機械翻訳方式が盛んに研究されている．NMTにおいては，原言語文を固定長ベクトルへ写像し，その固定長ベクトルから目的言語文を生成することから，意味的要素の翻訳に非常に優れており，SMTを上回る翻訳精度を達成している [1, 5, 18, 20, 30, 31, 44]．しかしながら，NMTの弱点の一つとして，扱える語彙に限りがある点が知られている．具体的には，扱う語彙のサイズの増加に伴い，NMTモデルの訓練および翻訳に要する時間が増す点が課題となっている．NMTにおいては，語彙辞書に含まれていない単語はすべて未知語トークンとして扱われるため，これが誤訳となる．日本語特許文をNMTによって中国語に翻訳した場合の誤り例を図 6.1 に示す．日本語入力文に対して，6.5.1 節で述べるベースラインNMTによる中国語翻訳文を，参照用中国語文と比較した．図 6.1 に示すように，略称「cmac」は語彙辞書に含まれていないため，出力文を生成する際に未知語として扱われ，未知語トークン $\langle unk \rangle$ に置き換えられて，誤訳となった．また，日本語単語「ブリッジ」の訳語である中国語単語 “桥架” も語彙辞書に含まれていないため，出力文を生成する際に未知語として扱われ，未知語トークン $\langle unk \rangle$ に置き換えられて，誤訳となった．さらに，入力文に含まれる数字「388」は語彙辞書に含まれていないため未知語として扱われ，結果的に翻訳されず訳抜けとなった[1]．

この問題に対して，これまでにも，NMTが扱える語彙の規模を拡大する方式についての研究がいくつか行われてきた．[42, 57, 28, 6] においては，訓練コーパス中の未知語をサーブワードもしく文字単位に分割することによって，未知語の語彙を減らす手法を提案している．[18] においては，大規模語彙を複数の部分集合に分割し，各語彙の確率を近似的に求めることによって，NMTが扱える語彙数を拡

[1] 入力文の内容の一部が翻訳されずに欠落することを，本論文では「訳抜け」(under-generation) と呼ぶが，この問題もNMTにおける重要な課題の一つである．この課題に対して，[33, 51, 11] においては,「訳抜け」に対処するNMTモデルが提案されている．

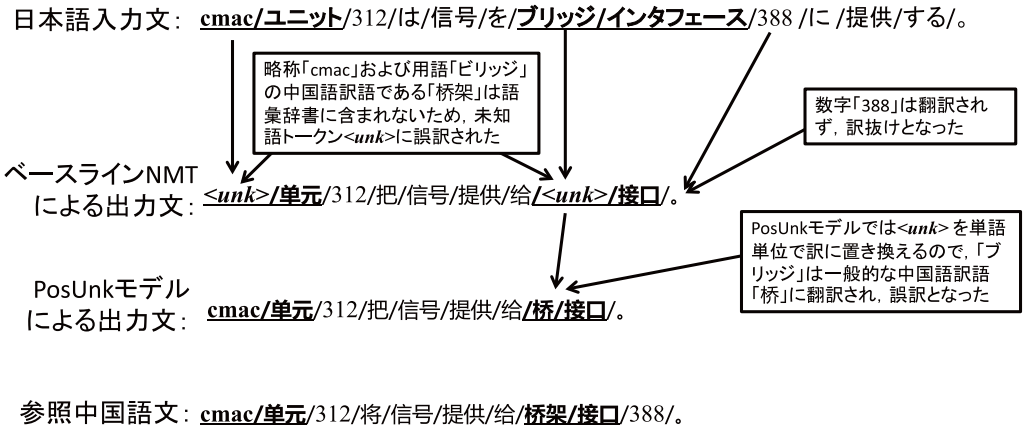


図 6.1: NMT によって日本語特許文を中国語に翻訳した場合の誤り例

大する方式が提案している．[24] においては，未知語を同義となる既知語に置き換えた後，NMT モデルを訓練する方式を提案している．また，[31] においては，二言語間の単語対応関係を記録した未知語トークンを導入し，出力文中のトークンを訳語に置き換える方式によって，NMT モデルにおける語彙の規模を拡大する手法を提案している．ここで，これらの先行研究は，いずれも，未知語となる単語をより細かい単位に分割することによって未知語の問題を回避した後，NMT によって翻訳する，もしくは，単語単位で未知語を訳語に置き換えるというアプローチとなっている．このため，未知語の翻訳の問題を，単語単位の構成的な翻訳の問題に帰着できる場合の対策にとどまっており，複合語的なフレーズの中でも，特に，その構成単語の訳語を構成的に組み合わせる方式に帰着できない非構成的な複合語フレーズ翻訳が取り扱えない点が弱点となっている．例えば，図 6.1 においては，ベースライン NMT において日本語単語「ブリッジ」の中国語訳語が未知語 $\langle unk \rangle$ となるのに対して，単語単位で未知語を訳語に置き換える PosUnk モデル [31] の NMT による翻訳結果においては，訓練文から学習された単語対応によって「ブリッジ」を一般的な中国語訳語 “桥” に置き換えて翻訳した．しかし，この中国語訳語は，日本語複合語フレーズ「ブリッジインタフェース」の参照訳 “桥架接口” における訳語とは異なっており，誤訳となっている．このように，日本語複合語フレーズ「ブリッジインタフェース」の中国語訳においては，構成的に「ブリッジ」の訳語を組み合わせる翻訳は不適切であり，非構成的な複合語フレーズ翻訳が不可避である．

以上の背景のもとで，本章においては，ニューラルネットワーク翻訳において，大規模フレーズ語彙に対応する方式について提案する．本章の提案手法においては，訓練用対訳文においてフレーズの二言語間対応の情報を収集し，二言語間で対応済みのフレーズ対訳対を同一のトークンに置き換えた後，NMT モデルの訓練を行う．翻訳時には，NMT モデルの語彙集合中の語彙部分に対しては，NMT モデルによる訳文生成がなされ，逆に，その他のフレーズまたは単語語彙部分に

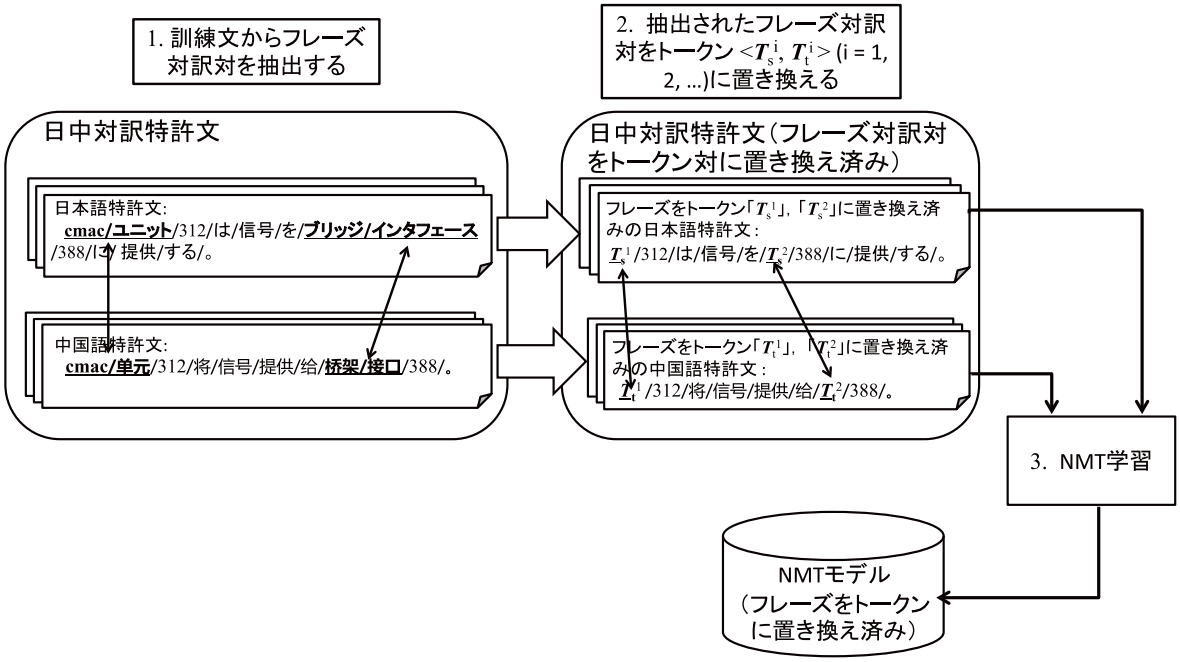


図 6.2: フレーズ対訳対をトークン対に置き換え済みの対訳文を用いたNMT モデルの訓練

対しては，SMT モデルによる翻訳がなされる．日中，中日，日英，英日の各方向の翻訳において評価を行い，提案手法の有効性を検証した．結果として，提案手法においては,「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の組み合わせによって最も高い翻訳性能を達成し，ベースラインであるSMT モデル，および，提案手法が適用されていないNMT モデルとの比較において，0.7 ポイント以上の BLEU の向上を達成できた．さらに，NMT の弱点である訳抜けの改善においては，提案手法が適用されていないNMT モデルによる訳抜けを約 30%減らすことができた．

6.2 大語彙フレーズに対応したNMT システム

6.2.1 大語彙フレーズに対応したNMT モデルの訓練

図 6.2に沿って，対訳文から抽出されたフレーズの対訳対をトークン対 $\langle T_i^s, T_i^t \rangle (i = 1, 2, \ldots)$ に置き換え，その対訳特許文を用いて，大規模フレーズ語彙に対応したNMT モデルを訓練する流れを以下に示す．

ステップ 1 まず，図 6.2 の「1. 訓練文からフレーズ対訳対を抽出する」において，6.3 節の手順によって対訳文 $\langle S_s, S_t \rangle$ の原言語文 S_s からフレーズ t_s^i を抽出する．そして，抽出されたフレーズ t_s^i に対して，3.2.2 節，および，3.2.3 節

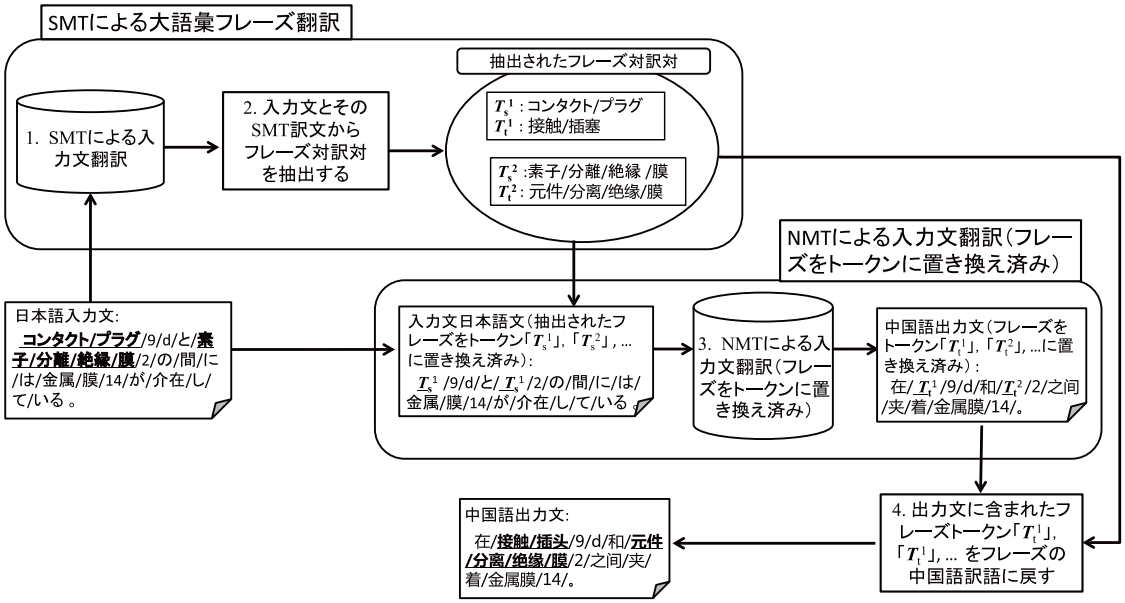


図 6.3: SMT による大語彙フレーズ翻訳とニューラルネットワークによる訳文生成

の手法によって目的言語文における訳語 t_t^i を推定し，フレーズ対訳対 $\langle t_s^i, t_t^i \rangle$ $(i = 1, 2, \ldots, k)$ を抽出する．ここでは，3.2.2 節の手法では訳語推定できないフレーズに対しては，3.2.3 節で述べた訳語推定手法を補完して訳語推定を行う [2] [3].

ステップ 2 次に，図 6.2 の「2. 抽出されたフレーズ対訳対をトークン対に置き換える」において，ステップ 1 の対訳文 $\langle S_s, S_t \rangle$ 中のフレーズ対訳対 $\langle t_s^1, t_t^1 \rangle$, $\langle t_s^2, t_t^2 \rangle, \ldots, \langle t_s^k, t_t^k \rangle$ を，原言語文 S_s における原言語フレーズ t_s^i $(i = 1, 2, \ldots, k)$ の出現順に，それぞれ，トークン対 $\langle T_s^1, T_t^1 \rangle$, $\langle T_s^2, T_t^2 \rangle$, ..., $\langle T_s^k, T_t^k \rangle$ に置き換える．

ステップ 3 最後に，図 6.2 の「3. NMT 学習」において，フレーズ対訳対をトークン対に置き換え済みの対訳文を用いて，フレーズをトークンに置き換え済みの NMT モデルを訓練する．

[2] ここで，フレーズ対訳対 $\langle t_s^i, t_t^i \rangle$ 中の原言語側のフレーズ t_s^i に対しては，6.3 節で述べるフレーズ抽出条件が適用されるが，目的言語側のフレーズ t_t^i に対しては，同様のフレーズ抽出条件の適用は行っていない．これは，原言語側のフレーズ候補に対してフレーズ抽出条件を適用した後，訳語の推定を行うことによって得られる訳語推定結果においては，既に目的言語側のフレーズ抽出条件が満たされていることが多いからである．実際に，予備実験において，目的言語側の訳語候補に対してもフレーズ抽出条件を適用した後，NMT モデルの訓練・評価を行った場合においても，翻訳性能に大きな差はなかった．

[3] 訳語 t_t^i が推定できないフレーズ t_s^i に対しては，以降の手順は行わず，この時点でフレーズ t_s^i を削除する．

6.2.2 大語彙フレーズに対応したNMTモデルを用いた翻訳

図 6.3 に沿って，6.2.1 節の手順によって訓練された，大規模フレーズ語彙に対応した NMT モデルを用いて，原言語文を目的言語訳文に翻訳する流れを以下に示す．

ステップ 1 まず，図 6.3 の上半分のうちの「1. SMT による入力文翻訳」において，原言語文 S_s に対して SMT モデルを適用し，SMT による訳文 S_{SMT} を生成する．

ステップ 2 次に，図 6.3 の「2. 入力文とその SMT 訳文からフレーズ対訳対を抽出する」において，6.3 節の手順によって入力文 S_s からフレーズ t_s^i を抽出する．そして，抽出されたフレーズ t_s^i に対して，3.2.2 節，および，3.2.3 節の手法によって，SMT による訳文 S_{SMT} における訳語 t_{SMT}^i を推定し，フレーズ対訳対 $\langle t_s^i, t_{SMT}^i \rangle (i = 1, 2, \ldots, k)$ を抽出する．ここでは，3.2.2 節の手法では訳語推定できないフレーズに対しては，3.2.3 節で述べた訳語推定手法を補完して訳語推定を行う [4]．

ステップ 3 ステップ 2 において抽出されたフレーズ対訳対 $\langle t_s^i, t_{SMT}^i \rangle$ $(i = 1, 2, \ldots, k)$ に対して，原言語文 S_s 中の原言語フレーズ $t_s^i (i = 1, 2, \ldots, k)$ をトークン $T_s^i (i = 1, 2, \ldots, k)$ に置き換える．

スッテプ 4 次に，図 6.3 の「3. NMT による入力文翻訳」において，フレーズ対訳対をトークン対に置き換えた後に訓練された NMT モデルを用いて，フレーズをトークンに置き換え済みの原言語文の訳文を生成する．翻訳結果の NMT 訳文においては，トークン T_s^i $(i = 1, 2, \ldots, k)$ はトークン T_t^i $(i = 1, 2, \ldots, k)$ に翻訳されており，原言語フレーズの訳語を挿入する位置が，訳文における各トークンの位置によって特定されている．

ステップ 5 最後に，図 6.3 の「4. 出力文に含まれたトークンをフレーズの訳語に戻す」において，ステップ 2 において抽出されたフレーズ対訳対 $\langle t_s^i, t_{SMT}^i \rangle$ $(i = 1, 2, \ldots, k)$ を用いて，NMT による訳文中のトークン T_t^i $(i = 1, 2, \ldots,)$ を，それぞれ，SMT による訳文 S_{SMT} から抽出した訳語 t_{SMT}^i $(i = 1, 2, \ldots, k)$ に置き換えて，最終的な訳文とする．

6.3 フレーズの抽出

本節では，前節においてトークンへの置き換え対象となる単言語フレーズを抽出する手法について述べる．

[4] 前節の NMT モデル訓練時の場合と同様に，訳語 t_{SMT}^i が推定できないフレーズ t_s^i に対しては，以降の手順は行わず，この時点でフレーズ t_s^i を削除する．

本節の手法においては，まず，訓練文および評価文の原言語側からフレーズを抽出するが，その具体的な手法としては，branching entropy を用いたフレーズ抽出 (6.3.1 節)，言語知識に基づき選定した名詞句フレーズの抽出 (6.3.2 節)，および，*C-value* を用いた名詞句部分集合のフレーズ抽出 (6.3.3)，の三種類の手法について述べる．

6.3.1 branching entropy を用いたフレーズ抽出

従来より，branching entropy を用いた手法は，文中におけるフレーズ分割 [19]，および，キーワードフレーズの抽出 [4] 等においてよく用いられてきた．それらの先行研究をふまえて，本論文では，left branching entropy (フレーズ候補の左に隣接する位置における branching entropy)，および，right branching entropy (フレーズ候補の右に隣接する位置における branching entropy) を用いてフレーズの境界を判定し，条件を満たすフレーズを抽出する．

フレーズ t に対して，フレーズ t の左に隣接する形態素・単語の集合を $V_l(t)$，フレーズ t の右に隣接する形態素・単語の集合を $V_r(t)$ とすると，left branching entropy および right branching entropy は次式で定義される．

$$H_l(t) = -\sum_{v \in V_l(t)} P_l(v|t) \log_2 P_l(v|t)$$

$$H_r(t) = -\sum_{v \in V_r(t)} P_r(v|t) \log_2 P_r(v|t)$$

ただし，形態素・単語の列 x の訓練文中における頻度を $f(x)$ として，条件付き確率 $P_l(v|t)$，および，$P_r(v|t)$ は，それぞれ次式で定義される

$$P_l(v|t) = \frac{f(v,t)}{f(t)} \qquad P_r(v|t) = \frac{f(t,v)}{f(t)}$$

以上の定義をふまえると，フレーズ t が別のより長いフレーズの部分列の場合には，t の左もしくは右には特定の形態素・単語が隣接するため，branching entropy の値が小さくなるという傾向を示す．図 6.1 の日本語形態素列「ブリッジ」を例として挙げると,「ブリッジ」は複合語「ブリッジインターフェイス」の一部であるため，その右に隣接する形態素の出現例の大部分は「インターフェイス」となるため,「ブリッジ」の branching entropy の値は小さくなる．一方，t そのものが他のフレーズの部分列ではなく，それ自身がフレーズを形成する場合，t の左右に隣接する形態素・単語の種類・頻度が多様となるため，branching entropy の値が大きくなるという傾向を示す．図 6.1 の日本語形態素列「ブリッジインターフェイス」を例として挙げると，左右に隣接する形態素・単語は,「を」,「388」等，その種類・頻度が多様であるため,「ブリッジインターフェイス」の branching entropy の値は大きくなる．

以上の議論をふまえて，本研究では，branching entropy の値に下限値を設定し，以下の条件をすべて満たすフレーズ t を抽出する．

(i) $H_l(t)$ および $H_r(t)$ の値は下限値以上であり，t の任意の部分列を t_{sub} とすると [5]，$H_l(t_{sub})$ および $H_r(t_{sub})$ の値は下限値未満である．

(ii) t には記号 [6] を含まない．

(iii) 頻度降順上位の形態素・単語をストップワードとして，t は，ストップワード中の高頻度形態素・単語を含まない [7]．

6.3.2 言語知識に基づき選定した名詞句フレーズの抽出

本論文では，日本語，中国語，および，英語の各言語において，言語知識に基づく品詞タグ付け・形態素解析ツールを用いて各形態素・単語に付与した品詞情報を利用して，各言語における名詞句を抽出するための品詞パターンを定義し，品詞パターンを満たす最長の形態素・単語列を名詞句とみなし，抽出対象フレーズとする．各言語における名詞句の品詞パターンの詳細を以下に述べる．

日本語 [58] における日本語の名詞句の抽出手法を改善して，本研究では，日本語の名詞句を以下のように定義する．

$$(\text{名詞} \mid \text{接頭辞} \mid \text{動詞} \mid \text{形容詞})^{+}\text{名詞}$$

ここで，日本語文の品詞タグ付けツールとして，品詞体系および形態素辞書として IPAdic を用いた Mecab を利用した．ただし，以下に該当する日本語フレーズを対象外とした．

(i) 表 6.1 の日本語ストップワードを含む日本語フレーズ．

(ii) 記号 [6] およびアラビア数字を含む日本語フレーズ．

(iii) 一形態素から構成され，文字長が 5 文字以下の日本語フレーズ．

[5] t_1 および t_2 を一つ以上の形態素・単語から構成されるフレーズとして，$t = t_1\ t_{sub}$，または，$t = t_{sub}\ t_2$，または，$t = t_1\ t_{sub}\ t_2$ のいずれかが成り立つ場合に，t_{sub} は t の部分列であるとし，$t \sqsupset t_{sub}$ と表記する．

[6] 1,215 種類の記号のストップワードリストを人手で作成して用いた．なお，6.3.2 節の「言語知識に基づき選定した名詞句フレーズの抽出」において，日本語，中国語，英語の三言語において記号を含むフレーズを抽出対象から除外する際にも，この記号のストップワードリストを三言語共通の記号として用いた．

[7] 本研究では，頻度降順の上位 100 形態素・単語をストップワードとした．

表 6.1: 言語知識に基づき選定した名詞句フレーズの抽出において用いるストップワードリスト

言語	出現位置	ストップワード	種類数
日本語	名詞句中の任意の位置	上記，前記，下記，当該，該，以上，以下，以外，以内，各種，場合，以後，これら，すべて，その他，その後，それぞれ，それら，ため，うち，する，そのもの，各々，個々，こと，ゆえ，まわり	27
	語頭	本，各，号，用，型，化，形，毎	8
	語尾	等，内，外，中，上，下，的，用，共，前，後，不	12
中国語	名詞句中の任意の位置	上述(上述した)，那样(あのように)，这样(このように)，也(も)，所述(に述べた)，并(それに)，且(それに)，并且(それに)，到(まで)，通过(に通じて)，此(この)，其(その)，之一(そのうち)，将(…を…)，而(しかし)，其中(そのうち)，进行(行う)，只有(のみ)，能够(できる)，又(また)，则(…は…)，形成(形成する)，该(当該)，是(…が)，特别是(特に)，合计(合計)，前述(前述した)，首先(まず)，使用(用いる)，下述(次にのべる)，后述(次にのべる)，从而(従って)，从其(それより)，但是(しかし)，于此(ここから)，以下(ここから)，然后(そして)，经过(経つ)，例如(例えば)，即(すなわち)，所得(得られた…)，如(のように)，作为(として)，因此(そのため)，稍许(すこし)，所以(だから)，至少(すくなくとも)，由此(これより)，更(さらに)，足够(足りる)，诸如(例えば)，然而(だが)，所谓(いわば)，后(後)	54
	語頭	是(…が…)，如(のように)，其(それ)，此(これ)，这(これ)，于(より)，而(それに)，仅(のみ)，由(より)	9
	語尾	–	0
英語	名詞句中の任意の位置	very, much, same, similar, great, numerous, several, year, just, good, such, other, another, first, second, third, fourth, fifth, next, lower, upper, invention, embodiment, section, step, left, right, correct, incorrect	29
	語頭	–	0
	語尾	–	0

中国語 [58] における中国語の名詞句の抽出手法を改善して，本研究では，中国語の名詞句を以下のように定義する．

$$((VA|JJ)|AD|(NR|NN|NT))^{+}(NR|NN|NT)$$

ここで，中国語文の品詞タグ付けツールとして，Chinese Penn Treebank を用いて訓練された Stanford Parser [43] を利用した．なお，日本語における品詞としては，$(VA|JJ)$ は形容詞に，AD は副詞に，$(NR|NN|NT)$ は名詞に，それぞれ対応する．ただし，以下に該当する中国語フレーズを対象外とした．

(i) 表 6.1 の中国語ストップワードを含む中国語フレーズ．

(ii) 記号[6] およびアラビア数字を含む中国語フレーズ．

(iii) 文字長が 3 文字以下の中国語フレーズ.

英語 [8] における英語の名詞句の抽出手法を改善して，本研究では，英語の名詞句を以下のように定義する.

$$((JJ|JJR|JJS)|(VB|VBG|VBN)|(NN|NNS|NNP))^{+}(NN|NNS|NNP)$$

ここで，英語文の品詞タグ付けツールとして，English Penn Treebank を用いて訓練された Stanford Parser [43] を用いた．なお，日本語における品詞としては，$(JJ|JJR|JJS)$ は形容詞に，$(VB|VBG|VBN)$ は動詞，$(NN|NNS|NNP)$ は名詞に，それぞれ対応する．ただし，以下に該当する英語フレーズを対象外とした.

(i) 表 6.1 の英語ストップワードを含む英語フレーズ.

(ii) 記号 [6] およびアラビア数字を含む英語フレーズ.

(iii) 一単語から構成される英語フレーズ.

6.3.3 *C-value* を用いた名詞句部分集合のフレーズ抽出

[8] においては，英語文テキストから名詞句を抽出することを目的として，言語知識を利用して品詞パータンによる制約を課した後，統計的尺度の一つとして *C-value* を定義して，*C-value* の値が上位となる名詞句を抽出する手法を提案している．本節では，[8] をふまえて，6.3.2 節において言語知識に基づき選定した名詞句フレーズをフレーズ候補集合 U として，U 中の各候補フレーズ $t(t \in U)$ の *C-value* の値を計算し，*C-value* の値が下限値を満たす候補フレーズを抽出対象フレーズとする．*C-value* の定式化においては，まず，フレーズ $t(t \in U)$ を部分列として含むフレーズの集合を次式の $U_{\sqsupset}(t)$ で定義した後,

$$U_{\sqsupset}(t) = \{t' \in U \mid t' \sqsupset t\} \tag{6.1}$$

フレーズ t の $C\text{-}value(t)$ を次式で定義する.

$$C\text{-}value(t) = \begin{cases} log_2 |t| \cdot f(t) & (U_{\sqsupset}(t) = \emptyset \text{の場合}) \\ log_2 |t| \left(f(t) - \frac{1}{|U_{\sqsupset}(t)|} \sum_{t' \in U_{\sqsupset}(t)} f(t') \right) & (\text{その他の場合}) \end{cases}$$

ただし，$|t|$ はフレーズ t の形態素長または単語長，$f(t)$ は全訓練文におけるフレーズ t の出現頻度を表す．$C\text{-}value(t)$ の値はフレーズ t のフレーズらしさを表し，この値が大きいほど t のフレーズらしさがより大きい．本論文では，$C\text{-}value(t)$ の値に下限値を設定し，6.3.2 節において抽出された名詞句集合 U のうち下限値を満たす名詞句の部分集合を選定して，フレーズ集合として抽出する.

表 6.2: 検証実験に用いられた対訳文の文数

	訓練用 対訳特許文	開発用 対訳特許文	評価用 対訳特許文
日中対訳文	998,054	2,000	2,000
日英対訳文	999,636	2,000	2,000

6.4 評価手順

6.4.1 データセットの詳細

本論文では，3.1.1 節および 3.1.2 節で述べた対訳特許文のうち，WAT2017 の特許翻訳タスクにおいて配布された 100 万文の日中対訳特許文，および，100 万文の日英対訳特許文を用いて，提案手法の検証実験を行った．日本語文に対しては，IPAdic を用いた Mecab による形態素解析を行い，一形態素を単語の単位とした．中国語文に対しては，Stanford Word Segment [46] によって形態素解析を行い，一形態素を単語の単位とした．英語文に対しては，Moses の Tokenizer ツール [8] を用いて，英語文中の記号部分の分割を行った．ここでは，Moses [22] による SMT モデル作成の際の制約から，一文中の形態素数・単語数が 100 を超える訓練用対訳文を除外した．検証実験において用いた日中・日英対訳特許文の文対数を表 6.2 に示す．

6.4.2 SMT モデルにおける設定

本節では，本章で用いる SMT モデルの設定について述べる．

まず，訓練用対訳特許文に対して句に基づく統計的機械翻訳モデルのツールキットである Moses [22](バージョン 2.1) を適用することにより，SMT モデルを訓練する．翻訳モデル訓練時には以下の設定を用いる．まず，3.2 節と同様に，パラメータ alignment を grow-diag-final-and として，単語対応およびフレーズ翻訳テーブルを作成する．また，フレーズ翻訳テーブルに含まれるフレーズの形態素数・単語数の上限 (パラメーター max-phrase-length) は 7 とする．さらに，パラメーター reordering を wbe-msd-bidirectional-fe として，語順並べ替え確率モデルを作成する．言語モデルの作成においては，訓練用対訳特許文の原言語文に対して Moses の KenLM ツール [14] を適用し，5-gram の言語モデルを作成する．そして，開発用対訳特許文を用いて，繰り返し回数 (パラメーター maximum-iterations) を 8 として SMT モデルのチューニングを行う．SMT モデルの訓練，および，チューニ

[8] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

ングにおけるその他の設定としては，デフォルトのものをそのまま用いる．また，SMT モデルを用いた翻訳時においては，1-best 訳文を SMT 訳文として用い，その他の設定はデフォルトのものをそのまま用いた．評価実験におけるベースライン SMT としても，本節で述べた SMT モデルをそのまま用いた．

6.4.3 大語彙フレーズに対応した NMT モデルにおける設定

本節では，6.2 節で述べた「大語彙フレーズに対応した NMT モデル」における設定の詳細について述べる．

フレーズ抽出およびトークンへの置き換え

6.2 節においては，NMT モデルの訓練 (6.2.1 節)，および，NMT モデルを用いた翻訳 (6.2.2 節) のいずれにおいても，6.3 節の手順によってフレーズを抽出した後，フレーズ対訳対のトークン対への置き換え (6.2.1 節)，もしくは，原言語フレーズのトークンへの置き換え (6.2.2 節) がなされる．本論文では，6.3 節におけるフレーズ抽出手法としては，

(a1) branching entropy を用いたフレーズ抽出 (6.3.1 節)

(a2) 言語知識に基づき選定した名詞句フレーズの抽出 (6.3.2 節)

(a3) *C-value* を用いた名詞句部分集合のフレーズ抽出 (6.3.3 節)

の三通りの手法を評価する．ここで，各フレーズの branching entropy の値，あるいは，*C-value* の値を算出するための確率パラメータは，訓練用対訳特許文に基づき算出する．また，branching entropy の下限値，および，*C-value* の下限値は，開発用対訳特許文に対する最適値として以下の値を用いる．

- branching entropy の下限値は，日中・中日方向は 5，日英・英日方向は 8 とする．
- *C-value* の下限値は，どの言語対においても 5 とする．

なお，本論文では，フレーズ対訳対のトークン対への置き換え (6.2.1 節) におけるステップ 2，および，原言語フレーズのトークンへの置き換え (6.2.2 節) におけるステップ 3 においてトークンへの置き換え対象とするフレーズの選定基準として，

(b1) **「未知語を含むフレーズ対訳対」が置き換え対象**
NMT モデルにおける語彙外となる未知語を原言語側もしくは目的言語側に含むフレーズに限定する場合

(b2) **「全フレーズ対訳対」が置き換え対象**
そのような限定をせず全フレーズを対象とする場合

の二通りの評価を行う．具体的には，本論文では，NMT モデルの訓練にあたって，原言語と目的言語の語彙数を，それぞれ，訓練用対訳特許文における頻度上位の 40,000 語とする．そこで，原言語および目的言語における NMT モデルの語彙集合を，それぞれ，V_s^{40K}，および，V_t^{40K} と表記する．すると，上述の二通りの場合分け (b1) および (b2) は，

(b1) **「未知語を含むフレーズ対訳対」が置き換え対象**
原言語フレーズ中に NMT モデルの語彙外となる単語 $w_s(\notin V_s^{40K})$ を少なくとも一つ含むフレーズ，もしくは，目的言語フレーズ中に NMT モデルの語彙外となる単語 $w_t(\notin V_t^{40K})$ を少なくとも一つ含むフレーズに限定する場合

(b2) **「全フレーズ対訳対」が置き換え対象**
そのような限定をせず全フレーズを対象とする場合

となる．

以上の (a1)，(a2)，(a3) の三通り，および，(b1)，(b2) の二通りの組み合わせ，合計 6 通りの設定において抽出されたフレーズ対訳対の異なり数，および，延べ数を表 6.3 に示す [9]．

NMT モデルの設定

本論文では，NMT モデルの実装においては，オープンソースの深層学習ツールキットである OpenNMT[10] の前身である seq2seq-attn ツール [11] を利用した．

原言語と目的言語の語彙数を，それぞれ，訓練用対訳特許文における頻度上位の 40,000 語とし，その他の語は，NMT モデルの語彙外となる未知語とする．6.2 節の手順の後，フレーズに対するトークンに置き換えられていない未知語は，「未知語を表すトークン」$\langle unk \rangle$ に置き換える．LSTM の隠れ層の階層数を 3 とし，符号化部の前向き・後向きの LSTM，復号化部の LSTM，および，入力単語の分散表現はいずれも次元数を 512 とする．また，NMT モデル訓練時には以下の設定を用いる．

[9]「*C-value* を用いた名詞句部分集合のフレーズ抽出 (6.3.3 節)」の場合のみ，NMT モデルを用いた翻訳時には，評価用対訳特許文の原言語中のフレーズ (特に，訓練用対訳特許文中に出現しないフレーズの場合) が，*C-value* の値として安定した値をとるのが難しいため，「*C-value* を用いた名詞句部分集合のフレーズ抽出 (6.3.3 節)」は，NMT モデルの訓練時のみ適用することとし，NMT モデルを用いた翻訳時には，「言語知識に基づき選定した名詞句フレーズの抽出 (6.3.2 節)」によるフレーズ候補集合をそのまま用いる．このため，表 6.3(b) 中では，「*C-value* を用いた名詞句部分集合のフレーズ抽出 (6.3.3 節)」欄を削除している．

[10] `http://opennmt.net/`

[11] `https://github.com/harvardnlp/seq2seq-attn`

表 6.3: 抽出されたフレーズ対訳対の異なり数 / 延べ数

(a) NMT モデルの訓練時 (訓練用対訳特許文の全体が対象)

抽出手法	置き換え対象	日中方向	中日方向	日英方向	英日方向
branching entropy を用いたフレーズ抽出 (6.3.1 節)	未知語を含むフレーズ対訳対	69,387 / 102,630	96,468 / 124,992	35,544 / 38,457	34,605 / 37,275
	全フレーズ対訳対	944,734 / 2,626,634	899,287 / 2,099,185	861,577 / 1,934,857	877,856 / 1,700,765
言語知識に基づき選定した名詞句フレーズの抽出 (6.3.2 節)	未知語を含むフレーズ対訳対	88,621 / 126,670	68,542 / 95,992	38,897 / 42,741	31,345 / 34,321
	全フレーズ対訳対	632,329 / 2,794,798	447,757 / 1,979,135	588,887 / 2,013,531	642,377 / 1,653,645
C-value を用いた名詞句部分集合のフレーズ抽出 (6.3.3 節)	未知語を含むフレーズ対訳対	39,294 / 72,049	11,095 / 27,321	11,796 / 13,668	2,828 / 3,518
	全フレーズ対訳対	361,922 / 2,462,980	214,684 / 1,685,096	236,116 / 1,592,719	216,190 / 1,152,191

(b) NMT モデルを用いた翻訳時 (評価用対訳特許文の原言語文全体が対象)

抽出手法	置き換え対象	日中方向	中日方向	日英方向	英日方向
branching entropy を用いたフレーズ抽出 (6.3.1 節)	未知語を含むフレーズ対訳対	247 / 260	417 / 382	221 / 249	230 / 246
	全フレーズ対訳対	5,378 / 6,639	4,918 / 5,654	3,993 / 4,827	3,468 / 4,041
言語知識に基づき選定した名詞句フレーズの抽出 (6.3.2 節)	未知語を含むフレーズ対訳対	378 / 419	448 / 486	288 / 359	171 / 181
	全フレーズ対訳対	4,802 / 6,859	3,666 / 5,051	3,622 / 5,125	3,121 / 3,902

1. モデル中のパラメータは，$[-0.1, 0.1]$ における一様分布からの乱数によって初期化する．
2. ミニバッチサイズを 128 とする．
3. モデル訓練時のエポック数を 10 とする．
4. 各パラメータの訓練時には確率的勾配降下法を用い，初期学習率を 1 とする．7 エポック目以降はエポックごとに学習率を半減する．
5. 勾配のノルムに対して，その上限値を 5 として正規化を行う [44]．
6. 隠れ層の各層間では確率 30%でドロップアウトを行う．

一方，NMT モデルを用いた訳文探索におけるビーム幅を 1 とした．また，評価実験におけるベースライン NMT としては，本節で述べた大語彙フレーズへの対応はしていないが，その他の設定については本節のものをそのまま用いた NMT モデルを用いた．

なお，GPU (GTX 1080) を 1 枚有するサーバマシンを用いた場合には，NMT モデルの訓練に要する時間は約 1 日である．

6.5 評価結果

6.5.1 文単位の翻訳性能の評価

自動評価

本論文では，文単位の翻訳精度の自動評価尺度として BLEU スコア [41] を用いる．各手法に対する自動評価の結果を表 6.4 に示す．全体的な傾向として，提案手法の中では，「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定における NMT モデルが最も高い BLEU スコアとなり [12]，ベースライン SMT，ベースライン NMT，および，単語単位で未知語を訳語に置き換える PosUnk モデル [31] の NMT の翻訳性能を上回った．

また，訓練コーパス中の未知語をサーブワードに分割することによって未知語の語彙を減らす手法を適用した NMT モデル [42, 57] との比較を行った．これらの手法のうち，[42] では，語彙数に基づく貪欲法によって語彙の単位を細分化する

[12] 文単位の翻訳性能の人手評価 (本節)，および，訳抜けの改善の評価 (6.5.2 節) のいずれにおいても，提案手法の中では，「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定における NMT モデルが相対的に最も高い性能であったため，6.5 節の表 6.6，表 6.7，および，表 6.8 の評価結果においては，「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定における NMT モデルによる評価結果のみを示す．

表 6.4: 自動評価の結果 (BLEU)

(a) ベースライン

手法	日中方向	中日方向	日英方向	英日方向
ベースライン SMT [22]	30.0	36.2	28.0	29.4
ベースライン NMT(2.3 節)	34.2	40.8	43.1	41.8
Sentence Piece NMT (原言語と目的言語の 語彙数はそれぞれ 30,000)	35.0	41.0	43.3	41.8
PosUnk モデル [31] による NMT	34.5	41.0	43.5	42.0

(b) 提案手法

フレーズ抽出手法	置き換え対象	日中方向	中日方向	日英方向	英日方向
branching entropy を用いたフレーズ抽出 (6.3.1 節)	未知語を含むフレーズ対訳対	**35.6**	**41.6**	**43.9**	**42.5**
	全フレーズ対訳対	34.2	39.9	40.0	40.1
言語知識に基づき選定した名詞句フレーズの抽出 (6.3.2 節)	未知語を含むフレーズ対訳対	35.3	41.5	43.7	**42.7**
	全フレーズ対訳対	35.5	41.5	40.8	40.1
C-value を用いた名詞句部分集合のフレーズ抽出 (6.3.3 節)	未知語を含むフレーズ対訳対	**35.7**	41.3	43.2	41.9
	全フレーズ対訳対	34.8	41.2	41.5	41.7

手法を提案している．一方，[57] および Sentence Piece ツール [13] においては，エントロピー最尤推定の目的関数を用いて語彙の単位を細分化する手法を提案している [14]．これらのうち，本節では，Sentence Piece ツールを用いて決めた語彙単位に対して，ベースライン NMT モデルと同一の設定での NMT モデル [15] を比較対象として，翻訳性能の評価を行った結果を表 6.4 「Sentence Piece NMT」に示す．この結果においても，提案手法のうちの「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定における NMT モデルの方が高い翻訳性能を達成した．

また，表 6.4(b) において，「「全フレーズ対訳対」が置き換え対象」の場合，および，「「未知語を含むフレーズ対訳対」が置き換え対象」の場合の評価結果を比較した結果においては，全体的な傾向として，「「未知語を含むフレーズ対訳対」が置き換え対象」の場合の方が高い翻訳性能を達成しており，その傾向は，日中・中日方向よりも日英・英日方向の方が顕著である．このことの主たる要因は，フレーズ対訳対抽出の際に，フレーズ内部の構文構造が二言語間で異なること等が原因で，フレーズ対訳対抽出の誤りが多く発生し，その結果，フレーズ対訳対をトークンに置き換えた後訓練された NMT モデルの翻訳性能が低下する点にある．フレーズ対訳対抽出の誤り箇所数は，「「全フレーズ対訳対」が置き換え対象」の場合の方が，「「未知語を含むフレーズ対訳対」が置き換え対象」の場合よりもはるかに多く，結果的に，その後訓練された NMT モデルの翻訳性能への影響が多く出ている．また，フレーズ内部の構文構造の二言語間の違いは，日中間よりも日英間の方が大きいことから，日中・中日方向よりも日英・英日方向の方が，「「全フレーズ対訳対」が置き換え対象」の場合と「「未知語を含むフレーズ対訳対」が置き換え対象」の場合との間の翻訳性能の差が大きくなった．

人手評価

本論文の人手評価においては，[36] における人手評価尺度である「一対評価」および「JPO 基準に基づく絶対評価」の二種類の人手評価尺度を用いる．評価用対訳特許文から無作為に抽出された 200 文を評価対象として，本論文の著者が評価を行う．

評価対象手法，および，ベースラインとなる手法との間の「一対評価」においては，評価対象手法による翻訳精度が，ベースラインとなる手法による翻訳精度

[13] https://github.com/google/sentencepiece

[14] [57] および Sentence Piece ツールの違いとして，[57] においては，語彙の単位は単語境界を超えない範囲の中から決められるが，Sentence Piece ツールにおいては，単語間の空白文字等の単語境界を超えて語彙の単位を決めることができる点が異なる．また，Sentence Piece ツールの開発者からの報告においては，[42] を適用した NMT モデルによる翻訳性能よりも高い性能が得られている．

[15] 原言語，および，目的言語の語彙数をそれぞれ 30,000 と設定したうえで，Sentence Piece ツールを用いて訓練用対訳特許文を語彙単位に分かち書きした後，NMT モデルの訓練を行う．

表 6.5: 内容の伝達レベルに対する JPO 評価基準

5:	すべての重要情報が正確に伝達されている．(100%)
4:	ほとんどの重要情報は正確に伝達されている．(80% ～ 100%未満)
3:	半分以上の重要情報は正確に伝達されている．(50% ～ 80%未満)
2:	いくつかの重要情報は正確に伝達されている．(20% ～ 50%未満)
1:	文意がわからない，もしくは正確に伝達されている重要情報はほとんどない．(20%未満)

を上回った文の数を W，評価対象手法による翻訳精度が，ベースラインとなる手法による翻訳精度を下回った文の数を L，評価対象手法による翻訳精度が，ベースラインとなる手法による翻訳精度と同等となった文の数を T として，一対評価のスコア (値の範囲は，$-100 \sim 100$) を次式で定義する．

$$score = 100 \times \frac{W - L}{W + L + T}$$

PosUnk モデル [31] による NMT，Sentence Piece NMT，および，提案手法のうちの「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定における NMT モデルをそれぞれ評価対象手法とした場合の「一対評価」結果を表 6.6 に示す．この結果においては，提案手法が最も高い評価結果となった．

「JPO 基準に基づく絶対評価」においては，表 6.5 の JPO 評価基準 [16] に基づき，各翻訳文に対して人手で $1 \sim 5$ の値の範囲のスコアを付与し，その平均を「JPO 基準に基づく絶対評価」のスコアとする．表 6.7 の評価結果に示すように，提案手法が最も高い評価結果となった．

6.5.2 訳抜けの改善の評価

翻訳文における訳抜けの問題の改善度合いを評価するために，評価用対訳特許文に対して，訳抜けした形態素・単語の数を人手で集計し，提案手法，および，比較対象の NMT モデルとの間で比較した．

まず，入力文 S_s，および，評価対象手法による翻訳文 S_t^o の対 $\langle S_s, S_t^o \rangle$ に対して，以下の条件を満たす形態素・単語 $w_s (w_s \in S_s)$ を訳抜けと定義して，人手でその数を集計する．

[16] https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/tokkyohonyaku_hyouka/01.pdf

表 6.6: 一対評価結果 (ベースライン NMT との比較，スコアの範囲: $-100 \sim 100$)

手法	日中方向	中日方向	日英方向	英日方向
PosUnk モデル [31] による NMT	13	12.5	9.5	14.5
Sentence Piece NMT (原言語と目的言語の語彙数はそれぞれ 30,000)	19.0	18	11.5	16
提案手法 (「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象)	**23.5**	**22.5**	**15.5**	**19**

w_s の訳語が翻訳文 S_t^o 中に存在するか否かを判定し，存在しない場合に訳抜けとみなす[17]．ただし，w_s が未知語トークン $\langle unk \rangle$ に翻訳された場合は，w_s の訳語は翻訳文 S_t^o 中に存在するとみなす．

集計結果を表 6.8 に示す．この結果においては，提案手法によって，ベースライン NMT における訳抜けのうちの約 30%が削減された．また，PosUnk モデルによる NMT，および，Sentence Piece NMT と比較しても，提案手法による訳抜けの数の方が少なくなっている．

Sentence Piece NMT においては，語彙の単位を細分化し，未知語を削減することによって，一部の訳抜けを改善したが，一文を構成する形態素・単語の数が多くなったことにより，NMT モデル自身における訳抜けの発生し易さが増し，訳抜けの原因となった．一方，提案手法においては，未知語を含み，複数形態素・単語から構成されるフレーズを一つのトークンに置き換えることによって，一文を構成する形態素・単語の数が抑えることができ，結果的に訳抜けの数を抑えられている．

6.5.3 名詞句翻訳性能の評価

本節では，本論文の大語彙フレーズに対応した NMT において，大語彙フレーズの典型例である名詞句の翻訳性能がどの程度改善したのかの評価を行う．まず，

[17]英語文の冠詞は通常日本語訳文中では陽には翻訳されないので，訳抜けとはみなさない．また，形態素・単語 w_s は直訳はされていないが，翻訳文中に w_s の意味が含まれている場合には w_s を訳抜けとはみなさない．

表 6.7: JPO 基準に基づく絶対評価結果 (スコアの範囲: 1 ～ 5)

手法	日中方向	中日方向	日英方向	英日方向
ベースライン SMT [22]	3.1	3.2	2.9	3.0
ベースライン NMT(2.3 節)	3.6	3.6	3.7	3.7
PosUnk モデル [31] による NMT	3.8	3.9	3.9	3.9
Sentence Piece NMT (原言語と目的言語の 語彙数はそれぞれ 30,000)	3.9	3.9	4.0	3.9
提案手法 (「branching entropy を用いたフレーズ抽出」 および「未知語を含む フレーズ対訳対」 が置き換え対象)	**4.1**	**4.1**	**4.2**	**4.1**

6.5.1 節の人手評価において評価対象とした 200 文に対して，全ての名詞句を人手で抽出し，評価対象名詞句とする．さらに，それらの評価対象名詞句の訳語名詞句を，200 文の評価用対訳特許文における参照用目的言語文中から抽出し，各評価対象手法に対してそれらの訳語名詞句の再現率を測定した．ここで，評価対象名詞句としては，(a) 構成形態素・単語として未知語を含むため，ベースライン NMT による翻訳が不可能である「未知語を含む評価対象名詞句」，(b) 「全評価対象名詞句」[18]，の二種類の集合を作成し，個別に評価を行った．これらの二種類の集合の作成手順を以下に示す.

(a) **未知語を含む評価対象名詞句**
6.4.3 節で定義された NMT モデルの語彙集合 V_s^{40K} に含まれない形態素・単語 $w_s(\notin V_s^{40K})$ を，構成形態素・単語として少なくとも一つ含む名詞句，もしくは，翻訳文中において，当該名詞句の訳語名詞句を人手で同定し，その訳語名詞句の構成形態素・単語として，NMT モデルの語彙集合 V_t^{40K} に含まれない形態素・単語 $w_t(\notin V_t^{40K})$ を少なくとも一つ含む名詞句に限定する場合

(b) **全評価対象名詞句**
そのような限定をせず全名詞句を対象とする場合

[18] このうちの大半は，未知語を含まない名詞句である．そのうち，構成形態素・単語の訳語の連結によって名詞句全体の訳語が合成不可能な非構成的な名詞句については，ベースライン NMT による翻訳が不可能であり,「全評価対象名詞句」の一部はそのような非構成的な名詞句から構成される.

表 6.8: 入力文における訳抜けの形態素・単語の数

手法	日中方向	中日方向	日英方向	英日方向
ベースライン NMT(2.3 節)	1,135	869	1,060	813
PosUnk モデル [31] による NMT	1,112	846	1,031	794
Sentence Piece NMT (原言語と目的言語の語彙数はそれぞれ 30,000)	871	717	796	683
提案手法 (「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象)	**736**	**581**	**655**	**571**

表 6.9: 名詞句翻訳性能の評価結果 (対象: 未知語を含む評価対象名詞句，ベースライン手法，異なり数に対しての再現率 / 延べ数に対しての再現率)

手法	日中方向	中日方向	日英方向	英日方向
ベースライン SMT [22]	28.1% (18/64) / 26.9% (18/67)	23.4% (15/64) / 23.9% (16/67)	35.4% (17/48) / 37.7% (20/53)	22.9% (11/48) / 26.4% (14/53)
ベースライン NMT(2.3 節)	3.1% (2/64) / 3.0% (2/67)	6.3% (4/64) / 6.0% (4/67)	8.3% (4/48) / 7.6% (4/53)	6.3% (3/48) / 5.7% (3/53)
PosUnk モデル [31] による NMT	3.1% (2/64) / 3.0% (2/67)	6.3% (4/64) / 6.0% (4/67)	8.3% (4/48) / 7.6% (4/53)	6.3% (3/48) / 5.7% (3/53)
Sentence Piece NMT (原言語と目的言語の語彙数はそれぞれ 30,000)	7.8% (5/64) / 7.5% (5/67)	9.4% (6/64) / 9.0% (6/67)	12.5% (6/48) / 13.2% (7/53)	12.5% (6/48) / 15.1% (8/53)

評価結果を表 6.9∼表 6.12 に示す．この結果においては，ベースライン SMT が最も高い再現率となり，提案手法はその次に高い再現率となった．この結果から

表 6.10: 名詞句翻訳性能の評価結果 (対象: 未知語を含む評価対象名詞句，提案手法，異なり数に対しての再現率 / 延べ数に対しての再現率)

フレーズ抽出手法	置き換え対象	日中方向	中日方向	日英方向	英日方向
branching entropy を用いたフレーズ抽出 (6.3.1 節)	未知語を含むフレーズ対訳対	**18.8% (12/64) / 17.9% (12/67)**	**20.3% (13/64) / 19.4% (13/67)**	**18.8% (9/48) / 17.0% (9/53)**	**18.8% (9/48) / 17.0% (9/53)**
	全フレーズ対訳対	17.2% (11/64) / 16.4% (11/67)	17.2% (11/64) / 17.9% (12/67)	14.6% (7/48) / 13.2% (7/53)	18.8% (9/48) / 17.0% (9/53)
言語知識に基づき選定した名詞句フレーズの抽出 (6.3.2 節)	未知語を含むフレーズ対訳対	17.2% (11/64) / 16.4% (11/67)	18.8% (12/64) / 19.4% (13/67)	14.6% (7/48) / 15.1% (8/53)	18.8% (9/48) / 17.0% (9/53)
	全フレーズ対訳対	18.8% (12/64) / 17.9% (12/67)	18.8% (12/64) / 19.4% (13/67)	14.6% (7/48) / 15.1% (8/53)	18.8% (9/48) / 17.0% (9/53)
C-value を用いた名詞句部分集合のフレーズ抽出 (6.3.3 節)	未知語を含むフレーズ対訳対	18.8% (12/64) / 17.9% (12/67)	18.8% (12/64) / 17.9% (12/67)	14.6% (7/48) / 13.2% (7/53)	16.7% (8/48) / 15.1% (8/53)
	全フレーズ対訳対	15.6% (10/64) / 14.9% (10/67)	17.2% (11/64) / 16.4% (11/67)	12.5% (6/48) / 11.3% (6/53)	16.7% (8/48) / 15.1% (8/53)

表 6.11: 名詞句翻訳性能の評価結果 (対象: 全評価対象名詞句，ベースライン手法，異なり数に対しての再現率 / 延べ数に対しての再現率)

手法	日中方向	中日方向	日英方向	英日方向
ベースライン SMT [22]	49.2% (208/423) / 47.4% (221/464)	47.4% (199/420) / 48.7% (227/466)	56.5% (198/350) / 56.9% (215/378)	53.6% (186/347) / 54.5% (208/382)
ベースライン NMT(2.3 節)	41.6% (176/423) / 41.2% (191/464)	43.1% (181/420) / 44.0% (205/466)	52.3% (183/350) / 52.4% (198/378)	49.3% (171/347) / 51.1% (195/382)
PosUnk モデル [31] による NMT	41.6% (176/423) / 41.2% (191/464)	43.1% (181/420) / 44.0% (205/466)	52.3% (183/350) / 52.4% (198/378)	49.3% (171/347) / 51.1% (195/382)
Sentence Piece NMT (原言語と目的言語の語彙数はそれぞれ 30,000)	43.3% (183/423) / 43.3% (201/464)	44.1% (185/420) / 44.6% (208/466)	48.3% (169/350) / 48.4% (183/378)	49.3% (160/347) / 48.4% (185/382)

分かるように，ベースライン SMT は，文単位の翻訳性能においては，自動評価・手動評価とも低い性能となっているが，本節の名詞句再現率においては，最も高い性能となった．また，提案手法の中では，「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定における NMT モデルが相対的に最も高い性能を達成した．

さらに，6.5.1 節の人手評価において評価対象とした 200 文に対して，提案手法のうち，「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定での翻訳文，および，「「言語知識に基づき選定した名詞句フレーズの抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定での翻訳文に対して，それぞれ，ベースライン NMT，PosUnk モデルによる NMT，および，Sentence Piece NMT と比較し，翻訳文においてトークン箇所のフレーズが正しく翻訳されたか否かを評価した．評価の結果を表 6.13 ～ 表 6.18 に示す。トークン箇所のフレーズに対して，「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定での翻訳文，および，「「言語知識に基づき選定した名詞句フレーズの抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定での翻訳文，のいずれも，ベースライン NMT および Sentence Piece NMT より多くのトークン箇

表 6.12: 名詞句翻訳性能の評価結果 (対象: 全評価対象名詞句，提案手法，異なり数に対しての再現率 / 延べ数に対しての再現率)

フレーズ抽出手法	置き換え対象	日中方向	中日方向	日英方向	英日方向
branching entropy を用いたフレーズ抽出 (6.3.1 節)	未知語を含むフレーズ対訳対	46.6% (197/423) / 46.1% (214/464)	46.0% (193/420) / 47.0% (219/466)	**55.4% (194/350)** / **55.8% (211/378)**	**51.9% (180/347)** / **53.1% (203/382)**
	全フレーズ対訳対	44.0% (186/423) / 43.1% (200/464)	43.6% (183/420) / 45.1% (210/466)	48.3% (169/350) / 48.7% (184/378)	46.7% (162/347) / 48.7% (186/382)
言語知識に基づき選定した名詞句フレーズの抽出 (6.3.2 節)	未知語を含むフレーズ対訳対	45.4% (192/423) / 44.4% (206/464)	46.0% (193/420) / 47.2% (220/466)	54.3% (190/350) / 54.8% (207/378)	50.4% (175/347) / 52.1% (199/382)
	全フレーズ対訳対	42.6% (180/423) / 41.8% (194/464)	44.1% (185/420) / 44.4% (207/466)	48.6% (170/350) / 48.7% (184/378)	48.7% (186/382) / 46.7% (162/347)
C-value を用いた名詞句部分集合のフレーズ抽出 (6.3.3 節)	未知語を含むフレーズ対訳対	**46.6% (197/423)** / **46.6% (216/464)**	**46.9% (197/420)** / **48.5% (226/466)**	53.4% (187/350) / 53.7% (203/378)	50.4% (175/347) / 52.1% (199/382)
	全フレーズ対訳対	40.4% (171/423) / 40.3% (187/464)	44.5% (187/420) / 45.7% (213/466)	51.4% (180/350) / 51.5% (195/378)	47.2% (164/347) / 49.2% (188/382)

表 6.13: 人手評価 200 文におけるトークン箇所翻訳性能の評価 (対象:「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」，比較手法: ベースライン NMT，日中方向/中日方向/日英方向/英日方向)

トークンに置き換えられたフレーズ	訳語は正解		訳語は不正解	
	比較手法の結果を改善	比較手法の結果と同様	比較手法の結果を改悪	比較手法の結果と同様
名詞句と完全一致	7/12/7/4	1/0/1/1	0/0/0/0	3/5/13/6
名詞句と部分一致	4/1/0/1	0/0/0/0	0/0/0/0	4/3/1/1
名詞句ではない	10/17/3/7	0/1/0/0	0/0/0/0	4/13/3/5
総計	21/30/10/12	1/1/1/1	0/0/0/0	11/21/17/12
合計	33/52/28/25			

表 6.14: 人手評価 200 文におけるトークン箇所翻訳性能の評価 (対象:「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」，比較手法: PosUnk モデルによる NMT，日中方向/中日方向/日英方向/英日方向)

トークンに置き換えられたフレーズ	訳語は正解		訳語は不正解	
	比較手法の結果を改善	比較手法の結果と同様	比較手法の結果を改悪	比較手法の結果と同様
名詞句と完全一致	7/12/7/4	1/0/1/1	0/0/0/0	3/5/13/6
名詞句と部分一致	4/1/0/1	0/0/0/0	0/0/0/0	4/3/1/1
名詞句ではない	6/14/1/4	4/4/2/3	0/1/0/0	4/12/3/5
総計	17/27/8/9	5/4/3/4	0/1/0/0	11/20/17/12
合計	33/52/28/25			

所のフレーズを正しく翻訳できたことがわかった．また，トークン箇所のフレーズを「名詞句と完全一致」のフレーズ,「名詞句と部分一致」のフレーズ (例えば,「走査/線/6406」のような名詞句と数字・記号のフレーズ，もしくは，名詞句の部分列)，および,「名詞句ではない」のフレーズ (例えば,「ステム/160」のような単一形態素と記号・記号のフレーズ，もしくは，数字・記号の列) に分類し，その内訳を表 6.13 ～ 表 6.18 に示す。「「言語知識に基づき選定した名詞句フレーズの抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」と比較してみると,「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の手法では，抽出されたフレーズ中で,「名詞句と完全一致」のフレーズおよび「名詞句と部分一致」のフレーズの割合が低く,「名詞句ではない」フレーズが多く抽出されたが，SMT によって正しく翻訳され，高い翻訳性能

表 6.15: 人手評価 200 文におけるトークン箇所翻訳性能の評価 (対象:「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」，比較手法: Sentence Piece NMT，日中方向/中日方向/日英方向/英日方向)

トークンに置き換えられたフレーズ	訳語は正解		訳語は不正解	
	比較手法の結果を改善	比較手法の結果と同様	比較手法の結果を改悪	比較手法の結果と同様
名詞句と完全一致	6/11/6/4	2/1/2/1	0/0/1/2	3/5/12/4
名詞句と部分一致	3/0/0/1	1/1/0/0	0/0/0/0	4/3/1/1
名詞句ではない	3/10/1/4	7/8/2/3	0/2/0/1	4/11/3/4
総計	12/21/7/9	10/10/4/4	0/2/1/3	11/19/16/9
合計	33/52/28/25			

表 6.16: 人手評価 200 文におけるトークン箇所翻訳性能の評価 (対象:「「言語知識に基づき選定した名詞句フレーズの抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」，比較手法: ベースライン NMT，日中方向/中日方向/日英方向/英日方向)

トークンに置き換えられたフレーズ	訳語は正解		訳語は不正解	
	比較手法の結果を改善	比較手法の結果と同様	比較手法の結果を改悪	比較手法の結果と同様
名詞句と完全一致	8/6/5/3	1/0/1/0	0/0/0/0	25/16/20/3
名詞句と部分一致	1/1/0/2	0/0/0/0	0/0/0/0	0/0/1/2
名詞句ではない	3/5/3/4	1/2/0/1	0/1/0/0	1/5/1/5
総計	12/12/8/9	2/2/1/1	0/1/0/0	26/21/22/10
合計	40/36/31/20			

を達成したことがわかった.

6.5.4 改善例

本節では，提案手法のうちの「「branching entropy を用いたフレーズ抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」の設定における NMT モデルによって，ベースライン手法の翻訳誤りを改善した例について述べる.

まず，図 6.4 および図 6.5 においては，提案手法によって，ベースライン NMT による未知語翻訳誤りを改善する例を示す．図 6.4 の例では，日本語形態素「焼入

表 6.17: 人手評価 200 文におけるトークン箇所翻訳性能の評価 (対象:「「言語知識に基づき選定した名詞句フレーズの抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」，比較手法: PosUnk モデルによる NMT，日中方向/中日方向/日英方向/英日方向)

トークンに置き換えられたフレーズ	訳語は正解		訳語は不正解	
	比較手法の結果を改善	比較手法の結果と同様	比較手法の結果を改悪	比較手法の結果と同様
名詞句と完全一致	8/6/5/3	1/0/1/0	0/0/0/0	25/16/20/3
名詞句と部分一致	1/1/0/2	0/0/0/0	0/0/0/0	0/0/1/2
名詞句ではない	2/4/1/2	2/3/2/3	0/1/0/0	1/5/1/5
総計	11/11/6/7	3/3/3/3	0/1/0/0	26/21/22/10
合計	40/36/31/20			

表 6.18: 人手評価 200 文におけるトークン箇所翻訳性能の評価 (対象:「「言語知識に基づき選定した名詞句フレーズの抽出」および「未知語を含むフレーズ対訳対」が置き換え対象」，比較手法: Sentence Piece NMT，日中方向/中日方向/日英方向/英日方向)

トークンに置き換えられたフレーズ	訳語は正解		訳語は不正解	
	比較手法の結果を改善	比較手法の結果と同様	比較手法の結果を改悪	比較手法の結果と同様
名詞句と完全一致	7/5/5/3	2/1/1/0	1/1/3/1	24/15/17/2
名詞句と部分一致	0/1/0/0	1/0/0/2	0/0/0/0	0/0/1/2
名詞句ではない	2/4/1/1	2/3/2/4	0/2/0/1	1/4/1/4
総計	9/10/6/4	5/4/3/6	1/3/3/2	25/19/19/8
合計	40/36/31/20			

れ」は未知語であるため，ベースライン NMT においては，日本語フレーズ「焼入れ剤」は未知語トークン $\langle unk \rangle$ に誤訳された．一方，提案手法では，日本語フレーズ「焼入れ剤」は，branching entropy の下限値を満たすためフレーズとして抽出され，フレーズトークンに置き換えられた．提案手法による翻訳文においては，フレーズトークンが SMT 訳文中の中国語訳語 “淬火剂” に置き換えられ，参照用訳文中の訳語と一致した．同様に，図 6.5 の例では，未知語を含む日本語フレーズ「脛骨トライアルシム」がトークンに置き換えられた後，提案手法による翻訳文においては，フレーズトークンが SMT 訳文中の英語訳語 “tibial trial shim” に置き換えられ，参照用訳文中の訳語と一致した．

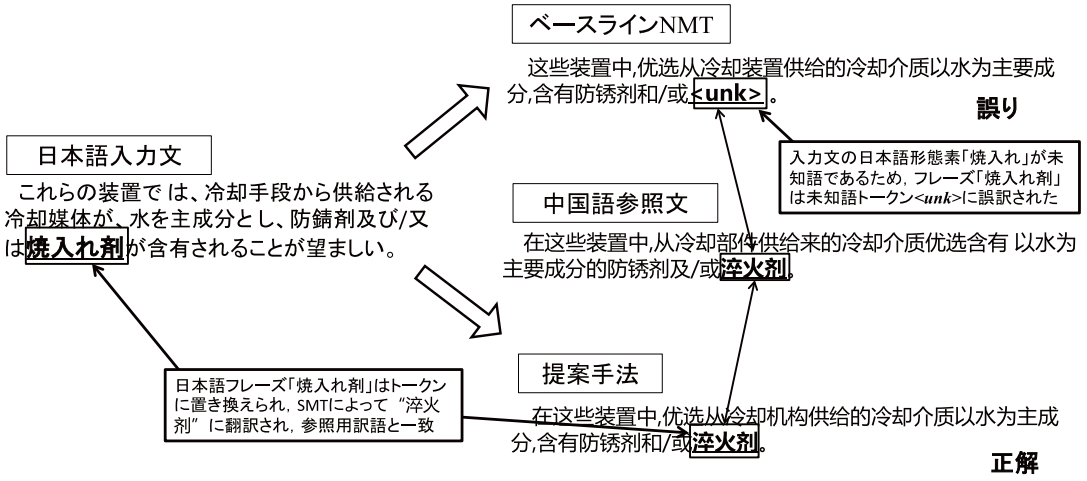


図 6.4: 提案手法による改善例 (対ベースライン NMT，日中翻訳，未知語翻訳誤り)

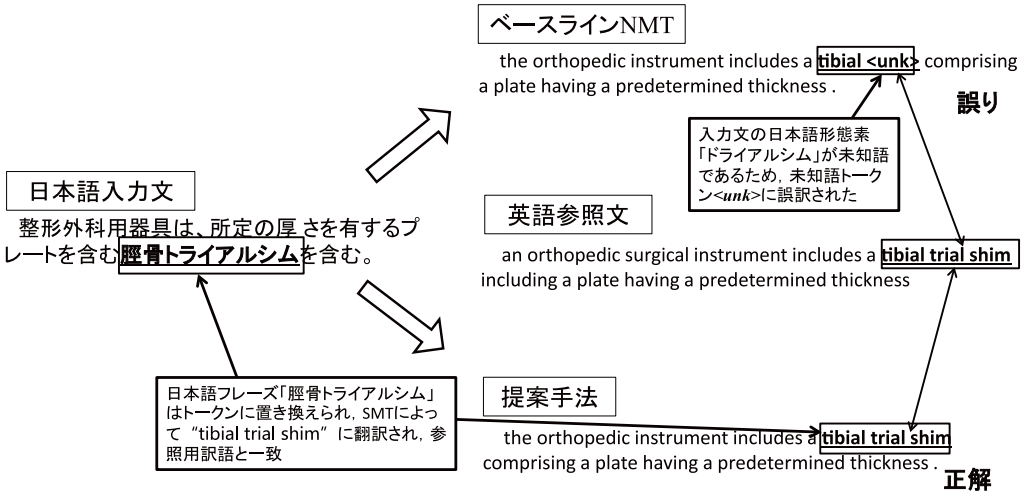


図 6.5: 提案手法による改善例 (対ベースライン NMT，日英翻訳，未知語翻訳誤り)

次に，図 6.6 および図 6.7 においては，提案手法によって，ベースライン NMT による訳抜け誤りを改善する例を示す．図 6.6 の例では，中国語単語 “涂覆” は未知語であるため，ベースライン NMT によって翻訳されず，中国語単語 “涂覆” に対応する訳語は翻訳文中には存在していない．一方，提案手法では，中国語単語 “涂覆” を含む中国語フレーズ “涂覆温度” は，branching entropy の下限値を満たすためフレーズとして抽出され，フレーズトークンに置き換えられた．提案手法による翻訳文においては，フレーズトークンが SMT 訳文中の日本語訳語「被覆温度」に置き換えられ，参照用訳文中の訳語と一致した．同様に，図 6.7 の例では，英語単語 “eukaryotic”，および，“promoters” の訳語である日本語形態素「プロモーター」が未知語となり，ベースライン NMT による翻訳文においては，“eukaryotic” を含むフレーズ “eukaryotic promoters” が翻訳されず訳抜けとなった．一方，提案手法においては，“eukaryotic promoters” がフレーズとして抽出され，SMT との

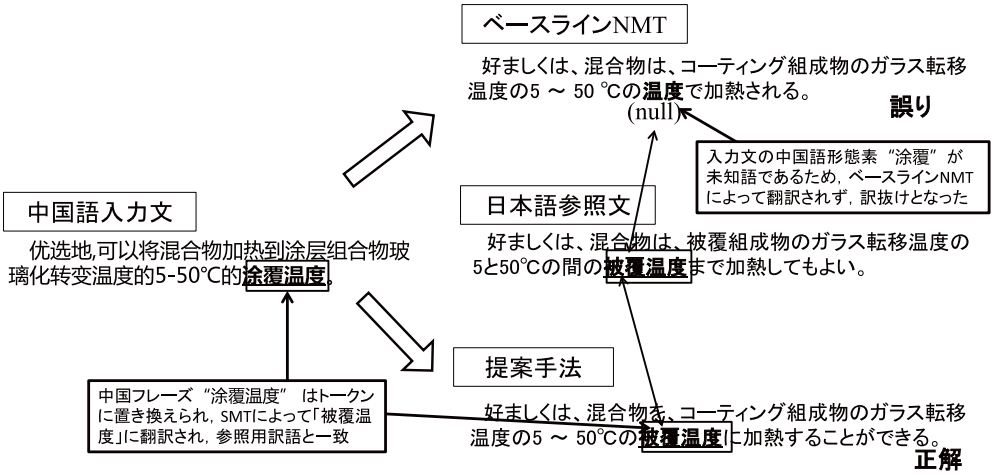


図 6.6: 提案手法による改善例 (対ベースライン NMT，中日翻訳，訳抜け誤り)

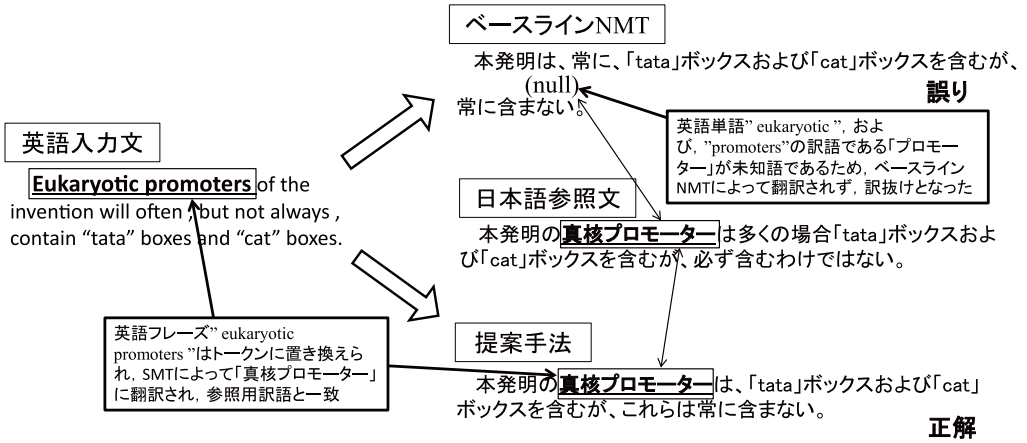


図 6.7: 提案手法による改善例 (対ベースライン NMT，英日翻訳，訳抜け誤り)

併用により正しく翻訳できた.

さらに，図 6.8∼図 6.10 においては，提案手法によって，三種類のベースライン手法であるベースライン NMT，PosUnk モデルによる NMT，および，Sentence Piece NMT の翻訳誤りを改善する例を比較する．図 6.8 では，日本語形態素「吸蔵」および日本語フレーズ「水素吸蔵」の中国語訳語である “储氢” は，NMT モデルの語彙集合に含まれない未知語であるため，ベースライン NMT による翻訳文においては，日本語フレーズ「水素吸蔵」は未知語トークン $\langle unk \rangle$ に誤訳された．一方，図 6.9 では，PosUnk モデルによる NMT による翻訳文においては，日本語未知語「吸蔵」に対応する未知語トークン $\langle unk \rangle$ が単語単位で中国語訳に置き換えられた結果，中国語 “氢嵌入” (「水素を埋め込む」を意味する) に誤訳された．また，図 6.10 では，Sentence Piece NMT による翻訳文においては，日本語形態素「吸蔵」が文字「吸」と「蔵」に分割され，日本語側で未知語とはならなかったが，NMT モデルによる翻訳過程において，高頻度な「吸」が優先され，低

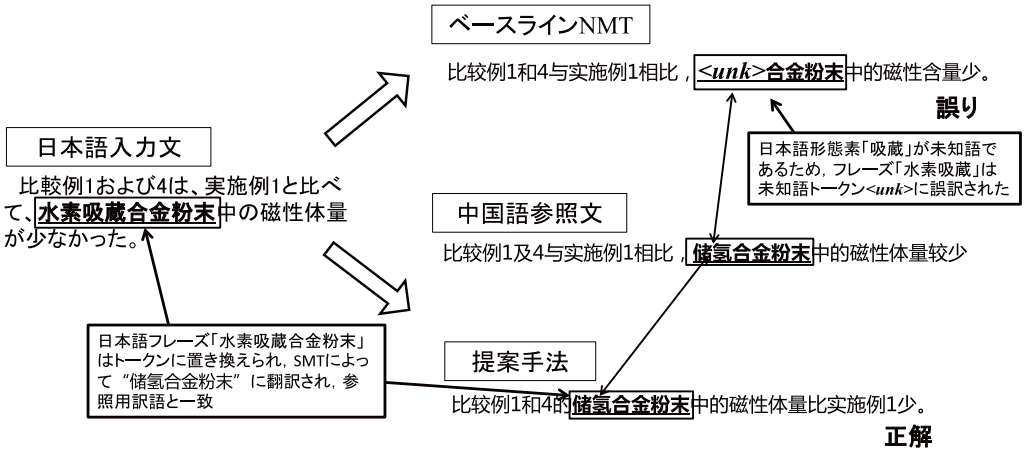


図 6.8: 提案手法による改善例 (対ベースライン NMT，日中翻訳，日本語入力文「比較例 1 および 4 は、実施例 1 と比べて、水素吸蔵合金粉末中の磁性体量が少なかった。」)

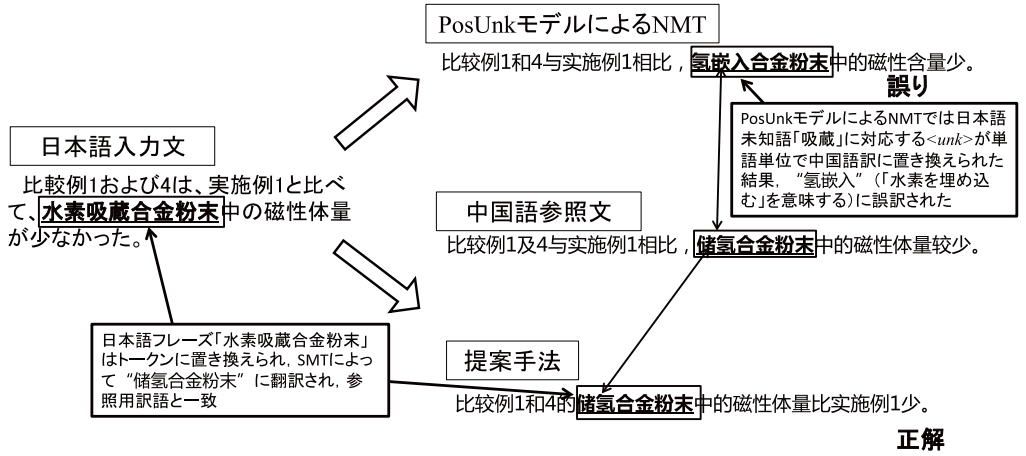


図 6.9: 提案手法による改善例 (対 PosUnk モデルによる NMT，日中翻訳，日本語入力文「比較例 1 および 4 は、実施例 1 と比べて、水素吸蔵合金粉末中の磁性体量が少なかった。」)

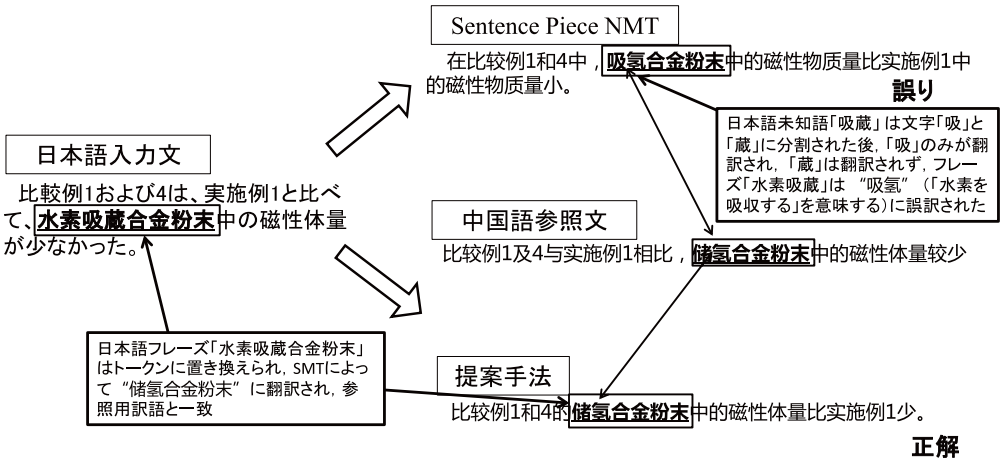


図 6.10: 提案手法による改善例 (対 Sentence Piece NMT，日中翻訳，日本語入力文「比較例 1 および 4 は、実施例 1 と比べて、水素吸蔵合金粉末中の磁性体量が少なかった。」)

頻度な「蔵」が翻訳されず,「水素吸蔵」が中国語 “吸氢” （「水素を吸収する」を意味する）に誤訳された．これらのベースライン手法に対して，提案手法による翻訳文においては，日本語形態素「吸蔵」を含む日本語フレーズ「水素吸蔵合金粉末」がフレーズとして抽出され，SMT 訳文中の中国訳語 “储氢合金粉末” に置き換えられ，この訳においては「吸蔵」は参照用訳文中の訳語と一致した．

6.6 関連研究

本論文に関連して，NMT が扱える語彙の規模を拡大する方式についての研究がいくつか行われてきた．[31] においては，二言語間の単語対応関係を記録した未知語トークンを導入し，出力文中のトークンを訳語に置き換える方式によって，NMT モデルにおける語彙の規模を拡大する手法を提案している．[24] においては，未知語を同義となる既知語に置き換えた後，NMT モデルを訓練する方式を提案している．しかし，[31, 24] においては，置き換え対象となるのが未知語が単語単位であるため，知語の翻訳の問題を，単語単位の構成的な翻訳の問題に帰着できる場合の対策にとどまっており，複合語的なフレーズの中でも，特に，その構成単語の訳語を構成的に組み合わせる方式に帰着できない非構成的な複合語フレーズ翻訳が取り扱えない点が弱点となっている．一方，本論文では，未知語を含むフレーズを置き換え対象としてトークンに置き換えた後，NMT モデルを訓練するアプローチをとっている．したがって，[31, 24] と比較すると，本論文の方式は，構成単語の訳語を構成的に組み合わせる方式に帰着できない非構成的な複合語フレーズ翻訳においてその威力を発揮する．この点については，前節の評価実験において，提案手法が，PosUnk モデル [31] による NMT を上回る性能を達成したことに

よって実証された．

また，[42, 28, 6, 57]，および，Sentence Pieceにおいては，単語をサーブワード単位又は文字単位に分割することによってNMTモデルが扱える語彙の規模を拡大する手法を提案している．このうち，[28, 6] は未知語のみを対象として分割を行う手法である．[28] では，未知語のみを文字単位に分割して，単語単位と文字単位の分割単位を併行して利用するNMTモデルを提案している．一方，[6] においては，未知語に対しては，文字単位での意味表現をNMTモデルの組み込む手法を提案している．また，[42, 57]，および，Sentence Pieceツールにおいては，文全体を対象として，文中の単語を構成文字列単位に分割することによって語彙数を削減する手法を提案している．特に，[42] においては，語彙数に基づく貪欲法によって語彙の単位を細分化する手法を提案している．また，[57] およびSentence Pieceツールにおいては，エントロピー最尤推定の目的関数を用いて語彙の単位を細分化する手法を提案している．これらの手法を用いたNMTモデルにおいては，フレーズはより小さい語彙の単位に分割された後，NMTモデルの訓練およびNMTモデルによる翻訳が行われる．一方，本論文の手法と，上述の手法との間の最も大きな相違点として，本論文の手法においては，未知語を含むフレーズが，フレーズ単位のままでSMTによって翻訳される点が挙げられる．本論文の手法の優位性については，前節の評価実験において，提案手法が，Sentence Piece NMTを上回る性能を達成したことによって実証された．

その他，本論文で対象としたフレーズの翻訳に関連して，[62] においては，NMTのn-best出力文をSMTによってランク付けする手法を提案している．具体的には，[42] の手法によって未知語をサブワード単位に分割した後，NMTモデルを訓練およびNMTモデルによる翻訳を行い，n-best翻訳文を生成する．そして，SMTのフレーズ翻訳テーブルを用いて，原言語文中のフレーズの翻訳精度の推定値スコアを算出し，このスコアを用いてn-best翻訳文の再ランク付けを行うことによって，翻訳性能を改善している．しかし，[62] の手法の欠点として，NMTのn-best出力文から最適な出力文を選択しているため，NMTによるn-best翻訳文のどの文においても適切な訳語が含まれない場合には，翻訳性能改善は困難である．実際に，前節の評価実験のうちの名詞句翻訳性能の評価においてSMTが最も高い性能を達成していることから，[62] の手法においても，上記の欠点に伴う一定の限界があると考えられる．

6.7 本章のまとめ

本章では，NMTモデルが大規模フレーズ語彙に対応できないという問題に対して一つの解決手法を提案した．提案手法では，訓練用対訳文においてフレーズ間の二言語対応の情報を収集し，二言語間で対応済みのフレーズ対訳対を同一のトークンに置き換えた後，NMTモデルの訓練を行う．NMTモデルによる翻訳時には，

NMT モデルの語彙集合中の語彙部分に対しては，NMT モデルによる訳文生成がなされ，一方，その他のフレーズまたは単語語彙部分に対しては，SMT モデルによる翻訳がなされる．日中，中日，日英，英日の各方向の翻訳において評価を行い，提案手法の有効性を検証した．結果として，ベースライン SMT モデル，および，ベースライン NMT モデルとの比較において，0.7 ポイント以上の BLEU の向上を達成できた．また，NMT の弱点の一つである訳抜けの改善においては，ベースライン NMT モデルによる訳抜けを約 30%削減することができた．今後の課題として，SMT による大語彙フレーズ翻訳と Sentence Piece に基づく NMT モデルを併用することにより翻訳性能を改善することが挙げられる．また，NMT モデルにおける訳抜けを改善するための既存の手法 (例えば，[11] 等) の枠組みを導入することにより，提案手法における訳抜けの改善をさらに促進することが挙げられる．

第7章　結論

本論文では，機械翻訳，特に，RBMT のパラダイムにおいて利用する大規模フレーズ翻訳知識を獲得する手法として，**対訳専門用語の同定**，および，**同義対訳専門用語の同定**の課題について論じた．さらに，NMT のパラダイムにおいて大規模フレーズ翻訳知識を利用することを目的として，**統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラル機械翻訳**の課題について論じた．

人手によって対訳辞書を作成するためには膨大な時間と労力を要するため，年々新しく作られる専門用語を迅速に専門用語対訳辞書に追加していくためには，自動もしくは半自動的に専門用語対訳辞書を構築する手法が必要である．この課題に対して，本論文では，SMT の重要な機能の一つとして，大規模フレーズ翻訳テーブルの自動生成機能に着目し，対訳文およびフレーズ翻訳テーブルを用いた訳語推定手法を利用して対訳専門用語の候補を収集した後，分類器学習によって適切な対訳専門用語を同定する方式によって，大規模フレーズ翻訳辞書作成支援方式を実現した．本論文の**対訳専門用語の同定**手法においては，特に，SVM の素性として，複数の対訳文から得られる素性を利用することによって，対訳専門用語同定において高い信頼度を達成することができた．実際に用いられた素性は，単言語情報を用いる素性と二言語情報を用いる素性に大別される．従来手法に対して新しく導入した素性としては，同一の日本語専門用語に対する中国語訳語候補の数 (f_7)，対訳文およびフレーズ翻訳テーブルを用いて訳語推定を行う際の文単位の句対応制約の違反のない対訳文の割合 (f_8)，要素合成法による訳語推定における翻訳確率 (f_9) 等の素性が挙げられる．評価実験においては，本研究において新しく導入したこれらの素性によって，従来手法の性能を改善することができた．さらに，本論文では,「中国語側が形態素単位のフレーズ翻訳テーブル」および「中国語側が文字単位のフレーズ翻訳テーブル」をそれぞれ独立に用いて対訳専門用語の獲得を行った．適合率を最大化する調整を行うことにより，中国語側が形態素単位の場合，および，文字単位の場合の両方において 90%以上の適合率を達成した．また，SVM による評価結果における誤り分析の結果においては，中国語側が形態素単位の場合と文字単位の場合との間では，誤りの傾向が異なっていた．このことから，今後，中国語側の区切り単位として，形態素および文字の二種類の単位を併用した上で SVM を適用することによって，日中対訳専門用語同定の性能の改善が期待できることが分かった．

次に，本論文では，対訳特許文から獲得した対訳専門用語の同義・異義関係を

同定する手法について研究を行った．この課題では，各対訳文を情報源として専門用語対訳対を同定する際に，それぞれの対訳文から同定された専門用語対訳対の間の関係性を考慮していない点に着目し，この問題点の解消に取り組んだ．この研究では，人手で選定した中心的対訳対に対する同義の可能性のある対訳対を網羅する同義対訳専門用語候補集合を生成するために，対訳特許文およびフレーズ翻訳テーブルを用いて，原言語・目的言語方向，および，目的言語・原言語方向，の両方向において専門用語の訳語推定を繰り返す方式を提案した．そして，生成した候補集合に対して分類器学習を適用することにより，対訳専門用語間の同義・異義関係を同定した．本論文の**同義対訳専門用語の同定**手法の研究においては，分類器学習における素性の組合わせの網羅的な評価および各素性の有効性に関する詳細な評価を行い，適合率最大の場合との間で有意差 (有意水準 5%) のない適合率となる素性の組合せを求め，性能を左右する重要な素性とした．その結果，性能を左右する重要な素性として，フレーズ翻訳テーブルから得られる訳語のうちの共通訳の割合 (f_{15})，および，専門用語の文字列のうちの非共有箇所に対してフレーズ翻訳テーブルから得られる訳語のうちの共通訳の割合 (f_{16}) の二つが得られた．この f_{15} および f_{16} は，フレーズ翻訳テーブルにおいてどの程度の割合で共通の訳語を持つかという情報と，単言語における同義関係の度合いとの間の相関に着目した素性である．評価実験においては，再現率が 25%以上という条件のもとで，約 90%の適合率を達成した．今後の課題として，再現率を改善するため，人手の介在を併用する半自動的な同義対訳専門用語の同定の枠組を開発することが重要であると考えられる．

また， 標準的な NMT モデルにおいては，扱える語彙数に限界があるため，モデル中では低頻度語を含むことができない点が問題となっている．このため，いかに流暢な訳文を生成できたとしても，特許文書の正確な翻訳に不可欠である新語・固有表現が未知語となって訳出されないという問題を抱えている．そこで，本論文では，NMT のパラダイムにおける大語彙の課題を解決するために，大規模フレーズ翻訳知識を利用する方式に取り組んだ．本論文の方式では，訓練用対訳文に対して SMT モデルを適用してフレーズ間の二言語対応の情報を収集し，二言語間で対応するフレーズ対訳対を同一のトークンに置き換えた後，NMT モデルの訓練を行う．NMT モデルによる翻訳時には，NMT モデル中に含まれる語彙部分に対しては NMT モデルによる訳文生成がなされ，一方，NMT モデル中に含まれない語彙，および，それらの語彙によって構成されるフレーズ部分に対しては，SMT モデルによる翻訳がなされる．本論文の**統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラル機械翻訳**方式においては，NMT による翻訳が容易ではないフレーズを SMT によって翻訳する方式によって，フレーズの翻訳において標準的な NMT モデルを上回る性能を達成できた．SMT によって翻訳する対象となるフレーズを抽出する手法としては，branching entropy の下限値の条件を満たし，かつ，未知語を含むフレーズを抽出する方式によって最も高い翻訳性能を達成することができた．

今後の課題の一つとして，第 6 章で述べた SMT による大語彙フレーズ翻訳との併用による NMT において，言語を横断する意味表現ベクトル [61, 29, 60] を利用することによって，SMT による大語彙フレーズ翻訳の性能を改善することが挙げられる．近年，深層学習技術の発展とともに，深層ニューラルネットワークを用いて，フレーズに対して言語を横断する意味表現ベクトルを求めて，各種タスクにおいて利用する手法が多数提案された．[61] においては，二言語再帰自己符号器 (bilingual recursive auto-encoder) の手法によって，フレーズの対訳対から言語を横断する埋め込みベクトルを学習する手法を提案している．[29] においては，対訳文の単語対応情報を利用して，二言語の埋め込みベクトルを同一ベクトル空間に写像する手法を提案している．[60] においては，原言語フレーズと目的言語フレーズの再帰自己符号器を統合することによって，意味表現モデルの学習を行う手法を提案している．これらの一連の手法においては，ベクトル空間において意味表現ベクトルが近い位置に存在するフレーズ組は，それらが単言語のフレーズ組の場合には同義関係にあり，一方，それらが言語を横断するフレーズ組の場合には対訳関係にある，という結果が得られる．実際に，これらの言語を横断する意味表現ベクトルを SMT に導入するタスクにおいては，SMT の翻訳性能を改善できたことが報告された [61, 29, 60]．一方，本研究では，6.5.3 節の評価結果において示したように，SMT によるフレーズ翻訳によって NMT の翻訳性能を改善できるが，それでもなお，対象となるフレーズの半数以上を誤訳している．そのため，この半数以上の誤訳フレーズの翻訳性能の改善が必要であり，言語を横断する意味表現ベクトルを導入することによってこの誤訳が改善すると期待される．

また，NMT は，流暢な目的言語文を生成できる点が長所であるが，その一方で，低頻度のフレーズや複合語の翻訳に弱いことと，訳抜けが発生することが弱点になっている．このうち，前者の弱点に対しては，本研究では，低頻度のフレーズをトークンに置き換えた NMT モデルを訓練するとともに，低頻度のフレーズを SMT によって翻訳するハイブリッド方式を採用した．しかし，本研究のこの手法においても，低頻度のフレーズをトークンに置き換えた NMT モデルにおいては，置き換え前のフレーズの語彙情報が反映されていない点が弱点となっている．この弱点を回避するためには，低頻度のフレーズをトークンに置き換えた NMT モデルを訓練するのではなく，トークンへの置き換えをする前の NMT モデルによる翻訳結果において，低頻度のフレーズの翻訳結果を，直接，SMT による翻訳結果に置き換える方式が有効であると期待される．また，NMT モデルを訓練する際の単語分割の単位としては，日本語，中国語のように，言語知識を組み込んだ形態素解析ツールとして高い性能のツールが利用可能な言語の場合を除くと，6.5 節の評価において比較対象とした Sentence Piece のように，統計情報のみに基づいて文字レベルでの語彙分割を行う方式が有望である．

一方，後者の訳抜けの弱点に対しては，第 6 章で述べた SMT による大語彙フレーズ翻訳との併用による NMT は，訳抜けの問題を直接改善することを意図した方式とはなっていないが，6.5.2 節の評価結果に示したように，ベースライン NMT

における訳抜けの問題を改善している．これは，低頻度のフレーズをトークンに置き換えたNMTモデルにおいて，一文の単語長が相対的に短く抑えられることにより，NMTモデル自身において訳抜けを抑制する効果があるためである．ここで，これまでのNMTにおける先行研究においては，訳抜けの問題を直接改善することを意図した方式もいくつか提案されている．例えば，[33, 51]においては，入力文中の各単語位置におけるカバレッジベクトルを逐次更新し，このカバレッジベクトルを参照して入力文のどの部分が翻訳済でどの部分が未翻訳であるかを示すことによって，未翻訳の部分が訳抜けのままにならないように翻訳過程を制御するカバレッジモデルを提案した．[11]においては，NMTモデルにおける注意機構の重み，および，逆翻訳時の確率をふまえて，出力文における訳抜けの度合いを推定するスコアを算出して，NMTモデルのn-bestの出力文において訳抜けが少ないと推定される出力文を選択する手法を提案した．また，[11]においては，NMTモデルにおける注意機構の重み，および，逆翻訳時の確率をふまえて，出力文における訳抜けの度合いを推定するスコアを算出して，NMTモデルのn-bestの出力文において訳抜けが少ないと推定される出力文を選択する手法を提案した．我々は，それらの研究のうち，特に[11]に着目し，第6章で述べたSMTによる大語彙フレーズ翻訳との併用によるNMTにおいて，[11]の方式による出力文中の訳抜け度合いの推定結果，および，実際の訳抜け数を測定し，いずれにおいてもベースラインNMTを改善することを示した[21]．このことから，第6章で述べたSMTによる大語彙フレーズ翻訳との併用によるNMTにおいて，[11]の訳抜け改善方式を組み込むことによって，訳抜けの問題をさらに改善することが期待される．

謝辞

本論文は筆者が筑波大学大学院システム情報工学研究科において行った研究成果をまとめさせて頂いたものです．本稿執筆の過程において，指導教員としてご指導を頂き，本論文を作成する上でも，多大なるご助言を賜わりましたシステム情報系宇津呂武仁教授に心より感謝いたします．本論文の審査過程において，ご助言とご指導を賜りました，システム情報系丸山勉教授，矢野博明教授，古賀弘樹教授，山本幹雄教授に感謝いたします．また，本論文において利用した日中パテントファミリーのデータを提供して頂いた日本特許情報機構 (JAPIO) の関係各位に感謝いたします

また，日常の議論を通じて多くの知識や示唆を頂いた自然言語処理研究室所属のみなさま，ありがとうございました．自然言語処理に関して，情報学の観点以外からのコメントや手助けをしてくれた友人らにも感謝します．

最後に，いついかなるときも私を支えてくれた家族に心から感謝します．

参考文献

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. 3rd ICLR*, 2015.

[2] D. Bouamor, N. Semmar, and P. Zweigenbaum. Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proc. 51st ACL*, pp. 759–764, 2013.

[3] C. Callison-Burch, P. Koehn, and M. Osborne. Improved statistical machine translation using paraphrases. In *Proc. HLT-NAACL*, pp. 17–24, 2006.

[4] Y. Chen, Y. Huang, S. Kong, and L. Lee. Automatic key term extraction from spoken course lectures using branching entropy and prosodic/semantic features. In *Proc. 2010 IEEE SLT Workshop*, pp. 265–270, 2010.

[5] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. EMNLP*, pp. 1724–1734, 2014.

[6] M. R. Costa-Jussà and J. A. R. Fonollosa. Character-based neural machine translation. In *Proc. 54th ACL*, pp. 357–361, 2016.

[7] M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. Improving the extraction of bilingual terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 5, No. 4, pp. 31:1–31:17, 2009.

[8] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the *C-value/NC-value* method. *International Journal on Digital Libraries*, Vol. 3, No. 2, pp. 115–130, 2000.

[9] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proc. 7th NTCIR Workshop Meeting*, pp. 389–400, 2008.

[10] P. Fung and L. Y. Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. 17th COLING and 36th ACL*, pp. 414–420, 1998.

[11] I. Goto and H. Tanaka. Detecting untranslated content for neural machine translation. In *Proc. 1st NMT*, pp. 47–55, 2017.

[12] R. Haque, S. Penkale, and A. Way. Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. In *Proc. 4th Computerm*, pp. 42–51, 2014.

[13] W. He, H. Wu, H. Wang, and T. Liu. Improve SMT quality with automatically extracted paraphrase rules. In *Proc. 50th ACL*, pp. 979–987, 2012.

[14] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn. Scalable modified Kneser-Ney language model estimation. In *Proc. ACL 2013*, pp. 690–696, 2013.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[16] F. Huang, Y. Zhang, and S. Vogel. Mining key phrase translations from Web corpora. In *Proc. HLT/EMNLP*, pp. 483–490, 2005.

[17] M. Itagaki, T. Aikawa, and X. He. Automatic validation of terminology translation consistency with statistical method. In *Proc. MT Summit XI*, pp. 269–274, 2007.

[18] S. Jean, K. Cho, Y. Bengio, and R. Memisevic. On using very large target vocabulary for neural machine translation. In *Proc. 28th NIPS*, pp. 1–10, 2014.

[19] Z. Jin and K. Tanaka-Ishii. Unsupervised segmentation of Chinese text by use of branching entropy. In *Proc. COLING/ACL 2006*, pp. 428–435, 2006.

[20] N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In *Proc. EMNLP*, pp. 1700–1709, 2013.

[21] Ryuichiro Kimura, Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Effect on reducing untranslated content by neural machine translation with a large vocabulary of technical terms. In *Proceedings of the 7th Workshop on Patent and Scientific Literature Translation*, pp. 9–20, September 2017.

[22] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

[23] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. HLT-NAACL*, pp. 127–133, 2003.

[24] X. Li, J. Zhang, and C. Zong. Towards zero unknown word in neural machine translation. In *Proc. 25th IJCAI*, pp. 2852–2858, 2016.

[25] B. Liang, T. Utsuro, and M. Yamamoto. Semi-automatic identification of bilingual synonymous technical terms from phrase tables and parallel patent sentences. In *Proc. 25th PACLIC*, pp. 196–205, 2011.

[26] D. Lin, S. Zhao, B. Van Durme, and M. Paşca. Mining parenthetical translations from the Web by word alignment. In *Proc. 46th ACL: HLT*, pp. 994–1002, 2008.

[27] B. Lu and B. K. Tsou. Towards bilingual term extraction in comparable patents. In *Proc. 23rd PACLIC*, pp. 755–762, 2009.

[28] M. Luong and C. D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proc. 54th ACL*, pp. 1054–1063, 2016.

[29] M. Luong, H. Pham, and C. D. Manning. Bilingual word representation with monolingual quality in mind. In *Proc. NAACL-HLT 2015*, pp. 151–159, 2015.

[30] M. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proc. EMNLP*, pp. 1412–1421, 2015.

[31] M. Luong, I. Sutskever, O. Vinyals, Q. V. Le, and W. Zaremba. Addressing the rare word problem in neural machine translation. In *Proc. 53rd ACL*, pp. 11–19, 2015.

[32] Y. Matsumoto and T. Utsuro. Lexical knowledge acquisition. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 24, pp. 563–610. Marcel Dekker Inc., 2000.

[33] H. Mi, B. Sankaran, Z. Wang, and A. Ittycheriah. Coverage embedding models for neural machine translation. In *Proc. EMNLP 2016*, pp. 955–960, 2016.

[34] E. Morin and A. Hazem. Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. In *Proc. 52nd ACL*, pp. 1284–1293, 2014.

[35] 森下洋平, 梁冰, 宇津呂武仁, 山本幹雄. フレーズテーブルおよび既存対訳辞書を用いた専門用語の訳語推定. 電子情報通信学会論文誌, Vol. J93–D, No. 11, pp. 2525–2537, 2010.

[36] T. Nakazawa, H. Mino, I. Goto, G. Neubig, S. Kurohashi, and E. Sumita. Overview of the 2nd workshop on Asian translation. In *Proc. 2nd WAT*, pp. 1–28, 2015.

[37] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[38] 奥村学, 渡辺太郎, 今村賢治, 賀沢秀人, N. Graham, 中澤敏明. 知識に基づく機械翻訳. 機械翻訳，1.2 節, 自然言語処理シリーズ, pp. 5–16. コロナ社.

[39] 奥村学, 渡辺太郎, 今村賢治, 賀沢秀人, N. Graham, 中澤敏明. 機械翻訳. 自然言語処理シリーズ. コロナ社, 2014.

[40] 奥村学, 渡辺太郎, 今村賢治, 賀沢秀人, N. Graham, 中澤敏明. 歴史. 機械翻訳，1.1 節, 自然言語処理シリーズ, pp. 2–4. コロナ社, 2014.

[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th ACL*, pp. 311–318, 2002.

[42] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proc. 54th ACL*, pp. 1715–1725, 2016.

[43] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng. Parsing with compositional vector grammars. In *Proc. ACL 2013*, pp. 455–465, 2013.

[44] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural machine translation. In *Proc. 27th NIPS*, pp. 3104–3112, 2014.

[45] 外池昌嗣, 宇津呂武仁, 佐藤理史. ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定. 自然言語処理, Vol. 14, No. 2, pp. 33–68, 2007.

[46] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proc. 4th SIGHAN Workshop on Chinese Language Processing*, pp. 168–171, 2005.

[47] 坪井裕太, 海野裕也, 鈴木潤. 機械翻訳. 深層学習による自然言語処理，5.1 節, 機械学習プロフェショナルシリーズ, pp. 122–132. 講談社, 2017.

[48] 坪井裕太, 海野裕也, 鈴木潤. 系列変換モデル. 深層学習による自然言語処理，3.4 節, 機械学習プロフェショナルシリーズ, pp. 72–90. 講談社, 2017.

[49] 坪井裕太, 海野裕也, 鈴木潤. 深層学習による自然言語処理. 機械学習プロフェショナルシリーズ. 講談社, 2017.

[50] T. Tsunakawa and J. Tsujii. Bilingual synonym identification with spelling variations. In *Proc. 3rd IJCNLP*, pp. 457–464, 2008.

[51] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li. Modeling coverage for neural machine translation. In *Proc. ACL 2016*, pp. 76–85, 2016.

[52] M. Utiyama and H. Isahara. A Japanese-English patent parallel corpus. In *Proc. MT Summit XI*, pp. 475–482, 2007.

[53] 宇津呂武仁. 機械翻訳の研究の現状—経験的アプローチを中心として—. 日本語学, Vol. 28-12, pp. 44–61, October 2009.

[54] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[55] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton. Grammar as a foreign language. In *Proc. 28th NIPS*, 2015.

[56] 渡辺太郎. ニューラルネットワークによる構造学習の発展. 人工知能, Vol. 31, No. 2, pp. 202–208, 2016.

[57] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. `http://arxiv.org/abs/1609.08144`, 2016. [Online; accessed 28-December-2017].

[58] W. Yang, J. Yan, and Y. Lepage. Extraction of bilingual technical terms for Chinese-Japanese patent translation. In *Proc. WAT 2016*, pp. 194–202, 2016.

[59] K. Yasuda and E. Sumita. Building a bilingual dictionary from a Japanese-Chinese patent corpus. In *Computational Linguistics and Intelligent Text Processing*, Vol. 7817 of *LNCS*, pp. 276–284. Springer, 2013.

[60] B. Zhang, D. Xiong, and J. Su. Battrae: Bidimensional attention-based recursive autoencoders for learning bilingual phrase embeddings. In *Proc. AAAI 2017*, pp. 3372–3378, 2017.

[61] J. Zhang, S. Liu, M. Zhou, and C. Zong. Bilingually-constrained phrase embeddings for machine translation. In *Proc. ACL 2014*, pp. 111–121, 2014.

[62] J. Zhang, M. Utiyama, E. Sumita, G. Neubig, and S. Nakamura. Improving neural machine translation through phrase-based forced decoding. In *Proc. IJCNLP 2017*, pp. 152–162, 2017.

研究業績リスト

査読付き論文雑誌

1. 龍梓, 董麗娟, 宇津呂武仁, 三橋朋晴, 山本幹雄. 日中対訳文を用いた同義対訳専門用語の同定手法. 情報処理学会論文誌, Vol. 56, No. 3, pp. 960–971, March 2015.

査読付き国際会議論文

1. Ryuichiro Kimura, Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Effect on reducing untranslated content by neural machine translation with a large vocabulary of technical terms. In *Proceedings of the 7th Workshop on Patent and Scientific Literature Translation*, pp. 9–20, September 2017.

2. Zi Long, Ryuichiro Kimura, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Neural machine translation model with a large vocabulary selected by branching entropy. In *Proceedings of the 16th Machine Translation Summit*, pp. 227–240, September 2017.

3. Teguh Budianto, Hyunwoo Oh, Yi Ding, Zi Long, and Takehito Utsuro. Identifying rush strategies employed in StarCraft II using support vector machines. In *Entertainment Computing — ICEC 2017, 16th IFIP TC 14 International Conference, Tsukuba City, Japan, September 18-21, 2017, Proceedings*, Vol. 10507 of *Lecture Notes in Computer Science*, pp. 357–361. Springer, September 2017.

4. Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Translation of patent sentences with a large vocabulary of technical terms using neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation*, pp. 47–57, December 2016.

5. Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Collecting bilingual technical terms from patent families of character-segmented Chinese sentences and morpheme-segmented Japanese sentences. In *Proceed-*

ings of the 6th Workshop on Patent and Scientific Literature Translation, pp. 68–80, October 2015.

6. Zi Long, Lijuan Dong, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Evaluating features for identifying Japanese-Chinese bilingual synonymous technical terms from patent families. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora — Building Resources for Machine Translation Research*, pp. 52–61, July 2015.

7. Zi Long, Lijuan Dong, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Identifying Japanese-Chinese bilingual synonymous technical terms from patent families. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora — Building Resources for Machine Translation Research*, pp. 49–54, May 2014.

8. Itsuki Toyota, Zi Long, Lijuan Dong, Takehito Utsuro, and Mikio Yamamoto. Compositional translation of technical terms by integrating patent families as a parallel corpus and a comparable corpus. In *Proceedings of 5th Workshop on Patent Translation*, pp. 16–23, September 2013.

著書

1. Lijuan Dong, Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Collecting bilingual technical terms from Japanese-Chinese patent families by SVM. In Kôiti Hasida and Ayu Purwarianti, editors, *Computational Linguistics, 14th International Conference of the Pacific Association for Computational Linguistics, PACLING 2015, Bali, Indonesia, May 19-21, 2015, Revised Selected Papers*, Vol. 593 of *Communications in Computer and Information Science*, pp. 251–262. Springer, February 2016.

全国大会，研究会報告，その他の研究発表

1. Zi Long, Ryuichiro Kimura, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Patent NMT integrated with large vocabulary phrase translation by SMT at WAT 2017. In *Proceedings of the 4th Workshop on Asian Translation*, pp. 110–118, November 2017.

2. 宇津呂武仁, 龍梓, 木村龍一郎, 山本幹雄. SMT による大語彙フレーズ翻訳との併用によるニューラルネットワーク機械翻訳. In 日本特許情報機構（編）, *Japio YEAR BOOK 2017*, pp. 290–297. November 2017.

3. Hyunwoo Oh, Teguh Budianto, Yi Ding, Zi Long, and Takehito Utsuro. Identifying the rush strategies in the game logs of the real-time strategy game StarCraft-II. 第 31 回人工知能学会全国大会論文集, May 2017.

4. Teguh Budianto, Hyunwoo Oh, Yi Ding, Zi Long, and Takehito Utsuro. An analysis on the rush strategies of the real-time strategy game StarCraft-II. 第 31 回人工知能学会全国大会論文集, May 2017.

5. 龍梓, 宇津呂武仁, 山本幹雄. Neural machine translation of patent sentences with large vocabulary technical terms. 日本特許情報機構（編）, 平成 28 年度 AAMT/Japio 特許翻訳研究会報告書「機械翻訳及び機械翻訳評価に関する研究及びシンポジウム報告」, pp. 11–20. March 2017.

6. Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Neural machine translation of patent sentences with a large vocabulary of technical terms. 言語処理学会第 23 回年次大会論文集, pp. 867–870, March 2017.

7. 王怡青, 龍梓, 土井俊弥, 韓炳材, 宇津呂武仁. ニューラルネットに基づく質問文生成モデルのクロスドメイン評価. 言語処理学会第 23 回年次大会論文集, pp. 998–1001, March 2017.

8. 宇津呂武仁, 龍梓, 山本幹雄. 日中パテントファミリーを利用した同義対訳専門用語の同定. In 日本特許情報機構（編）, *Japio YEAR BOOK 2016*, pp. 230–235. November 2016.

9. 龍梓, 宇津呂武仁, 山本幹雄. パテントファミリーを用いた日中対訳専門用語の同定. 日本特許情報機構（編）, 平成 27 年度 AAMT/Japio 特許翻訳研究会報告書「機械翻訳及び機械翻訳評価に関する研究及び海外調査」, pp. 20–26. March 2016.

10. 江原暉将, 長瀬友樹, 宇津呂武仁, 龍梓, 王向莉. 中国語特許文献の中日翻訳評価のためのテストセットの拡充. 日本特許情報機構（編）, 平成 27 年度 AAMT/Japio 特許翻訳研究会報告書「機械翻訳及び機械翻訳評価に関する研究及び海外調査」, pp. 40–42. March 2016.

11. 宇津呂武仁, 龍梓, 山本幹雄. 日中パテントファミリーを利用した専門用語訳語推定 — 分類器学習に基づく方式 —. In 日本特許情報機構（編）, *Japio YEAR BOOK 2015*, pp. 308–313. November 2015.

12. Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Identifying bilingual technical terms from Japanese-Chinese parallel patent sentences using SVM. In *the 15th China-Japan Natural Language Processing Joint Research Promotion Conference*, October 2015.

13. 龍梓, 董麗娟, 宇津呂武仁, 三橋朋晴, 山本幹雄. 日中パテントファミリーを用いた同義対訳専門用語同定手法の評価. 言語処理学会第 21 回年次大会論文集, pp. 728–731, March 2015.

14. 龍梓, 董麗娟, 宇津呂武仁, 山本幹雄. パテントファミリーにおける対訳文対非抽出部分を利用した専門用語訳語推定方式と統計的機械翻訳モデルとの間の比較・評価. 日本特許情報機構（編）, 平成 26 年度 AAMT/Japio 特許翻訳研究会報告書「機械翻訳及び機械翻訳評価に関する研究及びシンポジウム報告」, pp. 11–26. March 2015.

15. 董麗娟, 龍梓, 宇津呂武仁, 三橋朋晴, 山本幹雄. SVM を用いた日中パテントファミリーからの対訳専門用語収集. 言語処理学会第 21 回年次大会論文集, pp. 724–727, March 2015.

16. 宇津呂武仁, 董麗娟, 龍梓, 山本幹雄. 日中パテントファミリーを利用した専門用語訳語推定 — フレーズテーブルおよび対訳文対を利用する方式 —. In 日本特許情報機構（編）, *Japio YEAR BOOK 2014*, pp. 236–241. November 2014.

17. 宇津呂武仁, 董麗娟, 龍梓, 山本幹雄. パテントファミリーからの専門用語対訳辞書の構築. 第 3 回特許情報シンポジウム 論文資料集, pp. 45–53. November 2014.

18. Zi Long, Lijuan Dong, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Identifying bilingual synonymous technical terms from Japanese-Chinese patent families. In *the 14th China-Japan Natural Language Processing Joint Research Promotion Conference*, October 2014.

19. 龍梓, 董麗娟, 豊田樹生, 宇津呂武仁, 三橋朋晴, 山本幹雄. 日中パテントファミリーから抽出した対訳文を用いた同義対訳専門用語の同定. 言語処理学会第 20 回年次大会発表論文集, pp. 955–958, March 2014.

20. 董麗娟, 龍梓, 豊田樹生, 宇津呂武仁, 三橋朋晴, 山本幹雄. 日中パテントファミリーから抽出した対訳文を用いた専門用語の訳語推定. 言語処理学会第 20 回年次大会発表論文集, pp. 368–371, March 2014.

21. 豊田樹生, 龍梓, 董麗娟, 宇津呂武仁, 山本幹雄. パテントファミリーにおける対訳文対非抽出部分を利用した専門用語訳語推定方式の評価と分析. 日本特許情報機構（編）, 平成 25 年度 AAMT/Japio 特許翻訳研究会報告書「機械翻訳

及び機械翻訳評価に関する研究及びワークショップ報告」, pp. 15–31. March 2014.

22. Takehito Utsuro, Lijuan Dong, Zi Long, Itsuki Toyota, Tomoharu Mitsuhashi, and Mikio Yamamoto. Confident estimation of Japanese-Chinese technical term translation based on a phrase translation table trained with patent families. In *the 13th China-Japan Natural Language Processing Joint Research Promotion Conference*, October 2013.

23. 豊田樹生, 高橋佑介, 牧田健作, 龍梓, 董麗娟, 宇津呂武仁, 山本幹雄. 日英パテントファミリーにおける対訳文対非抽出部分を利用した専門用語訳語推定. 日本特許情報機構（編）, 平成 24 年度 AAMT/Japio 特許翻訳研究会報告書「機械翻訳及び辞書構築に関する研究及びシンポジウム・拡大評価部会報告」, pp. 2–9. March 2013.