修 士 論 文

単語境界が明示されていない言語を対象とした対訳辞書の自動構築

北陸先端科学技術大学院大学
情報科学研究科

王 馨竹

平成 29 年 3 月

修 士 論 文

単語境界が明示されていない言語を対象とした対訳辞書の自動構築

指導教員　白井清昭

審査委員主査　白井清昭
審査委員　飯田弘之
審査委員　池田心

北陸先端科学技術大学院大学
情報科学研究科

1510064 王 馨竹

提出年月: 平成 29 年 2 月

概 要

本研究は，大規模なパラレルコーパスから対訳辞書を自動的に獲得する新しい手法を提案する．特に，既存の形態素解析ツールの単語分割の誤りに影響されない手法の確立を目指す．対象とする言語は中国語と英語とし，中英対訳辞書を自動構築する．中国語の文については，単語分割ツールで文を単語に分割してから獲得される訳語対と, 文字の 1-gram と 2-gram に分割してから獲得される訳語対を組み合わせる．単語分割ツールを用いる従来手法では，単語分割の誤りが訳語対獲得の精度を低下させるという問題がある．また，中国語文を文字に分割する従来手法では，文字を単語に復元する際に正しい単語に復元されないという問題がある．提案手法ではこれらを組み合わせることで，お互いの誤りが補完され，訳語対獲得の精度が向上することが期待できる．また，中英対訳辞書は既にいくつか存在するが，辞書にない訳語も存在する．提案手法を評価する際, 自動獲得した訳語対のうち，既存の対訳辞書に含まれない新しい訳語対をどれだけ獲得できたかという観点でも評価する．

本研究の対訳辞書獲得手法における処理の流れを説明する．まず，中国語と英語のパラレルコーパスを用意する．ここでのパラレルコーパスとは，中国語の文と英語の文が, 文単位で対応付けられたコーパスである．次に，中国語と英語の文に対して前処理を行う. 英語文については lemmatization により各単語を原形に直す．中国語については 3 種類の前処理, すなわち単語分割，文字 1-gram への分割, 文字 2-gram への分割を行い，3 通りのパラレルコーパスを用意する．次に，3 種類の前処理が実施されたパラレルコーパスに対し，GIZA++ を用いて単語のアライメントを自動的に推定する．その結果から中英の訳語対の候補を抽出する. 結果として，3 種類の前処理が実施されたコーパスから 3 つの訳語対の候補の集合が得られる．最後に, これら 3 つの訳語対の候補の集合から，適切な訳語対を選択する．その結果，最終的な中英対訳辞書が得られる．

提案手法の有効性を評価する実験を行う．実験では，新聞記事のパラレルコーパスとして BBC news コーパスを，法律のパラレルコーパスとして Parallel Corpus of China ’ s Law Documents (PCCLD コーパス) を用いる．単語分割してから訳語対を獲得する手法，文字 1-gram に分割してから訳語対を獲得する手法，文字 2-gram に分割してから訳語対を獲得する手法の 3 つをベースラインとし，提案手法と比較する．それぞれの手法で計算された訳語対のスコアの大きい上位 100 件について，それが正しい訳語対であるかを判定し，その正解率を求める．提案手法の正解率は，新聞記事コーパスに対しては 0.94 であり,，法律のコーパスに対しては 0.95 であった．これらは 3 つのベースラインの正解率を上回った．また，出現頻度 5 以上の訳語対について，既存の中英対訳辞書に含まれない未知の訳語対の割合を新語獲得率と定義し，評価する．既存の中英対訳辞書として，約 110,000 の訳語対を持つ LDC English Chinese bilingual word lists を用いる．提案手法の新語獲得率は，新聞記事コーパスに対して 0.935，法律コーパスに対して 0.934 となった．これは 3 つのベースラインを 0.02 から 0.08 ポイント上回った．新聞記事と法律のコーパ

スを比較すると，提案手法とベースラインの差は法律コーパスの方が大きかった．これは，法律コーパスは新聞記事コーパスに比べて未知語が多く存在するが，提案手法は未知語をより多く獲得できたためと考えられる．上記の結果から，提案手法の有効性が確認された．

今後の課題を以下に述べる．現在，スコアが上位 100 件程度の訳語対しか評価していないため，スコアが下位の訳語対についても提案手法の有効性を検証する必要がある．また，精度だけではなく再現率の評価も必要である. さらに，中国語だけではなく，日本語や韓国語のような単語境界が明示されていない他の言語を対象に実験を行い，提案手法が言語に依らず有効であるかを調べたい．

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景

対訳辞書とは，単語とその別の言語における訳語の組からなる自然言語処理用知識である．対訳辞書は，機械翻訳，言語横断検索など，多言語情報処理に必要不可欠な知識である [1][2].

しかし，あらゆる単語を含む対訳辞書を構築することは難しい．日常生活においても新しい単語は日々生まれており，このような新語を全て含む対訳辞書を人手で整備することはほとんど不可能である．特定の分野のテキストで使われる専門用語も，対訳辞書に掲載されていないことが多い．また，テキストのジャンルによって，同じ単語でも訳語が異なることが知られている．一般的ではない特殊なジャンルで使われている訳語も対訳辞書に掲載されていないことがある．

また，対訳辞書は 1 つだけ用意すればよいというわけではない．現在，世界では 1000 を越える数の言語が使われている．しかし，その全ての言語の組に対して対訳辞書が作られているわけではない．特に，コーパスや辞書などの言語資源の整備が進んでいない言語については，対訳辞書の整備も遅れている．また，特定の分野やジャンルのテキストを翻訳するためには，その分野に特化した対訳辞書を作る必要がある．言い換えれば，ジャンル毎に専用の対訳辞書を用意することが望ましい．

以上の理由から，必要な全ての対訳辞書を人手で作成するのは困難であるし，あらゆる訳語を網羅的に収録した対訳辞書を人手で作成することもまた困難である．そのため，対訳辞書を自動構築する技術が必要とされている．実際，対訳辞書を自動獲得する多くの先行研究がある．それらの研究では，パラレルコーパスから対訳辞書を自動的に獲得することを試みたものが多い．

しかしながら，対訳辞書の自動獲得については以下のような問題点がある．中国語，日本語，韓国語，タイ語など，単語境界が明示されていない言語を対象とするときは，まず形態素解析 (単語分割) が前処理として行われる必要がある．しかし，単語分割の段階で誤りが生じたとき，正しい訳語対を獲得できない可能性がある．現在，形態素解析ツールの精度は一般に高いが，言語資源の整備が進んでいない言語については，単語分割の精度が悪かったり，あるいは公開されている単語分割ツールが存在しないこともある．また，専門分野の対訳辞書を構築する場合，専門分野のテキストに対する単語分割の誤りは無視できないほど多いと考えられる．

1.2　研究の目的

本研究は，大規模なパラレルコーパスから対訳辞書を自動的に獲得する新しい手法を提案する．特に，既存の形態素解析ツールの単語分割の誤りに影響されない手法の確立を目指す．対象とする言語は中国語と英語とし，中英対訳辞書を自動構築する．

中国語の文については，単語分割ツールで文を単語に分割してから獲得される訳語対と，1-gram と 2-gram に分割してから獲得される訳語対を組み合わせることにより，単語分割ツールの誤りの影響を軽減し，正確に訳語対を獲得することを狙う．また，中英対訳辞書は既にいくつか存在するが，先ほど述べたように，辞書にない訳語も存在する．提案手法を評価する際，自動獲得した訳語対のうち，既存の対訳辞書に含まれない新しい訳語対をどれだけ獲得できたかという観点でも評価する．

1.3　本論文の構成

本論文の構成は以下の通りである．2 章では，単語分割ならびに対訳辞書の自動獲得に関する先行研究を紹介し，先行研究と本研究の違いを述べる．3 章では，提案する対訳辞書の自動構築手法について詳しく述べる．4 章では，提案手法を用いて実際に対訳辞書を自動構築し，その品質を評価し，考察を行う．最後に，5 章で本論文のまとめと今後の課題について述べる．

第2章　関連研究

本章では関連研究について述べる．まず，2.1 節では，中国語文の単語分割のアルゴリズムを説明し，その一般的な問題点を論じる．次に，2.2 節では，対訳辞書をパラレルコーパスから自動獲得する先行研究を紹介する．最後に，2.3 節で本研究の特色について述べる．

2.1　中国語の単語分割

単語分割とは，単語境界が明示されていない文に対し，単語の境界を同定する処理である．与えられた文を文字列とみなし，それを単語を構成する部分文字列に分割する処理であるともいえる．中国語における単語分割の問題について考察する．英語の文では，単語の間にはスペースがあり，単語境界はスペースによって明示されている．一方，中国語の文では，文字あるいは段落や文章の区切りは明白であるが，単語の境界は明示されていない．したがって，中国語を対象とした自然言語処理システムの多くは，前処理として，文を単語に分割する処理が必要である．英語でも，これと類似した問題として，文 (単語列) を基本句の列に分割する問題がある．基本句への分割も難しい問題であるが，中国語の単語分割もまた難しい問題である．中国語の単語分割は，中国語を対象とした自然言語処理において，最も重要な技術と言える．中国語の形態素解析器あるいは単語分割ツールは，コーパスのような言語資源と同様に，自然言語処理のための重要なリソースである．また，単語分割ツールを作成するためには，中国語の文法における独自の特殊性を十分に考慮する必要がある．

2.1.1　中国語の単語分割アルゴリズム

中国語の単語分割アルゴリズムは，大まかに以下の 2 つに分けられる．

- 文字列の一致に基づく手法

  単語の辞書をあらかじめ用意する．もし文字列の一部分が辞書における単語の見出しと一致するとき，その部分列を単語とみなす．このような単語分割アルゴリズムは，ヒューリスティックルールを使用する [3]．例えば、「フォワード/リバース最大マッチング」や「長さ優先」などのヒューリスティクスがある．このアルゴリズムの利点は，高速であり，実装が容易であることである．単語分割のための計算量は，

文の長さを n としたとき，$O(n)$ となる．欠点は単語分割の曖昧性を正しく解消できないことがあるという点である．曖昧性の例を以下に挙げる．

独立自主和平等互利的原则
⟶ 独立自主/和/平等/互利/的/原则
⟶ 独立自主/和/平等互利/的/原则

この例では，6 番目から 9 番目の文字について，これを 6,7 番目の文字からなる単語と 8,9 番目の文字からなる単語に分割する可能性と，これらを 1 つの単語に分割する可能性の 2 つがあることを示している．文字列の一致に基づく手法はどちらかを選択するが，それが正しいとは限らない．もう一つの欠点は未知語に対処できないという点である．未知語とは単語辞書に登録されていない単語である．このアルゴリズムは，辞書における単語の見出しと一致するかで単語を同定しているため，辞書に登録されてない単語を認識できない．

- 統計的手法もしくは機械学習に基づく単語分割手法

  この手法では，中国語の文の複数の単語分割の候補に対してスコアを与えるモデルを設計する．そして，モデルパラメータが付与されたデータ（タグ付きコーパス）からモデルを自動的に学習する．与えられた文の単語を分割するとき，学習したモデルによって，様々な単語分割のスコアもしくは確率を計算し，最大確率を持つ単語分割を最終的な結果とする [4]．代表的なモデルとして，HIdden Markov Model (HMM) や Conditional Random Field (CRF) がある．機械学習に基づく単語分割アルゴリズムは，曖昧性解消の問題や未知語の問題に対応しやすいという点で，文字列の一致に基づく手法よりも優れている．しかし，大量のデータに対して人手で注釈を付与する必要があること，単語分割の速度が遅いこと，などの欠点もある．

2.1.2 jieba

中国語形態素解析器のひとつに jieba[1]がある．jieba は一般公開されている．以下，jieba の単語分割アルゴリズムの概要を説明する．

1. Trie ツリーの辞書に基づき，与えられた文内に出現する全ての単語を含む有向非循環グラフ (Directed Acyclic Graph; DAG) を生成する．

   jieba は dict.txt と呼ばれる辞書を持つ．その中には 20,000 以上の単語が含まれる．辞書中の単語は Trie ツリーとして保存されている．Trie ツリーとは，共通の接頭辞を持つ単語を同じパスで表わすような木構造のことである．Trie ツリーは接頭辞ツリーとも呼ばれる．Trie ツリーを用いることで，中国語の文に含まれる全ての辞書

[1] https://github.com/fxsjy/jieba

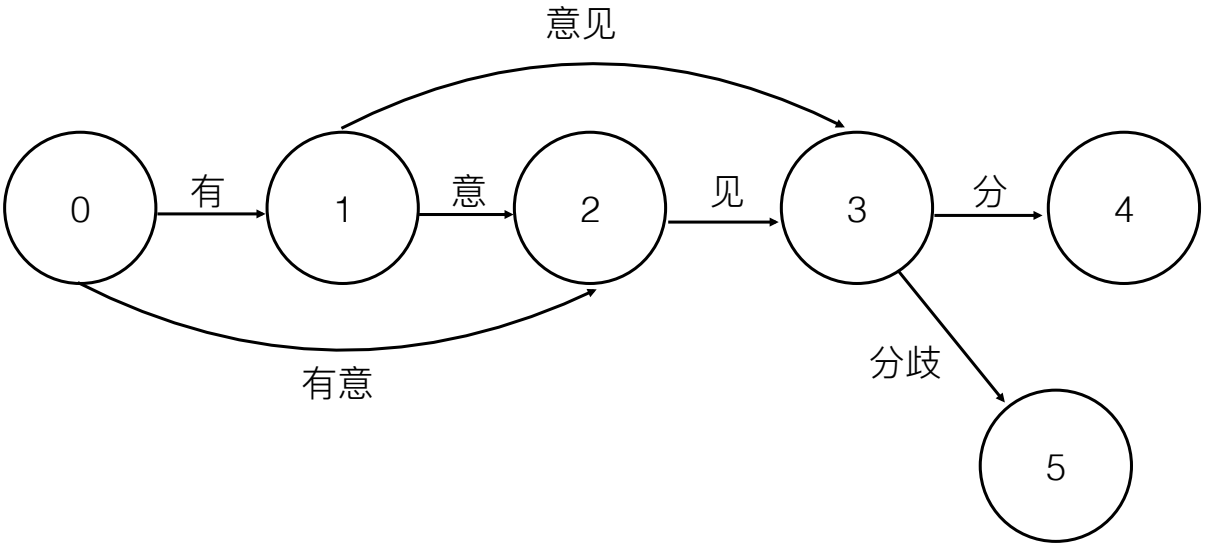

図 2.1: DAG の例

中の単語を高速に発見できる．解析対象とする中国語の文に対し，Trie ツリーの辞書を使って単語を検索し，DAG を生成する．DAG は可能な全ての単語分割を含んでいる．DAG を図 2.3 に示す．グラフ内のエッジは辞書に登録されている単語に対応し，エッジの始点と終点は単語の文中における開始位置と終了位置に対応する．この DAG は以下の中国語の文を入力とし，2 つの単語分割の候補を含んでいる．

有意见分歧
⟶ 有/意见/分歧
⟶ 有意/见/分歧

上記の最初の単語分割は，DAG における 0 → 1 → 3 → 5 というパスに対応し，2 番目の単語分割は 0 → 2 → 3 → 5 というパスに対応する．

2. 最も可能性が高いパスを見つけるために，ダイナミックプログラミングを使用する．具体的には，単語の出現頻度の和が最大となるような単語分割を見つける．

   ダイナミックプログラミングは，DAG における各ノードに対し，文末からそのノードに到達するパスの中で最大の確率を記録する．ノードの確率の計算は右 (文末) から左 (文頭) の順に行われる．新しいノードの最大確率を計算するときは，その右隣にあるノードの最大確率 (これは既に計算されている) の計算結果を利用する．これにより最大確率を持つ単語分割を高速に求めることができる．

3. 未知語を HMM モデルによって検出する．

未知語を検出する問題を系列ラベリング問題とみなす．中国語文におけるそれぞれの文字に対し，B,E,M,S のいずれかのラベルを与える．B は未知語の開始位置，E は未知語の終了位置，M は未知語の中間位置，S は 1 文字で構成される未知語を表わす．可能なひとつの系列ラベルに対し，その確率を HMM で推定する．また，Viterbi アルゴリズムを用いて，最大の確率を持つ系列ラベルを高速に求める．選択された系列ラベルを元に未知語を検出する．

HMM による未知語検出は訓練データに大きく依存する．一方，未知語が単語として世間一般に認められるためには，その単語が長い期間使われている必要がある．したがって，HMM を学習するためのコーパスは，長い年月に渡るテキストの集合を用意する必要がある．しかし，そのような通時性を持つコーパスを用意することは一般には難しいという問題点もある．

2.2 対訳辞書の自動獲得

既に述べたように，単語境界が明示されていない言語を処理する際には，単語分割が前処理として行われることが多い．その際，単語辞書を用いた簡単な文字列のマッチングで文を単語に分割したり，単語分割ツールが使われる．しかし，どの手法も完全に正しい単語分割が得られるわけではない．Xu らは，単語分割の誤りが機械翻訳の性能に悪影響を与える問題を指摘し，これに対する新しい対訳辞書の自動構築手法を提案した [5]．この研究は，機械翻訳の品質を最大化することと，翻訳システムの構築に要する人手作業を最小化することを目的とする．そのため，単語辞書や単語分割ツールを使わずに中国語テキストを単語に分割するための新しい方法を提案している．まず，単語分割されていないバイリンガルコーパスから統計的機械翻訳 (Statistical Machine Translation; SMT)[6] のモデルを訓練する．具体的には，GIZA++[7] を用いてモデルを訓練する．また，ここでは中国語の文を文字単位に分割する．次に，その結果を使用し，対訳関係にある文において，中国語文における複数の文字が英語文における同じ単語にマッピングされたとき，それらの文字を連結して中国語の単語を復元し，中国語単語と英単語の組を得る．この結果を利用することで，中国語の辞書 (単語のリスト) を自動的に作成できる．自動作成された中国語の辞書を基に，中国語の単語分割ツールを作成し，それを用いて中国語文の単語分割を行う．最後に，単語分割されたパラレルコーパスを用いて翻訳システムを再学習する．単語分割の処理を必要としないので，既存の単語分割ツールの誤りの影響を受けない．しかし，文字から単語を復元する際には誤りが生じる可能性がある．

Yasuda らは日中パラレルコーパスから対訳辞書を自動的に獲得する新しい手法を提案した [8]．図 2.2 にこの研究の概要を示す．まずは，日本語漢字と簡体字中国語との間の類似性に基づいて，日中パラレルコーパスから単語翻訳対 (Bilingual word pairs 1) を抽出する．次に，2 つの異なる統計的機械翻訳の訓練ツールを使用してフレーズテーブルを学習し，それらに共通して出現する単語訳語対 (Bilingual word pairs 2) を抽出する．最後に，前のステップによって得られた単語訳語対を用いて 2 種類の SMT システムを訓練し，そ

れを用いて Bilingual word pairs 3 と Bilingual word pairs 4 を得る．Bilingual word pairs 1 から 4 にかけて，獲得された単語訳語対の精度が 59.3%から 92.1%まで段階的に向上したと報告している．

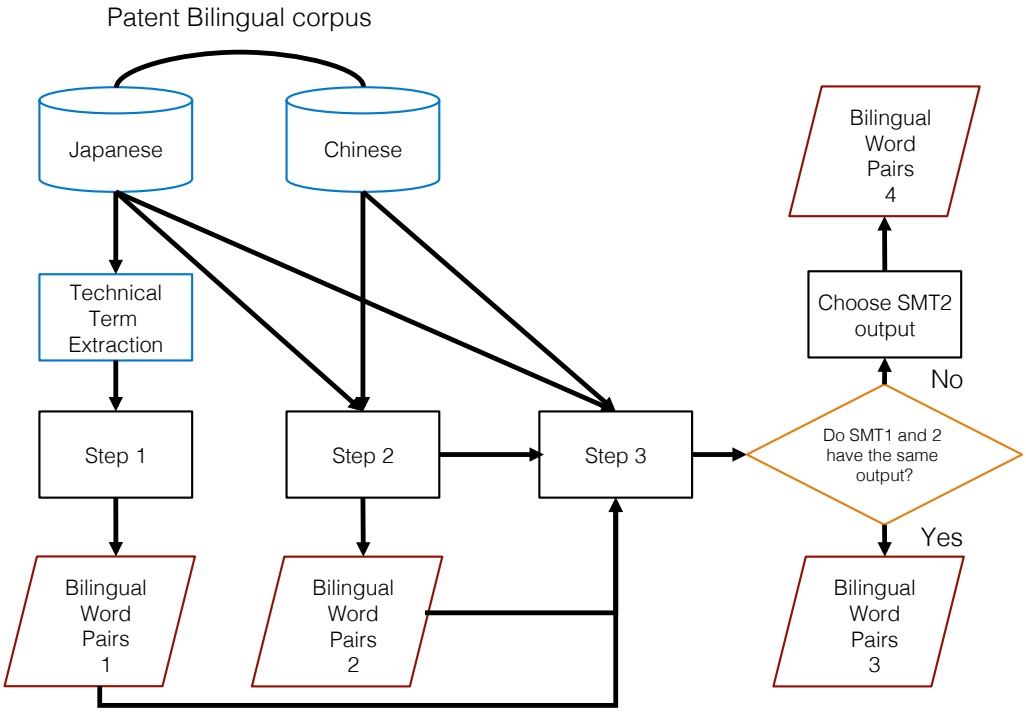

図 2.2: Yasuda らの研究の概要

北村らは，専門用語や定型表現の訳語は機械翻訳における翻訳の品質を決める重要な要因であり，それらの対訳を自動獲得することが求められていることから，パラレルコーパスから専門用語や定型表現の対訳表現を自動的に抽出する方法を提案した [9]．まずは，日英パラレルコーパスが与えられたとき，日本語文および英語文の形態素解析を行う．自立語を抽出し，その出現回数を求める．2 回以上出現する任意長の単語列を抽出する．次に，出現回数条件を設定し，閾値以上の出現回数を持った単語列を対象に，日本語単語列と英語単語列の単語列間の類似度を計算する．類似度が閾値より大きい単語列を対訳表現候補の集合に加え，そこから正しいと思われる対訳表現を選別し，データベースに登録する．最後に，これらの対訳表現を，対訳コーパスを参照して機能語を補うことで，対訳コーパスに出現した形に復元する．

張らは，日中対訳辞書を自動構築することを目的とし，品詞情報や漢字情報などに基づいて，日本語単語に対する中国語の訳語候補の妥当性を評価するための数多くのヒューリスティックを提案した [10]．

2.3 本研究の特色

本研究では，対訳辞書を自動構築する際，単語境界が明示されていない言語の文に対して，単語分割ツールで文を単語に分割してから獲得される訳語対と，Xu らの手法のように文字に分割してから獲得される訳語対を統合する点に特徴がある．単語分割ツールの誤りの影響を受けにくく，かつより正確に訳語対を獲得することを狙う．さらに，中国語の文を文字の 2-gram に分割してから訳語対を獲得することも試みる．

以上をまとめると，提案手法では，単語への分割，文字 1-gram への分割，文字 2-gram への分割という 3 種類の前処理を経てから訳語対を獲得し，その結果を統合する．これにより以下の効果が期待できる．

- 単語分割ツールを用いる手法と用いない手法を併用することで，単語分割の誤りと，文字を単語に復元する際の誤りが互いに補完され，訳語対獲得の精度が向上する．
- 複数の前処理を適用することにより，より多くの訳語対が獲得できる．
- より多くの訳語対を獲得することにより，その中に，既存の対訳辞書に含まれない新語や専門用語も多く含まれることが期待される．

第3章　提案手法

本章ではパラレルコーパスから訳語対を自動獲得する手法について述べる．まず，3.1 節で提案手法の概要を説明する．3.2 節では，パラレルコーパス中の文に対する前処理について述べる．3.3 節では，前処理されたパラレルコーパスにおける単語のアライメントを決める方法について述べる．3.4 節では，単語アライメントの結果から訳語対の候補を獲得する手法について述べる．最後に，3.5 節では，訳語対の候補から正しい訳語対を選別する手法について述べる．

3.1　概要

図 3.1 に提案手法の概要を示す．まず，中国語と英語のパラレルコーパスを用意する．ここでのパラレルコーパスとは，中国語の文と英語の文が，文単位で対応付けられたコーパスである．次に，中国語と英語の文に対して前処理を行う．英語文については lemmatization により各単語を原形に直す．中国語については 3 種類の前処理，すなわち単語分割，文字 1-gram への分割，文字 2-gram への分割を行い，3 通りのパラレルコーパスを用意する．次に，3 種類の前処理が実施されたパラレルコーパスに対し，GIZA++ を用いて単語のアライメントを自動的に推定する．その結果から中英の訳語対の候補を抽出する．結果として，3 種類の前処理が実施されたコーパスから 3 つの訳語対の候補の集合が得られる．最後に，これら 3 つの訳語対の候補の集合から，適切な訳語対を選択する．その結果，最終的な中英対訳辞書が得られる．

提案手法では，既存の形態素解析ツールの単語分割の誤りに影響されない手法，すなわち文を文字 1-gram ならびに文字 2-gram へ分割してから訳語対を獲得する手法と，ツールを用いて文を単語に分割してから訳語対を獲得する手法を併用する．これにより，互いの誤りが補完され，訳語対獲得の精度が向上することが期待できる．また，形態素解析ツールの誤りによって検出できなかった中国語の単語に対しても，文字 1-gram や 2-gram への分割の際には検出され，その英語の訳語を獲得できる可能性があるため，新語や専門用語の訳語対も獲得できる可能性が高まる．

なお，本研究では，文を文字 3-gram や 4-gram に分割してから訳語対を獲得する手法を検討した．その結果，以下のことがわかった．文字 3-gram への分割については，中国語の単語は 3 文字で構成されることが少ないことから，正しくない訳語対が数多く獲得された．文字 4-gram への分割については，獲得される訳語対が少なかった．したがって，本研究では文字 3-gram や 4-gram へ分割する手法は採用しないこととした．

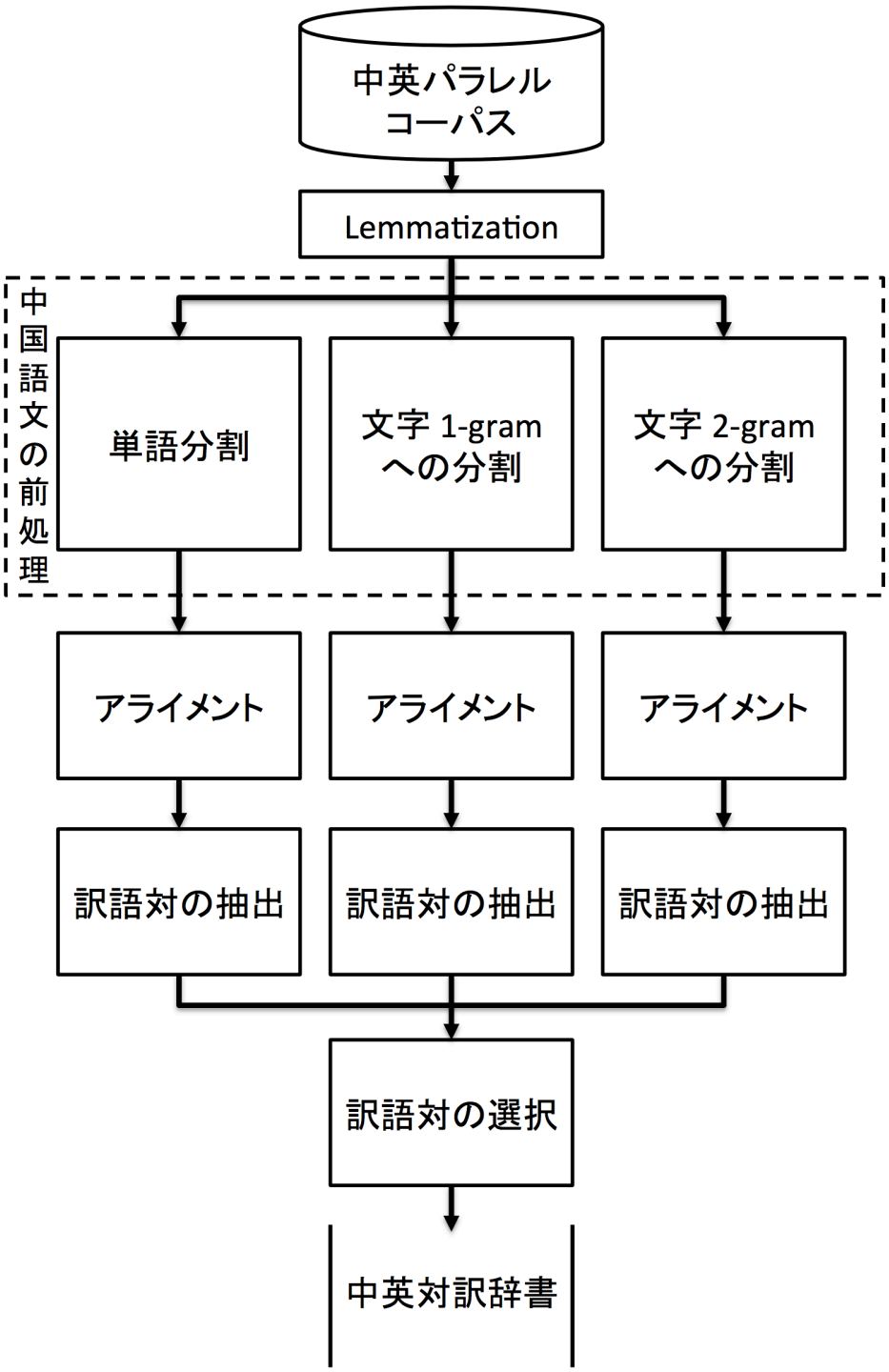
中英パラレル
コーパス
Lemmatization
中国語文の前処理
単語分割
文字 1-gram
への分割
文字 2-gram
への分割
アライメント
アライメント
アライメント
訳語対の抽出
訳語対の抽出
訳語対の抽出
訳語対の選択
中英対訳辞書

図 3.1: 提案手法の概要

3.2 前処理

中国語，英語のそれぞれの文に対して前処理を行う．英語文については，単語の正規化ために，lemmatization により各単語を原形に直す．中国語文については，単語境界が明示されていないため，文の単語への分割，文字 1-gram への分割，文字 2-gram への分割という 3 通りの前処理を行う．

3.2.1 英語文の前処理

英語文については lemmatization を行う．lemmatization とは, 文中に出現するそれぞれの単語を原型に直す処理である．英語は屈折語であり，単語は様々な形に活用する．以下に活用形と原型の例を示す．

dogs – dog, cats – cat
doing – do, done – do
better – good, best-good

1 行目は，名詞の複数形と単数形の例である．この場合，単数形が原型となる．2 行目は動詞の活用の例である．動詞は，過去形，過去分詞形，分詞形などに活用する．3 行目は形容詞の比較級，最上級の例である．lemmatization は対訳辞書を自動構築する際に重要な役割を果たす．lemmatization をしない場合，原型で出現する単語と活用形で出現する単語は異なる単語として扱われる．後述する単語アライメントでは，対訳の関係にある文中に数多く出現する中国語単語と英単語を対応付ける．活用語と原型の語を異なる単語として扱うと，中国語単語と英単語の共起頻度を正確に見積もることができない．例えば，dogs と dog が異なる単語として扱われたとする．パラレルコーパスにおいて，dogs や dog が中国語の「狗」(犬) とよく共起したとしても，dogs と「狗」，dog と「狗」の共起頻度は別々にカウントされる．そのため，dog と「狗」を訳語対として抽出できない可能性がある．lemmatization により，dogs を dog に正規化すれば，このような問題を避けることができる．このように，活用語と原型の語は同じ単語として取り扱われるべきである．

本研究では，lemmatization は Stanford CoreNLP[11] を用いて行う．以下の英語文を例に Stanford CoreNLP による lemmatization を説明する．

> Earlier the BBC was shown the extent of destruction rolled by the battles to control the city

図 3.2 はこの文を Standford CoreNLP で解析した結果である．3 行目以降は文中の 1 つの単語に対応する．Text は単語の出現形，CharacterOffsetBegin と CharacterOffsetEnd は単語の文中における開始位置と終了位置を，PartOfSpeech は単語の品詞を，Lemma は原型を表わす．この結果から Lemma の情報を取り出すと，次の文が得られる．

Sentence #1 (17 tokens):
Earlier the BBC was shown the extent of destruction rolled by the battles to control the city
[Text=Earlier CharacterOffsetBegin=0 CharacterOffsetEnd=7 PartOfSpeech=JJR Lemma=earlier]
[Text=the CharacterOffsetBegin=8 CharacterOffsetEnd=11 PartOfSpeech=DT Lemma=the]
[Text=BBC CharacterOffsetBegin=12 CharacterOffsetEnd=15 PartOfSpeech=NNP Lemma=BBC]
[Text=was CharacterOffsetBegin=16 CharacterOffsetEnd=19 PartOfSpeech=VBD Lemma=be]
[Text=shown CharacterOffsetBegin=20 CharacterOffsetEnd=25 PartOfSpeech=VBN Lemma=show]
[Text=the CharacterOffsetBegin=26 CharacterOffsetEnd=29 PartOfSpeech=DT Lemma=the]
[Text=extent CharacterOffsetBegin=30 CharacterOffsetEnd=36 PartOfSpeech=NN Lemma=extent]
[Text=of CharacterOffsetBegin=37 CharacterOffsetEnd=39 PartOfSpeech=IN Lemma=of]
[Text=destruction CharacterOffsetBegin=40 CharacterOffsetEnd=51 PartOfSpeech=NN
Lemma=destruction]
[Text=rolled CharacterOffsetBegin=52 CharacterOffsetEnd=58 PartOfSpeech=VBN Lemma=roll]
[Text=by CharacterOffsetBegin=59 CharacterOffsetEnd=61 PartOfSpeech=IN Lemma=by]
[Text=the CharacterOffsetBegin=62 CharacterOffsetEnd=65 PartOfSpeech=DT Lemma=the]
[Text=battles CharacterOffsetBegin=66 CharacterOffsetEnd=73 PartOfSpeech=NNS Lemma=battle]
[Text=to CharacterOffsetBegin=74 CharacterOffsetEnd=76 PartOfSpeech=TO Lemma=to]
[Text=control CharacterOffsetBegin=77 CharacterOffsetEnd=84 PartOfSpeech=VB Lemma=control]
[Text=the CharacterOffsetBegin=85 CharacterOffsetEnd=88 PartOfSpeech=DT Lemma=the]
[Text=city CharacterOffsetBegin=89 CharacterOffsetEnd=93 PartOfSpeech=NN Lemma=city]

図 3.2: Stanford CoreNLP の結果

earlier the BBC be show the extent of destruction roll by the battle to control the city

この例では，was が be に，shown が show に，rolled が roll に，battles が battle に変換されている．活用形を原型に直す他に，文頭が大文字の単語は全て小文字に変換される．例えば，Earlier は earlier に変換されている．ただし，常に大文字が使われる単語は変換されない．例えば,「I（私）」は文のどこに現われても大文字のままである．BBC のような固有名詞もそのまま変わらない．大文字のまま保たれる単語には他にも次のようなものがある．Russia，Chengdu，Jack のような地名．国名や人の名前などの固有名詞．Russian，Chinese のような言語や民族の名詞や形容詞．Sunday，August のような曜日，月の名前．CCTV のようなの略語．

3.2.2 中国語文の前処理

中国語文については，単語境界が明示されていないため，中国語の単語もしくは単語に相当する単位に分割する．具体的には以下の 3 通りの処理を行う．

1. 単語分割

   既存の単語分割ツールを用いて，文を単語に分割する．本論文では，単語分割ツールとして jieba を用いる．jieba の単語分割のアルゴリズムは 2.1.2 項で紹介した．以下に例を示す．

目睹了为争夺该市战斗所造成的破坏
→目睹 了为 争夺 该市 战斗 所 造成 的 破坏

2. 文字 1-gram への分割

   中国語の文を文字単位に分割する．例を以下に示す．

   目睹了为争夺该市战斗所造成的破坏
   →目 睹 了 为 争 夺 该 市 战 斗 所 造 成 的 破 坏

3. 文字 2-gram への分割

   中国語の文を 2 文字の単位に分割する．具体的には，中国語の文を n 個の文字列 $c_1c_2\cdots c_n$ とするとき，それに含まれる全ての文字 2-gram c_ic_{i+1} $(1 \leq i < n)$ の列に変換する．また，c_ic_{i+1} は i の昇順に並べる．例を以下に示す．

   目睹了为争夺该市战斗所造成的破坏
   →目睹 睹了了为 为争 争夺 夺该 该市 市战 战斗 斗所 所造 造成 成的 的破 破坏

3.3 単語アライメント

前処理された 3 種類のパラレルコーパスに対して単語のアライメントを推定する．単語のアライメントとは, 本研究では，英語文に出現する単語に対し, それに対応する中国語の単語もしくは単語に相当する単位を決定する処理である. 前処理によって，英単語と中国語の単語, 文字 1-gram もしくは 2-gram が対応付けられる．

本研究では，GIZA++を用いて単語のアライメントを推定する．GIZA++は，統計的機械翻訳で用いることを前提に，IBM モデルを実装した標準的な単語アライメントツールである．GIZA++はいくつかのステップを経てアライメントを決める．以下，各ステップを詳述する．

まず，plain2snt.out というコマンドを用いて，パラレルコーパスに出現する全ての単語に ID 番号を付与する．また，単語の頻度もカウントする．その結果，表 3.1 と表 3.2 に示すデータが得られる．

表 3.1 はパラレルコーパスに出現した単語のリストであり，単語 ID，それに対応する単語，その単語の出現頻度が出力されている．これは英語の単語リストであるが，中国語の単語リストも同様に作成される．表 3.2 は対訳文を表わす．このファイルでは，1 組の対訳文は 3 行で表わされる．1 行目は文の組の出現頻度である．2 行目は，源言語 (この例では英語) の文に出現する単語の ID のリストである．3 行目は，目標言語 (この例では中国語) の文に出現する単語の ID のリストである．

次に，snt2cooc.out というコマンドを使って共起ファイルを獲得する．その出力結果を表 3.3 に示す．この出力ファイルの各行は，原言語の単語の ID と目標言語の単語の ID の組である．ただし，パラレルコーパスにおいて対訳関係にある文に共に出現する単語の組のみが出力される．対訳関係にある文に一度も共に出現しない単語の組は出力されない．

表 3.1: plain2snt.out による単語リストの出力 (一部)

単語 ID	単語	単語の出現回数
2	the	88442
3	Basic	20
4	Law	699
5	of	48966

表 3.2: plain2snt.out による対訳文の出力 (一部)

文の出現回数	1
源言語の単語 ID	2 3 4 5 2 6 7 8 9 5 2 10 11 12 5 13
目標言語の単語 ID	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

表 3.3 の例の場合，源言語の単語 ID「3」は表 3.1 より Basic に該当するが，Basic を含む英訳と対訳関係にある文に出現した中国語単語の ID が「目標言語の単語 ID」の列に表示されている．

表 3.3: snt2cooc.out の出力 (一部)

源言語の単語 ID	目標言語の単語 ID
3	2
3	3
3	4
3	5
3	6
...	...
3	18

次に，mkcls で単語のクラスタリングを行う．ここでのクラスタリングとは，同じ品詞を持つと思われる単語をまとめてクラスタを作成する処理である．表 3.4 は，単語からクラスタ (品詞) への対応を表わす．このファイルでは単語はアルファベット順に並べられている．一方，表 3.5 は，品詞から単語への対応を表わしている．それぞれの品詞に対し，それに属する単語のリストが表示されている．

最後に，GIZA++というコマンドによって単語のアライメントを推定し，以下の出力ファイルを得る．

- T TABLE (Translation Table)

  T TABLE は，IBM Model 1 から Model 3 により作成された翻訳確率 $P(t|s)$ のデータである．s は源言語，t は目標言語の単語を表わす．T TABLE の例を表 3.6 に示す．T TABLE の各行は，源言語の単語 ID(s_{id})，目標言語の単語 ID(t_{id})，源言語の単語が目標言語の単語へ翻訳される確率 ($P(t_{id}|s_{id})$) で構成される．

表 3.4: mkcls による単語から品詞への対応 (一部)

単語 (アルファベット順)	品詞
Deputy	101
Design	84
Destruction	88
Development	60
Discovery	12

表 3.5: mkcls による品詞から単語への対応 (一部)

品詞	品詞に対応する単語
0:	$,
1:	
2:	ACC,Accurate,Airmen,Bureaux,COMPANY,Chihe,Chines,ELEMENT, EXECUTION,Economy,Hired,MANAGING,MARTIAL,METROLOGY, Martial,PARTNERSHIP,Provisional,TRANSIT,Unauthorized,abound, ambiguous,anthelmintic,article26and,diverse,judicature,lawbreaking, martial,oldage,pla

表 3.6: T TABLE (一部)

源言語の単語 ID	目標言語の単語 ID	$p(t\_id\|s\_id)$
0	2	0.00841958
0	4	0.0388686
0	5	0.00272237
0	6	0.000239207

- N TABLE (Fertility Table)

  N TABLE は，源言語の単語の繁殖数の確率分布のデータである．一般に，源言語の 1 つの単語は目標言語の複数の単語，あるいは 0 個の単語に翻訳される．0 個の単語に翻訳されるときは，源言語に対応する単語は目標言語側に存在しないことを表わす．繁殖数とは，源言語の単語に対応する目標言語の単語の数である．N TABLE の例を表 3.7 に示す．各行は，源言語の単語 ID($source\_token\_id$) と，その単語の繁殖数が i である確率 (p_i) で構成される．表 3.1 に示す通り，ID が 2 である単語は the であり，the は中国語の文に対応する単語がないことがほとんどなので，p_0 の確率が最も大きくなっている．

- A TABLE

  A TABLE は，源言語の文で特定の位置に出現する単語が，目標言語の文で特定の

表 3.7: N TABLE (一部)

$(source\_token\_id)$	(p_0)	(p_1)	(p_2)	(p_3)	(p_4)
2	0.972511	0.00132279	0.0245684	0.000164296	0.000893882
	(p_5)	(p_6)	(p_7)	(p_8)	(p_9)
	0.000219394	0.000151325	0.000121717	2.07883e-06	4.55978e-05

位置に出現する単語に翻訳される確率を表わす．A TABLE の例を表 3.8 に示す．i は源言語の文における単語の位置，j は目標言語の文における単語の位置，l は源言語の文の長さ (単語数)，m は目標言語の文の長さ (単語数)，$p(i|j,l,m)$ は，長さ l，m の源言語と目標言語の文があったとき，目標言語の j 番目の位置にある単語が源言語の i 番目の位置にある単語と対応関係にある確率である．

表 3.8: A TABLE の例 (一部)

i	j	l	m	$p(i\|j,l,m)$
0	1	1	100	0.000166204
1	1	1	100	0.999834
0	2	1	100	4.30877e-05
1	2	1	100	0.999957
0	3	1	100	0.0401887

- ALIGNMENT FILE

  ALIGNMENT FILE は，パラレルコーパスにおいて対訳関係にある文の組に対し，単語間の対応関係 (アライメント) を明示したファイルである．

本研究では，最終的には，GIZA++の出力として ALIGNMENT FILE を用いる．ALIGNMENT FILE の例を図 3.3, 3.4, 3.5 に示す．図 3.3 は前処理として単語分割を行ったとき，図 3.4 は前処理として文字 1-gram への分割を行ったとき，図 3.5 は前処理として文字 2-gram への分割を行ったときの結果の例である．これらのファイルでは 3 行で 1 つの文の組における単語アライメントを表わす．1 行目はヘッダである．2 行目は目標言語の文 (中国語の文) である．3 行目は源言語の文 (英語の文) である．3 行目において，英単語の間に続く括弧内の数字は，その英単語に対応する中国語の単語の位置を表わす．図 3.4 の stand({ 10 11 }) のように，1 つの単語に複数の数字が割り当てられているときは，中国語文の複数の単語 (この場合 10 番目と 11 番目の文字) に対応することを表わす．また，NULL は特別なシンボルで，英語文において対応する単語がない中国語文の単語の位置を表わす．図 3.3 では，4 番目の中国語の単語には，対応する英単語はない．

図 3.6, 3.7, 3.8 は，それぞれ図 3.3, 3.4, 3.5 の単語アライメントの結果をわかりやすく図示したものである．ただし，一部の単語アライメントは省略されている．

```
# Sentence pair (49) source length 17 target length 7 alignment score : 2.22789e-16
经 全国人民代表大会常务委员会 发回 的 法律 立即 失效
NULL ({ 4 }) any ({ }) law ({ }) return ({ }) by ({ 1 }) the ({ }) stand ({ 2 })
committee ({ 3 }) of ({ }) the ({ }) National ({ }) People ({ 5 }) be ({ })
 Congress ({ }) shall ({ }) immediately ({ 6 }) be ({ }) invalidate ({ 7 })
```

図 3.3: GIZA++による単語アライメントの出力 (単語分割)

```
# Sentence pair (49) source length 17 target length 23 alignment score : 4.05394e-32
经 全 国 人 民 代 表 大 会 常 务 委 员 会 发 回 的 法 律 立 即 失 效
NULL ({ 1 17 18 }) any ({ }) law ({ 19 }) return ({ 15 16 }) by ({ }) the ({ })
stand ({ 10 11 }) committee ({ 12 13 14 }) of ({ }) the ({ }) National ({ 2 3 })
People ({ 4 5 }) be ({ }) Congress ({ 6 7 8 9 }) shall ({ }) immediately ({ 20 21 })
be ({ }) invalidate ({ 22 23 })
```

図 3.4: GIZA++による単語アライメントの出力 (文字 1-gram)

```
# Sentence pair (49) source length 17 target length 22 alignment score : 1.7893e-44
经全 全国 国人 人民 民代 代表 表大 大会 会常 常务 务委 委员 员会 会发 发回 回的 的法 法律
 律立 立即 即失 失效
NULL ({ 18 }) any ({ }) law ({ }) return ({ }) by ({ }) the ({ }) stand ({ 9 10 11 })
 committee ({ 12 13 }) of ({ }) the ({ }) National ({ 1 2 3 4 }) People ({ }) be ({ })
Congress ({ 5 6 7 8 }) shall ({ }) immediately ({ 20 }) be ({ })
invalidate ({ 14 15 16 17 19 21 22 })
```

図 3.5: GIZA++による単語アライメントの出力 (文字 2-gram)

3.4 訳語対の抽出

アライメントの結果から，対応関係にある中国語の単語と英単語の組を訳語対の候補として抽出する．GIZA++によるアライメントは，1 つの英単語に対し，複数の中国語の単語もしくは文字 n-gram が対応付けられることがある．このとき，前処理によって以下の方法で訳語対の候補を抽出する．

- 中国語の文を単語分割した場合，文字 1-gram に分割した場合

  複数の中国語の単語もしくは文字を連結した文字列を中国語単語として訳語対の候補を抽出する．図 3.4(図 3.7) では，stand, committee, National, Park, Congress, immediately, invalidate は複数の文字に対応しているが，これらの文字を連結したものを中国語の単語とする．例えば，(People, 人民) という訳語対を得る．図 3.5 は，図 3.3(図 3.6) の結果から得られる訳語対の候補である．また，図 3.6 は，図 3.4(図 3.7) の結果から得られる訳語対の候補である．

  また，GIZA++による単語アライメントでは，源言語の単語が目標言語の文において連続して出現しない複数の単語に対応付けられることがある．例を図 3.11 に示す．これは前処理として中国語の文を文字 1-gram に分割している．同図における

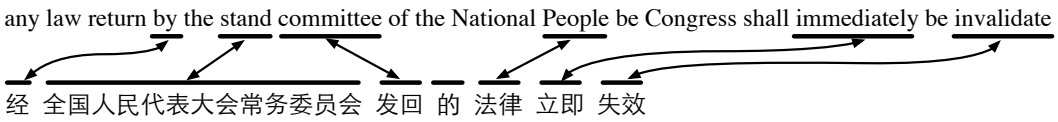

図 3.6: アライメントの例 (単語分割)

図 3.7: アライメントの例 (文字 1-gram)

図 3.8: アライメントの例 (文字 2-gram)

by 经
stand 全国人民代表大会常务委员会
committee 发回
People 法律
immediately 立即
invalidate 失效

図 3.9: 訳語対の候補の抽出例 (単語分割)

law 律
stand 常务
committee 委员会
National 全国
People 人民
Congress 代表大会
immediately 立即
invalidate 失效

図 3.10: 訳語対の候補の抽出例 (文字 1-gram)

「utility ({ 10 11 12 13 18 27 28 }) 」は，utility という単語が中国語の複数の文字に対応付けられているが，これらは文中で連続して出現していない．英単語が複数の不連続な中国語の単語 (もしくは文字 1-gram) と対応付けられるとき，中国語の単語 (もしくは文字 1-gram) の集合を連続する部分集合に分割し，それぞれの部分

集合を連結したものを中国語の訳語として訳語対を抽出する．このとき，1 つの英単語に対して複数の訳語対を得る．図 3.11 の utility の例では，中国語の文字のグループを { 10, 11, 12, 13}, { 18 }, { 27, 28 } の 3 つに分割し，以下に示す 3 つの訳語対を抽出する．

utility 公共事业
utility 付
utility 问题

Sentence pair (12) source length 24 target length 28 alignment score : 5.35551e-42
沙 特 称 将 削 减 燃 油 和 公 共 事 业 补 贴 来 对 付 创 纪 录 的 预 算 赤 字 问 题
NULL ({ 16 17 22 }) Saudi ({ 1 2 }) Arabia ({ }) say ({ 3 }) it ({ }) will ({ 4 }) cut ({ 5 6 }) fuel ({ 7 8 }) and ({ 9 }) utility ({ 10 11 12 13 18 27 28 }) subsidy ({ 14 15 }) in ({ }) the ({ }) country ({ }) as ({ }) part ({ }) of ({ }) measure ({ }) to ({ }) deal ({ }) with ({ }) a ({ }) record ({ 19 20 21 }) budget ({ 23 24 }) deficit ({ 25 26 })

図 3.11: 複数の不連続な単語に対応付けられる例

- 中国語の文を文字 2-gram に分割した場合

  複数の文字 2-gram が 1 つの英単語に対応付けられるときには訳語対の候補を抽出しない．したがって，抽出される訳語対において，中国語の単語は必ず 2 文字となる．図 3.5(図 3.8) では，immediately に対して 1 つの文字 2-gram が対応しているため，訳語対を抽出する．一方，stand や committee に対しては複数の文字 2-gram が対応するため，訳語対を抽出しない．この例から抽出される訳語対の候補は図 3.12 の通り 1 組だけである．

immediately 立即

図 3.12: 訳語対の候補の抽出例 (文字 2-gram)

これまでの手続きによって，TP_{seg}，TP_{1g}，TP_{2g} という 3 通りの訳語対の候補の集合が得られる．

TP_{seg} :
ツールによって単語分割された中国語文と英語文の組の集合 (パラレルコーパス) から抽出された訳語対の候補の集合

TP_{1g} :
文字 1-gram に分割された中国語文と英語文の組の集合 (パラレルコーパス) から抽出された訳語対の候補の集合

TP_{2g} :
文字 2-gram に分割された中国語文と英語文の組の集合 (パラレルコーパス) から抽出された訳語対の候補の集合

3.5 訳語対の獲得

訳語対の候補の中には正しくないものも多数含まれる．この中から正しい訳語対を選択する手法について述べる．

まず，以下に述べる簡単なヒューリスティクスを用いて，明らかに正しくない，あるいは獲得しても意味のない訳語対の候補を除外する．

- 英単語が付属語のとき

  ストップワードのリストを用意し，英単語がそのリストに登録されていれば，訳語対の候補を除外する．ストップワードとは，意味を持たない単語 (主に機能語) のリストである．例えば，英語では，“a”, “and”, “is”, “the” などの単語がストップワードに相当する．これらの単語は，対訳辞書のエントリとして獲得してもあまり意味がない．そのため，訳語対の候補から除外する．本研究で用いるストップワードの数は 448 である．

- 中国語の単語が 1 文字のとき

  中国語の単語が 1 文字のとき，中国語の単語分割が誤っている可能性が高い．また，単語分割が正しくても，1 文字の中国単語の多くは接頭辞，接尾辞，ストップワードである可能性が高く，訳語対を獲得しても意味がない．したがって，中国語の単語が 1 文字の訳語対の候補は除外する．除外される不適切な訳語対の例を図 3.13 に示す．

  chapter 第
  addition 除
  section 节
  Sino 中
  region 区
  accordance 依
  not 不
  judiciary 司
  decree 令

図 3.13: 中国語の単語が 1 文字の訳語対の例

- 英単語が数字のとき

4章で後述する評価実験において，法律のパラレルコーパスから訳語対を獲得した際，(1, 第1条) のように，英語の数字と中国語の条文の番号が対応付けられた訳語対が多く得られた．また，新聞のパラレルコーパスから訳語対を獲得した際，(2016,2016年) のように，英語の数字と中国語の日付が対応付けられた訳語対が多く得られた．これらは対訳辞書に収録するような適切な訳語対ではない．そのため，英単語が数字のとき，訳語対の候補を除外する．

- 英単語が6つ以上の中国語単語に対応付けられたとき

  1つの英単語が複数の中国語の単語に対応することがあるが，多くの中国語単語に対応付けられるときは誤りである可能性が高い．そのため，6つ以上の中国語単語に対応付けられる英単語があったとき，その全ての訳語対を除外する．図3.13に英単語が6つ以上の中国語単語に対応付けられた場合を示す．

shipwreck 沉没, 起浮清除,残骸,起浮,引起,是由于
transform 转化高技术,持有者,采用,单位,与原,实施,科技成果实施转化,单位合作实施, 与原单位, 转化,科技成果,合作,改造,科技成果转化活动,科技成果转化
weather 做好人工影响天气,订正,影响,订正气象法,跨地区跨部门, 灾害性天气警报气象法,天气, 灾害性天气,灾害性天气警报,气象法,人工影响天气
purchaser 收购要约期限内,方式,收购人还,达到百分之三十,书面报告并予公告,持有,十五日, 期限内,期限内收购人,方式收购,少于三十日,要约方式,收购期限内, 采取要约收购方式,收购人,采取, 同等条件,有效期限,数达到,收购方式,收购,收购要约,收购人
recording 录音录像制作者,制作录像制品, 著作权法施例,录音录像制品,录音录像, 录音录像制作者制作录音,录音制作者, 施例,制作录音制品,录音制品,著作权法,制品,制作,制作者, 录音

図 3.14: 6つ以上の中国語単語に対応する英単語の例

次に，残された訳語対の和集合をTPとする．すなわち，$TP = TP_{seg} \cup TP_{c1} \cup TP_{c2}$である．$TP$における訳語対$tp$に対するスコアを式(3.2)のように定義する．

$$Score(tp) = \max_{x \in \{seg, c1, c2\}} Score_x(tp) \tag{3.1}$$

$$Score_x(tp) = \begin{cases} \dfrac{O_x(tp)}{\sum_{tp \in TP_x} O_x(tp)} & \text{if } tp \in TP_x \\ 0 & \text{if } tp \notin TP_x \end{cases} \tag{3.2}$$

式(3.1)は，$TP_{seg}, TP_{c1}, TP_{c2}$のそれぞれにおける$tp$のスコアを$Score_{seg}, Score_{c1}, Score_{c2}$と定義し，その最大値を$tp$のスコアとすることを表わす．$Score_x$ (xはseg, $c1$, $c2$のいずれか)は式(3.2)のように定義する．$O_x(tp)$はTP_xにおけるtpの出現頻度であり，式(3.2)の分母はTP_xにおける全ての訳語対の候補の頻度の総和である．すなわち，$Score_x(tp)$はtpの相対出現頻度である．表3.9に訳語対のスコアの計算例を示す．

最後に，以下の2つの条件を満たすtpを訳語対として獲得し，対訳辞書を得る．

表 3.9: 訳語対のスコアの例

tp	$Score_{seg}(tp)$	$Score_{c1}(tp)$	$Score_{c2}(tp)$	$Score(tp)$
(people, 人民)	0.000991	0.164	0.248	0.248
(administrative, 行政)	0.0669	0.170	0.0636	0.170
(autonomous, 自治)	0	0.122	0.00621	0.122
(individual, 个人)	0.0376	0.0370	0.00593	0.0376
(work, 工作)	0.00714	0.0111	0.000488	0.0111
(export, 出口)	0.0199	0.0363	0.0112	0.0363

1. $Score_{seg}, Score_{c1}, Score_{c2}$ のうち 2 つ以上が 0 より大きい．すなわち，2 つ以上のパラレルコーパスから獲得された訳語対である．

2. $Score(tp)$ が閾値 T より大きい．

第4章　評価実験

本章では，提案手法の評価実験について述べる．まず，4.1 節ではで実験で使用したデータについて説明する．次に，4.2 節で実験の方法について述べる．次に，4.3 節で実験の結果について考察する．最後に，4.4 節では実際に獲得された訳語対の例を示す．

4.1　実験データ

訳語対を獲得する中英コーパスとして以下の 2 つを用いる．

- BBC news パラレルコーパス

  ウェブサイト[1] から，2015 年と 2016 年の中国語と英語の新聞記事の対訳をクロールして集めたパラレルコーパス．文の組の総数は 22,560 である．

  BBC では，英語でニュースを放送している．また，英語のニュースを中国語に翻訳者が翻訳して公開している．また，中国語と英語の新聞記事に対し，中国語文と英語文の対応関係を作業者が人手で与え，文のアライメントが付与された対訳コーパスを構築している．新聞記事は様々な話題を含み，使われる単語の種類も多い．さらに，2015 年と 2016 年の新聞記事を使っていることから，比較的新しい単語も多く含まれている．このような単語の中にはまだ既存の対訳辞書に載っていないものもある．さらに，新聞記事では人名や地名などの固有名詞も多く使われる．しかし，固有名詞は数が多いので，その訳語が既存の対訳辞書に載っていないことも多い．したがって，新聞記事のパラレルコーパスから対訳辞書を自動構築できれば，新語や固有名詞の対訳が新たに獲得されることになり，その意義は大きい．

- Parallel Corpus of China's Law Documents
  (PCCLD)

  中国語の法律とその訳文を集めたパラレルコーパス[2]．文の組の総数は 31,517 である．

  国家 (中国) 及び地域の法律や規制を集めたものである．法律に関する人々の義務や契約に関する文書も含まれる．収録されている地域の法律としては，香港の法律と台湾の法律がある．マカオの法律や規制の一部も収録されている．法令文書は特有

[1] http://www.kekenet.com/broadcast/bbc/
[2] http://corpus.usx.edu.cn/lawcorpus1/index.asp

の専門用語や定型表現が使われているため，既存の対訳辞書に含まれない未知語を多く含むと考えられる．また，法律のコーパスは法的な活動に関する知識や法令に関する知識を獲得するためのリソースとして重要であり，これから法律用語に関する対訳辞書を構築することの意義は大きい．

4.2 実験手順

4.1 節で述べた 2 つのパラレルコーパスから， 以下の 4 つの手法で訳語対を自動獲得し，その結果を比較する．

M_{seg} 中国語の文を単語分割ツールで単語に分割した後，訳語対を獲得する手法．

M_{c1} 中国語の文を 1 文字に分割した後，訳語対を獲得する手法．

M_{c2} 中国語の文を文字 2-gram に分割した後，訳語対を獲得する手法．

M_{pro} 提案手法．上記 3 つの手法を組み合わせて用いる手法．

それぞれの手法では，訳語対がスコアの降順に並べられる．M_{seg}, M_{c1}, M_{c2} では式 (3.2) が，M_{pro} では式 (3.1) が訳語対のスコアの定義である．スコアの上位 α 件の訳語対を獲得し，それから既存の中英対訳辞書に含まれる訳語対を削除する．以下，既存の中英対訳辞書に含まれる訳語対を「既知の訳語対」，辞書に含まれない新しい訳語対を「未知の訳語対」と呼ぶ．残された未知の訳語対の数が 100 件になるまで α の数を増やしていく．この 100 件の訳語対について，それらが正しいかを人手で判定し，精度を算出する．以下，この精度を P@100 と記す．また，α 件の訳語対の精度を P@100+X とし，これも評価基準とする．P@100 は未知の訳語対のみを，P@100+X は既知の訳語対 (X 件) と未知の訳語対 (100 件) の両方が評価の対象となる．

P@100 や P@100+X を算出する際に用いる既存の対訳辞書として，LDC English Chinese bilingual wordlists を用いる．この辞書に含まれる英単語の数は 56,071，訳語対の総数は 111,008 である．英中単語対応表には英単語と中国語訳語のみが記述され，英単語と中国語訳語の品詞情報はない．また，1 つの英単語は複数の中国語に対応することがある．例を図 4.1 に示す．この例では，extensions と extensity には 3 つの中国語の訳語が対応付けられている．このとき，英単語と全ての中国語の訳語の組を作成し，それを対訳辞書に追加する．また，元の組は削除する．これにより，1 つの中国語と 1 つ英語の訳語対からなる対訳辞書が作成される．図 4.2 は上記の変換処理を行った後の対訳辞書である．

さらに，各手法が未知の訳語対をどれだけ獲得できるかを評価するために，新訳語対獲得率 $\mathrm{R_{new}}$ を式 (4.1) のように定義する．

$$\mathrm{R_{new}} = \frac{\text{出現頻度 5 以上かつ未知の訳語対の数}}{\text{出現頻度 5 以上の訳語対の数}} \tag{4.1}$$

extensionality /外延性/
extensions /扩张/延长/外延/
extensity /广阔性/广大性/空间性/

図 4.1: LDC English Chinese bilingual wordlists(一部)

extensionality 外延性
extensions 扩张
extensions 延长
extensions 外延
extensity 广阔性
extensity 广大性
extensity 空间性

図 4.2: 加工された LDC English Chinese bilingual wordlists (一部)

未知の訳語対かを判定する際に用いる既存の中英対訳辞書は，同じく LDC English Chinese bilingual wordlists を利用する．R_{new} を算出する際には，正しくない訳語対も未知の訳語対とみなされていることに注意していただきたい．

4.3 実験結果

表 4.1 に 2 つのパラレルコーパスから抽出された訳語対候補の数を示す．ここでの訳語対の候補は，3.5 節で述べた 4 つのヒューリスティクスを適用する前のものである．法律のパラレルコーパスの方が新聞のパラレルコーパスと比べて文の組の数が多いが，獲得された訳語対の候補は新聞のパラレルコーパスの方が多い．新聞の方が法律よりも多様な単語が使われているためと考えられる．また，文字の 2-gram に分割した後に得られる訳語対の候補の数は，新聞のパラレルコーパスでは単語分割や文字 1-gram のケースよりも多いが，法律のパラレルコーパスでは逆に少なくなっている．表 4.2 は，3.5 節のヒューリスティクスを適用し，誤りと思われる訳語対の候補を削除した後の訳語対の候補の数を示している．新聞，法律とも訳語対の候補の数が減っているが，法律の方がより多くの訳語対が削除されていることがわかる．法律のコーパスでは，英語の数字と中国語の条文の番号の組が多く獲得され，これらがルールによって削除されたためと考えられる．

表 4.1: 獲得された訳語対の候補の数 (ルール適用前)

	単語分割	1-gram	2-gram
新聞	83390	80901	189510
法律	73152	65355	45054

表 4.2: 獲得された訳語対の候補の数 (ルール適用後)

	単語分割	1-gram	2-gram
新聞	41041	38785	50482
法律	18164	20013	14998

次に，新聞記事のパラレルコーパス (BBC news) から 4 つの手法によって獲得された訳語対の評価結果を表 4.3 に示す．提案手法 M_{pro} の P@100 と P@100+X は 3 つのベースライン手法 (M_{seg}, M_{c1}, M_{c2}) を上回る．したがって，ツールによって文を単語に分割してから訳語対を抽出するだけでなく，文字の 1-gram，2-gram に分割してから訳語対を抽出することで，訳語対獲得の精度が向上することが確認された．また，M_{c2} の正解率は他の手法と比べて低いが，これは獲得される訳語対が 2 文字の中国語と英単語の組に限定されているためと考えられる．

R_{new} についても，提案手法は他の 3 つのベースラインを上回る．ただし，M_{c2} との差はごくわずかである．しかし，M_{c2} の P@100 や P@100+X の値が低いことを考えると，M_{c2} の新訳語対獲得率が高いのは正しくない訳語対が多く取り出されているためと考えられる．したがって，提案手法は，未知の訳語対を獲得するという観点からもベースラインを上回る．

表 4.3: 実験結果 (新聞記事)

	P@100	P@100+X	R_{new}
M_{seg}	0.93	0.94	0.878
M_{c1}	0.93	0.94	0.873
M_{c2}	0.40	0.51	0.933
M_{pro}	0.94	0.95	0.935

法律のパラレルコーパス (PCCLD) から獲得された訳語対の評価結果を表 4.4 に示す．

表 4.4: 実験結果 (法律)

	P@100	P@100+X	R_{new}
M_{seg}	0.87	0.94	0.907
M_{c1}	0.94	0.95	0.896
M_{c2}	0.77	0.77	0.851
M_{pro}	0.95	0.96	0.934

この表から読み取れる傾向は表 4.3 に示した新聞コーパスの結果と同様である．すなわち，提案手法 M_{pro} の P@100 と P@100+X は 3 つのベースライン手法よりも高い．R_{new} についても，提案手法の 0.934 という値はベースラインを大きく上回る．

新聞記事コーパスと法律コーパスを比較すると，正解率の低い M_{c2} を除いて，法律の方が新聞記事と比べて $\mathrm{R_{new}}$ が高いもしくは同等であった．これは，法律のコーパスには専門用語が多く存在し，これらは既存の対訳辞書に含まれていないためと考えられる．P@100 について比較すると，M_{seg} は新聞コーパスの方が高いが，M_{c2} では法律コーパスの方が高く，他の 2 つは同等である．P@100+X について比較すると，M_{c2} では法律コーパスの方が新聞コーパスよりも高いが，それ以外は同等である．

4.4 獲得された訳語対の例

実験によって獲得された訳語対の例を示す．図 4.3，図 4.4，図 4.5 は，新聞のパラレルコーパスから，単語分割, 文字 1-gram への分割, 文字 2-gram への分割を前処理として獲得された訳語対の例である．図 4.6 は，同じく新聞のパラレルコーパスから，提案手法によって獲得された訳語対の例である．一方，図 4.7，図 4.8，図 4.9 は，法律のコーパスから，単語分割, 文字 1-gram への分割, 文字 2-gram への分割を前処理として獲得された訳語対の例である．図 4.10 は，同じく法律のパラレルコーパスから，提案手法によって獲得された訳語対の例である．これらの図の各行は，英単語，中国語の単語，訳語対のスコア，訳語対が正しいかどうか (○は正しい訳語対を，×は正しくない訳語対を表わす) を示している．また，各図は，それぞれの手法において，スコアの大きい上位 21 件の訳語対を示している．

英語	中国語	スコア	正解
President	总统	0.0147	○
US	美国	0.0113	○
news	新闻	0.0106	○
United	美国	0.0077	×
police	警方	0.0068	○
capital	首都	0.0056	○
city	城市	0.0048	○
country	国家	0.0048	○
Russia	俄罗斯	0.0048	○
international	国际	0.0047	○
UN	联合国	0.0047	○
political	政治	0.0046	○
syrian	叙利亚	0.0046	○
Greece	希腊	0.0045	○
Prime	总理	0.0043	○
News	新闻	0.0043	○
announce	宣布	0.0042	○
Iran	伊朗	0.0041	○
Britain	英国	0.0039	○
Trump	特朗普	0.0037	○
France	法国	0.0037	○

図 4.3: M_{seg} によって獲得された訳語対の例 (新聞)

英語	中国語	スコア	正解
President	总统	0.0161	○
news	新闻	0.0118	○
include	包括	0.0075	○
Obama	奥巴马	0.0074	○
World	世界	0.0070	○
Syria	叙利亚	0.0065	○
capital	首都	0.0061	○
international	国际	0.0060	○
UN	联合国	0.0058	○
political	政治	0.0056	○
United	美国	0.0055	×
british	英国	0.0055	○
Russia	俄罗斯	0.0053	○
syrian	叙利亚	0.0052	○
Iran	伊朗	0.0052	○
Greece	希腊	0.0049	○
Prime	总理	0.0047	○
News	新闻	0.0047	○
greek	希腊	0.0046	○
company	公司	0.0046	○
Britain	英国	0.0044	○

図 4.4: M_{c1} によって獲得された訳語対の例 (新聞)

英語	中国語	スコア	正解
firsttime	首次	0.0006	○
lastmonth	上月	0.0006	○
peacetalk	和谈	0.0006	○
WorldCup	界杯	0.0006	×
SouthAfrica	南非	0.0006	○
onSunday	周日	0.0006	○
Erdogan	多安	0.0006	×
Yanukovych	科维	0.0005	×
anti-government	反政	0.0005	○
civilwar	内战	0.0005	○
July	7月	0.0005	○
shia	叶派	0.0005	×
Nepal	泊尔	0.0005	×
humanrights	人权	0.0005	○
NATO	北约	0.0005	○
website	网站	0.0005	○
Vatican	蒂冈	0.0005	×
Thomas	托马	0.0004	×
Rio	里约	0.0004	○
onTuesday	周二	0.0004	○
Catholic	天主	0.0004	○

図 4.5: M_{c2} によって獲得された訳語対の例 (新聞)

英語	中国語	スコア	正解
President	总统	0.0161	○
US	美国	0.0160	○
report	报道	0.0130	○
news	新闻	0.0118	○
group	组织	0.0095	○
United	美国	0.0077	×
include	包括	0.0075	○
Obama	奥巴马	0.0074	○
World	世界	0.0070	○
Syria	叙利亚	0.0065	○
capital	首都	0.0061	○
official	官员	0.0060	○
city	城市	0.0060	○
international	国际	0.0060	○
UN	联合国	0.0058	○
islamic	伊斯兰	0.0057	○
political	政治	0.0056	○
support	支持	0.0055	○
british	英国	0.0055	○
leader	领导	0.0053	○
Russia	俄罗斯	0.0053	○

図 4.6: M_{pro} によって獲得された訳語対の例 (新聞)

英語	中国語	スコア	正解
shall	应当	0.2127	⭕
organ	机关	0.0392	⭕
measure	措施	0.0266	⭕
citizen	公民	0.0215	⭕
State	国家	0.0205	⭕
facility	设施	0.0183	⭕
design	设计	0.0183	⭕
directly	直接	0.0171	⭕
Hong	香港特别行政区	0.0146	✕
regulatory	监督管理机构	0.0127	⭕
refer	是指	0.0121	⭕
directly	直接责任人员	0.0114	✕
iv	第四章	0.0096	⭕
ii	第二章	0.0089	⭕
enjoy	享有	0.0087	⭕
iii	第三章	0.0080	⭕
measure	办法	0.0080	⭕
VI	第六章	0.0069	⭕
following	下列	0.0069	⭕
enact	制定	0.0068	⭕
aircraft	民用航空器	0.0067	⭕

図 4.7: M_{seg} によって獲得された訳語対の例 (法律)

英語	中国語	スコア	正解
people	人民	0.1648	○
state	国家	0.0940	○
Council	务院	0.0868	×
unit	单位	0.0794	○
autonomous	自治	0.0399	○
State	国家	0.0356	○
People	人民	0.0346	○
economic	经济	0.0339	○
Kong	香港	0.0292	×
entity	单位	0.0283	○
fix	下有	0.0262	×
violation	违反	0.0254	○
facility	设施	0.0226	○
design	设计	0.0219	○
National	全国	0.0208	○
social	社会	0.0201	○
development	开发	0.0171	○
development	发展	0.0171	○
stand	常务	0.0168	○
trust	信托	0.0159	○
meteorological	气象	0.0142	○

図 4.8: M_{c1} によって獲得された訳語対の例 (法律)

英語	中国語	スコア	正解
People	人民	0.0127	○
include	包括	0.0097	○
detention	刑或	0.0072	×
execution	执行	0.0043	○
refer	是指	0.0037	○
Enterprises	企业	0.0033	○
citizen	公民	0.0033	○
commercial	商业	0.0028	○
engage	从事	0.0026	○
environment	环境	0.0025	○
notice	通知	0.0024	○
forbid	禁止	0.0019	○
hazardous	危险	0.0019	○
designate	指定	0.0019	○
energy	能源	0.0018	○
deprivation	制或	0.0017	×
Trade	贸易	0.0015	○
support	支持	0.0014	○
processing	加工	0.0014	○
trust	信托	0.0014	○
rural	农村	0.0013	○

図 4.9: M_{c2} によって獲得された訳語対の例 (法律)

英語	中国語	スコア	正解
people	人民	0.2480	○
administrative	行政	0.1705	○
state	国家	0.0940	○
Council	务院	0.0868	×
unit	单位	0.0794	○
organ	机关	0.0662	○
provision	规定	0.0521	○
accordance	依照	0.0487	○
product	产品	0.0401	○
measure	措施	0.0382	○
individual	个人	0.0377	○
export	出口	0.0364	○
authority	机关	0.0363	○
People	人民	0.0346	○
institution	机构	0.0332	○
provide	提供	0.0321	○
Kong	香港	0.0292	×
order	责令	0.0287	○
imprisonment	徒刑	0.0286	×
formulate	制定	0.0285	○
criminal	刑事	0.0278	○

図 4.10: M_{pro} によって獲得された訳語対の例 (法律)

第5章　結論

5.1　まとめ

本論文は，大規模なパラレルコーパスから対訳辞書を自動的に獲得する新しい手法を提案した．提案手法は，既存の形態素解析ツールの単語分割の誤りに影響されにくいという特徴を持つ．対象とする言語は中国語と英語とし，中英対訳辞書を自動構築した．

まず，中国語と英語のパラレルコーパスを用意する．次に，中国語と英語の文に対して前処理を行う．英語文については lemmatization により各単語を原形に直す．単語境界が明示されていない中国語については 3 種類の前処理，すなわち単語分割，文字 1-gram への分割，文字 2-gram への分割を行い，3 通りのパラレルコーパスを用意する．次に，3 種類の前処理が実施されたパラレルコーパスに対し，GIZA++ を用いて単語のアライメントを行った．中国語の文を単語分割した場合と，文字 1-gram に分割した場合では，1 つの英単語と，それに対応する複数の中国語の単語もしくは文字を連結した文字列を中国語単語として，訳語対の候補を抽出する．中国語の文を文字 2-gram に分割した場合では，複数の文字 2-gram が 1 つの英単語に対応付けられるときには訳語対の候補を抽出しない. したがって, 抽出される訳語対において, 中国語の単語は必ず 2 文字となる．以上により，3 種類の前処理が行われたコーパスから 3 種類の訳語対の候補の集合を得る．最後に，訳語対の候補の中から，複数のパラレルコーパスから得られた訳語対で，かつ抽出される回数が多いものを選択し，最終的な中英対訳辞書を得た．このように，提案手法では, 単語分割，文字 1-gram への分割, 文字 2-gram への分割を前処理として併用することで，3 種類の手法のお互いの誤りを補完し，訳語対獲得の精度を向上させている．また，3 つの手法の併用により，多くの新語や専門用語の対訳対を新たに獲得できると期待される．

評価実験の結果，獲得した訳語対のうちスコアの上位 100 件を評価したところ，その抽出精度は 0.95 であった．これは，単語分割, 文字 1-gram への分割，文字 2-gram への分割を単独で実施したときの結果を上回ることを確認した．また，新しい単語が獲得される割合も，複数の前処理を併用することで向上した．

5.2　今後の課題

今後の課題を以下に述べる．現在，スコアが上位 100 件程度の訳語対しか評価していないため，スコアが下位の訳語対についても提案手法の有効性を検証する必要がある．また，精度だけではなく再現率の評価も必要である．提案手法は文単位の対応関係が付与さ

れたパラレルコーパスを必要とする．そのようなコーパスは，文間の対応を人手で付与する必要があるため，大量のデータを用意しづらい．これに対し，文単位でなく文書単位でアライメントされたパラレルコーパスはコンパラブルコーパスと呼ばれ，比較的容易に大量のデータを用意できる．したがって，コンパラブルコーパスに適用できるように提案手法を拡張することは意義が大きい．さらに，中国語だけではなく，日本語や韓国語のような単語境界が明示されていない他の言語を対象に実験を行い，提案手法が言語に依らず有効であるかを調べたい．

謝辞

本研究に際して，指導教官の白井清昭准教授には，研究に関する様々なご指導を賜りました．白井清昭准教授の下で学んだ2年間を誇りに思っています．心より感謝いたします．研究において貴重な意見を頂いた飯田弘之教授，池田心准教授に感謝致します．また, 本研究に関する多くの有益なご意見，ご助言をいただいた白井研究室に所属する学生の皆様に感謝致します．最後に，お世話になったすべての方々に感謝いたします．

参考文献

[1] Khang Nhut Lam, Feras Al Tarouti, and Jugal Kalita. Phrase translation using a bilingual dictionary and n-gram data: A case study from Vietnamese to English. In *Proceedings of NAACL-HLT*, pages 65–69, 2015.

[2] Liling Tan, Josef van Genabith, and Francis Bond. Passive and pervasive use of a bilingual dictionary in statistical machine translation. In *ACL 2015 Fourth Workshop on Hybrid Approaches to Translation*, 2015.

[3] Pak-kwong Wong and Chorkin Chan. Chinese word segmentation based on maximum matching and word binding force. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 200–203. Association for Computational Linguistics, 1996.

[4] Feng-wen Zhai, Feng-ling He, and Wan-li Zuo. Chinese word segmentation based on dictionary and statistics. *MINIMICRO SYSTEMS-SHENYANG-*, 27(9):1766, 2006.

[5] Jia Xu, Richard Zens, and Hermann Ney. Do we need chinese word segmentation for statistical machine translation? In *Proceedings of the Third SIGHAN Workshop on Chinese Language Learning*, pages 122–128, 2004.

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

[7] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[8] Keiji Yasuda and Eiichiro Sumita. Building a bilingual dictionary from a Japanese-Chinese patent corpus. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 276–284, 2013.

[9] 北村 美穂子 and 松本 裕治. 対訳コーパスを利用した対訳表現の自動抽出. **情報処理学会論文誌**, 38(4):727–736, 1997.

[10] 張 玉潔, 馬 青, and 井佐原 均. 英語を介した日中対訳辞書の自動構築. **自然言語処理**, 12(2):66–85, 2005.

[11] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.