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

WANG Xinzhu　　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{s1510064,kshirai}@jaist.ac.jp

1 はじめに

対訳辞書は機械翻訳を始め様々な多言語処理に必要となる言語資源である．しかし，対訳辞書の整備が進んでいない言語もあれば，特定の分野のテキストを翻訳するためにドメインに特化した対訳辞書が必要となることもあることから，対訳辞書を自動構築する技術が必要とされている．多くの先行研究では，パラレルコーパスから対訳関係にある単語の組を抽出する [1, 5, 6]．この際，日本語や中国語のような単語境界が明示されていない言語については，形態素解析 (単語分割) が前処理として行われる．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高いが，特定のドメインのパラレルコーパスやインターネット上のパラレルコーパスを解析の対象としたときには，単語分割の誤りが多くなることが予想される．

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，パラレルコーパスから機械翻訳のための対訳辞書を構築する手法を提案した [4]．この手法では，既存の単語分割ツールを使わず，中国語の文を文字に分割し，中英パラレルコーパスにおいて同じ英単語に対応する文字を組み合わせて中国語の単語を復元する．この手法は単語分割ツールを使わないためにその誤りの影響を受けないが，一方で中国語の文を文字に分割してから訳語対を獲得するため，文を単語に分割したときよりも獲得された訳語対の誤りが増えると考えられる．

本研究では，単語境界が明示されていない言語を対象とし，単語分割の誤りに頑健な対訳辞書の自動構築手法を提案する [3]．対象とする言語は中国語と英語である．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，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合わせることにより，単語分割ツールの誤りに影響を受けにくく，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．

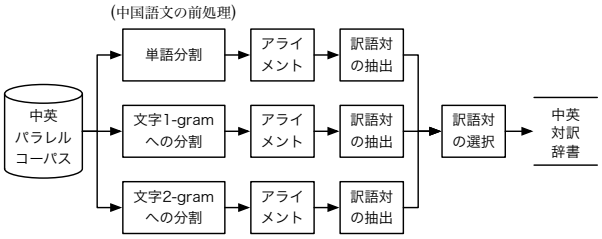

図 1: 提案手法の概要

2 提案手法

2.1 概要

図 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．パラレルコーパスでは文間の対応関係がわかっているものとする．次に，中国語と英語の文に対して前処理を行う．特に中国語については 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，パラレルコーパスにおける単語間のアライメントを自動的に決定し，その結果を基に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し，最終的な中英対訳辞書を得る．

提案手法では，単語分割ツールを使う手法と使わない手法 (文字 1-gram と 2-gram への分割) を併用することで，お互いの誤りを補完し，訳語対獲得の精度の向上を図る．なお，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用いることも検討した．しかし，文字 3-gram については，3 文字で構成される中国語の単語は少ないことから，正しくない訳語対が数多く獲得された．文字 4-gram 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

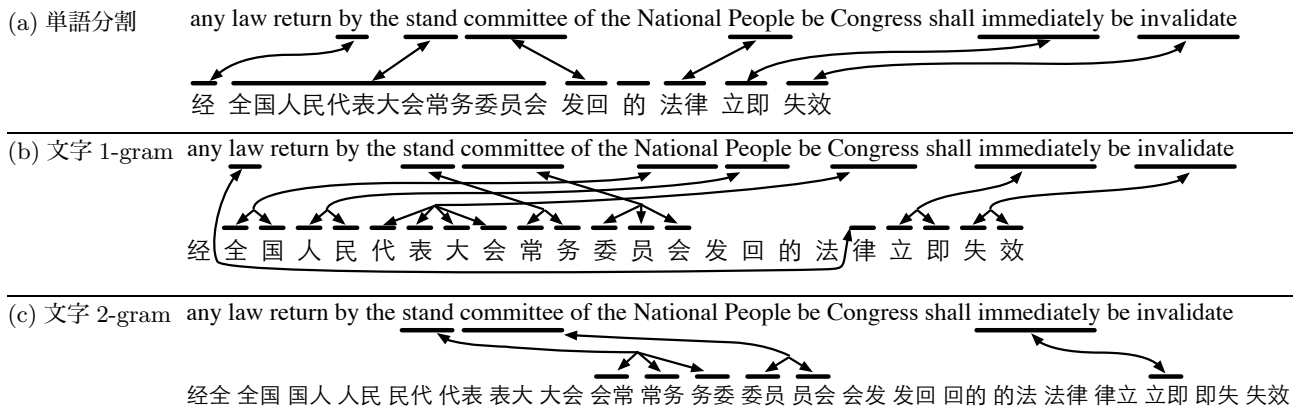

図 2: アライメントの例

文字 3-gram や 4-gram へ分割する前処理は行わないこととした．

2.2 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，lemmatization により各単語を原形に直す．本論文では Stanford CoreNLP[1]を用いた．

中国語文については，単語境界が明示されていないため，これを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[2] を用いた．
2. 文字 1-gram への分割
   中国語の文を文字単位に分割する．
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．例を以下に挙げる[3]．
   文化和社会事务
   → 文化 化和 和社 社会 会事 事务

2.3 アライメント

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自動的に決定する．単語のアライメントとは，ここ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を決定する処理である．ただし，前処理によっては，英単語と中国語の文字 1-gram もしくは 2-gram が対応付けられる．本研究では，GIZA++[2] を用いて単語のアライメントを決定した．アライメントの例を図 2 に示す．同図の (a) は前処理として単語に分割したとき，(b) は文字 1-gram に分割したとき，(c) は文字 2-gram に分割したときのアライメントを示している[4]．

2.4 訳語対の候補の抽出

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 2 (b) 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．
- 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 2 (c) 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

[1] http://stanfordnlp.github.io/CoreNLP/
[2] https://github.com/fxsjy/jieba
[3] この文の日本語訳は「文化・社会問題」である．
[4] 一部の単語のアライメントは省略されている．

対しては複数の文字 bi-gram が対応するため，訳語対を抽出しない．

訳語対の候補を抽出すると同時に，パラレルコーパス全体においてそれが抽出された回数，すなわち訳語対の候補の出現頻度も記録する．

これまでの手続きによって，3 通りの訳語対の候補の集合が得られる．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たパラレルコーパスから獲得された訳語対の候補の集合をそれぞれTP_{seg}，TP_{c1}，TP_{c2} とおく．

2.5 訳語対の選択

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．
- 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ため，それを含む訳語対の候補を除外する．
- 英単語が数字のとき
  後述の実験で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．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．

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 (2) のように定義する．

$$Score(tp) = \max_{x \in \{seg, c1, c2\}} Score_x(tp) \qquad (1)$$

$$Score_x(tp) = \begin{cases} \dfrac{O_x(tp)}{\sum_{tp \in TP_x} O_x(tp)} & \text{if } tp \in TP_x \\ 0 & \text{if } tp \notin TP_x \end{cases} \qquad (2)$$

式 (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 (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，$Score_x(tp)$ は tp の相対出現頻度である．

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．

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．

3 評価実験

3.1 実験手順

訳語対を獲得する中英コーパスとして以下の 2 つを用いた．

- BBC news パラレルコーパス
  ウェブサイト[5]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．
- 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[6]．文の組の総数は 31,517 である．

実験では以下の 4 つの手法を比較した．M_{seg}, M_{c1}, M_{c2} は，それぞれ中国語の文を単語に分割，文字 1-gram に分割，文字 2-gram に分割してから訳語対を獲得する手法である．M_{pro} は提案手法であり，上記 3 つの手法を組み合わせて訳語対を獲得する．

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．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

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

表 1: 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

表 2: 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

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．既存の辞書として，LDC English Chinese bilingual wordlists を用いた．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．

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 R_{new} を式 (3) のように定義する．

$$R_{new} = \frac{\text{出現頻度 5 以上かつ未知の訳語対の数}}{\text{出現頻度 5 以上の訳語対の数}} \quad (3)$$

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．

3.2 実験結果

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 1 に，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 2 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_{1g}, M_{2g}) 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．また，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 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．

4 おわりに

本論文では，単語境界が明示されていない言語を対象に，形態素解析の誤りに頑健な対訳辞書の自動構築手法を提案した．単語分割，文字 1-gram への分割，文字 2-gram への分割といった前処理によって 3 種類のパラレルコーパスを作成し，それぞれから訳語対を獲得した．評価実験の結果，複数の前処理を施したパラレルコーパスを利用することで，訳語対獲得の精度や新しい訳語が獲得される割合が向上したことを確認した．

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．

参考文献

[1] 北村美穂子, 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, Vol. 38, No. 4, pp. 727–736, 1997.

[2] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[3] Xinzhu Wang. 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. 修士論文, 北陸先端科学技術大学院大学, 2017.

[4] 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*, pp. 122–128, 2004.

[5] 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*, pp. 276–284, 2013.

[6] 張玉潔, 馬青, 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, Vol. 12, No. 2, pp. 66–85, 2005.