日本語-ベトナム語翻訳における考察

Nguyen Le An　　松本 啓之亮　　森 直樹
大阪府立大学 電気・情報系専攻 知能情報工学分野
an@ss.cs.osakafu-u.ac.jp

1 はじめに

近年の機械翻訳における研究では. 関連する言語資源の翻訳が研究されているが, データが少ない言語対や語順が異なる言語対に関する研究は極めて少ない. よって, 機械翻訳の質はヨーロッパ言語や英語-中国語などの言語対でのみ向上しており, 言語構造の大きく異なる他の言語間では機械翻訳の精度は高いとはいえない. その中の一つである日本語-ベトナム語の機械翻訳研究は発展途上であり, 翻訳された文が原文の意味とかけ離れてしまうケースが多い. そこで, 日本語-ベトナム語翻訳における質の良い機械翻訳システムを作る必要があると考えられる.

本研究では, より質の良い日本語-ベトナム語翻訳モデルの構築を目的とする. その前段階として, 本稿ではベースラインモデルである Moses 翻訳モデル [1] が日本語-ベトナム語翻訳の場合にどのような結果を出力するかを試す. そして, 構文トランスファ方式をベースとし, 構文解析ツールとして利用されている構文・構造解析システム KNP[2] を用いて, 日本語-ベトナム語の単文翻訳を処理する方法を提案し, 翻訳の結果を考察する.

2 Mosesによる日本語-ベトナム語翻訳学習

日本語-ベトナム語翻訳モデルを構築することは最終的な目標である. しかしながら, いきなり翻訳モデルを構築することは困難であるため, まず Moses 翻訳モデルで日本語-ベトナム語翻訳を検討する. また, 構築する翻訳モデルを評価するために, Moses 翻訳モデルをベースラインシステムとして比較する. よって, まず日本語-ベトナム語で Moses 翻訳モデルを適用した場合, どういう結果が出るかを実験する. 以下に実験の順序を示す.

2.1 訓練データ作成

データ取得

実験の訓練データとして TED Talks と Multilingual Bible Parallel Corpus(MBPC) を利用した. TED Talks から講演の字幕翻訳を日本語 1295 本, ベトナム語 1062 本を取得し, その中から共通な内容の文書を抽出して, ノイズを排除した後, 最終的に対訳文 91583 本が得られた.

また, MBPC から日本語-ベトナム語の対訳文 95836 本を取得した. MBPC は聖書の翻訳から作成された多言語の対訳コーパスであり, XML 形式で約 100 言語が含まれる. これらの対訳文を Moses の訓練データとして利用した.

形態素解析

用意された対訳データを形態素解析する. 出力フォーマットは一文一行の分かち書きされた語句のみである. 日本語は Mecab で, ベトナム語は vnTokenizer で形態素解析した.

Truecasing

Truecasing とはデータのスパース性を減らすために, 大文字小文字の区別をなくす処理である. 前ステップで形態素解析されたデータの単語の大文字と小文字を Truecasing する. 日本語の場合には必要がないが, ベトナム語は大文字と小文字の区別があるため, 必要となる. また, 頻繁に大文字になるものを置換しない.

Cleaning

訓練データの長い文章, 空の文章, 正しく整列されていない文章などを削除する処理である. Cleaning 処理では長文において時間がかかる性質があり, また長文を削除しても大きな影響は生じないため, 50 単語以上の文は削除した.

2.2 言語モデルの構築

言語モデルは出力言語の語順について, そのもっともらしさを確率として付与するモデルで, 出力言語の単言語データから学習する.

言語モデルはIRSTLM, SRILM, RandLM, KenLM, berkeleylm などがあるが, 本実験では KenLM を利用し, ベトナム語言語モデルを作成した. また, 計算速度を向上させるために, この言語モデルをバイナリ形式に変換した.

2.3 翻訳モデルの構築

翻訳モデルは訳語のもっともらしさを規定する統計モデルで, 対訳データから学習する. 実験では GIZA++ モデルを利用する. GIZA++ を適用することにより, フレーズ翻訳テーブルが作られる. GIZA++ とは単語の確率値の計算をするツールである. IBM 翻訳モデルの model1 から model5 に基づいて, 単語の対応関係の確率値を計算する.

図 1 にフレーズ翻訳テーブルの例を示す. この例では, 日本語の "新しい作品" をベトナム語"tac pham moi"に翻訳する確率は p(tac pham moi—新しい作品) = 0.8 である.

```
新しい作品 ||| tác phẩm mới ||| 0.8 ||| |||
続けている ||| đang tiếp tục ||| 0.5 ||| |||
```

図 1: フレーズ翻訳テーブルの例

2.4 Tuning

Tuning とはデータを最もモデルの重みが良くなるように調整する作業である. この重みパラメータを最適化する方法として, 一般的に Minimum Error Rate Training(MERT) が用いられる.

MERT は目的の評価関数を最大にするような翻訳結果を選択するために, 重みパラメータを Tuning する. Development data(dev データ) と呼ばれる試験的な翻訳データを使用し, 各文に対し上位 100 文程度の翻訳候補を出力する. その候補の中で重みを変えることで, より良い翻訳候補が上位にくるようにパラメータを調整する.

Ted Talks から取得したデータに dev データがなく, test データしかなかったため, Ted Talks の test データを半分に分け, 半分を test データ, 半分を dev データとした. これらの dev ファイルと test ファイルも分かち書き形式で形態素解析する.

表 1: Moses 学習結果の BLEU スコア

BLEU = 14.23, 45.4/17.4/9.0/5.7	
BP	1.000
ratio	1.041
hyp_len	10706
ref_len	10286

2.5 Testing

Testing では日本語の test データの一行ずつを入力として翻訳する. 入力された test データはフレーズに分けられ, フレーズ翻訳テーブルを参照し, フレーズベースでベトナム語に置き換えられる. 図 2 に入力データ, 図 3 に翻訳結果を示す.

翻訳結果から何ヶ所か日本語が翻訳されていない単語 (未知語) があるが, これは未知語が traing データに含まれていないからだと考えられる.

2.6 翻訳結果の評価

翻訳結果を BLEU[3] で評価した. 表 1 にその結果を示す. 0 が最低で, 100 が最高の値であるが, 実際の機械翻訳でも 50 を超えるものはほとんどない. また, スラッシュ区切りの 4 つの数値 45.4, 17.4, 9.0 5.7 がそれぞれ 1-gram, 2-gram, 3-gram, 4-gram のものである. hyp_len が出力文の単語の総数, ref_len が正解文の単語の総数, ratio は hyp_len と ref_len の比率である. 表 1 から BLEU スコアが 14.23 でやや低いため, Moses モデルより語順の異なる日本語-ベトナム語翻訳に適する翻訳モデルを構築する必要があると考えられる.

3 単文翻訳

前章の Moses モデルの考察より, 語順が異なる日本語-ベトナム語の翻訳に対しては, なんらかの手法により語順を変更しなければ更なる翻訳精度の向上は見込めないと考えられる. よって, この章では単文を対象とする翻訳処理について述べる.

```
370 初めて DNA の 構造 を 発見 し た 頃 、 コールド スプリング ハーバー で の 初 講義 で 物理 学者 の レオ シ ラード に 「 特許 を 取る つもり は ？ 」 と 聞か れ まし た
371 でも 彼 は 特許 法 を よく 知っ て い て 、 この 発見 で の 特許 取得 は 不可能 だ と 分かっ て い まし た 何 の 役 に も 立ちゃ し ませ ん から
372 少なくとも 特許 権 を 得 た だけ いい です ね 役に立つ こと も あり ます から
373 次に 、 彼ら は コード の 解読 を 研究 し まし た
374 いっぎに バイオテクノロジー 産業 が 出現 し まし た
375 でも 、 まだまだ 道のり は 長かっ た の です 子供 の 頃 に 私 を 捉え て い た 疑問 「 生まれ か 育ち か 」 という 問題 の 答え に たどり着く まで に は
376 この 上 の もの を 見る と 分かる よう に 癌 の ひどい もの は 遺伝子 が 挿入 さ れ たり 削除 さ れ たり し て い ます
377 だから DNA は かなり めちゃくちゃ な の です 反対 に もし 生存 可能 なら ば DNA は そこ まで めちゃくちゃ で は ない の です
```

図 2: 入力データ

```
370 lần đầu_tiên , tôi đã tìm ra cấu_trúc ADN . ハーバー スプリング , trong khi lần đầu_tiên trong bài giảng của một nhà vật_lý_học Leo ơi , ″ ông ? ″ và tôi đã hỏi tôi .
371 Nhưng ông ấy biết rõ luật sáng_chế , đó là chúng_tôi không_thể lấy bằng_sáng_chế đó , tôi đã biết rằng , cũng chẳng có_ích gì .
372 bằng_sáng_chế và bản_quyền , ít_nhất là hữu_ích . từ .
373 Và sau đó , họ là nghiên_cứu của đoạn mã .
374 có một ngành công_nghiệp công_nghệ_sinh_học .
375 Nhưng chúng_ta , từ thuở thơ_ấu của tôi , đó , tôi đã nói , ″ đang được sinh_ra , câu_hỏi là gì ? ″ ″ là một vấn_đề lớn của tôi là , cho_đến khi tôi đi đến câu trả_lời .
376 trên này , và những thứ bạn nhìn thấy những thứ tôi_tệ của bệnh ung_thư , nhưng nó được chèn thêm một gen , và được nó .
377 Vì_thế phân_tử ADN của bạn có khả_năng sống_còn , đó là ADN , nếu bạn đến không phải là nhiều .
```

図 3: 翻訳結果

3.1 提案手法

構文構造が異なる日本語-ベトナム語の翻訳において, 構文解析システム KNP と構文トランスファ方式を用いて, 単文翻訳を処理する方法を提案する. また, その翻訳結果を考察する. KNP は京都大学の黒橋・河原研究室で作られ, 構文解析システムとして公開されている.

図 4 に翻訳処理のフローチャートを示す. まず, 原言語 (日本語) を入力すると, KNP を用いて形態素解析, 構文解析され, 原言語の構文解析結果を得る. この時点でこの構文解析結果は原言語に依存した構文である. これを変換処理によって, 目標言語 (ベトナム語) に依存した構造に変換する. また, 修飾語が単語である場合は主要要素の後ろに配置する (ベトナム語は修飾語が主要要素の後ろに置かれる). そして, 対訳辞書と変換規則 [4] を利用し, これを翻訳する. 辞書には, 日本語-ベトナム語オンライン辞書 Soha を利用する. また, Soha 辞書にない単語を追加する拡張可能な対訳コーパスを構築する. Soha 辞書にはない単語が出現する場合は, 対訳コーパスにアクセスし, 訳語を検索する.

3.2 翻訳結果

表 2 に 50 単文を入力として, 自動評価尺度 BLEU で単文翻訳の翻訳結果を評価した結果を示す. 表 2 より, 翻訳結果の BLEU スコアは 32.59 でやや高い. ただし, 実験結果は単文の場合のみなので, 重文や複文の場合も考える必要がある.

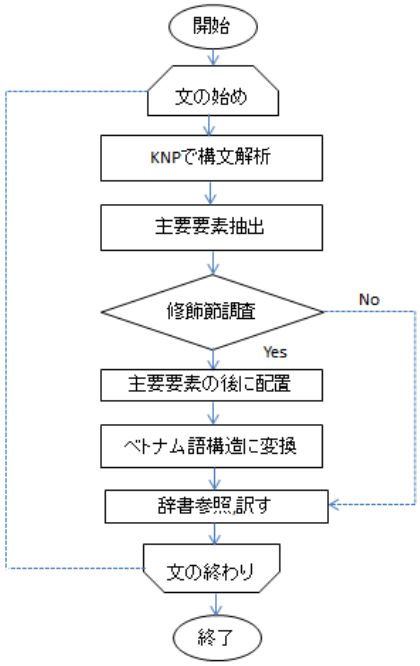

図 4: 処理のフローチャート

4 コーパスへの新語追加

Soha 辞書には日常で用いる言葉はほとんど含まれないため, この章では Soha 辞書にはない単語を構築されているコーパスに追加する方法を説明する. 近年, インターネットの普及などにより多くの新語が現れており, 特に Twitter でよく使われている. 新聞記事における言葉は正式な日本語であり, 新語は新聞記事ではあまり使用されないという特徴に基づいて, 新聞記事に現れる単語の頻度を計算することで, 新語を探しコーパスに追加した. 新聞つんどく及び Twitter4j を利用して, 新語を収集した. 新聞つんどくはネットで

表 2: 単文翻訳結果の BLEU スコア

BLEU = 32.59, 62.4/52.5/44.6/38.9	
BP	0.667
ratio	0.712
hyp_len	351
ref_len	493

公開されている読売, 朝日, 毎日, 日経, 産経の五大新聞の記事を自動的に収集し, データベースを構築するソフトであり, Twitter4j は Twitter のタイムラインからつぶやきを取得する API である.

まず, Twitter4j でツイートを, 新聞つんどくで約 3400 記事を集め, 新聞記事で出現するツイート単語の出現率を計算する. そして, 出現率が低いほど新語候補として選り分ける. 以下にその処理を示す.

1. いくつかのユーザのタイムラインからツイートを取得し, Sen で形態素解析する. また, ユーザのタイムラインを繰り返し更新することで, 新しいツイートを取得していく.

2. 新聞つんどくから約 3400 記事を取得し, Sen で形態素解析する.

3. 新聞記事に現れるツイート単語の出現率を計算する. N_{app} を新聞記事に出現するツイートの単語の出現回数, N_{all} を新聞記事の文字列の総数とすると, 出現率は以下の式で表される. また, 無駄な計算を避けるため, 記号や文字列などは除去した.

$$F_{\mathrm{n}} = \frac{N_{\mathrm{app}}}{N_{\mathrm{all}}} \tag{1}$$

4. 出現率の低い単語を抽出する. 任意の α より小さい出現率の単語すべてを新語の候補として選択する. 次に, これらの新語候補について Soha 辞書を参照し, ベトナム語の訳語を検索する. 候補に対応する訳語があれば候補は新語ではなく, Soha 辞書になければ候補は新語としてコーパスに登録する.

   実験では α を 0.05‰と設定し, ノイズを除いた後, 約 3128 語の新語の候補を取得した. そのうち, 0.05‰ より小さい出現率の単語は 1110 語であり, 35% であった. これらの候補単語を Soha 辞書と参照し, Soha 辞書にはない単語だけをコーパスに登録した. 図 5 に新語とそのベトナム語の例を示す.

$$0 < F_{\mathrm{n}} < \alpha \tag{2}$$

Japanese	Pronounce	Vietnamese
懐メロ	Natsumero	bài hát xưa
なう	Nau	bây giờ
コンティニュー	Continyu	tiếp tục
エステ	Esute	chăm sóc sắc đẹp
しょんどい	Shondoi	mệt thiệt
モテ	Mote	đào hoa
ヒトカラ	Hitokara	đi karaoke 1 mình
へー	He	vậy hả
チャリ	Chari	xe đạp
ワロタ	Warota	cười
ワクテカ	Wakuteka	mong ngóng
クレヒス	Kurehisu	thông tin thẻ ngân hàng
ハッキリ	Hakkiri	rõ ràng

図 5: 新語とそのベトナム語訳の例

5 まとめと今後の課題

本稿では翻訳モデル用の対訳コーパスを作成し, Moses 翻訳モデルによる日本語-ベトナム語翻訳の学習について説明した. また, 構文解析 KNP を用いた日本語-ベトナム語単文翻訳処理について説明した.

日本語-ベトナム語翻訳の精度を上げるために今後の課題として, 以下のことが挙げられる.

- 構文構造の変換に基づく翻訳モデルを構築する.
- 翻訳モデルに利用するため, より数の多い対訳コーパスを作成する.
- BLEU 以外の自動評価尺度を試す.

参考文献

[1] Moses Manual and Code Guide : http://www.statmt.org/moses/manual/manual.pdf

[2] 日本語構文・格解析システム KNP : http://nlp.ist.i.kyoto-u.ac.jp/index.php

[3] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu: Bleu: a Method for Automatic Evaluation of Machine Translation. *IBM Research Report RC22176*, W0109-022, 2001

[4] An Nguyen Le, Keinosuke Matsumoto, Naoki Mori: Japanese-Vietnamese Translation for Simple Sentences and Bilingual Corpus Improvement. *Proceedings of the Third Asian Conference on Information Systems ACIS 2014*, December, 1-3, 2012