# 日本語ーウイグル語ハイブリッド機械翻訳

PAERHATI ABUDUKADEER　MAIMITILI NIMATE　松尾 啓志　山本いずみ
名古屋工業大学創成シミュレーション工学専攻
parhat@matlab.nitech.ac.jp

## 1　初めに

近年英語、日本語、中国などのたくさん言語の間で機械翻訳が盛んに行われるとともに、種々の翻訳手法が開発された。コンピュータの発展、計算能力の上昇に伴い、機械翻訳でも数学モデル (特に統計的モデル) を用いて新たなシステム構築などの研究も行われていて、翻訳の性質もかなり上昇している。しかし、ウイグル語に関しては機械翻訳の研究の歴史が浅いということで、他言語に比べると翻訳に用いるコーパスなどの資源が少ないのが現状である。日本語とウイグル語は共に膠着言語に属し、文法構造が SOV 形である。表 1 に、日本語とウイグル語の相違点を示す。日本語ーウイグル語ルールベース機械翻訳で現在は各接辞が接合した時に母音と子音の変化の問題が生じ、人称語尾の対応も複雑である。それらの問題を統計翻訳で解決することができる。しかし、統計翻訳に必要な対訳コーパスが現状では容易には入手できないため、我々がそれらの問題を解決するために Mecab[1] を拡張したルールベース機械翻訳を提案し、そのルールベース機械翻訳と統計機械翻訳を組み合わせたハイブリッド機械翻訳の実装と実験を進めている。Mecab を拡張したルールベース機械翻訳に必要な対訳辞書とパターン辞書も同様に独自に実装した。統計翻訳に必要な対訳コーパスを約 1500 文作成し、言語モデルに関しては、約 6000 文のウイグル語文を作成し、実験を行った。以下のセクション 2 でルールベース翻訳と統計翻訳について簡単説明し、セクション 3 で日本語ーウイグル語ハイブリッド機械翻訳について提案手法を説明し、セクション 4 で実装と実験、そのあとセクション 5 でまとめと今後の課題について説明する。

表 1: 日本語とウイグル語の文法の相違点

| | 日本語 | ウイグル語 |
|---|---|---|
| SOV | ○ | ○ |
| 膠着言語 | ○ | ○ |
| 動詞の活用 | ○ | × |
| 人称語尾 | × | ○ |
| 母音と子音の変化 | △ | ○ |

## 2　ルールベース機械翻訳と統計機械翻訳

機械翻訳はルールベース機械翻訳と統計機械翻訳という大きい二つに分類される。

### 2.1　ルールベース機械翻訳

ルールベース機械翻訳 (RuleBase Machine Translation) は言語の文法関係を解析し、モデル化して、パターンを最初から作成しておき、そのパターンに従って、目的

の言語を生成する形で翻訳を行う手法である。この手法では、両言語の文法関係を十分にしることが必須となる。特に日本語と英語ののように、文法規則が異なる言語の間ではパターンを決めるのがもっと複雑で、システムを構築するのに、多数の言語スペシャリストによる長時間の作業が必要となる。当然、良質な翻訳規則が作成されれば、翻訳の質は高くなる。一方、汎用性が低いという問題もある。ルールベース翻訳方式を分類すると、直接変換、トランスファー方式、中間言語方式の三つがある。

### 2.2 統計機械翻訳

統計翻訳 (Statistical Machine Translation) は 1990 年代前半に IBM 研究所から提案されたもので、対訳コーパスを学習し、言語の間で翻訳モデルを自動的に生成する。対訳コーパスさえあれば、翻訳が可能となる。統計翻訳のメリットとしては、ルールベース翻訳に比べて、翻訳システム構築にかかる時間、作業が少ないこと、言語専門家を必要としないこと、汎用性の高いことである。しかし、問題となるのは、対訳コーパスである。対訳コーパスはすべての言語に対して十分に整備されている訳ではない。得に、英語など利用者が多い言語との間には、対訳コーパスは存在するが、ウイグル語に代表される利用者の少ない言語との間の対訳コーパスは十分に整備されていないのが現状である。コーパスの量が少ないと翻訳精度が低くなる。統計翻訳は単語に基づく翻訳モデルと句に基づく翻訳モデルに分類される。今は句に基づく翻訳モデルが研究の主流となっている。単語に基づく翻訳モデル対して、翻訳精度が高いということが主な理由である。

#### 2.2.1 基本概念

日本語の単語列 $\boldsymbol{j}$ が与えられた時、それた対する全ての組み合わせから、確率が最大になるウイグル語の単語列 $\hat{\boldsymbol{u}}$ を検索することで、翻訳を行う。統計翻訳は雑音のある通信路モデル (noizy channel model) によって表されている。これを Peter[2] らは提案し、以下がその基本式である。

$$\hat{\boldsymbol{u}} = \boldsymbol{argmax_u P(u|j)} \tag{1}$$

ベイズ定理に基づき式 (1) を以下のように変化することができる。

$$\boldsymbol{P(u|j)} = \frac{\boldsymbol{P(u,j)}}{\boldsymbol{P(j)}} = \frac{\boldsymbol{P(j|u)P(u)}}{\boldsymbol{P(j)}} \tag{2}$$

分母は $\boldsymbol{u}$ と独立していることから、求める $\hat{\boldsymbol{u}}$ は最大になる $\boldsymbol{u}$ を決定すると同じことで、$\boldsymbol{argmax_u P(j|u)P(u)}$ を求めればいい。結果的に $\hat{\boldsymbol{u}}$ は式 (3) により得ることができる。

$$\hat{\boldsymbol{u}} = \boldsymbol{argmax_u P(u|j)} \simeq \boldsymbol{argmax_u P(j|u)P(u)} \tag{3}$$

統計機械翻訳モデルは翻訳モデル、言語モデル、翻訳確率最大となる文を検索するデコーダから成り立っている。翻訳モデルは日本語とウイグル語の対訳コーパスから学習して作成される。言語モデルを目的言語であるウイグル語のコーパスから学習して作成される。デコーダは翻訳モデルと言語モデルを用いて、尤度の最も高いウイグル語文を生成する。

$\boldsymbol{P(j|u)}$ は翻訳モデル、$\boldsymbol{P(u)}$ は言語モデルという、実は [2] らはフランス語と英語の間の翻訳がベースになっているので、基本式では $\boldsymbol{P(e), P(f|e)}$ で表現している。我々は日本語とウイグル語の統計翻訳の研究をしていることから、式を $\boldsymbol{P(u), P(u|j)}$ で表現する。

## 3 日本語ーウイグル語ハイブリッド機械翻訳について

最近統計機械翻訳とルールベース機械翻訳を組み合わせた機械翻訳の研究も増えている。それぞれのメリットを活

かして、翻訳精度を向上させるという目的がある。このような機械翻訳手法をハイブリッド機械翻訳と言う。日本語とウイグル語の場合では、直接ルールベースに従った機械翻訳のシステムを構築した時、助詞と接辞の問題が複雑で、そのたびにルールを作るのも面度な作業となる。一方、統計翻訳でシステムを構築する場合、対訳コーパスの大きさが不十分であるという問題点がある。それらの弱点を二つの翻訳を組み合わせることにより、克服できることが期待される。日本語形態素解析があれば、日本語の文の処理が簡単になり、その解析で得られた各単語の品詞情報を基にして、辞書を引くことで簡単な日ーウ機械翻訳ができる。トランスファー方式で使う構文解析を用いる必要がなくなる。なぜなら、日本語とウイグル語の文法構造が共に ***SOV*** 形式であるから、形態素解析で出力され単語列に対して位置置換する必要はない。そこで本研究でも、まず最初に、ルールベースシステムの作成を行う。ここで日本語形態素解析 Mecab の Java version を拡張してルールベースシステムを作る。次に自分で作った対訳コーパスを用いて、統計翻訳を行う。最後に両方の翻訳結果を評価し、評価が高い文を出力する。以下に示す手順で行う。以下の図 1 に示す。

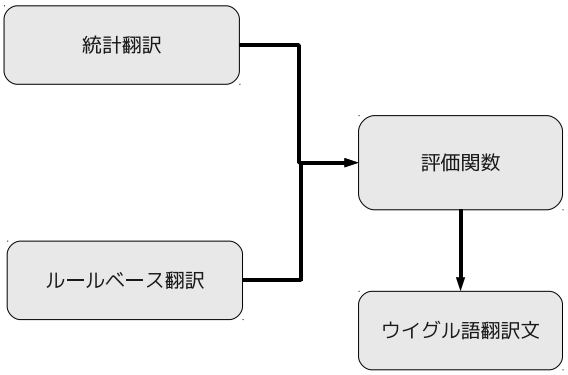


図 1: 日本語ーウイグル語ハイブリッド機械翻訳システム

- 日本語を分かち書けし、それらをリストに登録する。
- リストに登録された単語から対訳辞書を引いて、対訳リストを作る。
- 対訳リストからウイグル語文を生成する。
- パターン学習辞書からパターン解析し、文整形を行う。
- 統計翻訳によりウイグル語単語列を生成する。
- 二つのウイグル語単語列に対して評価を行う。
- 評価値が高い単語列を出力する。

## 3.1 Mecab を拡張した日ーウルールベース機械翻訳

日本語では単語と単語の間に空白がないため、機械翻訳で最初は日本語の分かち書けの処理が必要になる。小川 [3][4] らは、派生文法に基づいた独自の形態素解析を提案し、日本語とウイグル語の機械翻訳を進めている。我々は日本語の形態素解析に対して Mecab を用いた。Mecab は IPA 辞書と Juman 辞書を使う。

### 3.1.1 Mecab 出力フォーマットの設定

Mecab の出力フォーマットを自由に設定することができるため、本研究で解析結果を表層形、品詞、品詞細分類 1 だけを出力することにする。出力フォーマットは`-F\\n%m,%f[0],%f[1]\n` で設定した。この出力フォーマットで、表層形、品詞、品詞細分類 1 だけの情報を出力することが可能になる。

表 2: 日本語形態素出力情報

| TargetNode | TargetPart | TargetPartOf |
|---|---|---|
| 鳥 | 名詞 | 一般 |
| は | 助詞 | 係助詞 |
| 遠い | 形容詞 | 自立 |
| 所 | 名詞 | 非自立 |
| から | 助詞 | 格助詞 |
| 飛ん | 動詞 | 自立 |
| で | 助詞 | 接続助詞 |
| 来 | 動詞 | 非自立 |
| まし | 助動詞 | |
| た | 助動詞 | |

#### 3.1.2 対訳辞書から訳語を決定

- 日ーウ対訳辞書の作成：IPA 辞書に基づいて作成する。対訳辞書を IPA 辞書の CSV ファイルのように品詞ごとに別々のファイルにする。
- 訳語の生成: 検索される日本語の単語が対訳辞書から検索し、その単語の訳語と品詞を ***Hash*** に格納する。
- ***Hash*** から最初の ***Key*** から最後の ***Key*** まで出力する。ここまでの作業で、Replacement Translation が実現する。

表 3: 日ーウ対訳辞書ー動詞格納ファイル例

| | | | |
|---|---|---|---|
| 言う | VE | deyish | VE |
| 言わ | VE | di | VE |
| 言お | VE | di | VE |
| 言い | VE | di | VE |
| 言っ | VE | de | VE |
| 言え | VE | de | VE |
| 言え | VE | de | VE |

{ 鳥は飛んで来ました } という日本語の文に対して、表 4 で示したように、対訳が各辞書ファイルに格納されていることを分かる。この段階で最終的に得られる翻訳は簡単な

表 4: 日ーウ対訳辞書処理実例

| 日本語 | ウイグル語 | File.csv |
|---|---|---|
| 鳥 | qush | NU |
| は | ø | POPC |
| 飛ん | uchu | VE |
| で | p | COPOP |
| 来 | kel | VE |
| まし | ø | AUXVE |
| た | di | AUXVE |

置換翻訳で、以下のな訳語ができる。

qush ø uchu p kel ø di

この例で生成されたウイグル語単語列はまだ整形されていないことと、接辞の対応関係も良くないことが分かる。その問題を解決するために、ルールベースエンジンの作成を行う。

#### 3.1.3 ルールベースエンジンの作成

ルールベースエンジンの役割を以下に示す。

- 単語の前後関係から、接辞が接合する語幹を決める。
- ウイグル語での母音の弱化、脱落、差入などへの問題の対応
- 生成されたウイグル語の文に対して、人称語尾を正しく決める

qush ø | uchu p | kel ø di

最終結果

qush　uchup　keldi

この例で記号øは脱落し、uchu　p | kel　di らがお互いに接合することになった。

## 3.2　日ーウ統計翻訳実験

### 3.2.1　学習データの準備

日ーウ統計翻訳行う前、必ず対訳コーパスの準備を行う必要がある。

現在日本語とウイグル語の間に実験に使う対訳コーパスがないので、まず小規模な実験を行うために、最小限の対訳コーパスを自作した。日本語 1582 文を翻訳し、学習データとして扱った。対訳コーパスの一部を表 5 で示す。なお、日本語の文の間に空白がないため、最初は Mecab を用いて形態素単位で分割した。ウイグル語の場合単語間に空白があるので、形態素解析する必要がない。しかし、ウイグル語も膠着言語なので、単語に接辞が接合する場合がほとんどで、実験の結果から見ても本来日本語の翻訳されるはずの接辞がウイグル語に対応がない問題が多数発生した。

### 3.2.2　言語モデルの作成

言語モデルを *N*-gram モデルを用いて作成した。*N*-gram モデルの学習には SRILM [5] を用いる。日ーウ統計翻訳で言語モデルを作成する際に *N*-gram-count を 5 で設定する。ウイグル語言語モデルを作成した際に用いた文は 6063 文である。

### 3.2.3　翻訳モデルの作成

本研究で句に基づく翻訳モデル用いることで、最初は翻訳モデルを管理するフレーズテーブル (phrase table) を作成する。単語のアライメント (alignment) の計算には IBM モデル-4 を用いたシール GIZA++を用いる。GIZA++は学習データを双方向に対して、単語アライメントの計算を行う。”gorw-diag-final-and”で生じた単語列のアライメント対応関係表を表 7 で示す。　次に単語列アライメントから日本語単語列とウイグル語単語列のフレーズ対を得る。フレーズテーブルの作成には train-model.perl[6] を用いた。そのフレーズ対に対して翻訳確率を計算してフレーズテーブルを作成した。

表 6: *N*-gram で生じるウイグル語単語列

| *N*-gram *N* | count |
|---|---|
| *N*-gram 1 | 99677 |
| *N*-gram 2 | 481301 |
| *N*-gram 3 | 54033 |
| *N*-gram 4 | 33200 |
| *N*-gram 5 | 26425 |

### 3.2.4　デコーダの設定

デコーダは moses[6] を用いた。翻訳モデルの各パラメータの設定に関しては今回の実験で学習データとした日ーウ対訳文が小規模であるため、翻訳モデルの重みを 4 で設定した。対訳データの量が比較的少ないと言うことで言語モデルの重みを 3 に設定した。ほかのパラメータは大体 defult 値で設定した。

### 3.2.5　実験評価

通常実験の評価をコンピュータによる自動評価と人手による評価で行う。自動評価方法の代表的なものとして、BELU と METEOR が挙げられる。

表 5: 対訳コーパス例

| |
|---|
| 教室 は 知識 を 与える 。<br>sinip bilim beridu. |
| 知識 を 増やす の を 目標 に する 。<br>bilim ni kupeytishni nishan qilghan bolidu. |
| せっかく 与え た もの を 片端 から 、 捨て て しまっ て は 困る 。<br>ming teslikte berghen nersini ishletmestinla tashliwetsek yahshi emes. |
| 良く 覚え て おけ 。<br>isingde ching saqla. |
| 覚え て いる か どう か 、 ときどき 試験 を し て 調べる 。<br>este tutqan tutmighanliqni,daim imtahan elip sinap turidu. |
| 覚え て い なけれ ば 減点 し て 警告 する 。<br>este saqlimisaq numur tartip agahlanduridu. |

表 7: grow-diag-final-and の例

| | nimishqa | shundaq | uylaysen | dep | soridim | . |
|---|---|---|---|---|---|---|
| どうして | ⋆ | | | | | |
| そう | | | ⋆ | | | |
| 思う | | | ⋆ | | | |
| か | | | ⋆ | | | |
| 、 | | | | | | |
| 聞き | | | | | ⋆ | |
| まし | | | | | ⋆ | |
| た | | | | | ⋆ | |
| 。 | | | | | | ⋆ |

# 4 提案システム実装と実験結果

## 4.1 学習データのまとめ

表 8: 学習データのまとめ

| 種類 | 文 | 単語 |
|---|---|---|
| 統計翻訳モデル学習データ | 1582 | 23097 |
| 統計翻訳言語モデル学習データ | 6063 | 755441 |
| ルールベース翻訳対訳データ | | 4523 |

### 4.1.1 実装と実験結果

テストデータとして日本語 50 文を統計翻訳と Mecab を拡張したルールベース翻訳で翻訳した結果を表 9 で示す。

表 9: 実験結果

| | | |
|---|---|---|
| Baseline translation | 7(50) | 14% |
| Replacement translation | 19(50) | 38% |
| Rule base translation | 22(50) | 44% |

#### 4.1.2　ベスト翻訳を決める

本研究の目的である二つの翻訳システムからベスト翻訳文を決めることについて今回は、人手による評価を行った。本来は METEOR と BLEU を用いることがが望ましいが、これらによる評価は今後の課題とする。

## 5　まとめと今後の課題

今回日本語形態素解析 Mecab を拡張して、日ーウルールベース機械翻訳システムを作ることと日ーウ統計機械翻訳実験をし、二つシステムで得られた訳文から一番正しい文を決めることを試みた。ルールベース機械翻訳に対して、助詞と接辞の役割を決めるパターンをそのたびに作成することが困難なため、ウイグル語文生成にかかる部分のみパターンを作成した。一方、統計機械翻訳に関して対訳コーパスの量が不十分であるため、翻訳精度がとても低いという結果になった。しかし、統計機械翻訳でルールベース機械翻訳のように助詞と接辞の役割を決める問題は少ないことを本実験で確認した。翻訳モデルを作成した時に、日本語の学習文に対して形態素ごとに分割した。一方、ウイグル語の学習文に対して空白ごとに分割していたので、単語アライメントを計算した時に、助詞と接辞の多少外れがあることを確認した。それらの問題を解決するためには今後ウイグル語の文に対して形態素ごとに分割するか、もしくは日本語に対して空白ごとに分割するなどの対策が必要となる。さらにパターンの種類を引き続き増加することと対訳コーパスの拡張する課題もある。

## 参考文献


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

[2] Peter F.Brown, John Cocke, Stephen A.Della Pietra, Vincent J. Della Pietra, Frederick Jelink, John D. Lafferty, Robert L.Mercer, and Paul S.Roossin (1990). ”A Statistical Approach to Machine Translation.” ***Computational Linguistics***, 16(2), pp.7985

[3] 小川泰弘、ムフタル・マフスット、杉野花津江、外山勝彦、稲垣康善. ”派生文法に基づく日本語動詞句のウイグル語への翻訳 ”, 自然言語処理, Vol.7, No.3, pp.57-77, Jul. (2000).

[4] 小川泰弘、ムフタル・マフスット、外山勝彦、稲垣康善 (1999).”派生文法による日本語形態素解析.” 情報処理学会論文誌,40(3),1080-1090.

[5] http://www.speech.sri.com/projects/srilm/

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

表 10: 翻訳実例

| |
|---|
| 日本語　明日雨が降るそうです。 |
| Baseline translation　ete yamghur kup 降る shundaq .　× |
| Replacement translation　ata yamghur * yaghi dighandek tur * .　△ |
| Rulebase translation　ata yamghur yaghidighandek turdu.　◯ |
| 日本語　広島でおいしいお土産を買いましたから、来週持って行きます。 |
| Baseline translation　hiroshimadin setiwalghan sowghatni keler hapte elip barimen .　◯ |
| Replacement translation　hiroshima de temlik * sowghat ni setiwal * di din , kelar hepte eliwal p bar du .　△ |
| Rulebase translation　hiroshimade temlik sowghatni setiwaldidin , kelar hepte eliwalp bardu .　△ |
| 日本語　彼はどこから来た？ |
| Baseline translation　u yaqqa yerlerde yaghuz　× |
| Replacement translation　u * qeyer din kel di ?　△ |
| Rulebase translation　u qeyerdin keldi ?　◯ |
| 日本語　明日家へ帰らなければなりません。 |
| Baseline translation　ete oyi barsa 帰ら qilmisa bolmaydu .　× |
| Replacement translation　ata uy gha qayt mi sa boli may du .　△ |
| Rulebase translation　ata uygha qaytmisa bolimaydu .　△ |
| 日本語　北京を訪れ帰郷した時のことである。 |
| Baseline translation　beijinggha qilinghan seperdin qaytip keliwatqanda bolghan ish ibaret .　△ |
| Replacement translation　beijing ni kurup kel yurtqa qaytqan * di chag ning ish * bar .　△ |
| Rulebase translation　beijingni kurup kel yurtqa qaytqandi chagning ish bar .　△ |
| 日本語　来週私は弟の結婚式に参加します。 |
| Baseline translation　sowghatni keler hapte men 弟 ning toy bilen 参加 qilip qildighan .　× |
| Replacement translation　kelar hepte men * ini ning toy murasimi gha qatinishi * du .　△ |
| Rulebase translation　kelar hepte men inining toy murasimigha qatinishidu .　△ |