修士論文

形態素解析を用いた日本語・ウイグル語機械翻訳システムの開発および統計機械翻訳手法の基礎検討

Development of Japanese-Uyghur machine translation system using of Morphological analysis and basic study of Japanese-Uyghur statistical machine translation method

指導教員　　松尾 啓志 教授
津邑 公暁 准教授

名古屋工業大学大学院工学研究科
修士課程創成シミュレーション工学専攻
学生番号　20413579
PAERHATI ABUDUKADEER
松尾, 津邑研究室

2012 年 2 月 3 日

Introduction

In recent years machine translation has been actively carried out among the major languages such as English,Japanese and Chinese. Translation methods have also been developed in various forms with the development of the computer and incresing computational power, mathematical models in machine translation(statistical machine translation in particular)have also been studied for the construction of new systems for the translation and the accuracy of machine translation has increased considerably. however, in that short history of research about Uyghur machine translation , we have confirmed there and few resources and research results when compared to other languages.Both the Japanese and Uyghur are agglutinative languages, and their grammar structure is SOV.In Tabel1, we have summarised the similarities and differences between Uyghur and Japanese. we have been researching Japanese-Uyghur machine translation, promoting the implementation and experiments of hybird machine translation, which combinates rule-based machine translation which is an extension of Mecab for Japanese-Uyghur rule-based machine translation. we experimented by creating 2500 sentences for the required bilingual corpus of statistical machine translation and 7400 Uyghur sentences for a language model, In the following section 2 we briefly explain rule-based machine translation and statistical machine translation. After that, we describe our proposed method for Japanese-Uyghur hybrid machine translation. Then, we discuss our implementation, experiment,conclusions and future challenges.

目 次

表 目 次

図 目 次

1　序論

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[9] を用いたルールベース機械翻訳を提案し、ルールベース機械翻訳と統計機械翻訳の組み合わせを前提とする日ーウハイブリッド機械翻訳の実現に向いて、それらの実装と実験を進めている。Mecab を用い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。対訳辞書は約 5000 単語単位にした。統計翻訳に必要な対訳コーパスを約 2500 文作成し、言語モデルに関しては、約 6500 文のウイグル語文を作成し、実験を行った。

本論文の構成は以下の通りである、第二章では日本語とウイグル語に関して言語学上での関係について述べる。第三章では、機械翻訳システムについてルールベース翻訳と統計翻訳を取り上げて説明する。第四章では日本語ーウイグル語機械翻訳に関連した研究を紹介し、それまでの業績を少し検討する。第五章では日ーウハイブリット翻訳モデルを構築する手法について述べてた後、自分の提案手法を説明する。第六章では提案モデルの構築と実装手順を示す。第七章では本研究のまとめと今後の取り込むべき課題について検討する。

2　日本語とウイグル語の関係

本章では日本語とウイグル語の言語学上での関係を説明し、両言語に関して文法の構造の比較をする。

2.1　ウイグル語と日本語の関係

日本語は主に日本で使用されて、言語類型論上は、語順の点で SOV 形の言語に、形態論の点では膠着語に分類される。[1]

一方ウイグル語は主に中国新疆ウイグル自治区で住むウイグル人が使う言語で、テュルク諸語のチャガタイ語群に属する。言語類型論上は、語順の点で SOV 形の言語に、形態論

の点では膠着語に分類される。[2]

膠着語に分類された言語は、ある単語に接頭辞や接尾辞を付け加えることで、その単語

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	△	○

の文の中での文法関係を示す特徴を持つ。膠着語に分類される言語は、トルコ語、ウイグル語、ウズベク語、カザフ語等のテュルク諸語、日本語、朝鮮語、満州語、モンゴル語、フィンランド語、ハンガリー語、タミル語、エラム語、シュメール語などである。[3]
両言語ともに SOV 形の言語で、膠着語に分類されることから、文は「主語、修飾語、述語」の手順で形成され、語幹に接尾を付け加えることによって文全体の意味が変わって来る点ではよく似ている。本論文では日本語とウイグル語について、動詞の語形変化、形容詞の語形変化、膠着語の役割を果たす助詞と助動詞の関係、人称代名詞による文の変化、特にウイグル語人称語尾について詳しく説明し、機械翻訳におけるそれらの処理を述べる。

2.1.1 動詞の語形変化

動詞は名詞と並んで大体全ての自然言語が持つとされる基本的な品詞である。主に動作や状態や変化などを表す。ここで機械翻訳で動詞の変化をよく把握しないと翻訳精度が変わってしまうという点から、日本語とウイグル語の動詞の相違点を比較する。

日本語の動詞を形態により 3 種類に分ける。五段動詞、一段動詞、不規則動詞である。活用の形態により、五段活用、上一段活用、下一段活用、カ行変格活用、サ行変格活用に分類される。

表 2 で示したように、各形に対して動詞が変化し、それらに独自の接尾辞が接続される。

ウイグル語も膠着言語に分類される一方、日本語と違って動詞の活用の概念を用いていない。語の形成規則として派生文法を用いている。派生文法とは音韻規則に基づいて語幹に接辞をつけることによって新しい語が形成する方法である。ウイグル語では動詞の形成構造は { 動詞語幹+派生語尾+助動詞+人称助詞—格助詞+疑問を表す接尾 } である。動詞の第二語幹を派生する形は表 3 のようである。

表 2: 動詞の変化例

種類	基本型	未然型	未然ウ型	連用型	連用タ型	仮定型	命令型
五段	書く	書か	書こ	書き	書い	書け	書け
一段	食べる	食べ	食べよ	食べ		食べれ	食べよ
不規則	来る	来	来よ	来		来れ	来い

表 3: 動詞の形成規則

ウ動詞	形成	日訳	ウ動詞	形成	日訳
ugen	基本形	学ぶ +	ugen + ghuche	連用比較形	学ん + で
ugen + di	完了形	学び + た	ugen + mas	中止未完了形	学ば + ない
ugen + ma	否定形	学ば + ない	ugen + ele	可能形	学ぶ +
ugen + sa	条件形	学べ + ば	ugen + sun	三人称命令形	学ん + で
ugen + ay	意志形	学び + たい	ugen + iwal	連用状態形	学び + させ
ugen + gen	連体完了形	学び + た	ugen + ip	連体中止形	学ん + で
ugen + dur	使役形	学び + させ	ugen +il	受身形	学ば + れる
ugen + ghin	二人称願望形	学ん + で	ugen + ish	共同形	学び +
ugen + ghech	方向、理由形	学び + ながら	ugen + iwat	連体未完了	学ん + で
ugen + ghili	連用目的形	学ぶ +	ugen + ing	二人称命令形	学ん + で

2.1.2 形容詞の語形変化

形容詞は品詞の一種類で、日本語の場合形容詞と形容動詞がある。ウイグル語は形容詞だけである。形容詞も動詞と同じ活用をする。IPA 辞書 [4] でも一つの形容詞に対してすべての活用型を登録している。以下の表 4 では日本語形容詞活用型とウイグル訳の例を示す

表 4: 形容詞の変化例

日本語	ウイグル訳	活用型
赤い	qizil	基本形
赤し	qizil	文語基本形
赤から	qizil	未然ヌ接続
赤かろ	qizil	未然ウ接続
赤かっ	qizil	連用タ接続
赤く	qizil	連用テ接続
赤くっ	qizil	連用テ接続
赤けれ	qizil	仮定形
赤かれ	qizil	命令ｅ

2.1.3 助詞と接辞の対応関係

日本語とウイグル語が共に膠着言語に属するので、接辞と助詞の変化がとても重要である。対訳辞書作りにしても、統計翻訳実験でも助詞の動きが翻訳精度にかかる。助詞が語幹に接続し、その分を完全な文に変える。対訳辞書を作る際に接辞の訳が難題になる。IPA 辞書では助詞と接辞を別々のファイルにしてるので、ウイグル語の対訳もその形にする。

2.1.4 人称代名詞と人称語尾について

日本語とウイグル語の翻訳では人称代名詞とその語尾の対応関係が非常に重要である。日本語文が人称代名詞によって、品詞接続してくる接辞が変わらないに対して、ウイグル語では変わってしまう。一般のルールベース翻訳でもこれらの問題に対して独自のルールを作って解決する。以下の文ではその例を示している。

私 は 昨日 東京 から 来 まし た 。
men ø tunughun tokyo din kelø di m(一人称語尾, 単) .
あなた は 昨日 東京 から 来 まし た 。
sen ø tunughun tokyo din kel ø di ng(二人称語尾、単) .
彼 は 昨日 東京 から 来 まし た 。
u ø tunughun tokyo din kel ø di ø(三人称語尾、単) .
私たち は 昨日 東京 から 来 まし た 。
biz ø tunughun tokyo din kelø du q(一人称語尾, 複) .
あなたたち は 昨日 東京 から 来 まし た 。
sen ø tunughun tokyo din kel ø di nglar(二人称語尾、複) .
彼たち は 昨日 東京 から 来 まし た 。
u ø tunughun tokyo din kel ø di ø(三人称語尾、複) .

上の例で示したように日本語 {来ました} は代名詞が変わっても変わらない。しかし, ウイグル語では人称代名詞によって動詞 {kel} に接続してくる接尾が変わる。ウイグル語では各人称代名詞が独自の接辞を持つ。以下の表 5 でそらの対応関係を示す。

上の例で示したようにウイグル語の第一人称と第二人称では独自の接辞が語尾に付け加

表 5: 人称代名詞の対応関係

		単数型 (接辞)	複数型 (接辞)
一	日本語	私 (ø)	私たち (ø)
	ウイグル語	men((i)m,(i)men,(i)watimen)	biz((i)uq,(i)miz,(i)watimiz)
二	日本語	あなた (ø)	あなたたち (ø)
	ウイグル語	sen ((i)ng,(i)watisen,(i)sen)	silar ((i)nglar,(i)watisiler,(i)siler)
三	日本語	彼 (ø)	彼たち (ø)
	ウイグル語	u(wat(i))	ular(wat(i))

えられる。三人称は大体日本語と同じで変わりない。対訳辞書も人称接辞を考慮して作った。しかし、実際に日本語とウイグル語を派生文法に従って翻訳すると、ウイグル語の三人称にも独自の接辞があることになる、一方、日本語も各人称が三人称と同じ接辞を持つと考えられる。なお、実際のシステムでは派生文法ではなく普通の日本語が活用するという概念を前提に対訳辞書を作って、人称代名詞の処理に関して、ルールベースエンジンで処理をした。

そこで日本語の品詞とウイグル語の品詞をパラメータ化して、表 6 で示すようにまとめた。各パラメータが IPA 辞書の品詞情報を基に作っており、対訳辞書を作り際に、日本語と対訳の品詞情報をそのパラメータのようにして登録した。

機械翻訳システムを作成する際に、タグ付き品詞情報が必要である。各単語に品詞情

表 6: 品詞のパラメータ化

品詞	パラメタ	品詞	パラメタ
名詞	NU	代名詞	PRO
形容詞	ADJ	副詞	ADV
動詞	VE	助動詞	AUXVE
副助詞	APOP	終助詞	EPOP
格助詞	CPOP	係助詞	POPC
接助詞	COPOP	並立助詞	SAPOP
格助詞連語	CPOPCO	接頭詞	COP
名詞接頭詞	COPN	形容詞接頭詞	COPA
動詞接頭詞	COPV	類似頭詞	COPNO
接尾	SF	記号	SY
感動詞	INT	連体詞	ADP
形容動詞	ADVE	形容動詞ない	ADVENO
修飾語	O	人称語尾	PE
第一人称語尾	PEO	第二人称語尾	PES
第三人称語尾	PET		

報を持たせて、対訳辞書を作成した。

3 機械翻訳

本章では機械翻訳について述べる。特にルールベース翻訳と統計翻訳について説明する。

機械翻訳はルールベース翻訳と統計翻訳という大きい二つの種類に分類される。前者の言語の文法的関係を解析し、モデル化して、ルールを作って、そのルールを従って言語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文法関係をよく知ることが必要である。特に日英のような文法規則の違いが大きい言語間ではルールを決めるのがもっと複雑で、システムを構築するのには、たくさんのスペシャリストを必要とし、時間がかかってしまうケースが多い。翻訳規則をきちんと決めれば決めるほど翻訳の精度が高くなる。しかしながら、汎用性が低いという問題点もある。もう一つは統計翻訳である。統計翻訳に似ている用列ベース翻訳もあるが、現在は統計翻訳が主流になっている。統計翻訳は大量のデータを必要とする。文法スペシャリストを必要としないというメリットもある。汎

用性も高いので、対訳コーパスがあれば、どんな言語間でも翻訳ができる。データが大けれれば大きいほど翻訳の精度が高くなる。しかし、ウイグル語のように使う人が少ない言語に関しては、大量のコーパスが作られていないため、統計翻訳システムの構築が進んでいない状況である。独自でコーパスを作るにも時間がかなりかかるので、途中で断念してしまうケースも多い。

3.1 ルールベース翻訳

ルールベース翻訳方式を分類すると、直接変換方式、トランスファー方式、中間言語方式の三つがあると考えられる。

3.1.1 直接変換方式

直接変換方式は対訳辞書があれば簡単に翻訳を行う方式で、固定用語に関して使用するのが望ましい。単純に訳語を置き換えることで翻訳を行う仕組みで、文法語順が一緒の言語に対して簡単な句を翻訳できる。しかし、複文、修飾語を含んだ複雑な文章に対応できていない。

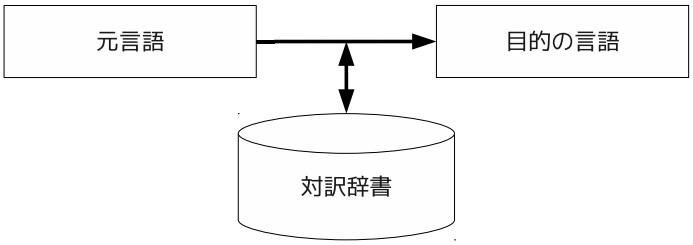


図 1: 直接変換方式

3.1.2 トランスファー方式

トランスファー方式では翻訳を行う文に対して形態素解析、構文解析、意味解析などの処理を行い、元言語から目的の言語に変換したあと、目的の言語の生成を行う。日英の

ような文法規則が違った言語に対して、上の手順で処理を行うが、日本語とウイグル語に対しては形態素解析まで処理を行って、その後文の生成をすれば、大体の翻訳が得られる。小川ら [5] が提案した派生文法による逐語翻訳でも構文解析を必要としなかった。

トランスファー方式は図２で示したとおりにソース言語解析部、構造変換部、目的言

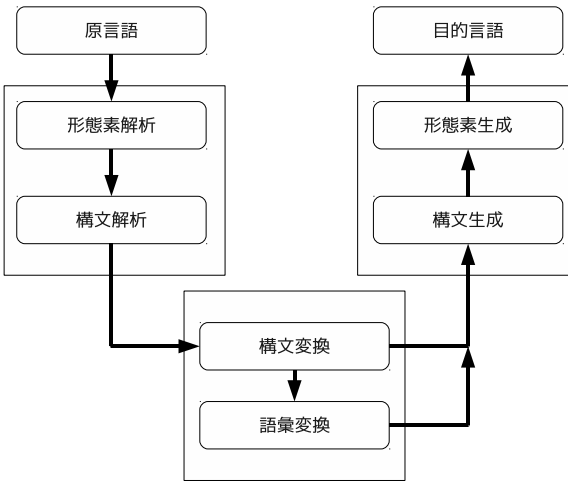


図 2: トランスファー方式

語生成部から成り立っている。以下で詳しく説明する。

(1)　ソース言語解析

ソース言語解析部は、通常形態素解析と構文解析からなる。日本語と多言語間の機械翻訳では最初は日本語を形態素解析で単語を分割しなければならない。一般に日本語を入力する時にスペースを必要としない、そのため言語処理の時に不便が生じる。そこで最初は形態素解析で日本語文字列を単語単位で分割しておくことが必要になる。分割された単語に関して品詞情報も付与することができるので、機械翻訳でも役に立つ。以下が日本語文を形態素解析した例である。形態素解析として Mecab[9] を使用する。Mecab は京都大学情報学研究科ー日本電信電話株式会社コミュニケーション科学基礎研究所を通じて開発されたオープンソース形態素解析ソフトである。

鳥は遠い所から飛んで来た。

```
鳥　　名詞, 一般,*,*,*,*, 鳥, トリ, トリ
は　　助詞, 係助詞,*,*,*,*, は, ハ, ワ
遠い　　形容詞, 自立,*,*, 形容詞・アウオ段, 基本形, 遠い, トオイ, トーイ
所　　名詞, 非自立, 副詞可能,*,*,*, 所, トコロ, トコロ
から　　助詞, 格助詞, 一般,*,*,*, から, カラ, カラ
飛ん　　動詞, 自立,*,*, 五段・バ行, 連用タ接続, 飛ぶ, トン, トン
で　　助詞, 接続助詞,*,*,*,*, で, デ, デ
来　　動詞, 非自立,*,*, カ変・来ル, 連用形, 来る, キ, キ
た　　助動詞,*,*,*, 特殊・タ, 基本形, た, タ, タ
。　　記号, 句点,*,*,*,*,。,。,。
EOS
```

ウイグル語は、もともと単語の間に空白があるので、形態素解析しなくても簡単に言語処理が可能だ。だが品詞が派生することから見ると、やはり形態素解析を行うことで翻訳精度がかなり上がると考えられる。

日本語と文法構造が違う言語間での機械翻訳に欠かせない処理として構文解析が挙げられる。構文解析で形態素解析から得られた単語の並び方から、言語の構造を表現している木構造を生成する。木構造の表現のしかたには二通りある。一つは Chomsky が提案した句構造規則 (phrase structure grammar)[10] で、もう一つは依存構造である。

句構造は名詞句や動詞句といったフレーズの集まりから言語構造を表現する方式である。例えば「鳥は遠い所から飛んで来た」という文に対して表 7 のような構造に成って、その木構造を図 3 で示す。

一方、依存構造では、単語間の係り受け関係を木構造で表現したものである。木構

表 7: 日本語句構造規則の例 1

S⟶ PP,VP
PP⟶ NP,P
VP⟶ PP,VP
VP⟶ V,TENS
NP⟶ N
N ⟶ 鳥, 遠い, 所
P ⟶ は, から, で
V⟶ 飛ん, 来
TENS⟶ た

造を構成するそれぞれのノードが単語になり、係り元単語が係り先単語の子ノードとなる

表 8: 非終段記号

S は文	VP は動詞句 (verb phrase)
NP は名詞句 (noun phrase)	PP は後置語句 (postpositional phrase)
P は後置語 (postposition)	V は動詞 (verb)
N は名詞 (noun)	TENS は時制 (tense)

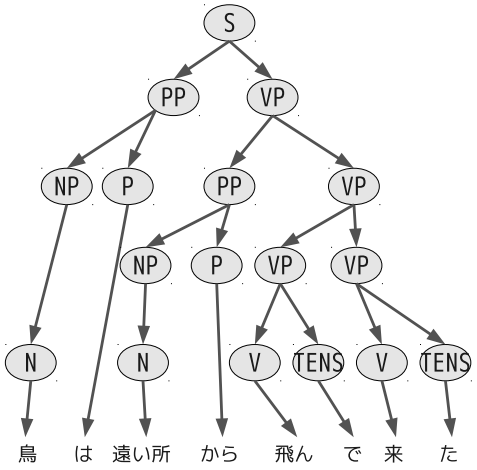


図 3: 句構造の例

ような構成される [11]。
(2) 構造変換
構造変換とは構文解析から得られる単語関係情報から、事前にデータベース化された構文変換規則を参照に目的の言語の構文構造に変換する処理である。
(3) 目的の言語生成
構文変換から得られる目的言語構文構造から、その言語文法規則に従って、間違った構文変換を正しい言語構造に変換する処理である。
日本語ーウイグル語間の機械翻訳では、同じ文法構造を持っているので、実は構文解析を必要としない。しかし、目的言語を生成するとき、接尾の多様さと助動詞が変化することから見ると、単語間の依存関係規則を作って、最後に出力される文字列の順番を決めることによって、いい翻訳結果につながる。

3.1.3 中間言語方式

中間言語方式とは、元言語や目的の言語とは独立した中間言語を設定し、元言語から中間言語へ翻訳してから、中間言語から目的の言語に翻訳すると言う仕組みである。これは多言語間翻訳に適した方法である。各言語に対して中間言語への変化処理と中間言語からの生成処理を記述すれば、任意の言語間での翻訳が可能になる。しかし、そのような共通的な中間言語を定めるのは難しいこともある。本研究でも中間言語方式について一切触れていないため、詳しい仕組みについても触れない。

3.2 統計翻訳

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機械翻訳手法で、対訳コーパスを学習し、言語間で翻訳を表すモデルを自動的に生成する。対訳コーパスさえ整えば、どんな言語の間でも翻訳ができる。統計翻訳のメリットとして、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と苦労が小さいこと、言語専門家を必要としないこと、汎用性の高いことが考えられる。しかし、対訳コーパスは整っているわけではない。たくさんの言語に関して、既に対訳コーパスがある一方、ウイグル語と他言語間との対訳コーパスがまだ整っていない。それで、ウイグル語の統計翻訳の研究が未だに進んでいない。コーパスの量が少ないと翻訳精度が低くなる。一般に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。現状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だ。

3.2.1 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) によって表される。これを Peter[12] 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、$\boldsymbol{argmax_u P(j|u)P(u)}$ を求めればよい。そして式が最終的に次の形になる。

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

図 4 で示したように、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

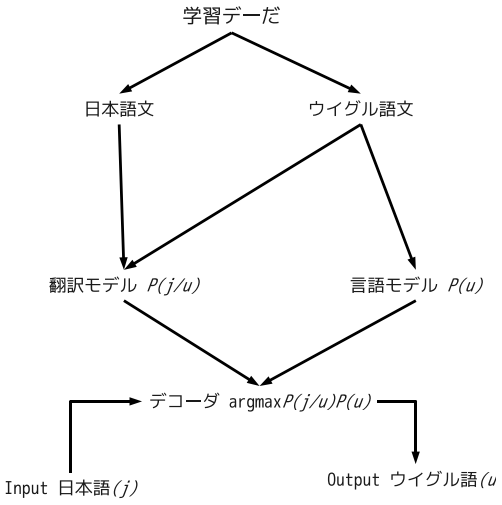


図 4: 統計翻訳処理フロー

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、[12] らはフランス語と英語の間の翻訳を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。

3.2.2 翻訳モデル

翻訳モデルは、原言語の単語列から目的言語単語列へ対訳コーパスを学習して確率的翻訳行うモデルである。大きく分けて、単語ベース翻訳モデルと句ベース翻訳モデルがある。現在句ベース翻訳モデルが主流になっている。理由として、単語ベース翻訳に比べて、翻訳精度が高いというメリットがある。句ベース翻訳モデル [13][14] は以下の式で表される。

$$\boldsymbol{p(\bar{f}_1^I|\bar{e}_1^I) = \Pi_{i=1}^I \phi(\bar{f}_i|\bar{e}_i) d(start_i - end_{i-1})} \tag{4}$$

式中の $\boldsymbol{I}$ は翻訳原言語 $\boldsymbol{f}$ の単語の連なりの数、$\boldsymbol{\bar{f}_1^I}$ はこれを句に分割したもの、$\boldsymbol{\bar{f}_i}$ は分割したそれぞれの句、$\boldsymbol{\bar{e}_i}$ は $\boldsymbol{\bar{f}_i}$ に対応した句、$\boldsymbol{a_i}$ は新たな翻訳する句の左端の位置、$\boldsymbol{b_{i-1}}$ は直前に翻訳した句の右端の位置である。ここで、$\boldsymbol{\phi(\bar{f}_i|\bar{e}_i)}$ を翻訳確率、$\boldsymbol{d(a_i - b_{i-1})}$ を歪み確率と呼ぶ。翻訳モデルはこれら二つの確率と関連する。

翻訳確率は、以下の式による相対確率で算出する。

$$\boldsymbol{\phi(\bar{f}|\bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_f count(\bar{f}, \bar{e})}} \tag{5}$$

相対確率は、それぞれの頻度に応じて確率を割り振ったものである。

歪み確率は、式 $\boldsymbol{d(start_i - end_{i-1}) = a^{|start_i - end_{i-1} - 1|}}$ で算出される。これは、翻訳する原言語の句の位置のずれに依存するモデルである。と言うのは、直前の翻訳した句の右端の位置と、次に翻訳する句の左端の位置の差の絶対値と関連する。

翻訳モデルの推定する手法として代表されるのは IBM モデルである。IMB モデルを EM アルゴリズムにより構築したシール GIZA ++ [15] によって最初単語対応を求める。求めた単語対応付けを使って対訳となる句を抽出する。最後に抽出した句の頻度から句の翻訳確率を求める。

3.2.3 言語モデル

言語モデルは、目的言語の単語列に対して、それらが起こる確率を付与するモデルである。日ウ翻訳で翻訳モデルで生成された翻訳候補からウイグル語として自然な文に

対して高い確率を与えることで選出する。言語モデルは、単語コーパスから学習される。言語モデルとして代表されるのは $\boldsymbol{N}$-gram モデルがある。$\boldsymbol{N}$-gram モデルは、単語列 $\boldsymbol{P(U_1^n) = u_1, u_2, ..., u_n}$ の $\boldsymbol{i}$ 番目の単語 $\boldsymbol{u_i}$ の生起確率 $\boldsymbol{P(u_i)}$ は直前の単語列 $\boldsymbol{u_{i-(N-1)}, u_{i-(N-2)}, ..., u_{i-1}}$ に依存するという仮説に基づいて提案されたモデルで以下は計算式である。

$$\boldsymbol{P(U_1^n) = \Pi_{i=1}^n P(u_i | u_{i-(N-1)}^{i-1})} \tag{6}$$

また、$\boldsymbol{P(u_i | u_{i-(N-1)}^{i-1})}$ の計算に以下の式を用いる。$\boldsymbol{count()}$ は単語列の出現数である。

$$\boldsymbol{P(u_i | u_{i-(N-1)}^{i-1}) = \frac{count(u_{i-(N-1)}^{i})}{count(u_{i-(N-1)}^{i-1})}} \tag{7}$$

以下が $\boldsymbol{N}$-gram モデル学習の例を表で表したもので、SRILM[16] を用いた。

表 6 で表したように左が単語 bu のあとに yerdin が来る確率である。真ん中は 2-gram

表 9: 2-gram の生成例

-2.778447	bu yerdin	-0.01282661
-2.235624	bu yerge	-0.06868306
-4.760924	bu yergimu	0

で生成された単語列、右は生成された単語列をスムージングによって生成され単語 bu の後に yerdin が来る確率である。

3.2.4 デコーダ

デコーダは翻訳モデルと言語モデルの確率が最大となる文を探索し、出力する仕組みであって、moses[17] が代表される。moses にはいくつかのパラメータを設定することができる。それらのパラメータをパラメータチューニング Minimum Error Rate Training(MERT) を行うことで最適値を求める。

moses のパラメータ例

- weight-l ... 言語モデルの重み (language model weights)
- weight-t ... 翻訳モデルの重み (translation model weights)
- weight-d ... 単語の移動の距離の重み (distortion(reordering)weight)
- weight-w ... 目的言語の長さに関するペナルティ (word penalty)
- distortion-limit ... フレーズの並び変えの範囲の制限値 (ditortion-limit)

4 日本語ーウイグル語の機械翻訳関連研究

本章では日本語ーウイグル語の機械翻訳に関連の研究を紹介して、成果と欠点を検討する。

4.1 日本語ーウイグル語の機械翻訳関連研究

日本語ーウイグル語の機械翻訳は名古屋大学の外山研究室で研究され、かなりの業績を出している。小川　泰弘, ムフタル・マフスット [5][6][8] らが膠着言語の共通の特徴に基づいて、派生文法に従った形態素解析をシステム (MAJO) を提案し、そのステムを利用して翻訳システムを構築した。インタネット公開している日本語ーウイグル語掲示板システム [7] もその手法で作成された。ムフタル・マフスット [8] らが最初に日本語の活用型に従って、対訳のウイグル語を辞書に登録する手法で助詞と助動詞のパラメータ化を推移グラフで求めるモデル提案して、翻訳行った。だが、この推移グラフは開始節点が動詞の活用型ごとに異なるため、一つの動詞の対して活用形の数だけ開始節点が必要である。また、一つの助動詞に対して複数の辺が対応しているため、実際に処理しにくいという理由で、小川らは派生文法に従って、助詞と助動詞の処理行う提案をして、翻訳を行った。表 7 でそれらの派生接尾の対応関係を示す。

図 5 はその手法による両言語間での翻訳例を示す.

図 5 で示したように入力文対して、形態素解析 MAJO で日本語を分割し、その後、対訳辞書を引くことで、訳語置換を行い、最後に生成されたウイグル語の文に対して整形を行う。これがシステムの主な流れである。派生文法で、一段活用動詞のように母音で終わる語幹を母音幹と呼ぶ。五段活用動詞のように子音で終わる語幹を子音幹と呼ぶ。接尾が接続される時、前の単語の末尾音節の母音か子音かによって、接尾の種類が変わる。また末尾の母音と子音の弱化、脱落、差入という問題とウイグル語の人称接尾と日本語人称接尾の違いで生じる問題が機械翻訳で最後出力するウイグル語の整形に大変な不便を与えてしまう。訳語置換を行った後に、ウイグル語整形ルールベースシステムを使って、正しい

表 10: 日本語ーウイグル語の派生接尾の対応

役割	日本語	ウイグル語	日本語例	ウイグル語例
使役	-(s)ase-	-guz-	書 k+ase-	yaz+guz-
受身	-(r)are-	-(i)l-	書 k+are-	yaz+il-
可能	-(r)e-	-(y)ala-	書 k+e-	yaz+ala-
丁寧	-(i)mas	-	書 k+imas-	yaz +
否定	-(a)na-	-ma-	書 k+ana-	yaz+ma-
希望	-(i)ta-	-gu-	書 k+ita-	yaz+gu-

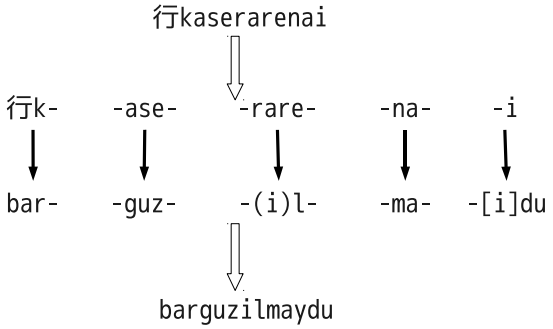


図 5: 派生文法による日本語ーウイグル語翻訳例

ウイグル語文を出力することを試みる。

4.2 ハイブリッド機械翻訳

最近統計翻訳とルールベース翻訳を合わせた機械翻訳の研究も増えている。それらのメリットを活かして、翻訳精度の向上が期待される。そのような翻訳をハイブリッド翻訳と言う。日本語とウイグル語の場合、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場合、派生文法の接尾問題が複雑で、いちいちルールを作るのも大変な作業で時間とコストがかかる。一方、統計翻訳でシステムの構築をする場合、対訳コーパスを前提とするので、最初はコーパスの整えることが条件となる。それらの弱点を二つの翻訳を合わせて克服することができる。これが本研究の目的で次節で詳しく述べる。

5 提案モデルの構築と実装

本章では提案モデルの構築及び実装について述べる。

5.1 ルールベース機械翻訳システムの開発手法と研究目的

本研究で形態素解析 Mecab を用いて独自の日本語ーウイグル語ルールベース機械翻訳システムの開発手法を提案する一方、日本語ーウイグル語統計機械翻訳に対して自作のコーパスを用いて実験を行うことで、これからのハイブリッド機械翻訳の可能性を観察することが目的である。

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情報を基に、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使わなくても済む。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た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、最初はルールベースシステムを作った。そこで日本語形態素解析 Mecab の Java version を用いてルールベース翻訳システムを作って、実装した。次に自作の対訳コーパスを用いて、統計翻訳を行った。最後に両方の翻訳結果を評価した。以下のような手順で行う

- 日本語を分かち書きし、それらをリストに登録して置く。
- リストに登録され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。

- パターンを解析し、生成されたウイグル語単語列に対して整形を行う。
- 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、自動評価手法を用いて評価を行う。
- 高い評価が与えられた単語列にを最終的に翻訳文をとして選ぶ

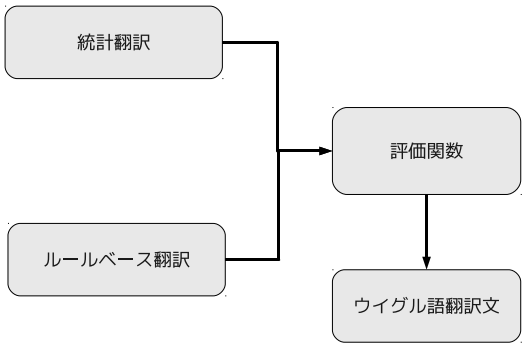


図 6: 日本語ーウイグル語機械翻訳システム

5.2 Mecab を用いた日ーウルールベース翻訳システムの提案

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きの処理が必要になる。そこで、日本語形態素解析として Mecab を使って日本語の分かち書きを行う。Mecab は辞書、コーパスを依存しない汎用的に設計されている。パラメータ推定に Conditional Random Fields(CRF) を用いており、他の形態素解析に比べて性能が向上していると考えられる。また、各種スクリプト言語でバインディングされている。本研究でも Java version を使う。Mecab は IPA 辞書と Juman 辞書を使う。二つとも CRF を用いてパラメータを推定する。解析された日本語文に対して以下のような出力情報を出す。

Mecab 出フォーマット

表層形　　品詞, 品詞細分類 1, 品詞細分類 2, 品詞細分類 3, 活用形, 活用型, 原形, 読み, 発音
日本　　名詞, 固有名詞, 地域, 国,*,*, 日本, ニッポン, ニッポン
行く　　動詞, 自立,*,*, 五段・カ行促音便, 基本形, 行く, イク, イク

システムの主な処理の流れを図 7 で示す。以下で各流れの処理を詳しく説明する。

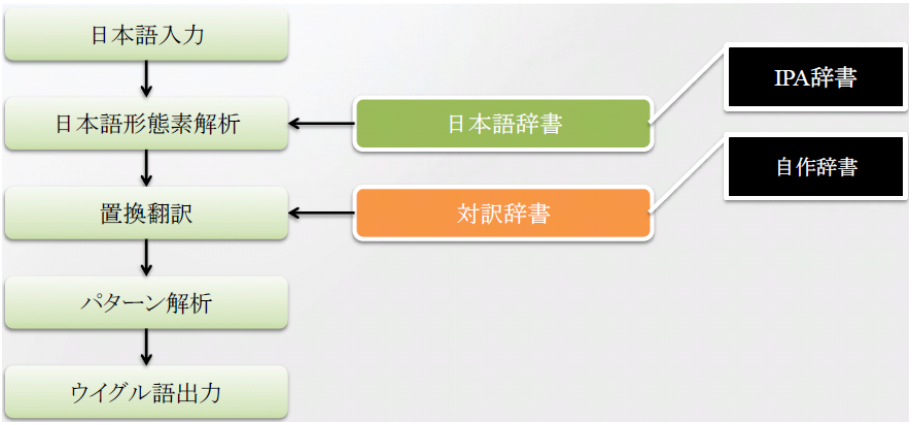


図 7: 日本語ーウイグル語ルールベースシステム

5.2.1 Mecab 出力フォーマットの設定

そこで、Mecab の出力フォーマットを自由に設定することができることから、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。それらをノード (日本語の単語) の基本情報としてリストに登録して置く。対訳辞書をそれらの情報を基にして引く。我々が必要としている単語とそれの品詞情報を得ることで、入力された日本語の前後関係に対して、値を設定することができることになり、辞書であらかじめ登録して置いた情報とマッチすることを前提にして、訳語を置換するコードを作る。出力フォーマットを`-F\\n%m,%f[0],%f[1]\n`で設定する。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 11 でその例を示す。

表 11 で示したように、出力された全ての形態素に対して品詞がある一方、品詞細分類 1 が全ての形態素にあると限られない。

5.2.2 日本語単語登録

Mecab で出力した単語とその単語の品詞情報をリストに登録して置く必要がある。そこで、***NodeTarget*** クラスを ***Java*** で自分で込んで、そのクラスにそれらの情報を登録するようにする。***NodeTarget*** クラスで三つ ***String*** 型の変数を作って、その変数の値としてそれらを登録して置く。表 12 で日本語形態素出力情報を示す

この例で示したように形態素 **鳥** とその品詞が最初にクラス ***NodeTarget*** の各変数に格納される。以降、毎回新しい形態素を読み込んだ時に新しい ***NodeTarget*** クラスが作成され、そのクラスの変数に新しい形態素とその品詞が格納され、形態素が終わりまでそ

表 11: 日本語形態素出力情報その 1

単語	品詞	品詞細分類 1
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	

表 12: 日本語形態素出力情報その 2

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	

ういう処理を行う。新しいクラスごとに ***List*** に格納される。主な処理が以下の step のようになる。

NodeTarget クラスの処理

1. Mecab で検出され形態素を CSV 形でリストに登録する, その形態素と品詞情報を一つの行ごとに読み込む
2. 読み込んだ一つの行の長さ (一行の形態素とその品詞の数) を測定し、それらを ***NodeTarget*** クラスの各変数に格納する
3. そのクラスが一回の変数を格納した段階でリストに登録される
4. 次に来た行に対して 1 からの処理を行う
5. 行がなくなったら、処理を終了

5.2.3 対訳辞書から訳語を決定

TargetNode を格納したリストから、単語品詞を基にして、その単語がどちらの対訳辞書に格納されているかを決定して、訳語を検索する。対訳辞書を Mecab で使用する ***IPA*** 辞書のように ***CSV*** データ構造のように作る。作る方法を 5.3 節で詳しく説明する。そこで表 3 で示してように ***IPA*** 辞書で登録された各単語の品詞の種類をパラーメタ化することによって、それぞれの品詞にマッチした形態素をそれらの辞書を引くことでウイグル語の単語列を取得する。以下例を持ってどのように辞書を引くかを述べる。

”鳥は遠い所から飛んで来ました”という文は ***NodeTarget*** クラスの各変数に格納されたあと、***NodeTarget*** クラスを格納する ***List*** を作る。そのあた ***List*** から最初の ***Elment*** から一つ一つ読み出す。

- $\boldsymbol{if\ w_i}$(鳥) ***is NU then open File.Nu***
- $\boldsymbol{if\ w_{i+1}}$(は) ***is POPC then open File.POPC***
- $\boldsymbol{if\ w_{i+2}}$(遠い) ***is ADJ then open File.ADJ***
- $\boldsymbol{if\ w_{i+3}}$(所) ***is NU then open File.Nu***
- $\boldsymbol{if\ w_{i+4}}$(から) ***is CPOP then open File.CPOP***
- $\boldsymbol{if\ w_{i+5}}$(飛ん) ***is VE then open File.VE***
- $\boldsymbol{if\ w_{i+6}}$(で) ***is COPOP then open File.COPOP***

- ***if*** w_{i+7}(来) ***is VE then open File.VE***
- ***if*** w_{i+8}(まし) ***is AUXVE then open File.AUXVE***
- ***if*** w_{i+9}(た) ***is AUXVE then open File.AUXVE***

辞書引きに対して、単語の品詞情報が決め手になる。それぞれの辞書に *ipadic* 辞書のように日ーウ語訳と品詞情報が格納されている。目的の辞書が決まったら、その辞書から対訳を検索し、見つかったら、その訳語を返す。なかったら *null* を返すようにし、これらの処理は *DictioaryFix* クラスで実装する。
{ 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が分かる。

表 13: 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

5.2.4 訳語生成

5.2.3 節で決定された訳語を出力結果として、*List* に格納して置いて、順番を付ける作業である。例えば、以下のような *List* が作られる。

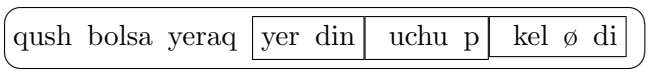
表 14 で示したように、日本語に対してウイグル語の訳がない場合は、Øで与え、リストに登録して置く。大体の場合動詞と助動詞、接続詞などの訳語は 1 対 1 ではないから、辞書から最も的した訳語を選ぶことが一般のルールベースシステムで難題の一つである。統計翻訳では一般に単語列の前後関係を確率的に求めることから、以上の問題を簡単に解決することができる。本研究でルールベース翻訳と統計翻訳を合わせる目的の一つもそれの問題を解決することである。*List* に格納されたウイグル語の単語列が順番どおりに出力されると、一つのウイグル語の文になる。ウイグル語と日本語は文法構造が同じであるから、構文解析を用いて訳語置換を行う必要はない。先の表 14 で示されたウイグル語の単語列を順番どおりに出力させると、以下のようになる。

qush bolsa yeraq yer din uchu p kel ø di

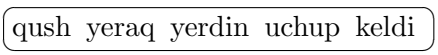
表 14: List に格納された訳語と品詞情報

StringKey(訳語)	StringValue(その品詞)
qush	NU
bolsa	POPC
yeraq	ADJ
yer	NU
din	CPOP
uchu	VE
p	COPOP
kel	VE
ø	AUXVE
di	AUXVE

それを見ても大体正しい順番になっている。しかし、ここで少しの問題が生じる。ウイグル語では助詞を前の単語に付け加えるので、先のような出力では不十分になる。それらを訳語の品詞情報を用いて、あらかじめ決めて置いたルールに基づいて解決する。訳語の品詞情報を分かれば、その品詞を持つ単語の前後に来る単語の接合状態を決める事が可能になる。接合可能な形態素を接合して文整形を行った後に理想的な出力形になる。それを以下で示す



最終結果



この例でウイグル語単語 bolsa は脱落し、yer din | uchu p | kel ø di らがお互いに接合することになった。次の節でそれをどのように実現させるかを説明する。

もう一つの問題が日本語では人称助詞がないことに対して、ウイグル語では人称助詞があることで、それらの処理も次の節で詳しく説明する。

5.2.5 ルールベースエンジンの作成

ルールベースを作ることで以下の問題を解決することができる。

1. 単語の前後関係から、接辞が接合する語幹を決める

2. 生成されたウイグル語の文にたいして、人称語尾を正しく決める

3. ウイグル語での文字の弱化、脱落、差入などの問題を解決する

1. 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

ウイグル語が日本語と違って、各人称代名詞が各自の人称語尾を持っている。それらが日ーウ機械翻訳で、必ず解決すべき問題である。表 5 で人称代名詞とその接辞の対応関係を示した。本研究で人称語尾を決定するルールが以下の二つの構造を持つ。

人称語尾決定ルール

- 構造 (1) PRO + O + VE + AUXVE + PE
  例： men + O + al + di + m
- 構造 (2) PRO + POP + O + VE + AUXVE + PE
  例： men + din + O + al + wal + di + ∗(PET)
  文の人称語尾を決めるアルゴリズムを以下に示す

```
METHOD getRulePro()      //人称語尾を決めるメソッド
    node for List⟨Node⟨String⟩⟩ w_id // node クラスに格納されているウイグル語の単語列
      if matcher(node(w_id),"PROX") := true // 単語列から人称代名詞を探して出す、あれば次の処理、なければ終わり
       if nextNode(w_id++) != "SY" // 単語列に終段記号が現れたら、ループを終了, なければ次の処理
         if nextNode(w_id++) := "PEX" // 単語列に含まれた人称語尾を確認
          removeElmementAt(nextNode(w_id++)) //確認された人称語尾を削除
          addLastIndex(newNode(w_PEX))// 適切な人称語尾を入り替える
         endif
        endif
        else
         break
      endif
    endfor
endmethod
```

人称語尾を決めるに、method getRulePro を使う。最初は単語列とそれらの品詞情報をゲットして、もし、それらの中に代名詞が見つかれば、その次の単語の品詞をチェックし、助動詞であれば、メソッドを終了する。そでなければ、文節の最後の単語を取り除いて、そこに、新しい単語として、人称語尾を追加して、メソッドを終了する。以上のステップで、人称語尾の決定が決まる。

5.2.4 節で述べたように、接辞がどの単語に接合されるかを決める問題に関してもルールを作って解決する。

2. 単語の前後関係から、接辞が接合する語幹を決める。

先の訳語の例で、品詞情報だけ出力すると

NU POPC ADJ	NU CPOP	VE COPOP	VE AUXVE AUXVE

ようになる。ウイグル語の文法規則によりあらかじめ各品詞関係をルール化して、そのルールに従って文の生成を行う。結果的に品詞間での関係に基づいて各接辞が前の単語に接合される可能性を探って、条件を満たせば接合が可能になるような仕組みなる。ウイグル語で助詞の種類がたくさんあるので、すべての接尾に対してルールを決めるのが難しい。そこで本研究では、最初は格助詞の処理をメインとして行った。ウイグル語と日本語の格助詞対応関係を表１５で示す。

ウイグル語単語列の中で格助詞が見つかった場合、その格助詞が前の単語と接合する

表 15: 格助詞対応表

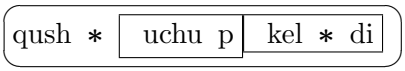
日本語	ウイグル語	日本語	ウイグル語
が	*	で	de,te,da,ta
の	ning	と	bilen
から	din,tin	より	din ,tin
を	ni	に	gha,ga,ge,ke
へ	gha,qa,ge,ke		

可能性が高いことが分かる。そこで文 $\boldsymbol{S}$ に対してルールを決めることを説明する。ここで文 $\boldsymbol{S}$ が $\boldsymbol{W_i}$ の単語列と空白から成り立っていると考える。

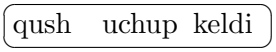
格助詞のルール例

- if key $\boldsymbol{W_i}$(ning) is $\boldsymbol{true}$ then Replace($\boldsymbol{W_{i-1} := W_{i-1} + W_i}$)
- if key $\boldsymbol{W_i}$(din,tin) is $\boldsymbol{true}$ then Replace($\boldsymbol{W_{i-1} := W_{i-1} + W_i}$)
- if key $\boldsymbol{W_i}$(ni) is $\boldsymbol{true}$ then Replace($\boldsymbol{W_{i-1} := W_{i-1} + W_i}$)
- if key $\boldsymbol{W_i}$(gha,qa,ge,ke) is $\boldsymbol{true}$ then Replace($\boldsymbol{W_{i-1} := W_{i-1} + W_i}$)
- if key $\boldsymbol{W_i}$(de,te,da,ta) is $\boldsymbol{true}$ then Replace($\boldsymbol{W_{i-1} := W_{i-1} + W_i}$)
- if key $\boldsymbol{W_i}$(bilen) is $\boldsymbol{true}$ then Replace($\boldsymbol{W_{i-1} := W_{i-1} + W_i}$)
- 詳しい処理が以下に示す
- 単語列 $\boldsymbol{W_i}$ が与えらた時、集合 $\boldsymbol{w \in S}$ から単語 $\boldsymbol{w_i}$ を呼び出して、その単語の品詞 $\boldsymbol{w_{id}}$ チェックする
- もし、$\boldsymbol{w_{id}}$ が $\boldsymbol{POP}$(格助詞) であれば、$\boldsymbol{w_{i-1}+w_i}$ を新しい文字列として、$\boldsymbol{w_{i-1}}$ に格納する
- もし、$\boldsymbol{w_{id}}$ が $\boldsymbol{VE}$(動詞) であって、$\boldsymbol{w_{i+1(id)}}$ が $\boldsymbol{AUXVE}$(助動詞) であれば,$\boldsymbol{w_{i+2}}$ の $\boldsymbol{id}$ をチェックし、もし $\boldsymbol{AUXVE}$ であれば、$\boldsymbol{w_i + w_{i+1} + w_{i+2}}$ を新しい文字列として、$\boldsymbol{w_i}$ に格納する
- それ以外であれば $\boldsymbol{w_i}$ に関して何もしない
- 最後に、未知語 $*$ を単語列から取り出す

主な作業は接合された文字列の間の空白を取り除くことであるので、格助詞 $\boldsymbol{W_i}$ が見つかることを条件として、その条件を満たせば、一つ前の空白を取り除いて、助詞とその空白前の $\boldsymbol{W_{i-2}}$ 単語を結合する。



最終結果



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

5.3 対訳辞書について

本節ルールベース翻訳システムを作り際に必要な辞書について述べる。辞書には日本語形態素解析が用いる IPA 辞書と日ーウ対訳辞書がある。

5.3.1 IPA 辞書

日本語形態素解析システム Mecab は通常 IPA 辞書と Juman 辞書のどちらかを使う。今回 IPA 辞書を使うことにした。IPA 辞書は IPA コーパスに基づき CRF でパラメータ推定した辞書である。IPA 辞書には日本語各単語の品詞情報を基づいて、CSV データ型で作られた辞書である。各品詞ごとに別々の CSV ファイルで保存される。以下の表で IPA 辞書ファイルを示す。

表 16 で示したよう各 CSV ファイルにはそのファイル名と同様の品詞を格納する。一

表 16: IPA 辞書ファイル

Adi.csv	Adnominal.csv	Adverb.csv	Auxil.csv	Conjunction.csv
Filler.csv	Interjection.csv	Noun.adjv.csv	Noun.adverbal.csv	Noun.place.csv
Noun.csv	Noun.nai.csv	Noun.name.csv	Noun.name.csv	Noun.number.csv
Nou.verbal.csv	Others.csv	Postp-col.csv	Postp.csv	Prefix.csv
Symbol.csv	Nou.other.csv	Noun.proper.csv	Verb.csv	Suffix.csv

つの単語に対して、表層形, 品詞, 品詞細分類 1, 品詞細分類 2, 品詞細分類 3, 活用形, 活用型, 原形, 読み, 発音との品詞情報を付与する。本研究でも日ーウ対訳辞書を品詞ごとに csv データとして別々のファイルに格納する。こうすることで対訳辞書引くアルゴリズムが簡単になり、手間を削減することが可能になる。

5.3.2 日ーウ対訳辞書

日ーウ対訳辞書を IPA 辞書の単語の品詞ごとに各ファイルにした。品詞情報を少なくし、単語とその品詞だけにした。以下の表１７でその例を示す。

表 18 で示したように各単語に対してその活動型に基づいた形でウイグル語の訳語と品詞情報を登録しておいた。形態素解析で得られた単語が日ーウ対訳辞書に登録され単語とマッチするような形にした。日本語を形態素ごとに分割すると細かい単語列が生成されて、日ーウ対訳辞書でも登録されている日本語とウイグル語に対して形態素ごとに登録した。

辞書ファイルが CSV データ形であることと IPA ファイルのかずと同じであること以外に、助動詞を格納したファイルに対して IPA 辞書と異なって各助動詞ごとに対訳ファイルを作った。こうする事で、辞書を引いた時の過ちを少なくすることを実現させた。以下の表 19 がその例を示す。

5.2.3 節で述べたように日本語の単語列を形態素解析を終えた後に、各単語の品詞の基に各辞書ファイルを引くことで、訳語を検索する仕組みであるので、辞書ファイルを品詞

表 17: 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

表 18: 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

表 19: 日ーウ対訳辞書ー助動詞格納ファイル

ファイル名	品詞	日本語	ウイグル訳
EPOP	終助詞	かしら	midu
APOP	副助詞	だって	disimu
SAPOP	並立助詞	とか	hem
POPC	係助詞	すら	mu
COPOP	接助詞	けど	emma
CPOP	格助詞	を	ni

ごとに登録した。

辞書に登録され単語全体で 15000 を越えた。ただ、活用しない名詞などが比較的多いので全体 7 割を占める。動詞と形容詞については、活用するごとに対訳を登録することで、一つの単語に対して少なくても 8 種類の訳語が登録されることになった。従って、全対訳辞書に対して、それらの占める割合が少ない。およそ動詞を 1000 単語と形容詞を 1000 単語ぐらいにした。

5.4 日ーウ統計翻訳手順と実験

5.4.1 学習データの準備

日ーウ統計翻訳行う前、対訳コーパスの適切な処理が必要である。各対訳文が極めて長いと翻訳モデルを学習することができないケースも多少ある。

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 2565 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 20 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。

表 20: 対訳コーパス例

教室 は 知識 を 与える 。 sinip bilim beridu.
知識 を 増やす の を 目標 に する 。 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 。 ming teslikte berghen nersini ishletmestinla tashliwetsek yahshi emes.
良く 覚え て おけ 。 isingde ching saqla.
覚え て いる か どう か 、 ときどき 試験 を し て 調べる 。 este tutqan tutmighanliqni,daim imtahan elip sinap turidu.
覚え て い なけれ ば 減点 し て 警告 する 。 este saqlimisaq numur tartip agahlanduridu.

5.4.2 *N*-gram 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 [16] を用いた。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 で設定した。ウイグル語言語モデルを作成した際に用いた文は 6563 文である。　23 表は N を 3

表 21: *N*-gram で生じる日本語単語列

N-gram N	count
N-gram 1	5885
N-gram 2	12842
N-gram 3	447
N-gram 4	167

表 22: *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

表 23: 日本語 3-gram 言語モデル例

$\boldsymbol{P(w_i \mid w_{i-1}, w_{i-2})}$	*3*-gram 単語列	*back-off smoothing* $\boldsymbol{P(w_i \mid w_{i-1}, w_{i-2})}$
-0.09820086	の 販売 機	-0.1389655
-0.05337211	自動 販売 機	-0.1675822
-0.1611529	を 買い まし	-0.4300385
-0.3339288	を 買っ て	-0.1124369
-0.2791271	を 貸し て	-0.1146558
-0.2230652	s 赤 繁	-0.1858207

にした時の日本語の言語モデルの例で、左の数値は日本語単語**の**の後に**販売、機**の来る確率を常用対数 $\mathbf{log_{10}}$ でとった値 $\mathbf{log_{10}}\, \boldsymbol{P(w_i \mid w_{i-1}, w_{i-2})}$ である。次に、*3*-gram で表された単語列**の　販売　機**、最後の数値は *back-off smoothing* で推定された、日本語単語**の**の

後に**販売、機**の来る確率を常用対数 $\log_{10}$ でとった値 $\log_{10} P(w_i|w_{i-1}, w_{i-2})$ である。表 24 は N を 3 にした時のウイグル語の言語モデルで生じる単語列とその確率の例で各数値のの意味が表 23 と同じである。

表 24: ウイグル語 3-gram 言語モデル例

$P(w_i\|w_{i-1}, w_{i-2})$	*3*-gram 単語列	*back-off smoothing* $P(w_i\|w_{i-1}, w_{i-2})$
-0.03815184	din ibaret .	-0.06479263
-0.03589532	tin ibaret .	-0.4018514
-0.008819266	bar idi .	-0.9247303
-0.01781972	kop idi .	-0.4018517
-0.01781972	qalghan idi .	-0.4018517
-0.01891517	qilghan idi .	-0.06479287

5.4.3 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。

1. 単語のアライメント (alignment) の計算

この計算には 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。ここで計算された日ーウ、ウー日の両方向の単語アライメントから、日ーウ、ウー日方向に 1:N の単語列アライメントを求める。この単語列アライメントは双方向の単語対応の和集合 (union) と積集合 (intersection) を利用してヒューリスティックスで求める [18]。通常の統計翻訳では和集合と積集合の中間ヒューリスティックスとして、”grow-diag”がある。”grow-diag”の最後の処理として”final”と”final-and”がある。”final-and”では,”final”に加えて、双方向共に単語対応がアライメントも用いる。本研究でも”grow-diag-final-and”を用いた。以下の表 25 で最初の単語アライメントの計算を示す。

”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 26 で示す。

次に単語列アライメントから、ヒューリスティックを用いて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[17](本研究で最終段階で行うウーウ統計翻訳も同じ手順で行うので代表として日ーウ統計翻訳の手順を説明することにした。) 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する。表２７は”grow-diag-final-and”で作成されたフレーズテーブルの表である。

表 25: 単語アライメントの計算

Sentence pair (1530) source length 6 target length 9 alignment score : 1.17518e-10
どうして そう 思う か 、 聞き まし た 。
NULL ({ }) nimishqa ({ 1 }) shundaq ({ }) uylaysen ({ 2 3 4 })
dep ({ }) soridim ({ 5 6 7 8 }) . ({ 9 })
Sentence pair (1531) source length 4 target length 5 alignment score : 1.34133e-05
常識 だ よ 。
NULL ({ }) ” ({ 1 }) adettiki ({ }) bilim ({ 2 }) ghu ({ 3 4 5 })
Sentence pair (1532) source length 8 target length 12 alignment score : 6.41833e-14
子ども の とき から 、 いつも 言わ れ た よ 。
NULL ({ }) kichik ({ 1 }) chaghlardin ({ 2 3 }) bashlap
({ 4 }) shudaq ({ 5 6 7 8 }}
dep ({ }) kelghen ({ 9 10 }) . ({ 11 }) ({ 12 })
Sentence pair (1533) source length 12
target length 16 alignment score : 1.09429e-18
わたし は どうして そんな 常識 が でき た か 、 調べ て み まし た 。
NULL ({ }) men ({ 1 2 }) nimishqa ({ 3 }) shundaq ({ 4 })
adettiki ({ }) bilim ({ 5 }) boldighan ({ 6 })
du ({ 7 9 }) dep ({ 12 }) , ({ 10 }) izdinip
({ 11 }) baqtim ({ 8 13 14 15 }) . ({ 16 })

表 26: 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					★	
。						★

表 27: ”grow-diag-final-and”で作成されたフレーズテーブル

5 人	5	0.142857 0.0074013 1 0.166667	7 1
5 人 でき	5	0.142857 1.62311e-05 1 0.166667	7 1
5 人 でき まし	5	0.142857 8.30543e-08 1 0.166667	7 1
5 人 でき まし た よ	5 boldi	1 2.00418e-09 1 0.0139109	1 1
5 人 でき まし た よ 。	5 boldi .	1 1.97954e-09 1 0.0135274	1 1
5 人 の	5 ademning	1 0.09375 1 0.00112323	1 1
5 人 の 生活 です	5 ademning turmushi	1 0.00765306 1 0.000109203	1 1
5 人 の 生活 です から	5 ademning turmushi bolghanliqtin	1 0.00382653 1 3.15616e-06	1 1
5 人 の 生活 です から 、	5 ademning turmushi bolghanliqtin ,	1 0.00353495 1 2.46764e-06	1 1

5.4.4 デコーダの設定

デコーダは moses[17] 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。

moses のパラメータ

- ttable-file ... 0 0 0 4
- lmodel-file ... 0 0 3
- ttable-limit ... 20
- weight-l ... 0.5000
- weight-t ... 0.20 0.20 0.20 0.20 0.20
- weight-d ... 0.3 0.3 0.3 0.3 0.3 0.3 0.3
- weight-w ... -1
- distortion-limit ... 6

5.4.5 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。

自動評価手法として、あらかじめ用意した翻訳正文と、機械翻訳で出力した翻訳結果を比較する方法がある。代表的なのは BLEU(Bilingual Evaluation Understudy)[19]、NIST(The National Institute of Standards and Technology)[20] が挙げられる。

自動翻訳評価指標 BLEU では、翻訳された文に関して、人手であらかじめ参照訳文を作っておいて、翻訳結果とその参照文を比較して、参照文に近ければ高いスコアを与える。

参照文を複数用意しておくと評価の精度が高まる。評価値が 0 から 1 の間に成るのが普通である。　BLEU スコアは次に式で求められる。

$$\boldsymbol{BLEU = BP_{BLEU} \times \exp(\sum_{n=1}^{N} w_n \log p_n)} \tag{8}$$

$\boldsymbol{p_n}$ は、翻訳文と参照文における N-gram の一致率を表している。以下の式で求められる。

$$\boldsymbol{p_n = \frac{\sum_{C \in (Candidates)} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in (Candidates)} \sum_{n-gram' \in C'} Count(n-gram')}} \tag{9}$$

$\boldsymbol{BP_{BLEU}}$ は翻訳文は参照文より短い場合にはペナルティを与える。$\boldsymbol{BP_{BLEU}}$ はそのペナルティであって、以下の式で求められる。

$$\boldsymbol{BP_{BLEU} = \begin{cases} 1 & (c > r) \\ e^{(1-r/c)} & (c \leq r) \end{cases}} \tag{10}$$

NIST は BLEU をベースに作られた自動評価指標である。式は以下のようである。

$$\boldsymbol{Score = \sum_{n=1}^{N} \frac{\sum_{all w_1 \cdots w_n that co-occur} Info(w_i \cdots w_n)}{\sum_{all w_1 \cdots w_n in sys output}(1)} \times \exp\left(\beta \log^2(min(\frac{L_{sys}}{L_{ref}}, 1))\right)} \tag{11}$$

ただし、

$$\boldsymbol{Info(w_1 \cdots w_n) = \log_2(\frac{\texttt{the count of occurrences of} w_1 \cdots w_{n-1}}{\texttt{the count of occurrences of} w_1 \cdots w_n})} \tag{12}$$

となる。　我々は BLEU と NIST を用いてルールベース機械翻訳と統計機械翻訳に関して自動評価を行うことと共に人手での評価も行った。次に章で実験評価について述べる。

6　システムの実装及び実験評価

本章では提案モデルを使って実際にシステムを作り、実装して、翻訳結果を評価する。

今回システム構築に必要だった対訳コーパス、言語モデルコーパス、単語辞書を表 28 でまとめた。

日本語の形態素解析ソフト Mecab を用いて開発したルールベース機械翻訳システムは置換翻訳 (Replacement Translation) とルールベース翻訳 (Rulebased Traslation) の二つの部分に分かれる。図 8,9,10 がそれぞれの段階での翻訳結果を示したシステムの外見である。最初がまだ翻訳ルールが決まっていなかった時の、単純の置換翻訳であって、その次がルールが決められた時の翻訳結果である。

表 29,30,31,32 と図 11,12,13,14 が統計翻訳とルールベース機械翻訳に関して自動評価指標 BLEU と NIST 用いて評価した時の結果である。自動評価指標 BLEU と NIST の実験の際に、N-gram 値を 1 から 5 までに設定して、単語列 (この場合 N-gram のことを指

表 28: 学習データのまとめ

種類	文	単語 (重複可能)
統計翻訳モデル学習データ	2565	29152
統計翻訳言語モデル学習データ	6563	755441
ルールベース翻訳対訳データ		5084

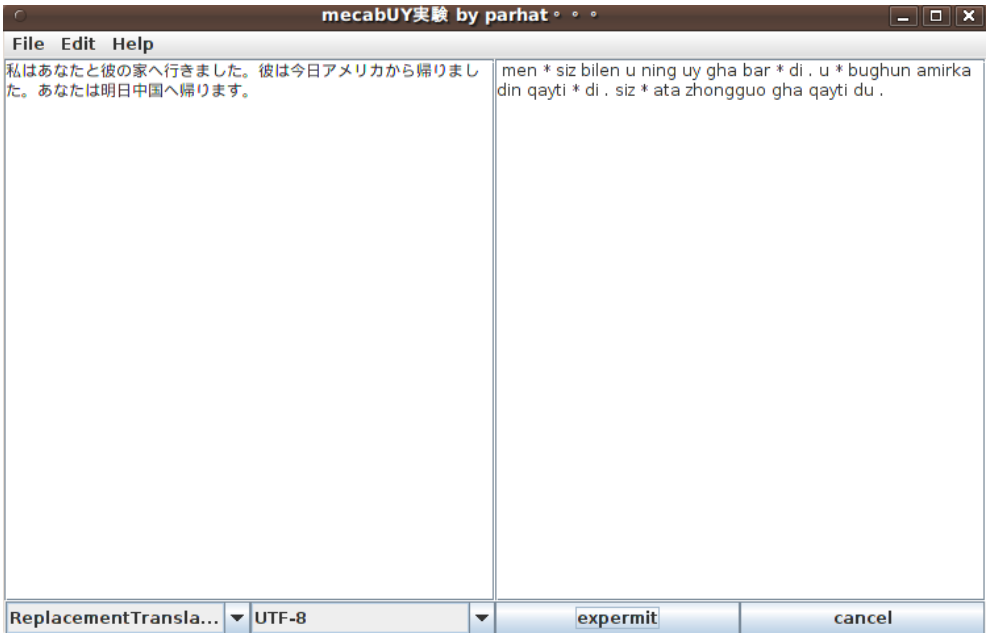


図 8: Replacement Translation System

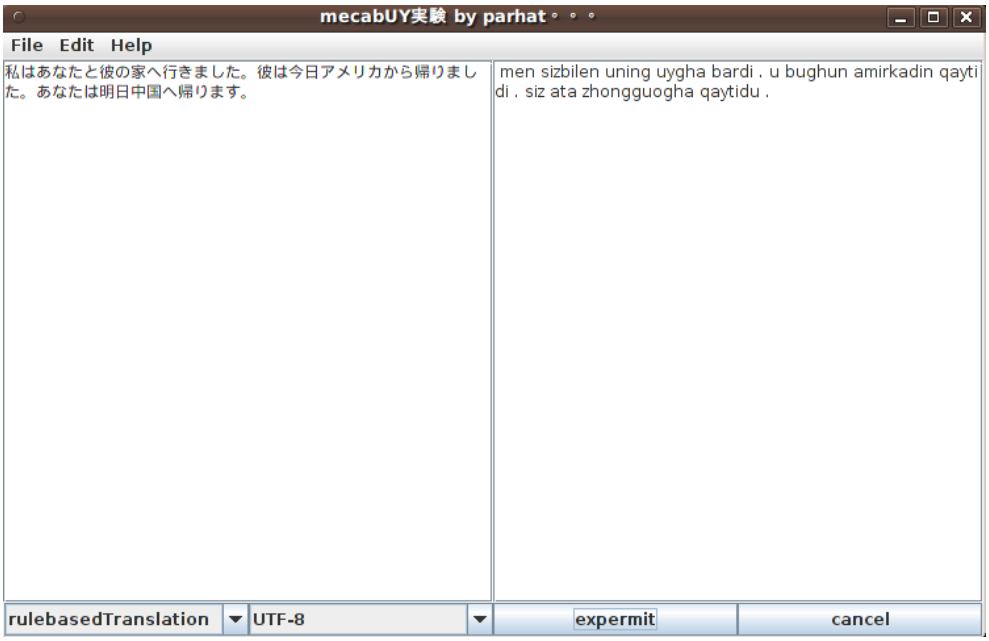


図 9: RuleBased Translation System

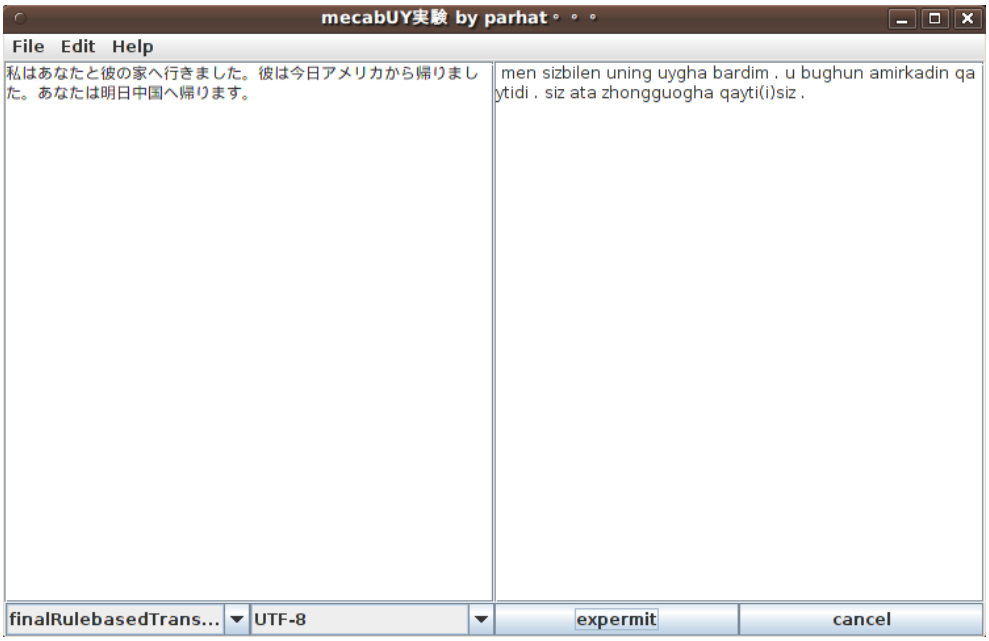


図 10: FinalRuleBased Translation System

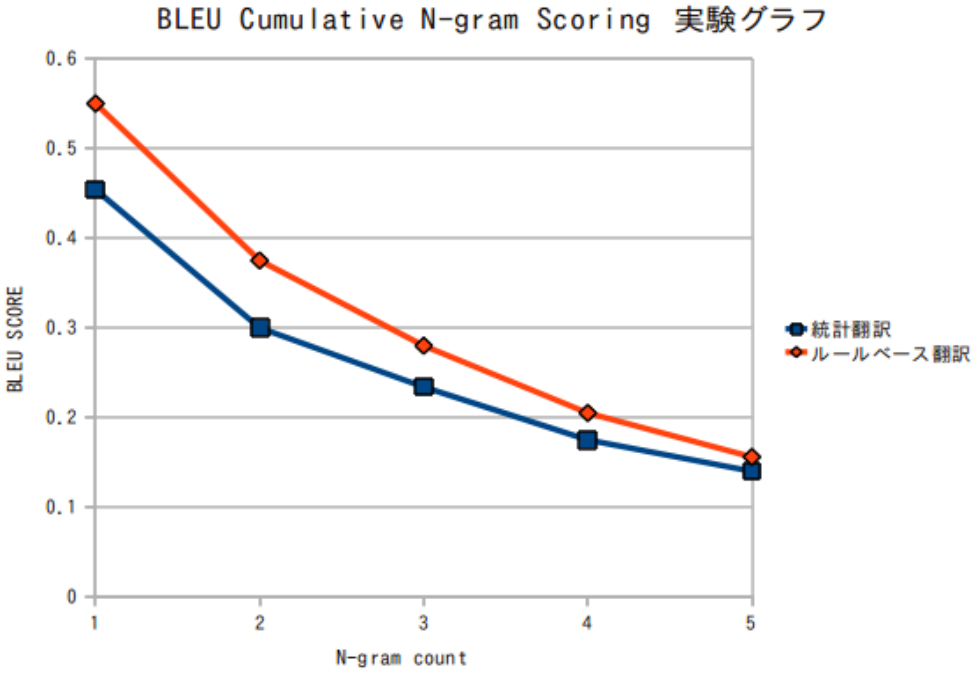


図 11: BLEU Cumulative N-gram Scoring 実験グラフ

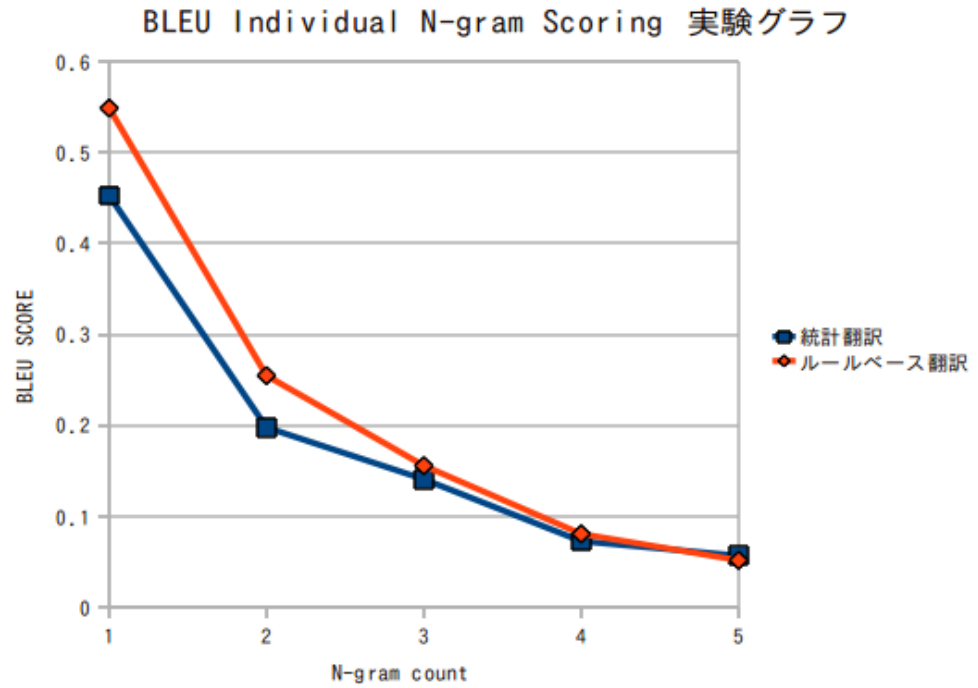


図 12: BLEU Individual N-gram Scoring 実験グラフ

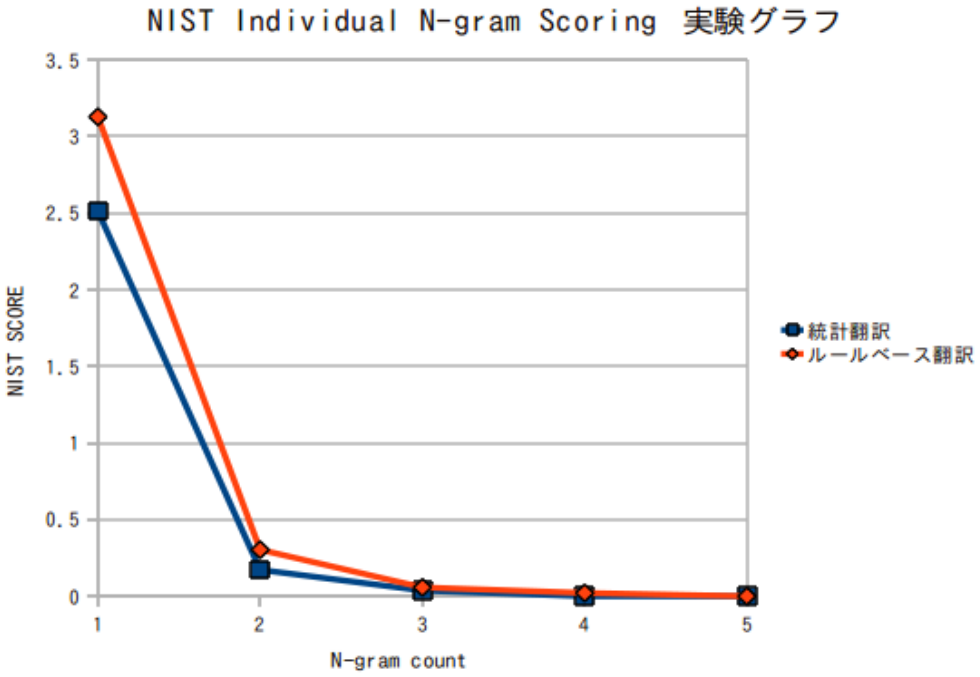


図 13: NIST Individual N-gram Scoring 実験グラフ

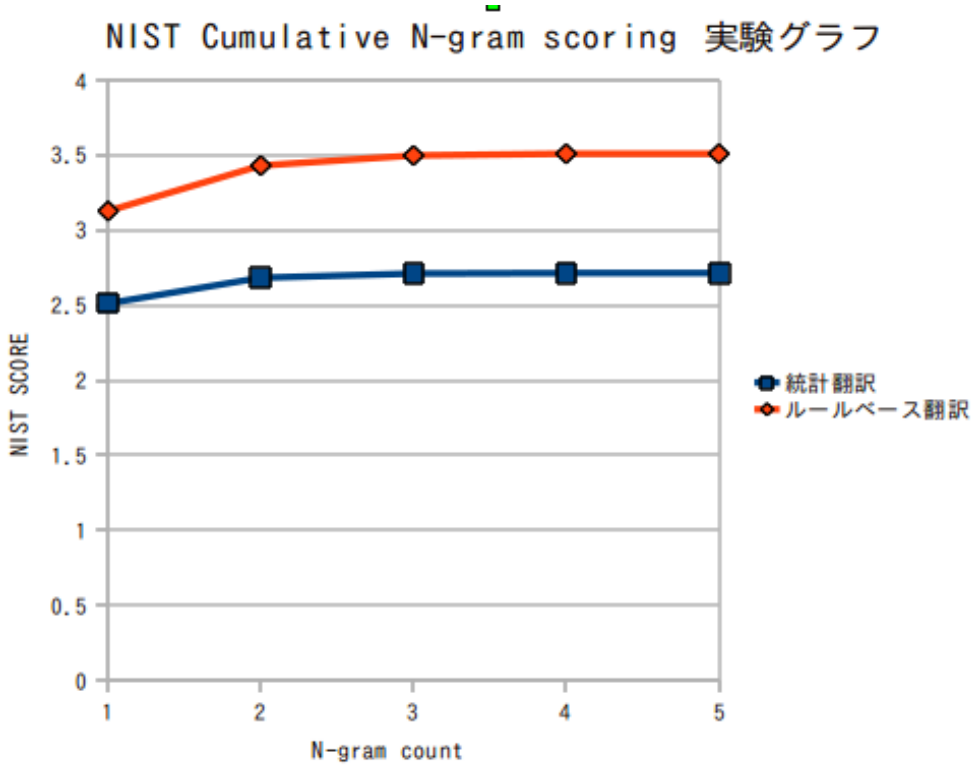


図 14: NIST Cumulative N-gram Scoring 実験グラフ

す) の重複を考慮した時と単語列の単一発生した時の BLEU 値と NIST 値を求めて、ルールベース翻訳結果と統計翻訳結果を比較した。今回開発したルールベース翻訳システムの翻訳結果が統計翻訳結果に比べてわずかに良いことが分かった。

実験結果が以下の表 33 で示した結果は人手で評価した時の結果であって、二つの翻

表 29: BLEU Individual SCORE 実験データ

BLEUScore(実験データ)	N=1	N=2	N=3	N=4
統計翻訳 (50)	0.453	0.198	0.142	0.072
ルールベース翻訳 (50)	0.549	0.255	0.156	0.081

表 30: BLEU Cumulative SCORE 実験データ

BLEUScore(実験データ)	N=1	N=2	N=3	N=4
統計翻訳 (50)	0.453	0.299	0.174	0.139
ルールベース翻訳 (50)	0.549	0.374	0.204	0.156

表 31: NIST Individual SCORE 実験データ

NISTScore(実験データ)	N=1	N=2	N=3	N=4
統計翻訳 (50)	2.513	0.172	0.037	0
ルールベース翻訳 (50)	3.129	0.304	0.057	0.021

訳でもそんなに高い翻訳結果が得られなかった。しかし、ルールベース機械翻訳は統計翻訳に比べると精度が高いことが分かった。

7　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

表 32: NIST Cumulative SCORE 実験データ

NISTScore(実験データ)	N=1	N=2	N=3	N=4
統計翻訳 (50)	2.513	2.684	2.722	2.722
ルールベース翻訳 (50)	3.129	3.432	3.511	3.511

表 33: 実験結果

翻訳種類	正しい文 (全テスト文)	パーセント
統計翻訳	7(50)	14%
置換翻訳	19(50)	38%
ルールベース翻訳	22(50)	44%

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。

8 感謝

本研究のために多大な御尽力を頂き, 日頃から熱心な御指導を賜った名古屋工業大学の松尾啓志教授, 津邑公暁准教授, 齋藤彰一准教授, 松井俊浩准教授、名古屋大学の外山勝彦准教授、小川康弘助教に深く感謝致します. また, 本研究の際に多くの助言, 協力をして頂いた松尾 津邑研究室、齋藤研究室ならびに名古屋大学の外山研究室の皆様に深く感謝致します.

参考文献

[1] http://ja.wikipedia.org/wiki/日本語

[2] http://en.wikipedia.org/wiki/Uyghurlanguage

[3] http://ja.wikipedia.org/wiki/膠着語

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

[5] 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).

[6] 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.

[7] 小川泰弘, 福田ムフタル, 外山勝彦,”日本語ーウイグル語翻訳掲示板システム,” 言語処理学会第 15 回年次大会講演論文集, pp.212-215, 鳥取大学, Mar. (2009).

[8] ムフタル・マフスット、外山勝彦、稲垣康善 ”日本語ーウイグル語機械翻訳における助動詞のパラメータ化による処理”, 電子情報通信学会, 信学技報,NLC94-13(1994-07)

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

[10] 志村賢治. ”自然言語処理の基礎 ”、サイエンス社、2005 年 4 月 10 日、初版第 3 刷発行

[11] 宮平 知博, 田添 英一, 武田 浩一, 渡辺 日出雄, 神山 淑朗. ”インターネット機械翻訳の世界 ”, 毎日コミュニケーションズ

[12] 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

[13] Philipp Koehn, Franz J. Och, and Daniel Marcu. ”Statistical phrase-based translation”.In Marti Hearst and Mari Ostendorf, editors, HLT-NAACL 2003: Main Proceedings,pp. 127133, Edmonton, Alberta, Canada, May 27 - June 1 2003. Association for ***Computational Linguistics***.

[14] Franz Josef Och and Hermann Ney. ”The alignment template approach to statistical machine translation.” In ***Computational Linguistics***, Vol. 30, pp. 417449, 2004.

[15] Franz Josef Och and Hermann Ney. ”A systematic comparison of various statistical alignment models.” In Computational Linguistics, Vol. 29, pp. 1951, 2003.

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

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

[18] 村上仁, 鏡味良太, 徳久雅人, 池原悟.”統計翻訳における人手で作成された大規模フレーズテーブルの効果”,Journal of natural language processing 17(4), 155-175, 2010-07-30 , 言語処理学会

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.”BLEU: a Method for Automatic Evaluation of Machine Translation”.Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[20] George Doddington. ”Automatic evaluation of machine translation quality using n-gram co-occurrence statistics”. In Proceedings of the HLT Conference, 2002.

A　付録　翻訳実例と BLEU スコア

以下のテーブルで統計翻訳とルールベース翻訳を行った時の各文 (50 のテスト文) に対しての実際翻訳結果である。これらの翻訳結果に関して人手で評価を行った。ここで記号 ◯ は翻訳された文は正しいこと、記号 △ は翻訳された文は部分的正しいこと、記号 × は翻訳され文は正しくないことを表す。表 35 で各 50 文に対して一つ一つ BLEU スコアを求めた時の結果である。

表 34: 翻訳実例

日本語	晩 ご飯 は とても おいしかっ た
統計翻訳	kechlik ご飯 bekmu おいしかっ bashlidi ×
ルールベース翻訳	axsham tamaq bekmu temlik △
日本語	ただ そこ に どう でも よく ない 事 が 一つ あっ た の です 。
統計翻訳	emma , u yerde qandaq emma daim yoq 事 bolsa bir idi . △
ルールベース翻訳	peqetla shuyer gha null likin null may ish * null bar di ning dur . △
日本語	声 は まだ 自分 の 声 の よう に は 聞こえ ない
統計翻訳	un bolsa tehi ozining un kuni digudek bolsa 聞こえ yoq ×
ルールベース翻訳	awaz * texi uz ning awaz ning dek gha * anglan may ×
日本語	この 子 は 何 か 大事 な もの を 持っ て いる 。
統計翻訳	zereklikige bolsa ayaligha 大事 nersilerni ichge alidu . ×
ルールベース翻訳	bu bala * nime mu muhim * nerse ni eliwal p watidu . △
日本語	父 の 病気 は 幸い 現状 維持 の まま だ 。
統計翻訳	父 ning keseller asasi 幸い exwali 維持 ning まま . ×
ルールベース翻訳	ataning kesel xudagha shukuri hazirqi ehwal saqlaqliqning petidi . △
日本語	奥さん の この 態度 が 自然 私 の 気分 に 影響 し て 来 まし た 。
統計翻訳	ayali bu 態度 bolsa tebi minig 気分 bilen 影響 qilip keldim . ×
ルールベース翻訳	xutunning muamile tebiet menning rohi haletgha teser yetkuz qilp kelidi . △
日本語	私 は その 晩 先生 の 宿 を 尋ね た
統計翻訳	men shu kechlik muellim ning 宿 ni 尋ね bashlidi ×
ルールベース翻訳	men axsham ustazning yataqni ziyaret qildim ○
日本語	私 は 歌 が 上手 です 。
統計翻訳	men naxshigha usta . ○
ルールベース翻訳	men naxsha ustadur . △
日本語	私 の 名前 は 田中 です 、あなた は 誰 です か ？
統計翻訳	men ismi 田中 iken 、あなた bolsa 誰 iken ？ △
ルールベース翻訳	menning isim tanaka(i)men , siz kim ？ ○
日本語	私 は 日本語 を 話せ ませ ん 。
統計翻訳	men yapon tilini 話せ bolmaydu . △
ルールベース翻訳	men yapontilini suzliyemay(i)men . ○
日本語	私 の 名前 は 山本 です 、 日本人 です 。
統計翻訳	men ismi yamamoto iken , yapunluq . △
ルールベース翻訳	menning isim yamamoto(i)men , yapunluqdur . △
日本語	10 月 から 3月 まで 雨 が 多い です 。
統計翻訳	10-aydin 3-ayghiche yamghur kup iken . ○
ルールベース翻訳	10 ay din uch ay ghiche yamghur * kup dur . △
日本語	日本 の 夏 は とても 熱い です 。
統計翻訳	yapunda yazda bek issiq . ○
ルールベース翻訳	yapunyening yaz bekmu issiq . ○
日本語	私 は たくさん の 歴史 関係 の 書籍 を 読み まし た 。
統計翻訳	men kup 歴史 munasiweti ning 書籍 ni 読み iken . △
ルールベース翻訳	men jiqning tarix munasewetning kitaplarni oqudi . △

日本語　あなた は 買い物 を どこ で 買い まし た か ？
統計翻訳　sizchu 買い物 bir yaqqa din setiwalghan ？　×
ルールベース翻訳　siz * nerse setiwalmaq ni qeyer de setiwal * di mu ？　×

日本語　彼女 は アメリカ に 留学 し た こと が あり ます 。
統計翻訳　emdi amirka bilen 留学 shagirtliqqa berdi .　×
ルールベース翻訳　qiz amirkagha bilim ashurush di ish bardu .　△

日本語　アメリカ 大統領 と 日本 の 総理 大臣 が 環境 問題 について 議論 し た 。
統計翻訳　amirka 大統領 bilen yapunning 総理 大臣 bolsa muhit mesile ning つい qoyup
議論 shagirtliqqa berdi .　×
ルールベース翻訳　amirka prizentibilen yapunyening ichki ishlar wezir muhit mesile
ghanispiten muzakire qildi .　△

日本語　明日 は 富士山 へ 行き ます 。
統計翻訳　ete bolsa 富士山 barsa barimen .　△
ルールベース翻訳　ata fujiteghigha bardu .　◯

日本語　先生 は 私 の この 問い に 答えよ う と は し なかっ た 。
統計翻訳　muellim bolsa minig bu 問い bilen 答えよ petinalmidi dep ish .　×
ルールベース翻訳　ustaz menning soalgha jawap ber qildi .　△

日本語　私 が 奥さん と 話し て いる 間 に 、 問題 が 自然 先生 の 事 から そこ へ 落ち て 来 た 。
統計翻訳　men bolsa ayali we suz bolup , mesile bolsa tebi muellim ning 事 bolghachqa
u barsa ghulap chushup kelgen .　×
ルールベース翻訳　men xutunbilen suzlip watidu chaghdagha , mesile tebiet ustazning
ishdin shuyergha chushp kelidi .　×

日本語　明日雨が降るそうです。
統計翻訳　ete yamghur kup 降る shundaq .　×
ルールベース翻訳　ata yamghur yaghidighandek turdu.　◯

日本語　私 自身 すでに そう だ と 告白 し て い た 。
統計翻訳　men 自身 alliqachan shundaq dep 告白 qilip qalghan idi .　△
ルールベース翻訳　men uzi alliqachan null di bilen izhar qilish qil p wati di .　×

日本語　父 の 病気 は 思っ た ほど 悪く は なかっ た 。
統計翻訳　父 ning keseller asasi oylighan dak oyghanmaq .　×
ルールベース翻訳　ataning kesel oylighandi nachar emas .　△

日本語　私 は 心 の うち で 父 と 先生 と を 比較 し て 見 た 。
統計翻訳　men bilen ning uyde , atam , muellim bilen ni 比較 qilip korup qaldi .　×
ルールベース翻訳　men yurekning ichide atabilen ustazbilen selishturup qilp kordim .　△

日本語　北京を訪れ帰郷した時のことである。
統計翻訳　beijinggha qilinghan seperdin qaytip keliwatqanda bolghan ish ibaret .　△
ルールベース翻訳　beijingni kurup kel yurtqa qaytqandi chagning ish bar .　△

日本語　今 まで 楽天的 に 傾い て い た 私 は 急 に 不安 に なっ た 。
統計翻訳　ayallarmu hazirghiche 楽天的 bilen 傾い qara men shundaq bilen ghelbisi bolup qaldi .　×
ルールベース翻訳　hazirghiche xoashal xoramliqgha erghip wetidi men birdinlagha xatirjemsizgha boldim .　△

日本語　先生 の 口元 に は 微笑 の 影 が 見え た 。
統計翻訳　muellim ning 口元 da 微笑 ning 影 見え bolidu .　×
ルールベース翻訳　ustazning eghizigha kulumsirigenning kulengge kurundi .　△

日本語　どれ くらい それ が 続い た の か わから ない 。
統計翻訳　どれ くらい shuyeri 続い bolghan ? わから bolmaydu . ×
ルールベース翻訳　qaysi chilik u dawamlashdining mu bilmay . ×

日本語　身体 は ひどく 消耗 し て い た
統計翻訳　身体 bolsa ひどく 消耗 hokum ×
ルールベース翻訳　beden bek serip qilp wetidi △

日本語　彼はどこから来た？
統計翻訳　u yaqqa yerlerde yaghuz ×
ルールベース翻訳　u qeyerdin keldi ? ○

日本語　更に 多く の 言葉 を 彼 は 求め て い た 。
統計翻訳　yenimu 多く ning 言葉 ni u 求め bolup qalghan idi . ×
ルールベース翻訳　yenimu jiqning tilni u kozlep wetidi . △

日本語　物語 を 語り たい という 意志 は たしかに ある 。
統計翻訳　物語 ni 語り dighen dighen 意志 bolsa たしかに ibaret . ×
ルールベース翻訳　hikayeni sozlep berghum bar digen irade heqiqetan bar . ○

日本語　文章 を 書く という 作業 に対して きわめて 謙虚 でも ある
統計翻訳　文章 ni 書く dighen 作業 に対して きわめて 謙虚 emma bir ×
ルールベース翻訳　maqale ni yaz digen meshxulat ghanispiten shuqeder kemter likin bar △

日本語　明日家へ帰らなければなりません。
統計翻訳　ete oyi barsa 帰ら qilmisa bolmaydu . ×
ルールベース翻訳　ata uygha qaytmisa bolimaydu . △

日本語　山本 という 男 に は どこ か はかり 知れ ない ところ が あっ た 。
統計翻訳　yamamoto dighen er da yaqqa ? はかり 知れ yoq utush bolghan . ×
ルールベース翻訳　yamamoto digen ogul bala gha * qeyer mu olchighili bil may jay * bar di . ×

日本語　彼 にとって は 偏見 も 真実 の 重要 な 要素 の ひとつ だっ た 。
統計翻訳　u nispiten 偏見 mu 真実 ning 重要 herhil 要素 ning ひとつ ishlaer iken . ×
ルールベース翻訳　u ghanispiten kemsitishmo heqiqetning muhim elmentining birdi . △

日本語　そう なろ う と 思え ば とことん 辛辣 に なる こと も でき た 。
統計翻訳　shundaq なろ petinalmidi dep oylisaq とことん 辛辣 bolghili bolidu . ×
ルールベース翻訳　shundaq bol oylisa adaqqiche neshterdekgha bolidu ishmo qilalaydi . ×

日本語　その 私生活 について 知る 人 は い ない 。
統計翻訳　shu 私生活 heqqide 知る ademler yoq . ×
ルールベース翻訳　shexsi turmush ghanispiten bildighan adem wetimay . △

日本語　ただ そう 言わ れれ ば 、 と 納得 できる ところ は あっ た
統計翻訳　emma , shundaq 言わ れれ , bilen 納得 できる utush bolsa bolghan ×
ルールベース翻訳　peqetla shundaq di p sa , bilen qanaet qilalay jay * bar di △

日本語　彼 は 遠い 所 から 走っ て 来 まし た 。
統計翻訳　u 遠い yaylarda bolghachqa yughurup keldim . ×
ルールベース翻訳　u yeraq urundin yugururp kelidi . ○

日本語　私 は 彼 が どの よう な 人間 な の か を 知り た がっ た 。
統計翻訳　men uning どの uhshap kishlerning herhil ? ni 知り がっ idi . ×
ルールベース翻訳　men * u * qaysi dek * insan * ning mu ni bil di null di . ×

日本語　奥さん は 私 を 静か な 人 、 大人しい 男 と 評し まし た 。
統計翻訳　ayalingiz men ni jimjit kishi , 大人しい er bilen 評し iken . ×
ルールベース翻訳　xutun menni jimhor adem , yowash ogul balabilen bahalandi . △

日本語　来週私は弟の結婚式に参加します。
統計翻訳　sowghatni keler hapte men 弟 ning toy bilen 参加 qilip qildighan . ×
ルールベース翻訳　kelar hepte men inining toy murasimigha qatinishidu . △

日本語　私 は 相 変ら ず 学校 へ 出席 し て い まし た 。
統計翻訳　men 相 変ら ず mektiptiki barsa 出席 qilip bolghan . △
ルールベース翻訳　men * burunqidek uzgur mey mektep gha yoqlimigha tuluq qatnishi qil p wati * di . △

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。
統計翻訳　hiroshimadin setiwalghan sowghatni keler hapte elip barimen . ○
ルールベース翻訳　hiroshimade temlik sowghatni setiwaldidin , kelar hepte eliwalp bardu . △

日本語　私 は 彼 に 向っ て 余計 な 仕事 を する の は 止せ と 言い まし た 。
統計翻訳　men uning bilen 向っ , 余計 herhil xizmet teylish bolsa 止せ didi . ×
ルールベース翻訳　men ugha qarap artuq ishni qil ning toxtatbilen didim . △

日本語　真相 を 知る こと は あなた を 傷つけ ます 。
統計翻訳　真相 ni bilghili bolsa siz ni 傷つけ . △
ルールベース翻訳　heqiqi ehwalni bildighan ish sizni jarahetlendu . △

日本語　仕事 は 終わっ た から 、 今 から 帰れる
統計翻訳　hizmet 終わっ bolghachqa , hazir bolghachqa 帰れる ×
ルールベース翻訳　ish * tughu di din , hazir din qaytiyalaydu △

日本語　与え られ た 才能 を できるだけ 大事 に 使う こと だ
統計翻訳　与え られ bolghan 才能 bir amal bar kupligen 大事 bilen istakanini ishletkili ×
ルールベース翻訳　ber aldi iqdidarni amalning beriche muhimgha ishlet ishdi △

日本語　夕方 の 空 に は まだ 雲 ひとつ 見え なかっ た 。
統計翻訳　夕方 ning asmandiki da tehi bulut ひとつ 見え . ×
ルールベース翻訳　kechqurunning asmangha texi bulut bir kurunmidi . △
